ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

‘’Μοντελοποίηση χαρακτηριστικών κυτταρικού πυρήνα από δεδομένα εικόνας για τη διάγνωση του όγκου του μαστού με μεθόδους μηχανικής μάθησης”

Παπαδόπουλος Μιχαήλ του Δημητρίου

Επιβλέπων Καθηγητής:

Βασιλείου Ευάγγελος

Χίος

Ιούνιος 2023

Πνευματικά δικαιώματα

Έχω διαβάσει και κατανοήσει τους κανόνες για τη λογοκλοπή και τον τρόπο σωστής αναφοράς των πηγών που περιέχονται στον οδηγό συγγραφής διπλωματικών εργασιών του ΤΜΟΔ. Δηλώνω ότι, από όσα γνωρίζω, το περιεχόμενο της παρούσας διπλωματικής εργασίας είναι προϊόν δικής μου δουλείας και υπάρχουν αναφορές σε όλες τις πηγές που χρησιμοποιούσα

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στο πανεπιστήμιο Αιγαίου για την αφοσίωση που επέδειξε και ειδικότερα στον επιβλέποντα καθηγητή μου κύριο Ευάγγελο Βασιλείου για την άριστη συνεργασία που είχαμε καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής εργασίας. Επίσης, θα ήθελα να εκφράσω ένα ακόμα μεγαλύτερο ευχαριστώ στην οικογένεια μου που με στήριξε κατά τη διάρκεια των σπουδών μου.

Πίνακας περιεχομένων

ΕΙΚΟΝΕΣ

ΠΙΝΑΚΕΣ

Περίληψη

Ελληνικά

Η πρόοδος της ιατρικής αλλά και οι διαδικασίες που χρησιμοποιούνται για την αξιολόγηση και θεραπεία νοσημάτων αναμφίβολα εξελίσσονται και αναπτύσσονται με καλπάζοντες ρυθμούς με τη βοήθεια της τεχνολογίας και της επιστήμης. Γιατροί και μηχανικοί δεδομένων χρησιμοποιούν μαθηματικά μοντέλα και αλγορίθμους μηχανικής μάθησης έτσι ώστε να επιτευχθούν οι επιθυμητοί στόχοι, αντλώντας τα απαραίτητα δεδομένα κυρίως από το ιστορικό των ασθενών. Είναι πλέον εύκολο να διαγνωστεί εάν ένας κυτταρικός όγκος είναι καλοήθης ή κακοήθης μέσω μοντέλων τα οποία έχουν εκπαιδευτεί από το ιστορικό προηγούμενων ασθενών. Σε γενικές γραμμές, η προσέγγιση της μηχανικής μάθησης βασίζεται στην προσπάθεια δημιουργίας αυτοματοποιημένων διαδικασιών οι οποίες εξελίσσονται με βάση το πρότυπο του τρόπου με τον οποίο ο άνθρωπος αποκτά τη γνώση.

Το προτεινόμενο θέμα διπλωματικής ασχολείται με την μοντελοποίηση χαρακτηριστικών κυτταρικού πυρήνα από δεδομένα εικόνας για τη διάγνωση του όγκου του μαστού με μεθόδους μηχανικής μάθησης. Η ιστοσελίδα Kaggle, έχοντας τον ρόλο της ‘’τράπεζας δεδομένων’’, παρέχει τα δεδομένα του πεδίου εφαρμογής τα οποία προκύπτουν και αντλούνται μέσω της ψηφιοποιημένης εικόνας στην οποία απεικονίζονται τα χαρακτηριστικά του κυτταρικού πυρήνα έπειτα από βιοψία δια λεπτής βελόνης (Fine Needle Aspirate - FNA) μάζας μαστού.

Εν συντομία, η βιοψία δια λεπτής βελόνης όγκου αποτελεί μια μέθοδο κατά την οποία περισυλλέγονται κύτταρα από όλα τα σημεία της βλάβης (όγκου) μέσω αναρρόφησης με την βοήθεια βελόνης 21g παρόμοια με αυτή της αιμοληψίας. Τα δεδομένα αυτά αποτελούνται από χαρακτηριστικά τα οποία αφορούν τα κύτταρα του σημείου της βλάβης (όγκου). Για παράδειγμα, οι αποστάσεις από το κέντρο σε σημεία της περιμέτρου, η υφή (η οποία μετριέται

στη κλίμακα του γκρι), ο αριθμός κοίλων τμημάτων του περιγράμματος, η περιοχή, η ομαλότητα κ.α.

Στόχος της εν λόγω διπλωματικής εργασίας είναι η μοντελοποίηση και ανάλυση των δεδομένων που προκύπτουν από την προαναφερθείσα εξέταση και έπειτα η εκπαίδευση αλγορίθμων μηχανικής μάθησης για την ταυτοποίηση και ταξινόμηση βλαβών (όγκων) ως καλοήθεις ή κακοήθης.

Για την υλοποίηση της διπλωματικής εργασίας χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης με επίβλεψη για την ανάλυση, μοντελοποίηση και αξιολόγηση των δεδομένων σε περιβάλλον R. Αρχικά, γίνεται έλεγχος για το αν στα δεδομένα εμπεριέχονται ελλιπής τιμές, το πώς κατανέμονται οι παρατηρήσεις της κάθε επεξηγηματικής μεταβλητής αλλά και το ποσοστό των κακοηθών όγκων. Επίσης, παρουσιάζονται κάποια περιγραφικά στατιστικά και κατασκευάζονται αντίστοιχα γραφήματα (barplots, boxplots, ιστογράμματα (histograms), γραφήματα, κ.α.) Στη συνέχεια, αφού χωριστούν τα δεδομένα εκπαίδευσης και δεδομένα ελέγχου, εκπαιδεύονται εναλλακτικοί αλγόριθμοι μηχανικής μάθησης με επίβλεψη για την αναγνώριση των πιο σημαντικών προσδιοριστικών παραγόντων του υπό μελέτη φαινομένου. Συμπληρωματικά, αξιολογείται η ικανότητα τους ως προς την ταξινόμηση των όγκων σε καλοήθης ή κακοήθης με την βοήθεια κατάλληλών μέτρων αξιολόγησης όπως η ακρίβεια (accuracy), η ευαισθησία, η εξειδίκευση (specificity), οι καμπύλες ROC, οι καμπύλες lift, οι καμπύλες precision vs recall, κ.α.. Τέλος, επιχειρείται μία σύγκριση των υπό μελέτη μεθόδων όσον αφορά στα ιδιαίτερα χαρακτηριστικά τους και την απόδοση τους στο συγκεκριμένο σύνολο δεδομένων.

Λέξεις – Κλειδιά: Μέθοδοι εποπτευόμενης μηχανικής μάθησης, αλγόριθμοι ταξινόμησης, Λογιστική Παλινδρόμηση, Δέντρα Αποφάσεων, Τυχαία Δάση, Μηχανές Διανυσμάτων Υποστήριξης, μέτρα αξιολόγησης κατηγοριοποίησης

Αγγλικά

The advancement of medicine and the procedures used to evaluate and treat diseases are undoubtedly evolving and developing at a galloping pace with the help of technology and science. Doctors and data engineers use mathematical models and machine learning algorithms to achieve the desired goals, drawing the necessary data mainly from patients' health record. It is now easy to diagnose whether a cellular tumor is benign or malignant through models trained from past patient's health record. In general, the machine learning approach is based on trying to create automated processes that evolve based on the pattern of how humans acquire knowledge.

The proposed thesis topic deals with the modeling of cell nucleus features from image data for breast tumor diagnosis using machine learning methods. The Kaggle website, acting as a

"database", provides the data of the field of application which is derived and extracted through the digitized image depicting the features of the cell nucleus after a Fine Needle Aspirate (FNA biopsy) of breast mass.

Briefly, fine-needle tumor biopsy is a method in which cells are collected from all parts of the lesion (tumor) by aspiration using a 21g needle similar to blood drawing. These data are characteristics of the cells at the point of lesion (tumor). For example, the distances from the center to points on the perimeter, the texture (which is measured in grayscale), the number of concave parts of the contour, the area, the smoothness, etc. Thesis' goal is to model and analyze the data resulting from the aforementioned examination and then train machine learning algorithms to identify and classify lesions (tumors) as benign or malignant.

For the implementation of the thesis, supervised machine learning algorithms are used for the analysis, modeling, and evaluation of the data in an R environment. Initially, a check is made for whether the data contain missing values, how the observations of each explanatory variable are distributed and also the percentage of malignant tumors. Also, some descriptive statistics are presented and corresponding graphs are constructed (barplots, violin plots, histograms, graphs, etc.). Then, after separating the training and control data, alternative supervised machine learning algorithms are trained to identify the most important determinants of the phenomenon under study. Additionally, their ability to classify tumors as benign or malignant is evaluated with the help of appropriate evaluation measures such as accuracy, sensitivity, specificity, ROC curves, lift curves, precision vs. recall, etc. Finally, a comparison of the methods under study is attempted in terms of their particular characteristics and their performance in the specific data set.

Εισαγωγή

Με την πρόοδο και την εξέλιξη της τεχνολογίας και των πληροφοριακών και επικοινωνιακών συστημάτων, είναι πλέον αναπόφευκτος ο πλήρης ψηφιακός μετασχηματισμός της

πληροφορίας. Η συμβολή των δεδομένων στην επιστημονική (και όχι μόνο) κοινότητα, έχει επιφέρει ραγδαία ανάπτυξη στις θετικές και τεχνολογικές επιστήμες όπως την ιατρική, την μηχανική δεδομένων κ.α.. Όντας λοιπόν ένα διεπιστημονικό πεδίο με κοινό παρονομαστή την ανάλυση των δεδομένων, λειτουργεί ως τροχοπέδη στην επίλυση προβλημάτων του εκάστοτε κλάδου μέσω μηχανικής μάθησης, της στατιστικής ανάλυσης και της τεχνητής νοημοσύνης.

Η ανάπτυξη των τεχνολογιών του διαδικτύου οδήγησε στην διεύρυνση και αύξηση του κοινού του και την έκρηξη της διαθεσιμότητας του αντίστοιχου περιεχομένου. Αυτό σημαίνει ότι στο διαδίκτυο διατίθενται μεγάλες ποσότητες δεδομένων για μεγάλο μέρος του παγκοσμίου πληθυσμού. Από τα δεδομένα αυτά, μέσα από κατάλληλη επεξεργασία, δύνανται να προκύψουν σημαντικά συμπεράσματα που μπορούν να ενισχύσουν τις διαδικασίες λήψης αποφάσεων. Το διαδίκτυο των πραγμάτων τροφοδοτεί τις υποδομές του διαδικτύου με δεδομένα που σχετίζονται με το περιβάλλον ή ζώντες οργανισμούς, μέσω καταλλήλων αισθητήρων που προσαρμόζονται σε διάφορες συσκευές.

Ο μεγάλος όγκος των δεδομένων που παράγονται και διαχέονται στο διαδίκτυο καθώς και ο ταχύς ρυθμός που αυτό συμβαίνει, καθιστά τις συμβατικές διεργασίες επεξεργασίας τους ανεπαρκείς για να υποστηρίξουν τις σύγχρονες διαδικασίες λήψης αποφάσεων. Στελέχη επιχειρήσεων λαμβάνουν υπόψιν την έκβαση μιας ανάλυσης δεδομένων έτσι ώστε να είναι ικανά για μία ορθολογική λήψη αποφάσεων. Με την παγκοσμιοποίηση να αυξάνει τις απαιτήσεις αποδοτικότητας των οργανισμών, προκειμένου να διατηρήσουν τη βιωσιμότητα τους, οι διαδικασίες λήψης αποφάσεων έχουν γίνει ζωτικής σημασίας για κάθε είδους οργανισμού καθώς μέσω αυτών χαράσσεται η στρατηγική και επιλέγονται οι τακτικές για την βελτίωση της θέσης τους στον ανταγωνισμό. Για παράδειγμα με την διεξαγωγή έρευνας σε πληθυσμό είναι εφικτό να εξορισχτούν δεδομένα τα οποία δίνουν τις απαραίτητες πληροφορίες για την σωστή λήψη αποφάσεων. Συνεπώς, εάν ως πόρισμα έρευνας προκύψει ότι οι Έλληνες καταναλωτές τείνουν να προτιμούν αυτοκίνητα με κινητήρα εσωτερικής καύσης, αποτελεί ο ορθολογική απόφαση από την πλευρά της αυτοκινητοβιομηχανία ηλεκτροκίνητων αυτοκινήτων να μην επενδύσει στην ελληνική αγορά.

Η μηχανική μάθηση είναι ένας σύγχρονος, αποδοτικός τρόπος για την επεξεργασία μεγάλων όγκων δεδομένων ώστε να παραχθούν αξιόπιστα συμπεράσματα. Πρόκειται για ένα είδος διαδικασιών τεχνητής νοημοσύνης (Artificial Intelligence - AI) που επιτρέπει στις εφαρμογές λογισμικού να προβλέπουν μέσα από αυτοματοποιημένες επεξεργασίες δεδομένων του παρελθόντος, τιμές χαρακτηριστικών αντικειμένων στο παρόν ή το μέλλον. Τα αποτελέσματα των διαδικασιών αυτών χρησιμοποιούνται για την δημιουργία συμπερασμάτων που χρησιμοποιούνται για να λαμβάνονται αποφάσεις. Οι αλγόριθμοι που υποστηρίζουν τέτοιου είδους διαδικασίες μπορούν να χρησιμοποιηθούν σε μία μεγάλη ποικιλία εφαρμογών και οποιασδήποτε κλίμακας. Αυτός είναι και ο βασικότερος λόγος για τον οποίο η μηχανική

μάθηση έχει γίνει ένας σημαντικός παράγοντας ενίσχυσης της ανταγωνιστικότητας των οργανισμών.

Σε γενικές γραμμές, η προσέγγιση της μηχανικής μάθησης βασίζεται στην προσπάθεια δημιουργίας αυτοματοποιημένων διαδικασιών οι οποίες εξελίσσονται με βάση το πρότυπο του τρόπου με τον οποίο ο άνθρωπος αποκτά τη γνώση. Καθώς η ανθρώπινη διαδικασία δημιουργίας της γνώσης είναι τέλεια και πολύπλοκή, η αντίστοιχη μηχανική, περιλαμβάνει διαφορετικές κατηγορίες τεχνικών και μεθοδολογιών, κάθε μία από τις οποίες είναι κατάλληλη να επιλύσει διαφορετικού τύπους προβλήματα, να διαχειριστεί διαφορετικού τύπου δεδομένα και να παράξει ανάλογα αποτελέσματα. Η καταλληλότητα τους ελέγχεται και αξιολογείται με βάση κατάλληλους μηχανισμούς ώστε σε κάθε πρόβλημα να χρησιμοποιούνται οι πλέον αποδοτικές. Εκτός αυτού, κατά την αξιολόγηση των εναλλακτικών μεθοδολογιών, εξετάζονται και διαφορετικές τιμές παραμέτρων και πως αυτές επηρεάζουν τα επίπεδα απόδοσης. Επομένως η μηχανική μάθηση αποτελεί ένα σύστημα που εξασφαλίζει υψηλή αποδοτικά των αποτελεσμάτων της.

Τα δεδομένα που χρησιμοποιούνται στις διαδικασίες μηχανικής μάθησης, είναι καθοριστικά για την αποδοτικότητα τους. Η ποσότητα και η ποιότητα τους θα πρέπει να είναι τέτοια που η επεξεργασία τους να δίνει αξιόπιστα αποτελέσματα. Τις περισσότερες φορές απαιτείται η προσαρμογή τους στις μορφές των εισόδων που μπορούν να διαχειριστούν οι σχετικοί αλγόριθμοι με την επεξεργασία τους από κατάλληλες διεργασίες. Από την άλλη μεριά, για κάθε περίπτωση που εφαρμόζονται οι διαδικασίες μηχανικής μάθησης, απαιτείται να παρέχονται τα αποτελέσματα σε συγκεκριμένες μορφές (πίνακες, διαγράμματα κλπ). Καθώς κάθε αλγόριθμος παρέχει τα αποτελέσματα που εξάγει σε συγκεκριμένες μορφές (πχ με την μορφή πιθανοτήτων για κάθε πιθανή κατηγορία ή με την μορφή συνεχών τιμών), επιλέγονται κατάλληλες αναπαραστάσεις τους προκειμένου να είναι όσο γίνεται πιο δηλωτικά όταν χρησιμοποιούνται σε διαδικασίες λήψης αποφάσεων.

Η μηχανική μάθηση εφαρμόζεται με επιτυχία στην ιατρική επιστήμη και πλέον απολαμβάνει σχεδόν καθολική αναγνώριση, για την αποτελεσματικότητα της, στον κλάδο της υγειονομικής περίθαλψης. Οι αλγόριθμοι της συμβάλλουν στην ανάλυση μεγάλων όγκων ιατρικών δεδομένων προκειμένου να διευκολύνεται η διαδικασία λήψης απόφασης, διαφορετικά επίπεδα και κατηγορίες σχετικών ζητημάτων. Υπό το πρίσμα αυτό, οι τρόποι με τον οποίο μπορεί να συμβάλλει η μηχανική μάθηση στην υποστήριξη της ιατρικής είναι οι εξής:

- Διάγνωση: Πρόκειται για την κυριότερη και πιο συχνά χρησιμοποιούμενη εφαρμογή της μηχανικής μάθησης στην ιατρική. Αφορά τον εντοπισμό και τη διάγνωση ασθενειών και παθήσεων μέσα από την μελέτη μεγάλων ποσοτήτων

δεδομένων τόσο από ιστορικά στοιχεία όσο και από στοιχεία ιατρικών φακέλων. Οι μηχανισμοί της μηχανικής μάθησης επιτρέπουν την ταχεία και με υψηλή ακρίβεια εκτίμηση των ασθενειών με αποτέλεσμα να λαμβάνονται έγκαιρα εύστοχες αποφάσεις για τις απαιτούμενες ενέργειες στις διαδικασίες θεραπείας.

- Ανακάλυψη και παρασκευή φαρμάκων: Η μηχανική μάθηση εφαρμόζεται στη διαδικασία ανακάλυψης φαρμάκων μέσα κυρίως από την εύρεση εναλλακτικών τρόπων για τη θεραπεία πολυπαραγοντικών ασθενειών. Στις περιπτώσεις αυτές χρησιμοποιούνται δεδομένα που σχετίζονται με την αποτελεσματικότητα διαφορετικών σκευασμάτων σε διαφορετικές ασθένειες.

- Εξατομικευμένη διάγνωση: Η τροποποίηση συμπεριφοράς είναι ένα σημαντικό μέρος της προληπτικής ιατρικής. Οι αλγόριθμοι της μηχανικής μάθησης μπορούν να εφαρμοστούν στην εξατομικευμένη μελέτη της επίδρασης των θεραπειών, με την εξέταση δεδομένων που αφορούν συγκεκριμένους ασθενείς. Επιπλέον μέσα από την παρακολούθηση του ιατρικού φακέλου του ασθενούς και τα πρότυπα ασθενειών καθίσταται εφικτός ο εύστοχος προσδιορισμός των ασθενειών και των χαρακτηριστικών τους. Αυτές οι δυνατότητες δίνουν την ευκαιρία στο ιατρικό προσωπικό να παρέχουν εύστοχες συμβουλές στους ασθενείς για προσαρμογή της συμπεριφοράς τους στις απαιτήσεις της θεραπείας τους.

- Διάγνωση με Ιατρική Απεικόνιση: Οι αλγόριθμοι της μηχανικής μάθησης μπορούν να εφαρμόζονται με επιτυχία και σε δομημένα και σε αδόμητα δεδομένα. Ως εκ τούτου παρέχουν τη δυνατότητα εξέτασης ιατρικών εικόνων και την ανάπτυξη αντίστοιχων προτύπων εντοπισμού ασθενειών. Αυτή η δυνατότητα αποκτά μεγαλύτερη αξία με την δυνατότητα που παρέχει πλέον η τεχνολογία για ταχεία παραγωγή και διαβίβαση υψηλής ποιότητας ιατρικών εικόνων.

- Έξυπνα Μητρώα Υγείας: Η μηχανική μάθηση εφαρμόζεται στην διατήρηση ενημερωμένων αρχείων υγείας, με την ανάπτυξη μεθόδων που χρησιμοποιούν διανυσματικές μηχανές και τεχνικές αναγνώρισης OCR.

- Κλινική Έρευνα: Η μηχανική μάθηση έχει πολλές εφαρμογές στον τομέα των κλινικών δοκιμών και της έρευνας. Η εφαρμογή προγνωστικών αναλύσεων που βασίζονται στους αλγορίθμους της, μπορούν να εφαρμοστούν σε δεδομένα από διάφορες πηγές όπως προηγούμενες επισκέψεις σε γιατρό ή και αναρτήσεις σε μέσα κοινωνικής δικτύωσης. Μπορούν επίσης να

παρακολουθούνται και δεδομένα σε πραγματικό χρόνο. Μέσα από την επεξεργασία των δεδομένων αυτών (σε συνδυασμό με αυτά που συγκεντρώνονται από συμβατικές πηγές) με κατάλληλους αλγορίθμους, παράγονται ισχυρά πρότυπα για την διάγνωσης και τη θεραπεία ασθενειών.

- Αξιοποίηση δεδομένων από ιατρικούς αισθητήρες: Οι μηχανισμοί της μηχανικής μάθησης έχουν τη δυνατότητα να αξιοποιούν σχετικά γρήγορα δεδομένα που προέρχονται από πληθυσμούς ασθενών. Τα πρότυπα που έχουν ήδη διαμορφωθεί χρησιμοποιούνται στις περιπτώσεις αυτές για την αξιολόγηση της πορείας της κατάστασης των ασθενών. Τα δεδομένα αυτά χρησιμοποιούνται επίσης και για την ανάπτυξη νέων προτύπων.
- Πρόβλεψη επιδημίας: Οι τεχνολογίες που βασίζονται στην τεχνητή νοημοσύνη και η μηχανική μάθηση χρησιμοποιούνται για την παρακολούθηση και την πρόβλεψη επιδημιών. Αυτό γίνεται με την εκμετάλλευση μεγάλων όγκων δεδομένων που συλλέγονται από δορυφόρους, ενημερώσεις μέσων κοινωνικής δικτύωσης σε πραγματικό χρόνο, αισθητήρες και άλλες δικτυωμένες πηγές. Τα τεχνητά νευρωνικά δίκτυα συμβάλλουν στη συλλογή αυτών των πληροφοριών, τη δημιουργία και αξιοποίηση προτύπων για την πρόβλεψη επιδημιών.

Τα τελευταία χρόνια έχουν προταθεί αρκετές τεχνικές για τη διάγνωση και πρόβλεψη του καρκίνου, που βασίζονται στη μηχανική μάθηση. Ο Cicchetti βασίστηκε σε νευρωνικά δίκτυα προκειμένου να εξετάζεται η πιθανότητα ύπαρξης καρκίνου του μαστού (Cicchetti, 1992). Έλληνες επιστήμονες εφάρμοσαν τεχνικές που βασίζονται στο θεώρημα του Bayes, se Νευρωνικά Δίκτυα, σε Support Vector Machines (SVMs), σε δένδρα απόφασης και Random Forests προκειμένου να εντοπίσουν με την μεγαλύτερη δυνατή ακρίβεια την πιθανότητα υποτροπής ασθενών με καρκίνο του στόματος (Exarchos, Goletsis, & Fotiadis, 2012). Οι συντάκτες της (Parka, και συν., 2013) εφάρμοσαν μοντέλα SVM, Νευρωνικών Δικτύων και ημιεποπετεύομενων τεχνικών, για να εκτιμήσουν την θνησιμότητα από τον καρκίνο του στήθους. Στις (Bottaci, και συν., 1997), (Philip, Pal, & Verma, 2020)εξετάστηκαν μέθοδοι που βασίζονταν σε Νευρωνικά Δίκτυα για την πρόγνωση και τη διάγνωση του καρκίνου. Η εφαρμογή της μηχανικής μάθησης σε ζητήματα διάγνωσης και πρόληψης του καρκίνου, φαίνεται ότι είναι ένα ανοικτό θέμα μελέτης για περισσότερο από δύο δεκαετίες.

Στην παρούσα εργασία, το επίκεντρο είναι η συμβολή της μηχανικής μάθησης στην πρόβλεψη της επικινδυνότητας και της εξέλιξης καρκινικών όγκων. Εξετάζεται το πως οι παραπάνω εφαρμογές της στην ιατρική, αξιοποιούνται στην προσπάθεια μείωσης της θνησιμότητας εξαιτίας του καρκίνου. Η μελέτη βασίζεται κατά ένα μεγάλο μέρος στην πρακτική αξιοποίηση

σχετικών δεδομένων με μηχανισμούς της μηχανικής μάθησης και την αξιολόγηση των αποτελεσμάτων της.

Το υπόλοιπο του παρόντος κειμένου έχει διαμορφωθεί ως εξής:

1. Κεφάλαιο 1: Στο πρώτο κεφάλαιο γίνεται μία θεωρητική προσέγγιση της μηχανικής μάθησης. Πραγματοποιείται μία επισκόπηση των μεθόδων της και των λειτουργιών της, με έμφαση σε αυτές που εμπίπτουν στην εποπτευόμενη μηχανικής μάθηση. Εξετάζονται επίσης οι τρόποι αξιολόγησης των αποτελεσμάτων των σχετικών αλγορίθμων.
2. Κεφάλαιο 2: Το κεφάλαιο αυτό είναι αφιερωμένο στο σύνολο των δεδομένων που θα χρησιμοποιηθεί για την πρακτική εφαρμογή των μεθόδων μηχανικής μάθησης. Εξετάζεται η προέλευσης τους, η πληροφορίες που εμπεριέχουν, η δομή τους καθώς και οι ενέργειες που χρειάζεται να πραγματοποιηθούν για την προεπεξεργασία τους και την προσαρμογή τους στις απαιτήσεις των αλγορίθμων που θα υλοποιηθούν.
3. Κεφάλαιο 3: Στο τρίτο κεφάλαιο εμπεριέχονται όλα τα περιγραφικά χαρακτηριστικά καθώς και η απεικόνιση των αποτελεσμάτων και οι πίνακες αυτών για την πλήρη κατανόηση τους.
4. Κεφάλαιο 4: Στο κεφάλαιο αυτό περιγράφεται η διαδικασία εφαρμογής των αλγορίθμων μηχανικής μάθησης σε ένα σύνολο δεδομένων. Εξετάζονται τέσσερις διαφορετικοί αλγόριθμοι: Λογιστική Παλινδρόμηση, Δένδρα Απόφασης, Random Forest και Support Vector Machines. Παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εφαρμογή του κάθε ενός από αυτά. Το τέλος του κεφαλαίου πραγματοποιείται μία σύγκριση της αποτελεσματικότητας της κάθε μεθόδου.
5. Συμπεράσματα: Το κείμενο κλείνει με την παράθεση των συμπερασμάτων που προκύπτουν από την εφαρμογή της μηχανικής μάθησης στην ιατρική. Γίνεται επίσης μία εκτίμηση για την δυναμική της στο μέλλον και το κατά πόσο θα συνεχίσει να προσφέρει λύσεις σε διάφορους τομείς της ανθρώπινης δραστηριότητας.

Κεφάλαιο 1: Μηχανική Μάθηση

Ορισμός – Χαρακτηριστικά

Η ανάλυση των δεδομένων που αφορούν το περιβάλλον ενός οργανισμού, είναι μία κομβική διαδικασία που εντάσσεται στις διαδικασίες λήψης αποφάσεων σε οργανισμούς κάθε μεγέθους και προσανατολισμού. Μέχρι πριν λίγα χρόνια, το βασικότερο ζήτημα που έπρεπε να αντιμετωπιστεί για να επιτευχθεί η αποδοτική ανάλυση των δεδομένων ήταν ο εντοπισμός των πιο κατάλληλων σε επαρκείς ποσότητες. Με την εκρηκτική ανάπτυξη των τεχνολογιών που σχετίζονται με το διαδίκτυο, η διαθεσιμότητα μεγάλων όγκων δεδομένων έχει αυξηθεί σε τέτοιο βαθμό που πλέον το ζήτημα που πρέπει να αντιμετωπιστεί με επιτυχία είναι η διαλογή τους και ο εντοπισμός εκείνων που μπορούν να οδηγήσουν σε αξιοποιήσιμα συμπεράσματα. Η επεξεργασία μεγάλων όγκων δεδομένων, είναι δύσκολο να αντιμετωπιστεί με συμβατικές μεθόδους κατά τις οποίες ο ανθρώπινος παράγοντας διαδραματίζει βασικό ρόλο. Παράλληλα η τεχνολογική ανάπτυξη στον τομέα της πληροφορικής, έχει καταστήσει την πρόσβαση σε μεγάλη υπολογιστική ισχύ αρκετά προσιτή για ένα ευρύ φάσμα οργανισμών αλλά και μεμονωμένων ατόμων. Για το λόγο αυτό αναζητήθηκαν τρόποι αυτοματοποίησης των σχετικών διαδικασιών, από τον επιστημονικό τομέα της τεχνητής νοημοσύνης. Μία από τις αποδοτικότερες προσεγγίσεις που χρησιμοποιήθηκαν είναι η Μηχανική Μάθηση (machine learning). Βασικό χαρακτηριστικό της προσέγγισης αυτής είναι ότι η ανάλυση των δεδομένων γίνεται με τη χρήση αναλυτικών μοντέλων που συνήθως στόχο έχουν την εκτίμηση ή την πρόβλεψη επερχόμενων καταστάσεων (Mullainathan & Spiess, 2017).

Η μηχανική μάθηση είναι μια μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί τη δημιουργία αναλυτικών μοντέλων. Αποτελεί κλάδο της τεχνητής νοημοσύνης. Βασίζεται στην παραδοχή ότι τα υπολογιστικά συστήματα μπορούν να καταστούν ικανά να μαθαίνουν μέσα από διαδικασίες που ομοιάζουν με τις αντίστοιχες διαδικασίες απόκτησης γνώσεων από τον άνθρωπο. Μέσα από τη μηχανική μάθηση τα συστήματα μπορούν να αναγνωρίζουν πρότυπα και να ενισχύουν την προσπάθεια λήψης ορθών αποφάσεων.

Οι μεθοδολογίες που χρησιμοποιούνται προέκυψαν από την αναγνώριση προτύπων και τη θεωρία ότι οι υπολογιστές μπορούν να αποκτούν εμπειρία και να ενεργούν χωρίς να είναι ρητά προγραμματισμένοι να εκτελούν συγκεκριμένες εργασίες. Το στοιχείο αυτό είναι σημαντικό καθώς τα αναλυτικά μοντέλα που κατασκευάζονται από τις διαδικασίες της μηχανικής μάθησης, μπορούν να προσαρμόζονται με την εφαρμογή νέων δεδομένων που περιγράφουν την εξέλιξη των τάσεων και των καταστάσεων. Η εφαρμογή των τρεχόντων δεδομένων στα μοντέλα που παράγονται από τις διαδικασίες μηχανικής μάθησης, παράγουν αξιόπιστες,

επαναλαμβανόμενες αποφάσεις και αποτελέσματα (Mechanistic Interpretation of Machine Learning Inference: A Fuzzy Feature Importance Fusion Approach, 2021).

Τα στάδια των μεθοδολογιών της μηχανικής μάθησης

Γενικά, κάθε διαδικασία μηχανικής μάθησης περιλαμβάνει μία σειρά από ενέργειες. Οι ενέργειες αυτές εξειδικεύονται περισσότερα για διαφορετικές μεθοδολογίες που μπορεί να ακολουθούνται. Οι ενέργειες αυτές είναι:

- Προσδιορισμός της σκοπιμότητας: Αρχικά θα πρέπει να τεθεί ο αντικειμενικό σκοπός της όλης διαδικασίας. Αυτός θα καθορίσει τις επιλογές που θα γίνουν σε κάθε ένα από τα επόμενα βήματα.
- Συλλογή Δεδομένων: Αναζητούνται και ανακτώνται τα δεδομένα τα οποία θα χρησιμοποιηθούν για την δημιουργία, εκπαίδευση και αξιολόγηση του μοντέλου πρόβλεψης. Χρησιμοποιούνται αξιόπιστες πηγές δεδομένων ή σχηματισμοί ολοκλήρωσης δεδομένων (data integration) προκειμένου να γίνει η συλλογή των καταλληλότερων δεδομένων για την παραγωγή μοντέλου με τη μέγιστη δυνατή αξιοπιστία.
- Προετοιμασία Δεδομένων: Τα δεδομένα που συγκεντρώθηκαν θα πρέπει να υποστούν προεπεξεργασία προκειμένου να πάρουν μορφή κατάλληλη για να αποτελέσουν είσοδο σε διαδικασίες μηχανικής μάθησης και ταυτόχρονα να συμβάλλουν στη δημιουργία ποιοτικού αποτελέσματος. Οι δραστηριότητες που περιλαμβάνονται συνήθως είναι: καθαρισμός από θόρυβο και ακραίες τιμές, αντιμετώπιση του φαινομένου ελλειπόντων τιμών, ανισορροπία του συνόλου ως προς την κατανομή των παρατηρήσεων σε κατηγορίες.
- Επιλογή Μοντέλου: Ανάλογα με τη σκοπιμότητα της διαδικασίας και τα χαρακτηριστικά των διαθεσίμων δεδομένων, επιλέγεται ο καταλληλότερος αλγόριθμος για την ανάπτυξη μοντέλου πρόβλεψης.
- Εκπαίδευσης Μοντέλου: Χρησιμοποιείται ένα μέρος του συνόλου δεδομένων για την εκπαίδευση του μοντέλου, σύμφωνα με τον αλγόριθμο που έχει επιλεγεί.
- Αξιολόγηση Μοντέλου: Ένα μέρος του συνόλου δεδομένων χρησιμοποιείται για την αξιολόγηση του μοντέλου που αναπτύχθηκε. Ειδικότερα για διαδικασίες εποπτευόμενης μηχανικής μάθησης εξετάζεται το κατά πόσο η πρόβλεψη είναι κοντά στο αναμενόμενο αποτέλεσμα.

- Χρησιμοποίηση του Μοντέλου: Αφού η αξιολόγηση του μοντέλου είναι ικανοποιητική, μπορεί να χρησιμοποιηθεί σε διαδικασίες πρόβλεψης (Yufeng, 2017).

Μεθοδολογίες Μηχανικής Μάθησης

Οι μεθοδολογίες που χρησιμοποιούνται στη μηχανική μάθηση για την παραγωγή αναλυτικών μοντέλων διακρίνονται στις ακόλουθες κατηγορίες:

- Εποπτευόμενη μηχανική μάθηση: Οι αλγόριθμοι που εντάσσονται σε αυτή την κατηγορία, δημιουργούν μια συνάρτηση που αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους. Για το σκοπό αυτό απαιτούνται σύνολα δεδομένων τα οποία περιγράφουν οντότητες και καταστάσεις με τη χρήση ενός πεπερασμένου αριθμού παραμέτρων (έστω ότι η i-στη παράμετρος είναι η pi). Σκοπός της όλης διαδικασίας είναι ο προσδιορισμός της τιμής μίας παραμέτρου (που επίσης περιλαμβάνεται στην περιγραφή των εγγραφών του συνόλου δεδομένων), έστω label. Στο σύνολο που χρησιμοποιείται για την παραγωγή του αναλυτικού μοντέλου, οι τιμές όλων των παραμέτρων (και της label) είναι γνωστές και αναζητείται μία συσχέτιση μεταξύ των τιμών των pi με την τιμή της label.
- Μη εποπτευόμενη μάθηση: Βασική διαφορά των αλγορίθμων αυτών σε σχέση με τους αλγορίθμους εποπτευόμενης μηχανικής μάθησης είναι ότι δεν παρέχεται η τιμή της παραμέτρου label. Αναζητείται ο τρόπος συσχέτισης των στοιχείων των συνόλων δεδομένων μεταξύ τους με βάση τις τιμές των παραμέτρων τους.
- Ημι-εποπτευόμενη μάθηση: Πρόκειται για αλγορίθμους που συνδυάζουν χαρακτηριστικά των αλγορίθμων τόσο της εποπτευόμενης όσο και της μη εποπτευόμενης μηχανικής μάθησης.
- Ενισχυτική μάθησης: Οι αλγόριθμοι αυτού του είδους προσαρμόζουν τη συμπεριφορά των υπολογιστικών συστημάτων με βάση τις συνθήκες που επικρατούν στο περιβάλλον τους. Οι συνθήκες αυτές περιγράφονται από ένα σύνολο παραμέτρων. Το ίδιο συμβαίνει με και τη συμπεριφορά των συστημάτων. Οι τιμές των δεύτερων αναπροσαρμόζονται ανάλογα με τις μεταβολές στις τιμές των πρώτων.
- Μετατροπή (Transduction): Οι αλγόριθμοι αυτού του είδους εξελίσσονται με παρόμοιο τρόπο σε σχέση με αυτούς της εποπτευόμενης μάθησης. Η διαφορά

είναι ότι δεν παράγουν μία ρητή αντιστοιχία μεταξύ παραμέτρων περιγραφής των εγγραφών και τιμής παραμέτρου – στόχο αλλά προβλέπουν τη διαμόρφωση των μελλοντικών καταστάσεων με βάση τα δεδομένα που εισάγονται, τις εξόδους που παράγονται και τις επικαιροποιημένες εισόδους. και τις νέες εισροές.

- Learning to learn: Είναι αλγόριθμοι οι οποίοι προσαρμόζουν τη συμπεριφορά του υπολογιστικού συστήματος, αποκλειστικά με βάση τα αποτελέσματα παρελθούσης συμπεριφοράς και των αποτελεσμάτων αυτής (Nasteski, 2017).

Στις επόμενες παραγράφους εξετάζονται αλγόριθμοι της εποπτευόμενης μηχανικής μάθησης.

Εποπτευόμενη μηχανική μάθηση

Η εποπτευόμενη μηχανική μάθηση (Supervised Machine Learning) είναι η κατηγορία αλγορίθμων μηχανικής μάθησης που εξετάζουν τα χαρακτηριστικά εγγραφών μεγάλων συνόλων δεδομένων και παράγουν μοντέλα στα οποία όταν εφαρμοστούν τα μελλοντικά δεδομένα, παράγονται προβλέψεις για μελλοντικές καταστάσεις. Οι αλγόριθμοι που χρησιμοποιούνται εξετάζονται ως προς την καταλληλότητα τους με βάση:

- Τη φύση και τον προσανατολισμό της έρευνας που η μηχανική μάθηση θα αποτελέσει μέρος
- Τη δομή των δεδομένων που θα χρησιμοποιηθούν
- Την ακρίβεια προβλέψεων που μπορεί να παρέχουν τα μοντέλα που παράγονται από κάθε έναν αλγόριθμό
- Η ταχύτητα με την οποία παράγονται τα μοντέλα πρόβλεψης
- Η ταχύτητα με την οποία τα μοντέλα πρόβλεψης παράγουν προβλέψεις (Akinsola, 2017)

Αλγόριθμοι Εποπτευόμενης μηχανικής μάθησης

Παλινδρόμηση – Κατηγοριοποίηση

Από τους βασικότερες διαδικασίες που εκτελούνται στο πλαίσιο της μηχανικής μάθησης είναι η κατηγοριοποίηση και η παλινδρόμηση. Κάθε μία από τις διαδικασίες αυτές επιλέγεται σε συνάρτηση με τη σκοπιμότητα της εργασίας μηχανικής μάθησης που υποστηρίζουν. Η κυριότερη διαφορά τους έγκειται στο γεγονός ότι οι διαδικασίες κατηγοριοποίησης επιχειρούν

να εντάξουν τα αντικείμενα σε ομάδες που χαρακτηρίζονται από κατηγορικές παραμέτρους ενώ η παλινδρόμηση επιχειρεί να υπολογίσει τις ακριβείς τιμές αριθμητικών παραμέτρων.

Στις επόμενες παραγράφους καταγράφονται τα βασικά τους χαρακτηριστικά.

Κατηγοριοποίηση

Η διαμέριση συνόλου οντοτήτων σε προκαθορισμένες ενότητες ονομάζεται κατηγοριοποίηση (classification). Είναι μία διαδικασία που χρησιμοποιείται σε πολλές διαφορετικές εφαρμογές στον πραγματικό κόσμο καθώς είναι απαραίτητη για τον καταμερισμό των πόρων στις διάφορες δραστηριότητες. Λεξικολογικά, ο όρος αναφέρεται σε μία συστηματική εργασία ένταξή οντοτήτων σε ομάδες με βάση ένα σύνολο κριτηρίων. Στη μηχανική μάθηση ως ταξινόμηση ορίζεται ως η διαδικασία αναγνώρισης, κατανόησης και ομαδοποίησης αντικειμένων και καταστάσεων σε προκαθορισμένες κατηγορίες. Για την αντιστοίχιση των αντικειμένων με τις προκαθορισμένες κατηγορίες, χρησιμοποιούνται οι τιμές που λαμβάνουν οι παράμετροι που τα χαρακτηρίζουν. Προκειμένου να διαφανούν οι αντιστοιχίσεις αυτές, χρησιμοποιούνται μεγάλα σύνολα που περιλαμβάνουν δεδομένα περιγραφής των αντικειμένων, για τα οποία είναι διαθέσιμη η πληροφορία σχετικά με το σε ποια κατηγορία εντάσσονται. Τα σύνολα αυτά υφίστανται επεξεργασία από εξειδικευμένους αλγορίθμους, ώστε να αναδειχθούν οι συσχετίσεις των παραμέτρων περιγραφής τους με την κατηγορία που εντάσσονται. Το αποτέλεσμα των αλγορίθμων αυτών είναι ένα σύνολο πιθανοτήτων που αφορά το ενδεχόμενο το αντικείμενο να εντάσσεται σε κάθε μία από τις διαθέσιμες κατηγορίες (Kotsiantis, Zaharakis, & Pintelas, 2007).

Οι διαδικασίες ταξινόμησης μπορούν να διακριθούν στις ακόλουθες κατηγορίες:

- Δυαδική Ταξινόμηση: Στις διαδικασίες αυτού του είδους, ο αριθμός των διαθεσίμων κατηγοριών είναι δύο. Συνήθως χρησιμοποιούνται για να δημιουργήσουν μοντέλα πρόβλεψης ένταξης των αντικειμένων σε μία κατηγορία αντικειμένων που ανήκουν σε φυσιολογικές καταστάσεις ή όχι. Μία τέτοιου είδους διαδικασία μπορεί να χρησιμοποιηθεί για τον έλεγχο ενός μηνύματος ηλεκτρονικού ταχυδρομείου για το αν είναι έγκυρο ή παραπλανητικό (Kumar & Srivastava, 2017). Οι αλγόριθμοι που συχνά χρησιμοποιούνται σε τέτοιου είδους διαδικασίες είναι η λογιστική παλινδρόμηση (Logistic Regression), ο Naïve Bayes, τα δένδρα απόφασης και τα Support Vector Machines (Soofi & Awan, 2017).

- Ταξινόμηση πολλαπλών κατηγοριών: Στις περιπτώσεις αυτών των διαδικασιών, οι διατιθέμενες κατηγορίες είναι περισσότερες από δύο. Τα αποτελέσματα των διαδικασιών αυτών επιστρέφουν τις πιθανότητες τα αντικείμενα να εμπίπτουν σε κάθε μία από τις διατιθέμενες κατηγορίες. Οι αλγόριθμοι που χρησιμοποιούνται συχνά σε τέτοιου είδους προβλήματα βασίζονται σε νευρωνικά δίκτυα και δένδρα απόφασης. Χρησιμοποιούνται επίσης συχνά η λογιστική παλινδρόμηση και τα Support Vector Machines (Aiolli & Sperduti, 2005). Σε ορισμένες περιπτώσεις μπορεί οι διαδικασίες αυτές να μεταπέσουν σε διαδικασίες δυαδικής ταξινόμησης όταν εξετάζεται η πιθανότητα ένα αντικείμενο να ανήκει σε μία συγκεκριμένη από τις κατηγορίες ή όχι ή η συγκριτική πιθανότητα να ανήκει μεταξύ δύο συγκεκριμένων κατηγοριών. Μία παραλλαγή των διαδικασιών αυτών είναι η περίπτωση όπου το αντικείμενο μπορεί να ταξινομηθεί σε περισσότερες από μία κατηγορίες. Χαρακτηριστικό παράδειγμα των διαδικασιών αυτών είναι οι διαδικασίες εντοπισμού αντικειμένων σε εικόνες. Σε μία εικόνα που εξετάζεται μπορεί να απεικονίζονται περισσότερα του ενός αντικείμενα, τα οποία μπορεί να κατατάσσονται σε διαφορετικές κατηγορίες. Στις περιπτώσεις αυτές συχνά χρησιμοποιούνται αλγόριθμοι που βασίζονται σε νευρωνικά δίκτυα (Singh, 2015).

Παλινδρόμηση

Η παλινδρόμηση, περιγράφει διαδικασίες που οι τιμές ενός συνόλου ανεξάρτητων μεταβολών συσχετίζονται συναρτησιακά με την τιμή που λαμβάνει μία άλλη εξαρτώμενη από αυτές. Η παλινδρόμηση που χρησιμοποιεί μία ανεξάρτητη μεταβλητή ονομάζεται μονομεταβλητή ανάλυση παλινδρόμησης ενώ η ανάλυση που χρησιμοποιεί περισσότερες από δύο ανεξάρτητες μεταβλητές ονομάζεται πολυμεταβλητή ανάλυση παλινδρόμησης. Ο κύριος στόχος της παλινδρόμησης είναι η κατασκευή ενός αποτελεσματικού μοντέλου για την πρόβλεψη των εξαρτημένων χαρακτηριστικών από μια δέσμη μεταβλητών ιδιοτήτων. Το αποτέλεσμα των διαδικασιών αυτών έχει τη μορφή διανύσματος, μεγέθους ίδιου με τον αριθμό των ανεξάρτητων μεταβλητών. Το διάνυσμα αυτό περιλαμβάνει συντελεστές που αντιστοιχούν σε κάθε μία από τις ανεξάρτητες μεταβλητές. Η εφαρμογή αυτών στις τιμές των παραμέτρων διαμορφώνει την τιμή της εξαρτημένης μεταβλητής με μία σχετικά καλή ακρίβεια (Kadam, Kanhere, & Mahindrakar, 2020).

Η παλινδρόμηση εξελίσσεται ως εξής:

- Οι παράμετροι που χαρακτηρίζουν τα αντικείμενα λαμβάνουν διαφορετικές τιμές σε κάθε βήμα.
- Μετά την απόδοση των τιμών, ελέγχεται το πως διαμορφώνεται η τιμή της εξαρτημένης μεταβλητής, με την εφαρμογή μίας συνάρτησης που λαμβάνει σαν ορίσματα τις τιμές των ανεξάρτητων μεταβλητών, και καταγράφεται η απόκλιση από την πραγματική τιμή (η οποία είναι διαθέσιμη).
- Οι τιμές και η συνάρτηση που καταγράφηκαν στο βήμα όπου παρατηρήθηκε η μικρότερη απόκλιση της υπολογισθείσας τιμής της εξαρτημένης σε σχέση με την πραγματική της, χρησιμοποιούνται για την παραγωγή ενός μοντέλου πρόβλεψης.

Οι τύποι διαδικασιών παλινδρόμησης είναι οι παρακάτω:

- Απλή γραμμική παλινδρόμηση: Είναι η τεχνική παλινδρόμησης στην οποία η ανεξάρτητη μεταβλητή έχει γραμμική σχέση με την εξαρτημένη μεταβλητή. Ο κύριος στόχος της απλής γραμμικής παλινδρόμησης είναι να ληφθούν υπόψη τα δεδομένα σημεία και να σχεδιαστεί η γραμμή η οποία προσαρμόζεται καλύτερα στις πραγματικές παρατηρήσεις.
- Πολλαπλή Γραμμική Παλινδρόμηση: Σε πολλές εφαρμογές, υπάρχουν περισσότεροι από ένας παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή. Τα μοντέλα πολλαπλής παλινδρόμησης περιγράφουν πώς μια μεμονωμένη εξαρτημένη μεταβλητή εξαρτάται γραμμικά από έναν αριθμό ανεξάρτητων μεταβλητών.
- Πολυωνυμική παλινδρόμηση: Με την τεχνική αυτή, μετατρέπονται τα αρχικά χαρακτηριστικά σε πολυωνυμικά ενός συγκεκριμένου βαθμού. Στην συνέχεια εκτελείται παλινδρόμηση με βάση τα προσαρμοσμένα χαρακτηριστικά.
- Support Vector Regression: Με τις τεχνικές αυτές προσδιορίζεται ένα υπερεπίπεδο με αρκετό περιθώριο έτσι ώστε ο μέγιστος αριθμός σημείων δεδομένων να είναι εντός αυτού.
- Παλινδρόμηση δέντρου αποφάσεων: Τα δένδρα αποφάσεων μπορούν να χρησιμοποιηθούν σε εργασίες παλινδρόμησης, όταν αναζητείται ο προσδιορισμός των χαρακτηριστικών που πρέπει να χρησιμοποιηθεί σε κάθε κόμβο του, ώστε να μειωθεί η τυπική απόκλιση μεταξύ υπολογιζόμενης και πραγματικής τιμής του στόχου.

- Random Forest Reggresion: Με τις τεχνικές αυτές εξετάζονται οι προβλέψεις που παράγονται από πολλαπλές παλινδρομήσεις δένδρων απόφασης (Kadam, Kanhere, & Mahindrakar, 2020).

Οι αλγόριθμοι που χρησιμοποιούνται συχνότερα σε διαδικασίες παλινδρόμησης είναι k-nearest neighbor και ο Support Vector Machine Regression. Αμφότεροι επιχειρούν να εντοπίσουν τη καλύτερη δυνατή συνάρτηση που πρέπει να εφαρμοστεί στις παραμέτρους περιγραφής των αντικειμένων, με βάση την απόσταση που προκύπτει μεταξύ της υπολογιζόμενης και της πραγματικής τιμής της εξαρτημένης μεταβλητής. Επιπλέον, οι αλγόριθμοι αυτοί καθορίζουν ένα επίπεδο ανοχής σφάλματος ακρίβειας προκειμένου να ολοκληρώνονται σε συντομότερο χρόνο.

Linear vs Non-Linear

Τα μοντέλα πρόβλεψης μπορεί να είναι γραμμικά ή μη γραμμικά. Ένα γραμμικό μοντέλο ακολουθεί μια συγκεκριμένη μορφή, όπου όλοι του οι όροι μπορεί να είναι είτε μία σταθερά ή μια παράμετρος πολλαπλασιασμένη με μια ανεξάρτητη μεταβλητή. Στην συνέχεια διαμορφώνεται μία εξίσωση μετά από την πρόσθεση όλων των όρων. Η εξίσωση αυτή έχει την παρακάτω μορφή:

$$Y = c + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

Τα μοντέλα αυτά είναι γραμμικά ως προς τις παραμέτρους τους. Αυτός σημαίνει τα γραμμικά μοντέλα μπορεί να περιλαμβάνουν στους όρους τους κάποιος που να είναι υψωμένοι σε μία δύναμη. Ένα παράδειγμα ενός τέτοιου μοντέλου θα μπορούσε να περιγράφεται από την ακόλουθη εξίσωση:

$$Y = c + b_1X_1^2 + b_2X_2 + \cdots + b_kX_k$$

Τα γραμμικά μοντέλα μπορούν επίσης να περιέχουν και αντίστροφους όρους για να ακολουθούν διαφορετικά είδη καμπυλών.

Εάν ένα μοντέλο δεν ακολουθεί τους κανόνες των γραμμικών, τότε είναι ένα μη γραμμικό μοντέλο. Στα μη γραμμικά μοντέλα περιλαμβάνεται μία ευρεία ποικιλία (Frost, 2022).

