

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Πτυχιακή Εργασία

**Μοντέλα διάγνωσης και πρόβλεψης ρίσκου επιβίωσης για τον καρκίνο του
πνεύμονα**

Μαθιουδάκης Εμμανουήλ Π2017027

Επιβλέπων: Θεμιστοκλής Έξαρχος

31 Μαΐου 2024

Επιβλέπων

Θεμιστοκλής Έξαρχος, Αναπληρωτής Καθηγητής,

Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Θεμιστοκλής Έξαρχος, Αναπληρωτής Καθηγητής,

Ιόνιο Πανεπιστήμιο

Παναγιώτης Βλάμος, Καθηγητής,

Ιόνιο Πανεπιστήμιο

Αριστείδης Βραχάτης, Επίκουρος Καθηγητής,

Ιόνιο Πανεπιστήμιο

Περίληψη

Συνεχώς παρατηρούμε τις καινοτόμες εφαρμογές και τεχνολογίες που προκύπτουν από την τεχνητή νοημοσύνη και τους συσχετιζόμενους με αυτή τομείς, αλλά πολλές φορές δυσκολευόμαστε να διακρίνουμε την ουσία αλλά και την προσφορά τους στην καθημερινή ζωή. Ένα παράγωγο αυτών των τεχνολογιών αποτελεί και η μηχανική μάθηση, με την χρήση της οποίας επιτυγχάνεται η δυνατότητα να προσδώσουμε παραπάνω «ικανότητες» σε μια υπολογιστική μηχανή από αυτές που ήδη διαθέτει. Παράδειγμα των ικανοτήτων αυτών είναι η παραγωγή αποφάσεων από ένα πρόγραμμα, αφού πρώτα αυτό τροφοδοτηθεί με τα κατάλληλα δεδομένα. Εφαρμογές αυτού του παραδείγματος χρησιμοποιούνται πλέον σε πολλούς τομείς της κοινωνίας, από τις διάφορες επιχειρήσεις που προσπαθούν να μεγιστοποιήσουν τις πωλήσεις τους, μέχρι και τα νοσοκομεία που επιθυμούν πιο έγκαιρες και έγκυρες διαγνώσεις για ζητήματα που απειλούν άμεσα την ανθρώπινη ζωή. Οι εφαρμογές αυτές ορίζονται ως συστήματα υποστήριξης αποφάσεων και αποτελούν κάποια από τα χρησιμότερα εργαλεία στα χέρια εξειδικευμένου προσωπικού. Ειδικότερα στην περίπτωση ασθενειών όπως τον καρκίνο του πνεύμονα τον οποίο πραγματευόμαστε εμείς, τέτοιου είδους συστήματα μπορούν να αποτρέψουν θάνασμα αποτελέσματα με την εγκυρότητα τους αλλά και την ταχύτητα που προσφέρουν κατά την παραγωγή αποφάσεων. Σήμερα υπάρχουν πολλές υλοποιήσεις τέτοιων εφαρμογών, τόσο σε πρακτικό όσο και θεωρητικό επίπεδο, οι οποίες είναι ικανές για την διάγνωση του καρκίνου του πνεύμονα μέσω της αναγνώρισης εικόνων που παράγονται από αξονικούς και μαγνητικούς τομογράφους, ελάχιστες όμως από αυτές χρησιμοποιούν απλούστερα δεδομένα που είναι πιο εύκολα προσβάσιμα. Εν συνεχεία, όλες οι διαθέσιμες υλοποιήσεις ασχολούνται μόνο με τη διάγνωση της ασθένειας και δεν διευρύνουν την λειτουργικότητα τους. Για τους παραπάνω λόγους αποφασίσαμε να συνεισφέρουμε στην προϋπάρχουσα βιβλιογραφία, δημιουργώντας το δικό μας σύστημα υποστήριξης αποφάσεων το οποίο πραγματεύεται όχι μόνο τη διάγνωση, αλλά ταυτόχρονα και την αξιολόγηση του ρίσκου επιβίωσης του ασθενή για παραπάνω από έναν χρόνο έπειτα από χειρουργική επέμβαση στην περίπτωση που βρέθηκε θετική διάγνωση. Αυτό επιτεύχθηκε έπειτα από αναλυτική μελέτη της ήδη υπάρχουσας βιβλιογραφίας και με την εύρεση και επεξεργασία δεδομένων που στη συνέχεια τροφοδοτήσαμε σε αλγορίθμους μηχανικής μάθησης ώστε να παράξουμε μοντέλα ικανά για την παραγωγή έγκυρων αποφάσεων. Το σύστημα αυτό θα στελεχωθεί σε μια διαδικτυακή εφαρμογή την οποία δημιουργήσαμε από την αρχή, ώστε να προσφέρει εύκολη πρόσβαση τόσο σε χρήστες με ιατρικό παρασκήνιο όσο και σε αυτούς χωρίς οποιαδήποτε εξειδίκευση. Τον στόχο αυτό τελικά τον πετύχαμε έπειτα από αρκετό πειραματισμό, με τη δημιουργία δύο μοντέλων με τη χρήση του αλγορίθμου Extreme Gradient Boosting στη γλώσσα Python, ο οποίος επέφερε μέση ισορροπημένη ακρίβεια ίση με 87% για το μοντέλο διάγνωσης του καρκίνου και 66% για το μοντέλο αξιολόγησης του ρίσκου επιβίωσης. Τέλος, δημιουργήσαμε την διαδικτυακή εφαρμογή που θα στελεχώνει τα μοντέλα αυτά με τη χρήση εργαλείων όπως η HyperText Markup Language, η Cascading Style Sheets, η Javascript και η Flask, έχοντας έτσι ολοκληρώσει τον στόχο μας για ένα σύστημα υποστήριξης αποφάσεων ικανό αλλά παράλληλα και εύκολο στη χρήση που θα προσφέρει έναν νέο απλούστερο τρόπο προσέγγισης στην ήδη υπάρχουσα βιβλιογραφία.

Πρόλογος και Ευχαριστίες

Στη σύγχρονη περίοδο που βιώνουμε, παρατηρούμε συνεχώς τις τεχνολογικές εξελίξεις που προκύπτουν, οι οποίες προσφέρουν σημαντική βοήθεια τόσο σε ατομικό αλλά και σε κοινωνικό επίπεδο. Η άνοδος της επιστήμης της πληροφορίας και της πληροφορικής που συνδέεται άμεσα με αυτή την τεχνολογική πρόοδο έχει οδηγήσει στην διείσδυση της στους περισσότερους επαγγελματικούς τομείς, προσφέροντας τους αμέτρητα πλεονεκτήματα. Αδιαμφισβήτητο είναι πως αυτή η αλληλεπίδραση θα συνεχιστεί για πολλές ακόμη δεκαετίες, χρίζοντας έτσι αναγκαία την αμοιβαία και ολοκληρωμένη συνεργασία μεταξύ των ειδικών του εκάστοτε κλάδου ώστε όλοι να επωφεληθούν από τα προνόμια που προκύπτουν.

Ένας από τους επηρεασμένους τομείς είναι αυτός της ιατρικής, ο οποίος μεριμνά για ένα από τα βασικότερα αγαθά της ζωής, την υγεία. Η επίτευξη αυτής όμως είναι μία περίπλοκη και δαπανηρή υπόθεση, καθώς οι γιατροί και τα νοσοκομεία καλούνται καθημερινά να αντιμετωπίσουν μαζικά πολλαπλές περιπτώσεις ασθενειών που συχνά απαιτούν χρόνο για την διάγνωση, χρόνο που για κάποιες περιπτώσεις είναι πολύτιμος. Σε αυτό το πρόβλημα προσπαθεί να βοηθήσει ο τομέας της πληροφορικής, ο οποίος τα τελευταία χρόνια παρέχει μέσω της καινοτόμας τεχνογνωσίας που διαθέτει πολλαπλές εφαρμογές και πληροφοριακά συστήματα που προσπαθούν να βελτιώσουν και να συντομεύσουν κάποιες από τις διεργασίες της ιατρικής που απαιτούν εγκυρότητα και ακρίβεια. Κανένα σύστημα, βέβαια, δεν θα καταφέρει να συναγωνιστεί με την εμπειρία των ειδικών, για αυτό και βασικός γνώμονας κατά την σχεδίαση και δημιουργία των εφαρμογών είναι η διευκόλυνση αυτών.

Με αυτή τη πτυχιακή εργασία στοχεύουμε στο να συνεισφέρουμε τόσο θεωρητικά αλλά και πρακτικά στο κομμάτι της πρόληψης, παρέχοντας στο κοινό ένα σύστημα που χρησιμοποιεί καινοτόμες τεχνολογίες και όγκο δεδομένων για την άμεση και ακριβή δημιουργία αποφάσεων, οι οποίες μπορούν να διευκολύνουν τον εκάστοτε επαγγελματία στην λήψη της δικής του. Αυτό είναι ιδιαίτερα χρήσιμο για περιπτώσεις όπως ο καρκίνος του πνεύμονα, με τον οποίο θα ασχοληθούμε, καθώς η διάγνωση του είναι συνήθως περίπλοκη και χρονοβόρα, παρότι είναι επιτακτική ανάγκη να προληφθεί έγκαιρα.

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή Θεμιστοκλή Έξαρχο που μου έδωσε την ευκαιρία αλλά και όλα τα εφόδια που χρειάστηκα για την διεκπεραίωση αυτής της πτυχιακής εργασίας που διασφαλίζει την ολοκλήρωση των σπουδών μου και την επίσημη είσοδο μου στον κλάδο της πληροφορίας και πληροφορικής.

Κέρκυρα, 31 Μαΐου 2024

Μαθιουδάκης Εμμανουήλ

Περιεχόμενα

A' Μελέτη Βιβλιογραφίας	1
A'.1 Εισαγωγή.....	1
A'.2 Ο καρκίνος του πνεύμονα στην Ελλάδα.....	2
A'.3 Ο ρόλος των κλινικών συστημάτων υποστήριξης αποφάσεων.....	2
A'.4 Βιβλιογραφική έρευνα.....	3
A'.5 Σύγκριση ερευνών.....	13
A'.6 Συμπεράσματα.....	16
B' Περιγραφή Δεδομένων	17
B'.1 Δεδομένα διάγνωσης.....	17
B'.2 Δεδομένα αξιολόγησης.....	17
Γ' Προεπεξεργασία Δεδομένων	19
Δ' Εκπαίδευση και αποτελέσματα	21
Δ'.1 Επιλογή αλγορίθμου εκπαίδευσης.....	21
Δ'.2 Μετρικές μοντέλου διάγνωσης.....	22
Δ'.3 Μετρικές μοντέλου αξιολόγησης.....	28
Ε' Υλοποίηση Διαδικτυακής Εφαρμογής	33
Ε'.1 Κατασκευή μοντέλων λήψης αποφάσεων.....	34
Ε'.2 Δημιουργία επιπέδου διεπαφής (Front-end).....	35
Ε'.3 Ανάπτυξη back-end επιπέδου.....	36
Ε'.4 Παρουσίαση διαδικτυακής εφαρμογής.....	37
ΣΤ' Συμπεράσματα	48
Παράρτημα A'	49
Βιβλιογραφία	76

Κεφάλαιο Α΄

Μελέτη Βιβλιογραφίας

Α΄.1 Εισαγωγή

Η πτυχιακή εργασία πραγματεύεται τα μοντέλα διάγνωσης του καρκίνου του πνεύμονα και της πρόβλεψης του ρίσκου επιβίωσης αυτού. Όπως θα δούμε παρακάτω, αυτά δημιουργήθηκαν με την χρήση της μηχανικής μάθησης, ενός υποπεδίου της τεχνητής νοημοσύνης που με βάση τα δεδομένα που δέχεται μπορεί να «μάθει» σε μια υπολογιστική μηχανή περίπλοκες ενέργειες που διαφορετικά δεν ήταν εφικτές όπως η αναγνώριση αντικειμένων μέσα από εικόνες, η δημιουργία αποφάσεων και ούτω καθεξής. Η προσπάθεια ανάπτυξης του συγκεκριμένου λογισμικού που συνδέεται έμμεσα με την αντιμετώπιση μιας ασθένειας την εντάσσει αυτόματα στο επιστημονικό πεδίο της βιοπληροφορικής, ενός σημαντικού κλάδου του τομέα της πληροφορικής που έχει καταγράψει σημαντική πρόοδο τα τελευταία χρόνια με τις καινοτόμες εφαρμογές που προσφέρει.

Βασικός γνώμονας για την διεκπεραίωση κάθε πτυχιακής εργασίας είναι η μελέτη και ανάλυση της βιβλιογραφίας που υπάρχει διαθέσιμη ώστε να δημιουργηθεί μια σφαιρική εικόνα για το αντικείμενο που πραγματεύεται αυτή. Συνεπώς στο κεφάλαιο αυτό θα πραγματοποιηθεί η ανάλυση πολλαπλών μελετών που χρησιμοποιούν παρόμοια μεθοδολογία με στόχο την σύγκριση αυτών με τα αποτελέσματα που προέκυψαν από την δική μας υλοποίηση. Στη συνέχεια θα ακολουθήσει ένα ακόμη κεφάλαιο στο οποίο θα καταγραφεί λεπτομερώς η μεθοδολογία της ανάπτυξης των μοντέλων διάγνωσης, τα εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη, καθώς και η ανάλυση των δεδομένων που χρησιμοποιήθηκαν και των αποτελεσμάτων που προέκυψαν. Παρότι υπάρχουν, όπως θα δούμε και στην έρευνα παρακάτω, πολλαπλές μελέτες που υλοποιούν παρόμοια μοντέλα με τον ίδιο αλλά και διαφορετικούς τύπους δεδομένων, τόσο για την διάγνωση του καρκίνου του πνεύμονα όσο και για την πρόβλεψη του ρίσκου επιβίωσης, καμία από αυτές δεν χρησιμοποιεί τον συνδυασμό των δύο σε μια διαδικτυακή εφαρμογή όπως αυτή που δημιουργήθηκε στα πλαίσια της συγκεκριμένης πτυχιακής εργασίας. Η λειτουργία αυτή παρέχει έτσι μια νέα δυνατότητα στον εκάστοτε χρήστη που δεν υπήρχε μέχρι στιγμής, κάτι που καθιστά την εφαρμογή καινοτόμα, καλύπτοντας ένα κενό που υπήρχε στην βιβλιογραφία.

Στόχος της εργασίας είναι να προσφέρει κάτι νέο στο πεδίο της βιοπληροφορικής και προπάντως στην κοινωνία, κάτι που θα μπορέσει να αποτελέσει μελλοντικά την βάση για περαιτέρω έρευνα και βελτίωση από το επιστημονικό κοινό και ίσως να οδηγήσει στην βελτίωση της ποιότητας ζωής ορισμένων συνανθρώπων μας. Με αυτό κατά νου, αποφασίσαμε να ασχοληθούμε με τον καρκίνο του πνεύμονα, μια ασθένεια δύσκολη στην αντιμετώπιση που αν δεν διαγνωστεί εγκαίρως μπορεί να επιφέρει μη αναστρέψιμες συνέπειες, όπως αποδεικνύει και η θέση του στην λίστα με τις κυριότερες αιτίες θανάτου παγκοσμίως. Λόγω της σοβαρότητας του βέβαια, έχει οδηγηθεί το επιστημονικό κοινό

στην μελέτη παραγόντων που συνδέονται με την ασθένεια, τους οποίους παρέχουν δωρεάν στο διαδίκτυο για περαιτέρω διερεύνηση. Έτσι λοιπόν, χρησιμοποιήσαμε και εμείς ένα κομμάτι αυτού του τεράστιου όγκου δεδομένων, το οποίο έπειτα από επεξεργασία αποτέλεσε την βάση για την δημιουργία των μοντέλων διάγνωσης και πρόβλεψης που αποτελούν τον πυρήνα της διαδικτυακής εφαρμογής που δημιουργήσαμε, η οποία λειτουργεί ως ένα σύστημα υποστήριξης αποφάσεων. Το σύστημα υποστήριξης (back-end) που περιλαμβάνει τα μοντέλα δημιουργήθηκε με την χρήση της γλώσσας προγραμματισμού Python, η οποία αποτελεί την κύρια προτίμηση των μηχανικών λογισμικού που χρησιμοποιεί την μηχανική μάθηση (Machine Learning), ενώ ο αλγόριθμος που χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων είναι ο Extreme Gradient Boost (XGBoost) ο οποίος έπειτα από σύγκριση με άλλους παρείχε καλύτερα αποτελέσματα. Για το επίπεδο διεπαφής χρήστη (front-end) δημιουργήθηκε μια εφαρμογή με την χρήση των γλωσσών προγραμματισμού Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) αλλά και της Javascript. Για την επίτευξη της επικοινωνίας μεταξύ front-end και back-end χρησιμοποιήσαμε το Flask, ένα μικροπλαίσιο διαδικτύου που είναι γραμμένο σε Python. Παραπάνω ανάλυση των τεχνικών λεπτομερειών θα συναντήσουμε στο δεύτερο κεφάλαιο της πτυχιακής εργασίας.

Α'.2 Ο καρκίνος του πνεύμονα στην Ελλάδα

Ως καρκίνος καθορίζεται μια ασθένεια που μπορεί να εμφανιστεί σε οποιοδήποτε όργανο του ανθρώπινου σώματος, όπου ασυνήθιστα κύτταρα αναπτύσσονται ραγδαία και ξεπερνούν το προβλεπόμενο μέγεθος τους, αποκτώντας έτσι την ικανότητα να «επιτίθενται» σε άλλα μέρη ή όργανα του σώματος. Στον καρκίνο του πνεύμονα υπάρχουν δύο διαφορετικοί τύποι, ο μη μικροκυτταρικός που επικρατεί στις περιπτώσεις εμφάνισης και ο μικροκυτταρικός ο οποίος μετατάσσεται πιο γρήγορα και η αντιμετώπιση τους εξαρτάται από το στάδιο, την εξάπλωση αλλά και την μορφή. Το συγκεκριμένο είδος καρκίνου παρουσιάζεται συχνότερα στις ηλικιακές ομάδες άνω των 40 ετών, με κύριο αίτιο πρόκλησης να είναι το κάπνισμα. [1] Ο καρκίνος του πνεύμονα στην Ελλάδα εμφανίζεται συχνότερα στους άνδρες παρά στις γυναίκες, με τις περιπτώσεις διάγνωσης να ανέρχονται στις 6781 για τους άνδρες το έτος 2020, τοποθετώντας την Ελλάδα στην όγδοη υψηλότερη θέση στον κόσμο σε ποσοστά εμφάνισης. [2] Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, για το έτος 2018 το σύνολο περιπτώσεων των διαφόρων μορφών καρκίνου στην Ελλάδα ήταν 67401, από τις οποίες ο καρκίνος του πνεύμονα είχε το μεγαλύτερο ποσοστό τόσο εμφάνισης όσο και θανάτων, με 14.8% και 25.1% αντίστοιχα. Αυτό αντιστοιχεί σε 9964 περιπτώσεις συνολικά. Εκτιμάται πως μέχρι το έτος 2040, με βάση την αύξηση που παρατηρείται από το παρελθόν μέχρι και σήμερα, οι συνολικές περιπτώσεις του καρκίνου του πνεύμονα θα ανέρθουν στις 13030, καθιστώντας την ανάγκη για πρόληψη επιτακτική. [3]

Α'.3 Ο ρόλος των κλινικών συστημάτων υποστήριξης αποφάσεων

Τη σημερινή εποχή ο κόσμος βρίσκεται καθημερινά αντιμέτωπος με προβλήματα που πρέπει να επιλυθούν. Πολλά από τα προβλήματα αυτά μπορεί να μην είναι τόσο σημαντικά και να μην έχουν μεγάλο αντίκτυπο πάνω του. Υπάρχει όμως μια κατηγορία προβλημάτων που απειλεί άμεσα την ζωή του ανθρώπου και πρέπει να προβλεφθεί έγκαιρα και να αντιμετωπιστεί από τους ειδικούς. Αυτά είναι τα ιατρικά προβλήματα που αντιμετωπίζονται από ειδικούς του τομέα ιατρούς. Οι ιατροί από την μεριά τους έρχονται καθημερινά αντιμέτωποι με αυτά τα προβλήματα και γνωρίζουν πως πρέπει να ληφθούν σωστές αποφάσεις για την επίλυσή τους. Για να γίνει όμως αυτό, οι ίδιοι πρέπει να επεξεργαστούν τεράστιους όγκους πληροφορίας για κάθε ασθενή, μεγάλο μέρος των οποίων μπορεί να μην συμβάλει καθόλου στην διαμόρφωση μιας εικόνας του τι συμβαίνει και τι «μέτρα» πρέπει να ληφθούν. Επίσης, μπορεί ένα μέρος της πληροφορίας να είναι και ανακριβές ή ακόμα και ψευδές. Η επεξεργασία αυτή πέρα από το ότι πρέπει να γίνει σωστά ώστε τα αποτελέσματα που προκύπτουν να είναι καλής ποιότητας, πρέπει παράλληλα να γίνει και σε μικρό χρονικό διάστημα καθώς ο χρόνος παίζει σημαντικό ρόλο στην αποφυγή άλλων προβλημάτων που μπορεί να εμφανισθούν κατά την διάρκεια. Οι ιατροί, λοιπόν, που καθημερινά έρχονται αντιμέτωποι πολλούς ασθενείς για τους οποίους πρέπει να ακολουθήσουν την ίδια διαδικασία έχει παρατηρηθεί πως χρειάζονται ένα σύστημα που θα τους παρέχει την ταχύτητα και ποιότητα που χρειάζονται ώστε να καταλήξουν στις κατάλληλες αποφάσεις.

Αυτό το εξασφαλίζουν τα κλινικά συστήματα υποστήριξης αποφάσεων (Clinical Decision Support Systems ή CDSS) τα οποία είναι κατασκευασμένα έτσι ώστε να επεξεργάζονται σε ελάχιστο χρόνο με μεγάλη ακρίβεια την πληροφορία που έχει εισαχθεί και να προτείνουν στους χειριστές τους αποφάσεις που θα φέρουν τα βέλτιστα αποτελέσματα. Με τη ραγδαία ανάπτυξη της τεχνολογίας και του λογισμικού που δημιουργείται, προκύπτουν μέρα με τη μέρα καινούργια καινοτόμα συστήματα αυτού του τύπου με αποδόσεις που μπορούν να κάνουν την διαφορά σε χιλιάδες περιπτώσεις ασθενών. Πολλές φορές ακόμα δεν χρειάζεται κάποια ειδική εξέταση για να τροφοδοτήσει το σύστημα με δεδομένα, καθώς πέρα από τα CDSS που χρησιμοποιούν ακτινογραφίες και μαγνητικές τομογραφίες ως δεδομένα, έχουν δημιουργηθεί και απλούστερα που δημιουργούν αποφάσεις μέσω κοινών συμπτωμάτων και συνηθειών που μπορεί να καταγράψει και ο ίδιος ο ασθενής για τον εαυτό του. Έτσι παρατηρούμε πως τα συστήματα υποστήριξης αποφάσεων γίνονται ολοένα και πιο προσβάσιμα, τόσο από εξειδικευμένο προσωπικό όσο και από τους ασθενείς. [4]

Α΄.4 Βιβλιογραφική έρευνα

Όπως αναφέρθηκε νωρίτερα, κάθε πτυχιακή εργασία πρέπει να περιέχει μια πλήρη βιβλιογραφική έρευνα που αφορά το αντικείμενο με το οποίο ασχολείται, έτσι ώστε να δημιουργήσει μια σφαιρική εικόνα στην αντίληψη του αναγνώστη και να συγκρίνει τα αποτελέσματα της με αυτά των άλλων επιστημόνων. Σε αυτή την ενότητα θα ασχοληθούμε με αυτό, αναλύοντας ένα μέρος των μελετών που υπάρχουν διαθέσιμες στο διαδίκτυο. Λόγω του τύπου δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων που δημιουργήθηκαν στα πλαίσια της εργασίας αυτής, τα οποία ήταν

δυναμικά με βάση συμπτώματα που έχει ο ασθενής αλλά και δεδομένα συμβολοσειράς (string), θα παρατεθούν μελέτες που χρησιμοποιούν παρόμοιους τύπους δεδομένων και μεθοδολογία.

A sustainable IoHT based computationally intelligent healthcare monitoring system for lung cancer risk detection:

Λόγω της ραγδαίας αύξησης των χρόνιων ασθενειών την σημερινή εποχή, παρατηρήθηκε από τους συγγραφείς της συγκεκριμένης έρευνας η ανάγκη για τη δημιουργία ενός μοντέλου διάγνωσης καρκίνου του πνεύμονα που θα είναι βιώσιμο και δεν θα αποτελεί βάρος για το περιβάλλον, με στόχο την ενσωμάτωση του στο Διαδίκτυο των Πραγμάτων Υγείας (Internet of Health Things ή IoHT), μιας καινοτόμας ιδέας που προσπαθεί με την ύπαρξη της να επιτρέψει την πιο άμεση επικοινωνία μεταξύ ασθενών και υπηρεσιών υγείας μέσω της συλλογής ιατρικών δεδομένων απομακρυσμένα ανά πάσα στιγμή. Για να δημιουργηθεί το μοντέλο διαλέχθηκαν αρχικά οι μεταβλητές που σχετίζονται περισσότερο με τον καρκίνο του πνεύμονα χρησιμοποιώντας τον αλγόριθμο Heuristic Greedy Best First Search (GBFS) και στην συνέχεια τα δεδομένα των μεταβλητών αυτών εισήχθησαν στον αλγόριθμο Random Forest ώστε αυτός να εκπαιδεύσει το μοντέλο στον διαχωρισμό των ασθενών που είχαν την ασθένεια και αυτών που δεν την απέκτησαν.

Παρότι το σύνολο των αρχικών δεδομένων αποτελούταν από μόνο 36 δείγματα, χρησιμοποιήθηκε τεχνική επαύξησης δεδομένων, η οποία δημιουργεί πολλαπλά ελαφρώς τροποποιημένα αντίτυπα, αυξάνοντας τον τελικό αριθμό σε 488 δείγματα. Οι κατηγορίες που περιέχει το σύνολο δεδομένων περιέχει κοινούς βιομετρικούς δείκτες, καθημερινές συνήθειες του ασθενή, απρόοπτα συμβάντα όπως απώλεια βάρους αλλά και γενετικές διαταραχές, τα περισσότερα σε μορφή ακέραιων αριθμών. Τα αποτελέσματα που προέκυψαν ήταν εντυπωσιακά, με την μέση τιμή της ακρίβειας (accuracy), της εξειδίκευσης (specificity), της ευαισθησίας (sensitivity) και του f-score να είναι 96.96%, 96.26%, 96.34% και 96.32% αντίστοιχα. Οι υψηλές τιμές των μετρικών αυτών καθιστούν το μοντέλο μια βιώσιμη επιλογή που θα μπορέσει να ανταπεξέλθει με μεγάλη ευστοχία στις σημερινές ανάγκες των ιατρικών υπηρεσιών. [5]

Early Detection of Lung Cancer Through Breath Analysis Using AdaBoost Ensemble Learning Method:

Στην μελέτη αυτή παρατίθενται οι λεπτομέρειες μιας πρώιμης έρευνας που έχει ως στόχο την έγκαιρη πρόβλεψη του καρκίνου του πνεύμονα με την χρήση βιοδεικτών που παρατηρούνται στην αναπνοή του ασθενή. Για την επίτευξη του στόχου αυτού οι ερευνητές δημιούργησαν έναν ηλεκτρονικό μετρητή που έχει την ικανότητα να καταγράφει μέσω των αισθητήρων του βιοδείκτες πτητικών οργάνικών ενώσεων από την μύτη του ασθενή κατά την εκπνοή. Στην πορεία χρησιμοποίησαν την συσκευή σε 25 ασθενείς, 10 από τους οποίους ήταν καρκινοπαθείς, με σκοπό τον διαχωρισμό των δειγμάτων αναπνοής που θα

χρησιμοποιούσαν για την εκπαίδευση του συστήματος. Συνεπώς, πέρα από κάποια κοινά χαρακτηριστικά των ασθενών όπως η ηλικία, το φύλο και το κάπνισμα, χρησιμοποιήθηκαν επίσης στην εκπαίδευση βιοδείκτες όπως η μεθανόλη, η αιθανόλη, το προπάνιο, το μεθάνιο, το διοξείδιο του άνθρακα και άλλα.

Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε η μέθοδος εκμάθησης AdaBoost Ensemble η οποία είναι μια μέθοδος εκμάθησης συνόλου και έχει δημιουργηθεί για να προσφέρει καλύτερες αποδόσεις σε ταξινομητές δυαδικών κατηγοριών. Παράλληλα εφαρμόστηκε και μια στατιστική μέθοδος μείωσης διαστάσεων στα δεδομένα, ενώ η αξιολόγηση έγινε με την χρήση 10-fold cross validation. Τα αποτελέσματα του μοντέλου ήταν ικανοποιητικά, με την μέση τιμή της ακρίβειας (accuracy) να είναι στο 76%, την εξειδίκευση (specificity) στο 80% και την ευαισθησία (sensitivity) στο 70%. Αξίζει να σημειωθεί πως χωρίς την μέθοδο μείωσης διαστάσεων το σύστημα έφερε χαμηλότερες αποδόσεις, με τα accuracy, sensitivity και specificity να βρίσκονται στο 72%, 70% και 73.3% αντίστοιχα. Τέλος, τονίζεται πως το σύστημα χρειάζεται περαιτέρω μελέτη και βελτίωση με περισσότερα δείγματα ώστε να είναι σε θέση να προβλέψει πιο εύστοχα τα πρώιμα στάδια του καρκίνου του πνεύμονα. [6]

Detection of lung cancer with electronic nose and logistic regression analysis:

Η εν λόγω έρευνα χρησιμοποιεί παρόμοια μεθοδολογία με αυτή που αναλύσαμε παραπάνω. Με την χρήση δηλαδή μιας ηλεκτρονικής συσκευής αναλύει δείγματα από τις ανάσες των ασθενών ώστε να ξεχωρίσει τους ασθενείς με καρκίνο του πνεύμονα από αυτούς που πάσχουν από άλλες ασθένειες των πνευμόνων. Ο λόγος που χρησιμοποιεί παρόμοια μεθοδολογία είναι επειδή έχει παρατηρηθεί σε άλλες μελέτες πως οι συγκεκριμένες συσκευές αποτελούν την βέλτιστη επιλογή για την αναγνώριση πρώιμων σταδίων του καρκίνου του πνεύμονα, χωρίς όμως να έχουν μελετηθεί πλήρως όλες οι πιθανές στατιστικές μέθοδοι ώστε να συγκριθούν τα αποτελέσματα και να επιλεγθεί η καλύτερη.

Σε αυτή την φάση λοιπόν χρησιμοποιείται η ανάλυση της λογιστικής παλινδρόμησης (Logistic Regression) πάνω σε 475 δείγματα ασθενών, 252 από τα οποία άνηκαν σε καρκινοπαθείς. Ανάμεσα στους εθελοντές που πήραν μέρος στην δειγματοληψία βρίσκονταν και άνθρωποι που έπασχαν και από πιο κοινές παθήσεις όπως πνευμονία και άσθμα, καθώς και άλλοι που ήταν υγιείς. Στη συνέχεια δημιουργήθηκαν δύο μοντέλα διάγνωσης, με το πρώτο να αφορά τους μη καπνιστές και το δεύτερο ασθενείς που κάπνιζαν. Και τα δύο μοντέλα επέφεραν πολύ υψηλά αποτελέσματα μετά την αξιολόγηση τους, με το πρώτο να επιτυγχάνει 96.2% sensitivity και 90.6% specificity, ενώ το δεύτερο 95.8% sensitivity και 92.3% specificity. Είναι λοιπόν σαφές το γεγονός πως η ανάλυση με Logistic Regression μπορεί να αποτελέσει μια ικανοποιητική επιλογή για μοντέλα που χρησιμοποιούν δείγματα αναπνοής ασθενών για να διαγνώσουν τα πρώιμα στάδια του καρκίνου του πνεύμονα. [7]

Development of a predictive model for early detection of lung cancer: SREAL-eLUNG study:

Η μελέτη αυτή αν και σύντομη, πραγματεύεται την δημιουργία εργαλείων με την χρήση μηχανικής μάθησης με απώτερο σκοπό την διάγνωση του καρκίνου του πνεύμονα, καθώς η διάγνωση του είναι μια περίπλοκη διαδικασία διότι η πάθηση συνήθως δεν παρουσιάζει ασυνήθιστα συμπτώματα κατά την αρχική της φάση. Χρησιμοποιήθηκαν δείγματα από 4332 πολίτες, 722 από τα οποία ανήκαν σε ασθενείς με καρκίνο του πνεύμονα, με κάποιες από τις μεταβλητές να αφορούν εργαστηριακές εξετάσεις, οικογενειακό ιστορικό με καρκίνο, κάπνισμα και άλλα. Για την επιλογή των κατηγοριών που θα χρησιμοποιούνταν κατά την εκπαίδευση εφαρμόστηκαν μέθοδοι επιλογής χαρακτηριστικών (feature selection methods) και αφού έγινε η επιλογή, το σύνολο των δεδομένων χωρίστηκε σε δεδομένα εκπαίδευσης και επαλήθευσης με ποσοστό 70% και 30% αντίστοιχα (train-test split).

Οι ερευνητές στην συνέχεια εκπαίδευσαν μοντέλα με τέσσερις διαφορετικούς αλγορίθμους, τον Logistic Regression, τον Decision Tree, τον Random Forest αλλά και τα νευρωνικά δίκτυα (Neural Network) και έγινε αξιολόγηση του καθενός με την χρήση 5-fold cross validation. Τα τελικά αποτελέσματα ήταν ικανοποιητικά, με τον αλγόριθμο Logistic Regression να επιτυγχάνει 77.9% sensitivity, 68.6% specificity και 0.79 AUC, τον Decision Tree 77.9% sensitivity, 68.1% specificity και 0.77 AUC, τον Random Forest 79.3% sensitivity, 68.3% specificity και 0.80 AUC και τα νευρωνικά δίκτυα 80% sensitivity, 68.3% specificity και 0.79 AUC. Τέλος, αναφέρεται πως ενώ τα μοντέλα θα μπορούσαν να μειώσουν το ρίσκο θνησιμότητας ασθενών έως και 20%, χρειάζονται ακόμη σημαντική βελτίωση για να ανταπεξέλθουν σε ένα πραγματικό σενάριο. [8]

Development of a risk prediction model for lung cancer: The Japan Public Health Center-based Prospective Study:

Με αφορμή τον αντίκτυπο που έχει το κάπνισμα στην άνοδο των περιπτώσεων καρκίνου του πνεύμονα τις τελευταίες δεκαετίες αποφάσισαν οι ερευνητές να δημιουργήσουν ένα μοντέλο πρόβλεψης του ρίσκου απόκτησης της πάθησης που δίνει μεγαλύτερη βαρύτητα σε λεπτομέρειες αυτής της συνήθειας, όπως την ποσότητα των τσιγάρων που καταναλώνονταν, την διάρκεια καπνίσματος και τον χρόνο που έχει περάσει από την οριστική διακοπή του. Για την εύρεση δεδομένων χρησιμοποιήθηκαν συνολικά 59161 δείγματα από μια μελέτη του δημοσίου κέντρου υγείας της Ιαπωνίας, από τα οποία διαλέχθηκαν συγκεκριμένες μεταβλητές που ήταν άμεσα συνδεδεμένες με την πάθηση, ενώ η δημιουργία του μοντέλου χωρίστηκε σε δύο μέρη. Αρχικά ξεκίνησε η διαδικασία επιλογής μεταβλητών με την χρήση του Cox proportional hazard regression τα αποτελέσματα της οποίας ελέγχθηκαν με την χρήση του Grambsch-Therneau test. Στη συνέχεια δημιουργήθηκε ένα ευέλικτο μοντέλο με την χρήση ενός linear predictor το οποίο αξιολογήθηκε με βάση 47501 δείγματα από μια άλλη έρευνα του δημοσίου κέντρου υγείας της Ιαπωνίας, επιφέροντας 0.772 μονάδες στο c-index. Οι ερευνητές συνεπώς κατάφεραν να δημιουργήσουν ένα ικανό μοντέλο που λαμβάνει υπόψη διάφορες λεπτομέρειες που αφορούν την

συνήθεια του καπνίσματος και όπως καταγράφουν έχει φανεί χρήσιμο στο να πείθει άτομα υψηλού ρίσκου εμφάνισης της πάθησης να ξεκινήσουν περαιτέρω διαδικασίες ελέγχου. [9]

A Prediction Model Based on Biomarkers and Clinical Characteristics for Detection of Lung Cancer in Pulmonary Nodules 1,2:

Η εν λόγω έρευνα έχει ως στόχο την δημιουργία ενός μοντέλου διάγνωσης του καρκίνου του πνεύμονα χρησιμοποιώντας πέρα από ραδιολογικές μεταβλητές και κάποιους βιοδείκτες που είναι άμεσα συνδεδεμένοι με την πάθηση. Κύριος λόγος της δημιουργίας του μοντέλου είναι το γεγονός πως ενώ οι αξονικές τομογραφίες μπορούν να μειώσουν την θνησιμότητα αναγνωρίζοντας την πάθηση, αυξάνουν παράλληλα τον αριθμό των ακαθόριστων πνευμονικών οζιδίων τα οποία μπορεί να είναι είτε καλοήγη είτε κακοήγη. Έτσι λοιπόν, συνδύασαν τα δεδομένα 137 ασθενών που ήταν επίσης καπνιστές και τα χρησιμοποίησαν μαζί με το multivariate logistic regression analysis για να δημιουργήσουν το επιθυμητό μοντέλο που θα μπορεί να ξεχωρίζει την καλοήθεια από την κακοήθεια.

Για την αξιολόγηση του μοντέλου χρησιμοποιήθηκε ένα σετ δεδομένων που αποτελούνταν από 111 δείγματα με πνευμονικά οζίδια. Τα αποτελέσματα που επιτεύχθηκαν ήταν αξιοσημείωτα, με το AUC να είναι ίσο με 0.91, το sensitivity 80.3% και το specificity 89.4%. Συνολικά, η έρευνα κατάφερε να παρουσιάσει και να δημιουργήσει έναν τρόπο διάγνωσης του καρκίνου του πνεύμονα που συμπεριλαμβάνει πέρα από τα συνηθισμένα δεδομένα και κάποια πιο εξειδικευμένα, αποδεικνύοντας πως ακόμα και κάποια ακαθόριστα πνευμονικά οζίδια μπορεί να παίξουν καθοριστικό ρόλο στην εμφάνιση του καρκίνου του πνεύμονα. [10]

Validating and Strengthen the Prediction Performance Using Machine Learning Models and Operational Research for Lung Cancer:

Με το παρόν άρθρο οι επιστήμονες προσπαθούν να αποδείξουν πως στην εποχή που ζούμε, όπου οι πρακτικές της τεχνητής νοημοσύνης χρησιμοποιούνται ολοένα και συχνότερα, υπάρχουν πιο ικανές μέθοδοι για την διάγνωση μιας δύσκολης ασθένειας όπως ο καρκίνος του πνεύμονα από τις εξετάσεις αίματος. Έτσι λοιπόν, στοχεύουν να δημιουργήσουν ένα μοντέλο διάγνωσης του καρκίνου του πνεύμονα με την χρήση της μηχανικής μάθησης, τροφοδοτώντας σε επτά διαφορετικούς αλγόριθμους ένα σετ δεδομένων με 309 δείγματα, καθένα από τα οποία έχει δεκαέξι μεταβλητές που περιγράφουν συμπτώματα, συνήθειες αλλά και γενικές πληροφορίες. Αξίζει να αναφέρουμε πως η συγκεκριμένη έρευνα χρησιμοποιεί το ίδιο σετ δεδομένων με αυτό που χρησιμοποιήθηκε για την δημιουργία του δικού μας μοντέλου διάγνωσης.

Οι αλγόριθμοι που χρησιμοποιούνται είναι οι IBk, AdaBoostML, LogitBoost, Random Forest, J48, Logistic Regression και Support Vector Machine. Κατά τους συγγραφείς η κύρια μετρική των αποτελεσμάτων που τους ενδιαφέρει είναι αυτή της ακρίβειας (accuracy), με τον αλγόριθμο Logistic Regression να πετυχαίνει τη μεγαλύτερη με 93.2% έπειτα από αξιολόγηση με 10-fold cross validation. Κάποια άλλα από τα λοιπά αποτελέσματα είναι: 0.899 precision και 0.903 recall για τον αλγόριθμο IBk, 0.926 precision και 0.929 recall για τον AdaBoostML, 0.901 precision και 0.909 recall για τον LogitBoost, 0.907 precision και 0.913 recall για τον Random Forest, 0.897 precision και 0.903 recall για τον J48, 0.929 precision και 0.932 recall για τον Logistic regression και τέλος 0.92 precision και 0.926 recall για τον αλγόριθμο Support Vector Machine. Συνοψίζοντας οι συγγραφείς τονίζουν πως υπάρχουν περιθώρια βελτίωσης της ακρίβειας του μοντέλου για το μέλλον με την χρήση μεθόδων όπως η επιλογή χαρακτηριστικών (feature selection), η εξισσορόπηση της τάξης (class balancing) αλλά και χρήση αλγορίθμων συνόλου (ensemble algorithms). [11]

Lung cancer prediction model using ensemble learning techniques and a systematic review analysis:

Με αφορμή τις μεγάλες διαστάσεις που έχει πάρει ο καρκίνος του πνεύμονα ως μια θανατηφόρα πάθηση αναζητούνται ολοένα και περισσότεροι τρόποι για την έγκαιρη διάγνωση του, η οποία στην συνέχεια θα παρέχει στον ασθενή περισσότερες πιθανότητες κατά την αντιμετώπιση. Ένας από τους τρόπους αυτούς είναι η χρήση της μηχανικής μάθησης για την δημιουργία συστημάτων που είναι ικανά να διαγνώσουν με ακρίβεια την πάθηση όσο ακόμη είναι νωρίς ώστε να προληφθεί η διόγκωση του προβλήματος. Σε αυτή την έρευνα λοιπόν έχουν ως στόχο οι μελετητές να χρησιμοποιήσουν αλγορίθμους συνόλου (ensemble algorithms) για τη δημιουργία ενός μοντέλου διάγνωσης, καθώς αυτοί έχει παρατηρηθεί πως επιφέρουν μεγαλύτερη ακρίβεια.

Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν τα ίδια δεδομένα που αναφέραμε και στην προηγούμενη έρευνα, αυτά δηλαδή που χρησιμοποιήσαμε και εμείς για την δημιουργία του ενός από τα μοντέλα. Αυτά περιγράφουν κυρίως συμπτώματα που μπορεί να βιώνει ο ασθενής, γενικές πληροφορίες αλλά και συνήθειες αυτού. Συνολικά υπήρχαν 309 δείγματα τα οποία δημιουργήθηκαν από την μέθοδο υπερδειγματοληψίας (oversampling) SMOTE και τροφοδοτήθηκαν σε τέσσερις αλγορίθμους, τους XGBoost, LightGBM, Bagging και AdaBoost. Τα διαφορετικά αυτά μοντέλα αξιολογήθηκαν με την χρήση του 10-fold cross validation και έπειτα από σύγκριση των αποτελεσμάτων ο αλγόριθμος XGBoost επέφερε τις καλύτερες μετρικές από τους τέσσερις, με το accuracy να βρίσκεται στο 94.42%, το precision στο 95.66% και το recall στο 94.46%. Αξίζει να παραθέσουμε για λόγους σύγκρισης και τις μετρικές που προέκυψαν από τους άλλους αλγορίθμους, με τον LightGBM να έχει 92.55% accuracy, 93.92% precision και 92.10 recall, τον AdaBoost 90.70% accuracy, 92.53% precision και 90.26% recall και τελευταίο τον Bagging με 89.76% accuracy, 91.88% precision και 89.35% recall. Τέλος, στο άρθρο αναφέρεται πως ενώ τα αποτελέσματα που προέκυψαν ήταν εξαιρετικά, θα

μπορούσαν να βελτιωθούν σημαντικά με την χρήση ενός μεγαλύτερου και πιο ισοροπημένου σετ δεδομένων. [12]

Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method:

Τα τελευταία χρόνια η βιβλιογραφία που αφορά την διάγνωση του καρκίνου του πνεύμονα μέσω μοντέλων μηχανικής μάθησης ασχολείται ως επί το πλείστον με μοντέλα αναγνώρισης εικόνας που χρησιμοποιούν δεδομένα από αξονικές και μαγνητικές τομογραφίες. Η έρευνα αυτή παρόλα αυτά έχει ως στόχο να αποδείξει πως μοντέλα που χρησιμοποιούν κειμενικά δεδομένα μπορούν να επιτύχουν εξίσου καλά αποτελέσματα στην διάγνωση της πάθησης χρησιμοποιώντας έναν συνδυασμό μοντέλων που χρησιμοποιούν τους K-Nearest Neighbours, Decision Tree και Neural Networks αλγορίθμους και μια μέθοδο συνόλου bootstrap aggregating ή bagging για την ενίσχυση της τελικής ακρίβειας. Στην ουσία, το σύστημα αυτό κατά τη λειτουργία του θα λαμβάνει τα δεδομένα, θα τα τροφοδοτεί στα εκπαιδευμένα μοντέλα και στη συνέχεια η πλειοψηφία των αποτελεσμάτων των μοντέλων θα παρουσιάζεται ως η τελική απόφαση.

Συνολικά χρησιμοποιήθηκαν 1000 δείγματα 23 μεταβλητές που σχετίζονταν με το ιστορικό του ασθενή, καθημερινές του συνήθειες, πληροφορίες που αφορούν τυχών συμπτώματα που μπορεί να έχει και άλλα. Αυτά αφού επεξεργάστηκαν για να αντιμετωπίσουν χαμένες τιμές, κανονικοποιήθηκαν με την χρήση του linear transformation αλγορίθμου και χωρίστηκαν σε δεδομένα εκπαίδευσης (80%) και επαλήθευσης (20%). Μετά την εκπαίδευση, το σύστημα αξιολογήθηκε με την χρήση πέντε διαφορετικών σετ δεδομένων, με τα αποτελέσματα να παραμένουν παρόμοια κάθε φορά. Αναλυτικότερα, το σύστημα συνολικά πέτυχε μια ακρίβεια 98%, ενώ τα μοντέλα όταν αξιολογήθηκαν αυτόνομα πέτυχαν 97%, 94% και 96% για τους Decision Tree, K-Nearest Neighbours και Neural Networks αντίστοιχα, κάτι που δείχνει πως το τελικό σύστημα αποτελεί έναν αξιόλογο τρόπο διάγνωσης που θα μπορούσε να ανταποκριθεί εξίσου καλά όπως τα συστήματα που χρησιμοποιούν αναγνώριση εικόνων. [13]

The Early Stage Lung Cancer Prognosis Prediction Model based on Support Vector Machine:

Λόγω της ραγδαίας αύξησης των ασθενών που έχουν διαγνωστεί με καρκίνο του πνεύμονα τις τελευταίες δεκαετίες έχει παρατηρηθεί η ανάγκη δημιουργίας ενός συστήματος που θα δίνει την δυνατότητα πρόβλεψης του χρόνου επιβίωσης του ασθενή. Αυτό επιδιώκουν να επιτύχουν οι συγγραφείς του άρθρου με την δημιουργία ενός μοντέλου που θα είναι ικανό να προβλέψει την επιβίωση του ασθενή για μια διάρκεια πέντε ετών με την χρήση του αλγόριθμου Support Vector

Machine (SVM) για την εκπαίδευση και του μοντέλου Cox-proportional hazard regression model (COX) για τον υπολογισμό.

Τα δείγματα που είχαν οι ερευνητές στην διάθεση τους ήταν συνολικά 174 με 11 κατηγορίες το καθένα, γεγονός που καθιστούσε δύσκολη την εκπαίδευση λόγω του μικρού όγκου. Για αυτό το λόγο χρησιμοποίησαν μια τεχνική υπερδειγματοληψίας της κατηγορίας με τα λιγότερα δεδομένα που ονομάζεται Borderline SMOTE (Borderline Synthetic Minority Over-Sampling Technique Algorithm). Αφού εκπαιδεύτηκε το μοντέλο με τον αλγόριθμο Support Vector Machine παρουσίασε accuracy ίσο με 85%, sensitivity 82% και specificity 86%, ενώ αυτό με τον COX πέτυχε 72% accuracy, 84% sensitivity και 10% specificity. Στα αποτελέσματα επίσης γίνεται αναφορά στις μετρικές που προέκυψαν πριν την χρήση της υπερδειγματοληψίας, οι οποίες φανερώνουν πως τα δεδομένα ήταν ανισόροπα, κάτι που οδήγησε τελικά και στην χρήση της SMOTE. Τέλος, για την περαιτέρω αξιολόγηση του μοντέλου δημιουργήθηκε ακόμη ένα που δημιουργεί προγνώσεις για δύο έτη αντί για πέντε, το οποίο παρουσίασε με τον SVM accuracy 89%, sensitivity 84% και specificity 92% ενώ με τον COX είχε accuracy 81%, sensitivity 12% και specificity 89%. Όπως παρατηρούμε, το δεύτερο μοντέλο που δημιουργήθηκε αρχικά για αξιολόγηση του πρώτου παρουσίασε καλύτερα αποτελέσματα από αυτό, με τον SVM να είναι ο βέλτιστος αλγόριθμος και στα δύο μοντέλα. [14]

Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest:

Ένα μεγάλο ζήτημα στις περιπτώσεις καρκίνου του πνεύμονα είναι πως δεν μπορεί να υπολογιστεί το ποσοστό επιτυχίας της αντιμετώπισης του, καθώς οι παραμέτροι που μπορεί να χρειάζονται για την δημιουργία αυτής της απόφασης μπορεί να είναι απρόβλεπτοι. Το γεγονός αυτό όμως δεν αποτρέπει τους συγγραφείς της μελέτης από το να δοκιμάσουν να δημιουργήσουν ένα μοντέλο που θα δώσει την ευκαιρία στους ασθενείς να κατανοήσουν την επιτυχία της θεραπείας τους υπολογίζοντας το ποσοστό επιβίωσης με την χρήση παραμέτρων όπως το στάδιο του καρκίνου, ο χρόνος θεραπείας που έχει επέλθει, το μέγεθος του όγκου, το φύλο και η ηλικία. Συνολικά χρησιμοποιήθηκαν 509 δείγματα περιπτώσεων, πάνω στα οποία εφαρμόστηκαν τεχνικές μείωσης και επιλογής χαρακτηριστικών στο στάδιο της προεπεξεργασίας, καθώς οι αλγόριθμοι μηχανικής μάθησης μπορούν να εκπαιδεύσουν με μεγαλύτερη επιτυχία τα μοντέλα όταν οι κατηγορίες των δεδομένων είναι λίγες.

Για να βεβαιωθούν οι ερευνητές πως το μοντέλο που θα δημιουργήσουν θα έχει τα βέλτιστα αποτελέσματα, αποφάσισαν να δημιουργήσουν συνολικά τέσσερα μοντέλα με τέσσερις διαφορετικούς αλγόριθμους εκμάθησης ώστε να επιλέξουν στην συνέχεια αυτό με τις καλύτερες μετρικές. Οι αλγόριθμοι που χρησιμοποιήθηκαν και τα αποτελέσματα που πέτυχε ο κάθε ένας ήταν τα εξής: ο Support Vector Machine (SVM) παρουσίασε 60.6% accuracy, 56% precision και 61% recall, ο Naive Bayes 56.7% accuracy, 57% precision και 57% recall, ο Decision Tree 50.4% accuracy, 51% precision και 50% recall ενώ ο Random Forest πέτυχε τις υψηλότερες μετρικές με 96.8% accuracy, 97% precision

και 97% recall. Επειδή όμως ο αλγόριθμος Random Forest έχει κάποια μειονεκτήματα, με το μεγαλύτερο να είναι ο μεγάλος χρόνος που χρειάζεται για να εκπαιδεύσει το μοντέλο, οι συγγραφείς τροποποίησαν τις χαμένες τιμές των δειγμάτων με την μέθοδο median / mode imputation, βρήκαν συσχετίσεις μεταξύ των κατηγοριών και εφάρμοσαν ξανά τον Random Forest παρέχοντας του το ψευδώνυμο “Improved Random Forest”. Οι τροποποιήσεις αυτές παρουσίασαν βελτιώσεις σε όλες τις μετρικές, με το accuracy να βρίσκεται πλέον στο 98.4% και το precision με το recall στο 98% , μειώνοντας επίσης σημαντικά και τον χρόνο που χρειάστηκε για την εκπαίδευση από 0.265 σε 0.164. [15]

Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting:

Λόγω της δυσκολίας διάγνωσης του καρκίνου του πνεύμονα στα αρχικά του στάδια και με τα προυπάρχοντα μοντέλα να μην αποδίδουν ικανοποιητικά, οδηγήθηκαν οι συγγραφείς της μελέτης αυτής στο να αναζητήσουν νέους τρόπους διάγνωσης της πάθησης. Έτσι αποφάσισαν να δοκιμάσουν να συσχετίσουν τα γονότυπα από 61 πολυμορφισμούς νουκλεοτιδίων (SNPs) με το ρίσκο απόκτησης του καρκίνου και να τα χρησιμοποιήσουν στην εκπαίδευση μοντέλων με την χρήση των αλγορίθμων Logistic Regression (LR) και Extreme Gradient Boosting (XGBoost), αφού πρώτα εφαρμόστηκε multivariate logistic regression για τον υπολογισμό της συσχέτισης μεταξύ των SNPs και του ρίσκου.

Ο αριθμός δειγμάτων που είχαν στην διάθεση τους οι συγγραφείς για την εκπαίδευση έφτανε τα 1979 και τα δύο μοντέλα αξιολογήθηκαν με την χρήση 10-fold cross validation, ενώ η κύρια μετρική σύγκρισης ήταν η Area Under the Curve (AUC). Για λόγους σύγκρισης δημιουργήθηκε επίσης ένα δεύτερο μοντέλο το οποίο δεν χρησιμοποιούσε τα SNPs κατά την διάγνωση, παρά μόνο κάποιες πληροφορίες για τους ασθενείς. Σημαντικό γεγονός αποτελεί η εύρεση δύο SNPs τα οποία παρουσίασαν πολύ υψηλή συσχέτιση με τον καρκίνο του πνεύμονα (TYMS rs3819102 και BAG6 rs1077393). Τα αποτελέσματα που παρουσιάστηκαν για το μοντέλο χωρίς τα SNPs είχαν AUC ίση με 0.703 με τον LR και 0.744 με τον XGBoost, ενώ τα αντίστοιχα του πλήρες μοντέλου που περιλάμβανε τα SNPs ήταν AUC ίση με 0.742 για τον LR και 0.759 για τον XGBoost. [16]

Optimized Deep Learning Model for Lung Cancer Prediction Using ANN Algorithm:

Με την άνοδο της τεχνητής νοημοσύνης ήρθε ο καιρός να εξελιχθούν και οι μέθοδοι διάγνωσης επικίνδυνων παθήσεων όπως είναι ο καρκίνος του πνεύμονα, ακόμη και σε περιοχές με μεγάλο αριθμό πληθυσμού και μικρά εισοδήματα. Αυτό είναι το βασικό κίνητρο των ερευνητών, οι οποίοι στοχεύουν στην δημιουργία ενός μοντέλου διάγνωσης του καρκίνου του πνεύμονα αρκετά ικανό ώστε να ανταπεξέλθει στις ανάγκες της καθημερινότητας των περιοχών αυτών. Συνολικά οι επιστήμονες

εκπαίδευσαν πέντε μοντέλα, καθένα από τα οποία χρησιμοποιεί έναν διαφορετικό αλγόριθμο τεχνητής νοημοσύνης για την εκμάθηση. Τα δεδομένα με τα οποία τροφοδότησαν τους αλγόριθμους είναι πλέον γνωστά και σε εμάς, καθώς είναι τα ίδια που χρησιμοποιήθηκαν και σε αυτή την πτυχιακή εργασία αλλά και σε έρευνες που αναλύσαμε παραπάνω.