Parametric vs Non-Parametric

Τα μοντέλα μηχανικής μάθησης μπορεί να είναι παραμετρικά ή μη παραμετρικά. Τα παραμετρικά μοντέλα είναι εκείνα που απαιτούν τον καθορισμό ορισμένων παραμέτρων για να μπορέσουν να χρησιμοποιηθούν για προβλέψεις, ενώ τα μη παραμετρικά μοντέλα δεν

βασίζονται σε συγκεκριμένες ρυθμίσεις παραμέτρων και επομένως συχνά παράγουν πιο ακριβή αποτελέσματα.

Η εκπαίδευση μοντέλων μηχανικής μάθησης αφορά την εύρεση μιας προσέγγισης συνάρτησης που έχει δημιουργηθεί με χρήση μεταβλητών εισόδου και της οποίας η έξοδος αντιπροσωπεύει τη μεταβλητή απόκρισης. Ο λόγος για τον οποίο ονομάζεται «προσέγγιση συνάρτησης» είναι επειδή υπεισέρχεται ένα σφάλμα σε σχέση με την τιμή της εξόδου συνάρτησης έναντι της πραγματικής τιμής. Η τιμή του σφάλματος μπορεί να μειωθεί με τη χρήση περαιτέρω χαρακτηριστικών και τεχνικών. Μια άλλη πτυχή αυτού του σφάλματος είναι μη αναγώγιμη, καθώς αντιπροσωπεύει το τυχαίο σφάλμα.

Η διαδικασία της εκτίμησης της συνάρτησης περιλαμβάνει τα ακόλουθα δύο βήματα:

- Προσδιορισμός της συνάρτησης.
- Προσδιορισμός των παραμέτρων της συνάρτησης σε περίπτωση που πρόκειται για γραμμική συνάρτηση.

Σε περίπτωση που η συνάρτηση που προσδιορίζεται είναι γραμμική, η εκπαίδευση των μοντέλων μηχανικής εκμάθησης καταλήγει στην εκτίμηση των αντίστοιχων παραμέτρων. Τέτοια μοντέλα ονομάζονται παραμετρικά μοντέλα μηχανικής μάθησης. Τα παραμετρικά μοντέλα είναι γραμμικά μοντέλα που περιλαμβάνουν τον προσδιορισμό των παραμέτρων.

Η κατασκευή μη παραμετρικών μοντέλων δεν κάνει σαφείς υποθέσεις σχετικά με τη λειτουργική μορφή. Αντίθετα, τα μη παραμετρικά μοντέλα μπορούν να θεωρηθούν ως η προσέγγιση συνάρτησης που πλησιάζει όσο το δυνατόν πιο κοντά στα σημεία δεδομένων. Το πλεονέκτημα έναντι των παραμετρικών προσεγγίσεων είναι ότι αποφεύγοντας την υπόθεση μιας συγκεκριμένης λειτουργικής μορφής, όπως το γραμμικό μοντέλο, τα μη παραμετρικά μοντέλα έχουν τη δυνατότητα να ταιριάζουν με ακρίβεια σε ένα ευρύτερο φάσμα πιθανών σχημάτων για την πραγματική συνάρτηση. Οποιαδήποτε παραμετρική προσέγγιση, περιλαμβάνει την πιθανότητα το αντίστοιχο μοντέλο να διαφέρει από την πραγματική συνάρτηση, με αποτέλεσμα οι προβλέψεις του να αποκλίνουν από την πραγματικότητα. Χαρακτηριστικό παράδειγμα τέτοιου είδους μοντέλων είναι τα νευρωνικά δίκτυα. Στην περίπτωση των παραμετρικών μοντέλων, γίνεται η υπόθεση που σχετίζεται με τη συναρτησιακή μορφή και εξετάζεται το γραμμικό μοντέλο. Σε περίπτωση μη παραμετρικών μοντέλων, δεν γίνεται υπόθεση για τη λειτουργική μορφή.

Τα παραμετρικά μοντέλα προσαρμόζονται πολύ πιο εύκολα από τα μη παραμετρικά μοντέλα, επειδή απαιτούν μόνο την εκτίμηση ενός συνόλου παραμέτρων. Στην περίπτωση μη παραμετρικού μοντέλου, χρειάζεται να εκτιμηθεί σε πρώτο χρόνο, κάποια αυθαίρετη συνάρτηση.

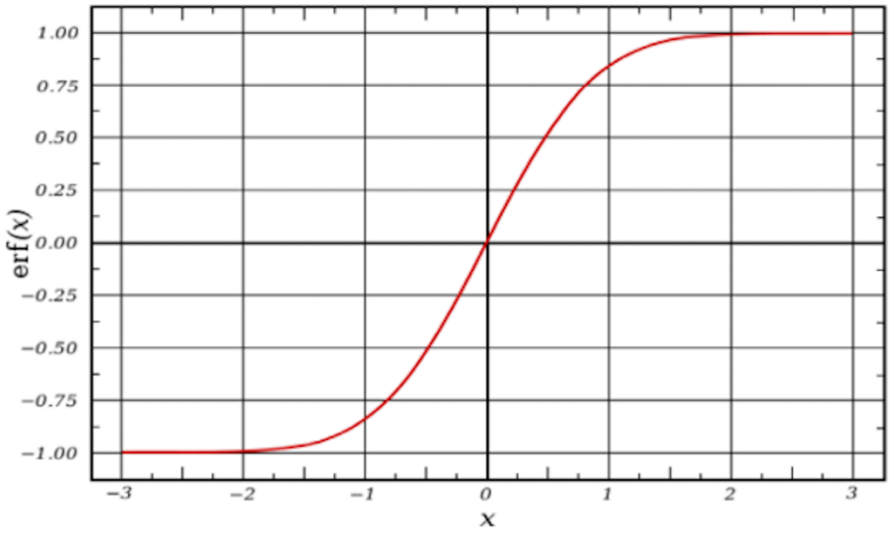
Τα παραμετρικά μοντέλα συχνά δεν ταιριάζουν με την άγνωστη συνάρτηση που προσπαθούμε να εκτιμήσουμε. Η απόδοση του μοντέλου είναι συγκριτικά χαμηλότερη από τα μη παραμετρικά μοντέλα. Οι εκτιμήσεις που γίνονται από τα παραμετρικά μοντέλα θα απέχουν περισσότερο από το να είναι αληθινές.

Τα παραμετρικά μοντέλα είναι ερμηνεύσιμα (σε αντίθεση με τα μη παραμετρικά μοντέλα) και δύνανται να χρησιμοποιηθούν για την παραγωγή συμπερασμάτων. Τα μη παραμετρικά μοντέλα είναι καταλληλότερα όταν ο κύριος στόχος είναι η πρόβλεψη ή η ερμηνεία καταστάσεων (Kumar A. , 2021).

Επισκόπηση Μεθόδων Μηχανικής Μάθησης

Logistic Regression

Η μέθοδος της λογιστικής παλινδρόμησης χρησιμοποιείται προκειμένου να αναπτυχθεί η συσχέτιση μίας εξαρτημένης μεταβλητής η οποία λαμβάνει κατηγορικές ή διακριτές αριθμητικές τιμές και μίας σειράς από ανεξάρτητες τυχαίες μεταβλητές των οποίων το πεδίο ορισμού είναι το σύνολο των πραγματικών αριθμών. Στην πράξη, η λογιστική παλινδρόμηση είναι μία γενίκευση της γραμμικής παλινδρόμησης, ώστε να μπορεί η εξαρτημένη μεταβλητή να ακολουθεί την εκθετική οικογένεια κατανομών. Ενώ στη γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων του δείγματος γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η εκτίμηση τους γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας. Με τη χρήση της μεθόδου αυτής, επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων, προκειμένου να οδηγήσουν στις πραγματικές τιμές της εξαρτημένης μεταβλητής. Με στατιστικούς όρους, η λογιστική παλινδρόμηση αναζητά την πιθανότητα εμφάνισης ενός ενδεχομένου, προσαρμόζοντας τα δεδομένα ελέγχου στη σιγμοειδή συνάρτηση, η γραφική παράσταση της οποία φαίνεται στην επόμενη εικόνα.



Εικόνα 1. 1 *Γραφική παράσταση της σιγμοειδούς συνάρτησης*

Η λογιστική παλινδρόμηση διακρίνεται σε τρεις κατηγορίες:

1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή: Συνήθως η εξαρτημένη μεταβλητή λαμβάνει δύο διαφορετικές τιμές, οπότε στόχος της μεθόδου είναι να υπολογίσει την πιθανότητα η παρατήρηση που περιγράφεται από τις εξαρτημένες μεταβλητές, να ανήκει σε μία από τις δύο κατηγορίες που αντιστοιχούν στις τιμές της εξαρτημένης. Η δυαδική λογιστική παλινδρόμηση αποτελεί μια διωνυμική εξίσωση στην οποία η μεταβλητή απόκρισης είναι το τυχαίο αποτέλεσμα εμφάνισης μιας από δύο καταστάσεις. Η μορφή της συνάρτησης της είναι:

$$f(y) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

όπου y είναι η μεταβλητή εισόδου και f(y) το αποτέλεσμα αυτής. Η μεταβλητή εισόδου λαμβάνει θετικές και αρνητικές τιμές. Το αποτέλεσμα της ωστόσο λαμβάνει τιμές στο διάστημα $[0, 1]$. Η μεταβλητή y αντανακλά τη συμπεριφορά μίας ομάδας ανεξαρτήτων μεταβλητών ενώ η $f(y)$ ορίζει το αποτέλεσμα της συμπεριφοράς αυτής. Η z που καθορίζει τη συμμετοχή της κάθε μεταβλητής στη συμπεριφορά της, υπολογίζεται από την παρακάτω σχέση:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

όπου β_0 είναι η κλίσης της καμπύλης παλινδρόμησης, β_i είναι οι συντελεστές παλινδρόμησης για τη μεταβλητή x_i. Θετική τιμή του συντελεστή δηλώνει ότι η αντίστοιχη μεταβλητή αυξάνει την πιθανότητα της πραγματοποίησης του ενδεχομένου ενώ αρνητική τη μειώνει. Όσο η τιμή απομακρύνεται από το 0, τόσο ισχυρότερη είναι η επίδραση της μεταβλητής στον καθορισμό του αποτελέσματος.

2. Τακτική (ordinal) μεταβλητή. Στις περιπτώσεις αυτές, η εξαρτημένη μεταβλητή μπορεί να πάρει τρείς ή περισσότερες διαφορετικές τιμές. Κάθε τιμή αντιστοιχεί σε κάποια κατηγορία. Μεταξύ των κατηγοριών αυτών ισχύει η έννοια της ανισότητας η οποία δημιουργεί μία κλιμάκωση τους. Οι πιθανότητες για την πραγματοποίηση κάθε ενός από τα ενδεχόμενα, εκφράζονται ως λόγος ζεύγους ακέραιων τιμών (odds). Ο αριθμητής είναι η πιθανότητα πραγματοποίησης του ενδεχομένου και ο παρονομαστής αντιστοιχεί στην πιθανότητα να μη συμβεί. Έτσι, αν p και $1-p$ οι αντίστοιχες πιθανότητες, ο λόγος των πιθανοτήτων είναι

$$\lambda = \frac{1}{1-p}$$

Με λογαρίθμηση της σχέσης αυτής θα είναι

$$logit(p) = \log\left(\frac{p}{1-p}\right) = z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Οι συντελεστές παλινδρόμησης υπολογίζονται με τη βοήθεια της εκτίμησης της μέγιστης πιθανοφάνειας[1] (Maximum Likelihood Estimate – MLE), από τη σχέση

$$L = \prod_{i=1}^{n} f(x_i\theta) = \sum_{i=1}^{n} logf(x_i\theta)$$

Όπου θ είναι παράμετρος της μεταβλητής.

3. Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυχοτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης: Έχει κοινά χαρακτηριστικά με την Τακτική, ως προς το πλήθος των διαφορετικών τιμών που μπορεί να πάρει η εξαρτημένη μεταβλητή. Οι κατηγορίες που ορίζονται σε αντιστοιχία με την εξαρτημένη μεταβλητή δεν παρουσιάζουν κάποια κλιμάκωση αλλά αποτελούν απλώς ένα χαρακτηρισμό του εξεταζόμενου αντικειμένου (Shalizi, 2012).

[1] Εκφράζει το πόσο μία μεταβλητή μπορεί να εκφράσει καλύτερα ένα αντικείμενο

Η λογιστική παλινδρόμηση χρησιμοποιείται για την εκτίμηση εξαρτημένων μεταβλητών σε αρκετά διαφορετικά επιστημονικά πεδία. Μερικές από τις κυριότερες εφαρμογές της είναι:

- Ο έλεγχος για την ύπαρξη μίας νόσου με βάση μία σειρά από αντικειμενικές μετρήσεις.
- Η πρόβλεψη επιλογών και τάσεων με βάση μία σειρά από δημογραφικά στοιχεία. (ηλικία, φύλο, φυλή, τόπος διαμονής, εισόδημα κτλ.)
- Πρόβλεψη έκβασης συγκεκριμένων διεργασιών με βάση τα δεδομένα των εισόδων τους.
- Πρόβλεψη βαθμού διείσδυσης στην αγορά για ένα νέο προϊόν.
- Πρόβλεψη για το κατά πόσο ένας δανειολήπτης θα φανεί συνεπείς στις υποχρεώσεις αποπληρωμής του δανείου του.

Δέντρα αποφάσεων

Οι διαδικασίες που βασίζονται στα δένδρα αποφάσεων εξελίσσονται με την διαδοχική λήψη αποφάσεων σε κάθε στάδιο τους. Η απόφαση σε κάθε στάδιο αφορά την πραγματοποίηση μίας επιλογής από ένα σύνολο διαθεσίμων. Η διαδικασία ολοκληρώνεται σε μία από τις έγκυρες τελικές καταστάσεις. Η εξέλιξη των αλγορίθμων των δένδρων απόφασης δημιουργεί μία δεντρική δομή, όπου κάθε αντικείμενο ακολουθεί ένα μονοπάτι από τη ρίζα προς έναν από τα φύλλα του δένδρου.

Ένα δέντρο περιλαμβάνει ρίζα, κλάδους και φύλλα. Η ίδια δομή ακολουθείται και στο δένδρο απόφασης. Κάθε εσωτερικός κόμβος αντιστοιχεί στην εξέταση της τιμής μίας ιδιότητας. Από κάθε εσωτερικό κόμβο εξέρχονται ένας ή περισσότεροι κλάδοι, ο οποίος αντιστοιχεί σε κάθε μία από τις τιμές που μπορεί να λάβει η ιδιότητα που εξετάζεται στον εσωτερικό κόμβο. Στα φύλλα του δένδρου τοποθετούνται οι κατηγορίες στις οποίες μπορεί να ενταχθεί κάθε ένα από τα αντικείμενα. Συνοπτικά τα δομικά στοιχεία ενός δένδρου απόφασης περιλαμβάνουν:

- Κόμβους Απόφασης: Πρόκειται για τους εσωτερικούς κόμβους του δένδρου που αντιστοιχούν στον έλεγχο μίας ιδιότητας των αντικειμένων.
- Κόμβους Φύλλα: Αντιστοιχούν στις εναλλακτικές κατατάξεις του κάθε αντικειμένου.
- Κλάδοι: Είναι οι σύνδεσμοι μεταξύ κόμβων που αντιστοιχούν Σύνδεσμοι μεταξύ κόμβων.

Η διαδικασία που περιεγράφηκε παραπάνω, ομοιάζει σε μεγάλο βαθμό με τις διαδικασίες λήψης αποφάσεων από ανθρώπους. Οι διαδικασίες αυτές εξετάζουν σε κάθε βήμα τις διαθέσιμες εναλλακτικές λύσεις και επιλέγουν την πιο συμφέρουσα, μέχρι να καταλήξουν στην τελική απόφαση.

Η αποδοτικότητα των δένδρων απόφασης είναι συνάρτηση της δομής τους. Όσο μεγαλύτερο είναι το βάθος τους, τόσο περισσότερο αργούν να καταλήξουν σε τελική απόφαση. Επομένως, η βασικότερη επιδίωξη είναι, το ύψος των δένδρων αποφάσεων να παραμένει όσο γίνεται πιο περιορισμένο. Ο περιορισμός του ύψους του δένδρου, περιορίζει ανάλογα και τις ιδιότητες που θα πρέπει να ελεγχθούν, προκειμένου να κατανεμηθεί στην κατάλληλη κατηγορία το κάθε αντικείμενο. Αυτό επιτυγχάνεται με τον προσδιορισμό της καταλληλότερης σειράς που θα πρέπει να εξεταστούν οι ιδιότητες του αντικειμένου. Η πιο κοινή μέθοδος για την επιλογή της καταλληλότερης ιδιότητας ελέγχου σε κάθε στάδιο, είναι ο υπολογισμός του πληροφοριακού κέρδους που προκύπτει με την κάθε επιλογή. Αν σε κάποιο από τα στάδια ανάπτυξης του δένδρου απόφασης απαιτηθεί ο προσδιορισμός της ιδιότητας που θα πρέπει να επιλεγεί στην συνέχεια, τα βήματα που ακολουθούνται είναι τα εξής:

- Υπολογίζεται η εντροπία[2] της πληροφορίας που περιλαμβάνει κάθε μία από τις επιλογές που διατίθενται για την επόμενη ιδιότητα που θα εξεταστεί. Η εντροπία της ιδιότητας της οποίας οι διαφορετικές τιμές του πεδίου ορισμού της είναι:

$$\{v_1, v_2, \dots, v_m),$$

και οι κατηγορίες που μπορεί να ενταχθεί το αντικείμενο είναι οι:

$$\{c_1, c_2, \dots c_n\},$$

δίνεται από τη σχέση:

$$H(X) = -\sum_{i=1}^{n} p_i lg p_i = \sum_{i=1}^{n} p_i lg \frac{1}{p_i}$$

Όπου pi είναι η πιθανότητα να τοποθετηθεί το αντικείμενο στην κατηγορία i.

- Από τον κόμβο που αντιστοιχεί στην εξεταζόμενη επιλογή, εξέρχονται m κλάδοι. Για κάθε κλάδο και τον αντίστοιχο κόμβο που διαμορφώνεται, υπολογίζονται με τον ίδιο τρόπο οι εντροπίες της πληροφορίας που περιλαμβάνουν.

[2] Η εντροπία είναι συνάρτηση του αριθμού των πιθανών καταστάσεων του συστήματος και ενός φυσικού μέτρου της βεβαιότητάς της τρέχουσας κατάστασης του.

- Υπολογίζεται ο μέσος όρος των τιμών της εντροπίας που υπολογίστηκε στο προηγούμενο βήμα.
- Αφαιρείται από την εντροπία του αρχικού κόμβου, ο μέσος όρος της εντροπίας των θυγατρικών κόμβων. Η διαφορά αυτή είναι το πληροφοριακό κέρδος που θα προκύψει αν επιλεγεί για έλεγχο η ιδιότητα αυτή. Όσο μεγαλύτερη είναι η διαφορά στην εντροπία μεταξύ πατρικού και θυγατρικών κόμβων, τόσο μειώνεται η αβεβαιότητα.
- Επιλέγεται η ιδιότητα η οποία παρουσιάζει το μεγαλύτερο πληροφοριακό κέρδος (Loh, 2011)

Σε περιπτώσεις όπου οι επιλεγμένες ιδιότητες μπορούν να λάβουν μία τιμή από ένα ευρύ σύνολο διαθεσίμων, τότε επιλέγεται η μετρική του λόγου του πληροφοριακού κέρδους, ο οποίος ορίζεται ως:

Αν

$$SI(X,a) = -p(y1|a)lg\ p(y1|a) - p(y2|a)lg\ p(y2|a) -\cdot - p(yc\ |a)\ lg\ p(yc\ |a)$$

Για ένα χαρακτηριστικό X και χαρακτηριστικό που ελέγχεται α το οποίο μπορεί να πάρει c διαφορετικές τιμές (το $y_i|a$ είναι το ποσοστό των οντοτήτων του συνόλου με συγκεκριμένη τιμή της α)

Ο λόγος κέρδους (gain ratio) τότε είναι:

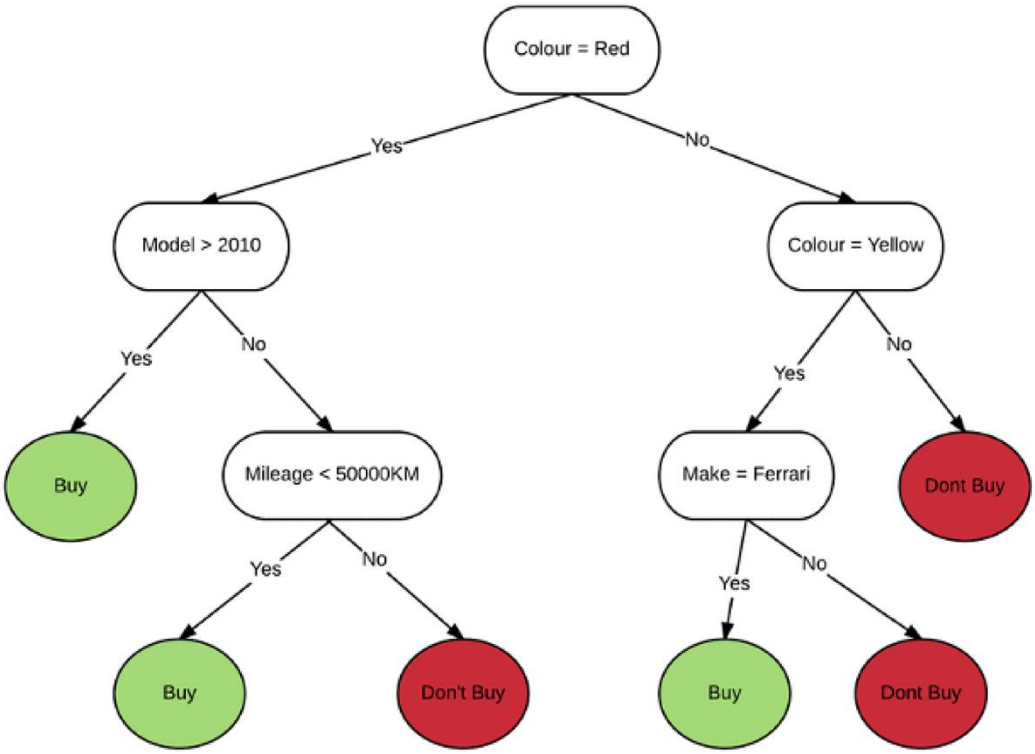
$$GR(X,A) = \frac{ΠΛΗΡΟΦΟΡΙΑΚΟ\ ΚΕΡΔΟΣ}{SI(X,a)}$$

Με τον τρόπο αυτό αποφεύγονται οι επιλογές ιδιοτήτων με υπερβολικά πολλές διακλαδώσεις.

Η αναδρομική διαδικασία εντοπισμού των καταλληλότερων χαρακτηριστικών ελέγχου, ολοκληρώνεται όταν:

- Προκύπτει κόμβος με αντικείμενα που ανήκουν σε μία κατηγορία. Στην περίπτωση αυτή ο στόχος της διαδικασίας έχει επιτευχθεί.
- Έχει ολοκληρωθεί ο έλεγχος όλων των ιδιοτήτων. Τότε ο κόμβος γίνεται φύλλο του δένδρου και η κατηγορία που του αντιστοιχίζεται είναι αυτή της πλειοψηφίας των αντικειμένων που περιλαμβάνει.
- Ο κόμβος που προκύπτει δεν περιλαμβάνει κανένα αντικείμενο. Τότε ο κόμβος γίνεται φύλλο και λαμβάνει τον χαρακτηρισμό της πλειοψηφίας του γονικού κόμβου (Louridas, 2020).

Στο παρακάτω σχήμα παρουσιάζεται η λειτουργία ενός δένδρου για την απόφαση αγοράς ενός αυτοκινήτου.



Εικόνα 1. 2 *Παράδειγμα δένδρου απόφασης*

Ο λήπτης απόφασης φθάνει στην αγορά αυτοκινήτου αν:

- Είναι κόκκινο και έχει κατασκευαστεί μετά το 2010
- Είναι κόκκινο και έχει διανύσει λιγότερα από 50.000km
- Είναι κίτρινο και Ferrari.

Στο δένδρο αυτό, η ιδιότητα με το μεγαλύτερο πληροφοριακό κέρδος είναι το χρώμα του αυτοκινήτου (για αυτό το λόγο και ελέγχεται σε πρώτο επίπεδο).

Random Forest

Το Random Forest είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης ο οποίος συνδυάζει την έξοδο πολλαπλών δέντρων αποφάσεων για να φτάσει σε ένα μόνο αποτέλεσμα. Μπορεί να χρησιμοποιηθεί τόσο σε προβλήματα ταξινόμησης όσο και παλινδρόμησης. Τα δέντρα απόφασης είναι κοινοί εποπτευόμενοι αλγόριθμοι μάθησης, που μπορεί να είναι επιρρεπείς σε προβλήματα, όπως μεροληψία και υπερπροσαρμογή. Ωστόσο, όταν πολλά δέντρα απόφασης σχηματίζουν ένα σύνολο στον αλγόριθμο τυχαίων δασών,

προβλέπουν πιο ακριβή αποτελέσματα, ιδιαίτερα όταν τα μεμονωμένα δέντρα δεν είναι συσχετισμένα μεταξύ τους.

Οι μέθοδοι μηχανικής μάθησης που βασίζονται σε σύνολα ταξινομητών και οι προβλέψεις τους συγκεντρώνονται για να προσδιορίσουν το πιο δημοφιλές αποτέλεσμα. Ο αλγόριθμος Random Forest ανήκει στην κατηγορία αυτή. Χρησιμοποιεί την τυχαιότητα χαρακτηριστικών για να δημιουργήσει ένα ασύνδετο δάσος δέντρων απόφασης.

Η τυχαιότητα χαρακτηριστικών, δημιουργεί ένα τυχαίο υποσύνολο χαρακτηριστικών, το οποίο εξασφαλίζει χαμηλή συσχέτιση μεταξύ των δέντρων αποφάσεων. Οι πιο γνωστές μέθοδοι συνόλου είναι το bagging, και το boosting. Με την πρώτη μέθοδο, ένα τυχαίο δείγμα δεδομένων σε ένα σύνολο εκπαίδευσης επιλέγεται. Αφού δημιουργηθούν πολλά δείγματα δεδομένων, αυτά τα μοντέλα στη συνέχεια εκπαιδεύονται ανεξάρτητα, και ανάλογα με τον τύπο της εργασίας, ο μέσος όρος ή η πλειονότητα αυτών των προβλέψεων αποδίδουν μια πιο ακριβή εκτίμηση. Αυτή η προσέγγιση χρησιμοποιείται συνήθως για τη μείωση της διακύμανσης μέσα σε ένα θορυβώδες σύνολο δεδομένων.

Η χρήση συνόλου ταξινομητών, είναι μια βασική διαφορά μεταξύ των δέντρων απόφασης και των τυχαίων δασών. Ενώ τα δέντρα απόφασης λαμβάνουν υπόψη όλες τις πιθανές διαιρέσεις χαρακτηριστικών, τα τυχαία δάση επιλέγουν μόνο ένα υποσύνολο αυτών των χαρακτηριστικών.

Οι αλγόριθμοι Random Forest εξετάζουν τρεις κύριες υπερπαραμέτρους:

- Το μέγεθος του κόμβου
- Τον αριθμό των δέντρων
- Τον αριθμό των χαρακτηριστικών που ελήφθησαν δειγματοληπτικά.

Ο αλγόριθμος Random Forest αποτελείται από μια συλλογή δέντρων αποφάσεων και κάθε δέντρο στο σύνολο αποτελείται από ένα δείγμα δεδομένων που προέρχεται από ένα σύνολο εκπαίδευσης με αντικατάσταση, που ονομάζεται δείγμα εκκίνησης. Από αυτό το δείγμα εκπαίδευσης, το ένα τρίτο θα χρησιμοποιηθεί για την δοκιμή της αποτελεσματικότητας του. Τυχαιότητα υπεισέρχεται και μέσω της χρήσης bagging χαρακτηριστικών, προσθέτοντας περισσότερη ποικιλομορφία στο σύνολο δεδομένων και μειώνοντας τη συσχέτιση μεταξύ των δέντρων αποφάσεων. Ανάλογα με τον τύπο του προβλήματος, ο προσδιορισμός της πρόβλεψης ποικίλλει. Για μια εργασία παλινδρόμησης, θα υπολογιστεί ο μέσος όρος των μεμονωμένων δέντρων απόφασης και για μια εργασία ταξινόμησης, η πλειοψηφία - δηλαδή η πιο συχνή κατηγορική μεταβλητή - θα δώσει την προβλεπόμενη κλάση. Στο τέλος, το τμήμα των δοκιμαστικών δεδομένων χρησιμοποιείται για την πραγματοποίηση δοκιμών.

Οι προϋποθέσεις για να έχει καλή απόδοση ο αλγόριθμος Random Forest είναι:

- Τα δεδομένα που χρησιμοποιούντα για την κατασκευή των μοντέλων πρόβλεψης πρέπει να περιλαμβάνουν υψηλό βαθμό σημασιολογίας.
- Οι προβλέψεις που γίνονται από τα μεμονωμένα δέντρα πρέπει να έχουν χαμηλές συσχετίσεις μεταξύ τους[3].

Τα πλεονεκτήματα που προκύπτουν από τη χρήση του αλγορίθμου Random Forest είναι τα παρακάτω:

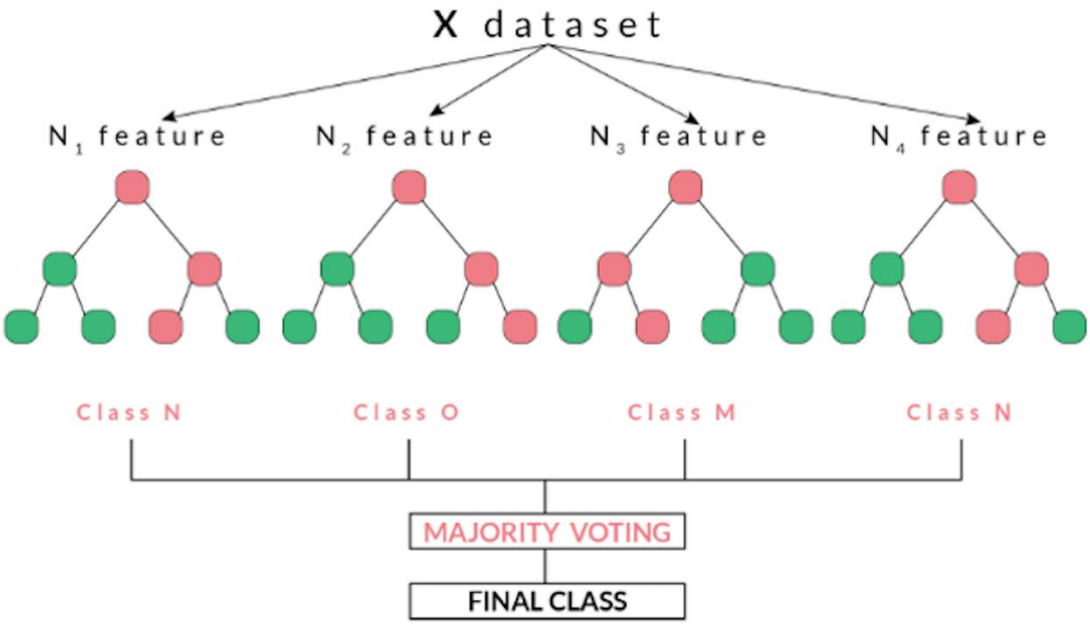
- Περιορίζεται η πιθανότητα υπερπροσαρμογής: Καθώς χρησιμοποιείται ένα σύνολο από δέντρα απόφασης, η πιθανότητα για υπερπροσαρμογή μειώνεται όσο αυξάνεται ο όγκος αυτού του συνόλου.
- Παρέχει ευελιξία: Μπορεί να φέρει εις πέρας εργασίες παλινδρόμησης και ταξινόμησης με υψηλό βαθμό ακρίβειας. Επιπλέον, η συσσώρευση χαρακτηριστικών καθιστά επίσης τον Random Forest ένα αποτελεσματικό εργαλείο για την εκτίμηση των τιμών που λείπουν, καθώς διατηρεί την ακρίβεια όταν λείπει ένα μέρος των δεδομένων.
- Ικανότητα προσδιορισμού της σημασίας χαρακτηριστικών: Διευκολύνει την αξιολόγηση της σημασίας ή της συνεισφοράς της μεταβλητής στο μοντέλο. Διατίθεται μετρικές για τον καθορισμό του πόσο μειώνεται η ακρίβεια του μοντέλου όταν εξαιρείται μια δεδομένη μεταβλητή ή με την μεταβολή της σύνθεσης του συνόλου δοκιμής.

Στα μειονεκτήματα της προσέγγισης των Random Forests είναι:

- Η ολοκλήρωση των εργασιών που τρέχουν σε αλγορίθμους Random Forest, αργεί, καθώς η επεξεργασία των δεδομένων πρέπει να περάσου από μία σειρά δένδρων αποφάσεων.
- Πρόκειται για αλγόριθμο απαιτητικό σε πόρους.
- Η ανάπτυξη και η εκτέλεση τους παρουσιάζουν υψηλή πολυπλοκότητα (Yang, 2022).

Στο παρακάτω σχήμα παρουσιάζεται σχηματικά η γενική λειτουργία των αλγορίθμων Random Forests.

[3] Γιατί με τον τρόπο αυτό μειώνονται και οι πιθανότητες συσχέτισης των λαθών.



Εικόνα 1. 3 *Παράδειγμα Random Forest*

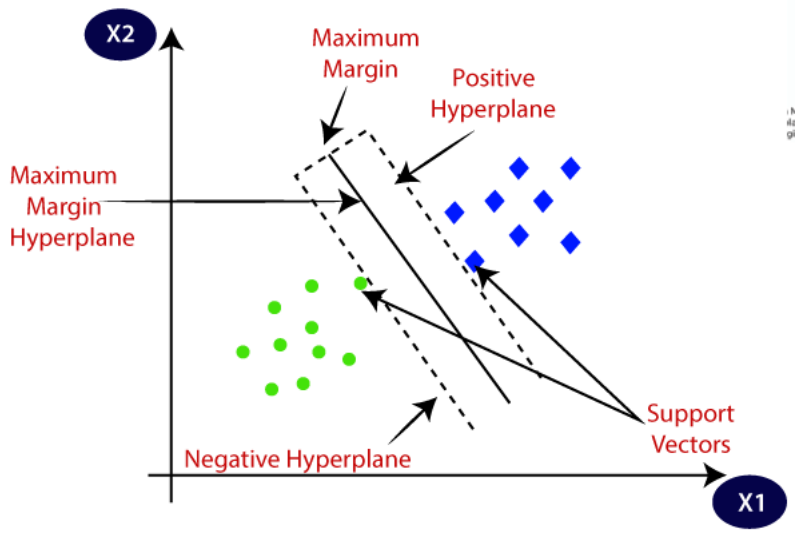
Support Vector Machines

Το Support Vector Machine, που συντομεύεται ως SVM, μπορεί να χρησιμοποιηθεί τόσο για εργασίες παλινδρόμησης όσο και για εργασίες ταξινόμησης. Ο στόχος του αλγόριθμου είναι να εντοπίσει ένα υπερεπίπεδο σε ένα χώρο N - διαστάσεων (N — ο αριθμός των χαρακτηριστικών) που ταξινομεί ευδιάκριτα τα σημεία που αντιστοιχούν σε εγγραφές. Για να διαχωριστούν δύο κατηγορίες σημείων δεδομένων, θα πρέπει να εξεταστούν ένας μεγάλος αριθμός πιθανών υπερεπιπέδων για επιλογή. Στόχος είναι να εντοπιστεί ένα επίπεδο που έχει το μέγιστο περιθώριο, δηλαδή τη μέγιστη απόσταση μεταξύ των σημείων δεδομένων και των δύο κατηγοριών. Η μεγιστοποίηση της απόστασης περιθωρίου παρέχει ένα βαθμό ασφαλείας – επιβεβαίωσης ότι μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη ακρίβεια.

Τα υπερεπίπεδα είναι τα όρια απόφασης. Αυτά ελέγχονται κατά τη διαδικασία ταξινόμησης των σημείων που αντιστοιχούν στα δεδομένα. Τα σημεία δεδομένων που εμπίπτουν σε κάθε πλευρά του υπερεπίπεδου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Η διάσταση του υπερεπίπεδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 2, τότε το υπερεπίπεδο είναι μια γραμμή ενώ αν ο αριθμός των χαρακτηριστικών εισόδου είναι 3, τότε το υπερεπίπεδο εκτείνεται στο δισδιάστατο επίπεδο (Bagchi, 2022).

Τα διανύσματα υποστήριξης είναι σημεία δεδομένων που βρίσκονται πιο κοντά στο υπερεπίπεδο και επηρεάζουν τη θέση και τον προσανατολισμό του. Χρησιμοποιώντας αυτά τα διανύσματα, επιδιώκεται η μεγιστοποίηση του περιθωρίου του ταξινομιτή. Διαγραφή των διανυσμάτων υποστήριξης προκαλεί αλλαγή στη θέση του υπερεπίπεδου. Με τη χρήση της μεθόδου των SVM, λαμβάνεται η έξοδος της γραμμικής συνάρτησης. Αν αυτή η έξοδος είναι μεγαλύτερη από 1, η αντίστοιχη παρατήρηση αποδίδεται σε μια κλάση. Αν η έξοδος είναι -1, τοποθετείται σε μία άλλη κλάση. προσδιορίζουμε ότι είναι με μια άλλη κλάση. Εφόσον οι τιμές κατωφλίου αλλάζουν καθορίζονται σε 1 και -1, το εύρος τω τιμών ενίσχυσης ακρίβειας των υπερεπιπέδων τοποθετείται στο διάστημα $[-1, 1]$, το οποίο λειτουργεί ως περιθώριο. Η μεγιστοποίηση του περιθωρίου μεταξύ των σημείων δεδομένων και του υπερεπίπεδου επιδιώκεται με τη χρήση της συνάρτησης απώλειας άρθρωσης (Ng, 2022).

Στην επόμενη εικόνα παρουσιάζονται τα κύρια χαρακτηριστικά των Support Vector Machines.



Εικόνα 1. 4 *Τα βασικά χαρακτηριστικά των Support Vector Machines*

Για τη χρήση των SVM σε προβλήματα πολλαπλών κλάσεων, δημιουργείται ένας δυαδικός ταξινομιτής για κάθε κατηγορία δεδομένων. Τα δύο αποτελέσματα κάθε ταξινομιτή θα είναι:

- Το σημείο δεδομένων ανήκει στην εξεταζόμενη κλάση.
- Το σημείο δεδομένων δεν ανήκει σε αυτήν την κλάση.

Οι αλγόριθμοι SVM μπορεί να είναι δύο ειδών:

- Γραμμικό SVM: Το γραμμικό SVM χρησιμοποιείται για γραμμικά διαχωρίσιμα δεδομένα. Πρόκειται για περιπτώσεις όπου εάν ένα σύνολο

δεδομένων μπορεί να ταξινομηθεί σε δύο κατηγορίες χρησιμοποιώντας μια ευθεία γραμμή.

- Μη γραμμικό SVM: Το μη γραμμικό SVM χρησιμοποιείται για μη γραμμικά διαχωρισμένα δεδομένα (Bagchi, 2022).

Μέτρα Αξιολόγησης Υποδείγματος

Μετά τη δημιουργία ενός μοντέλου πρόβλεψης, ακολουθεί η αξιολόγηση του. Στις εργασίες κατηγοριοποίησης το βασικό ζητούμενο είναι ένα μοντέλο πρόβλεψης να τοποθετεί την κάθε πρόβλεψη στη σωστή κλάση. Επομένως η ποιότητα του ελέγχεται κυρίως από την βαθμό που επιτυγχάνει αυτό το βασικό σκοπό. Πέραν αυτού, ανάλογα με τη σκοπιμότητα ανάπτυξης του κάθε μοντέλου, μπορεί να εξετάζονται και μία σειρά από άλλες παραμέτρους, προκειμένου να αξιολογηθεί η καταλληλότητα του μοντέλου να χρησιμοποιηθεί για το σκοπό που προορίζεται. Σε κάθε περίπτωση, για την αξιολόγηση ενός μοντέλου αξιολογούνται τα παρακάτω απόλυτα μέτρα:

- Για μία κατηγορία Α εξετάζεται πόσα δείγματα τοποθετήθηκαν στην κατηγορία Α και ανήκαν πραγματικά στην κατηγορία Α. Τα δείγμα αυτά χαρακτηρίζονται ως TRUE POSITIVE.
- Για μία κατηγορία Α εξετάζεται πόσα δείγματα τοποθετήθηκαν στην κατηγορία Α και δεν ανήκαν πραγματικά στην κατηγορία Α. Τα δείγμα αυτά χαρακτηρίζονται ως FALSE POSITIVE.
- Για μία κατηγορία Α εξετάζεται πόσα δείγματα δεν τοποθετήθηκαν στην κατηγορία Α και ανήκαν πραγματικά στην κατηγορία Α. Τα δείγμα αυτά χαρακτηρίζονται ως FALSE NEGATIVE.
- Για μία κατηγορία Α εξετάζεται πόσα δείγματα δεν τοποθετήθηκαν στην κατηγορία Α και δεν ανήκαν πραγματικά στην κατηγορία Α. Τα δείγμα αυτά χαρακτηρίζονται ως TRUE POSITIVE (Karimi, 2021).

Confusion Matrix

Με βάση τα μέτρα που αναφέρθηκαν παραπάνω κατασκευάζεται ο δισδιάστατος πίνακας σύγχυσης (Confusion Matrix) ως εξής:

- Κάθε γραμμή του πίνακα αντιστοιχεί στις προβλέψεις που πραγματοποιήθηκαν για κάθε μία από τις κατηγορίες.

- Κάθε στήλη του πίνακα αντιστοιχεί στα δείγματα που τοποθετούνται πραγματικά στην κάθε κατηγορία.

Σύμφωνα με αυτή την περιγραφή του πίνακα, στην κύρια διαγώνιο του, περιλαμβάνονται οι TRUE POSITIVE προβλέψεις για κάθε μία από τις κατηγορίες. Με εξαίρεση τα κελιά της κύριας διαγώνιου, όλα τα υπόλοιπα περιλαμβάνουν εσφαλμένες προβλέψεις. Σε κάθε γραμμή φαίνεται ποιες από τις προβλέψεις που αφορούσαν την κατηγορία που αντιστοιχεί σε αυτή, στην πραγματικότητα αντιστοιχεί στην κατηγορία της αντίστοιχης στήλης.

Ο πίνακας σύγχυσης παρέχει μία γενική συνοπτική εικόνα της απόδοσης του μοντέλου. Οι πληροφορίες που περιλαμβάνονται σε αυτόν, χρησιμοποιούνται για τον υπολογισμό πιο εξειδικευμένων μετρικών αξιολόγησης.

Έστω ότι χρησιμοποιείται ένα μοντέλο πρόβλεψης για τρεις κατηγορίες Α, Β και Γ. Σε κάθε μία από τις κατηγορίες αυτές ανήκουν 20, 30 και 50 δείγματα. Από τα 20 που ανήκουν στην κατηγορία Α, 12 προβλέφθηκε ότι ανήκουν όντως στην κατηγορία Α και από 4 στην κατηγορία Β και Γ. Από τα 30 που ανήκουν στην κατηγορία Β, 25 προβλέφθηκε ότι ανήκουν όντως στην κατηγορία Β, 1 στην Α και 4 στην κατηγορία Γ. Τέλος, από τα 50 που ανήκουν στην κατηγορία Γ, 35 προβλέφθηκε ότι ανήκουν όντως στην κατηγορία Γ, 10 στην Α και 5 στην κατηγορία Β (Karimi, 2021). Ο πίνακας σύγχυσης που αντιστοιχεί στο στιγμιότυπο αυτό είναι ο παρακάτω:

	Α	**Β**	**Γ**	**ΣΥΝΟΛΟ**
Α	12	4	4	20
Β	1	25	4	30
Γ	10	5	35	50
ΣΥΝΟΛΟ	23	34	43	100

Πίνακας 1. 1 *Παράδειγμα Πίνακα Σύγχυσης*

Ακρίβεια (Accuracy)

Η ακρίβεια ταξινόμησης είναι η βασικότερη μετρική αξιολόγησης των μοντέλων πρόβλεψης. Ορίζεται ως ο αριθμός των σωστών προβλέψεων διαιρεμένος με τον συνολικό αριθμό των προβλέψεων (Karimi, 2021).

$$Accuracy = \frac{Ορθές\ Προβλέψεις}{Σύνολο\ Προβλέψεων}$$

Στο παραπάνω παράδειγμα, η ακρίβεια του μοντέλου είναι το άθροισμα των True Positives για κάθε κλάση δια το πλήθος των συμμετεχόντων, δηλαδή:

$$\frac{72}{100} = 0.72$$

Ορθότητα (Precision)

Η μετρική αυτή εστιάζει στα σφάλματα τύπου I που αντιστοιχούν σε False Positive προβλέψεις. Οι τιμές που λαμβάνει η μετρική κινούνται στο διάστημα [0,1]. Τιμή κοντά στο 1 ότι οι προβλέψεις που αφορούν τη συγκεκριμένη κατηγορία είναι ακριβείς ή διαφορετικά ότι το μοντέλο έχει την ικανότητα να διακρίνει με επιτυχία τη συγκεκριμένη κατηγορία. Αντίθετα, τιμές κοντά στο 0 φανερώνουν αδυναμία του μοντέλου να προβλέψει με επιτυχία τη συγκεκριμένη κατηγορία. Για να έχει νόημα η χρησιμοποίηση της μετρικής για την αξιολόγηση του μοντέλου, θα πρέπει το δείγμα να είναι ισορροπημένο ως προς τις πραγματικές τιμές του χαρακτηριστικού – στόχος. Η τιμή του Precision δίνεται από την παρακάτω σχέση (Karimi, 2021).

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

Η ακρίβεια για κάθε κλάση είναι:

Α: $\frac{12}{23} = 0.52$

Β: $\frac{25}{34} = 0.73$

Γ: $\frac{35}{43} = 0.81$

Ανάκληση (Recall)

Η μέτρηση ανάκλησης εστιάζει σε σφάλματα τύπου II (FN) που αφορούν την αδυναμία του μοντέλου να προβλέψουν μία συγκεκριμένη κλάση. Οι τιμές που λαμβάνει η μετρική κινούνται στο διάστημα [0,1]. Τιμές κοντά στο 1 δείχνουν ότι το μοντέλο είναι ικανό να διακρίνει σωστά μία κατηγορία και να μη κάνει εσφαλμένες αρνητικές προβλέψεις. Οι τιμές κοντά στο 0

φανερώνουν αδυναμία να προβλέψει σωστά την κατηγορία. Η τιμή δίνεται από την παρακάτω σχέση.

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

Η ανάκληση για κάθε κλάση είναι:

Α: $\frac{12}{20} = 0.6$

Β: $\frac{25}{30} = 0.83$

Γ: $\frac{35}{50} = 0.7$

F1-Score

Πρόκειται για μία μετρική που συνδυάζει την ακρίβεια και την ανάκληση. Χρησιμοποιείται για να δώσει μία γενικότερη αξιολόγηση της ικανότητας του μοντέλου να προβλέψει τις κατηγορίες. Η τιμή του δίνεται από τη σχέση:

$$F1 = \frac{2}{\frac{1}{PRECISION} + \frac{1}{RECALL}}$$

Η μέγιστη τιμή που μπορεί να λάβει η μετρική είναι η μονάδα. Η τιμή αυτή αντιστοιχεί σε υψηλή ακρίβεια και υψηλή ανάκληση για την κατηγορία. Η μετρική αυτή δεν δίνει σαφείς πληροφορίες αν η τιμή της είναι χαμηλή.

Το F1-Score για κάθε κλάση είναι:

Α: 0.55

Β: 0.77

Γ: 0.76

AUROC (Εμβαδόν κάτω από την καμπύλη ROC)

Μια καμπύλη ROC (χαρακτηριστική καμπύλη λειτουργίας δέκτη) είναι ένα γράφημα που δείχνει την απόδοση ενός μοντέλου ταξινόμησης σε όλα τα σύνορα απόφασης. Αυτή η καμπύλη απεικονίζει δύο παραμέτρους:

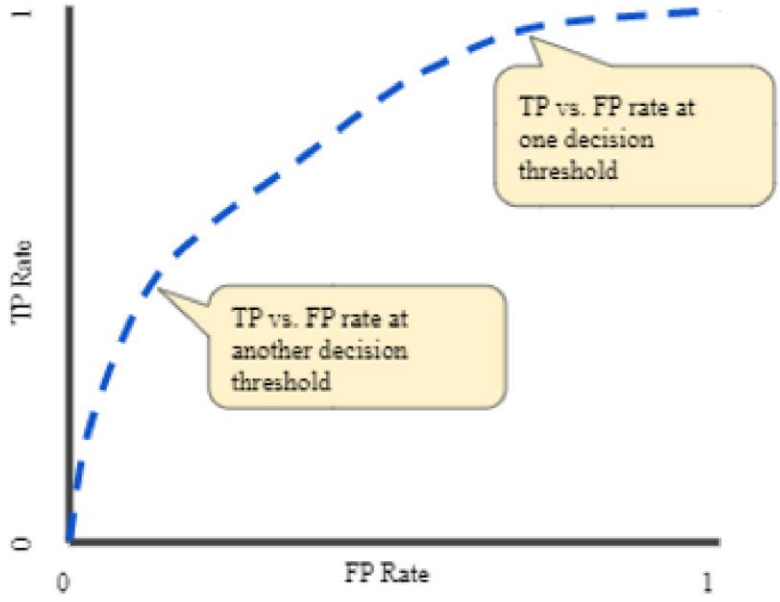
Το ποσοστό των True Positive προβλέψεων που εκφράζεται από τη σχέση

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

Το ποσοστό των False Positive προβλέψεων

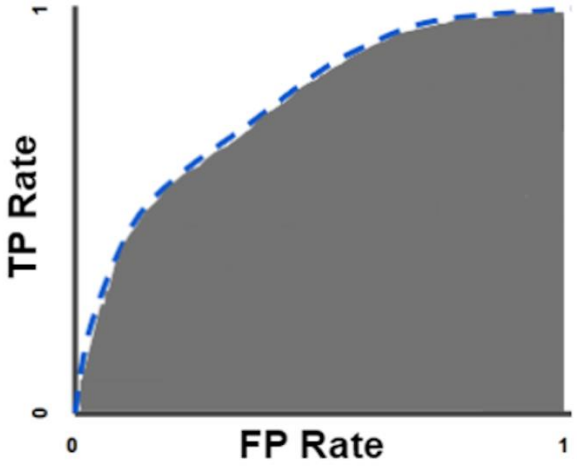
$$FPR = \frac{FP}{FP + TN}$$

Μια καμπύλη ROC απεικονίζει το TPR έναντι του FPR σε διαφορετικά σύνορα απόφασης. Η μείωση του ορίου ταξινόμησης ταξινομεί περισσότερα στοιχεία ως θετικά, αυξάνοντας έτσι τόσο τα False Positive όσο και τα True Positive. Το παρακάτω σχήμα δείχνει μια τυπική καμπύλη ROC.



Εικόνα 1. 5 *Παράδειγμα υπολογισμού ROC*

Η μετρική AUC μετρά ολόκληρη τη δισδιάστατη περιοχή κάτω από ολόκληρη την καμπύλη ROC από (0,0) έως (1,1). Στο παρακάτω διάγραμμα η περιοχή αυτή αντιστοιχεί στο γραμμοσκιασμένο τμήμα.