Αναλυτικότερα, το σετ δεδομένων αποτελείται από 309 δείγματα με 16 χαρακτηριστικά το καθένα, τα οποία έχουν αξιολογηθεί και βαθμολογηθεί με την χρήση του Gini index. Οι αλγόριθμοι μηχανικής μάθησης που εφαρμόστηκαν ήταν οι Bayes Net (BN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF) και Artificial Neural Network (ANN), οι οποίοι στην πορεία συγκρίθηκαν με βάση την μετρική ακρίβειας που πέτυχαν έπειτα από αξιολόγηση 10-fold cross validation. Στην περίπτωση του τελευταίου αλγόριθμου (ANN) δοκιμάστηκαν τρεις εφαρμογές με τρία επίπεδα (layers), σε κάθε μια από τις οποίες προστιθόταν ένα ακόμη layer κατά την εκπαίδευση. Τη βέλτιστη απόδοση παρουσίασε τελικά ο ANN με την χρήση ενός μόνο layer εκπαίδευσης με 92.23% accuracy, 91.9% precision και 92.2% recall, με τον RF να έρχεται στην δεύτερη θέση με 91.26% accuracy, 90.7% precision και 91.3% recall. [17]

Lung cancer prediction using multi-gene genetic programming by selecting automatic features from amino acid sequences:

Με την εν λόγω έρευνα οι συγγραφείς παρουσιάζουν έναν ενδιαφέρον και καινοτόμο τρόπο για την διάγνωση του καρκίνου του πνεύμονα όσο αυτός είναι ακόμα σε αρχικό στάδιο. Αναλύοντας μεταλλαγμένα γονίδια που σχετίζονται με τον καρκίνο προέκυψαν κάποιες ενδιαφέροντες πληροφορίες σε πρωτεϊνικές αλληλουχίες αμινοξέων τις οποίες χρησιμοποίησαν για να δημιουργήσουν έναν νέο πολυγονιδιακό γενετικό αλγόριθμο εκμάθησης. Συνολικά το σετ δεδομένων αποτελείται από 1000 δείγματα με 23 χαρακτηριστικά, τα οποία επέλεξαν τυχαία οι επιστήμονες μέσα από μεγαλύτερα σετ για να δημιουργήσουν ένα νέο ισορροπημένο, το 70% του οποίου τροφοδοτήθηκε στον αλγόριθμο για εκπαίδευση και το υπόλοιπο 30% για επαλήθευση. Ο αλγόριθμος εκπαίδευσης ήταν ένα multi gene Genetic Programming Tree (MGGP) και αξιολογήθηκε με την χρήση πολλαπλών feature spaces με βάση μετρικές όπως accuracy, sensitivity, specificity και άλλες. Τα βέλτιστα αποτελέσματα προήλθαν από ένα feature space με το όνομα Hybrid-FS2 το οποίο επέφερε 95.6% accuracy, 97.3% sensitivity και 94% specificity, καθιστώντας τον νέο αυτό αλγόριθμο πιο ικανό από προηγούμενους που έχουν παρουσιαστεί σε σχετικές έρευνες. [18]

Predicting Lung Cancer Survivability using Ensemble Learning Methods:

Στη συγκεκριμένη έρευνα οι επιστήμονες προσπαθούν με την χρήση μεθόδων συνόλου (ensemble methods) να βελτιώσουν τις αποδόσεις διάφορων μοντέλων που δημιουργήθηκαν για την πρόβλεψη

της επιβίωσης ασθενών με καρκίνο του πνεύμονα για 5 έτη μετά την διάγνωση ώστε να αποδείξουν πως οι τεχνικές αυτές αποτελούν ικανά μέσα ενίσχυσης. Οι μέθοδοι που χρησιμοποιήθηκαν ήταν οι Bagging, Dagging, AdaBoost, MultiBoosting και Random SubSpace σε μοντέλα που εκπαιδεύτηκαν με την χρήση οκτώ διαφορετικών αλγορίθμων εκμάθησης, οι οποίοι είναι οι RIPPER, Decision Stump, Simple Cart, C4.5, SMO, Logistic Regression, Bayes Net και Random Forest. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιείχε πριν την προεπεξεργασία συνολικά 643.924 εγγραφές με 149 χαρακτηριστικά, πάνω στις οποίες εφαρμόστηκαν τεχνικές αφαίρεσης δειγμάτων που δεν πληρούσαν τις προδιαγραφές, αντικατάσταση χαμένων τιμών, τεχνικές τυχαίας υποδειγματοληψίας (random undersampling) για την ισορρόπηση των κατηγοριών και άλλες, ώσπου τελικά κάθε μια από τις δύο βασικές κατηγορίες περιείχε 8.742 τιμές.

Οι μετρικές που λήφθηκαν υπόψιν κατά την σύγκριση των αποτελεσμάτων των αλγορίθμων και της βελτίωσης τους από τις μεθόδους συνόλου ήταν το accuracy και το Area Under ROC Curve (AUC). Συνοπτικά, η μέθοδος AdaBoost σημείωσε την μεγαλύτερη βελτίωση και στις δύο μετρικές, ανεβάζοντας τα accuracy των RIPPER από 88.88% σε 88.98%, Decision Stump από 81.21% σε 87.67% και SMO από 83.41% σε 87.16% αλλά και τα AUC των RIPPER από 91.5% σε 94.9%, Decision Stump από 81.2% σε 93.9%, J48 από 94.1% σε 94.9% και του SMO από 50% σε 92.1%. Συνεπώς, κάποιοι μέθοδοι συνόλου αποτελούν αρκετά ικανά μέσα βελτίωσης των μοντέλων πρόβλεψης της επιβίωσης των ασθενών, παρουσιάζοντας υποσχόμενα αποτελέσματα. [19]

Α΄.5 Σύγκριση ερευνών

Παρότι έγινε λεπτομερής ανάλυση πολλαπλών ερευνών στην προηγούμενη υποενότητα που σχετίζονται άμεσα με την προσέγγιση της συγκεκριμένης πτυχιακής εργασίας, είναι αναγκαίο να γίνει και μια σύγκριση ανάμεσα στις διαφορετικές τεχνικές που χρησιμοποιεί η καθεμία, έτσι ώστε να καταφέρουμε να διακρίνουμε πιθανές ελλείψεις, μειονεκτήματα ή ακόμη και πλεονεκτήματα της υπάρχουσας βιβλιογραφίας.

[]	Στόχος	Μέθοδος - Αλγόριθμοι	Δείγμα	Αποτελέσματα
[5]	Risk detection	- Heuristic Greedy Best First Search - Random Forest - Oversampling	488	96.96% accuracy 96.26% specificity 96.34% sensitivity 96.32% f-score
[6]	Early detection	- AdaBoost Ensemble 10-Fold Cross Validation	25	76% accuracy 80% specificity

				• 70% sensitivity
[7]	Detection	• Logistic Regression	475	• 96.2% accuracy • 90.6% specificity
[8]	Early detection	• Neural Network • Logistic Regression • Decision Tree • Random Forest	4332	• 77.9% accuracy • 68.6% specificity • 0.79 AUC
[9]	Risk prediction	• Linear Predictors • Cox proportional hazard regression	47501	• 0.772 c-index
[10]	Prediction	• Multivariate logistic regression analysis	137	• 0.91 AUC • 80.3% sensitivity • 89.4% specificity
[11]	Prediction	• IBk • AdaBoost • LogitBoost • Random Forest • J48 • Logistic Regression • Support Vector Machine • 10-Fold cross validation	309	• 93.2% accuracy • 0.929 precision • 0.932 recall
[12]	Prediction	• SMOTE oversampling • XGBoost • LightGBM • Bagging • AdaBoost • 10-Fold cross validation	309	• 94.42% accuracy • 95.66% precision • 94.46% recall
[13]	Prediction of cancer stage	• K-NN • Decision Tree • Neural Network • Bootstrap Aggregating • Linear Transformation	1000	• 98% accuracy
[14]	Early Diagnosis	• Support Vector Machine	174	• 89% accuracy • 84% sensitivity

		• Cox-proportional hazard regression model • Borderline SMOTE		• 92% specificity
[15]	Prediction of Survival Rate	• SVM • Naive Bayes • Decision Tree • Random Forest • Median / Mode imputation	509	• 98.4% accuracy • 98% precision • 98% recall
[16]	Risk prediction	• Logistic Regression • XGBoost • 10-Fold cross validation	1979	• 0.759 AUC
[17]	Prediction	• Bayes Net • Naive Bayes • Decision Tree • Random Forest • Neural Network • 10-Fold cross validation	309	• 92.23% accuracy • 91.9% precision • 92.2% recall
[18]	Prediction	• Multi gene Genetic Programming Tree	1000	• 95.6% accuracy • 97.3% sensitivity • 94% specificity
[19]	Prediction of Survivability	• Bagging • Dagging • AdaBoost • MultiBoosting • Random Subspace • RIPPER • Decision Stump • Simple Cart • C4.5 • SMO • Logistic Regression • Bayes Net • Random Forest	17484	• 88.98% accuracy • 94.9% AUC

Στον παραπάνω πίνακα απεικονίζεται η σύγκριση της βιβλιογραφίας που χρησιμοποιήθηκε στην προηγούμενη ενότητα. Στην πρώτη στήλη παρατίθεται η αναφορά της έρευνας, ενώ στην δεύτερη καταγράφεται ο στόχος του κάθε άρθρου καθώς δεν αποσκοπούν όλα στην διάγνωση του καρκίνου του πνεύμονα, αλλά παρατηρούνται και άλλα που στοχεύουν στην αξιολόγηση του ρίσκου επιβίωσης του ασθενούς ή στον υπολογισμό του ρίσκου απόκτησης της πάθησης. Η τρίτη στήλη απεικονίζει τους αλγορίθμους που χρησιμοποιούνται σε κάθε μελέτη, είτε αυτοί αποσκοπούν στην δημιουργία ενός μοντέλου, είτε βοηθούν στην κανονικοποίηση ή επαύξηση δεδομένων. Στην τέταρτη στήλη παρατίθεται ο αριθμός των δειγμάτων που αποτέλεσαν τελικά την βάση για την δημιουργία των μοντέλων κάθε έρευνας, έπειτα δηλαδή από όλες τις τεχνικές που μπορεί να υποβλήθηκαν τα σύνολα δεδομένων. Η τελευταία στήλη παρουσιάζει τα αποτελέσματα που παρουσίασαν τα μοντέλα της βιβλιογραφίας όπου χρησιμοποιούνται μετρικές όπως η ακρίβεια (accuracy), η ευαισθησία (sensitivity), η εξειδίκευση (specificity) και πολλές άλλες. Αξίζει να σημειωθεί πως σε έρευνες που δημιούργησαν παραπάνω από ένα μοντέλα παρατέθηκαν τα αποτελέσματα του βέλτιστου, καθώς αυτό θα χρησιμοποιηθεί ως μέτρο σύγκρισης με τα υπόλοιπα.

Α'.6 Συμπεράσματα

Με βάση τη βιβλιογραφική έρευνα που πραγματοποιήθηκε για τα μοντέλα μηχανικής μάθησης που προβλέπουν τον καρκίνο του πνεύμονα, προέκυψαν σημαντικά αποτελέσματα. Οι περισσότερες μελέτες εστίασαν στη χρήση διαφόρων μοντέλων μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα, ενώ άλλες χρησιμοποίησαν μέχρι και τεχνικές ανάλυσης της αναπνοής του ασθενή με την χρήση καινοτόμων συσκευών για την πρόβλεψη της νόσου με υψηλή ευστοχία. Αυτή η έρευνα επιβεβαιώνει την αξία των μοντέλων στον ευρύ τομέα της ιατρικής, προσφέροντας νέες προοπτικές για την πρόληψη και πρόβλεψη της νόσου. Παράλληλα, αποδεικνύεται η σημαντική συμβολή της τεχνητής νοημοσύνης στη βελτίωση της διάγνωσης και θεραπείας του καρκίνου του πνεύμονα, ανοίγοντας νέους ορίζοντες για τη μελλοντική έρευνα στον χώρο αυτόν. Σε αυτό το σημείο, έχοντας μελετήσει ένα ικανοποιητικό κομμάτι της βιβλιογραφίας θα προχωρήσουμε στο επόμενο κεφάλαιο, όπου και θα παραθέσουμε τη δική μας υλοποίηση, αναλύοντας τόσο τα δεδομένα που χρησιμοποιήσαμε όσο και τις διάφορες τεχνικές. Στη συνέχεια θα συγκρίνουμε τα αποτελέσματα ώστε να επιλέξουμε τα μοντέλα που ταιριάζουν περισσότερο στις απαιτήσεις μας και θα τα συμπεριλάβουμε στη δημιουργία μιας διαδικτυακής εφαρμογής, ικανής να διαγνώσει τόσο την ύπαρξη της νόσου αλλά και να αξιολογήσει την δυνατότητα επιβίωσης του ασθενή.

Κεφάλαιο Β'

Περιγραφή Δεδομένων

Η σημασία της εύρεσης ποιοτικών δεδομένων στα μοντέλα μηχανικής μάθησης είναι κρίσιμη για την επίτευξη υψηλής ακρίβειας και απόδοσης. Τα δεδομένα αποτελούν τη βάση πάνω στην οποία χτίζονται τα μοντέλα και κάθε «πρόβλημα» στην ποιότητά τους μπορεί να οδηγήσει σε ανεπιθύμητα αποτελέσματα. Τα δύο σετ δεδομένων που χρησιμοποιήσαμε για την υλοποίηση των μοντέλων βρέθηκαν στο Kaggle, μια διαδικτυακή πλατφόρμα που στεγάζει έναν τεράστιο όγκο δεδομένων αλλά και έργων, τα οποία κάθε χρήστης μπορεί ελεύθερα να ανεβάσει ή αντίστοιχα να κατεβάσει και να χρησιμοποιήσει ή να ελέγξει και να παραθέσει τις απόψεις και τους προβληματισμούς του.

Β'.1 Δεδομένα διάγνωσης

Στο πρώτο μοντέλο της εφαρμογής που δημιουργήσαμε το οποίο θα αφιερωθεί στην διάγνωση του καρκίνου του πνεύμονα χρησιμοποιήσαμε ένα σετ δεδομένων με διακόσια ογδόντα τέσσερα (284) παραδείγματα και συνολικά δεκαέξι (16) μεταβλητές, οι οποίες περιγράφουν γενικά χαρακτηριστικά του χρήστη καθώς και συμπτώματα από τα οποία μπορεί να πάσχει. Παραδείγματα αυτών αποτελούν το γένος, η ηλικία, το κάπνισμα, η κατανάλωση αλκοόλ, η δυσκολία κατάποσης, ο βήχας και άλλα. Η τελευταία μεταβλητή του σετ δεδομένων για κάθε παράδειγμα είναι το αν πάσχει από καρκίνο του πνεύμονα ή όχι, κάτι που χρειάζεται ο εκάστοτε αλγόριθμος μηχανικής μάθησης για να εκπαιδεύσει το μοντέλο. Όλα τα δεδομένα είναι σε δυαδική μορφή πέρα από την ηλικία και το γένος. Οι ηλικίες που συμπεριλαμβάνονται είναι από είκοσι ενός (21) έως και ογδόντα επτά (87) ετών και το ποσοστό των ανδρών που συμπεριλαμβάνονται στα δείγματα αγγίζει το 52% ενώ των γυναικών το 48%. Παρότι το σετ δεδομένων δεν είναι μεγάλο σε σύγκριση με αυτά άλλων ερευνών που χρησιμοποιούν χιλιάδες χαρακτηριστικά, είναι ικανοποιητικό για τη δημιουργία ενός αξιόπιστου μοντέλου λόγω της ισορροπίας των δεδομένων του. [20]

Β'.2 Δεδομένα αξιολόγησης

Στο δεύτερο σκέλος της διαδικτυακής εφαρμογής αξιολογείται η επιβίωση του ασθενή με τη μέθοδο χειρουργικής επέμβασης, ο οποίος όμως έχει ήδη «διαγνωσθεί» από το πρώτο μοντέλο με καρκίνο του πνεύμονα. Για τη δημιουργία του μοντέλου αυτού χρησιμοποιήθηκε ένα δεύτερο σετ δεδομένων, το οποίο ήταν ανερτημένο στην ίδια πλατφόρμα με το πρώτο, το Kaggle. Τα δεδομένα αυτά συλλέχθηκαν από το κέντρο θωρακοχειρουργικής του Βρότσλαβ (Wroclaw Thoracic Surgery Centre) με δείγματα που προέρχονται από ασθενείς που υποβλήθηκαν σε χειρουργεία πνεύμονα μεταξύ 2007 και 2011. Το

σετ αυτό αποτελείται από 470 δείγματα με 16 χαρακτηριστικά για το κάθε ένα, ενώ οι τύποι των μεταβλητών ποικίλλουν. Πιο συγκεκριμένα, τα χαρακτηριστικά αφορούν συμπτώματα και νόσους που είχαν διαγνωσθεί προτού το χειρουργείο, με κάποια παραδείγματα να αποτελούν ο βήχας πριν το χειρουργείο, ο πόνος, η δύσπνοια, η αιμόπτυση, η κατάσταση απόδοσης με βάση τον δείκτη Zubrod, το μέγεθος του αρχικού όγκου, οι περιφερειακές αρτηριακές παθήσεις και άλλα. Τελευταίο και κυριότερο χαρακτηριστικό είναι το αν ο ασθενής επέζησε για παραπάνω από έναν χρόνο μετά το χειρουργείο ή αν απεβίωσε, το οποίο θα αποτελέσει τη βάση της εκπαίδευσης του μοντέλου. Τα δεδομένα παρότι δεν περιείχαν χαμένες τιμές, είχαν κάποια διπλότυπα δείγματα, τα οποία εν τέλει αφαιρέθηκαν. [21] [22]

Κεφάλαιο Γ'

Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων αποσκοπεί στην καθαριότητα, την απομάκρυνση θορύβου, την αντιμετώπιση των χαμένων τιμών και την κανονικοποίηση, προετοιμάζοντας τα δεδομένα για βέλτιστη ανάλυση από τα μοντέλα μηχανικής μάθησης. Έτσι, η προσεκτική διαχείριση των δεδομένων και η αποτελεσματική τους επεξεργασία αποτελούν το θεμέλιο για την επιτυχημένη υλοποίηση μηχανικών μοντέλων με περισσότερη ακρίβεια. Κάθε σετ δεδομένων βέβαια προσδοκεί διαφορετική αντιμετώπιση, για να προσαρμοστεί στην «ιδανική» του μορφή. Στην δική μας περίπτωση, χρησιμοποιήθηκαν κοινές μέθοδοι και στα δύο σετ. Επίσης δημιουργήσαμε πολλαπλές διαφορετικές εκδοχές των συνόλων δεδομένων μας ώστε να τις εισάγουμε σε διάφορους αλγόριθμους και να συγκρίνουμε τα αποτελέσματα για να επιλέξουμε τα βέλτιστα. Πάνω σε αυτό θα αναφερθούμε και παρακάτω, παρατίθοντας εικόνες που αποδεικνύουν τις επιλογές μας.

Μια τεχνική που χρησιμοποιήσαμε ήταν η αντιμετώπιση χαμένων τιμών, η οποία αναγκαία μόνο στο dataset της διάγνωσης, καθώς το δεύτερο δεν περιείχε χαμένες τιμές. Η αντιμετώπιση των χαμένων τιμών είναι σημαντική στη μηχανική μάθηση για τη διασφάλιση της ακρίβειας και της αξιοπιστίας των μοντέλων. Οι τιμές αυτές μπορούν να προκύψουν από πολλούς λόγους, όπως σφάλματα στη συλλογή δεδομένων ή απλά τη φύση των δεδομένων. Η αντιμετώπισή τους μπορεί να γίνει με διάφορους τρόπους, όπως η απλή αφαίρεση των εγγραφών με χαμένες τιμές, η αναπλήρωση τους με το μέσο ή τη μέση τιμή των υπόλοιπων δεδομένων, ή η χρήση προηγούμενων ή επόμενων τιμών. Οι αποφάσεις για το πώς να χειριστούμε τις χαμένες τιμές επηρεάζουν την τελική απόδοση του μοντέλου και πρέπει να γίνονται με προσοχή, λαμβάνοντας υπόψη τη φύση των δεδομένων και τον σκοπό του αναλυτικού μοντέλου. Στην δική μας περίπτωση αφαιρέσαμε τις εγγραφές με τις χαμένες τιμές, καθώς δεν ήταν αρκετές ώστε να επηρεάσουν το τελικό μας αποτέλεσμα.

Η δεύτερη μέθοδος ήταν η αντιμετώπιση διπλότυπων τιμών. Η χρήση της κρίθηκε αναγκαία μόνο για το δεύτερο dataset, αυτό της αξιολόγησης, καθώς μόνο αυτή περιείχε διπλότυπες εγγραφές. Η διαχείριση διπλότυπων τιμών στη μηχανική μάθηση είναι σημαντική για την αποφυγή παραμετροποίησης και τη βελτίωση της ακρίβειας των μοντέλων. Οι διπλότυπες τιμές μπορεί να προκύψουν από πολλούς λόγους, όπως για παράδειγμα λόγω της ύπαρξης πολλαπλών εγγραφών για την ίδια παρατήρηση. Για τη διαχείρισή τους, όπως και με την αντιμετώπιση χαμένων τιμών, μπορούμε να χρησιμοποιήσουμε τεχνικές όπως η αφαίρεση των διπλότυπων εγγραφών, η συγχώνευσή τους ή η αντιμετώπισή τους ως ένα ξεχωριστό χαρακτηριστικό. Επιλέγοντας την κατάλληλη τεχνική, μπορούμε να διασφαλίσουμε την ομαλή λειτουργία του μοντέλου και την αξιοπιστία των αποτελεσμάτων. Γενικώς συνίσταται να μην διαγράφουμε τις διπλότυπες εγγραφές όταν με αυτό τον τρόπο αφαιρείται παραπάνω από τον μισό όγκο των συνολικών εγγραφών της επιθυμητής τιμής με τον μικρότερο όγκο,

ειδικά όταν έχουμε να κάνουμε με ανισόρροπα δεδομένα, όπως και στην περίπτωση μας. Παρόλα αυτά, εφόσον οι συνολικές διπλότυπες εγγραφές που περιείχε το δικό μας dataset ήταν ελάχιστες, καταλήξαμε στην διαγραφή τους.

Στη συνέχεια χρησιμοποιήσαμε την τεχνική spread subsample, η οποία αντιμετωπίζει το πρόβλημα της μη ισορροπίας δεδομένων, ένα συχνό φαινόμενο στα πραγματικά σύνολα δεδομένων όπου οι κλάσεις δεν έχουν ισορροπημένο αριθμό παραδειγμάτων. Κατά την εφαρμογή αυτής της τεχνικής, δημιουργήσαμε ένα υποσύνολο των δεδομένων με ισορροπημένο αριθμό παρατηρήσεων από κάθε κλάση. Αυτό επιτυγχάνεται με την τυχαία επιλογή μιας ισορροπημένης δειγματοληψίας από κάθε κλάση. Στη συνέχεια, τα επιλεγμένα δείγματα τα συνδυάσαμε για να δημιουργήσουμε το τελικό ισορροπημένο υποσύνολο δεδομένων. Ωστόσο, πρέπει να ληφθεί υπόψη ότι η χρήση αυτής της τεχνικής μπορεί να μειώσει τον αριθμό των δεδομένων που έχουμε στη διάθεσή μας, πράγμα που μπορεί να επηρεάσει την ικανότητα γενίκευσης του μοντέλου. Στη δική μας περίπτωση και λόγω μεγάλης εμφάνισης της μη ισορροπίας ανάμεσα στις κλάσεις και των δύο συνόλων δεδομένων, η τεχνική αυτή αποτέλεσε την «βάση» της προεπεξεργασίας, καθώς χωρίς αυτή οι μετρικές των μοντέλων μας δεν θα ήταν ικανοποιητικές και συνεπώς δεν θα μπορούσαμε να δημιουργήσουμε μια αποδεκτή εφαρμογή. Για την ακρίβεια δημιουργήσαμε πολλαπλά αντίτυπα των συνόλων δεδομένων μας, τα οποία είχαν πενήντα, εξήντα και εβδομήντα χαρακτηριστικά κλάσης αντίστοιχα για κάθε διαφορετικό σετ δεδομένων. Με αυτό τον τρόπο μπορέσαμε να βρούμε τον ιδανικό αριθμό χαρακτηριστικών που οδηγεί το μοντέλο μας στα καλύτερα αποτελέσματα.

Η τελευταία τεχνική που χρησιμοποιήσαμε ήταν η επιλογή χαρακτηριστικών, γνωστή και ως “Attribute Selection”, η οποία αποτελεί ένα σημαντικό στάδιο της διαδικασίας προεπεξεργασίας δεδομένων. Κατά την εφαρμογή αυτής της τεχνικής, επιλέξαμε τα σημαντικότερα χαρακτηριστικά από τα σύνολα δεδομένων μας, τα οποία παρέχουν τη μέγιστη συνεισφορά στην ακρίβεια και την απόδοση των μοντέλων. Οι τεχνικές επιλογής χαρακτηριστικών γενικότερα μπορεί να περιλαμβάνουν την ανάλυση της συσχέτισης μεταξύ των χαρακτηριστικών και της ετικέτας εξόδου, την εφαρμογή αλγορίθμων επιλογής χαρακτηριστικών ή τη χρήση μετρικών σημαντικότητας. Η επιλογή των σωστών χαρακτηριστικών βοηθά στην εξάλειψη της περιττής πληροφορίας και τη βελτίωση της απόδοσης του μοντέλου, καθώς μειώνει τη διάσταση των δεδομένων και ενισχύει την ικανότητά του να γενικεύει σωστά σε νέα δεδομένα. Πράγματι και στη δική μας περίπτωση, με την χρήση της επιλογής χαρακτηριστικών και στα δύο σύνολα δεδομένων μας και σε συνδυασμό με την τεχνική spread subsample που αναφέραμε πριν, καταφέραμε να βελτιώσουμε σημαντικά τις μετρικές και των δύο μοντέλων μας. Αυτό θα το διακρίνουμε και παρακάτω κατά την σύγκριση των διαφόρων συνόλων δεδομένων που δημιουργήσαμε, ώστε να δούμε ποιο προσφέρει τις καλύτερες μετρικές με τον αλγόριθμο που επιλέξαμε.

Κεφάλαιο Δ'

Εκπαίδευση και αποτελέσματα

Το πρώτο βήμα που πρέπει να γίνει μετά την προεπεξεργασία των δεδομένων είναι η εύρεση ενός αλγόριθμου, με τη χρήση του οποίου θα εκπαιδεύσουμε τα μοντέλα μας. Οι αλγόριθμοι μηχανικής μάθησης είναι αλγόριθμοι που επιτρέπουν στους υπολογιστές να μαθαίνουν από τα δεδομένα και να προβλέπουν μοντέλα και τάσεις, χωρίς να προγραμματίζονται εκ των προτέρων για συγκεκριμένες εργασίες. Οι αλγόριθμοι αυτοί αντλούν από το σύνολο των δεδομένων που τους παρέχονται, αναγνωρίζοντας πρότυπα και εξάγοντας σημαντικές πληροφορίες. Από την αναγνώριση προτύπων σε εικόνες μέχρι την πρόβλεψη τάσεων σε δεδομένα, οι αλγόριθμοι μηχανικής μάθησης έχουν εφαρμογές σε πολλούς τομείς, ενισχύοντας την απόδοση και την ακρίβεια σε διαφορετικά πεδία. Από την ιατρική έως την αυτοκινητοβιομηχανία και την χρηματοοικονομική ανάλυση, η εφαρμογή των αλγορίθμων μηχανικής μάθησης έχει μετασχηματίσει τον τρόπο με τον οποίο λειτουργούν οι οργανισμοί και η κοινωνία γενικότερα. Συνήθως χωρίζονται σε τρεις κύριες κατηγορίες: στην επιβλεπόμενη μάθηση, τη μη επιβλεπόμενη μάθηση και την ενισχυτική μάθηση. Στην επιβλεπόμενη μάθηση, οι αλγόριθμοι χρησιμοποιούν ετικετοποιημένα δεδομένα εισόδου-εξόδου. Στην μη επιβλεπόμενη μάθηση, οι αλγόριθμοι δεν διαθέτουν ετικέτες για τα δεδομένα εισόδου και επιδιώκουν να ανακαλύψουν τη δομή ή τα πρότυπα μέσα στα δεδομένα. Στην ενισχυτική μάθηση, ο αλγόριθμος εκπαιδεύεται να προβαίνει σε αποφάσεις με βάση την αλληλεπίδρασή του με ένα περιβάλλον. Σήμερα υπάρχουν εκατοντάδες αλγόριθμοι διαθέσιμοι, κάθε ένας από τους οποίους διαχειρίζεται καλύτερα κάποιο συγκεκριμένο τύπο δεδομένων, επιφέροντας έτσι καλύτερα αποτελέσματα. Υπάρχουν επίσης και διαφορετικές εκδοχές για κάθε αλγόριθμο, που είτε επιφέρουν αναβαθμίσεις στην προηγούμενη έκδοση, είτε δίνουν τη δυνατότητα καλύτερης διαχείρισης συνόλων δεδομένων.

Δ'.1 Επιλογή αλγόριθμου εκπαίδευσης

Για την περίπτωση μας χρειάστηκε να βρούμε έναν αλγόριθμο που να μπορεί να διαχειριστεί τόσο διαφορετικούς τύπους δεδομένων (δυαδικούς αλλά και κειμένους), όσο και σύνολα δεδομένων με μεγάλο ποσοστό ανισορροπίας. Συνολικά τροφοδοτήσαμε σε οκτώ διαφορετικούς αλγορίθμους τις πολλαπλές εκδοχές των συνόλων δεδομένων που δημιουργήσαμε, όπως αναφέραμε στο προηγούμενο υποκεφάλαιο. Οι αλγόριθμοι αυτοί είναι οι K-Nearest Neighbor (KNN), Gaussian Naive Bayes, Gradient Boosting, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine (SVM) και Extreme Gradient Boosting. Η μετρική η οποία θα αποτελέσει τη βάση σύγκρισης των αποτελεσμάτων μας θα είναι η μέση ισορροπημένη ακρίβεια ή average balanced accuracy, η οποία χρησιμοποιείται για να αξιολογήσει την απόδοση ενός μοντέλου σε περιπτώσεις μη ισορροπημένων κλάσεων. Υπολογίζεται ως ο μέσος όρος των ακριβειών για κάθε κλάση, εξισορροπώντας τις

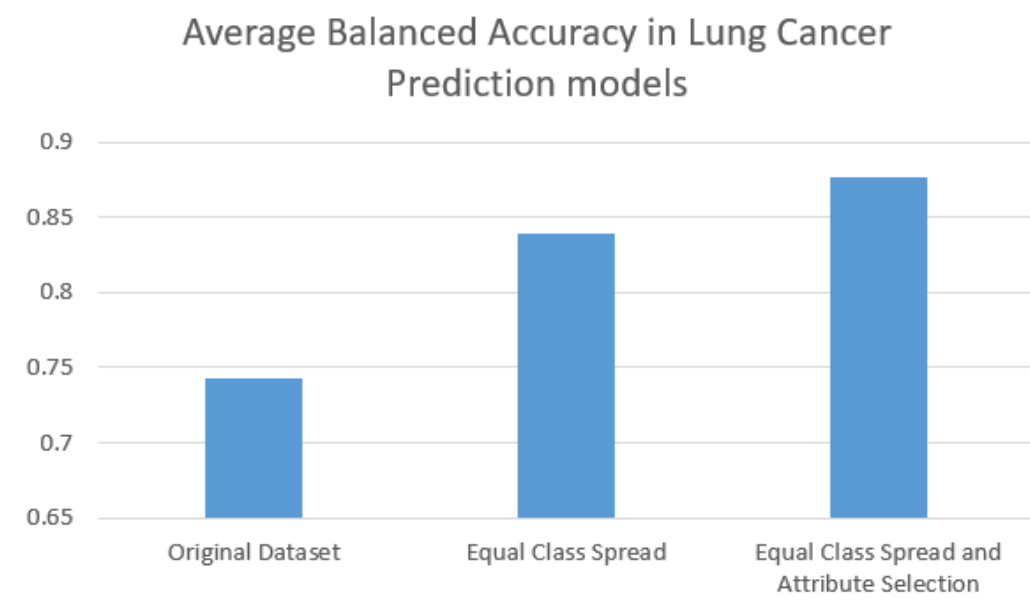
ανισορροπίες στο πλήθος των παρατηρήσεων ανά κλάση. Αυτό επιτρέπει στη μετρική να αξιολογήσει με μεγαλύτερη ακρίβεια την ικανότητα ενός μοντέλου να γενικεύει σε όλες τις κλάσεις του προβλήματος, ανεξαρτήτως της ανισορροπίας στα δεδομένα. Συνεπώς η υψηλή μέση ισορροπημένη ακρίβεια υποδηλώνει ότι το μοντέλο έχει καλή ικανότητα να ανταποκρίνεται σε όλες τις κλάσεις του προβλήματος, ενώ η χαμηλή μέση ισορροπημένη ακρίβεια μπορεί να υποδεικνύει προβλήματα γενίκευσης ή ανεπαρκή εκπαίδευση του μοντέλου, γεγονός που χρίζει τη μετρική αυτή ιδανική για την σύγκριση των δικών μας αποτελεσμάτων.

Τα μέγιστα αποτελέσματα που πέτυχε κάθε αλγόριθμος σε ποσοστό του average balanced accuracy για κάποιο από τα δύο σύνολα δεδομένων ήταν τα εξής: οι Gaussian Naive Bayes και SVM βρίσκονταν περίπου στο 50%, ο KNN πέτυχε ποσοστό μικρότερο του 60%, οι Gradient Boosting και Decision Tree κατάφεραν από 64% ο κάθε ένας, ενώ ο Random Forest και ο Logistic Regression έφτασαν το 65%. Παρότι κάποια από τα παραπάνω αποτελέσματα ήταν ικανοποιητικά, ιδίως αυτά που ξεπερνούσαν το 60%, υπήρξε ένας αλγόριθμος που επέφερε τα μεγαλύτερα αποτελέσματα και για τα δύο σύνολα δεδομένων. Αυτός ήταν ο Extreme Gradient Boosting, ο οποίος κατάφερε ένα 87% για το σετ της διάγνωσης και ένα 66% για το σετ δεδομένων της αξιολόγησης. Πιο αναλυτικά, ο αλγόριθμος Extreme Gradient Boosting (ή αλλιώς XGBoost) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται ευρέως σε πολλές εφαρμογές. Αυτός ο αλγόριθμος βασίζεται στην τεχνική του gradient boosting, η οποία επιδιώκει τη δημιουργία ενός ισχυρού μοντέλου με τη συνδυασμένη χρήση πολλών αδύναμων μοντέλων (weak learners). Έχει επίσης τη δυνατότητα να αντιμετωπίζει αποτελεσματικά τα προβλήματα μη ισορροπημένων δεδομένων όπως αυτά που διαθέτουμε. Ως εκ τούτου, ο XGBoost αποτελεί έναν ισχυρό και ευέλικτο αλγόριθμο που επιτυγχάνει εξαιρετική απόδοση σε πολύπλοκα προβλήματα μηχανικής μάθησης. Συνεπώς, είναι επόμενο να συνεχίσουμε την πορεία μας με την χρήση αυτού, δοκιμάζοντας παρακάτω περισσότερες διαφοροποιήσεις στην επεξεργασία των δεδομένων για να δούμε αν είναι εφικτή η περαιτέρω βελτίωση των μετρικών μας.

Δ'.2 Μετρικές μοντέλου διάγνωσης

Όπως αναφέραμε προηγουμένως, για κάθε σύνολο δεδομένων δημιουργήσαμε πολλαπλές εκδοχές με διαφορετική επεξεργασία στο κάθε ένα, με στόχο την εφαρμογή τους στον επιλεγμένο αλγόριθμο και την εύρεση των βέλτιστων αποτελεσμάτων. Εφόσον λοιπόν έχουμε αποφασίσει να προχωρήσουμε με τον αλγόριθμο Extreme Gradient Boosting (XGBoost), μένει να βρούμε την βέλτιστη εκδοχή κάθε ενός από τα σύνολα δεδομένων μας. Θα ξεκινήσουμε με το πρώτο σετ δεδομένων που αφορά την διάγνωση του καρκίνου του πνεύμονα. Για αυτό όπως θα δούμε και στην εικόνα παρακάτω δημιουργήσαμε δύο διαφορετικές περιπτώσεις. Για την πρώτη έγινε μόνο εξισορρόπηση των κλάσεων με την τεχνική Spread Subsample που περιγράψαμε προηγουμένως, η οποία βοηθά στην βελτιστοποίηση των αποτελεσμάτων σε δεδομένα χωρίς ισορροπία. Στη δεύτερη περίπτωση κάναμε το ίδιο με την πρώτη,

δηλαδή εξισορροπήσαμε τις κλάσεις με την ίδια τεχνική αλλά προσθέσαμε ακόμη ένα βήμα, το Attribute Selection, το οποίο περιόρισε τον αριθμό μεταβλητών, αλλά βελτίωσε σημαντικά τα αποτελέσματα. Παρακάτω βλέπουμε και την εικόνα που περιέχει την σύγκριση των αποτελεσμάτων με βάση το average balanced accuracy, με την πρώτη στήλη να ανήκει στο πρωταρχικό και ελαφρώς επεξεργασμένο σύνολο δεδομένων μας, τη δεύτερη σε αυτό με την εξισορρόπηση των κλάσεων και την τρίτη στον συνδυασμό των spread subsample και attribute selection. Όπως παρατηρούμε, και οι δύο τεχνικές μας προσδίδουν καλύτερα αποτελέσματα από ότι θα είχαμε χωρίς αυτές, με τον συνδυασμό των δύο να αποτελεί την βέλτιστη επιλογή. Άρα με αυτό το σύνολο δεδομένων θα δημιουργήσουμε το πρώτο μας μοντέλο για την διάγνωση του καρκίνου του πνεύμονα. Στο επόμενο υποκεφάλαιο θα παρατεθούν και οι υπόλοιπες μετρικές με γραφήματα.

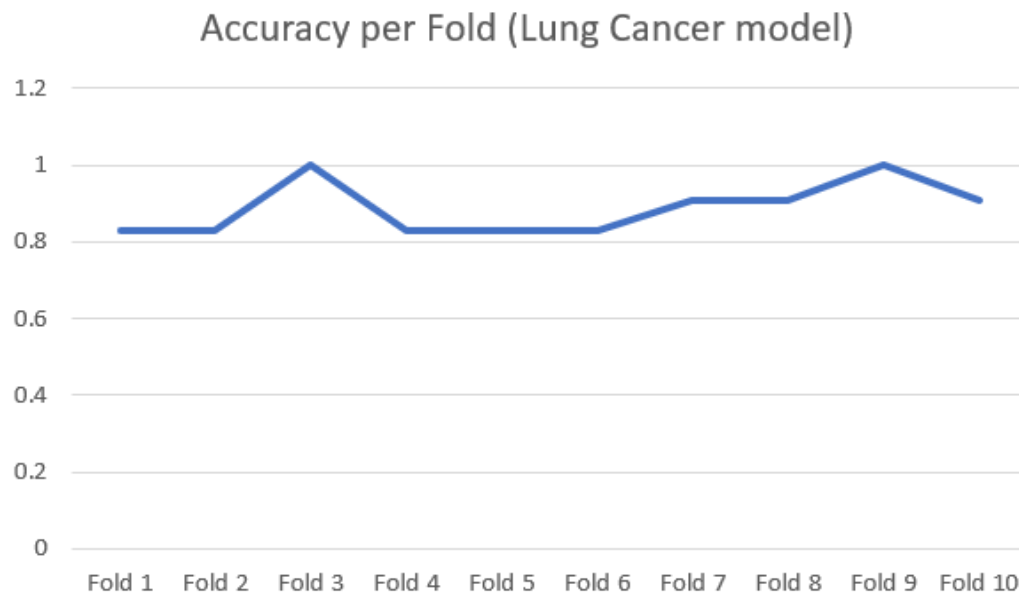


Εικόνα Δ'.1: Η μέση ισορροπημένη ακρίβεια για κάθε σύνολο δεδομένων του μοντέλου διάγνωσης.

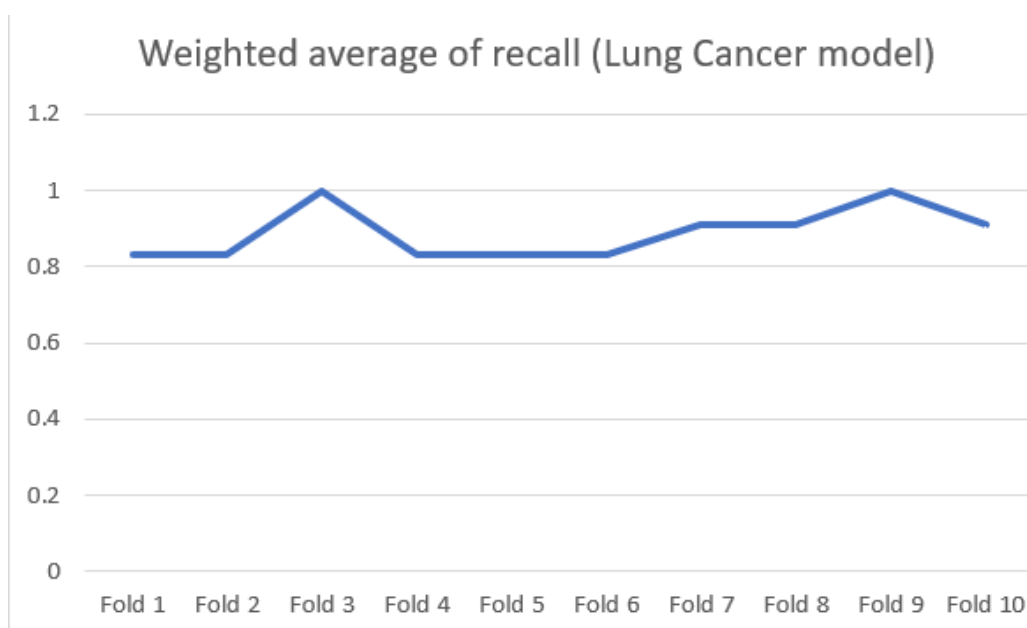
Πέρα από τη μέση ισορροπημένη ακρίβεια, κάθε μοντέλο αξιολογείται και από άλλες μετρικές, ώστε να εξακριβωθεί η ικανότητα του να αναγνωρίζει και να παίρνει σωστές αποφάσεις. Κάποιες από τις μετρικές αυτές είναι η ακρίβεια (accuracy), η ανάκληση (recall), η ακρίβεια (precision) αλλά και το F1 Score, με τις οποίες και θα ασχοληθούμε. Αναλυτικότερα, ως ακρίβεια (accuracy) στη μηχανική μάθηση αναφέρεται το ποσοστό των σωστών προβλέψεων που έκανε ένα μοντέλο σε σχέση με το συνολικό αριθμό των προβλέψεων που έκανε, δηλαδή τον συνολικό αριθμό παρατηρήσεων που το μοντέλο προέβλεψε σωστά, διαιρούμενο με το συνολικό αριθμό των παρατηρήσεων. Η ακρίβεια είναι μια καλή μετρική κυρίως όταν έχουμε δεδομένα που είναι ισορροπημένα σε όλες τις κλάσεις. Η ανάκληση (recall) από την άλλη, αναφέρεται στο ποσοστό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό των πραγματικών θετικών παρατηρήσεων, δηλαδή τον αριθμό των πραγματικά θετικών παρατηρήσεων που προβλέφθηκαν σωστά, διαιρούμενο με τον συνολικό αριθμό των

πραγματικών θετικών παρατηρήσεων. Στη συνέχεια, ως ακρίβεια (precision) θεωρείται το ποσοστό των σωστών θετικών προβλέψεων σε σχέση με το συνολικό αριθμό των θετικών προβλέψεων που έκανε το μοντέλο, δηλαδή τον αριθμό των πραγματικά θετικών παρατηρήσεων που προβλέφθηκαν σωστά, διαιρούμενο με τον συνολικό αριθμό των παρατηρήσεων που προβλέφθηκαν ως θετικές από το μοντέλο. Στο τέλος έχουμε το F1 score, το οποίο είναι ένας αριθμητικός δείκτης που συνδυάζει την ακρίβεια (precision) και την ανάκληση (recall) του μοντέλου σε ένα μόνο μέτρο, λαμβάνοντας υπόψη και τις δύο μετρικές. Υπολογίζεται ως ο μέσος όρος της ακρίβειας και της ανάκλησης, και παρέχει έναν καλό τρόπο για να αξιολογηθεί η συνολική απόδοση ενός μοντέλου, για αυτό και πήραμε την απόφαση να το συμπεριλάβουμε στα αποτελέσματα μας.

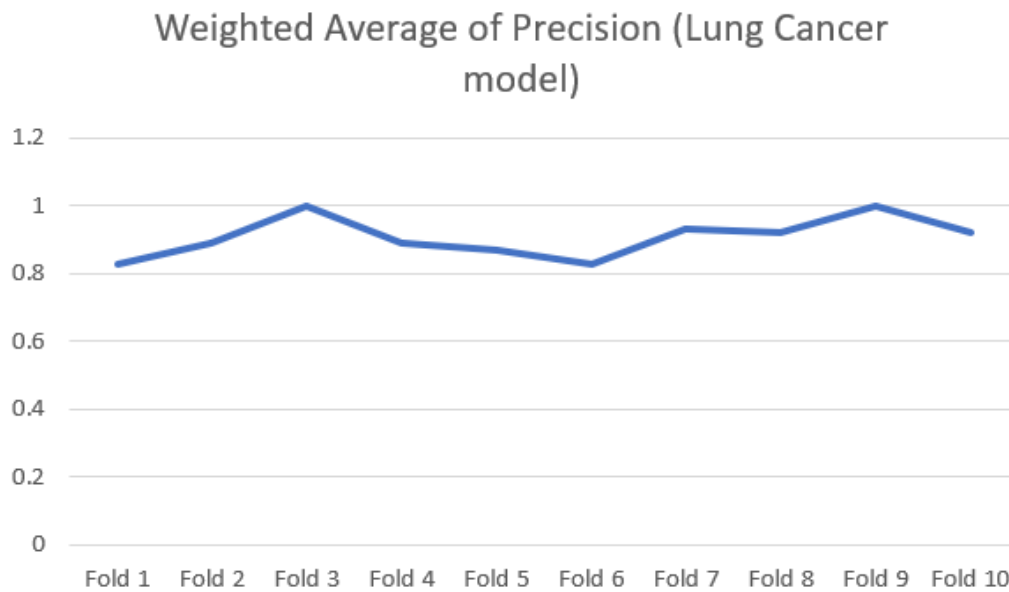
Για να υπολογιστούν όλες αυτές οι μετρικές όμως, χρειαζόμαστε ένα μέσο ή μια τεχνική αξιολόγησης του μοντέλου που πρόκειται να παραχθεί. Εδώ χρησιμοποιήσαμε την διασταύρωση K-fold (ή K-fold Cross Validation), η οποία είναι μια τεχνική αξιολόγησης της απόδοσης ενός μοντέλου μηχανικής μάθησης που βοηθά στην εκτίμηση της γενικευμένης ικανότητάς του να γενικεύει σε νέα δεδομένα. Σε αυτήν την τεχνική, τα δεδομένα διαιρούνται σε K ισότιμα μη-αναστρεπτικά σύνολα, ή "folds", στα οποία ένα από αυτά χρησιμοποιείται ως σύνολο ελέγχου και τα υπόλοιπα (K-1) σύνολα χρησιμοποιούνται για εκπαίδευση του μοντέλου. Η διαδικασία αυτή επαναλαμβάνεται K φορές, κάθε φορά με ένα διαφορετικό σύνολο ελέγχου. Οι μετρικές απόδοσης του μοντέλου υπολογίζονται συνολικά για όλα τα K σύνολα ελέγχου και στο τέλος υπολογίζεται ο μέσος όρος τους για να μπορούμε να προσδώσουμε ένα τελικό μέτρο απόδοσης. Το K-fold cross validation επιτρέπει τη χρήση όλων των δεδομένων για εκπαίδευση και αξιολόγηση του μοντέλου, βοηθώντας έτσι στην πιο αξιόπιστη εκτίμηση της απόδοσής του. Παράλληλα, υπάρχει μια τάση στη μηχανική μάθηση να επιλέγουμε τη διασταύρωση 10-fold γιατί προσφέρει μια καλή ισορροπία μεταξύ ακρίβειας και υπολογιστικού φόρτου. Η διασταύρωση αυτή διαιρεί τα δεδομένα σε 10 ισότιμα μη-αναστρεπτικά σύνολα, εκ των οποίων κάθε ένα χρησιμοποιείται ως σύνολο ελέγχου, ενώ τα υπόλοιπα 9 χρησιμοποιούνται για εκπαίδευση. Αυτό επιτρέπει να έχουμε μια αξιόπιστη εκτίμηση της απόδοσης του μοντέλου, αφού κάθε παρατήρηση χρησιμοποιείται τόσο για εκπαίδευση όσο και για αξιολόγηση. Ταυτόχρονα, η διαδικασία αυτή δεν απαιτεί υπερβολικό υπολογιστικό φόρτο, καθώς επαναλαμβάνεται μόνο 10 φορές. Παρακάτω ακολουθούν τα αποτελέσματα που προέκυψαν ανά fold για τα δύο μοντέλα μας με τη χρήση γραφημάτων, ξεκινώντας από αυτά του μοντέλου διάγνωσης του καρκίνου του πνεύμονα.



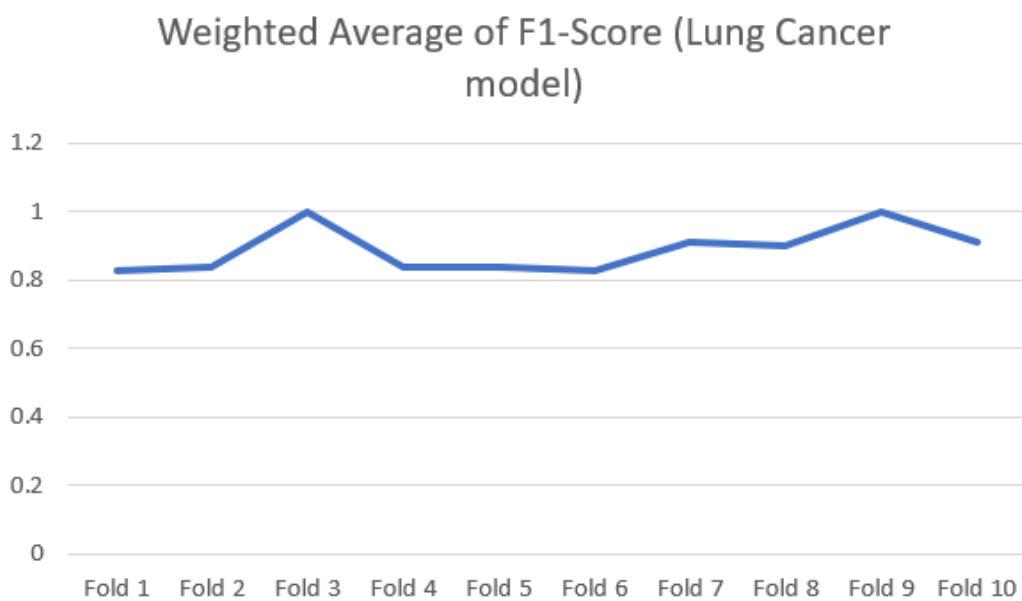
Εικόνα Δ'.2: Η ακρίβεια (accuracy) του μοντέλου διάγνωσης.



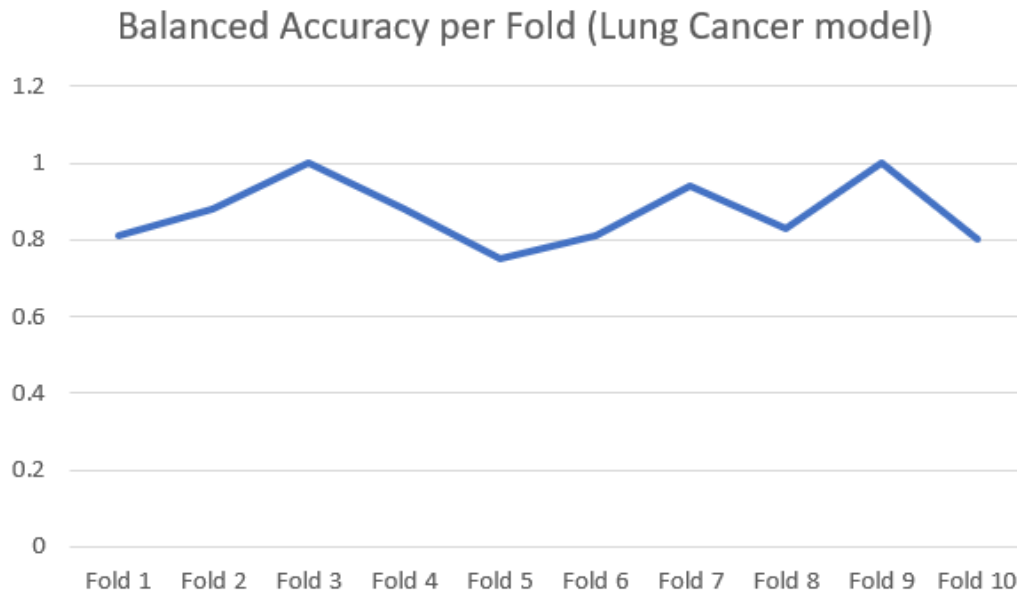
Εικόνα Δ'.3: Ο σταθμισμένος μέσος όρος της ανάκλησης (recall) του μοντέλου διάγνωσης.



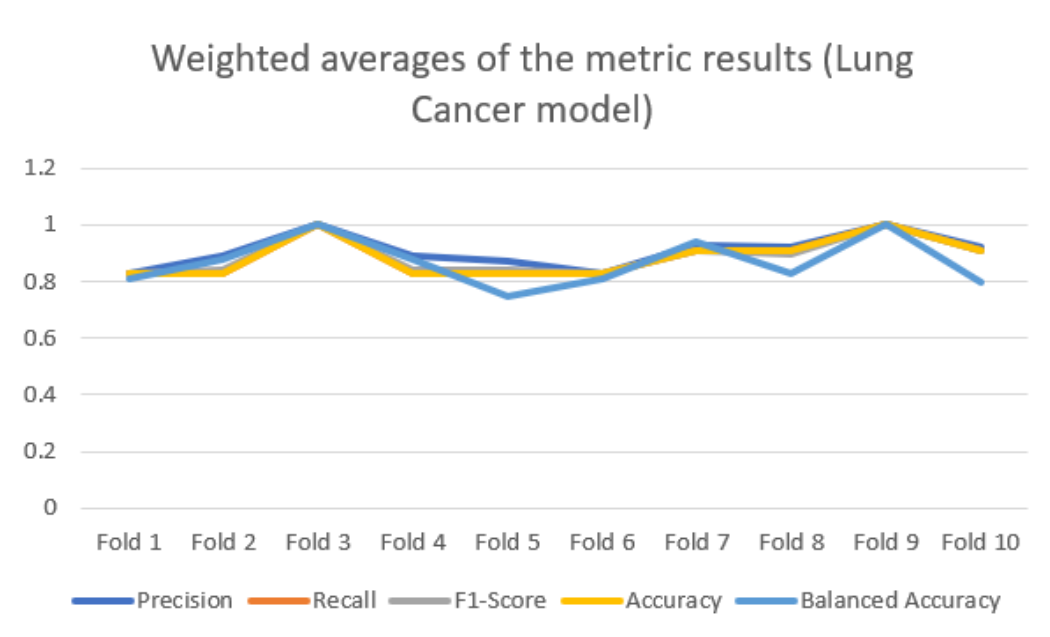
Εικόνα Δ'.4: Ο σταθμισμένος μέσος όρος της ακρίβειας (precision) του μοντέλου διάγνωσης.



Εικόνα Δ'.5: Ο σταθμισμένος μέσος όρος του F1-score του μοντέλου διάγνωσης.



Εικόνα Δ'.6: Η ισορροπημένη ακρίβεια (balanced accuracy) του μοντέλου διάγνωσης.