Εικόνα 1. 6 *Υπολογισμός AUC*

Η AUC παρέχει ένα συνολικό μέτρο απόδοσης σε όλα τα πιθανά όρια ταξινόμησης. Ένας τρόπος ερμηνείας της AUC είναι η πιθανότητα το μοντέλο να κατατάσσει ένα τυχαίο θετικό παράδειγμα υψηλότερα από ένα τυχαίο αρνητικό δείγμα. Η τιμή AUC κυμαίνεται από 0 έως 1. Ένα μοντέλο του οποίου οι προβλέψεις είναι 100% λανθασμένες έχει AUC 0,0. Αυτός του οποίου οι προβλέψεις είναι 100% σωστές έχει AUC 1,0.

Τα πλεονεκτήματα της μετρικής αυτής είναι:

- Είναι αμετάβλητη ως προς την κλίμακα. Μετρά πόσο καλά κατατάσσονται οι προβλέψεις, αντί για τις απόλυτες τιμές τους.
- Είναι αμετάβλητη ταξινόμηση-συνόρου απόφασης. Μετρά την ποιότητα των προβλέψεων του μοντέλου ανεξάρτητα από το όριο ταξινόμησης που επιλέγεται.

Confidence intervals (μέτρα εμπιστοσύνης)

Ένα διάστημα εμπιστοσύνης είναι μια μέθοδος που υπολογίζει ένα άνω και ένα κάτω όριο γύρω από μια εκτιμώμενη τιμή. Η πραγματική τιμή της παραμέτρου μπορεί να είναι εντός ή εκτός των ορίων αυτών. Τα διαστήματα εμπιστοσύνης είναι ένας τρόπος ποσοτικοποίησης της αβεβαιότητας μιας εκτίμησης. Σε ένα μοντέλο πρόβλεψης που προκύπτει από μία διαδικασία μηχανικής μάθησης μπορεί να χρησιμοποιηθεί για την εκτίμηση της ικανότητας του να παρουσιάσει μια δεδομένη απόδοση.

Συχνά, όσο μεγαλύτερο είναι το δείγμα από το οποίο αντλήθηκε η εκτίμηση, τόσο πιο ακριβής είναι η εκτίμηση και τόσο μικρότερο (καλύτερο) το διάστημα εμπιστοσύνης.

Μικρότερο διάστημα εμπιστοσύνης συνεπάγεται ακριβής εκτίμηση.

Μεγαλύτερο διάστημα εμπιστοσύνης συνεπάγεται λιγότερο ακριβής εκτίμηση.

Στις περιπτώσεις της ταξινόμησης, το διάστημα εμπιστοσύνης υπολογίζεται από τη σχέση

$$interval = z\sqrt{\frac{accuracy(1-accuracy)}{n}}$$

Όπου accuracy είναι η ακρίβεια που παρουσιάζει το μοντέλο πρόβλεψης, n το μέγεθος του δείγματος και z είναι ο αριθμός των τυπικών αποκλίσεων από την κατανομή Gauss[4].

Διακριτικότητα (specificity)

Η διακριτικότητα υπολογίζει αναλογία των αληθινών αρνητικών που προσδιορίζονται σωστά από το μοντέλο. Είναι ένα μέτρο που δείχνει πόσο ικανό είναι το μοντέλο να διακρίνει πότε ένα δείγμα δεν περιλαμβάνεται σε μία κλάση. Υψηλή ειδικότητα σημαίνει ότι το μοντέλο προσδιορίζει σωστά πότε ένα αντικείμενο δεν ανήκει σε μία κλάση. Η τιμή της διακριτικότητας δίνεται από τη παρακάτω σχέση

$$Specificity = \frac{TN}{TN + FP}$$

Η υψηλότερη τιμή διακριτικότητα σημαίνει υψηλή ικανότητα του μοντέλου να αποκλείει την ένταξη του δείγματος σε μία κλάση. Χαμηλή τιμή σημαίνει αδυναμία του μοντέλου να αποκλείσει την ένταξη του δείγματος σε κατηγορία που δεν ανήκει.

Η διακριτικότητα για κάθε κλάση είναι:

Α: $\frac{11}{19} = 0.58$

Β: $\frac{9}{14} = 0.64$

Γ: $\frac{8}{23} = 0.5$

[4] Συνήθως χρησιμοποιούνται οι αριθμοί 1.64 (με επίπεδο σημαντικότητας 90%), 1,96 (με επίπεδο σημαντικότητας 95%), 2,33 (με επίπεδο σημαντικότητας 98%), 2,58 (με επίπεδο σημαντικότητας 99%).

Prevalence

Η μετρική αυτή περιγράφει το ποσοστό των παρατηρήσεων μία κλάσης (είτε προβλέπονται σωστά είτε όχι) στο ολικό δείγμα. Η τιμή της δίνεται από την παρακάτω σχέση.

$$Prevelance = \frac{Αριθμός\ δειγμάτων\ κλάσης\ X}{Συνολικός\ αριθμός\ δειγμάτων}$$

Η ακρίβεια για κάθε κλάση είναι:

Α: $\frac{20}{100} = 0.2$

Β: $\frac{30}{100} = 0.3$

Γ: $\frac{50}{100} = 0.5$

Detection Rate

Πρόκειται για το μέτρο της ικανότητας του μοντέλου να προβλέπει σωστά ότι ένα αντικείμενο ανήκει σε δεδομένη κλάση. Η τιμή της μετρικής δίνεται από την παρακάτω σχέση.

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

Η μετρική Detection Rate για κάθε κλάση είναι:

Α: $\frac{12}{23} = 0.52$

Β: $\frac{25}{34} = 0.73$

Γ: $\frac{35}{43} = 0.81$

Ισορροπημένη ακρίβεια (Balanced accuracy)

Πρόκειται για ένα μέτρο που συνδυάζει την ευαισθησία και τη διακριτικότητα. Το μέτρο της δίνεται από την παρακάτω σχέση.

$$BA = \frac{Sensitivity + Specificity}{2}$$

Χρησιμοποιείται κυρίως σε περιπτώσεις όπου οι κλάσεις είναι δεν ισορροπημένες ως προς το πλήθος των δειγμάτων που περιλαμβάνουν. Όσο πιο κοντά είναι η ισορροπημένη ακρίβεια στο 1, τόσο καλύτερα το μοντέλο μπορεί να ταξινομήσει σωστά τις παρατηρήσεις (Karimi, 2021).

Η ισορροπημένη ακρίβεια για κάθε κλάση είναι:

A: 0.56

B: 0.78

Γ: 0.75

Miclassification

Πρόκειται για μία μετρική που αξιολογεί την αδυναμία του μοντέλου να προβλέψει σωστά. Το μέτρο της δίνεται από την παρακάτω σχέση (Karimi, 2021).

$$Misclassification = \frac{FN + FP}{Αριθμός\ Δειγμάτων}$$

Κεφάλαιο 3. Περιγραφική Ανάλυση Δεδομένων

Σε αυτό το κεφάλαιο θα γίνει προσπάθεια λήψης της γενικής εικόνας για τις μεταβλητές που περιλαμβάνονται στο συγκεκριμένο σύνολο δεδομένων. Θα πραγματοποιηθεί ανάλυση βάσει των επεξηγηματικών παραγόντων παραθέτοντας στοιχεία περιγραφικής στατιστικής αυτών προκειμένου να επιτευχθεί μια πιο συγκεκριμένη εικόνα του υπό μελέτη προβλήματος. Επίσης, θα μελετηθεί η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της μεταβλητής απόκρισης με τη χρήση διαγραμμάτων έτσι ώστε να γίνουν ακόμα πιο κατανοητά των διαφόρων μεταβλητών.

3.1 Περιγραφή προβλήματος

Στη συγκεκριμένη μελέτη περίπτωσης θα χρησιμοποιηθούν πραγματικά δεδομένα τα οποία αφορούν χαρακτηριστικά κυτταρικού πυρήνα από δεδομένα εικόνας τα οποία προκύπτουν από βιοψία με τη μέθοδο λεπτής βελόνης για τη διάγνωση του όγκου του μαστού. Η ιστοσελίδα Kaggle, έχοντας τον ρόλο της ‘’τράπεζας δεδομένων’’, παρέχει τα δεδομένα του πεδίου εφαρμογής τα οποία προκύπτουν και αντλούνται μέσω της ψηφιοποιημένης εικόνας στην οποία απεικονίζονται τα χαρακτηριστικά του κυτταρικού πυρήνα έπειτα από βιοψία δια λεπτής βελόνης (Fine Needle Aspirate - FNA) μάζας μαστού. Εν συντομία, η βιοψία δια λεπτής βελόνης όγκου αποτελεί μια μέθοδο κατά την οποία περισυλλέγονται κύτταρα από όλα τα σημεία της βλάβης (όγκου) μέσω αναρρόφησης με την βοήθεια βελόνης 21g παρόμοια με αυτή της αιμοληψίας. Τα δεδομένα αυτά αποτελούνται από χαρακτηριστικά τα οποία αφορούν τα κύτταρα του σημείου της βλάβης (όγκου). Το πρόβλημα το οποίο αντιμετωπίζεται είναι ένα κλασσικό πρόβλημα ταξινόμησης (classification) δύο κλάσεων, με στόχο την πρόβλεψη της επικρατούσας κλάσης των παρατηρήσεων, καθώς και ο εντοπισμός καταλυτικών παραγόντων που επιδρούν συνδυαστικά στο τελικό αποτέλεσμα.

3.2 Παρουσίαση δεδομένων

Το σετ δεδομένων το οποίο εξετάζεται αφορά αποτελέσματα βιοψίας δια λεπτής βελόνης 569 ασθενών τα οποία προέκυψαν από τη μοντελοποίηση χαρακτηριστικών κυτταρικού πυρήνα από δεδομένα εικόνας. Το συγκεκριμένο σύνολο δεδομένων χρησιμοποιείται προκειμένου να εξεταστεί ποιοι είναι οι σημαντικότεροι παράγοντες οι οποίοι καθορίζουν το αν ένας όγκος μαστού είναι καλοήθης ή κακοήθεις αλλά και για την ορθολογική ταξινόμηση αυτού σε άγνωστα δεδομένα βάσει της των εκάστοτε αποτελεσμάτων βιοψίας δια λεπτής βελόνης.

Προτού εφαρμοστεί οποιαδήποτε κατεργασία και ιδιαίτερα η περιγραφική ανάλυση στο σύνολο των δεδομένων χρήζει επιτακτικής ανάγκης ο έλεγχος του περιεχομένου και γενικότερα της εσωτερικής του δομής. Ως εκ τούτου είναι απαραίτητο ο ορισμός του τύπου της κάθε μεταβλητής αλλά και ο έλεγχος για ελλιπείς τιμές (missing values) στο σύνολο των δεδομένων. Όπως προκύπτει λοιπόν δεν υπάρχει κάποια ελλιπή τιμή.

Στο σύνολο των δεδομένων που πρόκειται να εξεταστεί υπάρχουν δύο κατηγορίες μεταβλητών. Η πρώτη είναι η κατηγορική μεταβλητή η οποία αντιστοιχεί σε ποιοτικά ή ονομαστικά δεδομένα. Οι κατηγορικές μεταβλητές μπορούν να λάβουν από μία έως έναν πεπερασμένο αριθμό πιθανών τιμών ή κατηγοριών. Οι κατηγορίες τυπικά ορίζονται από κάποιο είδος χαρακτηριστικού και χρησιμοποιούνται συνήθως για να περιγράψουν δημογραφικά ή άλλα μη αριθμητικά χαρακτηριστικά ενός πληθυσμού ή δείγματος. Παραδείγματα κατηγορικών μεταβλητών μπορεί να περιλαμβάνουν πράγματα όπως το φύλλο (αρσενικό ή θηλυκό), τη φυλή (Ευρωπαίος, Ασιάτης κ.ο.κ), τη θρησκεία (χριστιανός, μουσουλμάνος κ.ο.κ), τις πολιτικές πεποιθήσεις (καπιταλιστής, σοσιαλιστής κ.ο.κ) ή οποιοδήποτε άλλο χαρακτηριστικό που μπορεί να χωριστεί σε διακριτές κατηγορίες. Στην συγκεκριμένη διπλωματική το χαρακτηριστικό το οποίο διαχωρίζει σε δύο κατηγορίες το δείγμα, είναι ο τύπος του όγκου. Καλοήθης και κακοήθης.

Η δεύτερη κατηγορία είναι οι συνεχείς μεταβλητές. Οι συνεχείς μεταβλητές είναι ένας τύπος μεταβλητής που μπορεί να λάβει οποιαδήποτε τιμή εντός ενός καθορισμένου εύρους. Είναι μια ποσοτική μεταβλητή που μπορεί να μετρηθεί και να αναπαρασταθεί με έναν πραγματικό αριθμό. Οι συνεχείς μεταβλητές χρησιμοποιούνται συνήθως σε τομείς όπως η φυσική, η μηχανική και τα οικονομικά, όπου οι ακριβείς μετρήσεις και η ακριβής μοντελοποίηση συστημάτων είναι σημαντικές. Παραδείγματα συνεχών μεταβλητών μπορεί να περιλαμβάνουν πράγματα όπως βάρος, ύψος, θερμοκρασία, χρόνος και απόσταση.

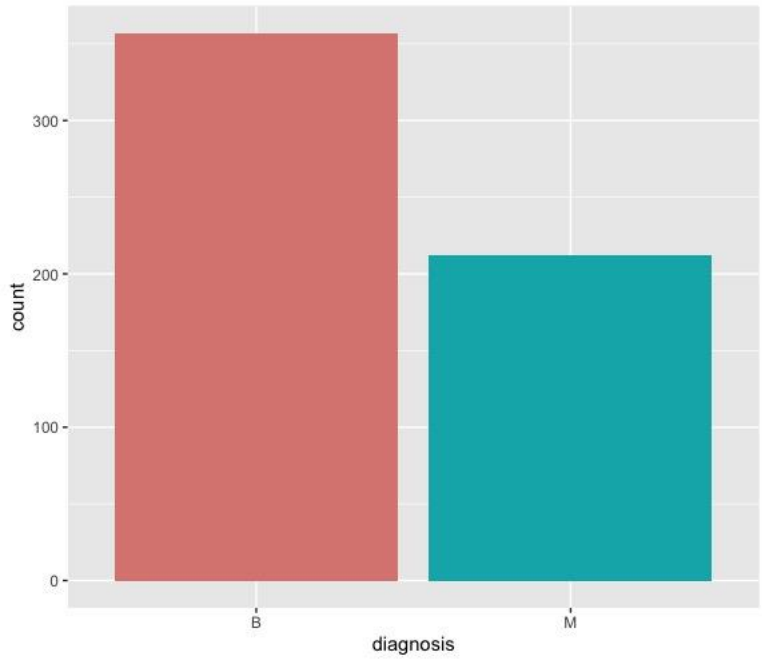
3.3 Ανάλυση ποιοτικών μεταβλητών

Στο σύνολο των δεδομένων που θα χρησιμοποιηθούν υπάρχει μια ποιοτική μεταβλητή η οποία είναι και η μεταβλητή απόκρισης η οποία παρουσιάζεται στον **πίνακα 3.1**.

Μεταβλητή	Επεξήγηση
diagnosis	αποτέλεσμα διάγνωσης

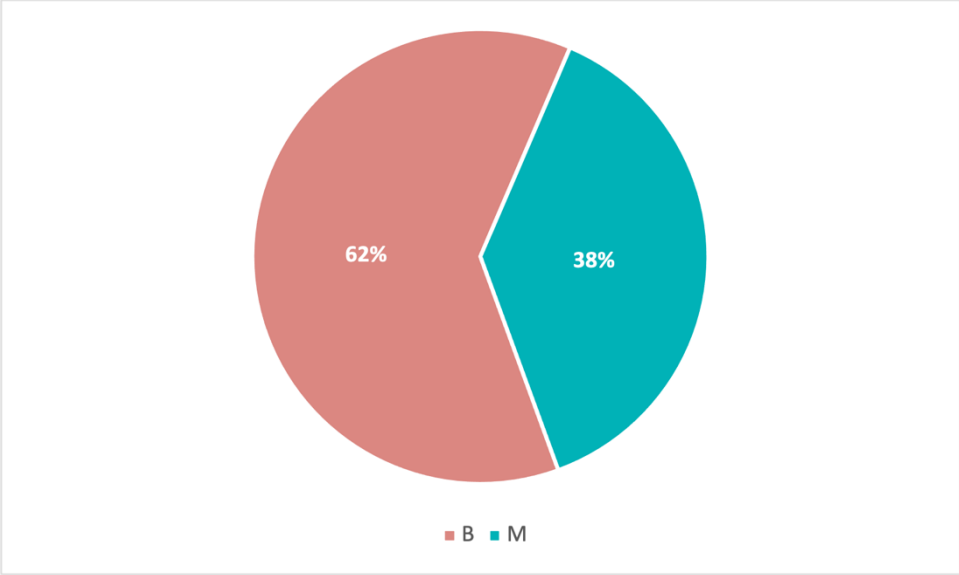
Πίνακας 3.1 *Κατηγορική μεταβλητή*

Ειδικότερα, η μεταβλητή ‘’diagnosis’’ είναι δίτιμη και ως εκ τούτου το σύνολο τιμών της αποτελείται από δύο τιμές (‘’B’’, ‘’M’’) οι οποίες αντιπροσωπεύουν την έκβαση της βιοψίας με την τιμή ‘’B’’ να προέρχεται από την αγγλική λέξη ‘’benign’’ που σημαίνει καλοήθης, ενώ η τιμή ‘’M’’ προέρχεται από την αγγλική λέξη ‘’malignant’’ που σημαίνει κακοήθης.



Εικόνα 3. 1 *Ραβδόγραμμα ποιοτικής μεταβλητής απόκρισης*

Ένα ραβδόγραμμα απεικονίζει το πλήθος των παρατηρήσεων που ανήκουν στη εκάστοτε κλάση. Στο συγκεκριμένο πρόβλημα και κατ’ επέκτασιν σύνολο δεδομένων, 357 παρατηρήσεις ανήκουν στην κλάση της καλοήθειας αποτελώντας το 62% του συνολικού δείγματος, ενώ 212 παρατηρήσεις ανήκουν στην κλάση της κακοήθειας αποτελώντας το υπόλοιπο 37%.



Εικόνα 3.2 *Ποσοστό καλοηθών και κακοηθών όγκων*

3.4 Ανάλυση συνεχών μεταβλητών

Σε αυτή την ενότητα θα περιγραφούν τα κυριότερα στατιστικά περιγραφικά μέτρα των μεταβλητών. Τέτοια μέτρα είναι η μέγιστη τιμή (max), όπου αντιπροσωπεύει την μέγιστη τιμή που λαμβάνει η συνεχής μεταβλητή, η ελάχιστη τιμή (min) όπου αντιπροσωπεύει την ελάχιστη τιμή που λαμβάνει η συνεχής μεταβλητή, το εύρος (range) το οποίο προκύπτει από την αφαίρεση της ελάχιστης από τη μέγιστη τιμή, η μέση τιμή (mean), τη διάμεσο όπου γι' αυτή την τιμή το δείγμα χωρίζεται ακριβώς στη μέση, η τιμή του πρώτου (1stQu.) και τρίτου (3stQu.) τεταρτημορίου, καθώς και η τυπική απόκλιση (SD).

Στη συνέχεια θα ακολουθήσει γραφική συμπεριφορά των αριθμητικών μεταβλητών αλλά και σχέση τους με το τελικό αποτέλεσμα για την πρόβλεψη του καρκίνου του μαστού. Θα γίνει χρήση των διαγραμμάτων κατανομής τα οποία παρουσιάζουν μια καθαρή εικόνα βάσει της οποίας δίνεται μια διαυγής εικόνα για την κατανομή των αριθμητικών μεταβλητών. Ειδικότερα, τα ιστογράμματα ουσιαστικά απεικονίζουν την κατανομή των τιμών μιας μεταβλητής, επισημαίνοντας το πλήθος των παρατηρήσεων που βρίσκεται εντός ενός εύρους τιμών. Επί προσθέτως, θα παρουσιαστούν τα γραφήματα πυκνότητας πιθανότητας έτσι ώστε να γίνει κατανοητό πως κατανέμονται οι τιμές των διαφόρων στατιστικών στοιχείων για κάθε κλάση του όγκου. Σε αρκετά γραφήματα πυκνότητας πιθανότητας οι τιμές του άξονας y ξεπερνούν την τιμή 1, αυτό δεν σημαίνει ότι η ίδια η μέτρηση πυκνότητας έχει υπερβεί την τιμή 1. Ο άξονας y μιας γραφικής παράστασης πυκνότητας αντιπροσωπεύει την πυκνότητα

πιθανότητας των δεδομένων, η οποία είναι το ύψος της καμπύλης σε ένα συγκεκριμένο σημείο στον άξονα x. Η συνολική επιφάνεια κάτω από την καμπύλη του διαγράμματος πυκνότητας πρέπει να είναι ίση με 1, επειδή αντιπροσωπεύει τη συνολική πιθανότητα όλων των πιθανών αποτελεσμάτων. Ωστόσο, το ύψος της καμπύλης σε ένα συγκεκριμένο σημείο του άξονα x μπορεί να είναι μεγαλύτερο από 1 εάν το εύρος του άξονα x είναι πολύ στενό ή εάν τα δεδομένα είναι πολύ στενά ομαδοποιημένα γύρω από μια μεμονωμένη τιμή. Αυτό συμβαίνει επειδή η μέτρηση πυκνότητας ορίζεται ως η πιθανότητα ανά μονάδα μέτρησης, επομένως ένα πολύ στενό εύρος μονάδων μέτρησης μπορεί να οδηγήσει σε υψηλή πυκνότητα πιθανότητας. Επίσης παρουσιάζονται τα violin plots στα οποία αναπαρίστανται τα κυριότερα περιγραφικά χαρακτηρίστηκα, δηλαδή η κατανομή και η πυκνότητα πιθανότητας των παρατηρήσεων ανάλογα με το τεταρτημόριο τους. Επίσης, με τα violin plots είναι εφικτό να εντοπιστούν εύκολα ακραίες τιμές (outliners). Τα violin plots ουσιαστικά αποτελούν μια παραλλαγή των θηκογραμμάτων (boxplots), η μονή διαφορά τους είναι ότι παρουσιάζουν και μια εκτίμηση της συνάρτησης πυκνότητας πιθανότητας για τα δεδομένα.

Για την καλύτερη κατανόηση των παραγόντων, θα τους κατανείμουμε σε τρεις κατηγορίες. Η πρώτη κατηγορία περιέχει τις μεταβλητές οι οποίες αντιπροσωπεύουν τις μέσες τιμές του κάθε παράγοντα. Η δεύτερη κατηγορία περιέχει τις μεταβλητές οι οποίες αντιπροσωπεύουν τις τυπικές αποκλίσεις του κάθε παράγοντα, για π και τέλος στην τρίτη κατηγορία βρίσκονται οι μεταβλητές οι οποίες αφορούν ακραίες αποκλίσεις, δηλαδή διακυμάνσεις, του κάθε παράγοντα. Ο μέσος όρος, η τυπική απόκλιση και το "χειρότερο", δηλαδή η διακύμανση (μέσος όρος των τριών μεγαλύτερων τιμών) αυτών των χαρακτηριστικών υπολογίστηκαν για κάθε εικόνα, με αποτέλεσμα να προκύψουν 30 χαρακτηριστικά. Για παράδειγμα, το 3^{o} σε σειρά χαρακτηριστικό αντιπροσωπεύει τη μέση ακτίνα του πυρήνα σε σημεία της περιμέτρου του, το 13^{o} σε σειρά χαρακτηριστικό αντιπροσωπεύει την τυπική απόκλιση της απόστασης του πυρήνα σε σημεία της περιμέτρου του και τέλος το 23^{o} σε σειρά χαρακτηριστικό αντιπροσωπεύει την διακύμανση της απόστασης του πυρήνα σε σημεία της περιμέτρου του.

Μεταβλητές	Επεξήγηση
radius_mean	μέση απόσταση του πυρήνα στα σημεία της περιμέτρου του
texture_mean	μέση τιμή υφής του κυτταρικού πυρήνα στην κλίμακα του γκρι
perimeter_mean	μέση τιμή περιμέτρου του κυτταρικού πυρήνα
area_mean	μέση τιμή της ακτίνας του κυτταρικού πυρήνα.
smoothness_mean	μέση τιμή απαλότητας του κυτταρικού πυρήνα
compactness_mean	μέση τιμή συμπαγότητας του κυτταρικού πυρήνα.
concavity_mean	μέση τιμή κοιλότητας του κυτταρικού πυρήνα
concave.points_mean	μέση τιμή του αριθμού των κοίλων σημείων του κυτταρικού πυρήνα
symmetry_mean	μέση τιμή συμμετρίας του κυτταρικού πυρήνα
fractal_dimension_mean	μέση τιμή μορφοκλασματικής διάστασης του κυτταρικού πυρήνα

Πίνακας 3.2 *Μεταβλητές μέσης τιμής παραγόντων*

Η δεύτερη κατηγορία (**πίνακας 3.3**) αντιπροσωπεύει τις τυπικές αποκλείσεις του κάθε παράγοντα όπως για παράδειγμα η μεταβλητή ‘’radius_se’’ αφορά την τυπική απόκλιση των αποστάσεων του πυρήνα στα σημεία της περιμέτρου του.

Μεταβλητές	Επεξήγηση
radius_se	τυπική απόκλιση απόστασης του πυρήνα στα σημεία της περιμέτρου του κυτταρικού πυρήνα
texture_se	τυπική απόκλιση υφής του κυτταρικού πυρήνα στην κλίμακα του γκρι
perimeter_se	τυπική απόκλιση περιμέτρου του κυτταρικού πυρήνα
area_se	τυπική απόκλιση της ακτίνας του κυτταρικού πυρήνα.
smoothness_se	τυπική απόκλιση του κυτταρικού πυρήνα
compactness_se	τυπική απόκλιση συμπαγότητας του κυτταρικού πυρήνα.
concavity_se	τυπική απόκλιση κοιλότητας του κυτταρικού πυρήνα
concave.points_se	τυπική απόκλιση του αριθμού των κοίλων σημείων
symmetry_se	τυπική απόκλιση συμμετρίας του κυτταρικού πυρήνα
fractal_dimension_se	τυπική απόκλιση μορφοκλασματικής διάστασης του κυτταρικού πυρήνα

Πίνακας 3. 3 *Μεταβλητές τυπικής απόκλισης παραγόντων*

Τέλος στην τρίτη κατηγορία (**πίνακας 3.4**) βρίσκονται οι μεταβλητές οι οποίες αφορούν ακραίες αποκλίσεις, δηλαδή διακυμάνσεις, του κάθε παράγοντα. Για παράδειγμα η μεταβλητή ‘’radius_worst’’ αφορά τη διακύμανση των αποστάσεων του πυρήνα στα σημεία της περιμέτρου.

Μεταβλητές	Επεξήγηση
radius_worst	διακύμανση απόστασης του πυρήνα στα σημεία της περιμέτρου
texture_worst	διακύμανση υφής του κυτταρικού πυρήνα στην κλίμακα του γκρι
perimeter_worst	διακύμανση περιμέτρου
area_worst	διακύμανση της ακτίνας του κυτταρικού πυρήνα.
smoothness_worst	διακύμανση του κυτταρικού πυρήνα
compactness_worst	διακύμανση συμπαγότητας του κυτταρικού πυρήνα.
concavity_worst	διακύμανση κοιλότητας.
concave.points_worst	διακύμανση του αριθμού των κοίλων σημείων
symmetry_worst	διακύμανση συμμετρίας
fractal_dimension_worst	διακύμανση μορφοκλασματικής διάστασης

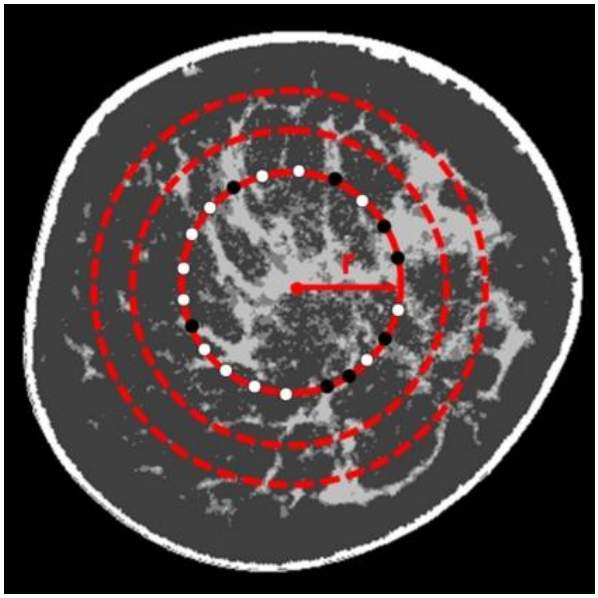
Πίνακας 3.4 *Μεταβλητές διακύμανσης παραγόντων*

3.4.1 Μεταβλητές μέσης τιμής παραγόντων

Οι μεταβλητές που αναφέρονται στις μέσες τιμές των παραγόντων του κυτταρικού πυρήνα είναι οι εξής:

1. Μέση τιμή ακτίνας (radius_mean)

   Το χαρακτηριστικό αυτό αφορά την απόσταση του πυρήνα σε σημεία της περιμέτρου του. Ως εκ τούτου η ανεξάρτητη μεταβλητή ‘’radius_mean’’ αντιπροσωπεύει τη μέση απόσταση του πυρήνα από σημεία της περιμέτρου του που προκύπτει από τα δεδομένα εικόνας. Το χαρακτηριστικό αυτό έχει ως μονάδα μέτρησης το πλήθος των εικονοστοιχείων (pixels).



Εικόνα 3. 3 *Αναπαράσταση ακτίνας κυτταρικού πυρήνα*

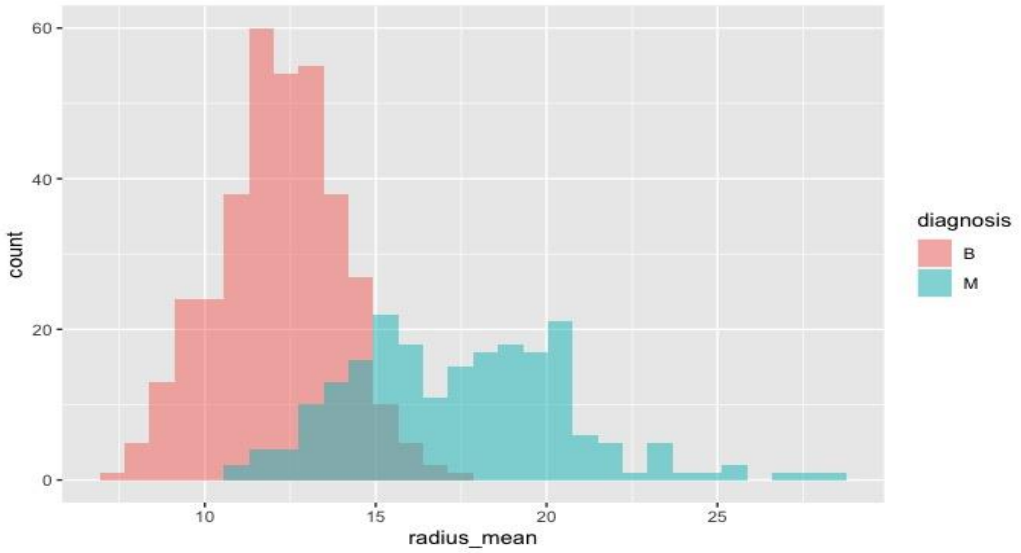
Πηγή: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3081868/

radius_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	6.981	11.080	12.200	12.147	13.370	17.850	1.781	10.869
Malignant	10.950	15.070	17.320	17.460	19.590	28.110	3.204	17.160

Πίνακας 3. 5 *Περιγραφικά μέτρα μέσης απόστασης του πυρήνα από σημεία της περιμέτρου*

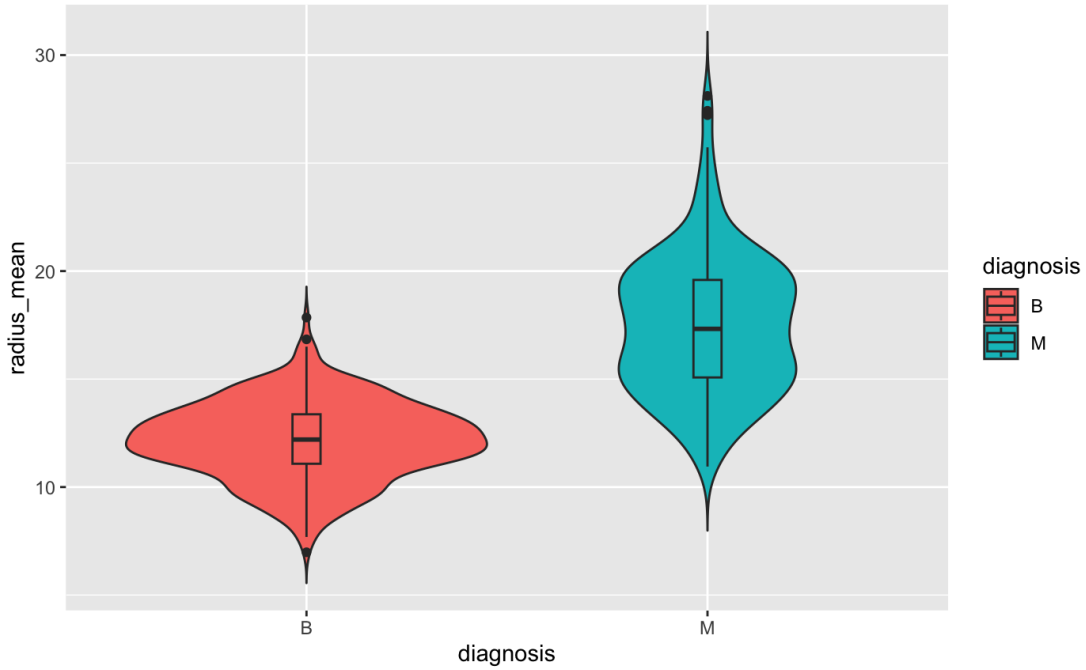
Παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης απόστασης της ακτίνας ίση με 6.981 και οι κακοήθεις ίση με 10.950. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 11.080 ενώ 15.07 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση απόσταση του πυρήνα σε σημεία της περιμέτρου τουλάχιστον ίση με 12.2 ενώ οι κακοήθεις όγκοι 17.32. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού μέσης ακτίνας ίση με 12.146 και 17.460 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης απόστασης της ακτίνας ίση με τουλάχιστον 13.370 και 19.590 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 17.85 για τους καλοήθεις όγκους και 28.110 για τους κακοήθεις αντιστοίχως. Τέλος, όσον

αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 10.869 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 17.160 κατανέμονται οι κακοήθεις όγκοι.Παρακάτω λαμβάνει χώρα ο διαχωρισμός των καλοηθών και κακοηθών όγκων καθώς ο διαχωρισμός αυτός αναπαρίσταται γραφικά συναρτήσει της κατανομής αλλά και της πυκνότητας πιθανότητάς του. Σύμφωνα με το ιστόγραμμα της **εικόνας 3.4** φαίνεται πως οι όγκοι οι οποίοι είναι ταξινομημένοι στην κλάση της καλοήθειας λαμβάνουν μικρότερες τιμές ως προς τη μέση απόσταση της ακτίνας τους σε σχέση με αυτούς που ανήκουν στην κλάση της κακοήθειας. Η μέση ακτίνα των περισσοτέρων καλοηθών όγκων είναι ευδιάκριτο πως απέχει περίπου 12 pixels ένατι των περισσοτέρων κακοηθών των οποίων απέχει περίπου 17 pixels.



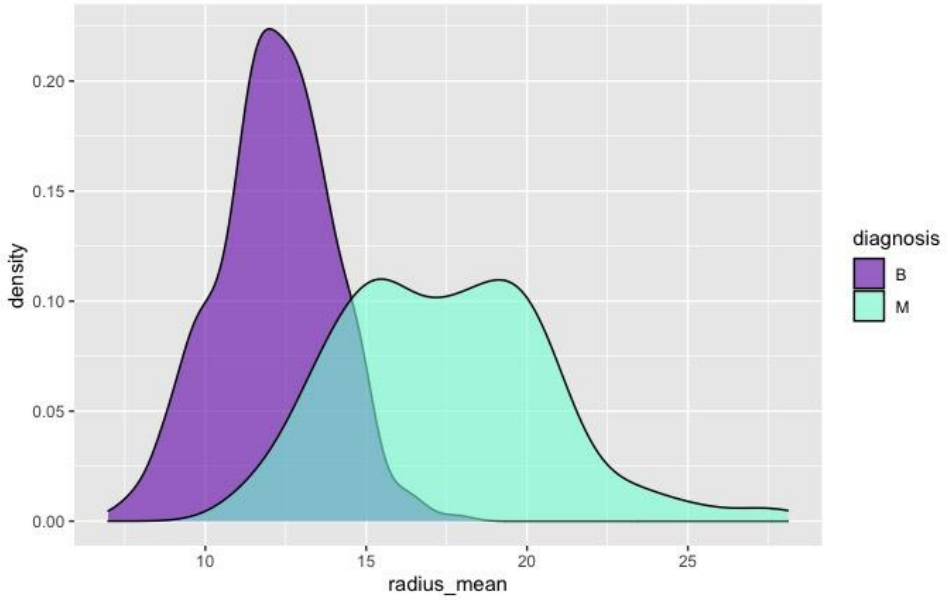
Εικόνα 3. 4 *Ιστόγραμμα μεταβλητής radius_mean*

Σύμφωνα με τα violin plots της **εικόνας 3.5** φαίνεται πως η διάμεσος της καλοήθειας βρίσκεται πιο κάτω σε σχέση με αυτή της κακοήθεια ενώ και στις δύο κλάσεις παρατηρούνται ακραίες τιμές οι οποίες κυμαίνονται στη τιμή του χαρακτηριστικού μέσης ακτίνας ίση με περίπου 18 για την καλοήθεια και 27 για την κακοήθεια.



Εικόνα 3. 5 *Violin plots της μεταβλητής radius_mean*

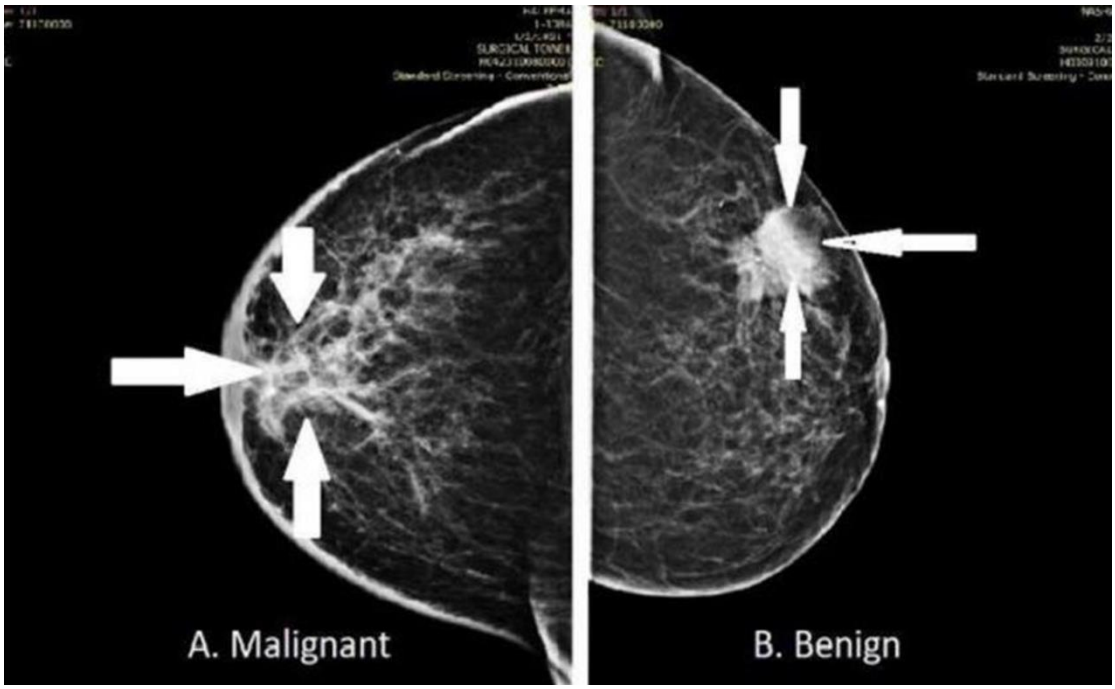
Βάσει του γραφήματος πυκνότητας πιθανότητας (**εικόνα 3.6**), για το διάστημα (10,14) παρουσιάζεται η μέγιστη πιθανότητα ένας όγκος να ανήκει στην κλάση της καλοήθειας ενώ για το διάστημα (15, 20) παρουσιάζεται η μέγιστη πιθανότητα να ανήκει στη κλάση της κακοήθειας.



Εικόνα 3. 6 *Γράφημα πυκνότητας πιθανότητας της μεταβλητής radius_mean*

2.Μέση τιμή υφής (texture_mean)

Η υφή, σύμφωνα με των πάροχο δεδομένων ‘’Kaggle’’, μετριέται στην κλίμακα του γκρι. Μια τιμή της κλίμακας του γκρι αντιπροσωπεύει την ένταση των αποχρώσεων σε κάθε pixel της εικόνας. Επομένως, η ανεξάρτητη μεταβλητή texture_mean, αντιπροσωπεύει τη μέση τιμή της υφής του κυτταρικού πυρήνα έχοντας ως μονάδα μέτρησης τις μονάδες της κλίμακας του γκρι.



Εικόνα 3. 7 *Απεικόνιση κακοήθους και καλοήθους όγκου βάσει των αποχρώσεων της κλίμακας του γκρι*

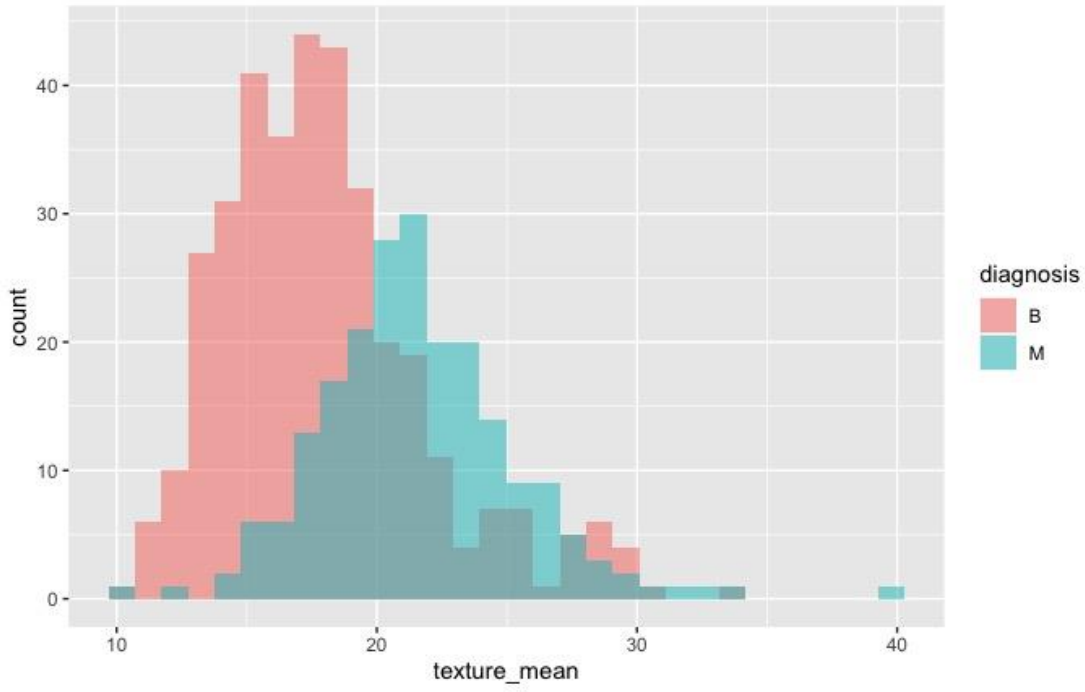
Στον **πίνακα 3.6**, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης υφής ίση με 9.71 και οι κακοήθεις ίση με 10.38. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 15.15 ενώ 19.33 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση ύφη τουλάχιστον ίση με 17.39 ενώ οι κακοήθεις όγκοι 21.46. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή υφής ίση με 17.91 και 21.60 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης υφής ίση με τουλάχιστον 19.76 και 23.77 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 33.81 για τους καλοήθεις όγκους και 39.28 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά

μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 10.87 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 28.90 κατανέμονται οι κακοήθεις όγκοι.

texture_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	9.71	15.15	17.39	17.91	19.76	33.81	1.78	10.87
Malignant	10.38	19.33	21.46	21.60	23.77	39.28	3.78	28.90

Πίνακας 3. 6 *Περιγραφικά μέτρα της μεταβλητής μέσης υφής*

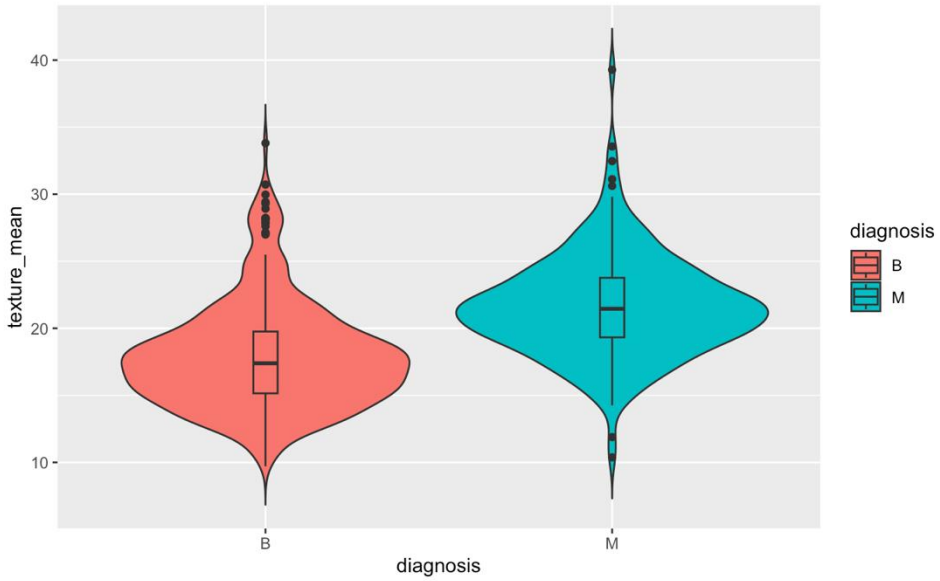
Στο ιστόγραμμα της μεταβλητής texture_mean το οποίο παρουσιάζεται στην **εικόνα 3.6** φαίνεται πως οι καλοήθεις όγκοι λαμβάνουν κατά κύριο λόγο μεγαλύτερες τιμές, έναντι των κακοηθών, με τις περισσότερες παρατηρήσεις να λαμβάνουν την τιμή περίπου 17.5, ενώ για την κλάση της κακοήθειας οι περισσότερες παρατηρήσεις λαμβάνουν την τιμή μέσης υφής περίπου ίση με 22.



Εικόνα 3. 8 *Ιστόγραμμα μεταβλητής texture_mean*

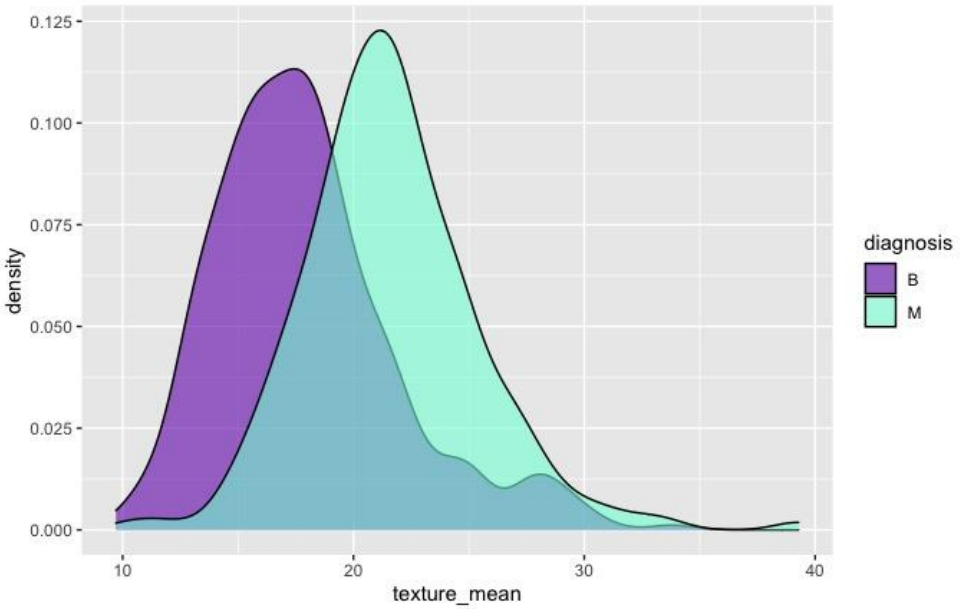
Οι παρατηρήσεις των δύο κλάσεων κατανέμονται εξίσου ομοιόμορφα όπως αποδεικνύεται και από τα violin plots της **εικόνας 3.7** όπου φαίνεται ότι οι παρατηρήσεις όσο για την καλοήθεια τόσο και για την κακοήθεια κατανέμονται γύρω από τη διάμεσο. Επίσης από τα violin plots

φαίνεται παρατηρούνται αρκετές ακραίες τιμές (outliners) και στις δύο κλάσεις οι οποίες κυμαίνονται για την καλοήθεια γύρο από την τιμή της μέσης ακτίνας ίση με 30, ενώ για την κακοήθεια από 30 έως και 40.



Εικόνα 3. 9 *Violin plots της μεταβλητής texture_mean*

Από το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.8**, εκμαιεύεται η πληροφορία για την αντιστοιχία των τιμών του χαρακτηριστικού μέσης υφή με την πιθανότητα ταξινόμησης στην εκάστοτε κλάση. Ειδικότερα, ένας όγκος του οποίο το πεδίο ορισμού των τιμών της μέσης υφής (15,17) εμφανίζει τη μεγαλύτερη πιθανότητα να ανήκει στην κλάση της καλοήθειας, ενώ για την κακοήθεια το πεδίο ορισμού των τιμών (20,25).