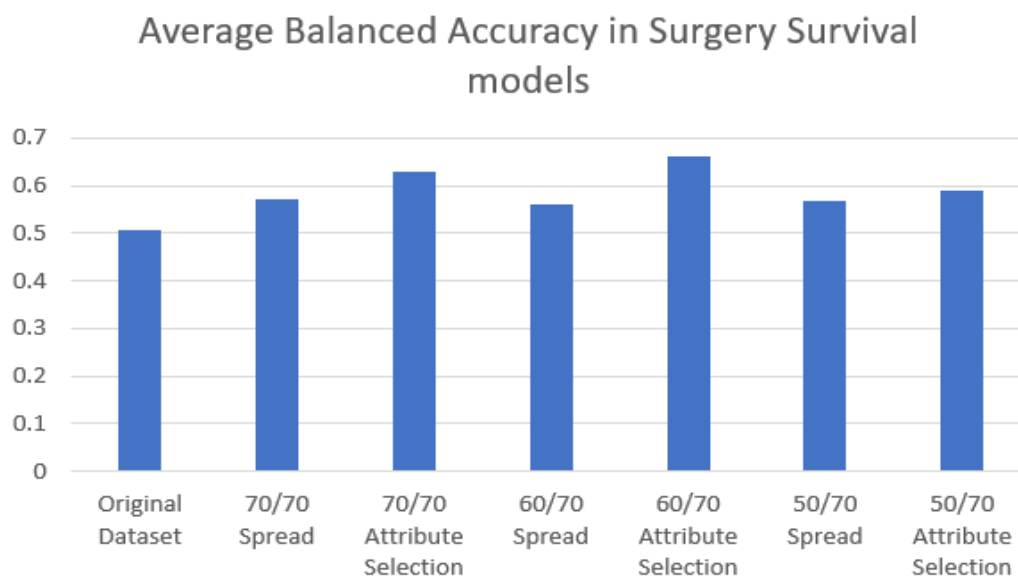


Εικόνα Δ'.7: Το σύνολο των σταθμισμένων μέσων όρων όλων των μετρικών του μοντέλου διάγνωσης.

Συνολικά μπορούμε να διακρίνουμε πολύ ικανοποιητικά αποτελέσματα μετά την αξιολόγηση του πρώτου μοντέλου που θα δημιουργήσουμε, με τους μέσους όρους των accuracy, recall, precision, f1-score και balanced accuracy να φθάνουν τα 80,5%, 88,8%, 90,8%, 88,7% και 87,7% αντίστοιχα. Με αυτές τις μετρικές μπορούμε πλέον να προχωρήσουμε στη δημιουργία του μοντέλου, καθώς βεβαιωνόμαστε για την αποτελεσματικότητά του.

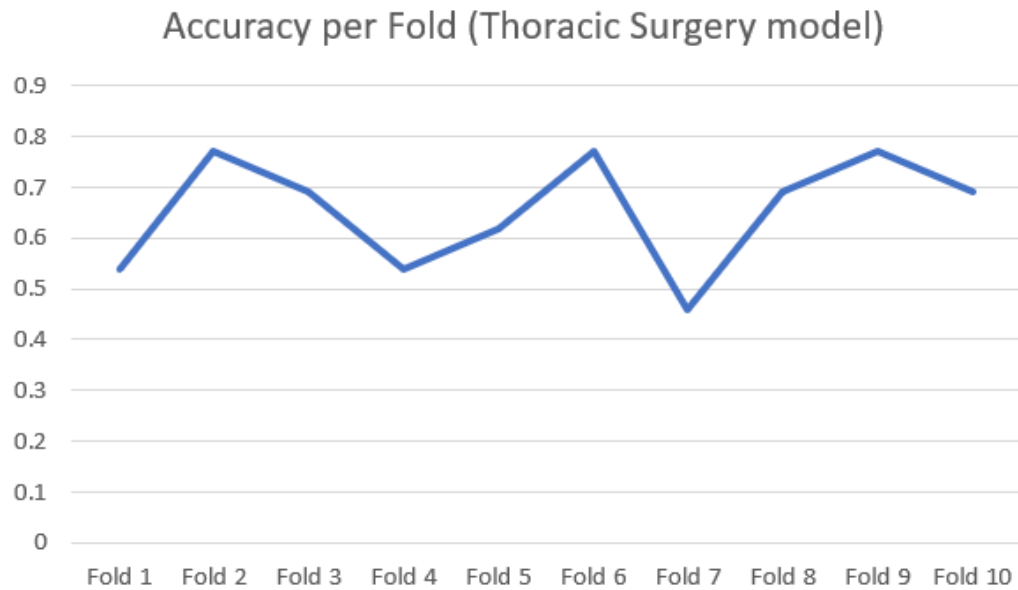
Δ'.3 Μετρικές μοντέλου αξιολόγησης

Την ίδια διαδικασία που είδαμε στο προηγούμενο υποκεφάλαιο θα ακολουθήσουμε και για το δεύτερο σετ δεδομένων, συγκεκριμένα αυτό της αξιολόγησης της επιβίωσης ενός ασθενή, ώστε να καταλήξουμε στο καλύτερο δυνατό μοντέλο. Παρατηρώντας το γράφημα παρακάτω μπορεί κανείς να διακρίνει πως σε αυτό το σύνολο δεδομένων δημιουργήσαμε περισσότερες εκδοχές, «παίζοντας» θα έλεγε κανείς με τις τιμές των δύο κλάσεων με την χρήση του spread subsample. Αυτό έγινε διότι το συγκεκριμένο σύνολο δεδομένων έχει πολύ μεγάλη απόκλιση μεταξύ των κλάσεων του, κάτι που το κάνει μη διαχειρίσιμο στην πρωταρχική του μορφή, όπως μπορούμε να δούμε και από το average balanced accuracy του που βρίσκεται στο 50%. Για αυτό αυξομειώσαμε λοιπόν τα σύνολα των δύο κλάσεων, όπως μπορούμε να διακρίνουμε και στην εικόνα, σε 70/70, 60/70 και 50/70 αντίστοιχα, χρησιμοποιώντας σε κάθε μια από αυτές και το attribute selection. Με αυτές τις τεχνικές καταφέραμε να βελτιστοποιήσουμε το αποτέλεσμα μας έως και 16% παραπάνω από το αρχικό, καθώς το σετ δεδομένων με τις κλάσεις στο 60/70 αλλά και τη χρήση της επιλογής χαρακτηριστικών αγγίζει το 66% average balanced accuracy.

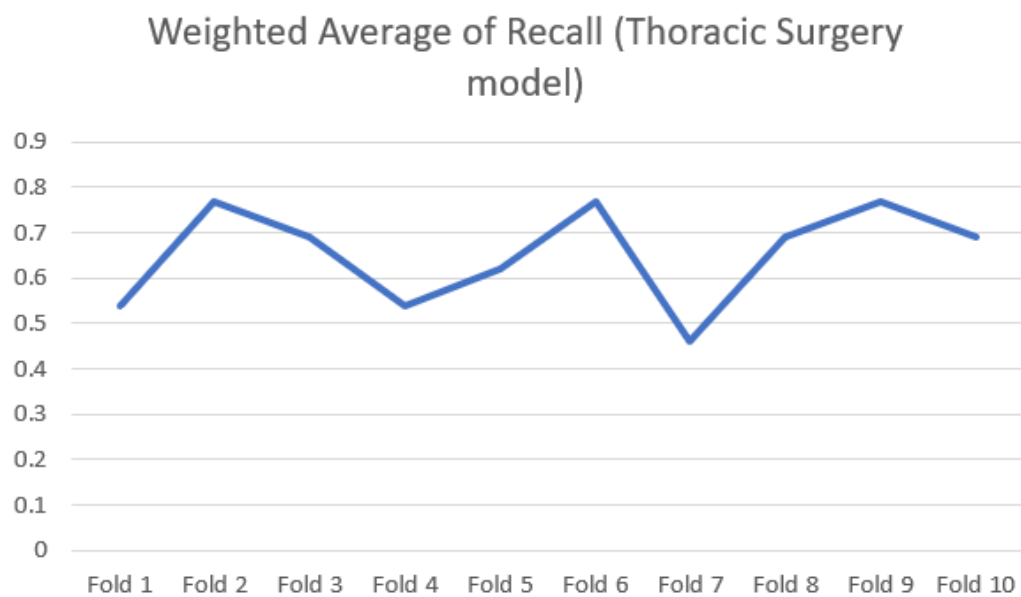


Εικόνα Δ'.8: Η μέση ισορροπημένη ακρίβεια για κάθε σύνολο δεδομένων του μοντέλου αξιολόγησης.

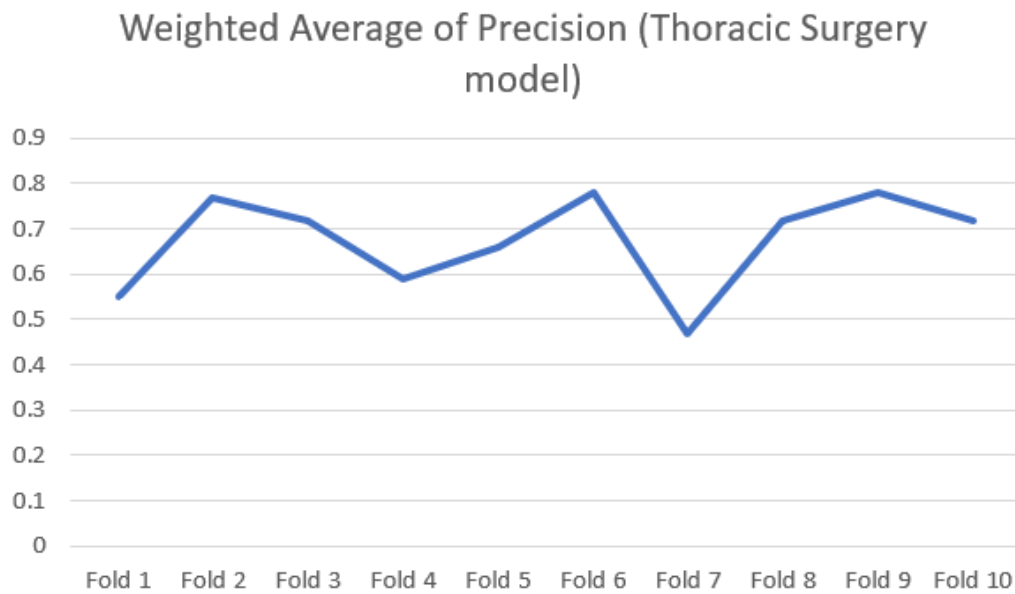
Έπειτα λοιπόν από την επιλογή του κατάλληλου σετ δεδομένων, οφείλαμε να υπολογίσουμε με τον ίδιο τρόπο και να παραθέσουμε και τις υπόλοιπες μετρικές και για το δεύτερο μοντέλο, αυτό της αξιολόγησης του ρίσκου επιβίωσης των ασθενών, με τα ίδια ή παρόμοια γραφήματα ώστε να μπορέσουμε να συγκρίνουμε τα δύο μοντέλα μας με ακρίβεια.



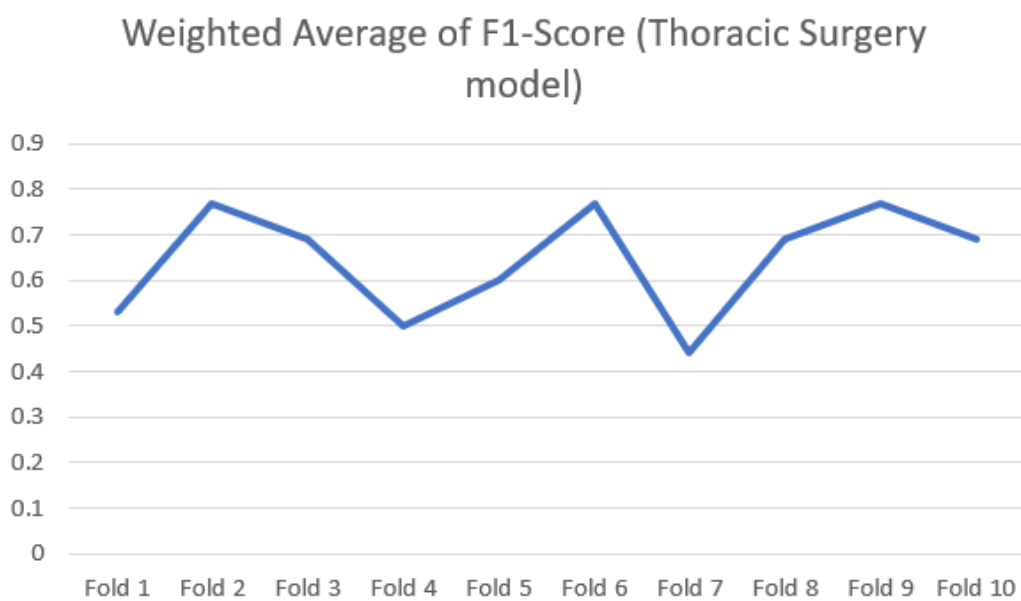
Εικόνα Δ'.9: Η ακρίβεια (accuracy) του μοντέλου αξιολόγησης.



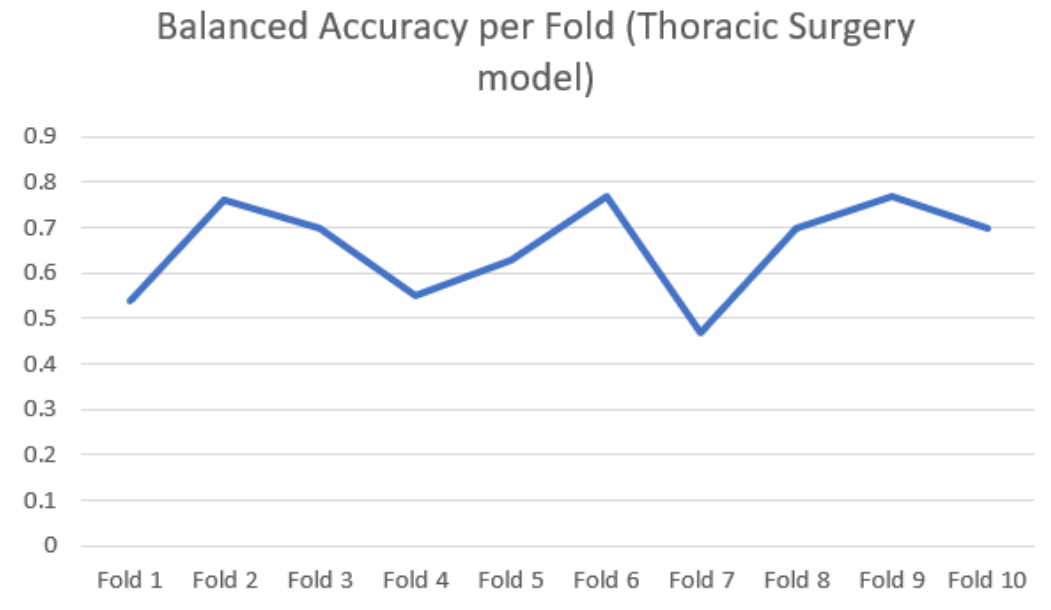
Εικόνα Δ'.10: Ο σταθμισμένος μέσος όρος της ανάκλησης (recall) του μοντέλου αξιολόγησης.



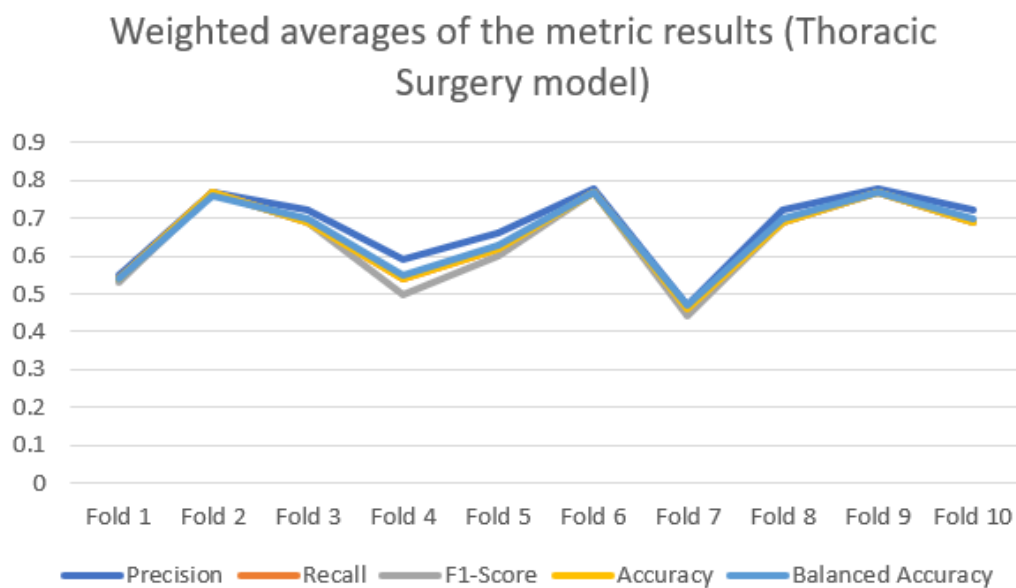
Εικόνα Δ'.11: Ο σταθμισμένος μέσος όρος της ακρίβειας (precision) του μοντέλου αξιολόγησης.



Εικόνα Δ'.12: Ο σταθμισμένος μέσος όρος του F1-score του μοντέλου αξιολόγησης.



Εικόνα Δ'.13: Η ισορροπημένη ακρίβεια (balanced accuracy) του μοντέλου αξιολόγησης.



Εικόνα Δ'.14: Το σύνολο των σταθμισμένων μέσων όρων όλων των μετρικών του μοντέλου αξιολόγησης.

Συγκριτικά με τα προηγούμενα αποτελέσματα, μπορούμε να διακρίνουμε ξεκάθαρα πως υπάρχουν διαφορές, με αυτά να είναι χαμηλότερα. Πιο συγκεκριμένα, για το δεύτερο μοντέλο της αξιολόγησης ρίσκου επιβίωσης ενός ασθενή με καρκίνο του πνεύμονα πετύχαμε μέσους όρους 65,4% εξίσου στα accuracy και recall, 67,6% στο precision, 64,5% στο f1-score και 66% στο balanced accuracy. Παρότι χαμηλότερα, τα αποτελέσματα παραμένουν ικανοποιητικά για τη δημιουργία του δεύτερου μοντέλου

μας, αν λάβει κανείς υπόψιν την δύσκολη «διαχειρισσιμότητα» του δεύτερου συνόλου δεδομένων, το οποίο περιείχε δύο κλάσσεις τελείως ανισόρροπες. Είμαστε λοιπόν έτοιμοι να προχωρήσουμε στην δημιουργία της εφαρμογής μας, έχοντας δύο ικανά μοντέλα για βάση, τα οποία μπορούν να αναγνωρίσουν με επιτυχία τα ζητούμενα μας.

Κεφάλαιο Ε΄

Υλοποίηση διαδικτυακής εφαρμογής

Μια διαδικτυακή εφαρμογή είναι μια εφαρμογή που λειτουργεί μέσω του Παγκόσμιου Ιστού (World Wide Web) και παρέχει δυνατότητες και λειτουργίες στους χρήστες μέσω ενός προγράμματος περιήγησης (browser). Αυτές οι εφαρμογές μπορεί να περιλαμβάνουν διάφορα είδη λειτουργικότητας, όπως αποθήκευση και επεξεργασία δεδομένων, επικοινωνία με άλλους χρήστες μέσω των κοινωνικών δικτύων, πρόσβαση σε πληροφορίες, αγορές προϊόντων και παροχή υπηρεσιών, παραδείγματα των οποίων χρησιμοποιούμε καθημερινά στην σημερινή εποχή. Οι εφαρμογές αυτές μπορούν να λειτουργήσουν σε οποιαδήποτε συσκευή με πρόσβαση στο διαδίκτυο, όπως υπολογιστές, κινητά τηλέφωνα και τάμπλετ και αποτελούν σημαντικό μέρος της σύγχρονης διαδικτυακής εμπειρίας. Η δημιουργία μιας διαδικτυακής εφαρμογής απαιτεί συνήθως πολλά βήματα και διάφορες τεχνολογίες. Ο τρόπος κατασκευής εξαρτάται από τη φύση της εφαρμογής και τις απαιτήσεις του έργου. Τα βασικά βήματα για την ανάπτυξη ενός συστήματος υποστήριξης αποφάσεων περιλαμβάνουν αρχικά την ανάπτυξη των τελικών μοντέλων που θα χρησιμοποιηθούν για την παραγωγή αποφάσεων. Αυτά θα αποτελέσουν τη βάση της εφαρμογής. Έπειτα έχουμε τον προγραμματισμό της εφαρμογής. Σε αυτό το στάδιο, αναπτύσσεται ο κώδικας της εφαρμογής με τη χρήση διαφόρων γλωσσών προγραμματισμού και πλατφόρμων ανάπτυξης, όπως οι HTML, CSS, JavaScript και άλλες. Η διαδικασία αυτή είναι γνωστή και ως front end development, καθώς κατά τη διάρκεια της δημιουργούμε το μπροστινό επίπεδο της εφαρμογής, αυτό με το οποίο θα έρχεται σε επαφή ο χρήστης κάθε φορά που θα επισκέπτεται την εφαρμογή. Τυπικά το επόμενο βήμα είναι η δημιουργία μιας βάσης δεδομένων όπου θα γίνεται η αποθήκευση και διαχείριση των δεδομένων της εφαρμογής. Κάποια από τα εργαλεία και τις γλώσσες που χρησιμοποιούνται συνήθως στο βήμα αυτό είναι τα MySQL, PostgreSQL και MongoDB. Στη δική μας περίπτωση αυτό το βήμα είναι περιττό καθώς η εφαρμογή μας δεν χρειάζεται προς στιγμήν να αποθηκεύσει κάποιο μεγάλο όγκο δεδομένων, ούτε να κρατήσει ιστορικό του ασθενή. Η πληροφορία είναι μικρή και μεταδίδεται απευθείας για επεξεργασία ώστε να παραχθεί η απόφαση και έπειτα διαγράφεται. Στη συνέχεια παράγεται το back end τμήμα της εφαρμογής, το οποίο είναι υπεύθυνο για την σύνδεση των μοντέλων με το κομμάτι της διεπαφής, ώστε να μπορέσουν αυτά να λάβουν την πληροφορία του χρήστη και να παράξουν και να εμφανίσουν το αποτέλεσμα. Για την ανάπτυξη του back end υπάρχουν διάφορα δημοφιλή εργαλεία που θα αναφέρουμε περαιτέρω παρακάτω. Στο τέλος έχουμε τη δοκιμή της λειτουργικότητας της εφαρμογής, όπου ελέγχουμε πως όλα έχουν δημιουργηθεί σωστά και δεν υπάρχουν λάθη, λειτουργικά ή ακόμη και λογικά. Κατά τη διάρκεια κάθε σταδίου χρησιμοποιήσαμε πλατφόρμες ανάπτυξης και εργαλεία που μας βοήθησαν στη διαχείριση του κώδικα, των εξαρτήσεων και των διαδικασιών ανάπτυξης. Το κύριο περιβάλλον ανάπτυξης (IDE) όμως στο οποίο βασιστήκαμε ήταν το Visual Studio

Code της Microsoft, το οποίο υποστηρίζει πολλαπλές γλώσσες και διαθέτει τα κατάλληλα plug in που μας ήταν απαραίτητα.

Ε΄.1 Κατασκευή μοντέλων λήψης αποφάσεων

Η δημιουργία μοντέλων μηχανικής μάθησης συνήθως ακολουθεί ένα συγκεκριμένο σύνολο μεθόδων και εργαλείων. Αρχικά, η διαδικασία ξεκινά με τη συλλογή και τον καθαρισμό των δεδομένων, προετοιμάζοντας τα για την είσοδο στο μοντέλο. Έπειτα, επιλέγεται ένας αλγόριθμος μηχανικής μάθησης και τα δεδομένα διαιρούνται σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Στη συνέχεια, το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και αξιολογείται στο σύνολο ελέγχου για να μετρήσει την απόδοσή του. Τα παραπάνω βήματα όχι μόνο τα αναφέραμε, αλλά τα είδαμε και στη πράξη σε προηγούμενα υποκεφάλαια. Τα εργαλεία που χρησιμοποιούνται συχνά περιλαμβάνουν τη γλώσσα προγραμματισμού Python και βιβλιοθήκες όπως η scikit-learn, η TensorFlow και η PyTorch. Άλλα εργαλεία περιλαμβάνουν περιβάλλοντα ανάπτυξης (IDEs) όπως τα Jupyter Notebooks αλλά και πλατφόρμες που δεν απαιτούν τη χρήση κώδικα όπως το Weka. Στην πράξη πειραματιστήκαμε και εμείς με διάφορα εργαλεία, όπως τα Weka και τα Jupyter Notebooks, αλλά καταλήξαμε στην δημιουργία των μοντέλων μέσω του Visual Studio Code, όπου με τη χρήση της γλώσσας προγραμματισμού Python δημιουργήσαμε όλα τα κομμάτια που αποτελούν την εφαρμογή μας, όπως θα δούμε και στο παράρτημα που θα αναρτηθεί παρακάτω, το οποίο θα περιέχει τον κώδικα που γράψαμε για κάθε βήμα με τη μορφή των snippets.

Πιο συγκεκριμένα, για την ανάπτυξη και των δύο μοντέλων μας χρησιμοποιήσαμε τις βιβλιοθήκες pandas, numpy, sklearn και pickle, οι οποίες είναι πολύ δημοφιλείς στον κόσμο της επιστημονικής υπολογιστικής και της ανάπτυξης λογισμικού με Python. Αρχικά, η βιβλιοθήκη pandas περιέχει δομές δεδομένων και λειτουργίες για την εύκολη και αποτελεσματική ανάλυση και επεξεργασία δεδομένων. Χρησιμοποιείται ευρέως για τη διαχείριση δεδομένων σε πίνακες και χρονοσειρές, επιτρέποντας την εκτέλεση διαφόρων λειτουργιών όπως φιλτράρισμα, επιλογή, ομαδοποίηση και συνδυασμός δεδομένων. Η βιβλιοθήκη NumPy από την άλλη παρέχει υψηλής απόδοσης πίνακες και προσδιορισμένες λειτουργίες για αριθμητικούς υπολογισμούς. Χρησιμοποιείται ευρέως για την εκτέλεση πράξεων μεγάλου όγκου δεδομένων και την ανάλυση αριθμητικών δεδομένων. Η βιβλιοθήκη scikit-learn μας προσδίδει ένα ευρύ φάσμα αλγορίθμων μηχανικής μάθησης και εργαλείων για την ανάπτυξη μοντέλων μηχανικής μάθησης. Περιλαμβάνει αλγορίθμους για ταξινόμηση, παλινδρόμηση, συσταδοποίηση, εκτίμηση, επεξεργασία δεδομένων και άλλα. Από αυτή χρησιμοποιήσαμε τη μέθοδο train_test_split, η οποία βρίσκεται στο πακέτο model_selection, για να διαιρέσουμε τα σύνολα δεδομένων μας σε υποσύνολα εκπαίδευσης και ελέγχου κατά τη δημιουργία των μοντέλων. Αυτό μας επιτρέπει να αξιολογήσουμε την απόδοση ενός μοντέλου μηχανικής μάθησης σε ξεχωριστά σύνολα δεδομένων εκπαίδευσης και ελέγχου, χωρίς να επηρεάσουμε την ανεξαρτησία των δεδομένων ελέγχου

από τα δεδομένα εκπαίδευσης. Αυτό βοηθά στην αποφυγή του overfitting (υπερεκπαίδευσης) και στη διασφάλιση της γενικευσιμότητας του μοντέλου. Τέλος, με τη χρήση της βιβλιοθήκης pickle, η οποία παρέχει λειτουργίες για την αποθήκευση και τη φόρτωση δεδομένων Python σε αρχεία, αποθηκεύσαμε τα μοντέλα μηχανικής μάθησης μετά την εκπαίδευσή τους, προκειμένου να μπορούν να χρησιμοποιηθούν σε μελλοντικές εφαρμογές χωρίς την ανάγκη εκ νέου εκπαίδευσης.

Ε΄.2 Δημιουργία επιπέδου διεπαφής (Front-end)

Το "front-end" στην ανάπτυξη εφαρμογών αναφέρεται στο μέρος μιας εφαρμογής ή ιστοσελίδας που αλληλεπιδρά με τον χρήστη και προβάλλει πληροφορίες. Αυτό περιλαμβάνει τα γραφικά, τα κουμπιά, τις φόρμες και άλλα στοιχεία που βλέπει και αλληλεπιδρά με αυτά ο χρήστης κατά τη χρήση της εφαρμογής. Η καλή σχεδίαση και η χρήση βέλτιστων πρακτικών front-end είναι σημαντικές τόσο για μια ευχάριστη και αποτελεσματική εμπειρία του χρήστη, όσο και για την επιτυχημένη συλλογή των πληροφοριών που χρειάζεται ο εκάστοτε δημιουργός για να παράξει ένα αποτέλεσμα. Για την ανάπτυξη του επιπέδου αυτού χρησιμοποιούνται συνήθως οι HTML, CSS και JavaScript, ενώ υπάρχουν και άλλες γλώσσες προγραμματισμού ή frameworks ήδη υπάρχοντων όπως το React, το Angular και το Swift που προσφέρουν περισσότερες επιλογές κατά τη δημιουργία δυναμικών διαδικτυακών εφαρμογών. Στα πλαίσια αυτής της πτυχιακής εργασίας βέβαια δεν θεωρήσαμε αναγκαία τη χρήση «bleeding edge» τεχνολογιών για την δημιουργία της εφαρμογής μας, οπότε περιοριστήκαμε στις πρώτες τρεις γλώσσες που αναφέραμε.

Πιο συγκεκριμένα, η πρώτη γλώσσα που χρησιμοποιήσαμε είναι η HTML (HyperText Markup Language) που χρησιμοποιείται για τη δημιουργία και τη διαμόρφωση ιστοσελίδων. Αποτελεί τον βασικό τρόπο με τον οποίο οι ιστοσελίδες παρουσιάζουν πληροφορίες, δομούνται και επικοινωνούν με τους χρήστες. Η HTML χρησιμοποιεί "tags" ή ετικέτες για να περιγράψει τη δομή του περιεχομένου της ιστοσελίδας, όπως κείμενο, εικόνες, συνδέσμους και πολλά άλλα. Οι ετικέτες HTML είναι συχνά τοποθετημένες μεταξύ αγκύλων και χρησιμοποιούνται για να δηλώσουν τον τύπο του περιεχομένου. Στην εφαρμογή μας, την οποία θα παρουσιάσουμε πλήρως με φωτογραφικό υλικό σε επόμενο υποκεφάλαιο, χρησιμοποιήσαμε κυρίως ετικέτες για τη παρουσίαση κειμένου αλλά και φόρμες, οι οποίες ήταν απαραίτητες για την συλλογή των δεδομένων που θα τροφοδοτήσουν τα μοντέλα μας. Συνολικά δημιουργήσαμε έξι διαφορετικά αρχεία HTML, ένα για κάθε «εκδοχή» της εφαρμογής μας (κεντρική σελίδα, φόρμες διάγνωσης και αξιολόγησης καρκίνου, σελίδες αποτελεσμάτων), τα οποία όμως όπως θα δούμε παρακάτω περιέχουν και άλλες γλώσσες.

Σε συνδυασμό με την HTML χρησιμοποιήσαμε και τη CSS (Cascading Style Sheets), η οποία είναι μια γλώσσα στυλιστικών φύλλων που χρησιμοποιείται για την παρουσίαση και τη μορφοποίηση του περιεχομένου που δηλώνεται σε μια ιστοσελίδα HTML. Η CSS ορίζει την εμφάνιση, το layout και το

στυλ των στοιχείων HTML, όπως κείμενο, εικόνες, πίνακες και άλλα. Χρησιμοποιώντας τη CSS, τροποποιήσαμε το χρώμα, τη γραμματοσειρά, το περίγραμμα, το πλάτος, το ύψος και πολλά άλλα χαρακτηριστικά των στοιχείων που θα παρουσιαστούν στην εφαρμογή μας ώστε να κάνουμε την εμφάνιση της πιο ελκυστική, ευανάγνωστη αλλά και συνάμα επαγγελματική. Δεν μείναμε όμως εκεί, καθώς θεωρήσαμε απαραίτητη και τη χρήση της JavaScript, μιας προγραμματιστικής γλώσσας που χρησιμοποιείται για την προσθήκη δυναμικής λειτουργικότητας και αλληλεπίδρασης σε ιστοσελίδες. Είναι μια από τις τρεις κύριες τεχνολογίες που χρησιμοποιούνται για τη δημιουργία ιστοσελίδων, μαζί με την HTML και την CSS. Τη χρησιμοποιήσαμε για να δημιουργήσουμε διάφορα δυναμικά στοιχεία στη σελίδα, όπως τις αλληλεπιδραστικές μας φόρμες, animation κατά την αλλαγή μεταξύ των σελίδων, καθώς και αλλαγές στο περιεχόμενο της σελίδας χωρίς ανανέωση. Συνεπώς, συνδυάζοντας τα τρία αυτά εργαλεία μέσα στα έξι διαφορετικά αρχεία με τις σελίδες μας, καταφέραμε να ολοκληρώσουμε τη δημιουργία του επιπέδου διεπαφής της εφαρμογής μας, κάτι που μας επιτρέπει να προχωρήσουμε στο επόμενο βήμα, τη δημιουργία του back-end επιπέδου που θα ενώσει τα μοντέλα μας με την πληροφορία που παρέχει το front-end επίπεδο.

Ε΄.3 Ανάπτυξη back-end επιπέδου

Η ανάπτυξη του back end επιπέδου αφορά τη δημιουργία και τη συντήρηση του μέρους της μας εφαρμογής που δεν είναι ορατό στον χρήστη, αλλά παρέχει τη λειτουργικότητα που απαιτείται για την εκτέλεση της εφαρμογής. Αυτό περιλαμβάνει τη διαχείριση των αιτημάτων των χρηστών, την επεξεργασία δεδομένων και τη δημιουργία δυναμικού περιεχομένου που παρουσιάζεται στον χρήστη. Για την ανάπτυξη αυτού του επιπέδου υπάρχουν πολλά καινοτόμα εργαλεία αλλά και πλαίσια εργασίας (frameworks) που καθιστούν τον προγραμματισμό πολύπλοκων εφαρμογών απλό, με δημοφιλή παραδείγματα να είναι το Django της Python, το Laravel της PHP και άλλα. Επειδή λοιπόν και εμείς χρησιμοποιήσαμε την Python για την δημιουργία των μοντέλων μας και μας ήταν γνώριμη, αποφασίσαμε να χρησιμοποιήσουμε ένα πλαίσιο εργασίας βασισμένο σε αυτή, με κύριες επιλογές τα Django και Flask, τα οποία είναι δύο από τα δημοφιλέστερα frameworks για την ανάπτυξη web εφαρμογών στη γλώσσα προγραμματισμού Python. Και τα δύο πλαίσια είναι ισχυρά και ευέλικτα, αλλά υπάρχουν κάποιες βασικές διαφορές μεταξύ τους. Πιο συγκεκριμένα, το Django είναι ένα ολοκληρωμένο framework που παρέχει πολλές έτοιμες λύσεις και λειτουργίες για την ανάπτυξη web εφαρμογών, έχει ενσωματωμένες λύσεις για τη διαχείριση των χρηστών και των διαχειριστικών περιβαλλόντων και χρησιμοποιείται συνήθως για ανάπτυξη μεγάλων εφαρμογών με σύνθετες απαιτήσεις, ενώ το Flask είναι πιο ελαφρύ και ευέλικτο, δεν παρέχει πολλές έτοιμες λύσεις όπως το Django, αλλά επιτρέπει τη διαμόρφωση μιας εφαρμογής ανάλογα με τις εκάστοτε ανάγκες. Το Flask είναι καταλληλότερο για μικρότερες εφαρμογές που είναι πιο απλές στη δομή τους και για αυτό το λόγο το επιλέξαμε σαν το κατάλληλο framework για την περίπτωση μας.

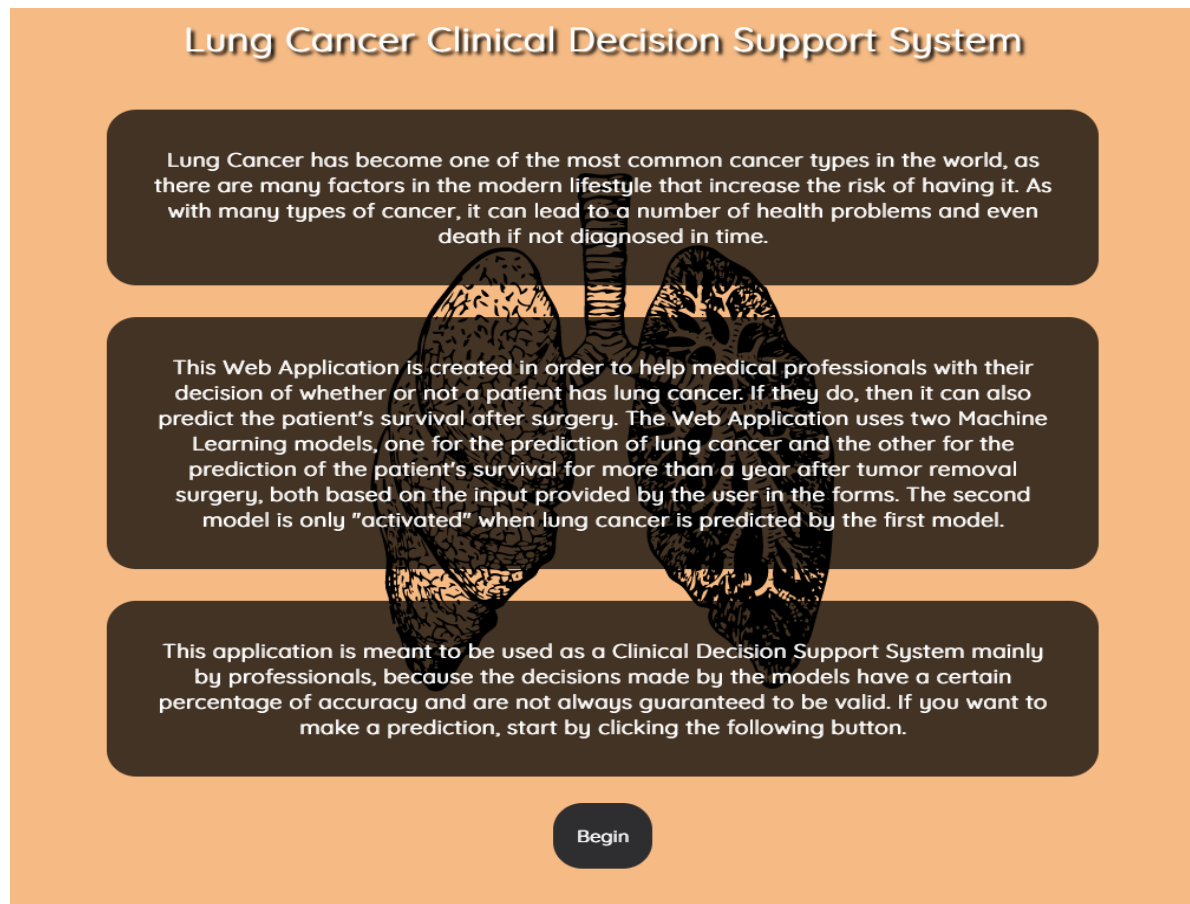
Για την δημιουργία του back end χρειάζεται να εισάγουμε κάποιες βιβλιοθήκες στον κώδικα που θα μας βοηθήσουν μετέπειτα. Αυτές είναι η pickle για την διαχείριση των μοντέλων, η numpy για την διαχείριση των δεδομένων που θα παραχθούν, αλλά και η flask η οποία περιέχει έτοιμες συναρτήσεις όπως οι request, render_template, redirect και url_for τις οποίες εισάγουμε για να εκμεταλλευτούμε στη συνέχεια. Αρχικά, αφού καθορίσουμε την εφαρμογή μας και κάνουμε load τα μοντέλα μας, θα δημιουργήσουμε τρεις «δίοδους», κάθε μία από τις οποίες θα περιέχει μια συνάρτηση που θα αφορά διαφορετικό τομέα της εφαρμογής. Αναλυτικότερα, η πρώτη δίοδος μας είναι αυτή που θα βλέπει ο χρήστης με το που εισέρχεται στην εφαρμογή, το homepage δηλαδή, μέσα στο οποίο παρατίθενται πληροφορίες για τη λειτουργία της εφαρμογής και το σκοπό δημιουργίας της. Από εκεί ο χρήστης έχει μόνο μία επιλογή, να συνεχίσει στην επόμενη σελίδα, αυτή που περιέχει το πρώτο μοντέλο της διάγνωσης του καρκίνου του πνεύμονα και με αυτή του την επιλογή ενεργοποιείται η μέθοδος “GET”, η οποία με απλά λόγια ζητάει από τον server στον οποίο λειτουργεί η εφαρμογή την ανάκτηση κάποιων δεδομένων. Με την «ενεργοποίηση» της μεθόδου αυτής επιστρέφεται στον χρήστη η δεύτερη δίοδος της εφαρμογής, δηλαδή η εμφάνιση της φόρμας συμπλήρωσης συμπτωμάτων που δημιουργήσαμε. Όταν ο χρήστης ολοκληρώσει την εισαγωγή των δεδομένων και επιλέξει να συνεχίσει ενεργοποιείται η επόμενη μέθοδος “POST”, μέσω της οποίας αποστέλλονται τα δεδομένα που εισήγαγε ο χρήστης πίσω στον server για επεξεργασία και ανάλογα με το αποτέλεσμα του επιστρέφεται μια διαφορετική σελίδα από αυτές που δημιουργήσαμε, είτε αυτή του θετικού αποτελέσματος, είτε αυτή του αρνητικού. Εδώ να σημειώσουμε πως κάθε σελίδα παρουσιάζεται στον χρήστη με τη χρήση της render_template συνάρτησης του Flask. Στην περίπτωση αρνητικού αποτελέσματος δίνεται η επιλογή επιστροφής στην αρχική σελίδα της εφαρμογής, ενώ στην περίπτωση θετικού αποτελέσματος επιστρέφεται στον χρήστη το δεύτερο σκέλος της εφαρμογής ή η επόμενη δίοδος, δηλαδή η φόρμα της αξιολόγησης του ρίσκου επιβίωσης του. Εδώ ακολουθείται η ίδια διαδικασία όπως και πριν, με τη μόνη διαφορά πως για το κάθε αποτέλεσμα του επιστρέφουμε μια διαφορετική σελίδα που του εξηγεί την απόφαση που έλαβε το μοντέλο. Από εκεί και πέρα, ή μόνη επιλογή που δίνεται στον χρήστη είναι να επιστρέψει στο πρώτο στάδιο της εφαρμογής, αυτό της διάγνωσης του καρκίνου. Με αυτό τον τρόπο καταφέραμε να δημιουργήσαμε μια απλή διαδικτυακή εφαρμογή η οποία διαθέτει τις λειτουργίες ενός συστήματος υποστήριξης αποφάσεων με πολύ βατά βήματα. Ο κώδικας που δημιουργήσαμε θα παρατεθεί στο σύνολο του στο παρακάτω παράρτημα, ενώ το επόμενο βήμα θα είναι να περιηγηθούμε στην εφαρμογή με τη χρήση εικόνων και περιγραφών, ώστε αυτό να αποτελέσει ένα είδος εγχειρίδιου της.

Ε΄.4 Παρουσίαση διαδικτυακής εφαρμογής

Μετά την λεπτομερή ανάλυση του τρόπου δημιουργίας της εφαρμογής καταλήξαμε σε ένα τελικό προϊόν ικανό για αυτά που στοχεύαμε, δηλαδή την ακριβή διάγνωση και αξιολόγηση του καρκίνου του πνεύμονα και του ρίσκου επιβίωσης του ασθενή. Παράλληλα δημιουργήσαμε ένα εύχρηστο προϊόν και

απλό στη χρήση, γεγονός που επιτρέπει τη χρήση του από οποιονδήποτε διαθέτει σύνδεση στο ίντερνέτ, ανεξαρτήτως εκπαίδευσης ή οικονομικού υπόβαθρου. Στο υποκεφάλαιο αυτό θα παρουσιάσουμε βήμα βήμα την διεπαφή του χρήστη με την εφαρμογή από την στιγμή που θα την ανοίξει, μέχρι και τη στιγμή που θα την κλείσει. Παράλληλα θα αναλύσουμε και τις διάφορες πτυχές των αποτελεσμάτων που μπορεί να δεχθεί ο χρήστης, καθώς και τη σημασία κάθε δεδομένου που του ζητείται.

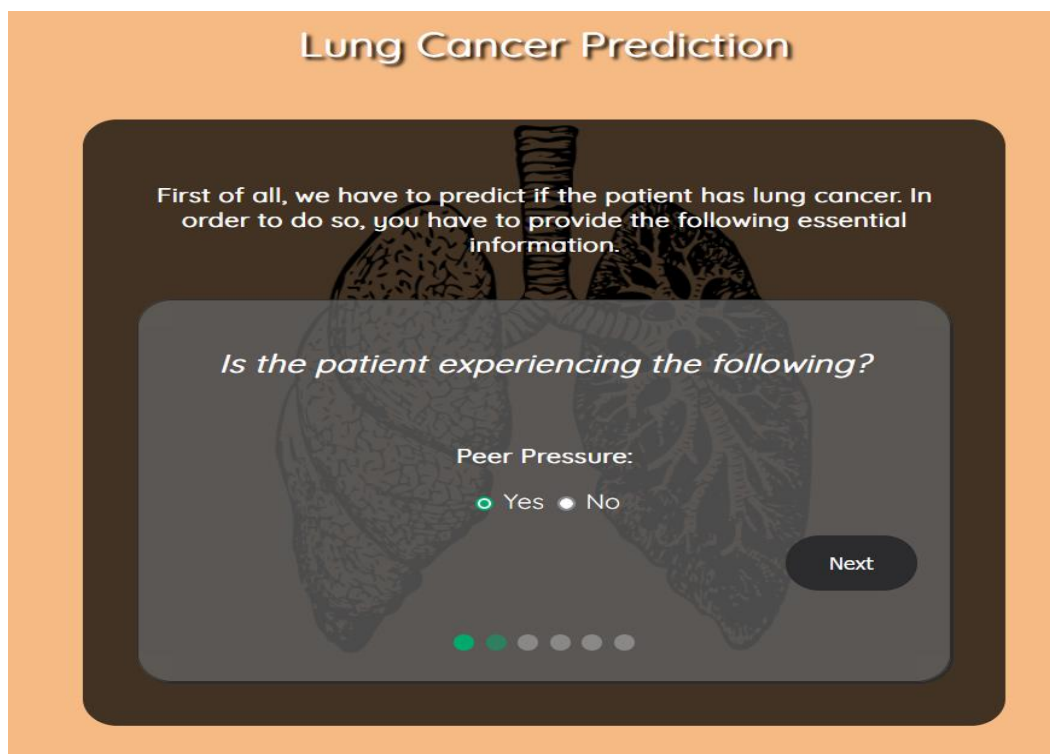
Το πρώτο πράγμα που παρουσιάζεται στον χρήστη σε κάθε εφαρμογή που χρησιμοποιεί είναι μια αρχική σελίδα, στην οποία παρατίθενται χρήσιμες πληροφορίες για τον λόγο ύπαρξης της εφαρμογής, το πρόβλημα που διευθετεί, καθώς και διάφορες άλλες πληροφορίες για το τι πρόκειται να ακολουθήσει. Έτσι και στη δική μας δημιουργήσαμε ένα homepage με συνοπτικές πληροφορίες τόσο για την ασθένεια, όσο και τον σκοπό της δημιουργίας της αλλά και τους πιθανούς χρήστες, όπως μπορούμε να δούμε και στην παρακάτω εικόνα. Επίσης υπάρχει σε αυτή και το κουμπί εκκίνησης της λειτουργίας της, η οποία οδηγεί στο πρώτο μοντέλο διάγνωσης του καρκίνου του πνεύμονα εάν χρησιμοποιηθεί.



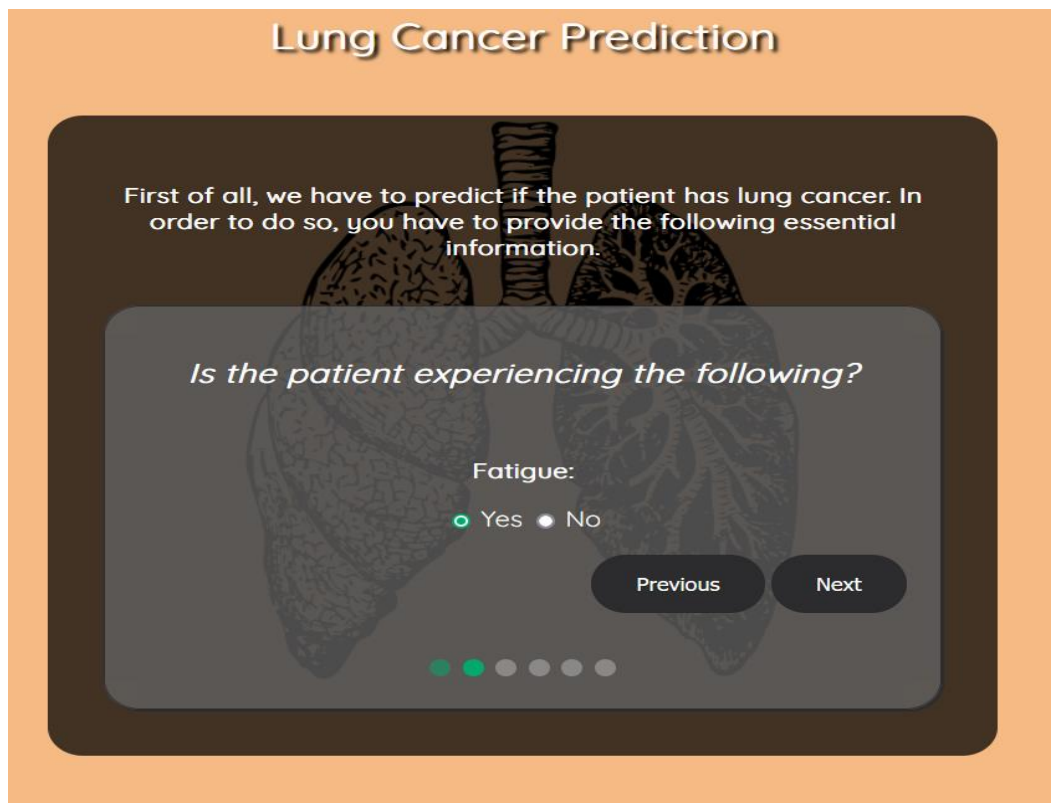
Εικόνα Ε΄.1: Η αρχική σελίδα της διαδικτυακής μας εφαρμογής.

Μετά τη χρήση του κουμπιού εκκίνησης (“Begin”) που παρατέθηκε στην αρχική σελίδα, ο χρήστης οδηγείται στο πρώτο σκέλος της εφαρμογής, τη διάγνωση του καρκίνου του πνεύμονα. Για να γίνει όμως η διάγνωση, ο χρήστης καλείται να εισάγει κάποια δεδομένα στην φόρμα που έχουμε

δημιουργήσει, τα οποία αφορούν συμπτώματα που μπορεί να αισθάνεται τη δεδομένη χρονική στιγμή. Οι επιλογές που μπορεί να επιλέξει είναι τύπου Boolean, δηλαδή θετικές ή αρνητικές για κάθε ένα από τα έξι συμπτώματα που του παρουσιάζονται. Αφού επιλέξει την απάντηση του, του δίνεται η δυνατότητα συνέχειας στην επόμενη ερώτηση ή η επιστροφή στην προηγούμενη, σε περίπτωση που αλλάξει γνώμη ή κάνει κάποιο λάθος. Αυτό επιτυγχάνεται με τη χρήση κουμπιών συνέχειας (“Next”) ή επιστροφής (“Previous”), όπως θα δούμε παρακάτω. Αφού απαντήσει και στην τελευταία ερώτηση όμως, το κουμπί συνέχειας αλλάζει σε κουμπί υποβολής των δεδομένων (“Submit”), ώστε τα δεδομένα να εισαχθούν στο μοντέλο και να παραχθεί η απόφαση από αυτό. Πιο αναλυτικά, τα συμπτώματα για τα οποία καλείται ο χρήστης να επιλέξει μια απάντηση είναι τα εξής: το αν δέχεται πίεση από ομοτίμους του (peer pressure), η αίσθηση κούρασης (fatigue), η ύπαρξη αλλεργιών (allergy), ο συριγμός (wheezing), δηλαδή μια αναπνοή που συνέεται με έναν έντονο ήχο σαν αυτό σφυρίχτρας και συνήθως υποδηλώνει την παρουσία αναπνευστικών προβλημάτων, την κατανάλωση αλκοόλ (alcohol consuming) και τέλος την δυσκολία κατάποσης (swallowing difficulty). Στην περίπτωση που κάποια επιλογή είναι πιο εξειδικευμένη, όπως για παράδειγμα αυτή του wheezing, έχουμε δημιουργήσει στην εφαρμογή κουτάκια πληροφοριών που φανερώνουν την επεξήγηση, εάν ο χρήστης περάσει τον κέρσορα του από αυτά. Τέλος, εάν η απόφαση του μοντέλου είναι αρνητική, τότε του δίνεται η επιλογή της επιστροφής στην αρχική σελίδα με τη χρήση ενός κουμπιού (“Go Back”). Οι εικόνες που θα ακολουθήσουν παρουσιάζουν τόσο τις ερωτήσεις που αφορούν το πρώτο μοντέλο με τη σειρά που θα εμφανιστούν στον χρήστη, όσο και την περίπτωση αρνητικής απάντησης από το μοντέλο. Την περίπτωση της θετικής απάντησης θα την αναλύσουμε μετέπειτα στο δεύτερο σκέλος της εφαρμογής.



Εικόνα Ε΄.2: Η αρχή της φόρμας με την επιλογή της πίεσης από ομοτίμους.



Lung Cancer Prediction

First of all, we have to predict if the patient has lung cancer. In order to do so, you have to provide the following essential information.

Is the patient experiencing the following?

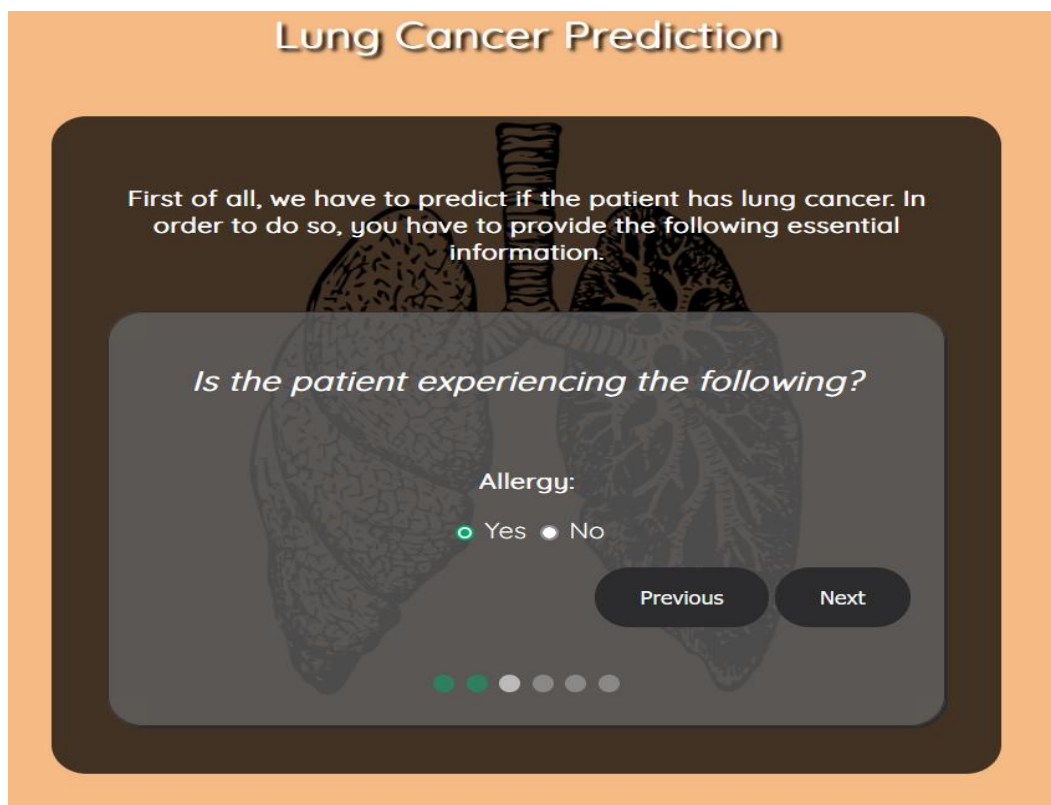
Fatigue:

☒ Yes ☐ No

Previous Next

Progress indicator: 1st step active, 2nd step active, 3rd step, 4th step, 5th step, 6th step.

Εικόνα Ε΄.3: Η δεύτερη επιλογή που αφορά το αίσθημα κούρασης.



Lung Cancer Prediction

First of all, we have to predict if the patient has lung cancer. In order to do so, you have to provide the following essential information.

Is the patient experiencing the following?