Εικόνα 3. 10 *Διάγραμμα πυκνότητας πιθανότητας της μεταβλητής texture_mean*

3.Μέση τιμή περιμέτρου (perimeter_mean)

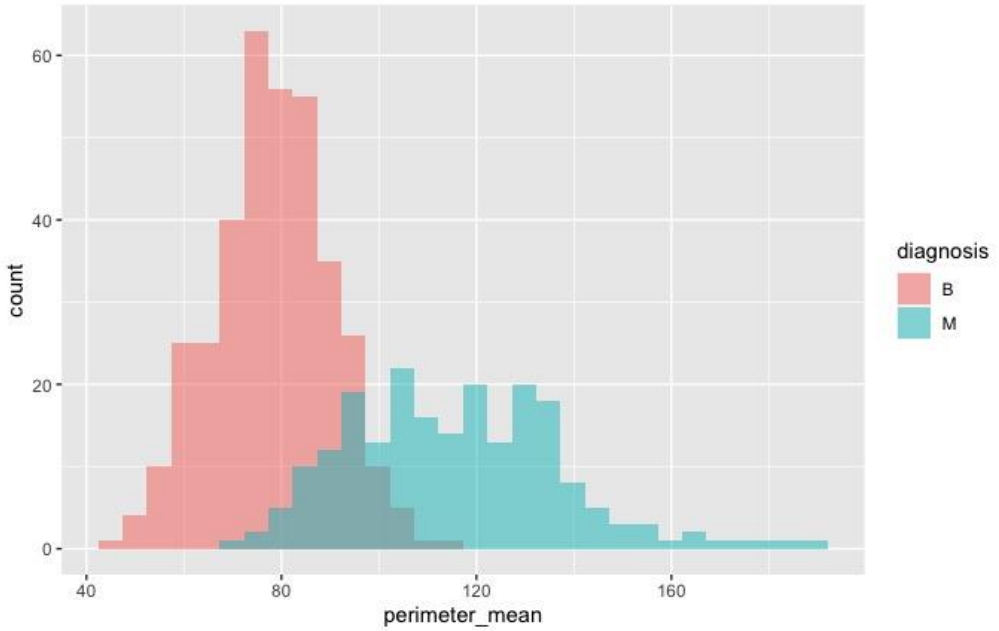
Το χαρακτηριστικό της περιμέτρου αφορά τη συνολικής απόστασης του ορίου του κυτταρικού πυρήνα και έχει ως μονάδα μέτρησης το πλήθος των εικονοστοιχείων (pixels). Τα βασικά περιγραφικά μέτρα της μεταβλητής δίνονται στον **πίνακα 3.7**. Ειδικότερα, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης περιμέτρου ίση με 43.79 και οι κακοήθεις ίση με 71.9. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 70.87 ενώ 98.75 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση περίμετρο τουλάχιστον ίση με 78.18 ενώ οι κακοήθεις όγκοι 114.2. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή χαρακτηριστικού ίση με 78.08 και 115.37 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης περιμέτρου ίση με τουλάχιστον 86.10 και 129.9 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 114.6 για τους καλοήθεις όγκους και 188.5 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο

κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 70.81 ενώ σε μεγαλύτερο το οποίο ισούται με 116.6 κατανέμονται οι κακοήθεις όγκοι.

perimeter_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	43.79	70.87	78.18	78.08	86.10	114.60	11.81	70.81
Malignant	71.90	98.75	114.20	115.37	129.90	188.50	21.85	116.60

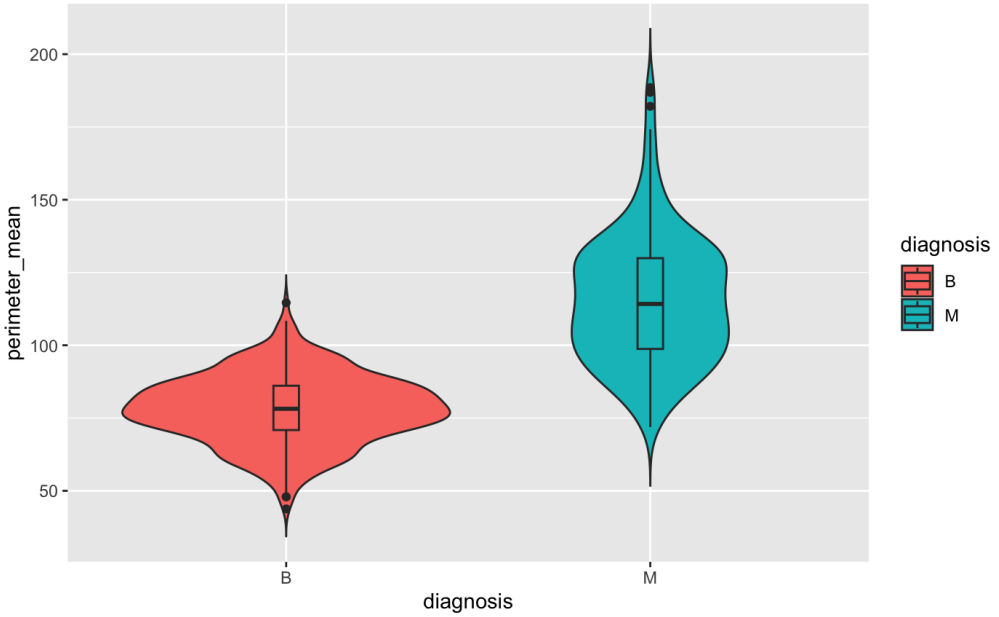
Πίνακας 3. 7 *Περιγραφικά μέτρα της μέσης περιμέτρου του κυτταρικού πυρήνα*

Βάσει του ιστόγραμμα της εικόνας 3.9, η καλοήθεις όγκοι λαμβάνουν αρκετά μεγαλύτερες έναντι των κακοηθών. Επίσης, το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνει την τιμή μέσης περιμέτρου περίπου ίση με 75 ενώ το μεγαλύτερο ποσοστό των κακοηθών όγκων περίπου ίση με 110.



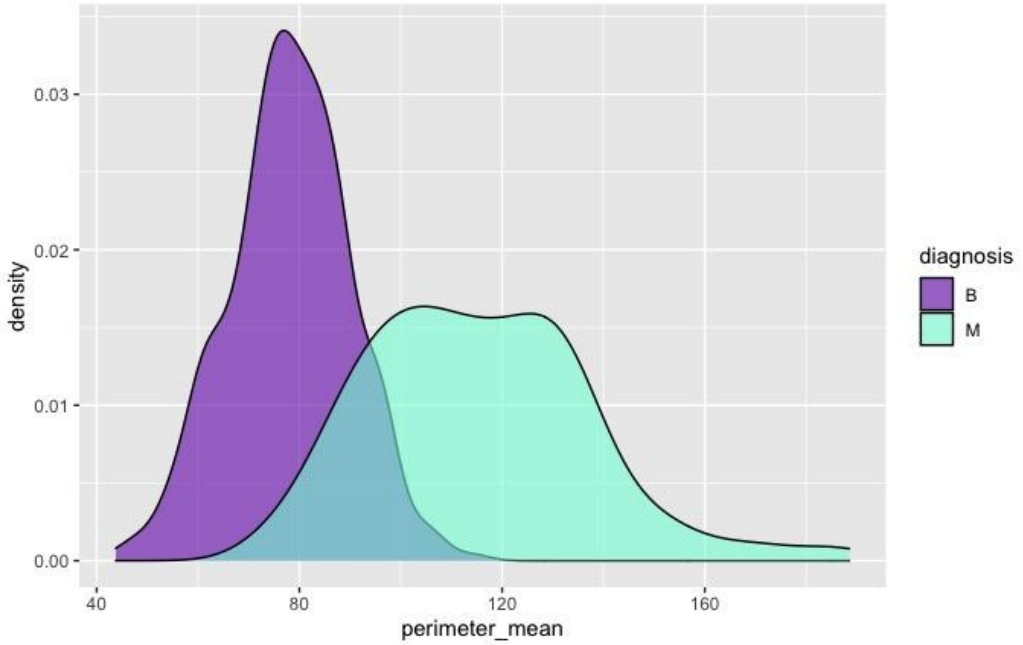
Εικόνα 3. 11 *Ιστόγραμμα μεταβλητής perimeter_mean*

Από τα violin plots της **εικόνας 3.10**, παρατηρείται πως η διάμεσος των παρατηρήσεων της καλοήθειας βρίσκεται χαμηλότερα έναντι της κακοήθειας. Επίσης, οι ακραίες τιμές της καλοήθειας είναι ελάχιστες και βρίσκονται κοντά και στην ελάχιστη αλλά και στη μέγιστη τιμή της μέσης περιμέτρου ενώ της κακοήθειας παρατηρούνται κοντά στη μέγιστη τιμή και κυμαίνονται περίπου στην τιμή μέσης περιμέτρου ίσης με περίπου 170.



Εικόνα 3. 12 *Violin plots για τη μεταβλητή perimeter_mean*

Το γράφημα πυκνότητας της **εικόνας 3.11** για το χαρακτηριστικό της μέσης περιμέτρου υποδεικνύει πως για το πεδίο ορισμού των μεταβλήτων του χαρακτηριστικού αυτού (70, 85) παρουσιάζεται η μέγιστη πιθανότητα ένας όγκος να ανήκει στην κλάση της καλοήθειας, ενώ αντίστοιχα για το πεδίο ορισμού (100, 130) στην κακοήθεια.



Εικόνα 3. 13 *Γράφημα πυκνότητας πιθανότητας της μεταβλητής perimeter_mean*

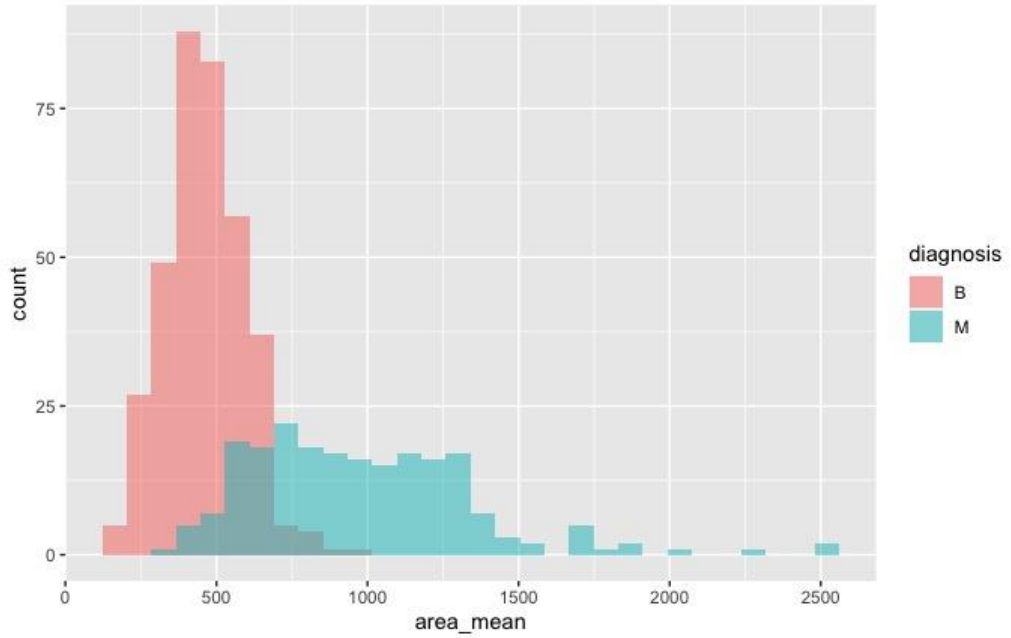
4.Μέση τιμή μεγέθους του κυτταρικού πυρήνα (area_mean)

Το συγκεκριμένο χαρακτηριστικό μετρά τη μέσο πλήθος των εικονοστοιχείων (pixels) στο εσωτερικό του ορίου με την προσθήκη του μισού πλήθους των εικονοστοιχείων στην περίμετρο, για τη διόρθωση του σφάλματος που προκαλείται από την ψηφιοποίηση. Ως εκ τούτου το χαρακτηριστικό αυτό αφορά το μέγεθος του κυτταρικού πυρήνα και έχει ως μονάδα μέτρησης $pixels^2$. Στον **πίνακα 3.8**, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσου μεγέθους ίση με 143.5 και οι κακοήθεις ίση με 361.6. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 378.2 ενώ 705.3 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέσο μέγεθος τουλάχιστον ίσο με 458.4 ενώ οι κακοήθεις όγκοι 932. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή για το μέσο μέγεθος ίση με 462.8 και 978.4 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης μεγέθους ίση με τουλάχιστον 551.1 και 1203.8 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 992.1 για τους καλοήθεις όγκους και 2501 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 134.2 ενώ σε πολύ μεγαλύτερο το οποίο ισούται με 367.9 κατανέμονται οι κακοήθεις όγκοι.

area_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	143.5	378.2	458.4	462.8	551.1	992.1	134.2	848.6
Malignant	361.6	705.3	932	978.4	1203.8	2501	367.9	2139.4

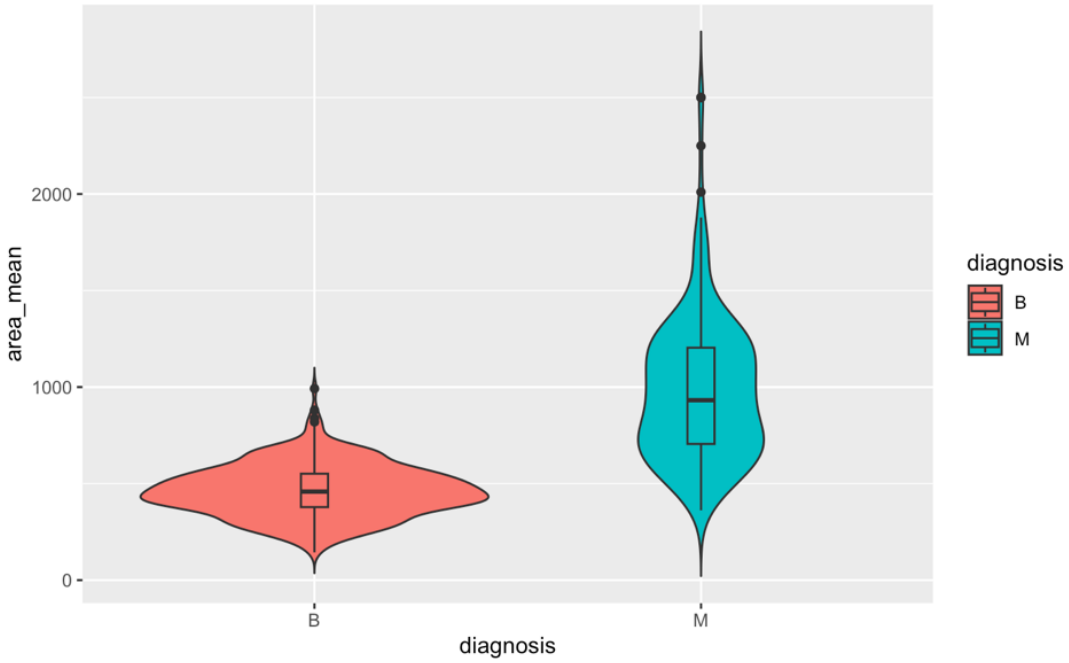
Πίνακας 3. 8 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού μέσου μεγέθους του κυτταρικού πυρήνα*

Παρακάτω στην **εικόνα 3.12**, παρουσιάζεται το ιστόγραμμα της μεταβλητής ‘’area_mean’’, όπου απεικονίζονται οι κατανομές της καλοήθειας και της κακοήθεις, με την καλοήθεια να ναι εμφανώς ξεκάθαρο ότι λαμβάνει τις μικρότερες σε αντίθεση με την κακοήθεια. Οι περισσότεροι καλοήθεις όγκοι λαμβάνουν την τιμή μέσου μεγέθους ίση με 375, οι κακοήθεις περίπου ίση με 750.



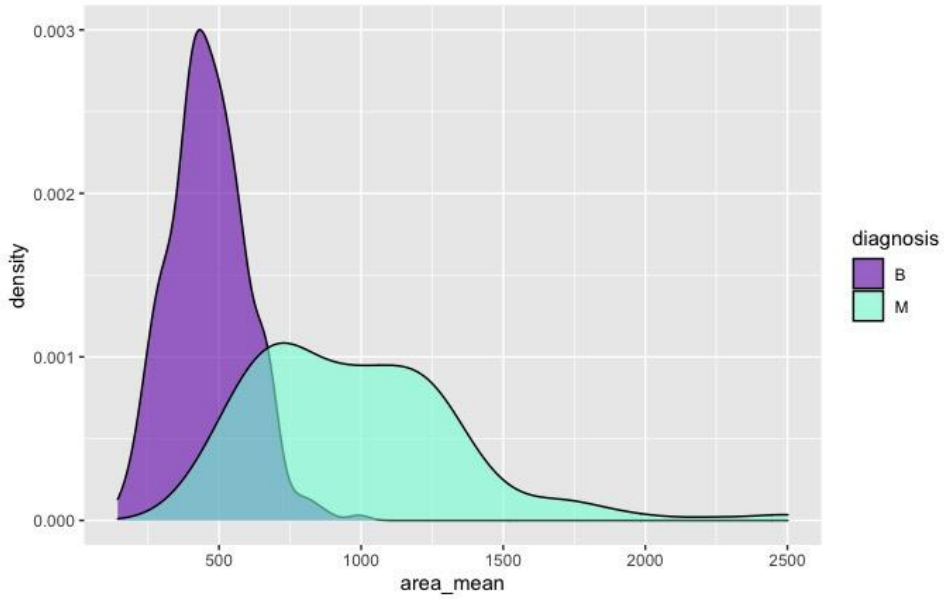
Εικόνα 3. 14 *Ιστόγραμμα μεταβλητής area_mean*

Έπειτα, στην **εικόνα 3.13**, παρουσιάζονται τα violin plots, όπου φαίνεται ξεκάθαρα ότι όλες οι παρατηρήσεις της καλοήθειας που αφορούν το μέσο μέγεθος του κυτταρικού πυρήνα κατανέμονται σε πολύ μικρότερο εύρος σχετικά με την κακοήθεια ενώ βάσει του διαγράμματος πυκνότητας πιθανότητας φαίνονται οι πιθανότητες που αντιστοιχούν στις αντίστοιχες παρατηρήσεις του μέσου μεγέθους του κυτταρικού πυρήνα για κάθε κλάση των όγκων.



Εικόνα 3. 15 *Violin plots για τη μεταβλητή area_mean*

Με αρκετά μικρό εύρος τιμών του χαρακτηριστικό μέσου μεγέθους του κυτταρικού πυρήνα κατανέμονται οι τιμές αυτού συναρτήσει των πιθανοτήτων για κάθε κλαση. Από το γράφημα πυκνότητας πιθανότητας (**εικόνα 3.14)** φαίνεται πως για τις τιμές του μέσου μεγέθους του κυτταρικού πυρήνα περίπου 450 με 470 παρουσίαζεται η μέγιστη πιθανότητα να αντιστοιχεί στην κλάση της καλοήθειας, ενώ για τις τιμές 700 εώς 1250, στην κλάση της κακοήθειας.



Εικόνα 3. 16 *Γράφημα πυκνότητας πιθανότητας της μεταβλητής area_mean*

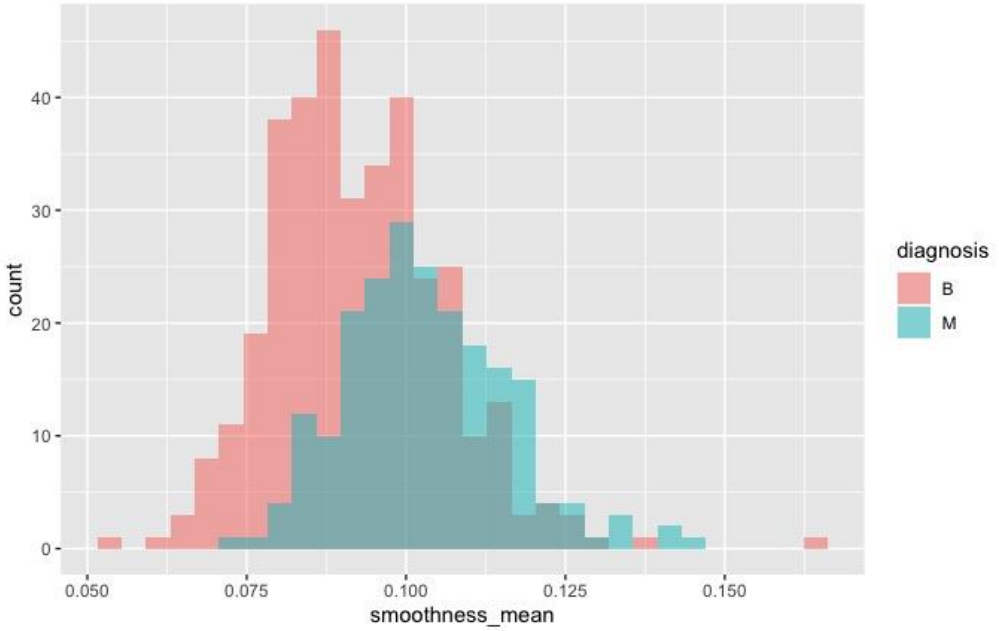
5. Μέση τιμή απαλότητας (smoothness_mean)

Οι όγκοι και οι μάζες του μαστού εμφανίζονται συνήθως με τη μορφή πυκνών περιοχών στις μαστογραφίες. Οι καλοήθεις μάζες έχουν γενικά λεία, στρογγυλά και καλά περιεγραμμένα όρια, σε αντίθεση με τους κακοήθεις όγκους, οι οποίοι συνήθως έχουν κηλιδωμένα, τραχιά και θολά όρια. (Homer, 1996). Στον **πίνακα 3.9**, αναγράφονται τα κυριότερα χαρακτηριστικά του γνωρίσματος της μέσης τιμής απαλότητας Αναλυτικότερα, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης απαλότητας ίση με 0.05 και οι κακοήθεις ίση με 0.07. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.08 ενώ 0.09 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση ύφη του χαρακτηριστικού μέσης απαλότητας του κυτταρικού πυρήνα τουλάχιστον ίση με 0.09 ενώ οι κακοήθεις όγκοι 0.1. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή για το χαρακτηριστικό μέσης απαλότητας ίση με 0.09 και 0.1 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης απαλότητας ίση με τουλάχιστον 0.1 και 0.11 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.16 για τους καλοήθεις όγκους και 0.14 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.1 όπως επίσης και αυτών των κακοηθών.

smoothness_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.05	0.08	0.09	0.09	0.10	0.16	0.01	0.11
Malignant	0.07	0.09	0.10	0.10	0.11	0.14	0.01	0.07

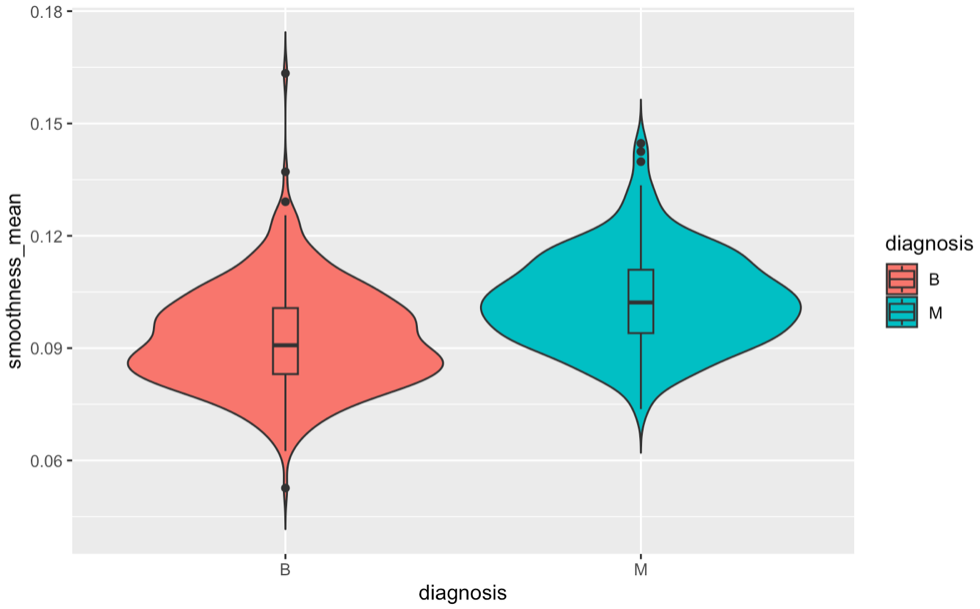
Πίνακας 3. 9 *Περιγραφικά μέτρα της μεταβλητής smoothness_mean*

Έπειτα, στην εικόνα 3.15, αναπαρίσταται το ιστόγραμμα της μεταβλητής που αφορά το γνώρισμα της μέσης απαλότητας διαχωρίζοντας τους όγκους ανάλογα την κλάση στην οποία ανήκουν. Οι περισσότεροι καλοήθεις όγκοι παρουσιάζουν μέση απαλότητα περίπου ίση με 42



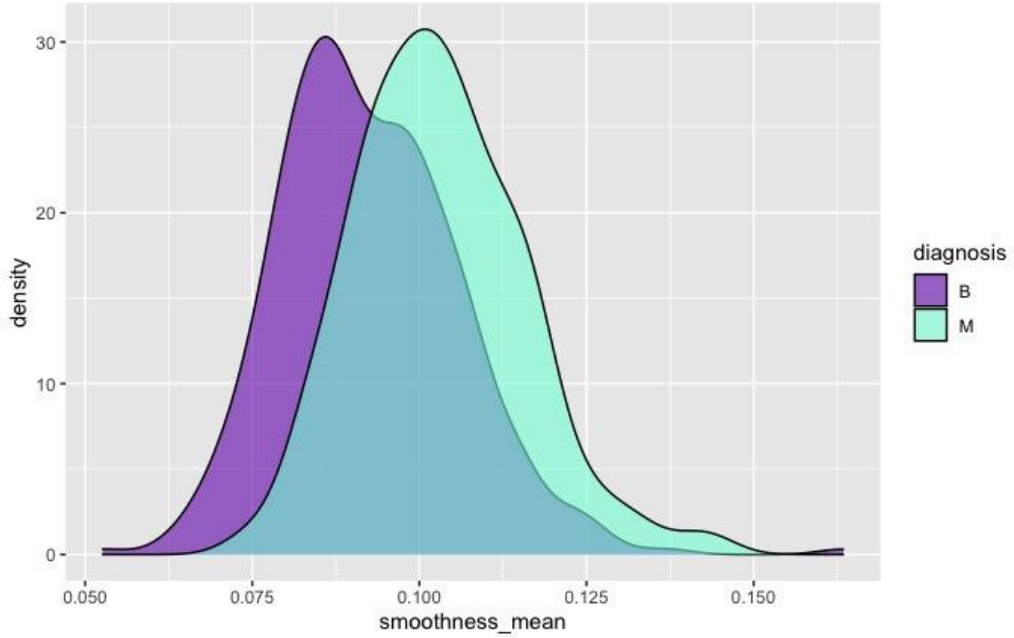
Εικόνα 3. 17 *Ιστόγραμμα μεταβλητής smoothness_mean*

Στην εικόνα 3.16, φαίνονται τα violin plots για κάθε κλάση τα οποία υποδεικνύουν πως κατανέμονται ομοιόμορφα με την καλοήθεια να έχει μία ιδιαίτερα ακραία τιμή κοντά στο 4^{ο} τεταρτημόριο, καθώς και η πιθανότητα να ανήκει ο όγκος σε κάθε κλάση ανάλογα με τη μέση τιμή της απαλότητας στο διάγραμμα πυκνότητας πιθανότητας της εικόνας 3.17.



Εικόνα 3. 18 *Violin plots για τη μεταβλητή smoothness_mean*

Από το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.17**, φαίνεται πως για το πεδίο ορισμού των τιμών μέσης (0.08, 0.087) παρουσιάζεται η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της καλοήθειας, ενώ για το πεδίο ορισμού (0.09, 0.11) στην κλάση της κακοήθειας.



Εικόνα 3. 19 *Γράφημα πυκνότητας πιθανότητας της μεταβλητής smoothness_mean*

6. Μέση τιμή συμπαγότητας (compactness_mean)

Η ανεξάρτητη μεταβλητή ‘’compactness_mean’’ αφορά το γνώρισμα της μέσης απαλότητας των όγκων του κάθε ασθενή. Η περίμετρος και το εμβαδόν συνδυάζονται χρησιμοποιώντας τον τύπο:

$$compactness = \left(\frac{perimeter^2}{area - 1}\right)$$

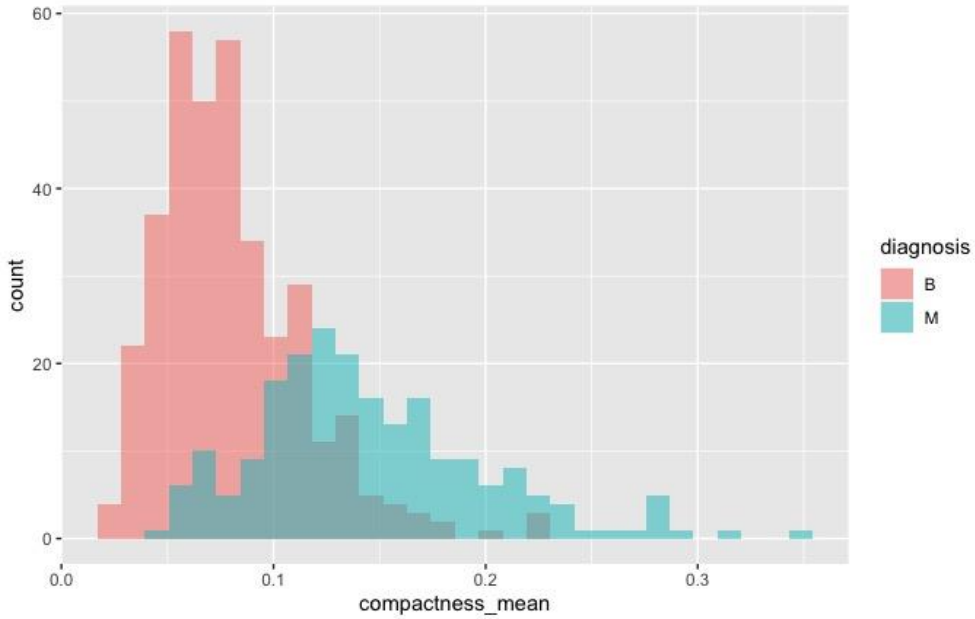
όπου ‘’perimeter’’ η περίμετρος και όπου ‘’area’’ το μέγεθος του κυτταρικού πυρήνα, για να λάβουμε ένα μέτρο συμπαγότητας της περιοχής των κυτταρικών πυρήνων. Στον **πίνακα 3.10** αναφέρονται τα κυριότερα χαρακτηριστικά του γνωρίσματος της μέσης απαλότητας βάσει των οποίων προκύπτει πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης συμπαγότητας ίση με 0.02 και οι κακοήθεις ίση με 0.50. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.06 ενώ 0.11 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση συμπαγότητα τουλάχιστον ίση με 0.08 ενώ οι κακοήθεις όγκοι 0.13. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού μέσης συμπαγότητας ίση με 0.08 και 0.15 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης συμπαγότητας ίση με τουλάχιστον 0.1 και 0.17 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.22 για τους καλοήθεις όγκους και 0.35 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.2 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 0.3 κατανέμονται οι κακοήθεις όγκοι.

compactness_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.02	0.06	0.08	0.08	0.10	0.22	0.03	0.20
Malignant	0.05	0.11	0.13	0.15	0.17	0.35	0.05	0.30

Πίνακας 3.10 *Περιγραφικά μέτρα της μεταβλητής μέσης συμπαγότητας*

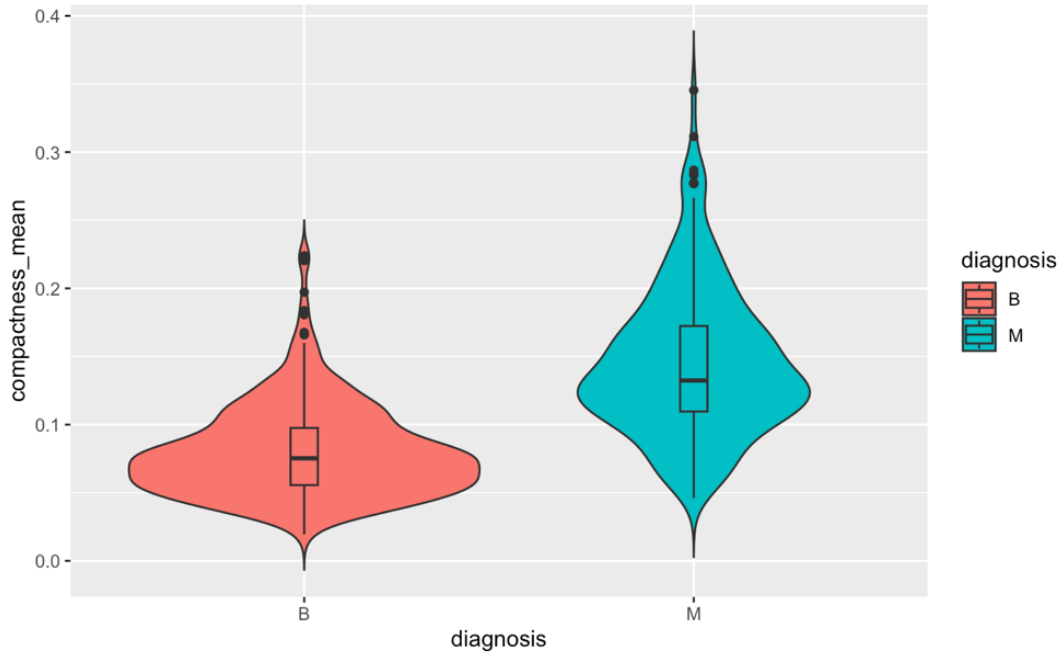
Στην **εικόνα 3.18** όπου παρουσιάζεται το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’compactness_mean’’, είναι ευδιάκριτο πως όγκοι οι οποίοι έχουν υψηλές τιμές για το

γνώρισμα της μέσης απαλότητας είναι πιθανότερο να ανήκουν στην κλάση της κακοήθειας. Επίσης, οι περισσότερες τιμές της καλοήθειας, είναι συγκεντρωμένες και κατανεμημένες σε μικρότερο εύρος τιμών.



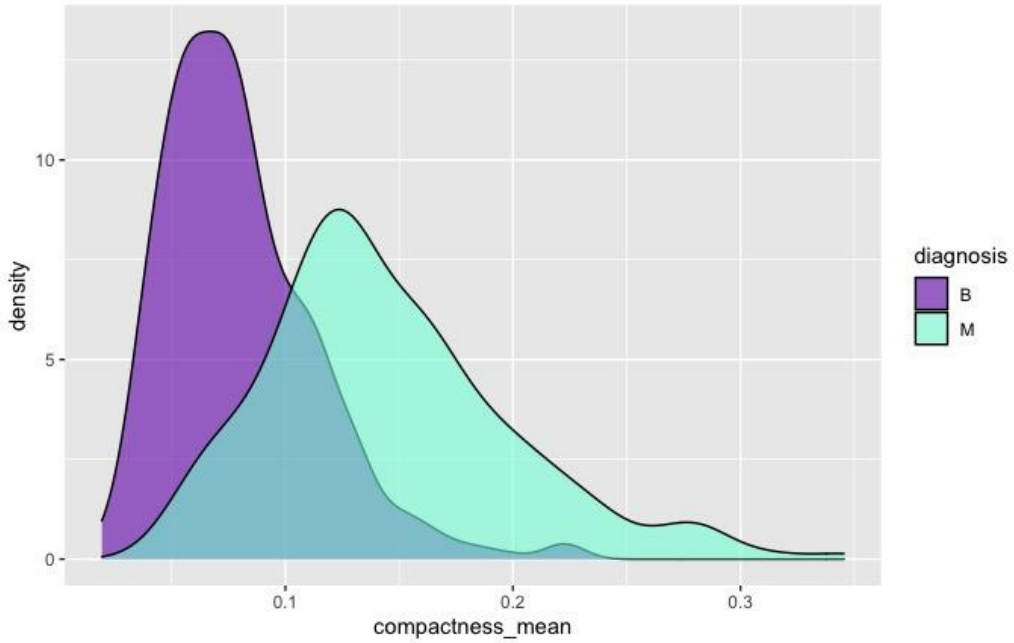
Εικόνα 3.20 *Ιστόγραμμα ανεξάρτητης μεταβλητής compactness_mean*

Επίσης, όσον αφορά τα violin plots της **εικόνας 3.19**, φαίνεται πως τόσο οι καλοήθεις όσο και οι κακοήθεις όγκοι έχουν αρκετές ακραίες τιμές. Για την καλοήθεια οι ακραίες τιμές κυμαίνονται γύρω από την τιμή του χαρακτηριστικού μέσης συμπαγότητας ίση με 0.2 ενώ για την κακοήθεια γύρω από την τιμή 0.3.



Εικόνα 3.21 *Violin plots της μεταβλητής compactness_mean*

Όσον αφορά την κατανομή των πιθανοτήτων (**εικόνα 3.20)**, για την κλάση της καλοήθειας, για το πεδίο ορισμού των τιμών του χαρακτηριστικού μέσης συμπαγότητας του κυτταρικού πυρήνα (0.05, 0.07) είναι πιθανότερο ο εκάστοτε όγκος να ανήκει στην κλάση της καλοήθειας, ενώ για το πεδίο ορισμού (0.11, 0.12) είναι πιθανότερο να ανήκει στην κλάση της κακοήθειας.



Εικόνα 3. 22 *Γράφημα πυκνότητας πιθανότητας της μεταβλητής compactness_mean*

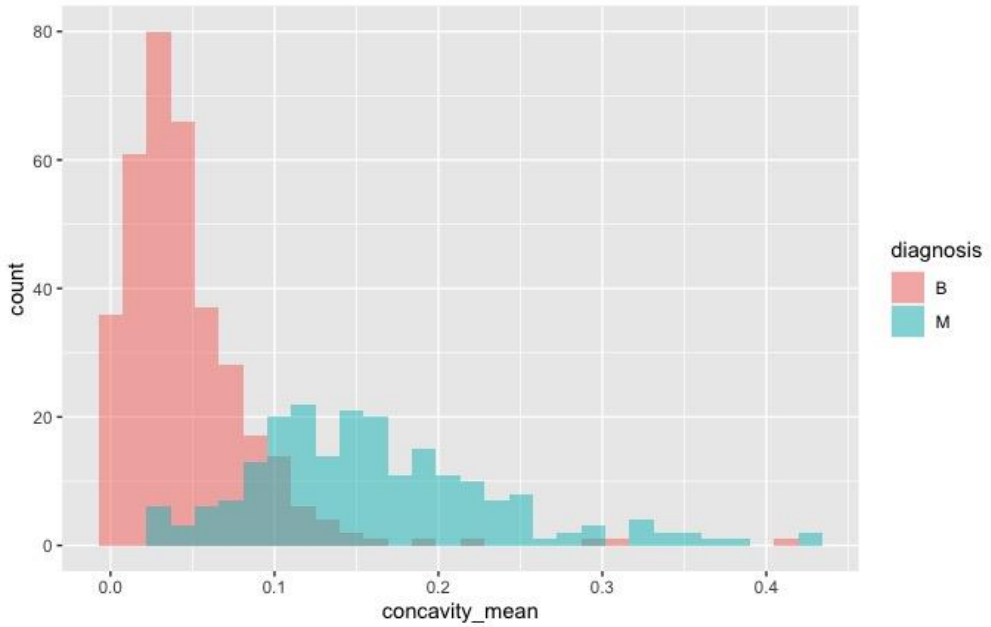
7. Μέση τιμή κοιλότητας του κυτταρικού πυρήνα (concavity_mean)

Μια υψηλή κοιλότητα σημαίνει ότι το όριο του κυτταρικού πυρήνα έχει εσοχές υποδηλώνοντας ότι είναι τραχύ παρά ομαλό. Τα βασικά περιγραφικά χαρακτηριστικά του χαρακτηριστικού που αφορά τη μέση κοιλότητα του κυτταρικού πυρήνα έχουν ως εξής. Αρχικά, από τον **πίνακα 3.11**, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης κοιλότητας ίση με 0 και οι κακοήθεις ίση με 0.02. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.02 ενώ 0.11 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση κοιλότητα τουλάχιστον ίση με 0.04 ενώ οι κακοήθεις όγκοι 0.15. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού μέσης κοιλότητας ίση με 0.05 και 0.16 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης κοιλότητας ίση με τουλάχιστον 0.06 και 0.2 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.41 για τους καλοήθεις όγκους και 0.43 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.41 ενώ σε ελάχιστα μικρότερο το οποίο ισούται με 0.40 κατανέμονται οι κακοήθεις όγκοι.

concavity_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.00	0.02	0.04	0.05	0.06	0.41	0.04	0.41
Malignant	0.02	0.11	0.15	0.16	0.20	0.43	0.08	0.40

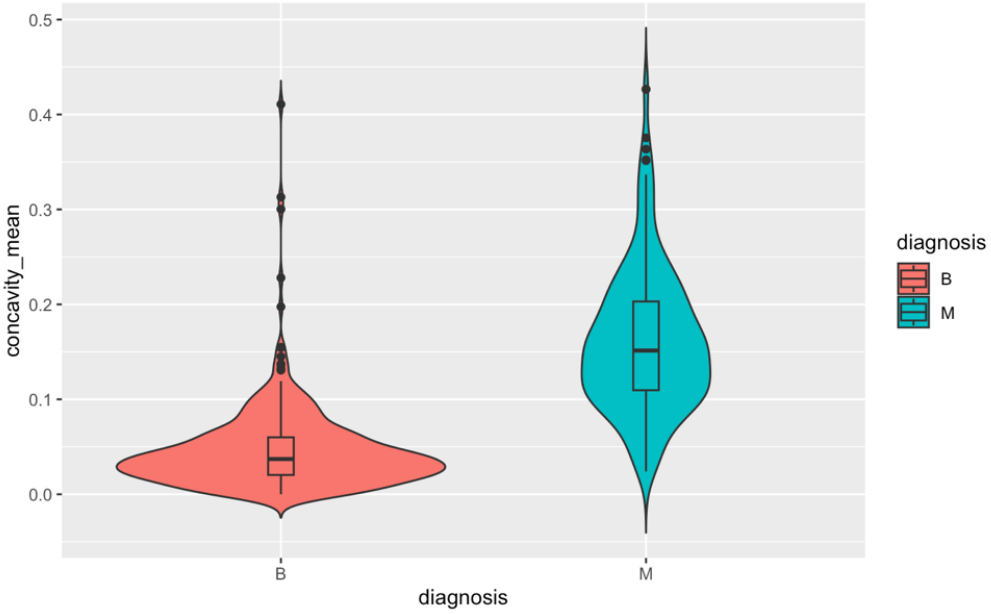
Πίνακας 3. 11 *Περιγραφικά μέτρα της μεταβλητής μέση τιμή κοιλότητας*

Στην **εικόνα 3.21** παρουσιάζεται το ιστόγραμμα της μεταβλητής ‘’concavity_mean’’, όπου υποδεικνύει ότι η καλοήθεια λαμβάνει κατά κύριο λόγο μικρότερες τιμές σε σχέση με την κακοήθεια, καθώς είναι κατανεμημένες σε πολύ μικρότερο εύρος σε σχέση με την κακοήθεια. Οι περισσότερες παρατηρήσεις για την κλάση τα καλοήθειας λαμβάνου τιμή για το χαρακτηριστικό μέσης κοιλότητας περίπου ίση με 0.03, ενώ για την κακοήθεια τιμή ίση περίπου με 0.13.



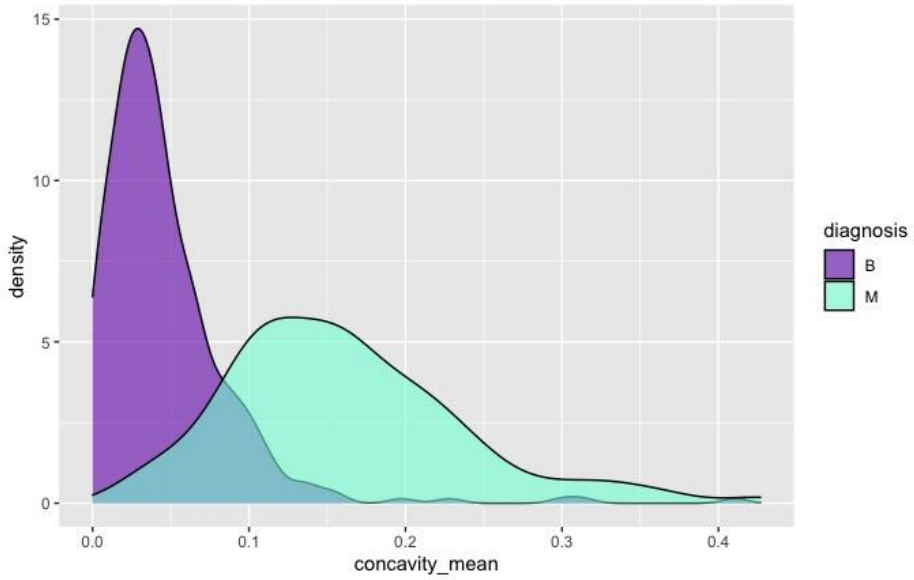
Εικόνα 3. 23 *Ιστόγραμμα της μεταβλητής concavity_mean*

Σύμφωνα με το violin plot της **εικόνας 3.21,** στην κλάση της καλοήθειας ανήκουν αρκετές ακραίες τιμές σε σχέση με αυτές που ανήκουν στην κλάση της κακοήθειας οι οποίες κυμαίνονται γύρω από την τιμή μέσης κοιλότητας ίση με 0.03 ενώ για την κακοήθεια γύρω από την τιμή ίση με 0.4



Εικόνα 3. 24 *Violin plots για τη μεταβλητή concavity_mean*

Στην **εικόνα 3.23**, παρουσιάζεται το γράφημα πυκνότητας πιθανότητας όπου υποδεικνύει την πιθανότητα κατά την οποία ένας όγκος ανήκει στην εκάστοτε κλάση βάσει της τιμής που λαμβάνει το χαρακτηριστικό της μέσης κοιλότητας. Ως εκ τούτου ένας όγκος παρουσιάζει τη μέγιστη πιθανότητα να ναι καλοήθεις για την τιμή μέσης κοιλότητας του κυτταρικού πυρήνα ίση με περίπου 0.25, ενώ για την κακοήθεια τιμή εντός του πεδίου ορισμού (0.1, 0.15).



Εικόνα 3. 25 *Γράφημα πυκνότητας πιθανότητας της μεταβλητής concavity_mean*

8.Μέσο πλήθος κοίλων τμημάτων (concave.points_mean)

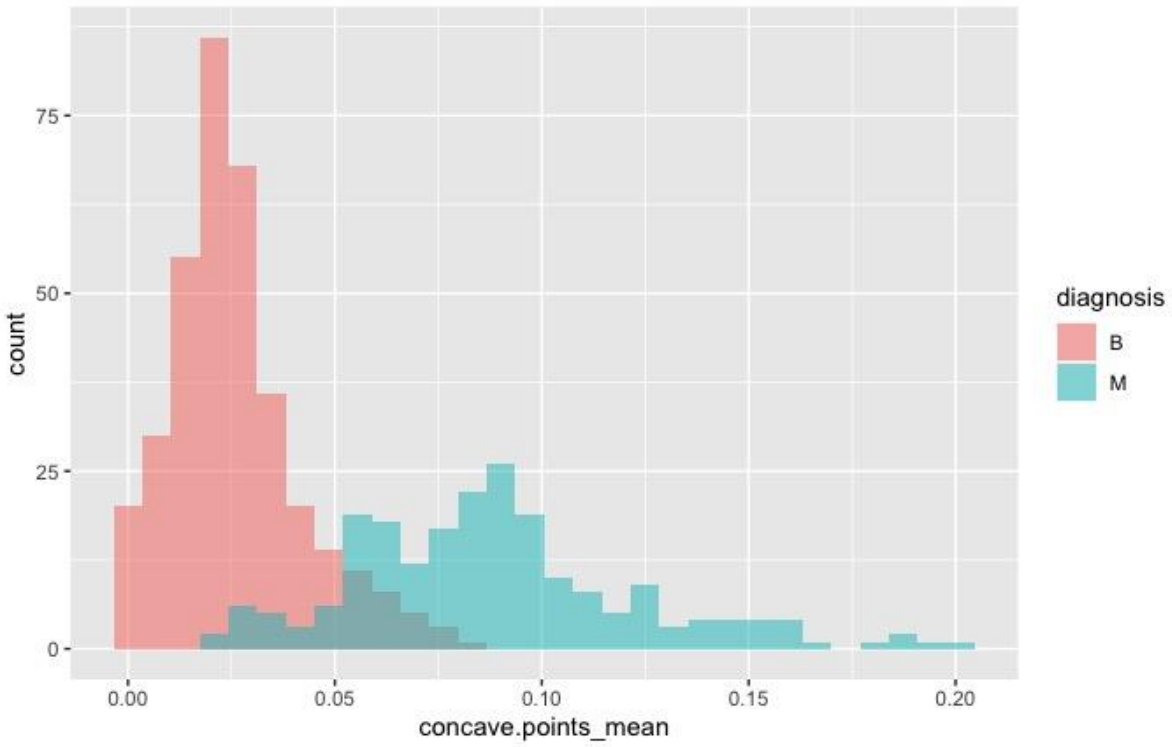
Το χαρακτηριστικό αυτό αφορά τον μέσο αριθμό κοίλων τμημάτων του περιγράμματος του κυτταρικού πυρήνα. Στον **πίνακα 3.12**, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσου πλήθους κοίλων τμημάτων ίση με 0 και οι κακοήθεις ίση με 0.02. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.02 ενώ 0.06 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέσο πλήθος κοίλων τμημάτων τουλάχιστον ίσο με 0.02 ενώ οι κακοήθεις όγκοι 0.09. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού μέσου πλήθους κοίλων τμημάτων ίση με 0.03 και 0.09 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσου πλήθους κοίλων τμημάτων ίση με τουλάχιστον 0.03 και 0.1 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.09 για τους καλοήθεις όγκους και 0.2 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το

εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.9 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 0.18 κατανέμονται οι κακοήθεις όγκοι.

concave.points_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.00	0.02	0.02	0.03	0.03	0.09	0.02	0.09
Malignant	0.02	0.06	0.09	0.09	0.10	0.20	0.03	0.18

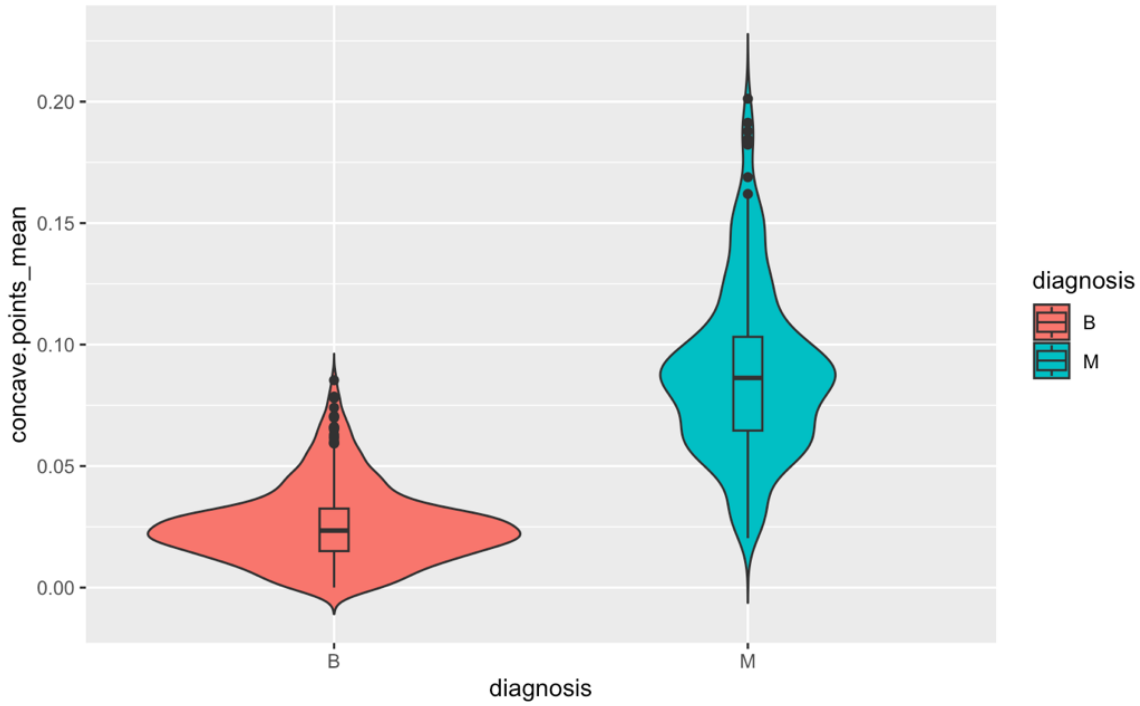
Πίνακας 3. 12 *Περιγραφικά μέτρα της μεταβλητής μέσου πλήθους κοίλων τμημάτων*

Στην **εικόνα 3.24**, παρουσιάζεται το ιστόγραμμα της μεταβλητής ‘’concave.points_mean’’. Όπου φαίνεται ουσιαστικά το πλήθος των παρατηρήσεων που ανήκουν σε κάθε κλάση για κάθε τιμή του χαρακτηριστικού του μέσου αριθμού κοίλων τμημάτων. Παρατηρείται πως για ακόμα μια φόρα οι όγκοι οι οποίοι ανήκουν στην κλάση της καλοήθειας κατανέμονται κυρίως στις μικρές τιμές του χαρακτηριστικού που αφορά τον μέσο αριθμό κοίλων τμημάτων ενώ οι κακοήθεια λαμβάνει τις μεγαλύτερες. Άρα όσο μεγαλύτερος είναι ο μέσος αριθμός κοίλων τμημάτων ενός όγκου τόσο πιθανότερο είναι να είναι κακοήθεις.