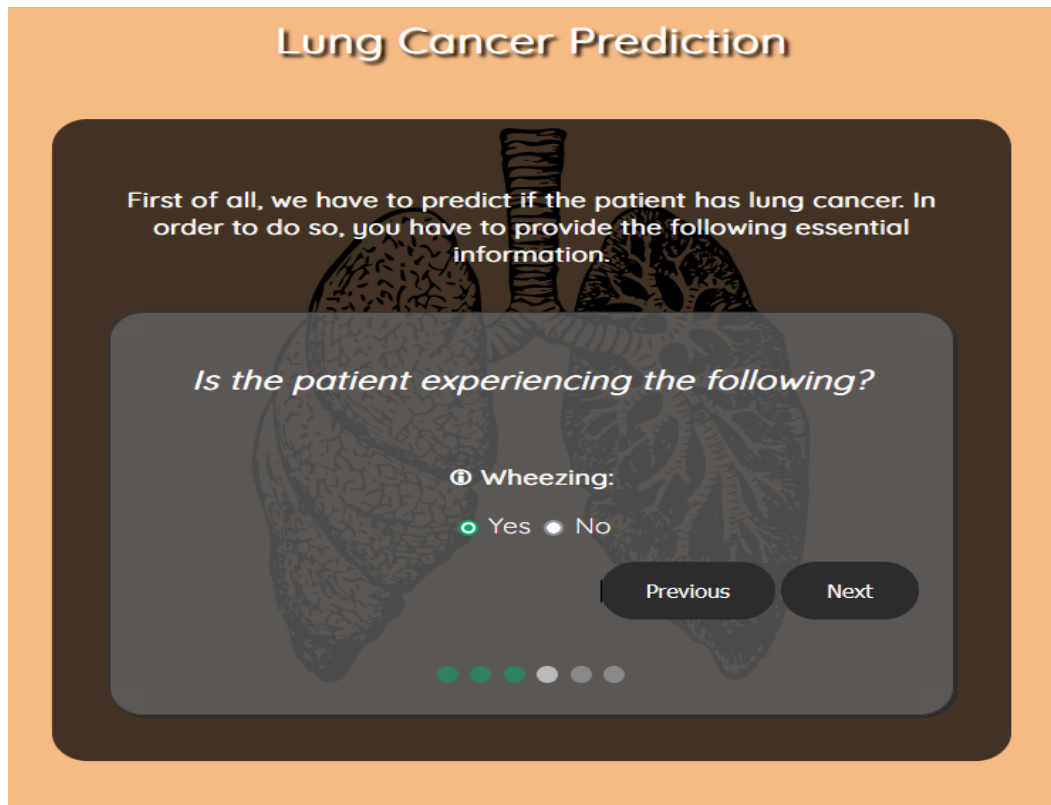
Allergy:

☒ Yes ☐ No

Previous Next

Progress indicator: 1st step active, 2nd step active, 3rd step, 4th step, 5th step, 6th step.

Εικόνα Ε΄.4: Η τρίτη επιλογή που αφορά την ύπαρξη αλλεργιών.



Lung Cancer Prediction

First of all, we have to predict if the patient has lung cancer. In order to do so, you have to provide the following essential information.

Is the patient experiencing the following?

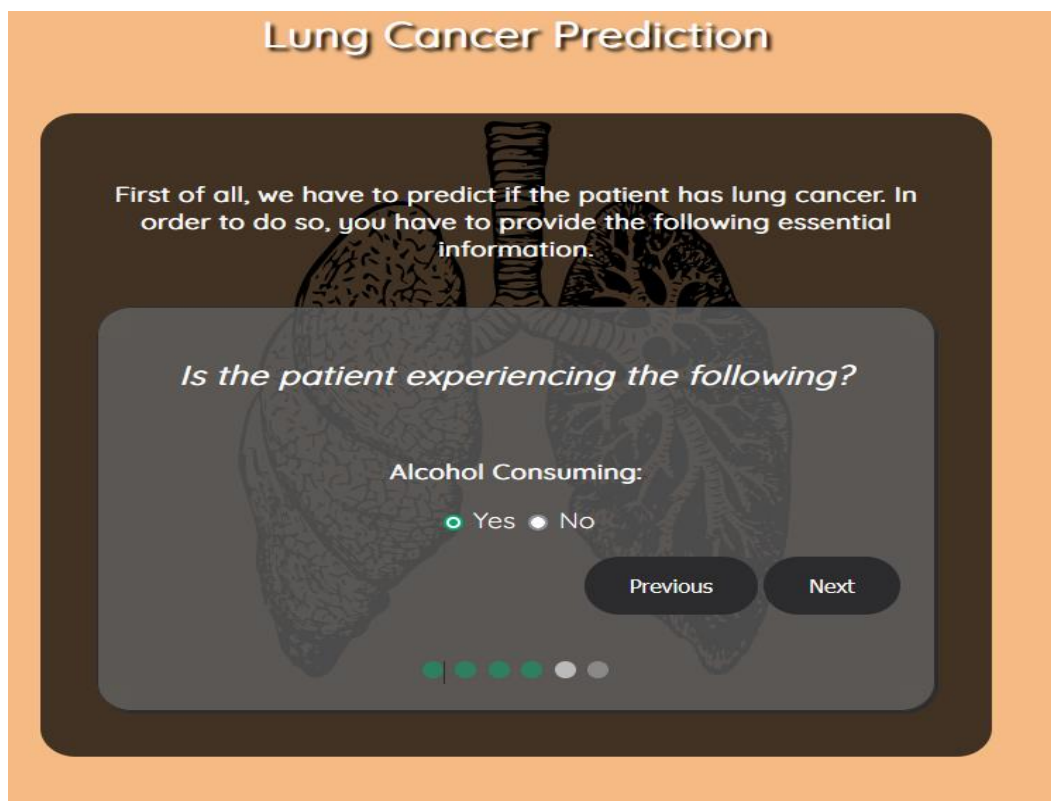
④ Wheezing:

☒ Yes ☐ No

Previous Next

Progress indicator: 4 green dots, 2 grey dots.

Εικόνα Ε΄.5: Η τέταρτη επιλογή που αφορά την ύπαρξη συριγμού.



Lung Cancer Prediction

First of all, we have to predict if the patient has lung cancer. In order to do so, you have to provide the following essential information.

Is the patient experiencing the following?

Alcohol Consuming:

☒ Yes ☐ No

Previous Next

Progress indicator: 5 green dots, 1 grey dot.

Εικόνα Ε΄.6: Η πέμπτη επιλογή που αφορά την κατανάλωση αλκοόλ.

Lung Cancer Prediction

First of all, we have to predict if the patient has lung cancer. In order to do so, you have to provide the following essential information.

Is the patient experiencing the following?

Swallowing Difficulty:

☒ Yes ☐ No

Previous Submit

Progress indicator: 6 dots, 5th dot is active.

Εικόνα Ε΄.7: Η έκτη επιλογή που αφορά την δυσκολία κατάποσης.

No Lung Cancer detected

Based on the information provided in the previous form, the "subject" does not qualify as a potential Lung Cancer patient. Do not forget that this decision could be a false negative, as there is always room for error.

If you wish to make another entry, please click the following button.

Go back

Εικόνα Ε΄.8: Η περίπτωση όπου το αποτέλεσμα είναι αρνητικό.

Έχοντας δει λοιπόν τον τρόπο λειτουργίας του πρώτου σκέλους της διαδικτυακής μας εφαρμογής αλλά και ένα πιθανό παραγόμενο αποτέλεσμα αυτής, μπορούμε να προχωρήσουμε στην ανάλυση του επόμενου. Υπενθυμίζουμε πως το δεύτερο σκέλος της εφαρμογής ενεργοποιείται μόνο όταν έχουμε θετική πρόβλεψη κατά τη διάγνωση του καρκίνου του πνεύμονα, καθώς δεν υφίσταται αξιολόγηση του ρίσκου επιβίωσης χωρίς την ύπαρξη της ασθένειας. Όπως θα παρατηρήσουμε λοιπόν και στις παρακάτω εικόνες, όταν εντοπίζεται θετικό αποτέλεσμα από το πρώτο μοντέλο, παράγει η εφαρμογή μια σελίδα στην οποία αναφέρει το θετικό αποτέλεσμα, αλλά παρέχει ταυτόχρονα και τη δεύτερη φόρμα συμπλήρωσης, ώστε να συλλέξει από τον χρήστη τα νέα δεδομένα που χρειάζεται ώστε να λάβει το επόμενο μοντέλο μια απόφαση. Η δομή και εμφάνιση της φόρμας είναι ολόιδια με την προηγούμενη, με την μόνη αλλαγή να βρίσκεται στα ζητούμενα της φόρμας, τα οποία αλλάζουν και είναι ελαφρώς πιο εξειδικευμένα. Λόγω της φύσης των δεδομένων αυτών ίσως να καθίσταται δύσκολη η χρήση του δεύτερου σκέλους της εφαρμογής από έναν χρήστη που είναι ασθενής και όχι εξειδικευμένος επιστήμονας, εάν πρώτα δεν έχει λάβει τα σχετικά δεδομένα από τον ιατρό του. Αναλυτικότερα, τα δεδομένα που ζητά είναι η διάγνωση (“Diagnosis”) η οποία είναι ένας ειδικός συνδυασμός κωδικών ICD-10 και έχει επτά διαφορετικές επιλογές για πρωτοπαθείς, δευτεροπαθείς καθώς και πολλαπλούς όγκους, εάν αυτοί υπάρχουν, η δύσπνοια πριν την επέμβαση (“Dyspnoea before surgery”), το μέγεθος του όγκου (“Size of original tumor”) το οποίο υπολογίζεται βάση το σύστημα TNM που χρησιμοποιείται για την περιγραφή του μεγέθους ενός όγκου. Αυτό διαθέτει τέσσερις διαφορετικές επιλογές (OC11, OC12, OC13, OC14) με το “OC11” να είναι το μικρότερο και το “OC14” το μεγαλύτερο. Τέλος, έχουμε την ύπαρξη σακχαρώδη διαβήτη δεύτερου τύπου (“Type 2 Diabetes Mellitus”) καθώς και την ύπαρξη άσθματος στον ασθενή (“Asthma”). Μετά την επιλογή του κουμπιού υποβολής των δεδομένων υπάρχουν δύο διαφορετικές εκδοχές αποτελέσματος. Είτε η απόφαση με βάση την κατάσταση του ασθενή θα τον κρίνει ικανό να επιβιώσει για παραπάνω από έναν χρόνο μετά το χειρουργείο, είτε το αντίθετο. Για κάθε πιθανό αποτέλεσμα θα γίνει ανακατεύθυνση του χρήστη στην αντίστοιχη σελίδα που θα του περιγράψει την απόφαση και θα του δίνει την επιλογή να επιστρέψει στο πρώτο σκέλος της εφαρμογής με τη χρήση ενός κουμπιού επιστροφής (“Return”), όπως θα δούμε και παρακάτω.

Surgery Survival Prediction

Unfortunately, the patient has Lung Cancer.

You can now estimate the survival of the patient for more than a year in case they choose to remove the cancerous tumour with surgery. In order to do so, you need to follow the same process as before, but this time by providing more specialized information about the patient in the following form.

Please fill in the data below:

① Diagnosis:

☒ DGN1 ☐ DGN2 ☐ DGN3 ☐ DGN4 ☐ DGN5 ☐ DGN6 ☐ DGN8

Next

Progress indicator: 1st dot is filled, 2nd-5th are empty.

Εικόνα Ε΄.9: Εκκίνηση του δεύτερου σκέλους έπειτα από θετικό αποτέλεσμα – επιλογή διάγνωσης.

Surgery Survival Prediction

Unfortunately, the patient has Lung Cancer.

You can now estimate the survival of the patient for more than a year in case they choose to remove the cancerous tumour with surgery. In order to do so, you need to follow the same process as before, but this time by providing more specialized information about the patient in the following form.

Please fill in the data below:

Dyspnoea before surgery:

☒ Yes ☐ No

Previous Next

Progress indicator: 1st dot is filled, 2nd-5th are empty.

Εικόνα Ε΄.10: Η δεύτερη επιλογή της δύσπνοιας πριν την επέμβαση.

Surgery Survival Prediction

Unfortunately, the patient has Lung Cancer.

You can now estimate the survival of the patient for more than a year in case they choose to remove the cancerous tumour with surgery. In order to do so, you need to follow the same process as before, but this time by providing more specialized information about the patient in the following form.

Please fill in the data below:

Size of original tumour:

☒ OC11 ☐ OC12 ☐ OC13 ☐ OC14

Previous Next

Εικόνα Ε΄.11: Η τρίτη επιλογή του μεγέθους του αρχικού όγκου.

Surgery Survival Prediction

Unfortunately, the patient has Lung Cancer.

You can now estimate the survival of the patient for more than a year in case they choose to remove the cancerous tumour with surgery. In order to do so, you need to follow the same process as before, but this time by providing more specialized information about the patient in the following form.

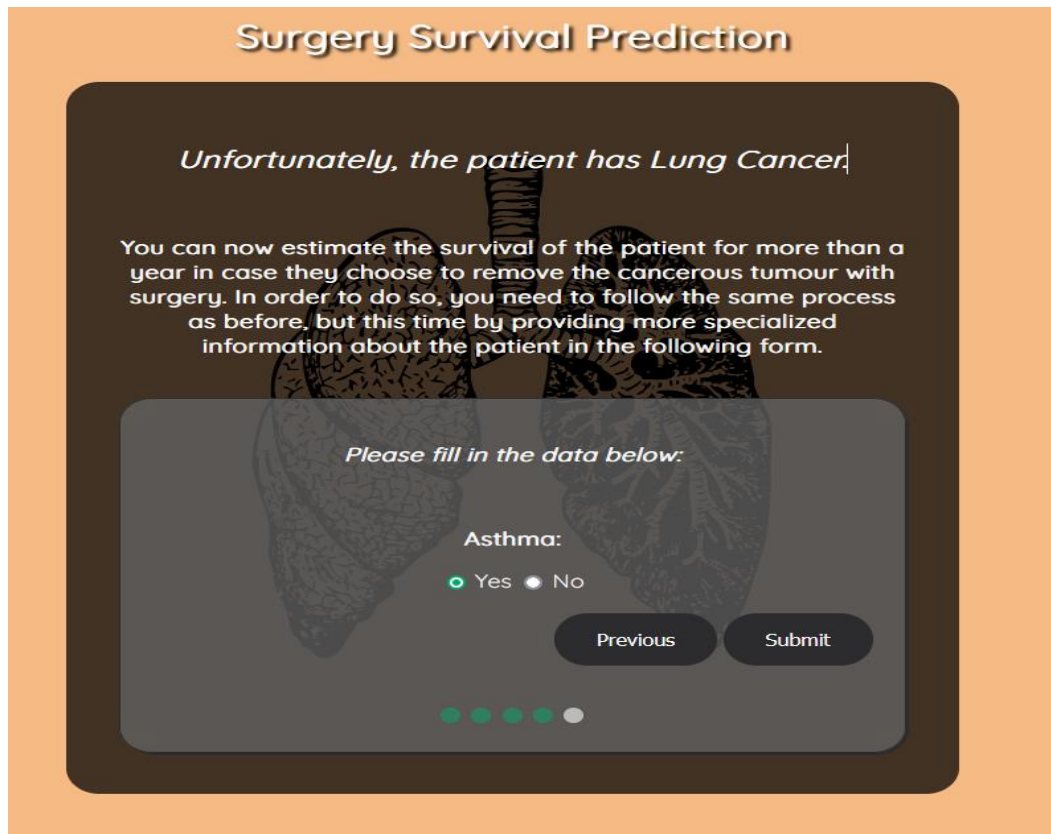
Please fill in the data below:

Type 2 Diabetes Mellitus:

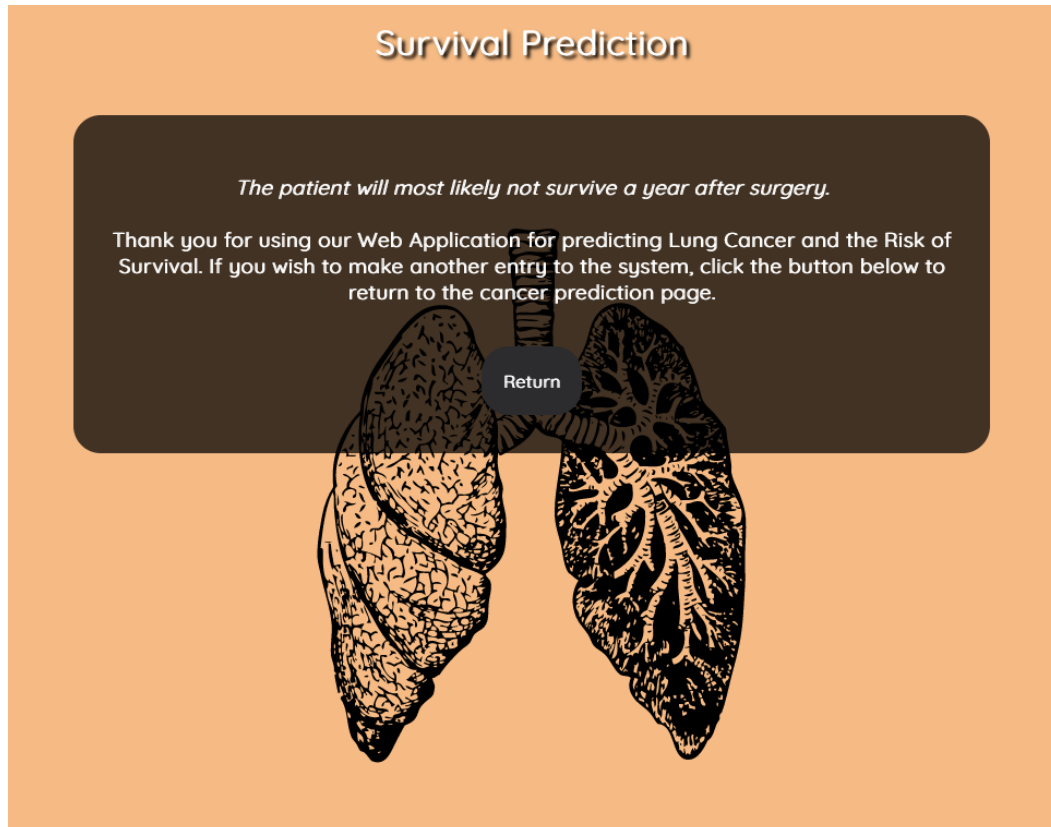
☐ Yes ☒ No

Previous Next

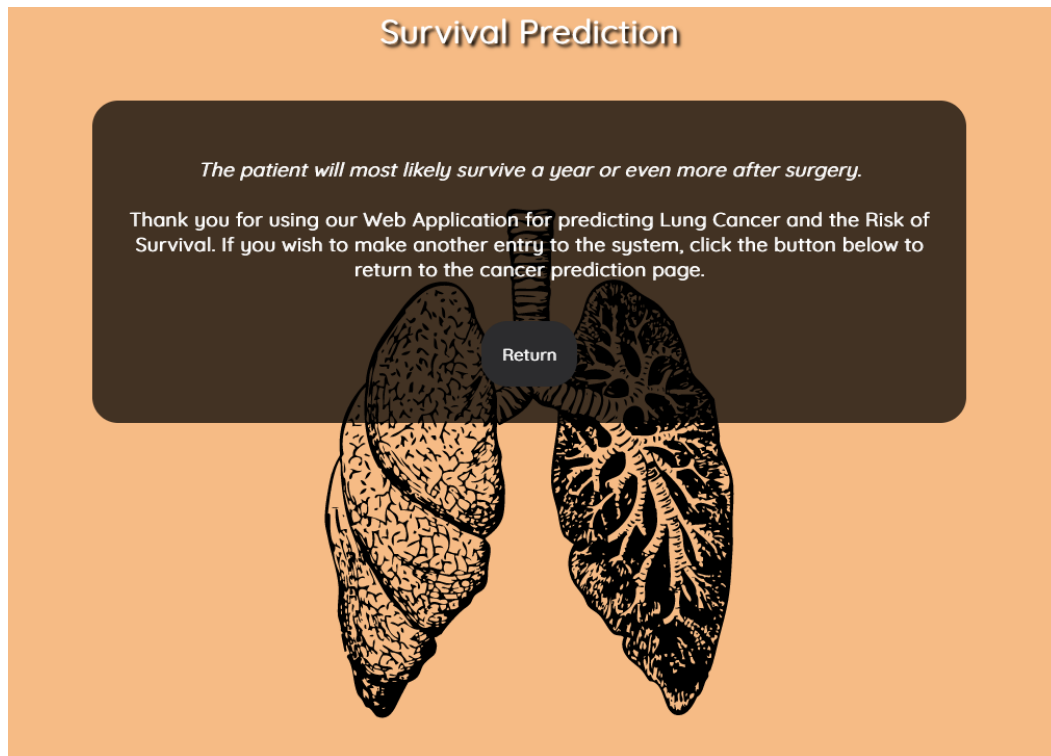
Εικόνα Ε΄.12: Η τέταρτη επιλογή του σακχαρώδη διαβήτη δεύτερου τύπου.



Εικόνα Ε΄.13: Η πέμπτη και τελευταία επιλογή του άσθματος.



Εικόνα Ε΄.14: Η περίπτωση αρνητικού αποτελέσματος επιβίωσης του δεύτερου μοντέλου.



Εικόνα Ε΄.15: Η περίπτωση θετικού αποτελέσματος επιβίωσης του δεύτερου μοντέλου.

Κεφάλαιο ΣΤ'

Συμπεράσματα

Σε αυτό το σημείο ολοκληρώνεται και η παρουσίαση της ανάπτυξης της διαδικτυακής μας εφαρμογής, γεγονός που σημάνει το τέλος αυτής της πτυχιακής εργασίας. Συνολικά, δημιουργώντας μία εφαρμογή από το σημείο της λήψης της ιδέας έως και το σημείο της ολοκλήρωσης της ήταν μια σημαντική εμπειρία που μας εξέθεσε σε πολλή πληροφορία, πολλά εργαλεία, διάφορους τρόπους ανάπτυξης, πολλαπλά προβλήματα αλλά και λύσεις, εφόδια χρήσιμα για τη μελλοντική τελειοποίηση τόσο αυτής της εφαρμογής, όσο και νέων που μπορεί να προκύψουν. Οι επιλογές είναι ατελείωτες όταν πραγματεύεται κανείς με τέτοιου είδους ιδέες, ειδικότερα στον τομέα της μηχανικής μάθησης με τον οποίο ασχολούμαστε, που συμμετέχει όλο και πιο συχνά σε καινοτόμες τεχνολογίες και λύσεις που παρουσιάζονται στον κλάδο της βιοπληροφορικής. Αυτό απέδειξε και η βιβλιογραφική μας έρευνα και σύγκριση των μελετών, στην οποία είδαμε από πρώτο χέρι πως ενώ σε πολλές μελέτες χρησιμοποιούνται παρόμοιες μεθόδους, δημιουργούνται παράλληλα καινοτόμα συστήματα και συσκευές που επιφέρουν σημαντικές αλλαγές στον κλάδο. Έπειτα παρατηρήσαμε την πληθώρα δεδομένων και τεχνικών επεξεργασίας τους που υπάρχουν διαθέσιμες στο διαδίκτυο, με τις οποίες μπορεί κανείς να διαμορφώσει έτσι τα δεδομένα που τον εξυπηρετούν ώστε να παράξει μοτίβα που αντικατοπτρίζουν συνήθειες, αίτια και άλλα σημαντικά παράγωγα. Στη συνέχεια καταφέραμε να υλοποιήσουμε την ιδέα μας με την χρήση εργαλείων που ενώ είναι διαθέσιμα εδώ και δεκαετίες, παραμένουν αποτελεσματικά και χρήσιμα για αυτούς που επιθυμούν να δημιουργήσουν κάτι ουσιαστικό. Παρότι φαινομενικά απλή, η εφαρμογή μας πετυχαίνει τον στόχο της και δίνει την ευκαιρία σε εξειδικευμένο και μη προσωπικό να χρησιμοποιήσει δύο απολύτως ικανά μοντέλα τόσο για τη διάγνωση του καρκίνου του πνεύμονα, όσο και για την αξιολόγηση του ρίσκου επιβίωσης του ασθενή, όλα αυτά αλληλεπιδρώντας με μια εφαρμογή που είναι εύκολη και εργονομική για τον χρήστη. Σίγουρα βέβαια υπάρχουν περιθώρια για μια μελλοντική συνέχιση της ανάπτυξης της εφαρμογής ή της τελειοποίησης της, τόσο με την προσθήκη νέων λειτουργιών, όσο και με την βελτιστοποίηση των υπάρχοντων μοντέλων ή την δημιουργία νέων ικανότερων. Στην πορεία υπάρχει διαθέσιμο ένα παράρτημα όπου βρίσκεται όλος ο κώδικας που δημιουργήσαμε, τόσο για τα μοντέλα, όσο και για το κάθε διαφορετικό κομμάτι της εφαρμογής, καθώς και τις βιβλιογραφικές πηγές από τις οποίες αντλήσαμε χρήσιμες πληροφορίες για την έρευνα μας.

Παράρτημα Α΄

Παράθεση κώδικα

Σε αυτό το παράρτημα θα καταχωρήσουμε το σύνολο του κώδικα που γράφθηκε για τη δημιουργία των μοντέλων, των διάφορων σελίδων του επιπέδου διεπαφής (front-end) καθώς και του επιπέδου back-end, με τη σειρά που αυτά αναλύθηκαν παραπάνω.

Κώδικας δημιουργίας μοντέλου διάγνωσης καρκίνου του πνεύμονα:

```
import pandas as pd
import numpy as np
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
import pickle

data = pd.read_csv("dfMissSpreadAS.csv")

X = data.values
y = data["LUNG_CANCER"].values
X = np.delete(X, 6, axis=1)

clf = XGBClassifier()
clf.fit(X, y)

pickle.dump(clf, open('LCmodel.pkl', 'wb'))

model = pickle.load(open('LCmodel.pkl', 'rb'))
print(model.predict([[1,1,1,1,1,1]]))
```

Κώδικας δημιουργίας μοντέλου αξιολόγησης ρίσκου επιβίωσης του καρκίνου του πνεύμονα:

```
import pandas as pd
import numpy as np
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
import pickle

data = pd.read_csv("60tsAS.csv")

data['DGN'] = data['DGN'].map( {'DGN1':1, 'DGN2':2, 'DGN3':3, 'DGN4':4,
'DGN5':5, 'DGN6':6, 'DGN8':8,} )
data['PRE14'] = data['PRE14'].map( {'OC11':11, 'OC12':12, 'OC13':13,
'OC14':14} )
```

```

X = data.values
y = data["Risk1Yr"].values
X = np.delete(X, 5, axis=1)

clf = XGBClassifier()
clf.fit(X, y)

pickle.dump(clf, open('TSmodel.pkl', 'wb'))

model = pickle.load(open('TSmodel.pkl', 'rb'))
print(model.predict([[4,0,11,0,0]]))

```

Κώδικας αρχικής σελίδας της εφαρμογής:

```

<!DOCTYPE html>
<html>
<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Lung Cancer CDSS</title>

    <style>

        @import
url('https://fonts.googleapis.com/css2?family=Quicksand&display=swap');

        body {
            position: center;
            text-align: center;
            max-width: 850px;
            margin: auto;
            padding: auto;
            font-family: 'Quicksand', sans-serif;
            background-image: url('/static/bodyclr.png');
            background-repeat: no-repeat;
            background-position: center center;

        }
        h1, h2, h3, h4, p{
            color: white;
            text-align: center;
        }

        .border {
            border-radius: 25px;

```

```

        background-color: rgba(0, 0, 0, 0.734);
        padding: 35px;
        margin: 5px;
    }

    .fade-in {
        animation: fadeIn 3s;
    }

    @keyframes fadeIn {
    0% { opacity: 0; }
    100% { opacity: 1; }
    }

    .button {

        border: none;
        border-radius: 25px;
        color: white;
        padding: 16px 32px;
        text-align: center;
        text-decoration: none;
        display: inline-block;
        font-size: 16px;
        margin: auto;
        transition-duration: 0.4s;
        cursor: pointer;
    }

    .button1 {
        background-color: rgb(44, 44, 46);
        color: rgb(255, 255, 255);
    }

    .button1:hover {
        background-color: #474545;
    }

</style>
</head>

<body>
    <div class="fade-in">
        <h1 style="text-shadow: 3px 3px 3px black;">Lung Cancer Clinical
Decision Support System</h1>

```

```
<br>

<h3 class="border">Lung Cancer has become one of the most common
cancer types in the world, as there
    are many factors in the modern lifestyle that increase the risk of
having it. As with many types of
    cancer, it can lead to a number of health problems and even death
if not diagnosed in time.
</h3>

<br>

<h3 class="border">This Web Application is created in order to
help medical professionals with their decision of whether or not a patient has
    lung cancer. If they do, then it can also predict the
patient's survival after surgery. The Web Application uses two Machine
Learning models, one for the prediction of lung cancer and the other for the
prediction of the patient's survival for more than a year
    after tumor removal surgery, both based on the input provided
by the user in the forms. The second model is only "activated" when lung
cancer is predicted by the first model.
</h3>

<br>

<h3 class="border">
    This application is meant to be used as a Clinical
Decision Support System mainly by professionals, because
    the decisions made by the models have a certain percentage
of accuracy and are not always guaranteed to be valid.
    If you want to make a prediction, start by clicking the
following button.
</h3>

<br>

<a href="{{url_for('predict_lung_cancer')}}" class="button button1"
style="padding: 0px 20px;"><h4>Begin</h4></a>

</div>

</body>
</html>
```

Κώδικας πρώτου σκέλους της εφαρμογής - φόρμα που αφορά τη διάγνωση του καρκίνου:

```
<!DOCTYPE html>
<html>
<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Lung Cancer CDSS</title>

    <style>

        @import
url('https://fonts.googleapis.com/css2?family=Quicksand&display=swap');

        body {
            position: center;
            text-align: center;
            max-width: 850px;
            margin: auto;
            padding: auto;
            font-family: 'Quicksand', sans-serif;
            background-image: url('/static/bodyclr.png');
            background-repeat: no-repeat;
            background-position: center center;

        }

        h1, h2, h3, h4, p{
            color: white;
            text-align: center;
        }

        .border {
            border-radius: 25px;
            background-color: rgba(0, 0, 0, 0.734);
            padding: 35px;
            margin: 5px;

        }

        .borderform {
            border-radius: 25px;
```

```
        border: 1px;
        padding: 4%;
        border-style: solid;
        background-color: rgba(99, 99, 99, 0.764);
        border-color: rgb(44, 44, 46);
        box-shadow: 2px 2px rgb(44, 44, 46);
    }
    .fade-in {
        animation: fadeIn 3s;
    }

    @keyframes fadeIn {
        0% { opacity: 0; }
        100% { opacity: 1; }
    }

    .button {

        border: none;
        border-radius: 25px;
        color: white;
        padding: 16px 32px;
        text-align: center;
        text-decoration: none;
        display: inline-block;
        font-size: 16px;
        margin: auto;
        transition-duration: 0.4s;
        cursor: pointer;
    }

    .button1 {
        background-color: rgb(44, 44, 46);
        color: rgb(255, 255, 255);
    }

    .button1:hover {
        background-color: #474545;
    }

    #regForm {

        margin: 50px auto;
        padding: 40px;
        width: 70%;
    }
```

```
        min-width: 300px;
    }

    input {
        border-radius: 25px;
        background: white;
        border: 5px solid #FFF;
        accent-color: #04AA6D;
    }

    label {
        color: white;
        font-family: 'Quicksand', sans-serif;
        font-size: larger;
    }

    input.invalid {
        background-color: #ff0000;
    }

    .tab {
        display: none;
    }

    .step {
        height: 15px;
        width: 15px;
        margin: 0 2px;
        background-color: #bbbbbb;
        border: none;
        border-radius: 50%;
        display: inline-block;
        opacity: 0.5;
    }

    .step.active {
        opacity: 1;
    }

    .step.finish {
        background-color: #04AA6D;
    }

    .tooltip {
        position: relative;
        display: inline-block;

    }
```

```

.tooltip .tooltiptext {
    visibility: hidden;
    width: 120px;
    background-color: rgb(44, 44, 46);
    color: rgb(255, 255, 255);
    text-align: center;
    border: 1px solid black;
    border-radius: 6px;
    padding: 5px 0;
    position: absolute;
    z-index: 1;
}

.tooltip:hover .tooltiptext {
    visibility: visible;
}

</style>
</head>

<body>
    <div class="fade-in">

        <h1 style="text-shadow: 3px 3px 3px black;">Lung Cancer
        Prediction</h1>

        <form id="regForm" class="border" action="/predict_lung_cancer"
        method="post">

            <h3>First of all, we have to predict if the patient has lung
            cancer. In order to do so, you have to provide
                the following essential information.
            </h3>

            <br>

            <div class="borderform">

                <h2><em>Is the patient experiencing the following?</em></h2>

                <br>

                <div class="tab"><h3>Peer Pressure:</h3>

                    <input type="radio" id="yes" name="PEER_PRESSURE"
                    value="1" checked required>
                    <label for="yes" method="post">Yes</label>

```

```

        <input type="radio" id="no" name="PEER_PRESSURE"
value="0">
        <label for="no" method="post">No</label>

</div>

<div class="tab"><h3>Fatigue:</h3>

        <input type="radio" id="yes" name="FATIGUE" value="1"
checked required>
        <label for="yes" method="post">Yes</label>

        <input type="radio" id="no" name="FATIGUE" value="0">
        <label for="no" method="post">No</label>

</div>

<div class="tab"><h3>Allergy:</h3>

        <input type="radio" id="yes" name="ALLERGY" value="1"
checked required>
        <label for="yes" method="post">Yes</label>

        <input type="radio" id="no" name="ALLERGY" value="0">
        <label for="no" method="post">No</label>

</div>

<div class="tab">

        <h3>
            <div class="tooltip">&#x1F6C8;
                <span class="tooltiptext">Wheezing is a high-
pitched whistling sound made while breathing.</span>
            </div>
            Wheezing:
        </h3>

        <input type="radio" id="yes" name="WHEEZING" value="1"
checked required>
        <label for="yes" method="post">Yes</label>

        <input type="radio" id="no" name="WHEEZING" value="0">
        <label for="no" method="post">No</label>

</div>

```

[illegible]

```

        </div>

    </form>

</div>

<script>
    var currentTab = 0;
    showTab(currentTab);

    function showTab(n) {
        var x = document.getElementsByClassName("tab");
        x[n].style.display = "block";
        if (n == 0) {
            document.getElementById("prevBtn").style.display = "none";
        } else {
            document.getElementById("prevBtn").style.display = "inline";
        }
        if (n == (x.length - 1)) {
            document.getElementById("nextBtn").innerHTML = "Submit";
        } else {
            document.getElementById("nextBtn").innerHTML = "Next";
        }
        fixStepIndicator(n)
    }

    function nextPrev(n) {
        var x = document.getElementsByClassName("tab");
        if (n == 1 && !validateForm()) return false;
        x[currentTab].style.display = "none";
        currentTab = currentTab + n;
        if (currentTab >= x.length) {
            document.getElementById("regForm").submit();
            return false;
        }
        showTab(currentTab);
    }

    function validateForm() {
        var x, y, i, valid = true;
        x = document.getElementsByClassName("tab");
        y = x[currentTab].getElementsByTagName("input");
        if (valid) {
            document.getElementsByClassName("step")[currentTab].className
+= " finish";
        }
        return valid;
    }

```

```

    }

    function fixStepIndicator(n) {

        var i, x = document.getElementsByClassName("step");
        for (i = 0; i < x.length; i++) {
            x[i].className = x[i].className.replace(" active", "");
        }
        x[n].className += " active";
    }
</script>
</body>
</html>

```

Κώδικας σελίδας αρνητικού αποτελέσματος της διάγνωσης καρκίνου του πνεύμονα:

```

<!DOCTYPE html>
<html>
<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Lung Cancer CDSS</title>

    <style>

        @import
url('https://fonts.googleapis.com/css2?family=Quicksand&display=swap');

        body {
            position: center;
            text-align: center;
            max-width: 850px;
            margin: auto;
            padding: auto;
            font-family: 'Quicksand', sans-serif;
            background-image: url('/static/bodyclr.png');
            background-repeat: no-repeat;
            background-position: center center;
            background-attachment: fixed;

        }

        h1, h2, h3, h4, p{
            color: white;
            text-align: center;

```

```
}

.border {
  border-radius: 25px;
  background-color: rgba(0, 0, 0, 0.734);
  padding: 35px;
  margin: 5px;
}

.fade-in {
  animation: fadeIn 3s;
}

@keyframes fadeIn {
  0% { opacity: 0; }
  100% { opacity: 1; }
}

.button {

  border: none;
  border-radius: 25px;
  color: white;
  padding: 16px 32px;
  text-align: center;
  text-decoration: none;
  display: inline-block;
  font-size: 16px;
  margin: auto;
  transition-duration: 0.4s;
  cursor: pointer;
}

.button1 {
  background-color: rgb(44, 44, 46);
  color: rgb(255, 255, 255);
}

.button1:hover {
  background-color: #474545;
}

</style>
</head>
```

```

<body>
  <div class="fade-in">

    <h1 style="text-shadow: 3px 3px 3px black;">No Lung Cancer
detected</h1>

    <br>

    <div class="border">

      <h3>Based on the information provided in the previous form, the
"subject" does not qualify as a potential Lung Cancer patient. Do not forget
that this decision
        could be a false negative, as there is always room for error.
      <br><br>
      If you wish to make another entry, please click the following
button.
    </h3>

    <br>

    <a href="{{url_for('predict_lung_cancer')}}" class="button button1"
style="padding: 0px 20px;"><h4>Go back</h4></a>

    </div>

  </div>

</body>
</html>

```

Κώδικας δεύτερου σκέλους της εφαρμογής που αφορά το θετικό αποτέλεσμα μετά τη διάγνωση αλλά και την αξιολόγηση ρίσκου επιβίωσης:

```

<!DOCTYPE html>
<html>
<head>

  <meta charset="UTF-8">

  <meta name="viewport" content="width=device-width, initial-scale=1.0">

  <title>Lung Cancer CDSS</title>

  <style>

    @import
url('https://fonts.googleapis.com/css2?family=Quicksand&display=swap');

```

```
body {
  position: center;
  text-align: center;
  max-width: 850px;
  margin: auto;
  padding: auto;
  font-family: 'Quicksand', sans-serif;
  background-image: url('/static/bodyclr.png');
  background-repeat: no-repeat;
  background-position: center center;
}

h1, h2, h3, h4, p{
  color: white;
  text-align: center;
}

.border {
  border-radius: 25px;
  background-color: rgba(0, 0, 0, 0.734);
  padding: 5px;
  margin: 5px;
}

.borderform {
  border-radius: 25px;
  border: 1px;
  padding: 4%;
  border-style: solid;
  background-color: rgba(99, 99, 99, 0.764);
  border-color: rgb(44, 44, 46);
  box-shadow: 2px 2px rgb(44, 44, 46);
}

.fade-in {
  animation: fadeIn 3s;
}

@keyframes fadeIn {
  0% { opacity: 0; }
  100% { opacity: 1; }
}

.button {
```

```
border: none;
border-radius: 25px;
color: white;
padding: 16px 32px;
text-align: center;
text-decoration: none;
display: inline-block;
font-size: 16px;
margin: auto;
transition-duration: 0.4s;
cursor: pointer;
}

.button1 {
background-color: rgb(44, 44, 46);
color: rgb(255, 255, 255);
}

.button1:hover {
background-color: #474545;
}

#regForm {

margin: 25px auto;
padding: 40px;
width: 70%;
min-width: 300px;
}

input {
border-radius: 25px;
background: white;
border: 5px solid #FFF;
accent-color: #04AA6D;
}

label {
color: white;
font-family: 'Quicksand', sans-serif;
font-size: large;
}

.tab {
display: none;
}
```

```
.step {
    height: 15px;
    width: 15px;
    margin: 0 2px;
    background-color: #bbbbbb;
    border: none;
    border-radius: 50%;
    display: inline-block;
    opacity: 0.5;
}

.step.active {
    opacity: 1;
}

.step.finish {
    background-color: #04AA6D;
}

.tooltip {
    position: relative;
    display: inline-block;

}

.tooltip .tooltiptext {
    visibility: hidden;
    width: 120px;
    background-color: rgb(44, 44, 46);
    color: rgb(255, 255, 255);
    text-align: center;
    border: 1px solid black;
    border-radius: 6px;
    padding: 5px 0;
    position: absolute;
    z-index: 1;
}

.tooltip:hover .tooltiptext {
    visibility: visible;
}

</style>
</head>

<body>
    <div class="fade-in">
```

```
<h1 style="text-shadow: 3px 3px 3px black;">Surgery Survival
Prediction</h1>
```

```
<form id="regForm" class="border" action="/survival" method="post">
```

```
<h2><em>Unfortunately, the patient has Lung Cancer.</em></h2>
```

```
<br>
```

```
<h3>
```

You can now estimate the survival of the patient for more than a year in case they choose to remove the cancerous tumour with surgery. In order to do so, you need to

follow the same process as before, but this time by providing more specialized information about the patient in the following form.

```
</h3>
```

```
<br>
```

```
<div class="borderform">
```

```
<h3><em>Please fill in the data below:</em></h3>
```

```
<br>
```

```
<div class="tab">
```

```
<h3>
```

```
<div class="tooltip">&#x1F6C8;
```

Specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any.

```
</div>
```

```
Diagnosis:
```

```
</h3>
```

```
<input type="radio" id="DGN1" name="DGN" value="1" checked
required>
```

```
<label for="DGN1" method="post">DGN1</label>
```

```
<input type="radio" id="DGN2" name="DGN" value="2">
```

```
<label for="DGN2" method="post">DGN2</label>
```

```
<input type="radio" id="DGN3" name="DGN" value="3">
```

```
<label for="DGN3" method="post">DGN3</label>
```

```
<input type="radio" id="DGN4" name="DGN" value="4">
```

```
<label for="DGN4" method="post">DGN4</label>
```

```

<input type="radio" id="DGN5" name="DGN" value="5">
<label for="DGN5" method="post">DGN5</label>

<input type="radio" id="DGN6" name="DGN" value="6">
<label for="DGN6" method="post">DGN6</label>

<input type="radio" id="DGN8" name="DGN" value="8">
<label for="DGN8" method="post">DGN8</label>

</div>

<div class="tab"><h3>Dyspnoea before surgery:</h3>

  <input type="radio" id="yes" name="PRE9" value="1" checked
required>
  <label for="yes" method="post">Yes</label>

  <input type="radio" id="no" name="PRE9" value="0">
  <label for="no" method="post">No</label>

</div>

<div class="tab">

  <h3>
    <div class="tooltip">&#x1F6C8;
      <span class="tooltiptext">T in clinical TNM - size
of the original tumour, from OC11 (smallest) to OC14 (largest).</span>
    </div>
    Size of original tumour:
  </h3>

  <input type="radio" id="OC11" name="PRE14" value="11"
checked required>
  <label for="OC11" method="post">OC11</label>

  <input type="radio" id="OC12" name="PRE14" value="12">
  <label for="OC12" method="post">OC12</label>

  <input type="radio" id="OC13" name="PRE14" value="13">
  <label for="OC13" method="post">OC13</label>

  <input type="radio" id="OC14" name="PRE14" value="14">
  <label for="OC14" method="post">OC14</label>

</div>

<div class="tab"><h3>Type 2 Diabetes Mellitus:</h3>

```

```

        <input type="radio" id="yes" name="PRE17" value="1"
checked required>
        <label for="yes" method="post">Yes</label>

        <input type="radio" id="no" name="PRE17" value="0">
        <label for="no" method="post">No</label>

    </div>

    <div class="tab"><h3>Asthma:</h3>

        <input type="radio" id="yes" name="PRE32" value="1"
checked required>
        <label for="yes" method="post">Yes</label>

        <input type="radio" id="no" name="PRE32" value="0">
        <label for="no" method="post">No</label>
    </div>

    <br>

    <div style="overflow:auto;">
        <div style="float:right;">
            <button type="button" class="button button1"
id="prevBtn" onclick="nextPrev(-1)">Previous</button>
            <button type="button" class="button button1"
id="nextBtn" onclick="nextPrev(1)">Next</button>
        </div>
    </div>

    <div style="text-align:center;margin-top:40px;">
        <span class="step"></span>
        <span class="step"></span>
        <span class="step"></span>
        <span class="step"></span>
        <span class="step"></span>
    </div>
</div>
</form>

</div>

<script>
    var currentTab = 0;
    showTab(currentTab);

    function showTab(n) {

```

```

    var x = document.getElementsByClassName("tab");
    x[n].style.display = "block";
    if (n == 0) {
        document.getElementById("prevBtn").style.display = "none";
    } else {
        document.getElementById("prevBtn").style.display = "inline";
    }
    if (n == (x.length - 1)) {
        document.getElementById("nextBtn").innerHTML = "Submit";
    } else {
        document.getElementById("nextBtn").innerHTML = "Next";
    }
    fixStepIndicator(n)
}

function nextPrev(n) {
    var x = document.getElementsByClassName("tab");
    if (n == 1 && !validateForm()) return false;
    x[currentTab].style.display = "none";
    currentTab = currentTab + n;
    if (currentTab >= x.length) {
        document.getElementById("regForm").submit();
        return false;
    }
    showTab(currentTab);
}

function validateForm() {
    var x, y, i, valid = true;
    x = document.getElementsByClassName("tab");
    y = x[currentTab].getElementsByTagName("input");
    if (valid) {
        document.getElementsByClassName("step")[currentTab].className
+= " finish";
    }
    return valid;
}

function fixStepIndicator(n) {
    var i, x = document.getElementsByClassName("step");
    for (i = 0; i < x.length; i++) {
        x[i].className = x[i].className.replace(" active", "");
    }
    x[n].className += " active";
}
</script>
</body>
</html>

```

Κώδικας σελίδας θετικού αποτελέσματος της αξιολόγησης ρίσκου επιβίωσης από τον καρκίνο του πνεύμονα:

```
<!DOCTYPE html>
<html>
<head>

  <meta charset="UTF-8">

  <meta name="viewport" content="width=device-width, initial-scale=1.0">

  <title>Lung Cancer CDSS</title>

  <style>

    @import
url('https://fonts.googleapis.com/css2?family=Quicksand&display=swap');

    body {
      position: center;
      text-align: center;
      max-width: 850px;
      margin: auto;
      padding: auto;
      font-family: 'Quicksand', sans-serif;
      background-image: url('/static/bodyclr.png');
      background-repeat: no-repeat;
      background-position: center center;
      background-attachment: fixed;

    }

    h1, h2, h3, h4, p{

      color: white;
      text-align: center;
    }

    .border {

      border-radius: 25px;
      background-color: rgba(0, 0, 0, 0.734);
      padding: 35px;
      margin: 5px;
    }

    .fade-in {

      animation: fadeIn 3s;
```

```
}

@keyframes fadeIn {

    0% { opacity: 0; }
    100% { opacity: 1; }
}

.button {

    border: none;
    border-radius: 25px;
    color: white;
    padding: 16px 32px;
    text-align: center;
    text-decoration: none;
    display: inline-block;
    font-size: 16px;
    margin: auto;
    transition-duration: 0.4s;
    cursor: pointer;
}

.button1 {

    background-color: rgb(44, 44, 46);
    color: rgb(255, 255, 255);

}

.button1:hover {

    background-color: #474545;
}

</style>
</head>

<body>
    <div class="fade-in">

        <h1 style="text-shadow: 3px 3px 3px black;">Survival Prediction</h1>

        <br>

        <div class="border">
```

```

        <h3>
        <em>The patient will most likely survive a year or even more after
surgery.</em>

        <br><br>

        Thank you for using our Web Application for predicting Lung
Cancer and the Risk of Survival. If you wish to make another entry to the
system,
        click the button below to return to the cancer prediction
page.

    </h3>

    <br>

    <a href="{{url_for('predict_lung_cancer')}}" class="button
button1" style="padding: 0px 20px;"><h4>Return</h4></a>

    </div>

</div>

</body>
</html>

```

Κώδικας σελίδας αρνητικού αποτελέσματος της αξιολόγησης ρίσκου επιβίωσης από τον καρκίνο του πνεύμονα:

```

<!DOCTYPE html>
<html>
<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Lung Cancer CDSS</title>

    <style>

        @import
url('https://fonts.googleapis.com/css2?family=Quicksand&display=swap');

        body {
            position: center;
            text-align: center;
            max-width: 850px;
            margin: auto;

```

```
padding: auto;
font-family: 'Quicksand', sans-serif;
background-image: url('/static/bodyclr.png');
background-repeat: no-repeat;
background-position: center center;
background-attachment: fixed;

}
h1, h2, h3, h4, p{
color: white;
text-align: center;
}

.border {
border-radius: 25px;
background-color: rgba(0, 0, 0, 0.734);
padding: 35px;
margin: 5px;
}

.fade-in {
animation: fadeIn 3s;
}

@keyframes fadeIn {
0% { opacity: 0; }
100% { opacity: 1; }
}

.button {

border: none;
border-radius: 25px;
color: white;
padding: 16px 32px;
text-align: center;
text-decoration: none;
display: inline-block;
font-size: 16px;
margin: auto;
transition-duration: 0.4s;
cursor: pointer;
}

.button1 {
background-color: rgb(44, 44, 46);
color: rgb(255, 255, 255);
```

```

    }

    .button1:hover {
        background-color: #474545;
    }

</style>
</head>

<body>
    <div class="fade-in">
        <h1 style="text-shadow: 3px 3px 3px black;">Survival Prediction</h1>

        <br>
        <div class="border">
            <h3>
                <em>The patient will most likely not survive a year after
surgery.</em>

                <br><br>

                Thank you for using our Web Application for predicting Lung
Cancer and the Risk of Survival. If you wish to make another entry to the
system,

                click the button below to return to the cancer prediction
page.

            </h3>

            <br>

            <a href="{{url_for('predict_lung_cancer')}}" class="button
button1" style="padding: 0px 20px;"><h4>Return</h4></a>

        </div>

    </div>

</body>
</html>

```

Κώδικας επιπέδου back-end – σύνδεση όλων των παραπάνω:

```
import flask
from flask import Flask, request, render_template, redirect, url_for
import pickle
import numpy as np
import time

app = Flask(__name__)

lc_model = pickle.load(open('models/LCmodel.pkl', 'rb'))
ts_model = pickle.load(open('models/TSmodel.pkl', 'rb'))

@app.route('/')

def home():
    return render_template('home.html')

@app.route('/predict_lung_cancer/', methods=['GET', 'POST'])
def predict_lung_cancer() -> "html":
    if request.method == 'GET':
        return render_template('cancer_form.html')
    else:
        int_lc_features = [int(x) for x in request.form.values()]
        lc_features = [np.array(int_lc_features)]
        lc_output = lc_model.predict(lc_features)
        if (lc_output==0):
            return render_template('cancer_false.html')
        else:
            return render_template('survival_form.html')

@app.route('/survival', methods=['GET', 'POST'])
def survival() -> "html":
    if request.method == 'GET':
        return render_template('survival_form.html')
    elif request.method == 'POST':
        int_ts_features = [int(x) for x in request.form.values()]
        ts_features = [np.array(int_ts_features)]
        ts_output = ts_model.predict(ts_features)
        if (ts_output==0):
            return render_template('survival_false.html')
        else:
            return render_template('survival_true.html')

if __name__ == "__main__":
    app.run()
```

Βιβλιογραφία

- [1] Division of Cancer Prevention and Control, Centers for Disease Control and Prevention, "What is Lung Cancer?", July 31, 2023, "https://www.cdc.gov/cancer/lung/basic_info/what-is-lung-cancer.htm"
- [2] Lung Cancer Statistics, World Cancer Research Fund International, 2020, "<https://www.wcrf.org/cancer-trends/lung-cancer-statistics/>"
- [3] "https://cdn.who.int/media/docs/default-source/country-profiles/cancer/grc-2020.pdf?sfvrsn=f9aa454c_2&download=true"
- [4] Power, D., "What is the need for decision support?" DSS News, Vol. 10, No. 11, May 31, 2009; updated April 29. 2017.
- [5] Sushruta Mishra, Hiren Kumar Thakkar, Pradeep Kumar Mallick, Prayag Tiwari, Atif Alamri, "A sustainable IoHT based computationally intelligent healthcare monitoring system for lung cancer risk detection", Sustainable Cities And Society, Volume 72, September 2021
- [6] Binson V A, Dr. M Subramoniam, Dr. Ragesh G K, Dr. Ajay Kumar, "Early Detection of Lung Cancer Through Breath Analysis Using AdaBoost Ensemble Learning Method", 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), IEEE, September 2021
- [7] Madara Tirzīte, Māris Bukovskis, Gunta Strazda, Normunds Jurka, Immanuels Taivans, "Detection of lung cancer with electronic nose and logistic regression analysis", Journal of Breath Research, IOP Publishing, November 2018
- [8] O.J. Juan Vidal, M.E. Gas Lopez, C. López Gómez, A. Loras Monfort, J.L. Valles Pardo, S. Palanca Suela, N. Mancheño Franch, E. Ansotegui, V. Calvo Medina, C. Muñoz Nuñez, F.J. Celada Alvarez, B. Martínez Sanchís, S. Marset García, J. Castejón Esteban, C. Cano Cerviño, J. Medina Alvarez, G. Garcia Rubia, A. Callejo Mellen, B. Valdivieso Martínez, "Development of a predictive model for early detection of lung cancer: SREAL-eLUNG study", Journal of Thoracic Oncology, Vol. 18 No. 4S, April 2023
- [9] Hadrien Charvat, Shizuka Sasazuki, Taichi Shimazu, Sanjeev Budhathoki, Manami Inoue, Motoki Iwasaki, Norie Sawada, Taiki Yamaji, Shoichiro Tsugane, JPHC Study Group, "Development of a risk prediction model for lung cancer: The Japan Public Health Center-based Prospective Study", Cancer Science of the official journal of the Japanese Cancer Association, February 2018

- [10] Jie Ma, Maria A. Guarnera, Wenxian Zhou, HongBin Fang, Feng Jiang, "A Prediction Model Based on Biomarkers and Clinical Characteristics for Detection of Lung Cancer in Pulmonary Nodules 1,2", *Translational Oncology*, vol. 10, Issue 1, February 2017
- [11] Deepak Rawat, "Validating and Strengthen the Prediction Performance Using Machine Learning Models and Operational Research for Lung Cancer", *International Conference on Data Science and Information System*, IEEE, 2022
- [12] Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis", *World AI IoT Congress*, IEEE, 2022
- [13] DendiGayathri Reddy, Emmidi Naga Hemanth Kumar, DesiredreddyLohithSaiCharan Reddy, Monika P, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method", *1st International Conference on Advances in Information Technology*, IEEE, 2019
- [14] Zhuqing Cai, Zhulian Yu, Haiyu Zhou, Zhenghui Gu, "The Early Stage Lung Cancer Prognosis Prediction Model based on Support Vector Machine", *23rd International Conference on Digital Signal Processing (DSP)*, IEEE, 2018
- [15] Pranamita Nanda, Dr. N. Duraipandian, "Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest", *Fifth International Conference on Inventive Computation Technologies (ICICT-2020)*, IEEE, 2020
- [16] Yutao Li, Zixiu Zou, Zhunyi Gao, Yi Wang, Man Xiao, Chang Xu, Gengxi Jiang, Haijian Wang, Li Jin, Jiucun Wang, Huai Zhou Wang, Shicheng Guo, Junjie Wu, "Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting", *Cancer Medicine*, PubMed Central, December 2022
- [17] Deepak Rawat, Meenakshi, Lokesh Pawar, Guarav Bathla, Ravi Kant, "Optimized Deep Learning Model for Lung Cancer Prediction Using ANN Algorithm", *Third International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2022
- [18] Mohsin Sattar, Abdul Majid, Nabeela Kausar, Muhammad Bilal, Muhammad Kashi, "Lung cancer prediction using multi-gene genetic programming by selecting automatic features from amino acid sequences", *Computational Biology and Chemistry*, Volume 98, June 2022
- [19] Ali Safiyari, Reza Javidan, "Predicting Lung Cancer Survivability using Ensemble Learning Methods", *Intelligent Systems Conference*, IEEE, 2017

[20] Mysar Ahmad Bhat, "Lung Cancer", "<https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>"

[21] Manu Siddhartha, "Thoracic Surgery Dataset", "<https://www.kaggle.com/datasets/sid321axn/thoracic-surgery>"

[22] Maegan Londa, "Thoracic Surgery", "<https://rpubs.com/melonda/190569>"