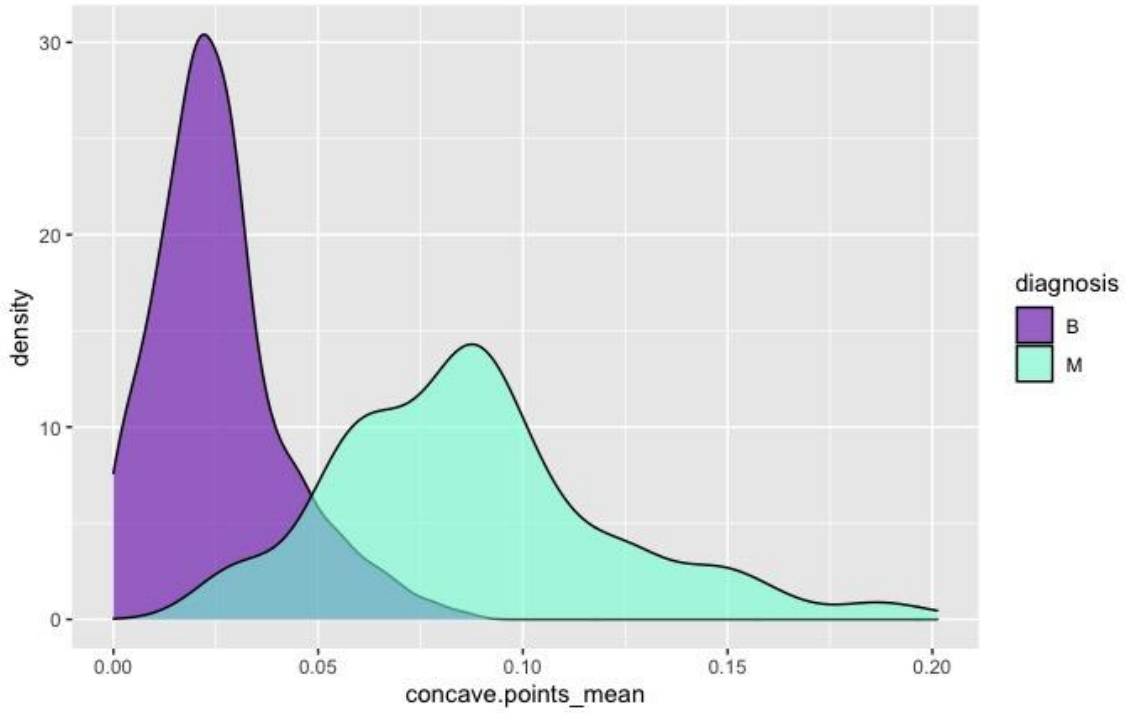
Εικόνα 3. 26 *Ιστόγραμμα ανεξάρτητης μεταβλητής concave.points_mean*

Έπειτα, στα violin plots, φαίνονται οι κατανομές που ακολουθούν οι παρατηρήσεις των δύο κλάσεων καθώς και τις ακραίες τους τιμές. Τόσο η κλάση της καλοήθειας όσο και η κλάση της κακοήθειας



Εικόνα 3. 27 *Violin plots της ανεξάρτητης μεταβλητής concave.points_mean*

Ενώ, στην **εικόνα 3.26,** παρουσιάζεται το γράφημα πυκνότητας πιθανότητας όπου υποδεικνύει την πιθανότητα κατά την οποία ένας όγκος ανήκει στην εκάστοτε κλάση βάσει της τιμής που λαμβάνει το χαρακτηριστικό του μέσου αριθμού κοίλων τμημάτων. Ειδικότερα, για την τιμή 0.025 του χαρακτηριστικό του μέσου αριθμού κοίλων τιμημάτων του κυτταρικού πυρήνα παρουσιάζεται η μέγιστη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας, ενώ για την κλάση της κακοήθειας μεγαλύτερη πιθανότητα να ανήκει ο όγκος στη συγκεκριμένη κλάση όταν η τιμή του χαρακτηριστικού ανήκει στο πεδίο ορισμού (0.075, 0.1).



Εικόνα 3. 28 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής concave.points_mean*

9.Μέση τιμή συμμετρίας

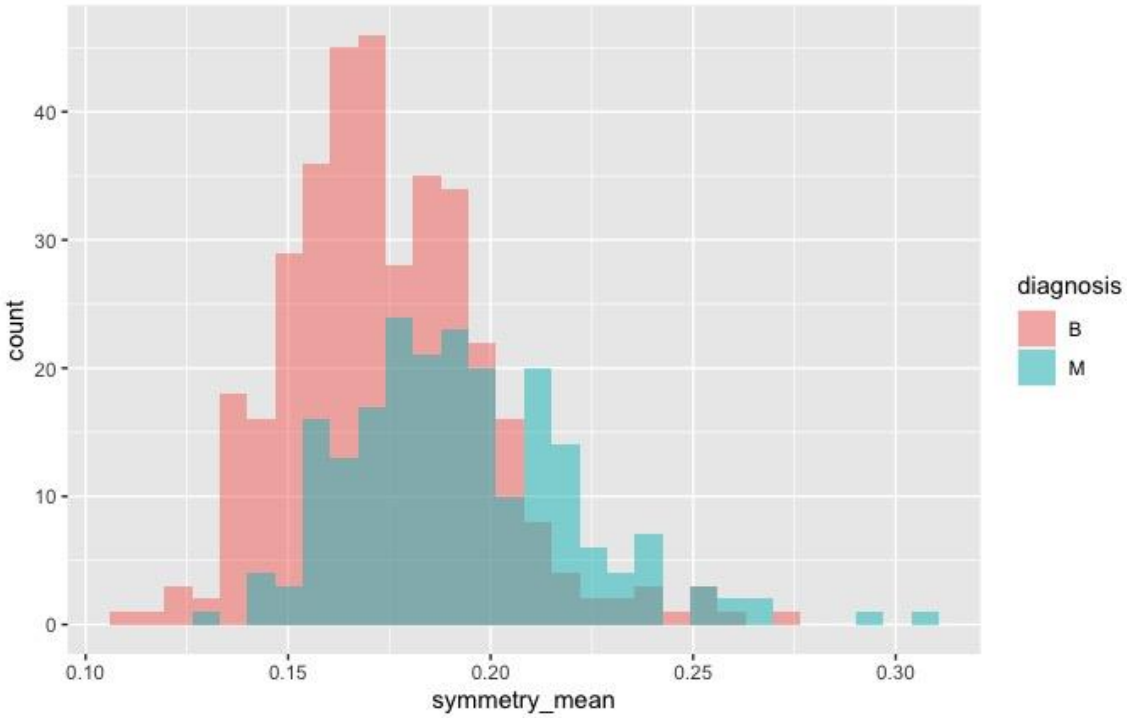
Η συμμετρία προκύπτει έπειτα από τον εντοπισμό της μεγαλύτερης ευθείας δύο οριακών σημείων η οποία διασχίζει το κέντρο του πυρήνα. Τα βασικά περιγραφικά μέτρα του χαρακτηριστικού της μέσης συμμετρίας παρουσιάζονται στον **πίνακα 3.13**. Αναλυτικότερα, παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης συμμετρίας ίση με 0.11 και οι κακοήθεις ίση με 0.03. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.16 ενώ 0.17 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση συμμετρία τουλάχιστον ίση με 0.17 ενώ οι κακοήθεις όγκοι 0.19. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού μέσης συμμετρίας ίση με 0.17 και 0.19 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης συμμετρίας ίση με τουλάχιστον 0.19 και 0.21 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.27 για τους καλοήθεις όγκους και 0.3 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι

παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και των δύο κλάσεων ισούται με 0.17.

symmetry_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.11	0.16	0.17	0.17	0.19	0.27	0.02	0.17
Malignant	0.13	0.17	0.19	0.19	0.21	0.30	0.03	0.17

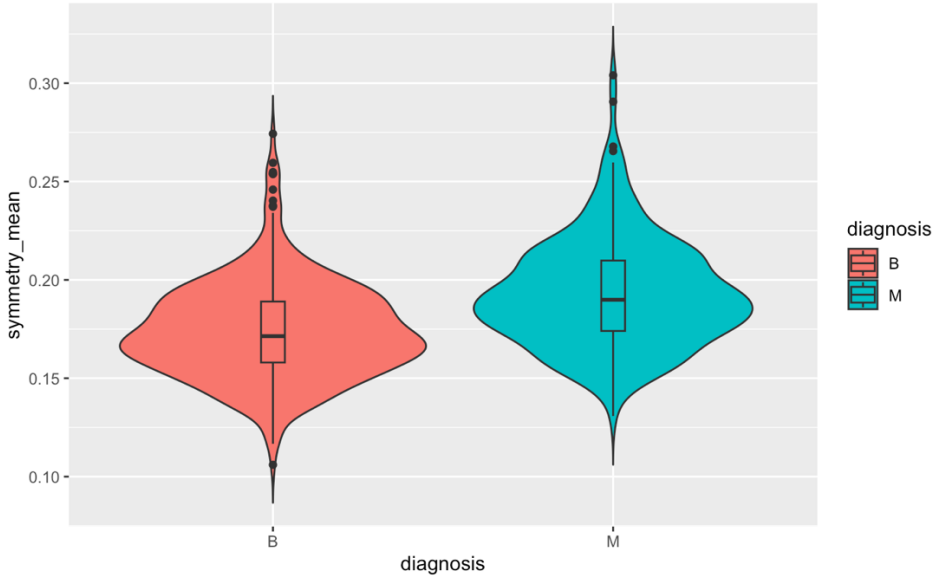
Πίνακας 3. 13 *Τα βασικά περιγραφικά μέτρα του χαρακτηριστικού της μέσης συμμετρίας*

Στην **εικόνα 3.27**, παρουσιάζεται το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’symmetry_mean’’. Παρατηρείται ότι οι περισσότεροι όγκοι οι οποίοι ανήκουν στην κλάση της καλοήθειας παρουσιάζουν μέση συμμετρία περίπου ίση με 0.17 ενώ οι περισσότεροι κακοήθεις όγκοι παρουσιάζουν τιμή μέσης συμμετρίας περίπου ίση με 0.18.



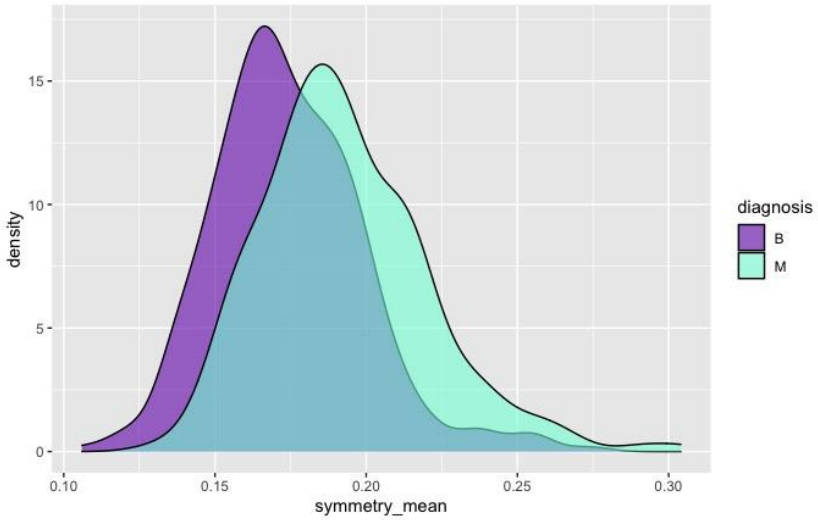
Εικόνα 3. 29 *Ιστόγραμμα ανεξάρτητης μεταβλητής symmetry_mean*

Από τα violin plots της **εικόνας 3.28**, παρατηρούμε ότι διάμεσος της κακοήθειας βρίσκεται πιο ψηλά σε σχέση με την καλοήθεια συναρτήσει του χαρακτηριστικού μέσης συμμετρίας. Επίσης, οι έκπτωτες τιμές της καλοήθειας κυμαίνονται γύρω από την τιμή μέσης συμμετρίας 0.25 ενώ της κακοήθειας με 0.27 παρουσιάζοντας μεγαλύτερο εύρος.



Εικόνα 3. 30 *Violin plots της ανεξάρτητης μεταβλητής symmetry_mean*

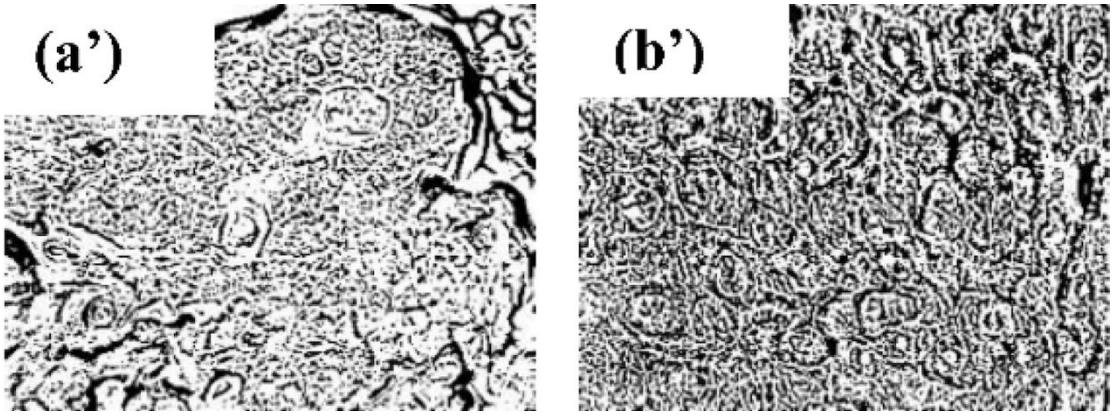
Στη συνέχεια, παρατηρώντας το γράφημα πυκνότητας πιθανότητας του γραφήματος της **εικόνας 3.29**, είναι ευδιάκριτο πως για το εύρος τιμών (0.16, 0.17) του χαρακτηριστικού που αντιπροσωπεύει τη μέση συμμετρία του κυτταρικού πυρήνα είναι πιθανότερο ο εξεταζόμενος όγκος να ανήκει στην κλάση της καλοήθειας ενώ για το εύρος (0.175, 0.2) στην κακοήθεια.



Εικόνα 3. 31 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής symmetry_mean*

10. Μέση τιμή μορφοκλασματικής διάστασης (fractal_dimension_mean)

Με σκοπό την αποδοτικότερη μύηση στην έννοια που αντιπροσωπεύει το χαρακτηριστικό της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα, είναι απαραίτητες οι αναφορές αυτής. Αναλυτικότερα, η μορφοκλασματική διάσταση, ενός αντικειμένου και συγκεκριμένα ενός βιολογικού ιστού αποτελεί ένα μέτρο το οποίο υποδηλώνει το πόσο παρόμοια παραμένει η δομή του έπειτα από αλλαγές των διαστάσεων του. Ένας βιολογικός ιστός από τη φύση του αποτελεί σχήμα μορφοκλασματικής διάστασης του οποίου η μορφοκλασματική διάσταση αλλάζει συναρτήσει του σταδίου του καρκίνου (Elkington, Adhikari, & Pradhan, 2022)



Εικόνα 3. 32 *Ενδεικτική απεικόνιση της μορφοκλασματικής διάστασης ενός ιστού από την αρχή (a'), έως το τρίτο στάδιο καρκίνου (b')*

Η μορφοκλασματική διάσταση (fractal dimension) υπολογίζεται από την «προσέγγιση της ακτογραμμής». Η περίμετρος του πυρήνα μπορεί να μετρηθεί χρησιμοποιώντας διαφορετικά μήκη ραβδιών μέτρησης. Καθώς αυτό το μήκος αυξάνεται, το συνολικό μήκος της μετρούμενης «ακτογραμμής» μειώνεται λόγω μικρότερης ακρίβειας της μέτρησης.

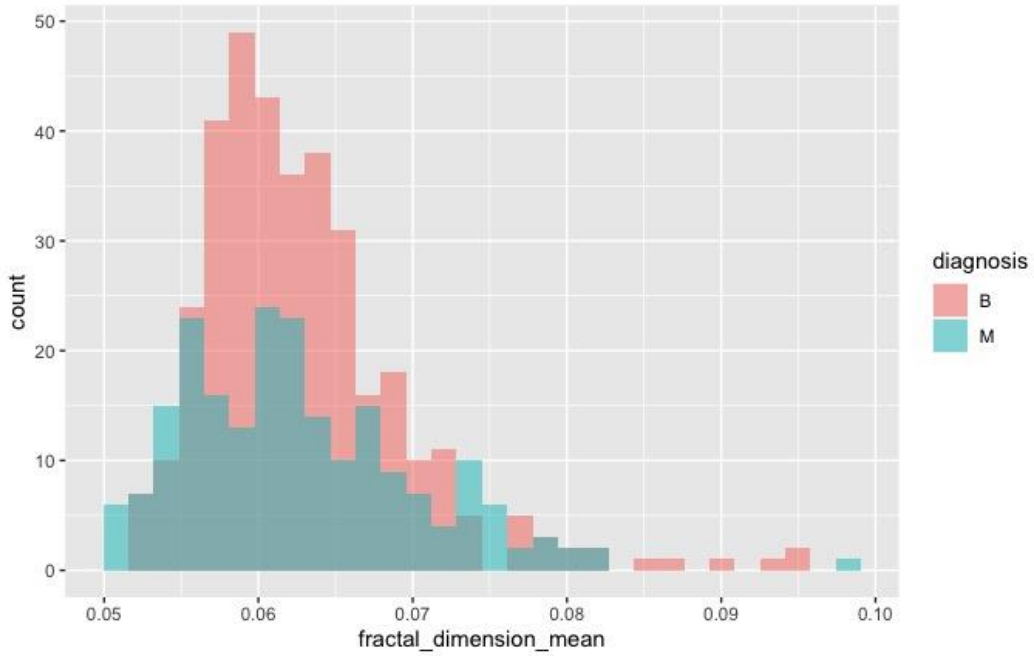
Τα βασικά περιγραφικά μέτρα του χαρακτηριστικού μέση τιμή μορφοκλασματικής διάστασης παρουσιάζονται στον **πίνακα 3.14**. Παρατηρείται πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή μέσης μορφοκλασματικής διάστασης ίση με 0.0519και οι κακοήθεις ίση με 0.0499. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.0585 ενώ 0.0566 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν μέση τιμή μορφοκλασματικής διάστασης του κυτταρικού πυρήνα ίση με 0.0615 ενώ οι κακοήθεις όγκοι 0.0616. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού μέσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα ίση με 0.0629 και 0.0627 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή μέσης

μορφοκλασματικής διάστασης του κυτταρικού πυρήνα ίση με τουλάχιστον 0.0658 και 0.0671 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.0958 για τους καλοήθεις όγκους και 0.0974 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.439 ενώ σε ελάχιστα μικρότερο το οποίο ισούται με 0.0475 κατανέμονται οι κακοήθεις όγκοι.

fractal_dimension_mean	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0519	0.0585	0.0615	0.0629	0.0658	0.0958	0.0067	0.0439
Malignant	0.0499	0.0566	0.0616	0.0627	0.0671	0.0974	0.0076	0.0475

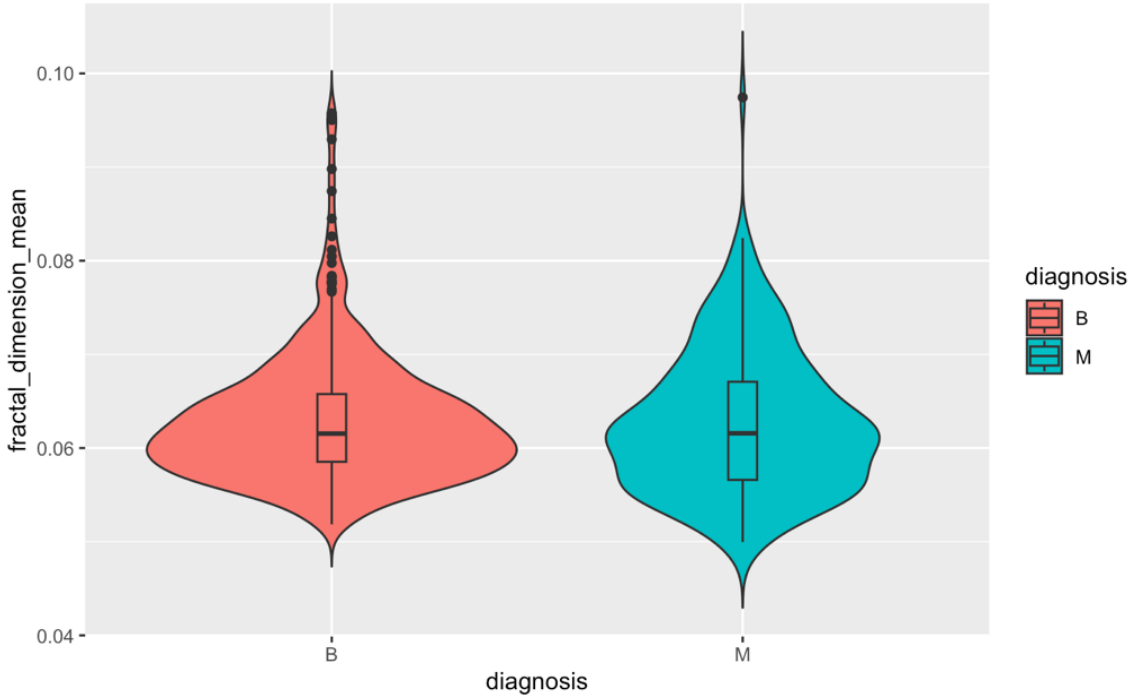
Πίνακας 3. 14 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού μέσης τιμής μορφοκλασματικής διάστασης.*

Έπειτα, ακολουθεί το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’fractal_dimension_mean’’ στην **εικόνα 3.30**. Σύμφωνα λοιπόν με το ιστόγραμμα, οι όγκοι που ανήκουν στην κλάση της καλοήθειας λαμβάνουν περισσότερο την τιμή για το χαρακτηριστικό μέσης μορφοκλασματικής διάστασης ίση με περίπου 0.058, ενώ αυτοί που ανήκουν στην κλάση της κακοήθειας ίση με περίπου 0.061.



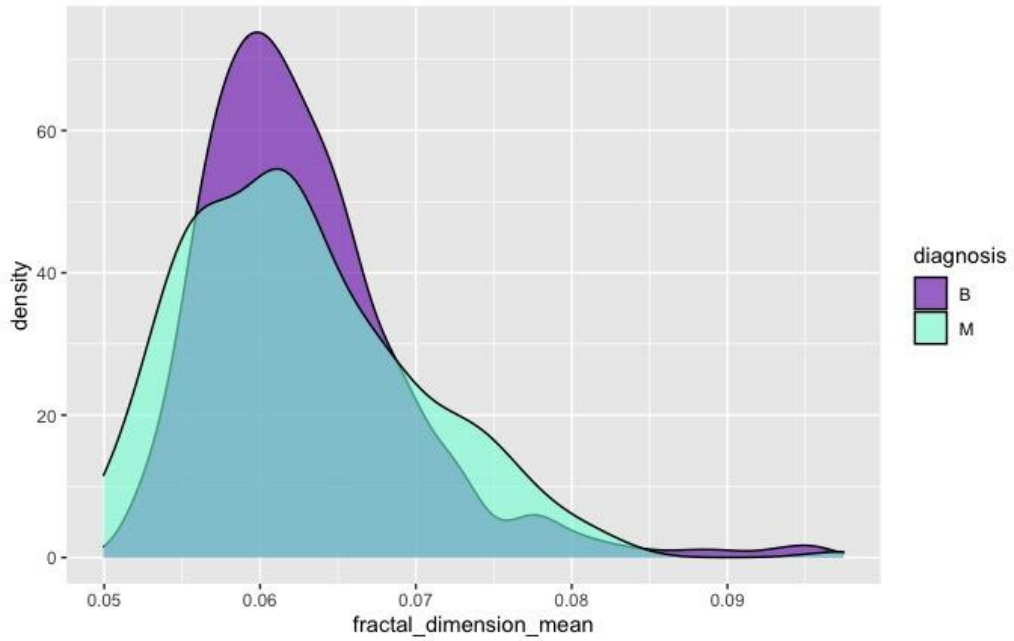
Εικόνα 3.33 *Ιστόγραμμα ανεξάρτητης μεταβλητής fractal_dimension_mean*

Όσον αφορά τα violin plots της **εικόνας 3.31**, φαίνεται πως η κλάση της καλοήθειας έχει αρκετές ακραίες παρατηρήσεις στο εύρος τιμών μέσης μορφοκλασματικής διάστασης (0.078, 10) σε αντίθεση με την κλάση της κακοήθειας στην οποία παρατηρείται μια μόνο ακραία παρατήρηση κοντά στην τιμή 0.98.



Εικόνα 3. 34 Violin plots για την ανεξάρτητη μεταβλητή

Επίσης, από το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.32**, για την τιμή μέσης μορφοκλασματικής διάστασης ίση με περίπου 0.06 αντιστοιχεί η μέγιστη πιθανότητα να ταξινομηθεί ο εξεταζόμενος όγκος στην κλάση της καλοήθειας, ενώ για την τιμή περίπου ίση με 0.062 να ταξινομηθεί στην κλάση της κακοήθειας.



Εικόνα 3. 35 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής fractal_dimension_mean*

Στη συνέχεια ακολουθούν οι ανεξάρτητες μεταβλητές οι οποίες αφορούν την τυπική απόκλιση των χαρακτηριστικών του κυτταρικού πυρήνα (**πίνακας 3.3**).

3.4.2 Μεταβλητές τυπικής απόκλισης παραγόντων

Ακολουθεί η δεύτερη κατηγορία στην οποία ανήκουν τα χαρακτηριστικά του κυτταρικού πυρήνα η οποία αντιπροσωπεύει τις τυπικές αποκλίσεις για κάθε χαρακτηρηηριστικό που προκύπτει από τις μετρήσεις μέσω δεδομένων εικόνας.

11. Τυπική απόκλιση ακτίνας (radius_se).

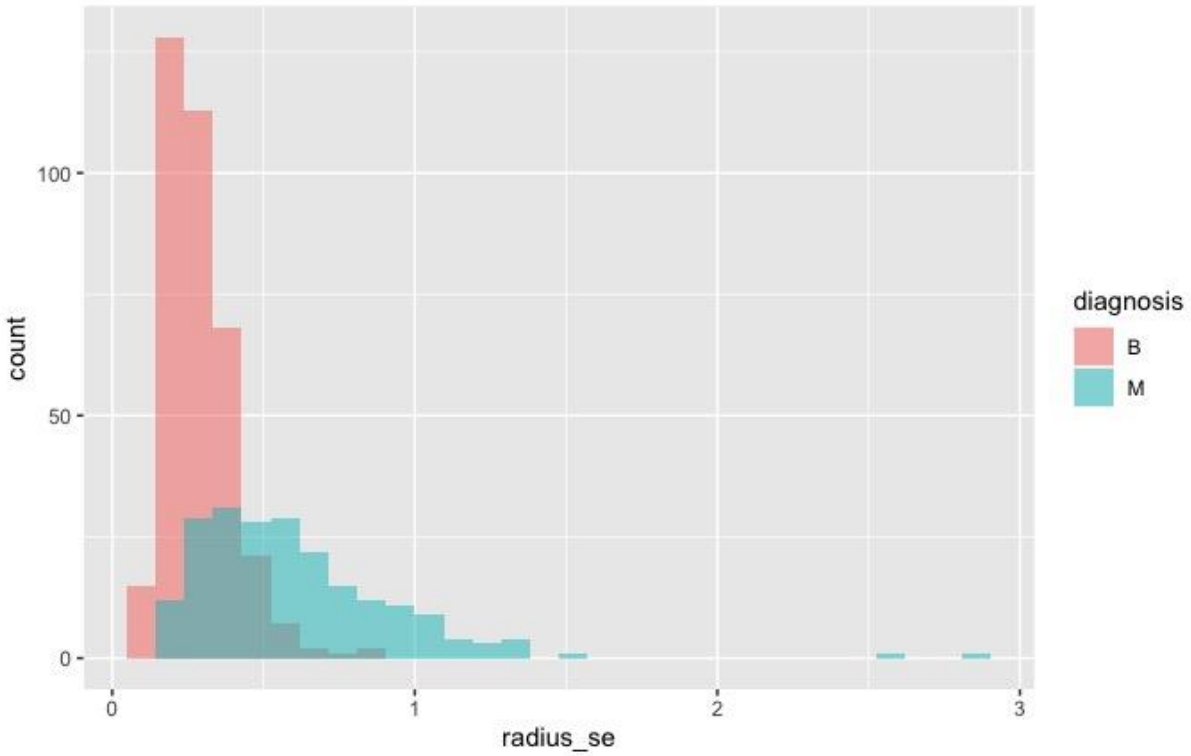
Στον **πίνακα 3.15** αναγράφονται τα βασικά περιγραφικά μέτρα του χαρακτηριστικού που αφορά την τυπική απόκλιση της ακτίνας του πυρήνα σε σημεία της περιμέτρου του. Ειδικότερα, οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης απόστασης του πυρήνα σε σημεία της περιμέτρου του ίση με 0.11 και οι κακοήθεις ίση με 0.19. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.21 ενώ 0.39 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπικής απόκλισης απόστασης του πυρήνα σε σημεία της περιμέτρου του τουλάχιστον ίσο με 0.26 ενώ οι κακοήθεις όγκοι 0.55. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης απόστασης του πυρήνα σε σημεία της περιμέτρου του ίση με 0.28 και 0.61 οι όγκοι οι οποίοι

ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης απόστασης του πυρήνα σε σημεία της περιμέτρου του ίση με τουλάχιστον 0.34 και 0.76 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.88 για τους καλοήθεις όγκους και 2.87 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.77 ενώ αρκετά μεγαλύτερο το οποίο ισούται με 3.21 κατανέμονται οι κακοήθεις όγκοι.

radius_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.11	0.21	0.26	0.28	0.34	0.88	0.11	0.77
Malignant	0.19	0.39	0.55	0.61	0.76	2.87	0.48	3.21

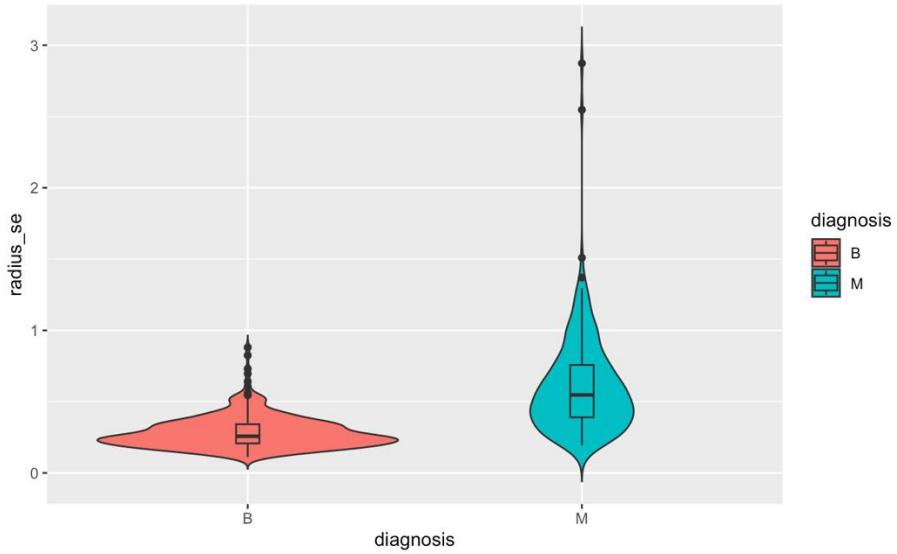
Πίνακας 3. 15 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης της ακτίνας.*

Στη συνέχεια ακολουθεί το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’radius_se’’(**εικόνα 3.33)**, το οποίο υποδεικνύει τις κατανομές των παρατηρήσεων των δύο κλάσεων ανάλογα με την τιμή που λαμβάνει το χαρακτηριστικό τυπικής απόκλισης της ακτίνας. Είναι ευδιάκριτο πως οι παρατηρήσεις που αφορούν τους καλοήθεις όγκους κατανέμονται σε ένα ‘’στενό’’ και ιδιαίτερα περιορισμένο εύρος σε αντίθεση με αυτό των κακοηθών. Ειδικότερα, από το ιστόγραμμα προκύπτει πως οι περισσότεροι καλοήθεις όγκοι παρουσιάζουν τυπική απόκλιση της ακτίνας του πυρήνα σε σημεία της περιμέτρου του περίπου ίση με 0.2 ενώ οι περισσότεροι κακοήθεις όγκοι περίπου ίση με 0.4.



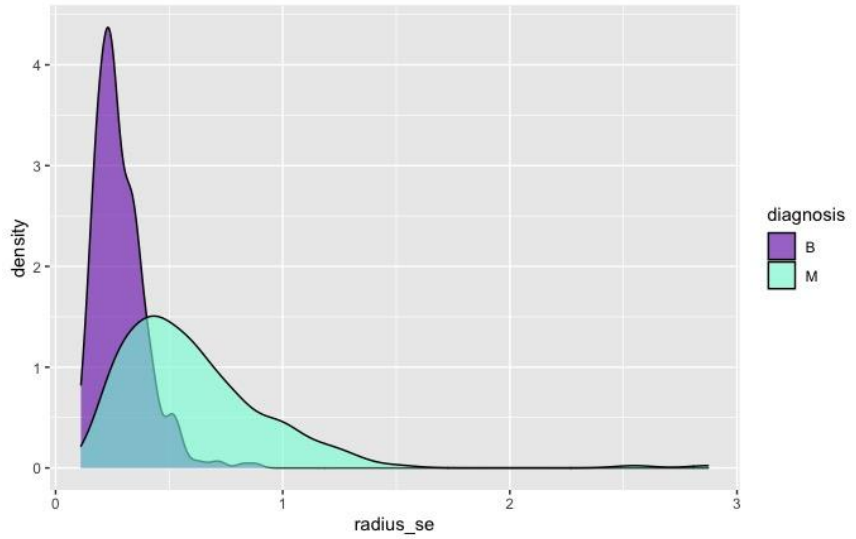
Εικόνα 3. 36 *Ιστόγραμμα ανεξάρτητης μεταβλητής radius_se*

Έπειτα, ακολουθούν τα violin plots της ανεξάρτητης μεταβλητής radius_se των δύο κλάσεων στην **εικόνα 3.35**, όπου φαίνεται εξίσου το πως κατανέμονται οι παρατηρήσεις των δύο κλάσεων παρέχοντας επίσης και την πληροφορία των ακραίων τιμών τους. Ειδικότερα, οι ακραίες τιμές των καλοηθών όγκων είναι περισσότερες από αυτές των κακοηθών και ως εκ τούτου κυμαίνονται γύρω από την τιμή της τυπικής απόκλισης της ακτίνας του πυρήνα σε σημεία της περιμέτρου του περίπου ίση με 0.75. Ωστόσο για την κλάση της κακοήθειας, ενώ οι ακραίες παρατηρήσεις είναι πολύ λιγότερες, παρουσιάζουν ένα ιδιαίτερα μεγαλύτερο εύρος. Πιο συγκεκριμένα παρατηρούνται δύο γύρω από την τιμή του χαρακτηριστικού ίση με περίπου με 1.5 και άλλες δύο γύρω από την τιμή περίπου ίση με 2.7.



Εικόνα 3. 37 *Violin plots της ανεξάρτητης μεταβλητής radius_se*

Τέλος, όσον αφορά το χαρακτηριστικό που αφορά την τυπική απόκλισης της ακτίνας του κυτταρικού πυρήνα σε σημεία της περιμέτρου του ακολουθεί το γράφημα πυκνότητας πιθανότητας (**εικόνα 3.36**). Σύμφωνα με το γράφημα λοιπόν εκμαιεύεται η πληροφορία που υποδεικνύει πως ένας όγκος του οποίου η τυπική απόκλιση της απόστασης του πυρήνα σε σημεία της περιμέτρου του ίση με περίπου 0.3 είναι πιθανότερο να ανήκει στην κλάση της καλοήθειας, ενώ για την τιμή του χαρακτηριστικού ίση με περίπου 0.5 είναι πιθανότερο να ανήκει στην κλάση της κακοήθειας.



Εικόνα 3. 38 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής radius_se*

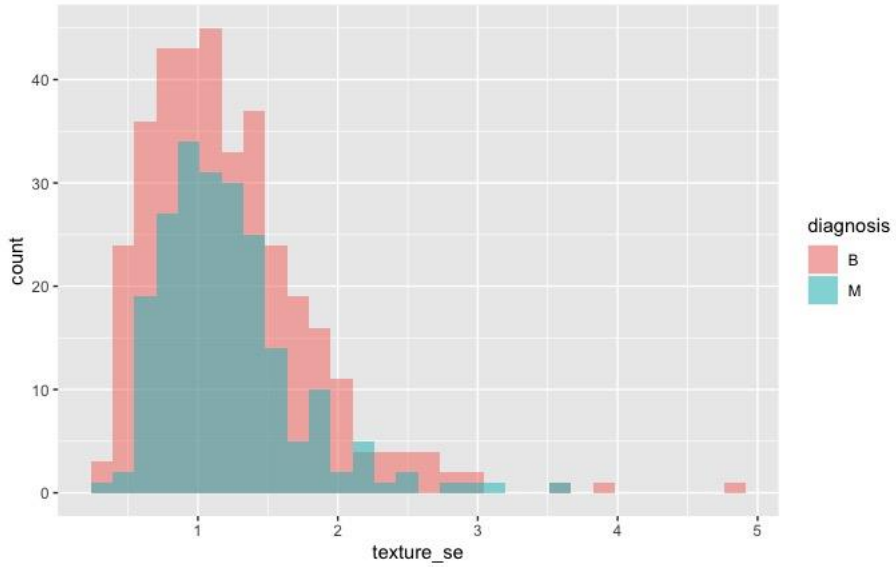
12.Τυπική απόκλιση υφής

Στη συνέχεια ακολουθεί το χαρακτηριστικό που αφορά την τυπική απόκλιση της υφής του κυτταρικού πυρήνα στην κλίμακα του γκρι, βασικά μέτρα του οποίου παρουσιάζονται στον **πίνακα 3.16**. Ειδικότερα, οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης υφής ίση με 0.36 και οι κακοήθεις ίση με 0.362. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ ίση με 0.8 ενώ ίση 0.89 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπικής απόκλισης υφής ίση με 1.11 ενώ οι κακοήθεις όγκοι 0.1. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης υφής ίση με 1.22 και 1.21 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης υφής ίση με τουλάχιστον 1.49 και 1.43 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 4.89 για τους καλοήθεις όγκους και 3.57 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 4.52 ενώ αρκετά μεγαλύτερο το οποίο ισούται με 20.65 κατανέμονται οι κακοήθεις όγκοι.

texture_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.36	0.80	1.11	1.22	1.49	4.89	0.59	4.52
Malignant	0.362	0.89	1.10	1.21	1.43	3.57	2.57	20.65

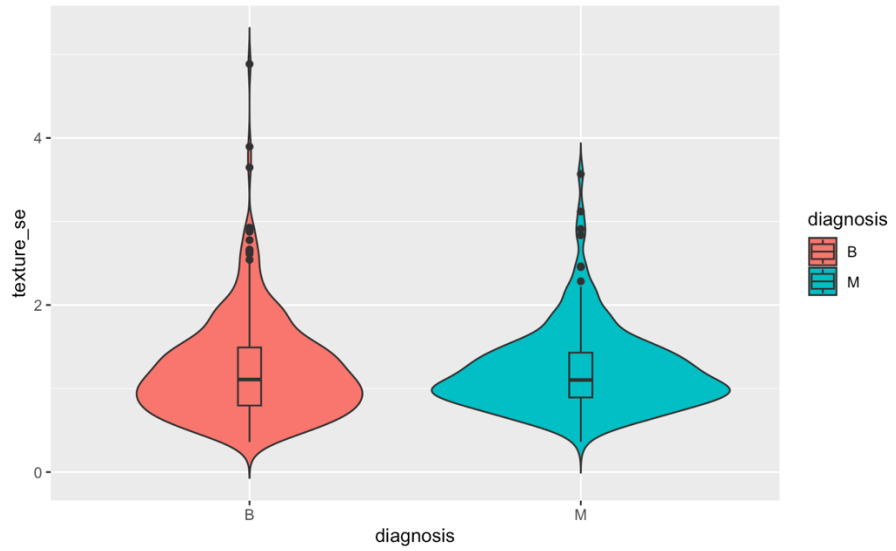
Πίνακας 3. 16 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης της υφής.*

Έπειτα παρουσιάζεται το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’texture_se’’ **(εικόνα 3.37)**, υποδεικνύοντας το κάθε πλήθος το οποίο αντιστοιχεί στην εκάστοτε τιμή της τυπικής απόκλισης της υφής του κυτταρικού πυρήνα ανάλογα στην κλάση στην οποία ανήκει. Πιο συγκεκριμένα, το μεγαλύτερο πλήθος των καλοηθών όγκων αντιστοιχούν στην τιμή του χαρακτηριστικού της τυπικής απόκλισης της υφής ίση με περίπου 1.2 ενώ για την κακοήθεια περίπου ίση με 0.8.



Εικόνα 3. 39 *Ιστόγραμμα ανεξάρτητης μεταβλητής texture_se*

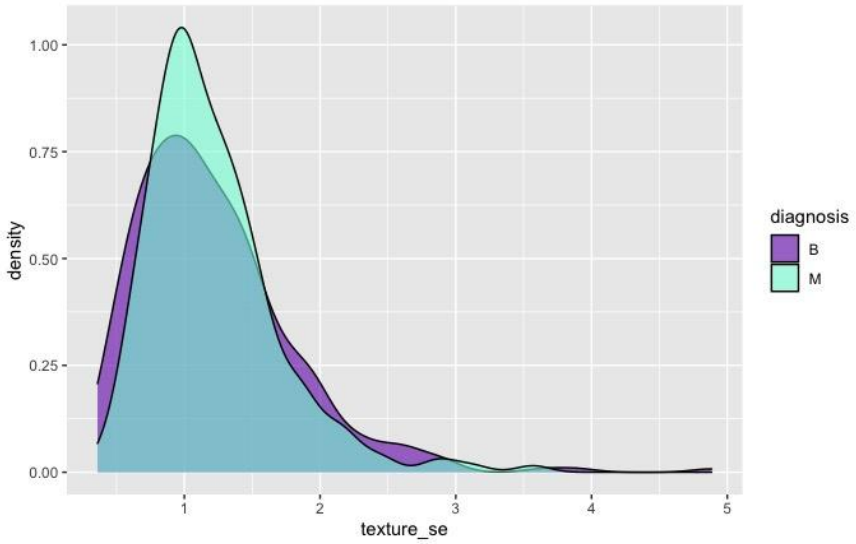
Στην **εικόνα 3.37**, παρουσιάζονται τα violin plots βάσεις των οποίων αντιλαμβάνεται πλήρως οι κατανομές των παρατηρήσεων, η πυκνότητα πιθανότητας αλλά και οι ακραίες τιμές κάθε κλάσης. Πιο συγκεκριμένα οι δύο κλάσεις έχουν παρόμοιες κατανομές με τις διαμέσους τους να βρίσκονται σχεδόν στο ίδιο επίπεδο. Ωστόσο οι ακραίες τιμές του χαρακτηριστικού τυπικής απόκλισης της υφής για τους καλοήθεις όγκους κατανέμονται σε πολύ μεγαλύτερο εύρος σε σχέση με αυτές της κακοήθειας.



Εικόνα 3. 40 *Violin plots της ανεξάρτητης μεταβλητής texture_se*

Στο γράφημα πυκνότητας της ανεξάρτητης μεταβλητής ‘’texture_se’’, είναι ευδιάκριτο πως για την τιμή της τυπικής απόκλισης της υφής του κυτταρικού πυρήνα ίση με περίπου 1,

παρουσιάζεται η μέγιστη πιθανότητα ένας όγκος τόσο στην καλοήθεια όσο και στην κακοήθεια.



Εικόνα 3. 41 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής texture_se*

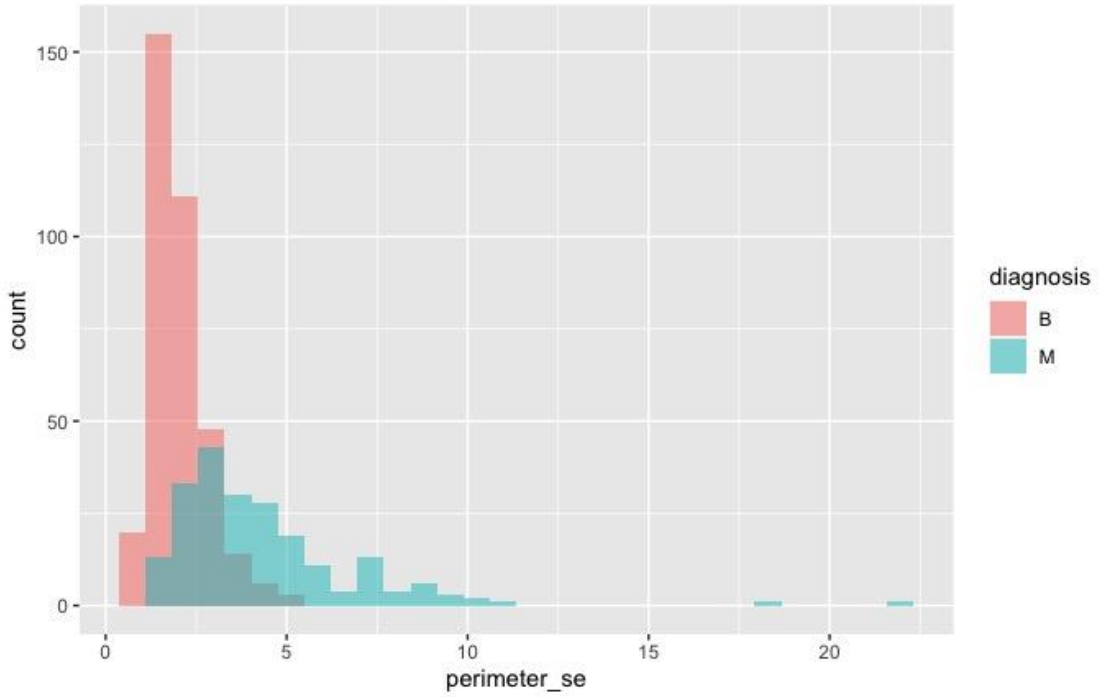
13.Τυπική απόκλιση περιμέτρου του κυτταρικού πυρήνα (perimeter_se).

Στη συνέχεια σειρά έχει το χαρακτηριστικό που αφορά την τυπική απόκλιση της περιμέτρου του κυτταρικού πυρήνα, βασικά μέτρα του οποίου παρουσιάζονται στον **πίνακα 3.17** όπου είναι ευδιάκριτο πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης της περιμέτρου ίση με 0.76 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 1.33. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 1.45 ενώ 2.72 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση υφής τουλάχιστον ίση με 1.85 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 3.68. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης υφής ίση με 2.00 και 4.32 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης υφής ίση με τουλάχιστον 2.39 και 5.21 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 5.12 για τους καλοήθεις όγκους και 21.98 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 4.36 ενώ αρκετά μεγαλύτερο το οποίο ισούται με 20.65 κατανέμονται οι κακοήθεις όγκοι

perimeter_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.76	1.45	1.85	2.00	2.39	5.12	0.77	4.36
Malignant	1.33	2.72	3.68	4.32	5.21	21.98	2.57	20.65

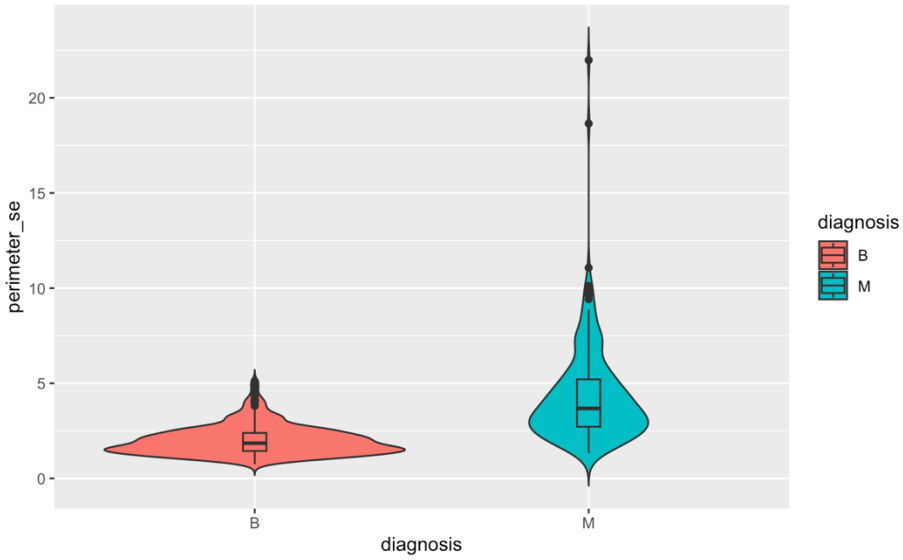
Πίνακας 3. 17 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης της περιμέτρου.*

Στην **εικόνα 3.39**, παρουσιάζεται το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’perimeter_se’’, όπου αντλείται η πληροφορία της αντιστοιχίας του μεγαλύτερου πλήθους παρατηρήσεων των καλοηθών όγκων στην τιμή του χαρακτηριστικού τυπικής απόκλισης της περιμέτρου ίση με περίπου 1.3 και 2.6 για των κακοηθών.



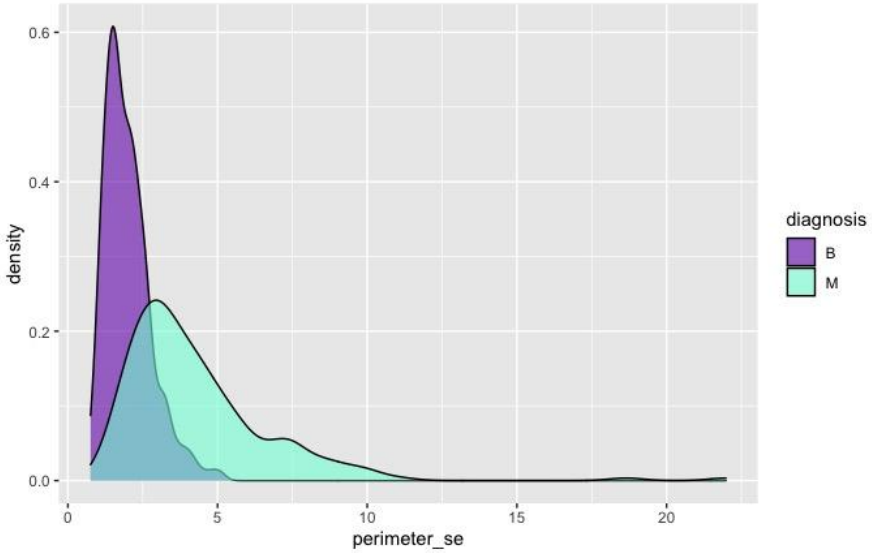
Εικόνα 3. 42 *Ιστόγραμμα ανεξάρτητης μεταβλητής perimeter_se*

Σύμφωνα με τις πρόσθετες πληροφορίες που παρέχουν τα violin plots της **εικόνας 3.40**, φαίνονται τόσο οι κατανομές των παρατηρήσεων της κάθε κλάσης αλλά και οι ακραίες τιμές αυτών. Οι ακραίες τιμές της καλοήθειας κατανέμονται σε πολύ μικρότερο εύρος σε σχέση με αυτές της κακοήθειας κυμαίνονται γύρω από την τιμή της μέσης περιμέτρου ίση με περίπου 5. Όσον αφορά την κλάση της κακοήθειας, παρατηρούνται δύο γύρω από την τιμή 20 και αρκετές γύρω από την τιμή 10 της μέσης περιμέτρου.



Εικόνα 3. 43 *Violin plots ανεξάρτητης μεταβλητής perimeter_se*

Επιπροσθέτως, από το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.41**, προκύπτει πως για τη τιμή της τυπικής απόκλισης ίση με περίπου 1.3 παρουσιάζεται η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της καλοήθειας, ενώ για την τιμή της τυπικής απόκλισης της περιμέτρου περίπου ίση με 2.6 παρουσιάζεται η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 44 *Γράφημα πυκνότητας πιναθότητας ανεξάρτητης μεταβλητής perimeter_se*

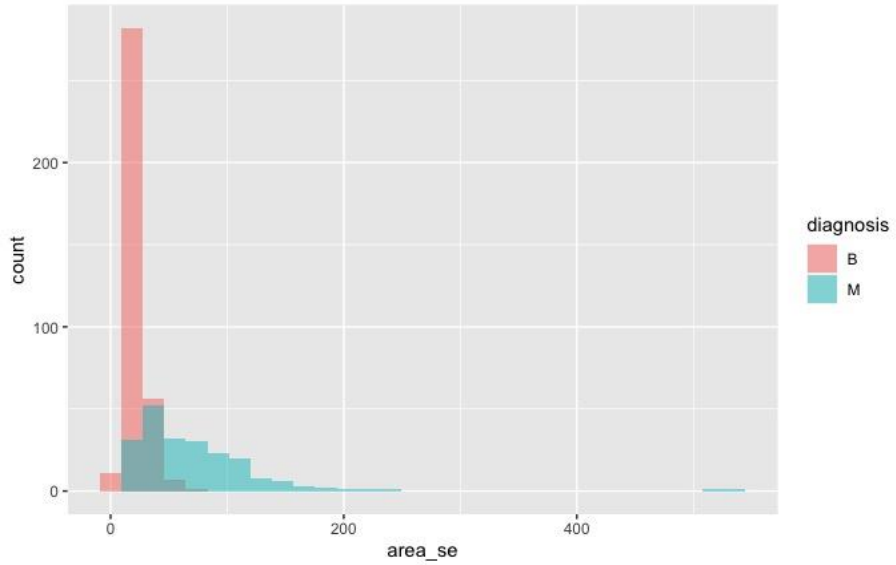
14.Τυπική απόκλιση όγκου (area_se)

Επόμενο χαρακτηριστικό του κυτταρικού πυρήνα είναι αυτό που αφορά την τυπικής απόκλισης του μεγέθους του κυτταρικού πυρήνα. Στον **πίνακα 3.18**, αναγράφονται τα κυριότερα περιγραφικά μέτρα του χαρακτηριστικού. Βάσει αυτού λοιπόν προκύπτει πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης του μεγέθους τους ίση με 6.8 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 13.99. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 15.26 ενώ 35.76 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση του μεγέθους τους τουλάχιστον ίση με 19.63 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 58.45. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης του μεγέθους τους ίση με 21.14 και 72.67 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης μεγέθους ίση με τουλάχιστον 25.03 και 94 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 77.11 για τους καλοήθεις όγκους και 61.36 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 70.31 ενώ πολύ μεγαλύτερο το οποίο ισούται με 528.21 κατανέμονται οι κακοήθεις όγκοι.

area_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	6.80	15.26	19.63	21.14	25.03	77.11	8.84	70.31
Malignant	13.99	35.76	58.45	72.67	94.00	542.20	61.36	528.21

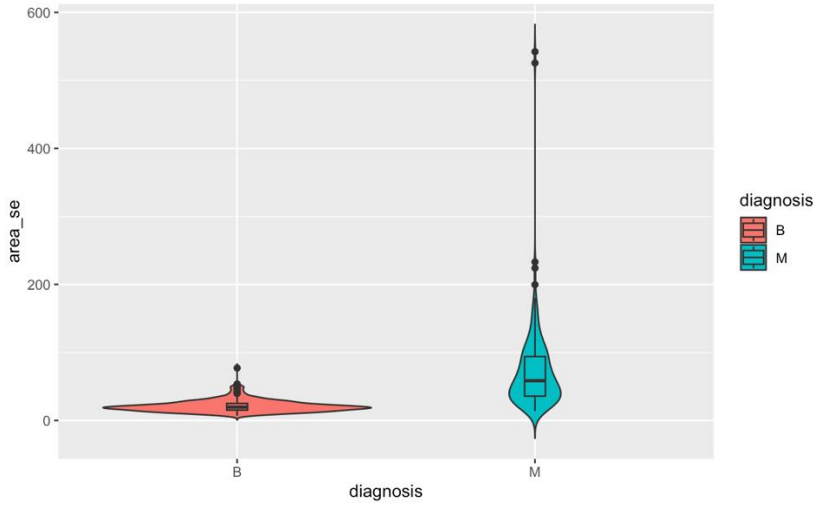
Πίνακας 3. 18 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης του μεγέθους*

Από το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’area_se’’ είναι ευδιάκριτο το πως κατανέμονται οι παρατηρήσεις του χαρακτηριστικού για κάθε κλάση. Ειδικότερα, φαίνεται πως για την κλάση της καλοήθειας οι παρατηρήσεις κατανέμονται σε ιδιαίτερα ‘’στενό’’ εύρος με τις περισσότερες παρατηρήσεις να ανήκουν στην τιμή του χαρακτηριστικού περίπου με 15 ενώ για την κλάση της κακοήθειας περίπου με 35.



Εικόνα 3. 45 *Ιστόγραμμα ανεξάρτητης μεταβλητής area_se*

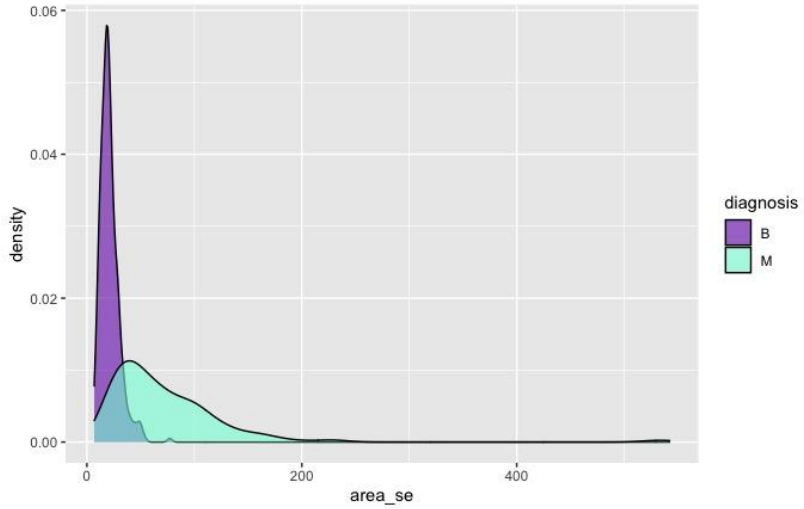
Στην **εικόνα 4.43**, παρουσιάζονται τα violin plots της ανεξάρτητης μεταβλητής ‘’area_se’’, όπου διακρίνονται οι ακραίες παρατηρήσεις της κάθε κλάσης με αυτές τις καλοήθειας να κυμαίνονται γύρω από την τιμή της τυπικής απόκλισης ίση με περίπου 70. Για την κλάση της κακοήθειας παρατηρούνται τρείς ακραίες τιμές γύρω από την τιμή 210 και άλλες 2 γύρω από την τιμή 550.



Εικόνα 3. 46 *Violin plots της ανεξάρτητης μεταβλητής area_se*

Στην εικόνα 3.44, απεικονίζεται το γράφημα πυκνότητας πιθανότητας το οποίο υποδεικνύει πως ένας όγκος με τυπική απόκλιση μεγέθους ίση με περίπου 21, παρουσιάζει τη μέγιστη

πιθανότητα να ανήκει στην κλάση της καλοήθειας ενώ για την τιμή ίση με περίπου 35 στην κακοήθεια.



Εικόνα 3. 47 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής area_se*

15.Τυπική απόκλιση απαλότητα

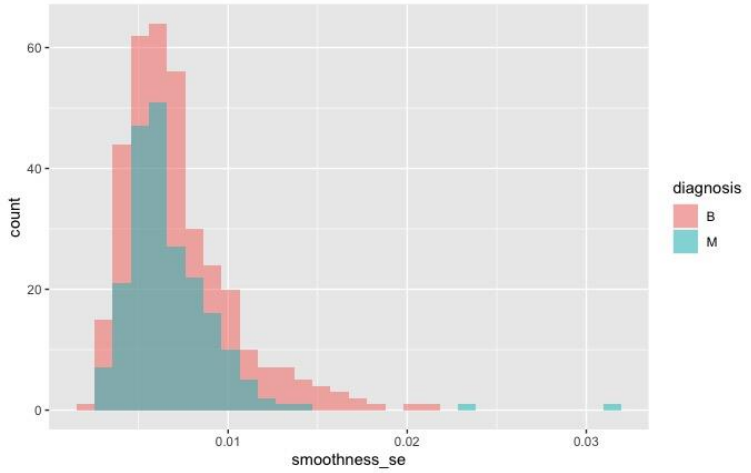
Το επόμενο χαρακτηριστικό αφορά την τυπική απόκλιση του χαρακτηριστικού απαλότητας του κυτταρικού πυρήνα βασικά περιγραφικά μέτρα του οποίου παρουσιάζονται παρακάτω στον **πίνακα 4.19**. Βάσει αυτού λοιπόν προκύπτει πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης απαλότητας ίση με 0.0017 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.0026. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.0052 ενώ 0.005 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση απαλότητας τουλάχιστον ίση με 0.0065 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.0062. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης απαλότητας ίση με 0.0071 και 0.0067 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης απαλότητας ίση με τουλάχιστον 0.0085 και 0.0079 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.021 για τους καλοήθεις όγκους και 0.031 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι

παρατηρήσεις των καλοηθών όγκων ισούται με 0.020 ενώ ελάχιστα μεγαλύτερο το οποίο ισούται με 0.028 κατανέμονται οι κακοήθεις όγκοι.

smoothness_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0017	0.0052	0.0065	0.0071	0.0085	0.021	0.0031	0.020
Malignant	0.0026	0.005	0.0062	0.0067	0.0079	0.031	0.003	0.028

Πίνακας 3. 19 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης της απαλότητας*

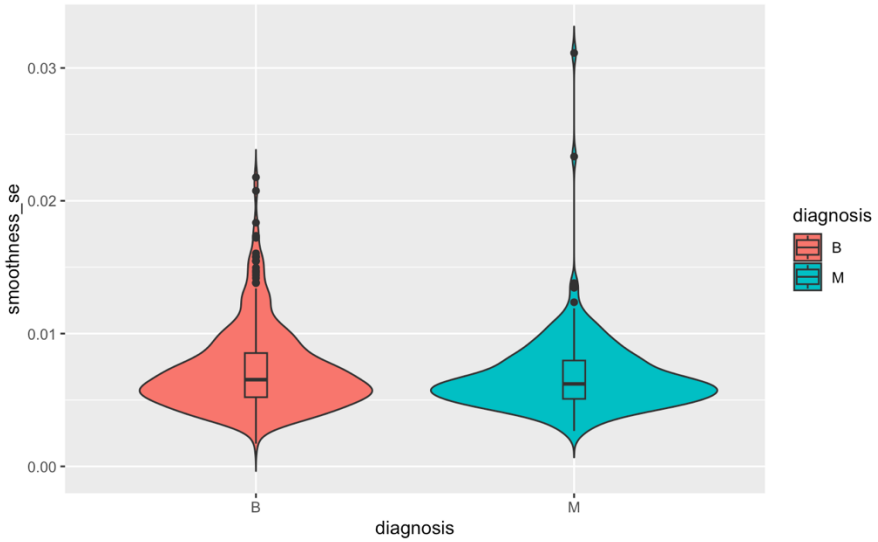
Έπειτα, στην **εικόνα 3.45**, ακολουθεί το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’smoothness_se’’ στο οποίο απεικονίζεται γραφικά το πλήθος των όγκων της κάθε κλάσης συναρτήσει κάθε τιμής του χαρακτηριστικού τυπικής αποκλίσεις απαλότητας. Ειδικότερα, οι περισσότεροι καλοήθεις όγκοι παρουσιάζουν τυπική απόκλιση απαλότητας ίση με περίπου 0.006 όπως επίσης και οι κακοήθεις με τις δύο κλάσεις να κατανέμονται ομοιόμορφα και με παρόμοια κατανομή.



Εικόνα 3. 48 *Ιστόγραμμα ανεξάρτητης μεταβλητής smoothness_se*

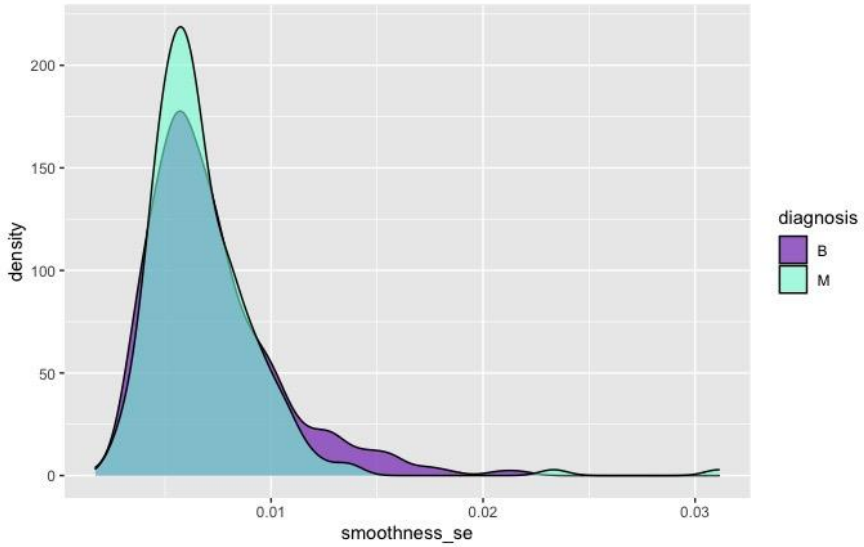
Παρακάτω στην **εικόνας 3.46** στα violin plots της ανεξάρτητης μεταβλητής ‘’smoothness_mean’’ απεικονίζονται οι κατανομές των δύο κλάσεων καθώς επίσης είναι ευδιάκριτες και οι ακραίες τους τιμές. Πιο συγκεκριμένα, οι ακραίες τιμές της καλοήθειας κυμαίνονται κυρίως γύρω από την τιμή της τυπικής απόκλισης της απαλότητας περίπου ίση με 0.005, για την κακοήθειας οι ακραίες της τιμές έχουν πολύ μεγαλύτερο εύρος έναντι της καλοήθειας με τρείς παρατηρήσεις να κυμαίνονται γύρω από την τιμή του χαρακτηριστικού

περίπου ίση με 0.013, μία γύρω από την τιμή 0.023 και μία κοντά στην τιμή 0.032. Επίσης, η διάμεσος της καλοήθειας φαίνεται σύμφωνα με τα violin plots ότι είναι περίπου στο ίδιο επίπεδο σε σχέση με αυτό της κακοήθειας.



Εικόνα 3. 49 *Violin plots της ανεξάρτητης μεταβλητής smoothness_se*

Τέλος, όσον αφορά την ανεξάρτητη μεταβλητή ‘’smoothhness_se’’, στην **εικόνα 3.47** παρουσιάζεται το γράφημα πυκνότητας πιθανότητας. Είναι προφανές και λογικό λοιπόν πως για το εύρος τιμών της τυπική απόκλισης της απαλότητας 0.003 με 0.008 είναι πολύ πιθανότερο να ανήκει στην κλάση της κακοήθειας έναντι της κακοήθειας.



Εικόνα 3. 50 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής smoothness_se*

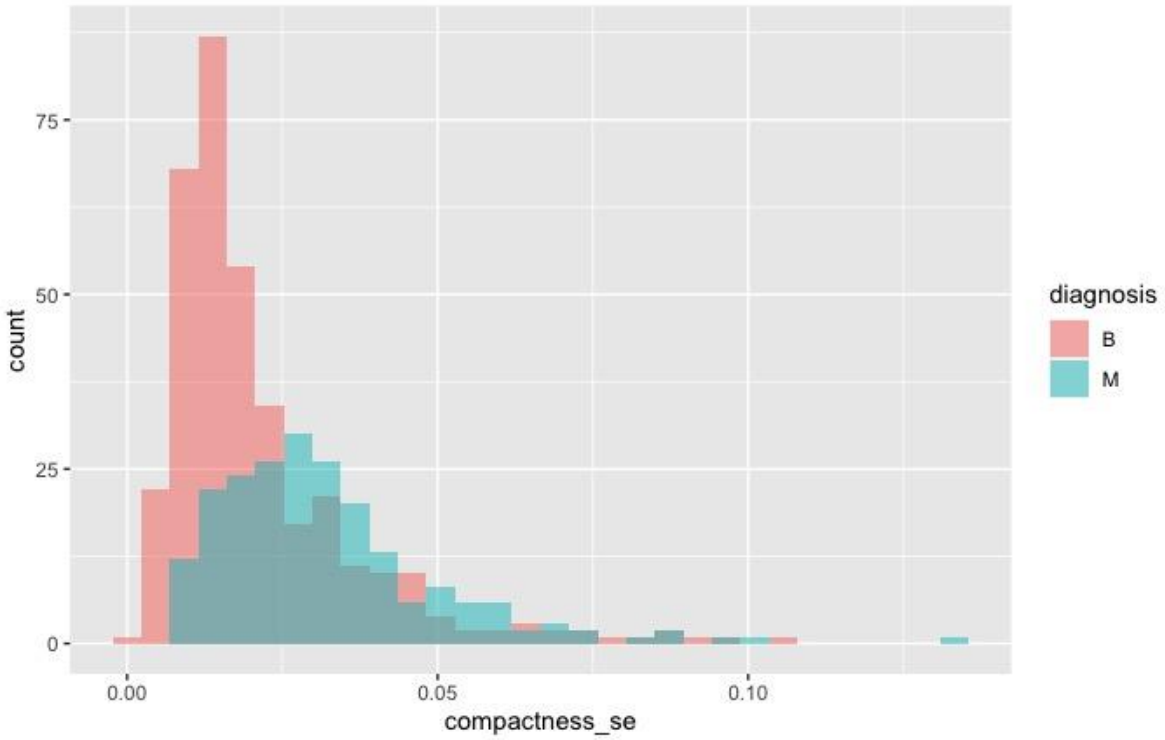
16.Τυπική απόκλιση συμπαγότητας (compactness_se)

Η επόμενη ανεξάρτητη μεταβλητή αφορά το χαρακτηριστικό της τυπικής απόκλισης της συμπαγότητας του κυτταρικού πυρήνα του οποίου τα κύρια περιγραφικά παρουσιάζονται παρακάτω στον **πίνακα 3.20.** Σύμφωνα με τον πίνακα λοιπόν προκύπτει πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης συμπαγότητας ίση με 0.002 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.008. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.011 ενώ 0.02 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση συμπαγότητας τουλάχιστον ίση με 0.016 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.029. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης συμπαγότητας ίση με 0.021 και 0.032 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης συμπαγότητας ίση με τουλάχιστον 0.026 και 0.039 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.106 για τους καλοήθεις όγκους και 0.135 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.104 ενώ ελάχιστα μεγαλύτερο το οποίο ισούται με 0.127 κατανέμονται οι κακοήθεις όγκοι.

compactness_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.002	0.011	0.016	0.021	0.026	0.106	0.016	0.104
Malignant	0.008	0.020	0.029	0.032	0.039	0.135	0.018	0.127

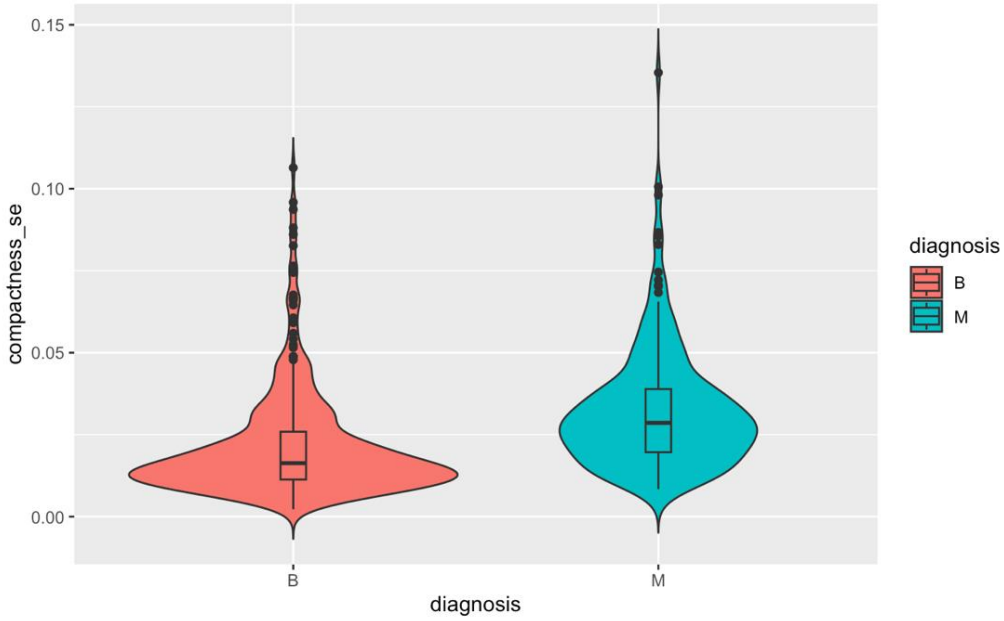
Πίνακας 3. 20 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης της συμπαγότητας*

Παρακάτω στην **εικόνα 3.48**, παρουσιάζεται το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’compactness_se’’, όπου σύμφωνα με αυτό οι περισσότερες παρατηρήσεις της καλοήθειας λαμβάνουν την τιμή της τυπικής απόκλισης του κυτταρικού πυρήνα περίπου ίση με περίπου 0.012 ενώ για την τιμή περίπου ίση με 0.027 αντιστοιχούν οι περισσότερες τιμές στην κλάση της κακοήθειας.



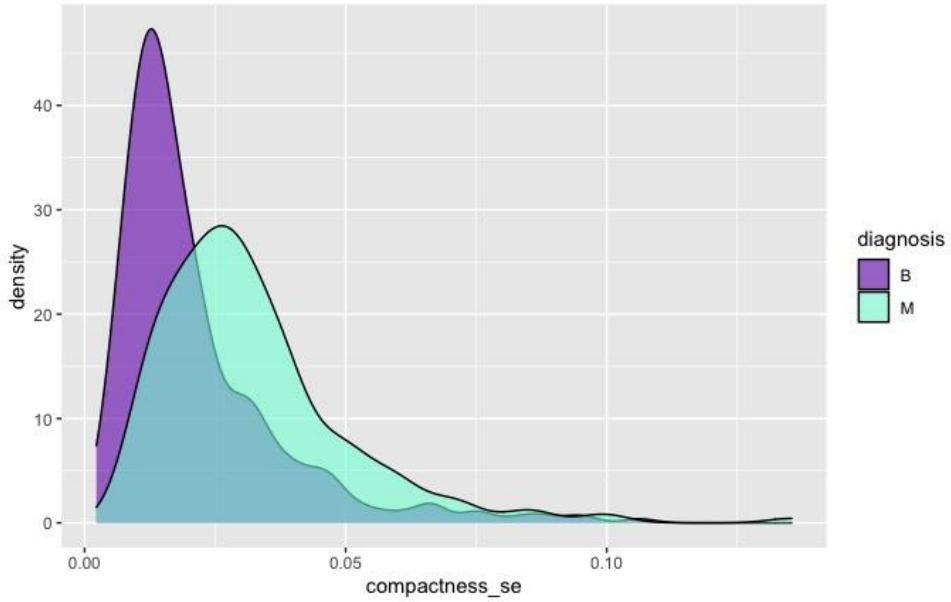
Εικόνα 3. 51 *Ιστόγραμμα ανεξάρτητης μεταβλητής compactness_se*

Από τα violin plots της ανεξάρτητης μεταβλητής ‘’compactness_se’’, όπου είναι ευδιάκριτη η κατανομή της κάθε κλάσης όπως επίσης και η πυκνότητα της, παρατηρούνται επίσης και οι ακραίες τιμές της. Για κάθε κλάση αντιστοιχούν αρκετές ακραίες τιμές συναρτήσει της τιμής του χαρακτηριστικού της τυπικής απόκλισης της συμπαγότητας με αυτές της καλοήθειας να ναι λιγότερες ως προς το πλήθος σε σχέση με αυτές της καλοήθειας όμως κατανέμονται σε μικρότερο εύρος από αυτό της κακοήθειας.



Εικόνα 3. 52 *Violin plots ανεξάρτητης μεταβλητής compactness_se*

Στο γράφημα πυκνότητας της εικόνας 3.50, απεικονίζεται γραφικά η πυκνότητα πιθανότητας του χαρακτηριστικού της τυπικής απόκλισης της συμπαγότητας όπου για την τιμή του χαρακτηριστικού ίση με περίπου 0.012 εμφανίζεται η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της καλοήθειας ενώ για την τιμή περίπου ίση με 0.025 στην κλάση της κακοήθειας.



Εικόνα 3. 53 *Γράφημα πυκνότητας πιθανότητας ανεξάρτητης μεταβλητής compactness_se*

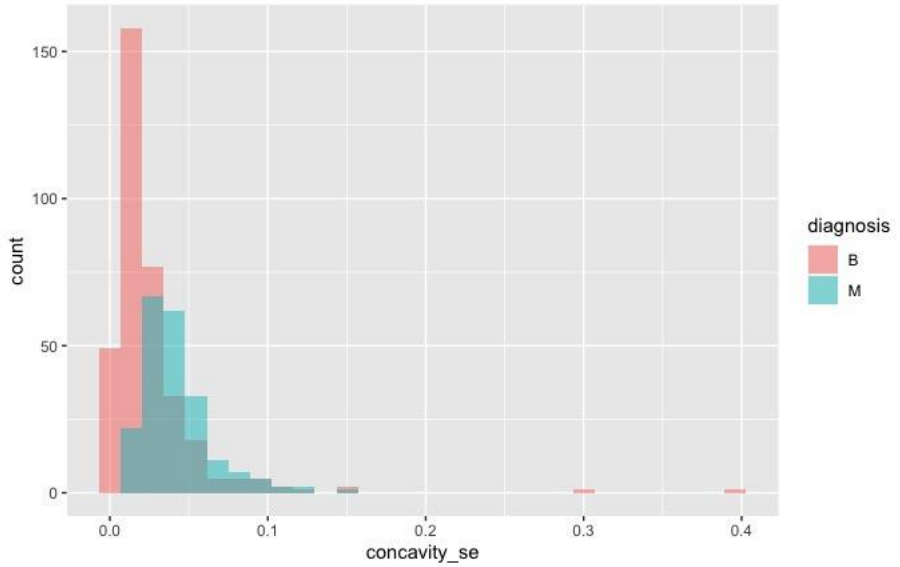
17. Τυπική απόκλιση της κοιλότητας του κυτταρικού πυρήνα (concavity_se)

Το επόμενο χαρακτηριστικό του κυτταρικού πυρήνα τα οποίο θα εξεταστεί αφορά την τυπική απόκλιση της κοιλότητας του κυτταρικού πυρήνα. Σύμφωνα με τον **πίνακα 3.21** λοιπόν προκύπτει πως οι καλοήθεις όγκοι παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης κοιλότητας ίση με 0 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.011. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.011 ενώ 0.027 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση κοιλότητας τουλάχιστον ίση με 0.018 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.037. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης κοιλότητας ίση με 0.026 και 0.042 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή τυπικής απόκλισης κοιλότητας ίση με τουλάχιστον 0.031 και 0.050 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.396 για τους καλοήθεις όγκους και 0.144 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.396 ενώ ελάχιστα μικρότερο το οποίο ισούται με 0.133 κατανέμονται οι κακοήθεις όγκοι.

concavity_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.000	0.011	0.018	0.026	0.031	0.396	0.033	0.396
Malignant	0.011	0.027	0.037	0.042	0.050	0.144	0.022	0.133

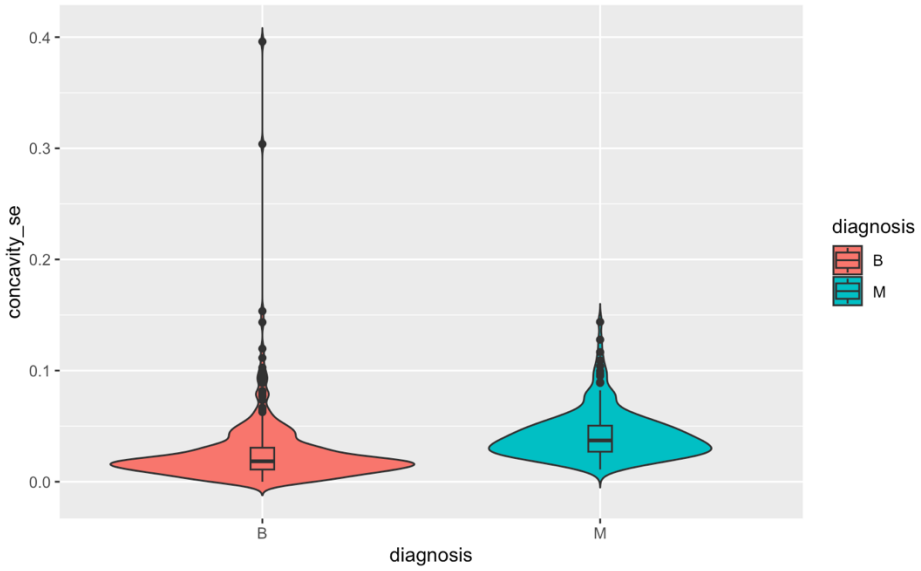
Πίνακας 3. 21 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης κοιλότητας*

Στη συνέχεια από το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’concavity_se’’ στην **εικόνα 3.51**, είναι ευδιάκριτο πως από τις μετρήσεις της κοιλότητας των όγκων οι περισσότεροι καλοήθεις παρουσιάζουν τυπική απόκλιση ίση με περίπου 0.02 ενώ οι κακοήθεις περίπου ίση με 0.03.



Εικόνα 3. 54 *Ιστόγραμμα ανεξάρτητης μεταβλητής concavity_se*

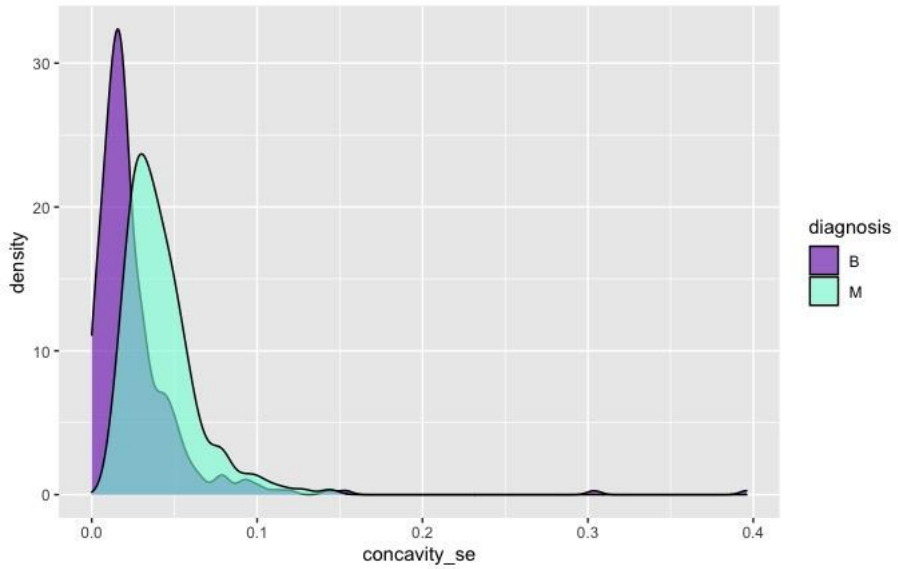
Επίσης, από το violin plot της κλάσης της καλοήθειας παρατηρείται πως οι παρατηρήσεις συναρτήσει της τυπικής απόκλισης του χαρακτηριστικού κατανέμονται σε ιδιαίτερα μικρό εύρος ενώ οι ακραίες τιμές κυμαίνονται οι περισσότερες γύρω από την τιμή 0.1 ενώ οι υπόλοιπες σε πολύ μεγαλύτερο εύρος με την πιο ακραία να βρίσκεται κοντά στην τιμή της τυπικής απόκλισης της κοιλότητας ίση με περίπου 0.4. Για την κλάση της καλοήθειας, οι ακραίες τιμές της είναι κοντά συγκεντρωμένες και κατανεμημένες γύρω από την τιμή του χαρακτηριστικού ίσο με περίπου 0.1.



Εικόνα 3. 55 *Violin plots ανεξάρτητης μεταβλητής concavity_se*

Τέλος, όσον αφορά τα περιγραφικά χαρακτηριστικά της ανεξάρτητης μεταβλητής που αφορά το χαρακτηριστικό τυπικής απόκλισης κοιλότητας του κυτταρικού πυρήνα

παρουσιάζεται το γράφημα πυκνότητας πιθανότητας της. Πιο συγκεκριμένα, ένας όγκος με τυπική απόκλιση κοιλότητας ίση με περίπου 0.2 παρουσιάζει μεγαλύτερη πιθανότητα να ανήκει στην κλάση της καλοήθειας, ενώ με τιμή χαρακτηριστικού ίσο με περίπου 0.3 παρουσιάζει μεγαλύτερη πιθανότητα να ανήκει στην κλάση της κακοήθειας.



Εικόνα 3. 56 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής concavity_se*

18. Τυπική απόκλιση πλήθους κοίλων τμημάτων (concave.points_se)

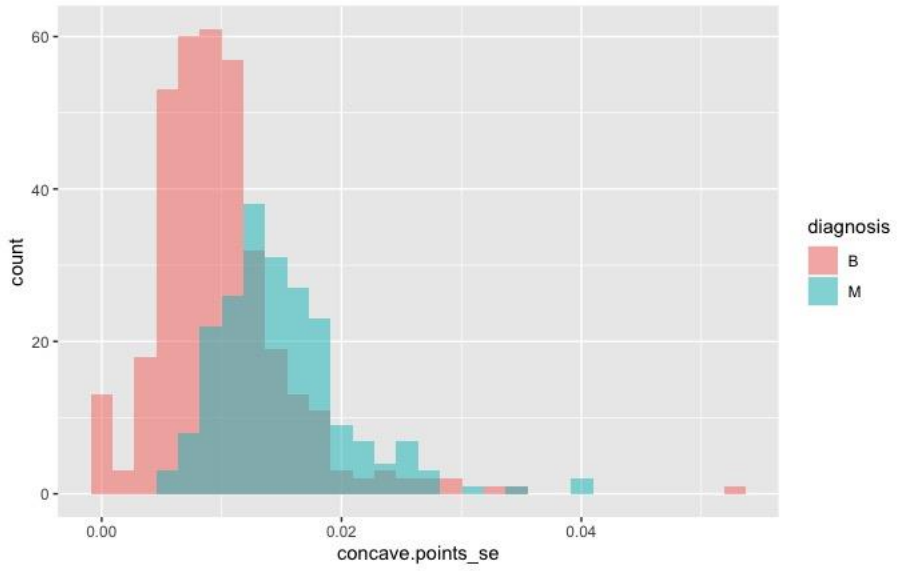
Αυτό το χαρακτηριστικό αφορά την τυπική απόκλιση του πλήθους των κοίλων τμημάτων που προκύπτουν από τις μετρήσεις μέσω των δεδομένων εικόνας. Από τον **πίνακα 3.22,** εκμαιεύονται οι πληροφορίες που αφορούν τα βασικά περιγραφικά μέτρα της εν λόγω ανεξάρτητης μεταβλητής Οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης του πλήθους κοίλων τμημάτων ίση με 0 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.0052. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.0064 ενώ 0.0114 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση του πλήθους κοίλων τμημάτων τουλάχιστον ίση με 0.0091 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.0142. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης του πλήθους κοίλων τμημάτων ίση με 0.0099 και 0.0151 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τυπικής απόκλισης του πλήθους κοίλων τμημάτων ίση με τουλάχιστον 0.0119 και 0.0175 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των

παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.528 για τους καλοήθεις όγκους και 0.0409 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.0528 ενώ ελάχιστα μικρότερο το οποίο ισούται με 0.0357 κατανέμονται οι κακοήθεις όγκοι.

concave.points_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0000	0.0064	0.0091	0.0099	0.0119	0.0528	0.0057	0.0528
Malignant	0.0052	0.0114	0.0142	0.0151	0.0175	0.0409	0.0055	0.0357

Πίνακας 3. 22 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης κοιλότητας*

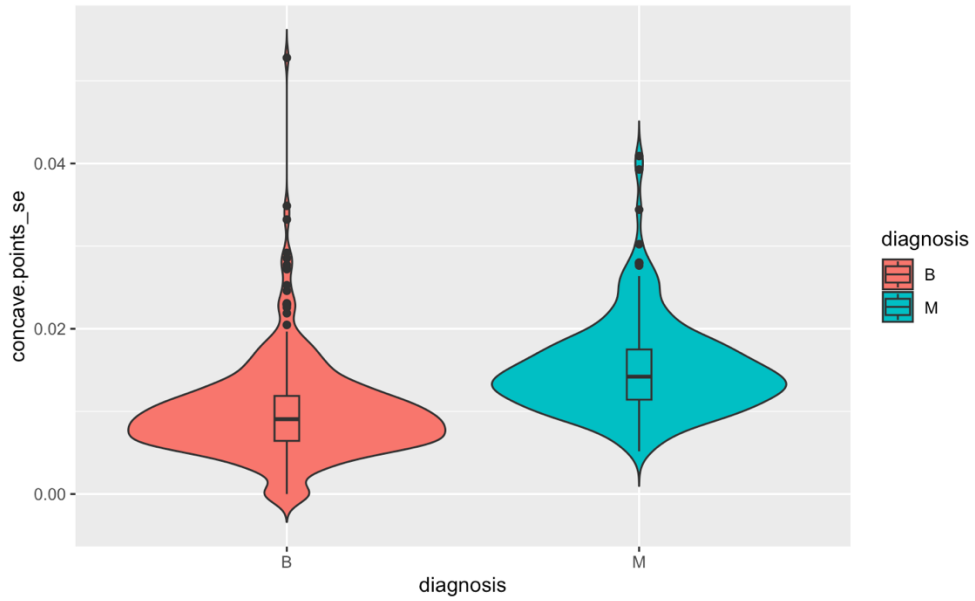
Έπειτα, σύμφωνα με το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’concave.points_se’’ που βρίσκεται στην **εικόνα 3.54**, αποδεικνύεται πως οι περισσότερες παρατηρήσεις που ανήκουν στην κλάση της καλοήθειας ανήκουν στην τιμή του χαρακτηριστικού ίση με περίπου 0.008, ενώ οι περισσότερες που ανήκουν στην κλάση της κακοήθειας αντιστοιχούν στην τιμή περίπου ίση με 0.014.



Εικόνα 3. 57 *Ιστόγραμμα ανεξάρτητης μεταβλητής concave.points_se*

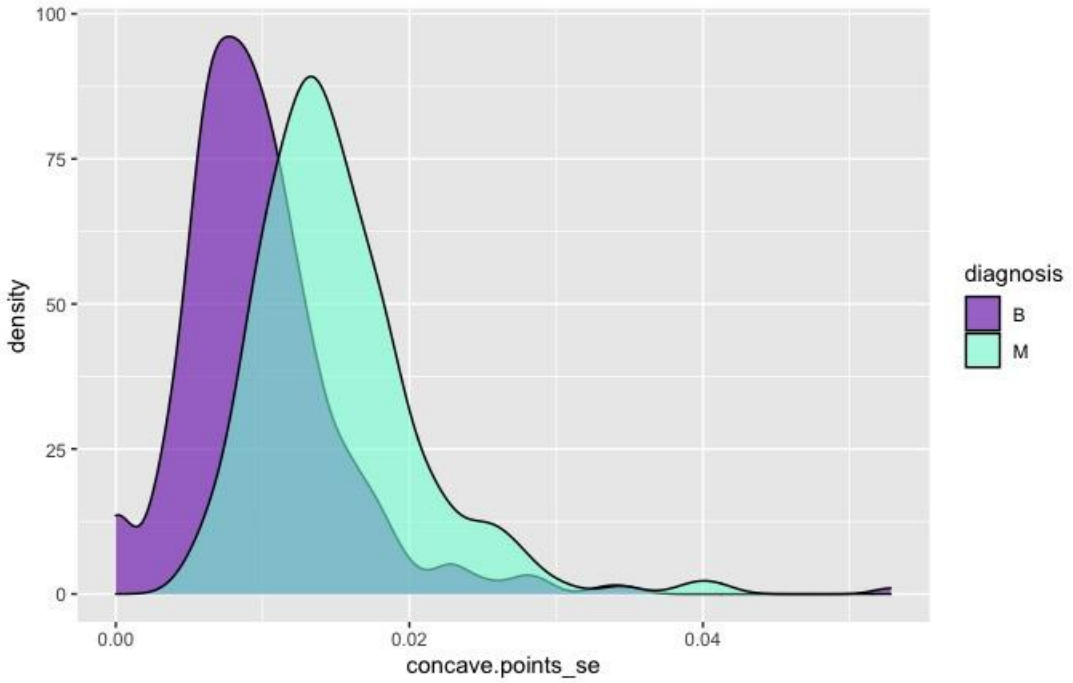
Από τα violin plots, της **εικόνας 3.55**, παρατηρούνται οι ακραίες τιμές των δύο κλάσεων με αυτές της καλοήθειας να κατανέμονται σε μεγαλύτερο εύρος απ’ ότι αυτές της κακοήθειας,

των οποίων οι περισσότερες κυμαίνονται γύρω από την τιμή 0.03 ενώ παρατηρείται μία κοντά στην τιμή 0.05. Όσον αφορά την κλάση της κακοήθειας παρατηρούνται μερικές ακραίες τιμές γύρω από την τιμή περίπου ίση με 0.03 και τις υπόλοιπες κοντά στην τιμή 0.04.



Εικόνα 3. 58 *Violin plots της ανεξάρτητης μεταβλητής concave.points_se*

Τέλος, ακολουθεί το γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής ‘’concave.points_se’’, βάσει του οποίου προκύπτει το συμπέρασμα πως για την τιμή της τυπικής απόκλισης των αριθμών κοίλων τμημάτων ίση με περίπου 0.013 παρουσιάζεται η μέγιστη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας ενώ για την τιμή περίπου ίση με 0.017 παρουσιάζεται η μέγιστη πιθανότητα να ανήκει στην κλάση της κακοήθειας.



Εικόνα 3. 59 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής concave.points_se*

19. Τυπική απόκλιση συμμετρίας

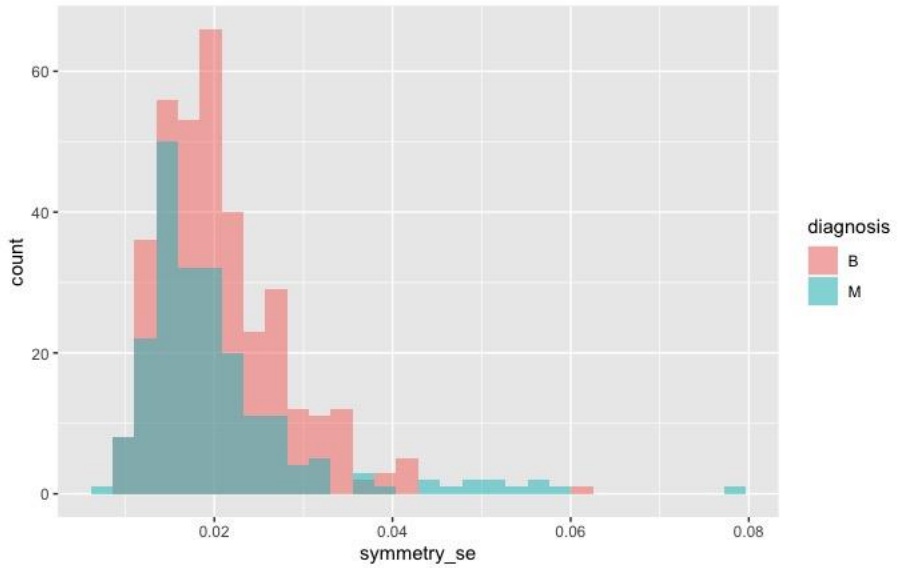
Στη συνέχεια ακολουθεί το χαρακτηριστικό που αφορά την τυπική απόκλιση της συμμετρίας του κυτταρικού πυρήνα το οποίο προκύπτει από τα δεδομένα εικόνας. Τα βασικά περιγραφικά μέτρα του εν λόγω χαρακτηριστικού παρουσιάζονται στον **πίνακα 3.23**. Οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης συμμετρίας ίση με 0.01 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.008. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.0016 ενώ 0.015 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση συμμετρίας τουλάχιστον ίση με 0.019 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.017. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης συμμετρίας ίση με 0.021 και 0.020 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τυπικής απόκλισης συμμετρίας ίση με τουλάχιστον 0.0024 και 0.022 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.062 για τους καλοήθεις όγκους και 0.079 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το

εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.052 ενώ ελάχιστα μεγαλύτερο το οποίο ισούται με 0.071 κατανέμονται οι κακοήθεις όγκοι

symmetry_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.010	0.016	0.019	0.021	0.024	0.062	0.007	0.052
Malignant	0.008	0.015	0.017	0.020	0.022	0.079	0.010	0.071

Πίνακας 3. 23 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης συμμετρίας*

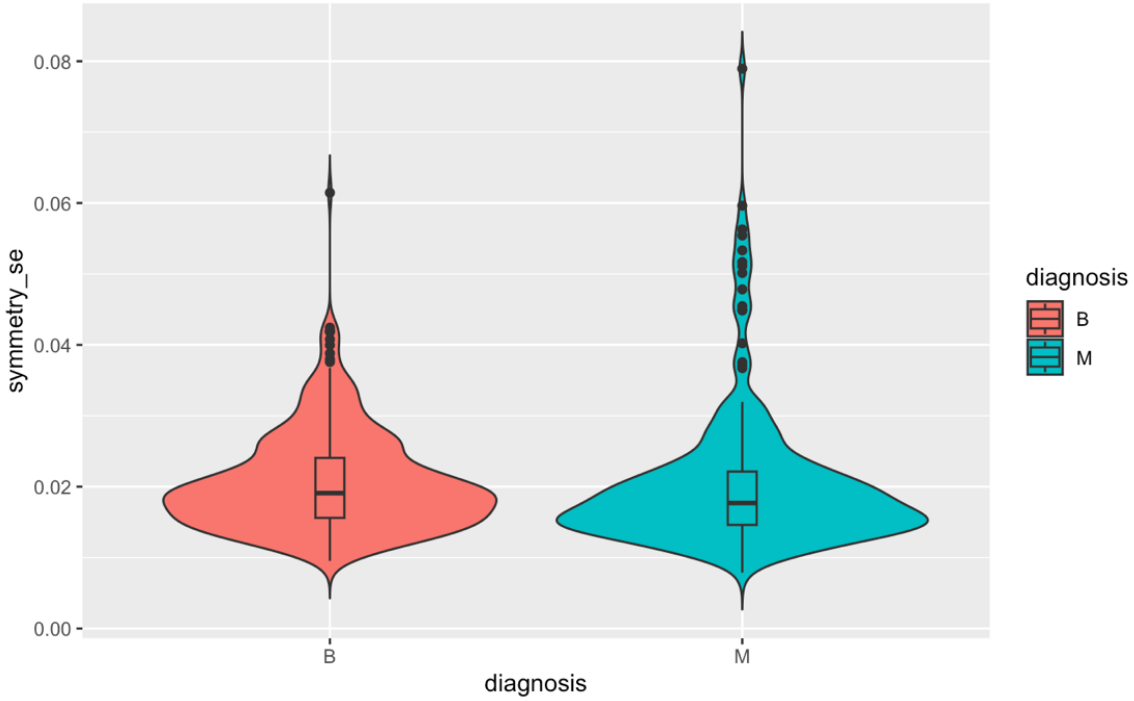
Όσον αφορά το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’symmetry_se’’, το οποίο βρίσκεται παρακάτω στην **εικόνα 3.57**, παρέχει την πληροφορία η οποία υποδεικνύει πως οι περισσότεροι καλοήθεις όγκοι λαμβάνουν την τιμή περίπου ίση με 0.02 , ενώ οι περισσότεροι κακοήθεις όγκοι λαμβάνουν την τιμή 0.015 .



Εικόνα 3. 60 *Ιστόγραμμα ανεξάρτητης μεταβλητής symmetry_se*

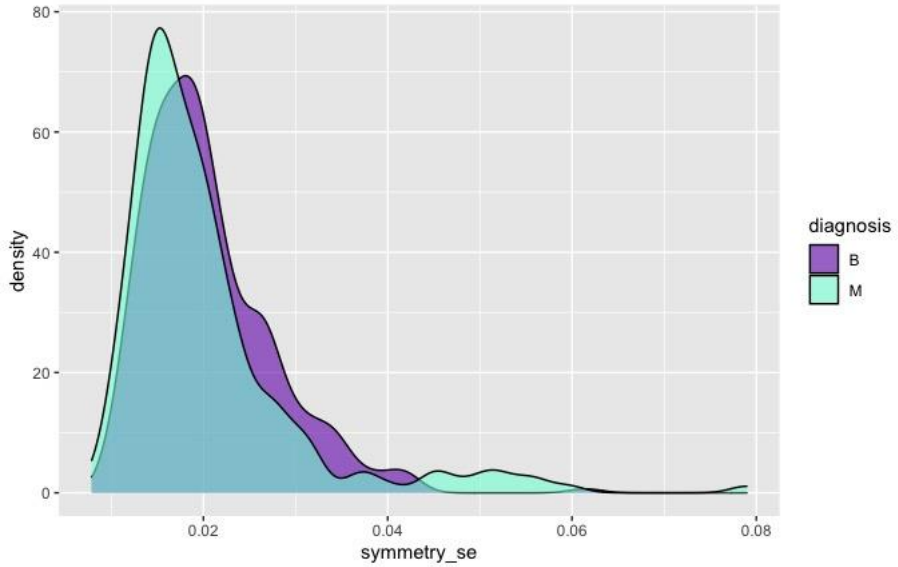
Από τα violin plots (**εικόνα 3.58)** της ανεξάρτητης μεταβλητής ‘’symmetry_se’’, τα οποία προηγούνται του ιστογράμματος, είναι ευδιάκριτα τα τεταρτημόρια σύμφωνα με τα οποία κατανέμονται οι παρατηρήσεις της κάθε κλάσης όπως επίσης και οι ακραίες του τιμές. Πιο συγκεκριμένα η κλάση της καλοήθειας αρκετές ακραίες τιμές οι οποίες κυμαίνονται γύρω από την τιμή της τυπικής απόκλισης της συμμετρίας ίση με 0.4 ενώ μια ακόμα μεμονωμένη ακραία παρατήρηση βρίσκεται κοντά στην τιμή 0.06. Παρατηρώντας την τιμή της κακοήθειας

φαίνονται αρκετές ακραίες τιμές με ιδιαίτερα μεγάλο εύρος το οποίο κυμαίνεται στο διάστημα (0.028, 0.08).



Εικόνα 3. 61 *Violin plots της ανεξάρτητης μεταβλητής symmetry_se*

Έπειτα ακολουθεί το γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής ‘’symmetry_se’’ (**εικόνα 3.59)** υποδεικνύοντας πως για την τιμή του χαρακτηριστικού της τυπικής απόκλισης της συμμετρίας του κυτταρικού πυρήνα περίπου ίση με 0.15, είναι πιθανότερο ένας όγκος να ανήκει στην κλάση της καλοήθειας, ενώ για τιμή περίπου ίση με 0.18, να ανήκει στην κλάση της κακοήθειας.



Εικόνα 3. 62 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής symmetry_se*

20.Τυπική απόκλιση μορφοκλασματικής διάστασης (fractal_dimension_se)

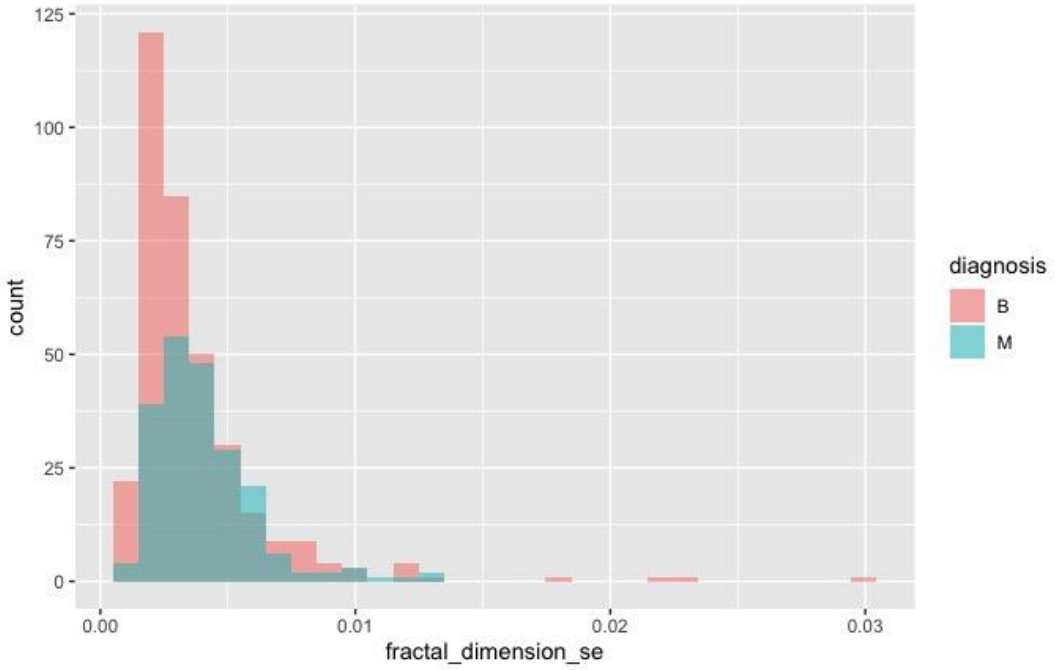
Τελευταίο χαρακτηριστικό που αφορά τις τυπικές αποκλίσεις των μετρήσεων του κυτταρικού πυρήνα μέσω δεδομένων εικόνας αποτελεί αυτό της τυπικής απόκλισης της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα. Όπως προκύπτει και από τον **πίνακα 3.24**, οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή τυπικής απόκλισης μορφοκλασματικής διάστασης ίση με 0.0009 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.0011. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.00210 ενώ 0.0027 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση μορφοκλασματικής διάστασης τουλάχιστον ίση με 0.0028 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.00370. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού τυπικής απόκλισης μορφοκλασματικής διάστασης ίση με 0.0036 και 0.0041 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τυπικής απόκλισης μορφοκλασματικής διάστασης ίση με τουλάχιστον 0.0042 και 0.0049 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.0298 για τους καλοήθεις όγκους και 0.0128 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο

κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.028 ενώ ελάχιστα μικρότερο το οποίο ισούται με 0.011 κατανέμονται οι κακοήθεις όγκοι.

fractal_dimension_se	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.00090	0.00210	0.00280	0.00360	0.00420	0.02980	0.00294	0.02895
Malignant	0.00110	0.00270	0.00370	0.00410	0.00490	0.01280	0.00204	0.01175

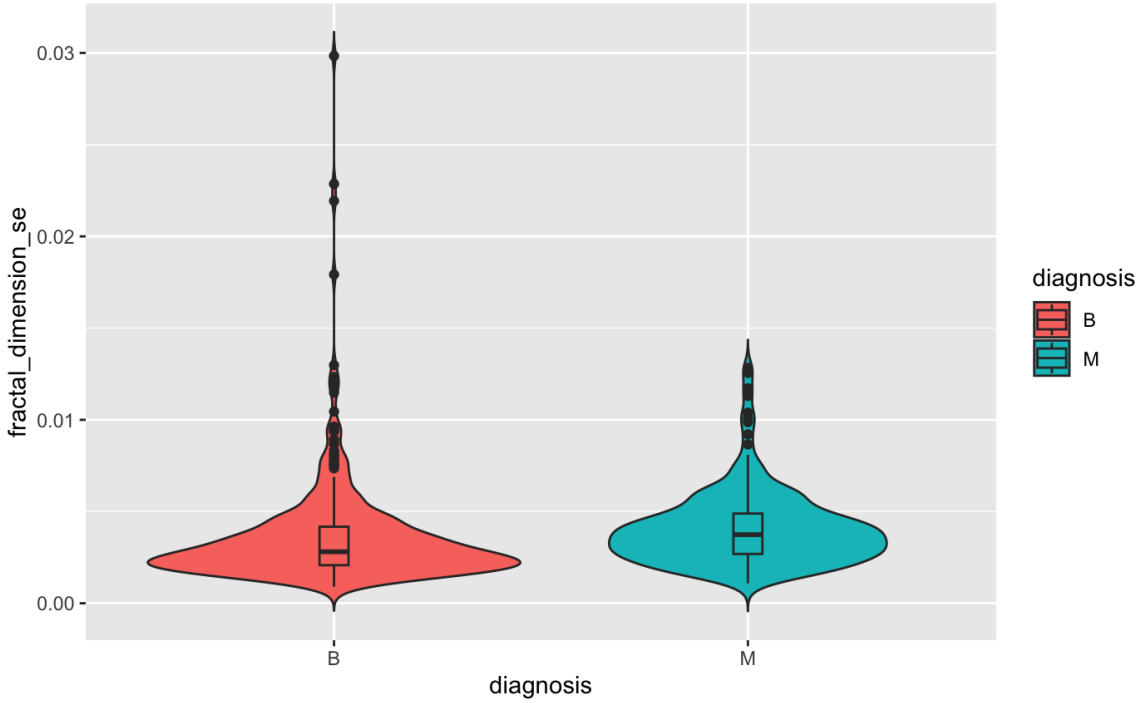
Πίνακας 3. 24 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού τυπικής απόκλισης μορφοκλασματικής διάστασης*

Παρατηρώντας το ιστόγραμμα της ανεξάρτητης μεταβλητής fractal_dimension_se, είναι αντιληπτό πως το μεγαλύτερο ποσοστό των καλοηθών όγκων παρουσιάζουν τυπική απόκλιση μορφοκλασματικής διάστασης περίπου ίση με 0.002 ενώ το μεγαλύτερο ποσοστό των κακοηθών όγκων παρουσιάζουν τιμή περίπου ίση με 0.003.



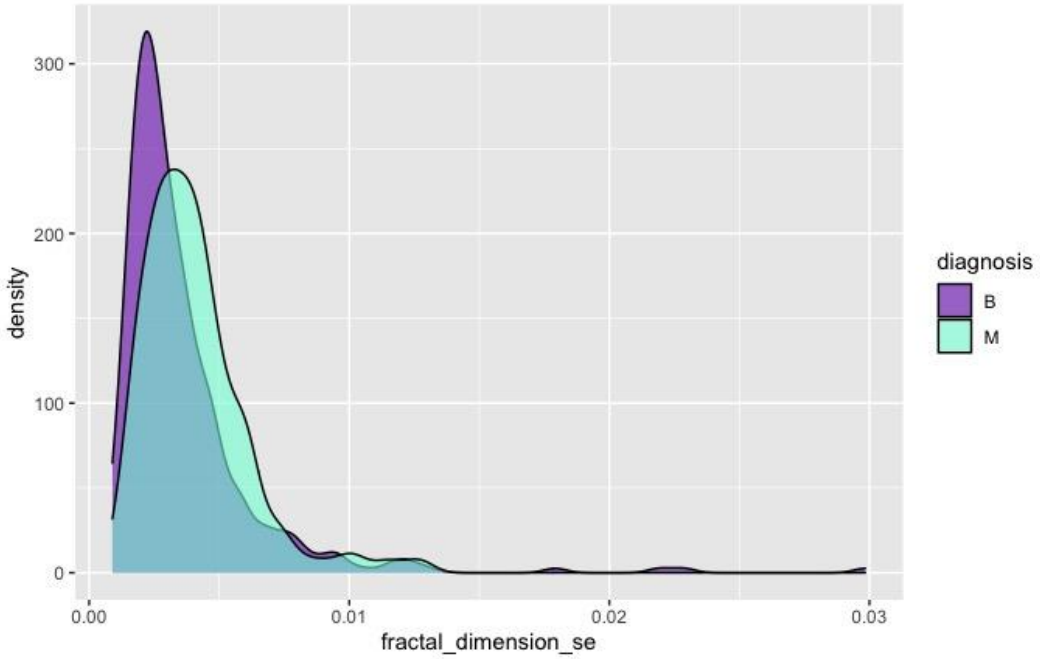
Εικόνα 3. 63 *Ιστόγραμμα ανεξάρτητης μεταβλητής fractal_dimension_se*

Επίσης από τα violin plots της **εικόνας 3.61**, γίνεται αντιληπτό πως τόσο στην κλάση της καλοήθειας όσο και στην κλάση της κακοήθειας εμπεριέχονται ακραίες τιμές. Ειδικότερα οι ακραίες τιμές που ανήκουν στην κλάση της καλοήθειας είναι αριθμητικά περισσότερες και κατανέμονται σε πολύ μεγαλύτερο εύρος από αυτές της κακοήθειας. Πιο συγκεκριμένα οι περισσότερες κυμαίνονται γύρω από την τιμή της τυπικής απόκλισης της μορφοκλασματικής διάστασης ίση με 0.01 ενώ αυτή η οποία αποτελεί και την πιο ‘’απομακρυσμένη’’ λαμβάνει την τιμή 0.03. Οι ακραίες τιμές της κακοήθειας είναι πιο ‘’στενά’’ κατανεμημένες και κυμαίνονται γύρω από την τιμή του χαρακτηριστικού ίση με 0.01.



Εικόνα 3. 64 *Violin plots ανεξάρτητης μεταβλητής fractal_dimension_se*

Τέλος, όσον αφορά το χαρακτηριστικό της τυπικής απόκλισης της μορφοκλασματικής διάστασης, εκμαιεύεται η πληροφορία πως για την τιμή του χαρακτηριστικού ίση με περίπου 0.002 παρουσιάζεται η μέγιστη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας, ενώ για την τιμή περίπου ίση με 0.003 ο όγκος είναι πιθανότερο να είναι κακοήθης.



Εικόνα 3. 65 *Γράφημα πυκνότητας πιθανότητας ανεξάρτητης μεταβλητής fractal_dimension_se*

3.4.2 Μεταβλητές διακύμανσης παραγόντων

Η τελευταία κατηγορία στην οποία ανήκουν τα χαρακτηριστικά του κυτταρικού πυρήνα αφορά τις ‘’χειρότερες’’ τιμές ή αλλιώς διακυμάνσεις που προκύπτουν από τις μετρήσεις μέσα από τα δεδομένα εικόνας.

21. Διακύμανση ακτίνας.

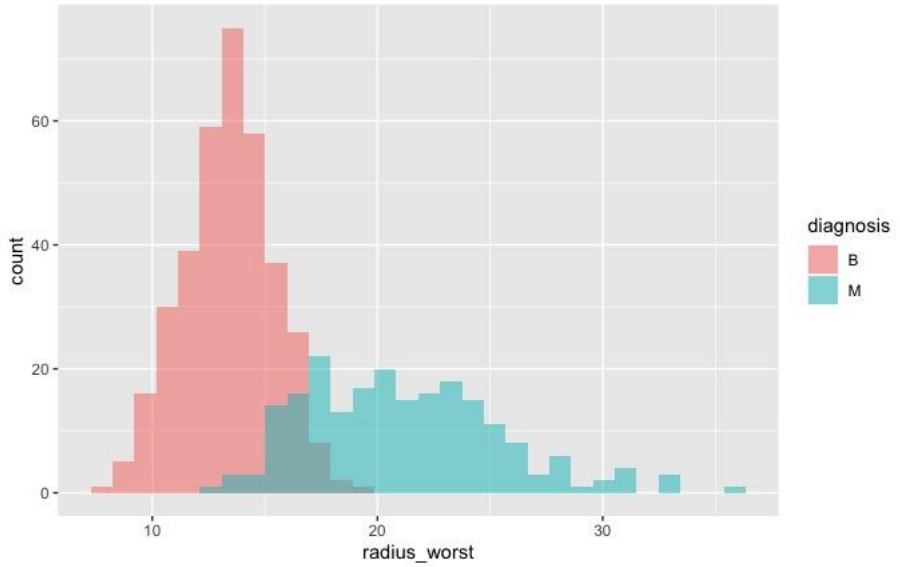
Το εν λόγω χαρακτηριστικό αφορά την διακύμανση που προκύπτει μετρώντας την απόσταση της ακτίνας του πυρήνα σε σημεία της περιμέτρου του. Αναλυτικότερα, η ελάχιστη τιμή του χαρακτηριστικού που παρατηρείται όσον αφορά την κλάση της καλοήθειας ισούται με 7.93, Όπως προκύπτει και από τον **πίνακα 3.25**, οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης της ακτίνας ίση με 7.93 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 12.84. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 12.80 ενώ 17.73 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση ακτίνας τουλάχιστον ίση με 13.35 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 20.59. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης της ακτίνας ίση με 13.38 και 21.13 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των

καλοηθών όγκων λαμβάνει τιμή διακύμανσης της ακτίνας ίση με τουλάχιστον 14.80 και 23.81 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 19.82 για τους καλοήθεις όγκους και 36.04 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 11.89 ενώ σε αρκετά μεγαλύτερο το οποίο ισούται με 23.2 κατανέμονται οι κακοήθεις όγκοι.

radius_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	7.93	12.08	13.35	13.38	14.80	19.82	1.98	11.89
Malignant	12.84	17.73	20.59	21.13	23.81	36.04	4.28	23.20

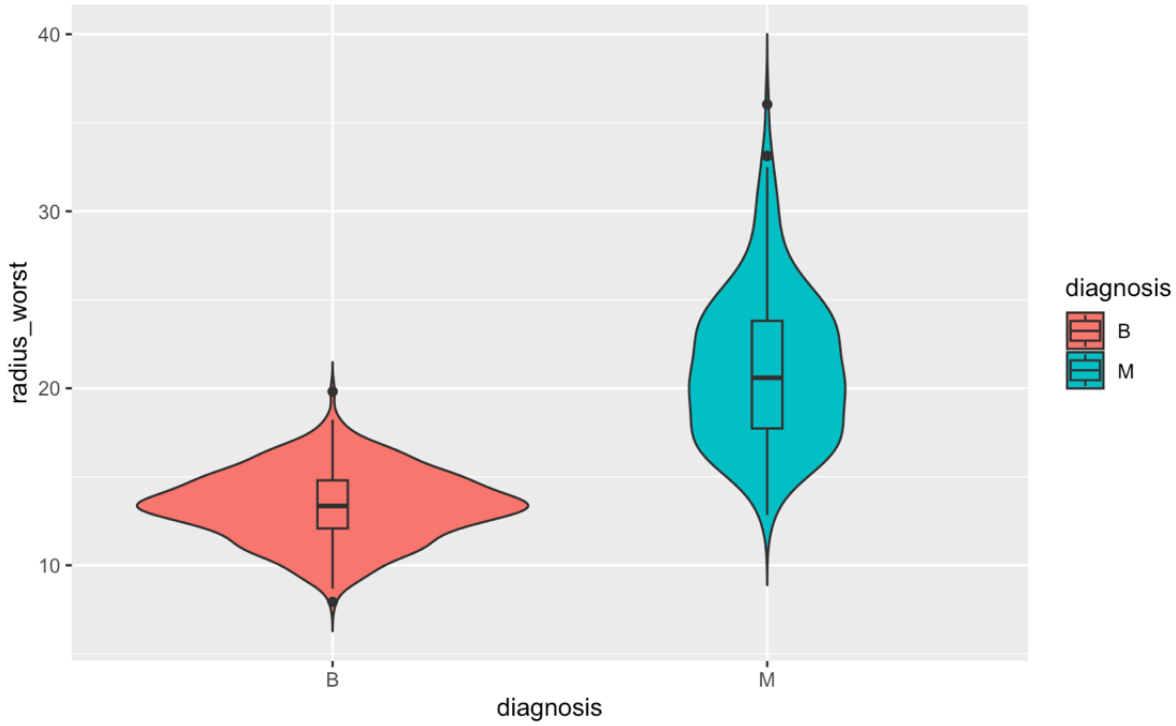
Πίνακας 3. 25 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης ακτίνας*

Έπειτα από το ιστόγραμμα της ανεξάρτητης μεταβλητή ‘’radius_worst’’ το οποίο παρέχει την πληροφορία της αντιστοιχίας του πλήθους των παρατηρήσεων για κάθε κλάση με την εκάστοτε τιμή του χαρακτηριστικού της διακύμανσης της ακτίνας. Πιο συγκεκριμένα το μεγαλύτερο ποσοστό των παρατηρήσεων όσον αφορά την κλάση της καλοήθειας αντιστοιχεί στην τιμή της μεταβλητής ‘’radius_worst’’ ίση με περίπου 13, ενώ το μεγαλύτερο ποσοστό των όγκων που αφορούν την κακοήθεια αντιστοιχεί στην τιμή περίπου ίση με 17.



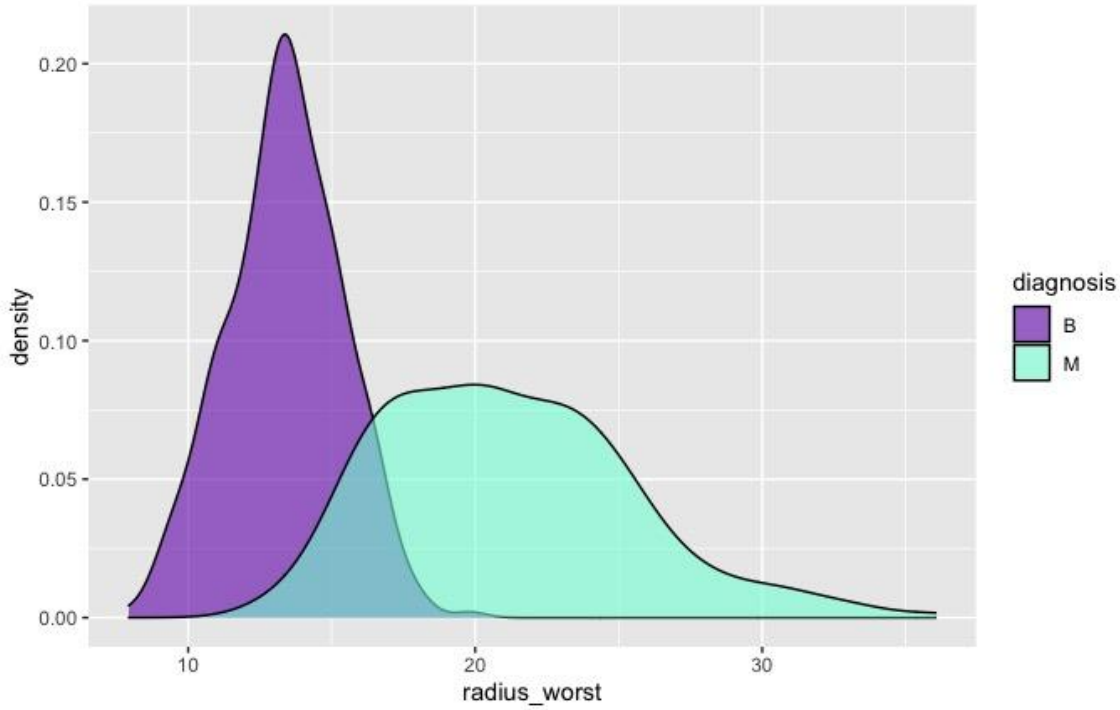
Εικόνα 3. 66 *Ιστόγραμμα ανεξάρτητης μεταβλητής radius_worst*

Στην συνέχεια παρατηρούνται οι κατανομές των παρατηρήσεων για κάθε κλάση στα violin plots της **εικόνας 3.67,** βάσει των τεταρτημόριων τους καθώς είναι εμφανείς και οι ακραίες τους τιμές. Πιο συγκεκριμένα φαίνεται πως η κλάση της καλοήθειας βρίσκεται πιο κάτω σε σχέση με την κακοήθεια παρατηρώντας δύο ακραίες τιμές κοντά στο πρώτο και τέταρτο τεταρτημόριο σε αντίθεση με την κλάση της κακοήθειας στην οποία παρατηρούνται δύο ακραίες τιμές μόνο στο τέταρτο τεταρτημόριο.



Εικόνα 3. 67 *Violin plots της ανεξάρτητης μεταβλητής radius_worst*

Τέλος στο γράφημα πυκνότητας πιθανότητας του σχήματος της **εικόνας 3.68**, πως για την τιμή διακύμανσης της ακτίνας περίπου ίση με 13 είναι πιθανότερο ένας όγκος να ανήκει στην κλάση της καλοήθεια όπως επίσης για την τιμή διακύμανσης της ακτίνας περίπου ίση με 20 είναι πιθανότερο ένας όγκος να ανήκει στην κλάση της κακοήθειας.



Εικόνα 3. 68 *Γράφημα πυκνότητας πιθανότητας ανεξάρτητης μεταβλητής radius_worst*

22.Διακύμανση υφής (texture_worst)

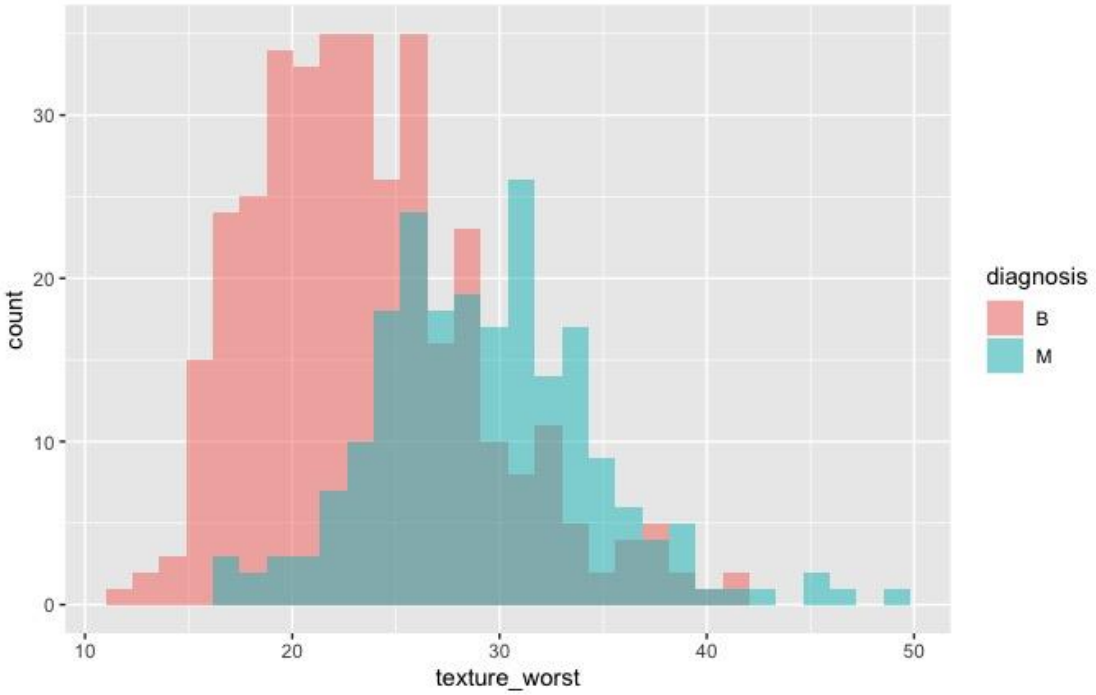
Η επόμενη ανεξάρτητη μεταβλητή, της οποίας τα περιγραφικά μέτρα θα αναλυθούν, είναι αυτή που αφορά τη διακύμανση του χαρακτηριστικού της υφή που προκύπτει από τις μετρήσεις. Τα βασικά περιγραφικά μέτρα της εν λόγω ανεξάρτητης μεταβλητής παρουσιάζονται στον **πίνακα 3.26**. Πιο συγκεκριμένα οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης της υφής ίση με 12.02 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 16.67. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 19.58 ενώ 25.78 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση υφής τουλάχιστον ίση με 22.82 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 28.95. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης της υφής ίση με 23.52 και 29.32 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης της υφής ίση με τουλάχιστον 26.51 και 32.69 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 41.78 για τους καλοήθεις όγκους και 49.54 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και

η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 29.76 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 32.87 κατανέμονται οι κακοήθεις όγκοι.

texture_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	12.02	19.58	22.82	23.52	26.51	41.78	5.49	29.76
Malignant	16.67	25.78	28.95	29.32	32.69	49.54	5.43	32.87

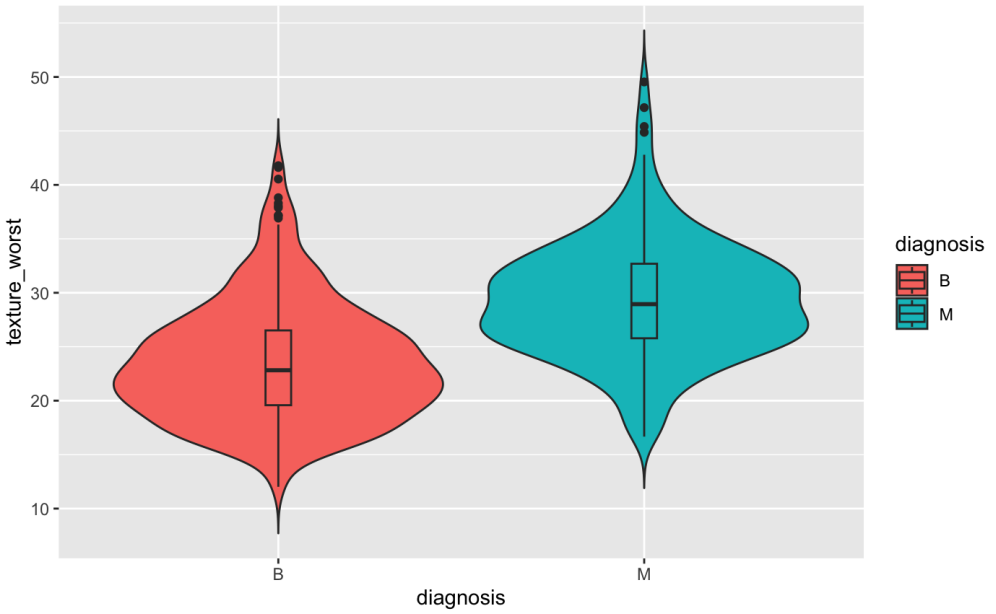
Πίνακας 3. 26 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της υφής*

Έπειτα, στην **εικόνα 3.69**, παρουσιάζεται το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’texture_worst’’, όπου σύμφωνα με αυτό παρατηρείται πως το μεγαλύτερο ποσοστό των καλοηθών όγκων κατανέμονται στο εύρος της τιμή του χαρακτηριστικού (19, 23), ενώ το μεγαλύτερο ποσοστό των κακοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού περίπου ίση με 31 .



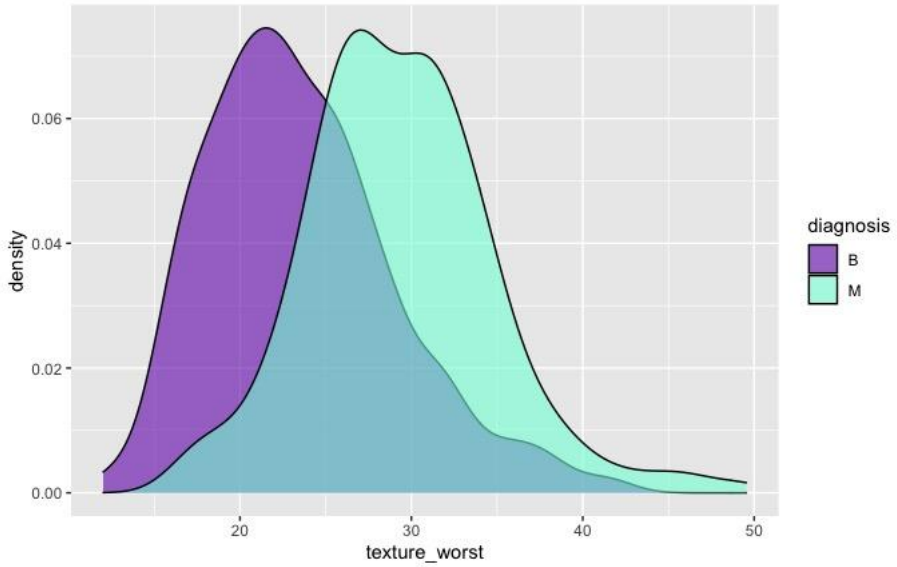
Εικόνα 3. 69 *Ιστόγραμμα ανεξάρτητης μεταβλητής texture_worst*

Παρακάτω, στην **εικόνα 3.69**, προηγούνται τα violin plots των δύο κλάσεων της ανεξάρτητης μεταβλητής ‘’texture_worst’’, όπου βάσει αυτόν εκμαιεύονται οι πληροφορίες ότι οι κατανομές των παρατηρήσεων των δύο κλάσεων είναι περίπου πανομοιότυπες συναρτήσει των τεταρτημόριων, με αυτή της καλοήθειας να παρουσιάζει διάμεσο και κατ’ επέκτασιν παρατηρήσεις που λαμβάνουν κατά κύριο λόγο μικρότερες σε σχέση με αυτές τις κακοήθειας. Επίσης, το πλήθος των ακραίων τιμών είναι περίπου ίσο με αυτές τις καλοήθειας να κυμαίνονται γύρω από την τιμή του χαρακτηριστικού διακύμανσης της υφής περίπου ίση με 40, ενώ γύρω από την τιμή περίπου ίση με 45 κυμαίνονται οι ακραίες τιμές της κλάσης της κακοήθειας.



Εικόνα 3. 70 *Violon plots της ανεξάρτητης μεταβλητής texture_worst*

Τέλος, όσον αφορά την ανεξάρτητη μεταβλητή ‘’texture_worst’’, σειρά έχει το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.71**, όπου είναι εμφανές η αντιστοιχία των τιμών του χαρακτηριστικού διακύμανσης της υφής συναρτήσει των πιθανοτήτων να αντιστοιχεί στην εκάστοτε κλάση. Πιο συγκεκριμένα, για την τιμή του χαρακτηριστικού περίπου ίση με 22 , παρουσιάζεται η μέγιστη πιθανότητα να αντιστοιχεί στην κλάση της καλοήθειας, ενώ για την τιμή του χαρακτηριστικού περίπου ίση με 27, παρουσιάζεται η μέγιστη πιθανότητα να αντιστοιχεί ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 71 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής "texture_worst"*

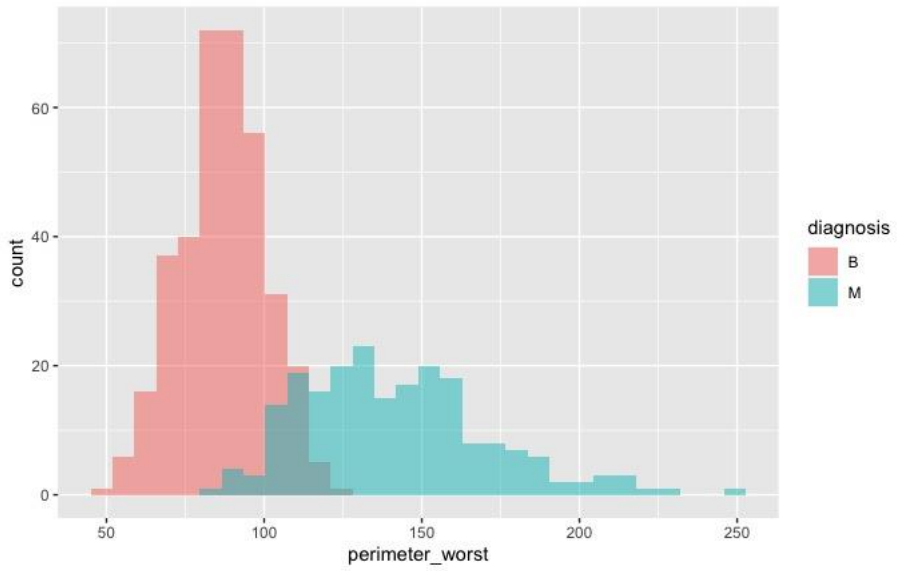
23. Διακύμανση περιμέτρου (perimeter_worst)

Το επόμενο χαρακτηριστικό αφορά τη διακύμανση της περιμέτρου των όγκων που προκύπτει από τις μετρήσεις μέσω των δεδομένων εικόνας. Πιο συγκεκριμένα, τα βασικά περιγραφικά μέτρα της συγκεκριμένης ανεξάρτητης μεταβλητής αναγράφονται στον **πίνακα 3.27**, όπου βάσει αυτού προκύπτει ότι οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης της περιμέτρου ίση με 5.41 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 85.1. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 78.27 ενώ 119.3 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση υφής τουλάχιστον ίση με 86.92 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 138. Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης της περιμέτρου ίση με 87.01 και 141.4 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης της περιμέτρου ίση με τουλάχιστον 96.59 και 159.8 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 127.1 για τους καλοήθεις όγκους και 251.20 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 76.69 ενώ σε πολύ μεγαλύτερο το οποίο ισούται με 166.1 κατανέμονται οι κακοήθεις όγκοι.

texture_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	50.41	78.27	86.92	87.01	96.59	127.10	13.53	76.69
Malignant	85.10	119.30	138.00	141.40	159.80	251.20	29.46	166.10

Πίνακας 3. 27 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της περιμέτρου*

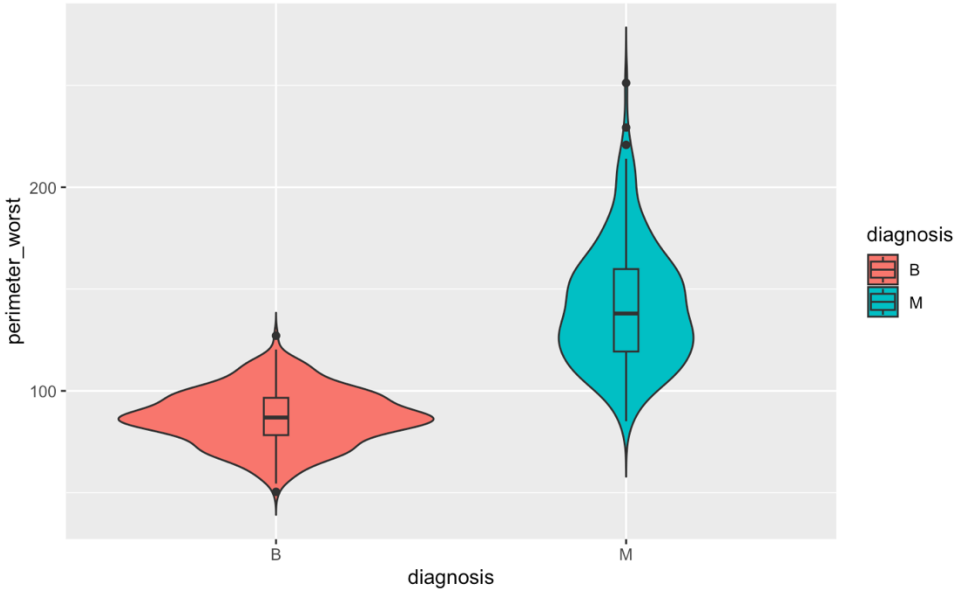
Βάσει των πληροφοριών οι οποίες παρέχονται από το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’perimeter_worst’’ της **εικόνας 3.72**, προκύπτει ότι το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού που αφορά την διακύμανση της περιμέτρου του κυτταρικού πυρήνα περίπου ίση με 85, όπως επίσης το μεγαλύτερο ποσοστό των κακοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού περίπου ίση με 130.



Εικόνα 3. 72 *Ιστόγραμμα της ανεξάρτητης μεταβλητής perimeter_worst*

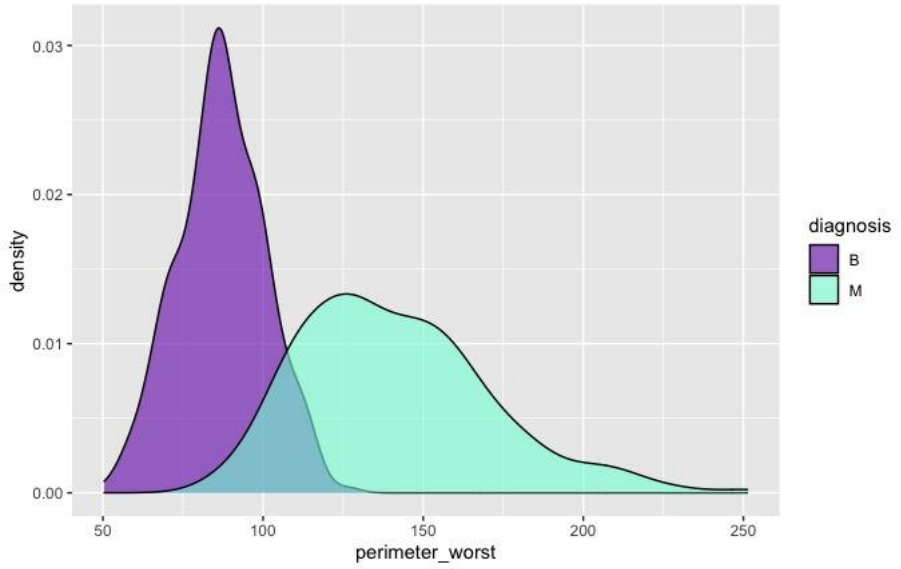
Έπειτα, από τα violin plots της **εικόνας 3.72**, φαίνεται το πως κατανέμονται οι παρατηρήσεις συναρτήσει των τιμών της διακύμανσης της περιμέτρου σε κάθε τεταρτημόριο όπως επίσης και οι ακραίες τιμές για κάθε κλάση. Γίνεται λοιπόν αντιληπτό ότι οι παρατηρήσεις της καλοήθεια παρουσιάζουν μικρότερη διακύμανση της περιμέτρου του κυτταρικού πυρήνα και ως εκ τούτου η διάμεσος της να βρίσκεται χαμηλότερα σε σχέση με αυτή της κακοήθειας. Επίσης, στην κλάση της καλοήθειας παρατηρούνται δύο ακραίες τιμές. Μία κοντά στην ελάχιστη τιμή και μία στη μέγιστη τιμή διακύμανσης της περιμέτρου του κυτταρικού πυρήνα,

ενώ στην κλάση της κακοήθειας παρατηρούνται 3 ακραίες τιμές οι οποίες κυμαίνονται στο εύρος (200, 250)



Εικόνα 3. 73 *Violin plots της ανεξάρτητης μεταβλητής perimeter_worst*

Τέλος, όσον αφορά την ανεξάρτητη μεταβλητή ‘’perimeter_worst’’, σειρά έχει το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.74**, όπου είναι εμφανές η αντιστοιχία των τιμών του χαρακτηριστικού διακύμανσης της περιμέτρου συναρτήσει των πιθανοτήτων να αντιστοιχεί στην εκάστοτε κλάση. Πιο συγκεκριμένα, για την τιμή του χαρακτηριστικού περίπου ίση με 87, παρουσιάζεται η μέγιστη πιθανότητα να αντιστοιχεί στην κλάση της καλοήθειας, ενώ για την τιμή του χαρακτηριστικού περίπου ίση με περίπου 75, παρουσιάζεται η μέγιστη πιθανότητα να αντιστοιχεί ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 74 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής perimeter_worst*

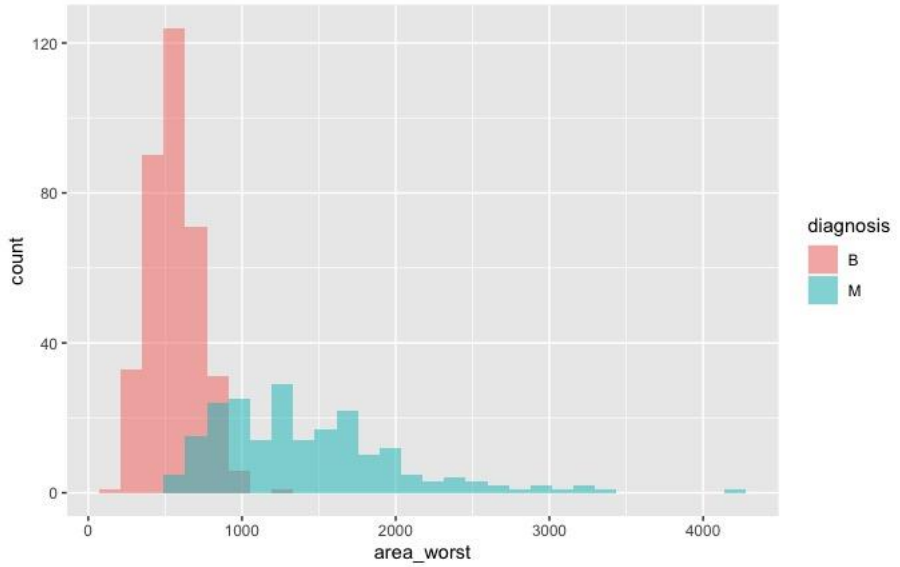
24. Διακύμανση μεγέθους του κυτταρικού πυρήνα (area_worst)

Το επόμενο χαρακτηριστικό αφορά τη διακύμανση του μεγέθους των όγκων που προκύπτει από τις μετρήσεις μέσω των δεδομένων εικόνας. Πιο συγκεκριμένα, τα βασικά περιγραφικά μέτρα της συγκεκριμένης ανεξάρτητης μεταβλητής αναγράφονται στον **πίνακα 3.28**, όπου βάσει αυτού προκύπτει ότι οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης μεγέθους ίση με 185.2 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 508.1. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 447.1 ενώ 970.3 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση μεγέθους τουλάχιστον ίση με 547.4 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 1303 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης μεγέθους ίση με 558.9 και 1422.3 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης μεγέθους ίση με τουλάχιστον 670 και 1712.8 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 1210 για τους καλοήθεις όγκους και 4254 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 1024.8 ενώ σε πολύ μεγαλύτερο το οποίο ισούται με 3745.9 κατανέμονται οι κακοήθεις όγκοι.

area_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	185.20	447.10	547.40	558.90	670.00	1210.00	163.60	1024.80
Malignant	508.10	970.30	1303.00	1422.30	1712.80	4254.00	597.97	3745.90

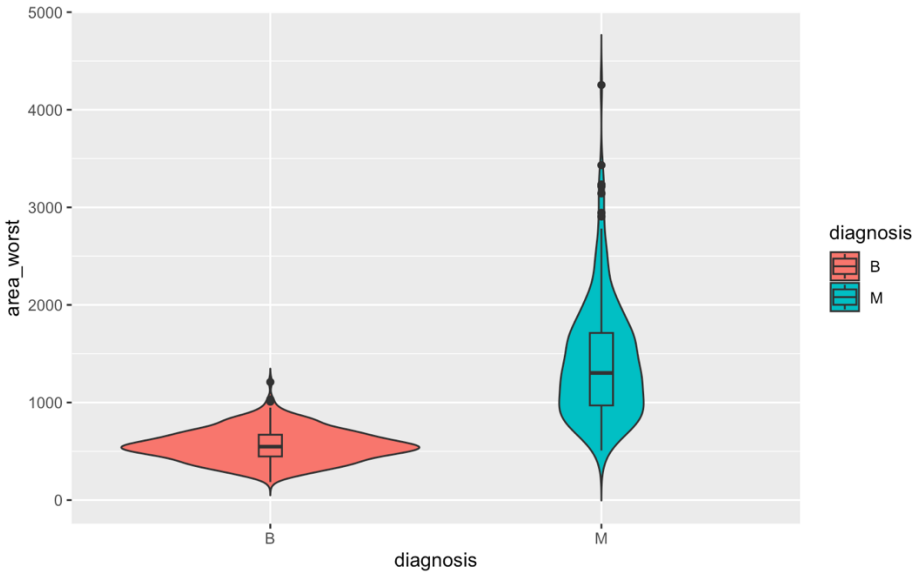
Πίνακας 3. 28 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης του μεγέθους*

Σύμφωνα με τις πληροφορίες που παρέχει το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’area_worst’’ (**εικόνα 3.75)**, το μεγαλύτερο ποσοστό των όγκων παρουσιάζουν διακύμανση μεγέθους περίπου 550, ενώ το μεγαλύτερο ποσοστό των κακοηθών όγκων παρουσιάζουν διακύμανση μεγέθους περίπου 1300.



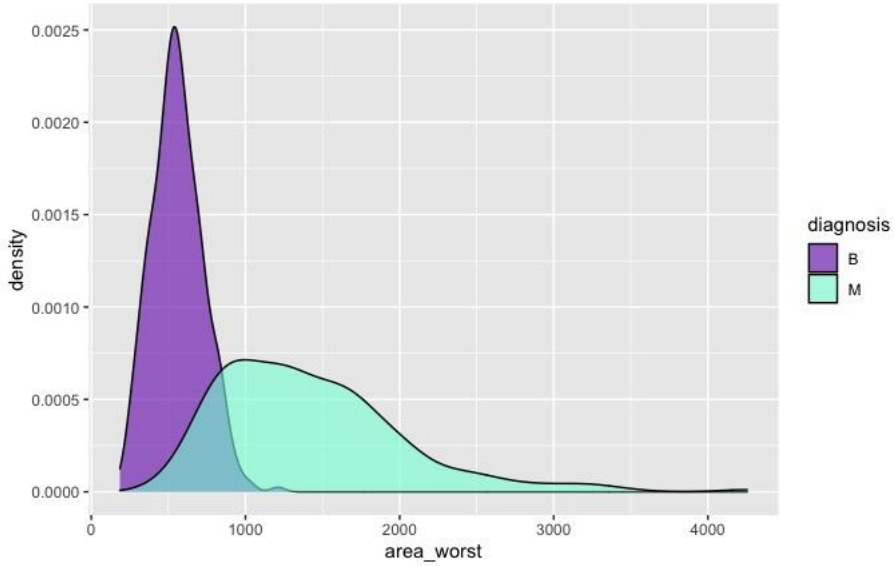
Εικόνα 3. 75 *Ιστόγραμμα της ανεξάρτητης μεταβλητής area_worst*

Επίσης τα violin plots (**εικόνα 3.76)** της ανεξάρτητης μεταβλητής υποδεικνύει πως το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων είναι πολύ μικρότερο και λαμβάνει τιμές διακύμανσης μεγέθους μικρότερες από αυτές που λαμβάνουν οι κακοήθεις όγκοι. Επίσης, στην κλάση της καλοήθειας παρατηρούνται μερικές ακραίες τιμές κοντά στο τέταρτο τεταρτημόριο, όπως επίσης παρατηρείται και στην κλάση της κακοήθειας με τη μόνη διαφορά ότι υπερέχουν όσον αφορά το πλήθος και κατανέμονται σε μεγαλύτερο εύρος.



Εικόνα 3. 76 *Violin plots της ανεξάρτητης μεταβλητής area_worst*

Επίσης, όσον αφορά το γράφημα πυκνότητας υποδεικνύει ότι για τις τιμές διακύμανσης μεγέθους του κυτταρικού πυρήνα οι οποίες κυμαίνονται στο εύρος (500, 510), παρουσιάζεται μεγαλύτερη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας. Ωστόσο, για τις τιμές της διακύμανσης μεγέθους του κυτταρικού πυρήνα οι οποίες κυμαίνονται γύρω από την τιμή του χαρακτηριστικού ίση με 100, αντιστοιχεί η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 77 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής*

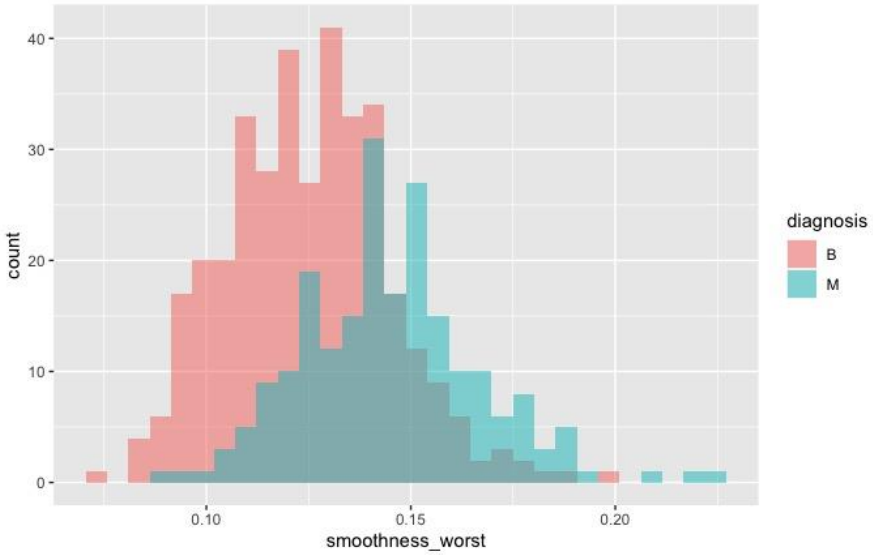
25. Διακύμανση απαλότητας

Το επόμενο χαρακτηριστικό αφορά την διακύμανση της απαλότητας του κυτταρικού πυρήνα που προκύπτει από την ανάλυση δεδομένων εικόνας. Ειδικότερα, στον **πίνακα 3.29**, παρουσιάζονται τα βασικά περιγραφικά μέτρα του χαρακτηριστικού το οποίο παρουσιάζει ότι οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης απαλότητας ίση με 0.0712 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.0882. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.1104 ενώ 0.1305 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση απαλότητας τουλάχιστον ίση με 0.1254 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.1435 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης απαλότητας ίση με 0.1250 και 0.1435 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης απαλότητας ίση με τουλάχιστον 0.1376 και 0.1560 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.2006 για τους καλοήθεις όγκους και 0.2226 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.1294 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 0.1344 κατανέμονται οι κακοήθεις όγκοι.

smoothness_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0712	0.1104	0.1254	0.1250	0.1376	0.2006	0.0200	0.1294
Malignant	0.0882	0.1305	0.1435	0.1449	0.1560	0.2226	0.0219	0.1344

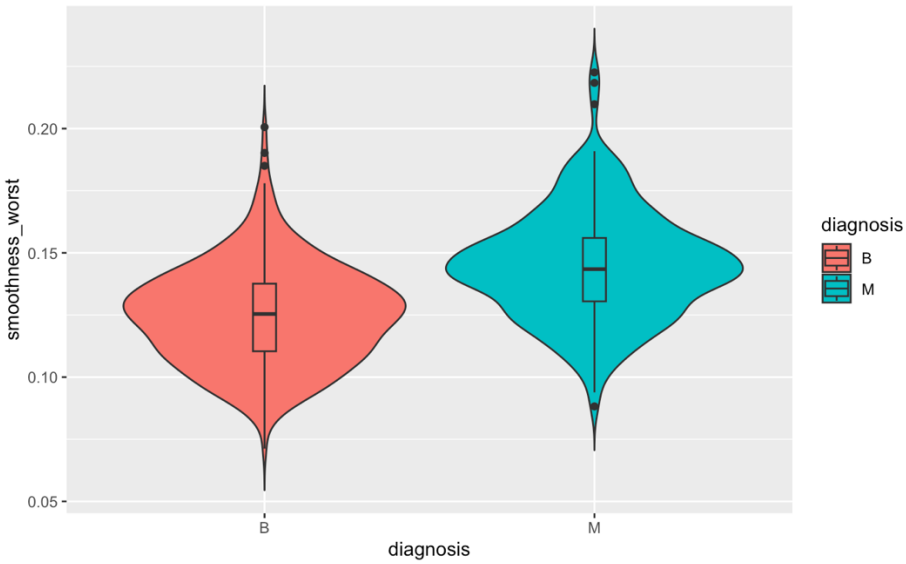
Πίνακας 3. 29 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της απαλότητας*

Σύμφωνα με τις πληροφορίες που παρέχει το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’smoothness_worst’’ (**εικόνα 3.78)**, το μεγαλύτερο ποσοστό των όγκων παρουσιάζουν διακύμανση μεγέθους περίπου 0.15, ενώ το μεγαλύτερο ποσοστό των κακοηθών όγκων παρουσιάζουν διακύμανση μεγέθους περίπου 1300.



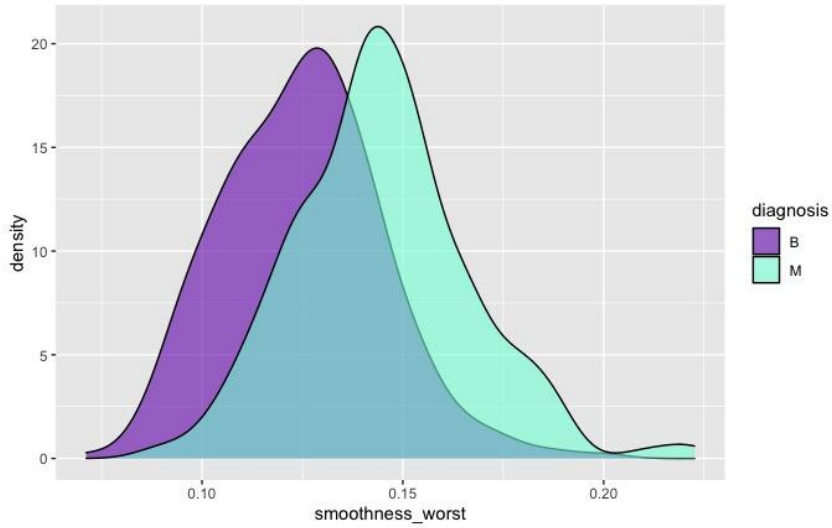
Εικόνα 3. 78 *Ιστόγραμμα ανεξάρτητης μεταβλητής smoothness_worst*

Επίσης τα violin plots **(εικόνα 3.76)** της ανεξάρτητης μεταβλητής ‘’smoothness_worst’’ υποδεικνύουν πως το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων είναι παρόμοιο με αυτό των κακοηθών και λαμβάνει τιμές διακύμανσης απαλότητας μικρότερες από αυτές που λαμβάνουν οι κακοήθεις όγκοι. Επίσης, στην κλάση της καλοήθειας παρατηρούνται μερικές ακραίες τιμές κοντά στο τέταρτο τεταρτημόριο, όπως επίσης παρατηρείται και στην κλάση της κακοήθειας.



Εικόνα 3. 79 *Violin plots της ανεξάρτητης μεταβλητής smoothness_worst*

Επιπλέον, όσον αφορά το γράφημα πυκνότητας πιθανότητας του χαρακτηριστικού διακύμανσης της απαλότητας (**εικόνα 3.80)** προκύπτει ότι για τις τιμές του χαρακτηριστικού αυτού οι οποίες κυμαίνονται στο εύρος (0.12, 0.13), παρουσιάζεται μεγαλύτερη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας. Ωστόσο, για τις τιμές της διακύμανσης απαλότητας του κυτταρικού πυρήνα οι οποίες κυμαίνονται γύρω από την τιμή του χαρακτηριστικού περίπου 0.14, αντιστοιχεί η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 80 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής smoothness_worst*

26.Διακύμανση συμπαγότητας (compactness_worst)

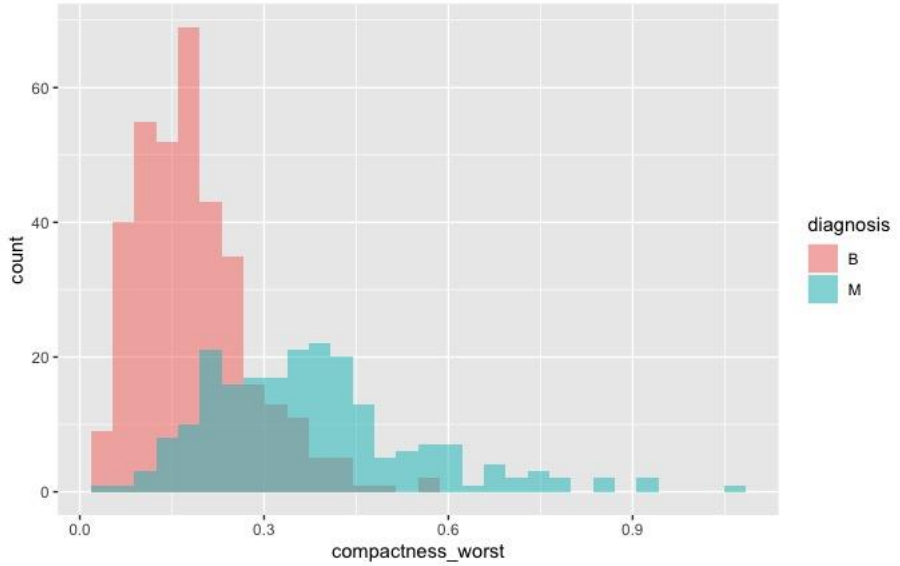
Το χαρακτηριστικό αυτό αφορά τη διακύμανση της σκληρότητας του κυτταρικού πυρήνα που προκύπτει από της μετρήσεις μέσω δεδομένων εικόνας. Ειδικότερα, στον **πίνακα 3.30**, παρουσιάζονται τα βασικά περιγραφικά μέτρα του χαρακτηριστικού το οποίο παρουσιάζει ότι οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης συμπαγότητας ίση με 0.0273 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.0513. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.1120 ενώ 0.2445 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση συμπαγότητας τουλάχιστον ίση με 0.1698 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.3564 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση

τιμή του χαρακτηριστικού διακύμανσης σκληρότητας ίση με 0.1827 και 0.3748 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης συμπαγότητας ίση με τουλάχιστον 0.2302 και 0.4479 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.5849 για τους καλοήθεις όγκους και 1.0580 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.5576 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 1.0076 κατανέμονται οι κακοήθεις όγκοι.

compactness_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0273	0.1120	0.1698	0.1827	0.2302	0.5849	0.0922	0.5576
Malignant	0.0513	0.2445	0.3564	0.3748	0.4479	1.0580	0.1704	1.0067

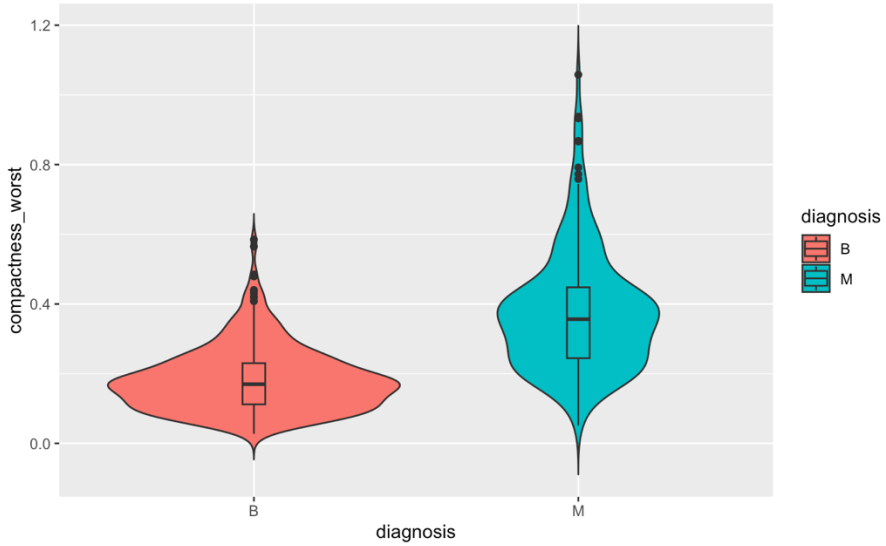
Πίνακας 3. 30 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της συμπαγότητας*

Βάσει των πληροφοριών τις οποίες παρέχει το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’compactness_worst’’ της **εικόνας 3.81**, προκύπτει ότι το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού που αφορά την διακύμανση της συμπαγότητας του κυτταρικού πυρήνα περίπου ίση με 1.8, όπως επίσης το μεγαλύτερο ποσοστό των κακοηθών όγκων κατανέμονται γύρω από το εύρος τιμών $(3.5, 4)$.



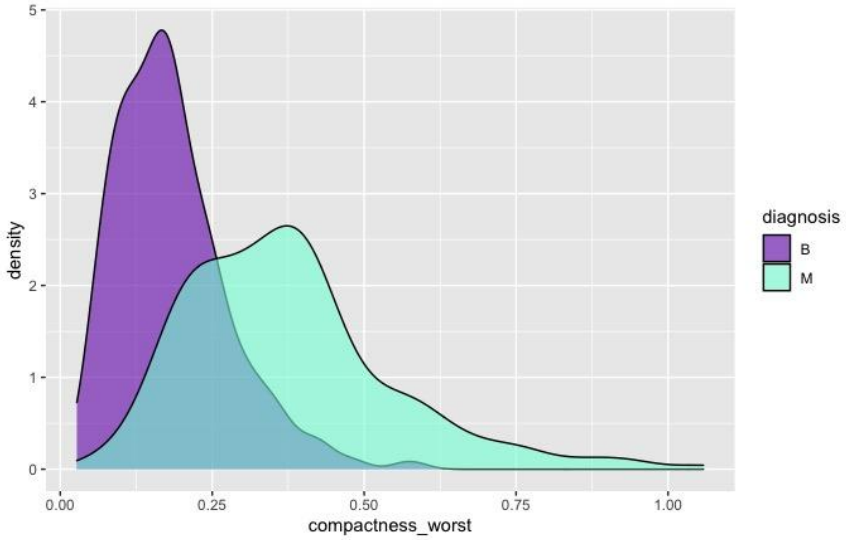
Εικόνα 3. 81 *Ιστόγραμμα ανεξάρτητης μεταβλητής compactness_worst*

Επίσης τα violin plots **(εικόνα 3.82)** της ανεξάρτητης μεταβλητής compactness_worst, υποδεικνύουν πως το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων βάσει των τεταρτημόριων είναι μικρότερο από αυτό στο οποίο κατανέμονται οι κακοήθεις, όπως άλλωστε υποδεικνύουν τα βασικά περιγραφικά μέτρα του **πίνακα 3.30**. Επίσης, στην κλάση της καλοήθειας παρατηρούνται μερικές ακραίες τιμές κοντά στο τέταρτο τεταρτημόριο, όπως επίσης παρατηρείται και στην κλάση της κακοήθειας με τη διαφορά ότι κατανέμονται σε μεγαλύτερο εύρος.



Εικόνα 3. 82 *Violin plots της ανεξάρτητης μεταβλητής compactness_worst*

Επιπλέον, όσον αφορά το γράφημα πυκνότητας πιθανότητας του χαρακτηριστικού διακύμανσης συμπαγότητας **(εικόνα 3.83),** προκύπτει ότι για τις τιμές του χαρακτηριστικού αυτού οι οποίες κυμαίνονται στο εύρος $(0.11, 0.16)$, παρουσιάζεται μεγαλύτερη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας. Ωστόσο, για τις τιμές της διακύμανσης απαλότητας του κυτταρικού πυρήνα οι οποίες κυμαίνονται γύρω από την τιμή του χαρακτηριστικού περίπου 0.37, αντιστοιχεί η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 83 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής compactness_worst*

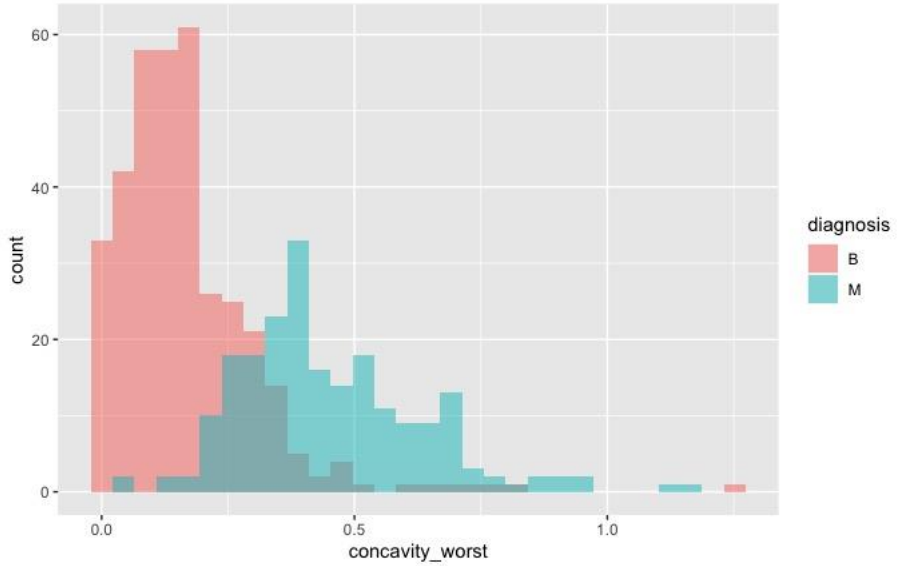
27.Διακύμανση κοιλότητας κυτταρικού πυρήνα (concavity_worst)

Το επόμενο χαρακτηριστικό, με σκοπό της περιγραφικής του ανάλυσης, αφορά τη διακύμανση της κοιλότητας που προκύπτει από τις μετρήσεις. Ειδικότερα, όπως φαίνεται και από τον **πίνακα 3.31**, οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης κοιλότητας ίση με 0 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.024. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού το πολύ 0.0771 ενώ 0.3264 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση κοιλότητας τουλάχιστον ίση με 0.1412 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.4049 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης σκληρότητας ίση με 0.1827 και 0.3748 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης κοιλότητας ίση με τουλάχιστον 0.2216 και 0.5562 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 1.252 για τους καλοήθεις όγκους και 1.17 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 1.252 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 1.1460 κατανέμονται οι κακοήθεις όγκοι.

concavity_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0000	0.0771	0.1412	0.1662	0.2216	1.2520	0.1404	1.2520
Malignant	0.0240	0.3264	0.4049	0.4506	0.5562	1.1700	0.1815	1.1460

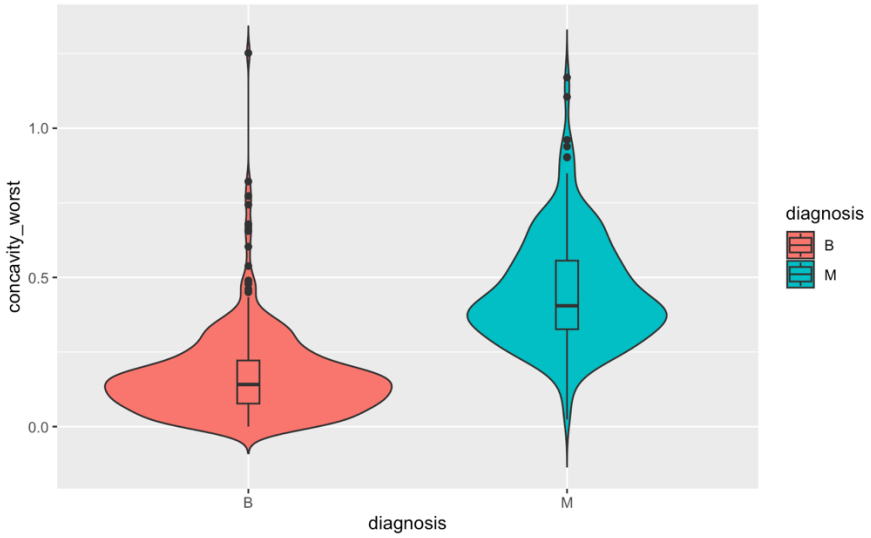
Πίνακας 3.31 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της κοιλότητας κυτταρικού πυρήνα*

Βάσει των πληροφοριών τις οποίες παρέχει το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’concavity_worst’’ της **εικόνας 3.84**, προκύπτει ότι το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού που αφορά την διακύμανση της κοιλότητας του κυτταρικού πυρήνα ίση με περίπου 0.13, όπως επίσης το μεγαλύτερο ποσοστό των κακοηθών όγκων παρουσιάζουν τιμή διακύμανσης κοιλότητας κυτταρικού πυρήνα περίπου ίση με 0.37.



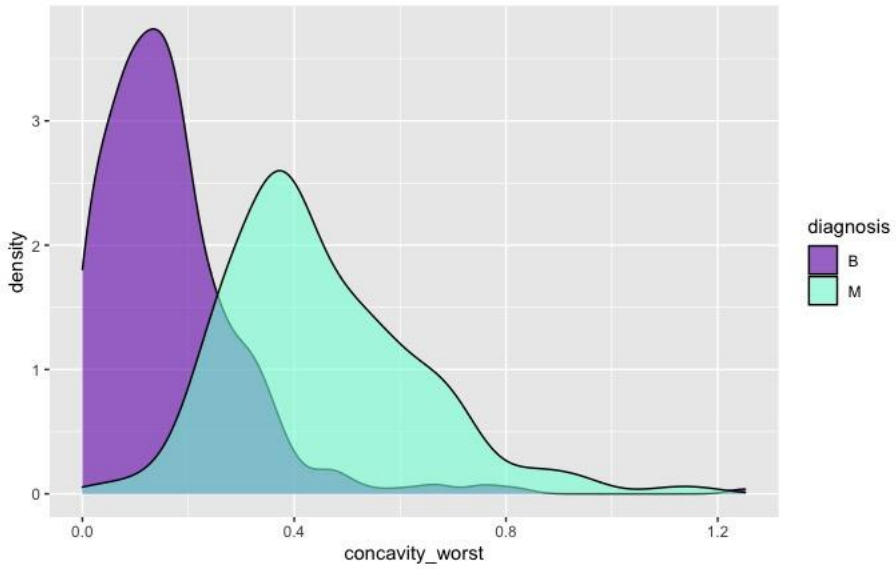
Εικόνα 3. 84 *Ιστόγραμμα της ανεξάρτητης μεταβλητής concavity_worst*

Όσον αφορά τα violin plots της **εικόνας 3.85**, είναι ξεκάθαρο πως η κλάση της καλοήθειας λαμβάνει κατά κύριο λόγο μεγαλύτερες τιμές για το χαρακτηριστικό του κυτταρικού πυρήνα που αφορά τη διακύμανση της κοιλότητας. Ως εκ τούτου η διάμεσος της όπως φαίνεται και από το σχήμα βρίσκεται χαμηλότερα σε σχέση με αυτή της καλοήθειας. Επίσης, βάσει του violin plot της κλάσης της καλοήθειας είναι εμφανές πως υπάρχει μεγαλύτερο πλήθος ακραίων τιμών σε σχέση με αυτό της κακοήθειας οι οποίες κατανέμονται σε πολύ μεγαλύτερο εύρος.



Εικόνα 3. 85 *Violin plots της ανεξάρτητης μεταβλητής concavity_worst*

Πέρα από την κατανομή των παρατηρήσεων συναρτήσει του πλήθους που αντιστοιχούν στις εκάστοτε τιμές του χαρακτηριστικού, εξίσου σημαντική είναι και η περιγραφική ανάλυση που αφορά την πυκνότητα πιθανότητας της ανεξάρτητης μεταβλητής ‘’concavity_worst’’. Πιο συγκεκριμένα το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.86**, προκύπτει πως ο για την τιμή διακύμανσης της κοιλότητας περίπου ίση με 0.17 ενός κυτταρικού πυρήνα είναι πιθανότερο ο όγκος να ανήκει στην κλάση της καλοήθειας. Επίσης, για την τιμή διακύμανσης της κοιλότητας ενός κυτταρικού πυρήνα ίση με περίπου 0.37, αντιστοιχεί η μέγιστη πιθανότητα ο όγκος να ανήκει στην κλάση της κακοήθειας



Εικόνα 3. 86 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής concavity_worst*

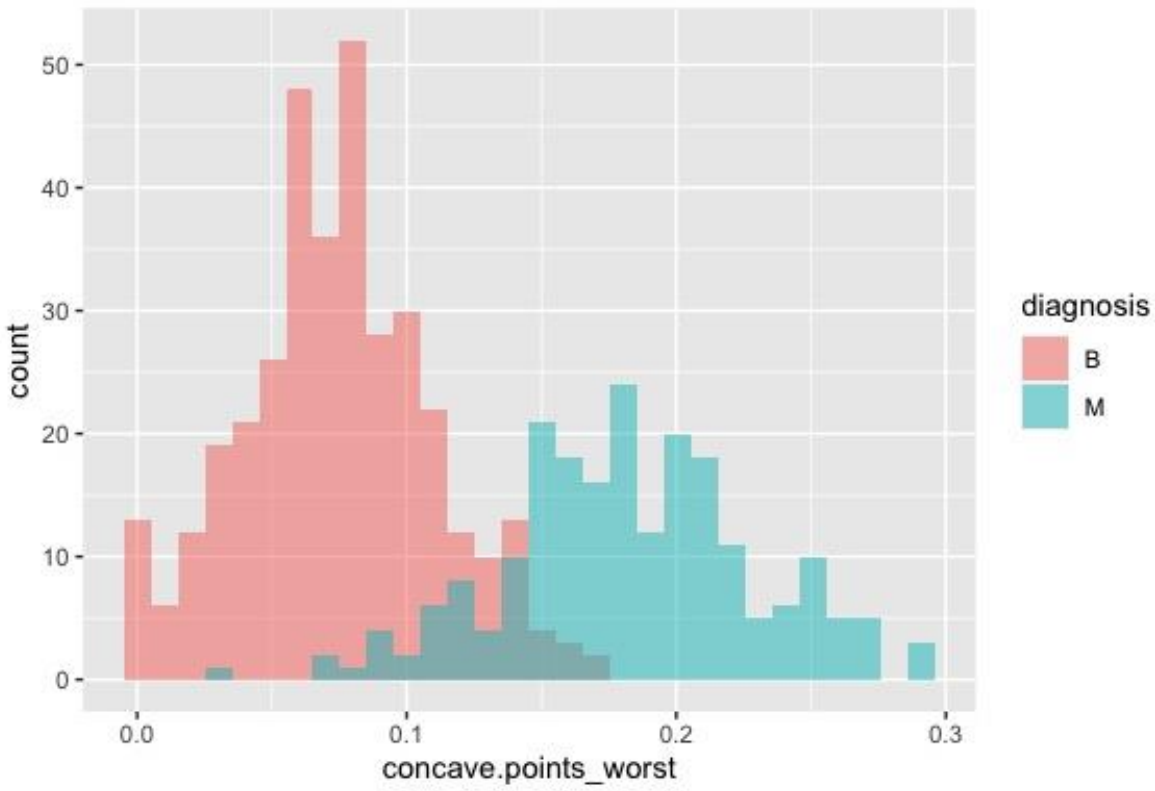
28.Διακύμανση πλήθους κοίλων τμημάτων (concave.points_worst)

Το χαρακτηριστικό αυτό αφορά τη διακύμανση, που προκύπτει από τις μετρήσεις μέσω δεδομένων εικόνας, του πλήθος κοίλων τμημάτων ενός κυτταρικού πυρήνα. Ειδικότερα, όπως φαίνεται και από τον **πίνακα 3.32**, οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα ίση με 0 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.029. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού κυτταρικού πυρήνα το πολύ 0.0510 ενώ 0.1528 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα τουλάχιστον ίση με 0.0743 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.1820 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα ίση με 0.0744 και 0.1822 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα ίση με τουλάχιστον 0.0975 και 0.2107 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.1750 για τους καλοήθεις όγκους και 0.2910 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.175 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 0.262 κατανέμονται οι κακοήθεις όγκοι.

concave.points_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0000	0.0510	0.0743	0.0744	0.0975	0.1750	0.0358	0.1750
Malignant	0.0290	0.1528	0.1820	0.1822	0.2107	0.2910	0.0463	0.2620

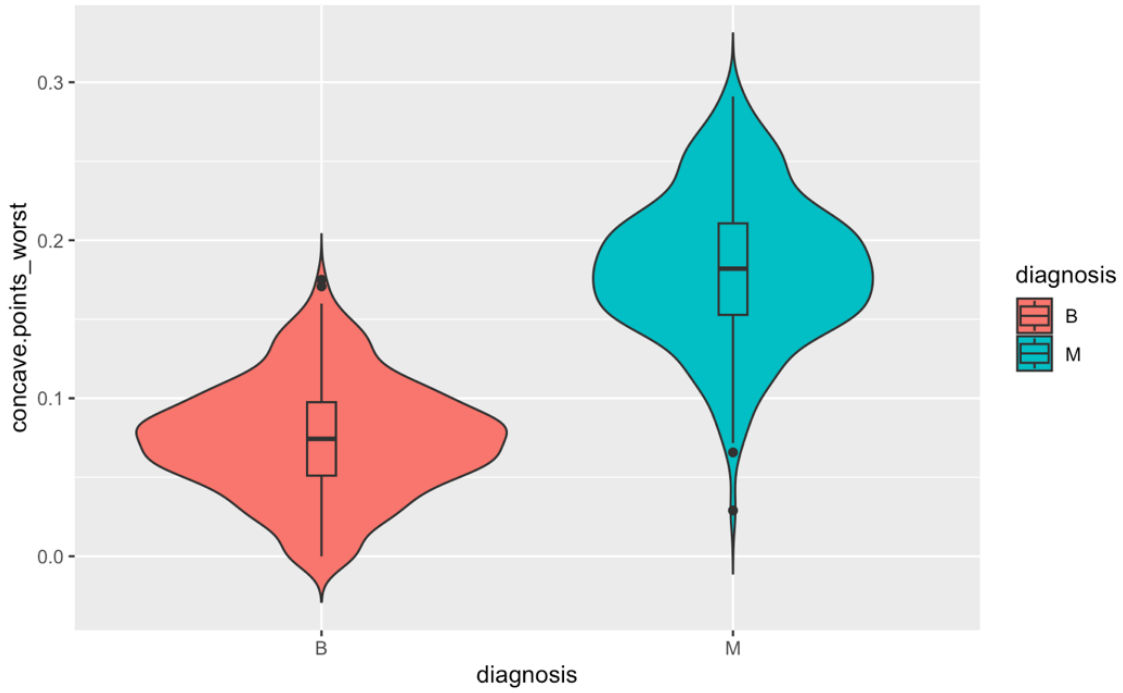
Πίνακας 3.32 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης του πλήθους κοίλων τμημάτων κυτταρικού πυρήνα*

Όπως προκύπτει από το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’concavity_worst’’ της **εικόνας 3.87**, προκύπτει ότι το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού που αφορά την διακύμανση του πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα ίση με περίπου 0.075, όπως επίσης το μεγαλύτερο ποσοστό των κακοηθών όγκων παρουσιάζουν τιμή διακύμανσης του πλήθους κοίλων τμημάτων κυτταρικού πυρήνα περίπου ίση με 0.175.



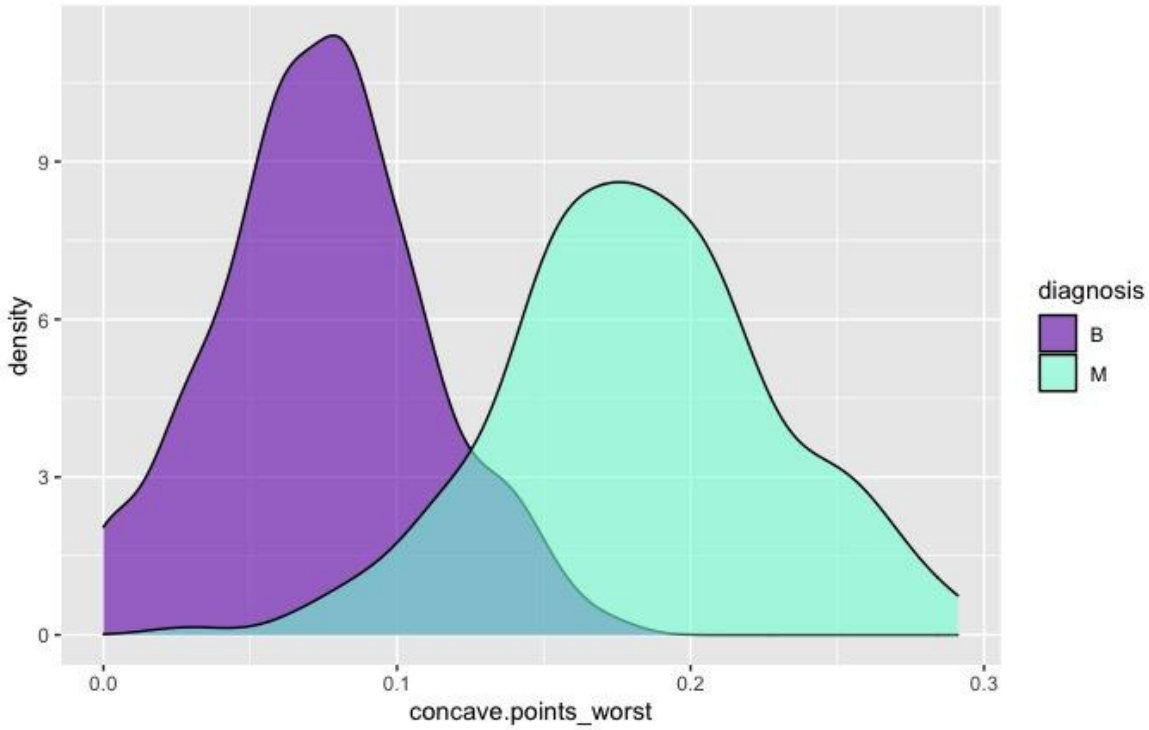
Εικόνα 3. 87 *Ιστόγραμμα ανεξάρτητης μεταβλητής concave.points_worst*

Έπειτα, παρατηρώντας τα violin plots της ανεξάρτητης μεταβλητής ‘’concave.points_worst’’, προκύπτει ότι οι κατανομές των παρατηρήσεων των δύο κλάσεων είναι περίπου πανομοιότυπες με αυτή της καλοήθειας να λαμβάνει μικρότερες τιμές διακύμανσης του πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα σε σχέση με την κλάση της κακοήθειας, ενώ στις παρατηρήσεις των καλοηθών όγκων παρατηρείται μία ακραία τιμή στο τέταρτο τεταρτημόριο ενώ το αντίθετο συμβαίνει στην κλάση της κακοήθειας όπου παρατηρούνται δύο ακραίες τιμές κοντά στο πρώτο τεταρτημόριο.



Εικόνα 3. 88 *Violin plots της ανεξάρτητης μεταβλητής concave.points_worst*

Πέρα από την κατανομή των παρατηρήσεων συναρτήσει του πλήθους που αντιστοιχούν στις εκάστοτε τιμές του χαρακτηριστικού, εξίσου σημαντική είναι και η περιγραφική ανάλυση που αφορά την πυκνότητα πιθανότητας της ανεξάρτητης μεταβλητής ‘’concavity_worst’’. Πιο συγκεκριμένα το γράφημα πυκνότητας πιθανότητας της **εικόνας 3.89**, προκύπτει πως ο για την τιμή διακύμανσης του πλήθους κοίλων τμημάτων του κυτταρικού πυρήνα περίπου ίση με 0.075 ενός κυτταρικού πυρήνα είναι πιθανότερο ο όγκος να ανήκει στην κλάση της καλοήθειας. Επίσης, για την τιμή διακύμανσης του πλήθους κοίλων τμημάτων ενός κυτταρικού πυρήνα ίση με περίπου 0.17, αντιστοιχεί η μέγιστη πιθανότητα ο όγκος να ανήκει στην κλάση της κακοήθειας



Εικόνα 3. 89 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής concave.points_worst*

29. Διακύμανση συμμετρίας του κυτταρικού πυρήνα (symmetry_worst)

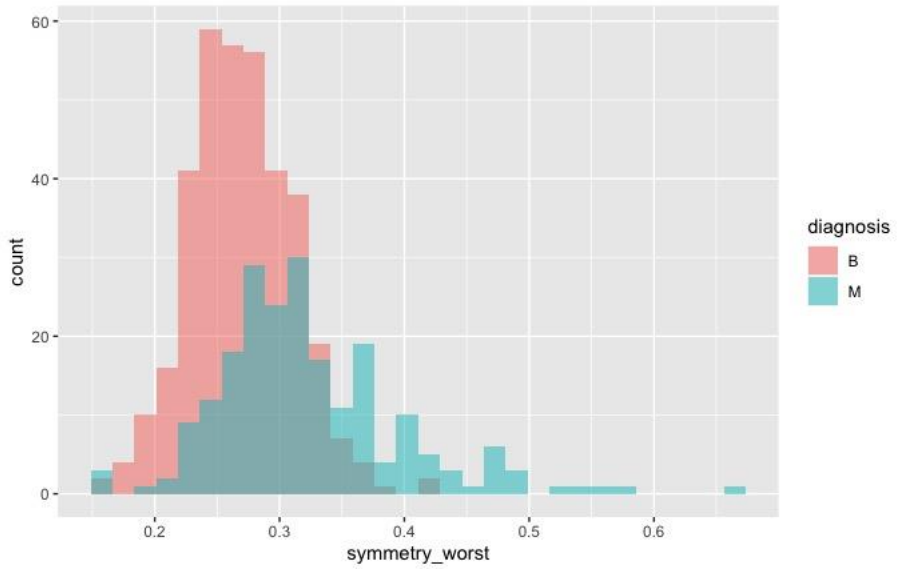
Το χαρακτηριστικό αυτό αφορά τη διακύμανση, που προκύπτει από τις μετρήσεις μέσω δεδομένων εικόνας, της συμμετρίας ενός κυτταρικού πυρήνα. Ειδικότερα, όπως φαίνεται και από τον **πίνακα 3.33**, οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης συμμετρίας του κυτταρικού πυρήνα ίση με 0.1566 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.1565. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού του κυτταρικού πυρήνα το πολύ 0.2406 ενώ 0.2765 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανση συμμετρίας του κυτταρικού πυρήνα τουλάχιστον ίση με 0.2687 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.3235 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης συμμετρίας του κυτταρικού πυρήνα ίση με 0.2702 και 0.3235 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης συμμετρίας του κυτταρικού πυρήνα ίση με τουλάχιστον 0.2983 και 0.3592 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές

0.4228 για τους καλοήθεις όγκους και 0.6638 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.2662 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 0.5073 κατανέμονται οι κακοήθεις όγκοι.

symmetry_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.1566	0.2406	0.2687	0.2702	0.2983	0.4228	0.0417	0.2662
Malignant	0.1565	0.2765	0.3103	0.3235	0.3592	0.6638	0.0747	0.5073

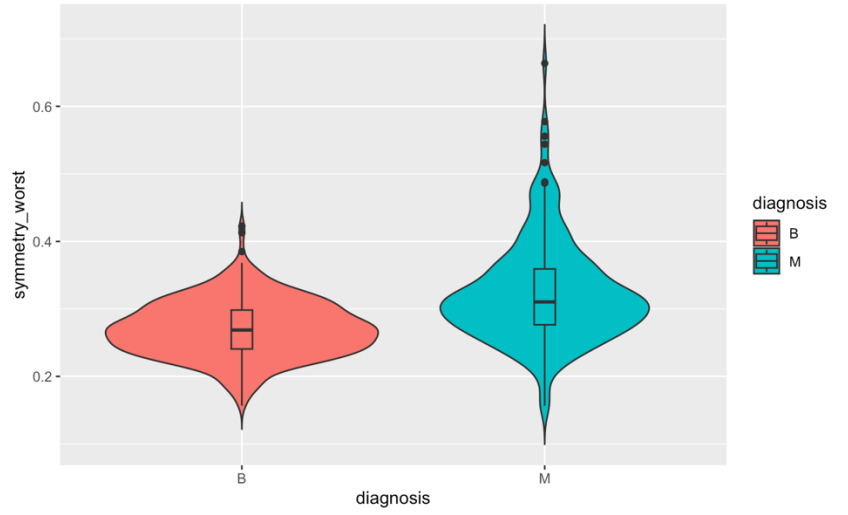
Πίνακας 3. 33 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της συμμετρίας του κυτταρικού πυρήνα*

Όπως προκύπτει από το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’symmetry_worst’’ της **εικόνας 3.90**, προκύπτει ότι το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού που αφορά την διακύμανση συμμετρίας του κυτταρικού πυρήνα η οποία κυμαίνεται στο εύρος (0.13, 0.17), όπως επίσης το μεγαλύτερο ποσοστό των κακοηθών όγκων παρουσιάζουν τιμή διακύμανσης συμμετρίας του κυτταρικού πυρήνα περίπου ίση με 0.32.



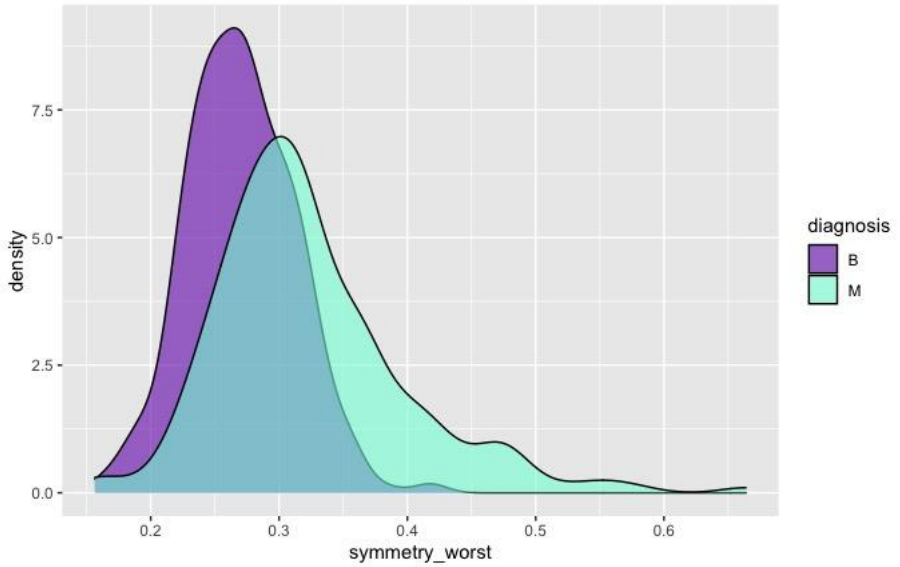
Εικόνα 3. 90 *Ιστόγραμμα της ανεξάρτητης μεταβλητής symmetry_worst*

Επίσης, τα violin plots **(εικόνα 3.91)** της ανεξάρτητης μεταβλητής symmetry_worst, υποδεικνύουν πως το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων βάσει των τεταρτημόριων είναι μικρότερο από αυτό στο οποίο κατανέμονται οι κακοήθεις, όπως άλλωστε υποδεικνύουν και τα βασικά περιγραφικά μέτρα του **πίνακα 3.33**. Επίσης, στην κλάση της καλοήθειας παρατηρούνται μερικές ακραίες τιμές κοντά στο τέταρτο τεταρτημόριο, όπως επίσης παρατηρείται και στην κλάση της κακοήθειας με τη διαφορά ότι κατανέμονται σε μεγαλύτερο εύρος.



Εικόνα 3. 91 *Violin plots της ανεξάρτητης μεταβλητής symmetry_worst*

Επιπλέον, όσον αφορά το γράφημα πυκνότητας πιθανότητας του χαρακτηριστικού διακύμανσης της απαλότητας (**εικόνα 3.92)** προκύπτει ότι για τις τιμές του χαρακτηριστικού αυτού οι οποίες κυμαίνονται στο εύρος (0.25, 0.26), παρουσιάζεται μεγαλύτερη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας. Ωστόσο, για τις τιμές της διακύμανσης συμμετρίας του κυτταρικού πυρήνα οι οποίες κυμαίνονται γύρω από την τιμή του χαρακτηριστικού περίπου ίση με 0.3, αντιστοιχεί η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της κακοήθειας.



Εικόνα 3. 92 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής symmetry_worst*

30. Διακύμανση μορφοκλασματικής διάστασης κυτταρικού πυρήνα (fractal_dimension_worst)

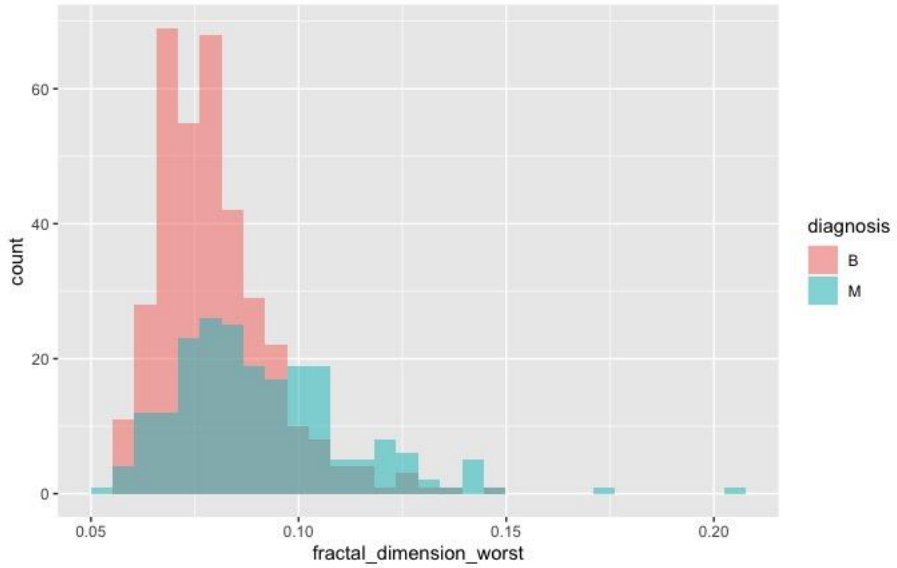
Η τελευταία ανεξάρτητη μεταβλητή της οποίας θα αναλυθούν τα βασικά περιγραφικά μέτρα αποτελεί η διακύμανση που προκύπτει από τις μετρήσεις μέσω δεδομένων εικόνας της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα. Πιο συγκεκριμένα από τον **πίνακα 3.34**, προκύπτει ότι οι καλοήθεις όγκοι λοιπόν παρουσιάζουν ελάχιστη τιμή διακύμανσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα ίση με 0.0552 ενώ οι κακοήθεις παρουσιάζουν ελάχιστη τιμή ίση με 0.0550. Επίσης, το 25% των όγκων της καλοήθειας λαμβάνουν τιμή χαρακτηριστικού του κυτταρικού πυρήνα το πολύ 0.0701 ενώ 0.0763 το 25% των όγκων της κακοήθειας. Επιπροσθέτως, το 50% των καλοηθών όγκων παρουσιάζουν διακύμανσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα τουλάχιστον ίση με 0.0771 ενώ οι κακοήθεις όγκοι τουλάχιστον ίση με 0.0876 Ακόμη, οι όγκοι που ανήκουν στην κλάση της καλοήθειας, παρουσιάζουν μέση τιμή του χαρακτηριστικού διακύμανσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα ίση με 0.794 και 0.0915 οι όγκοι οι οποίοι ανήκουν στην κλάση της κακοήθειας. Επιπλέον, το 25% των καλοηθών όγκων λαμβάνει τιμή διακύμανσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα ίση με τουλάχιστον 0.0854 και 0.1026 το 25% των κακοήθων όγκων. Ως μέγιστες τιμές των παρατηρήσεων που προκύπτουν από τις μετρήσεις αποτελούν οι τιμές 0.1486 για τους καλοήθεις όγκους και 0.0216 για τους κακοήθεις αντιστοίχως. Τέλος, όσον αφορά τα βασικά περιγραφικά μέτρα, αποτελούν το εύρος στο οποίο κατανέμονται οι παρατηρήσεις και η τυπική απόκλιση που

προκύπτει από τις μετρήσεις. Πιο συγκεκριμένα το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων ισούται με 0.0934 ενώ σε λίγο μεγαλύτερο το οποίο ισούται με 0.1525 κατανέμονται οι κακοήθεις όγκοι.

fractal_dimension_worst	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	range
Benign	0.0552	0.0701	0.0771	0.0794	0.0854	0.1486	0.0138	0.0934
Malignant	0.0550	0.0763	0.0876	0.0915	0.1026	0.2075	0.0216	0.1525

Πίνακας 3. 34 *Βασικά περιγραφικά μέτρα του χαρακτηριστικού διακύμανσης της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα*

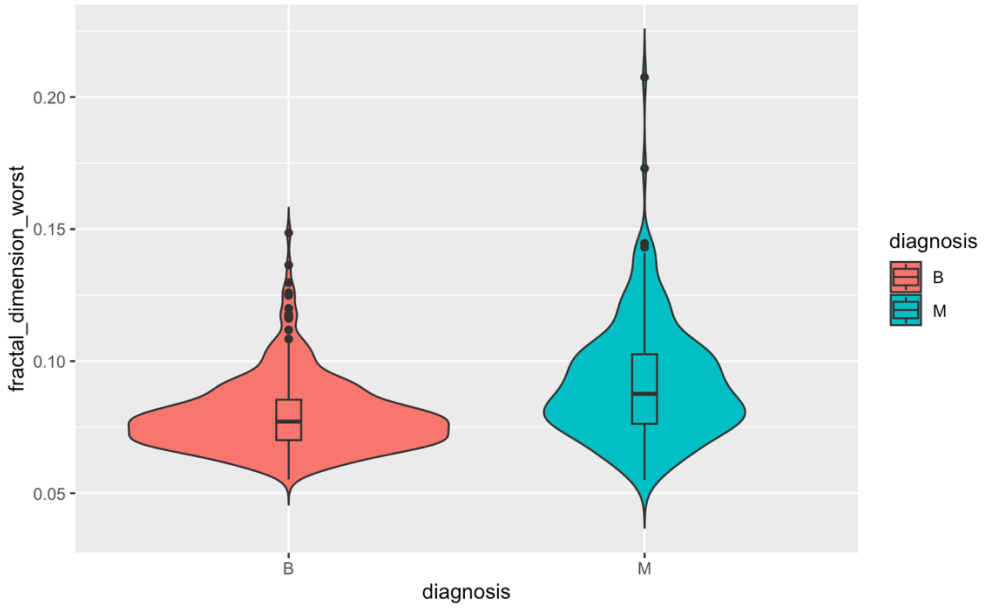
Βάσει των πληροφοριών τις οποίες παρέχει το ιστόγραμμα της ανεξάρτητης μεταβλητής ‘’fractal_dimension_worst’’ της **εικόνας 3.93**, προκύπτει ότι το μεγαλύτερο ποσοστό των καλοηθών όγκων λαμβάνουν την τιμή του χαρακτηριστικού που αφορά την διακύμανση της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα περίπου ίση με 0.07, όπως επίσης το μεγαλύτερο ποσοστό των κακοηθών όγκων κατανέμονται γύρω από το εύρος τιμών (0.075, 0.08).



Εικόνα 3.93 *Ιστόγραμμα της ανεξάρτητης μεταβλητής της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα*

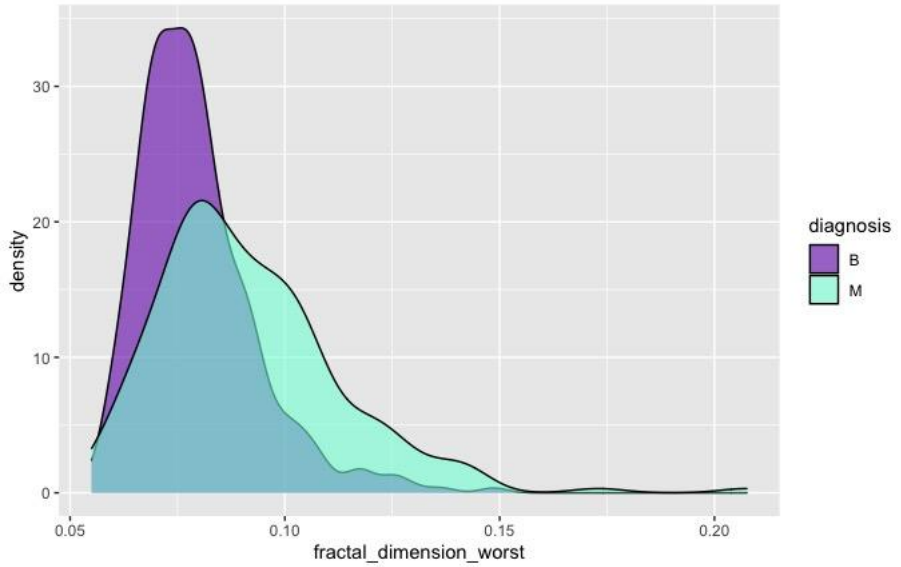
Επίσης το violin plot **(εικόνα 3.94)** της ανεξάρτητης μεταβλητής ‘’fractal_dimension_worst’’ υποδεικνύουν πως το εύρος στο οποίο κατανέμονται οι παρατηρήσεις των καλοηθών όγκων

είναι αρκετά μικρότερο και λαμβάνει τιμές διακύμανσης μορφοκλασματικής διάστασης κυτταρικού πυρήνα μεγαλύτερες από αυτές που λαμβάνουν οι κακοήθεις όγκοι. Επίσης, στην κλάση της καλοήθειας παρατηρούνται μερικές ακραίες τιμές κοντά στο τέταρτο τεταρτημόριο, όπως επίσης παρατηρείται και στην κλάση της κακοήθειας με τη μόνη διαφορά ότι υπερέχουν αριθμητικά όσον αφορά το πλήθος όπως επίσης ότι κατανέμονται και σε μεγαλύτερο εύρος.



Εικόνα 3. 94 *Violin plots της ανεξάρτητης μεταβλητής της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα*

Τέλος, όσον αφορά το γράφημα πυκνότητας πιθανότητας του χαρακτηριστικού διακύμανσης της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα (**εικόνα 3.5)** προκύπτει ότι για τις τιμές του χαρακτηριστικού αυτού οι οποίες κυμαίνονται στο εύρος (0.07, 0.08), παρουσιάζεται μεγαλύτερη πιθανότητα ο όγκος να ανήκει στην κλάση της καλοήθειας. Ωστόσο, για τις τιμές της διακύμανσης της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα οι οποίες κυμαίνονται γύρω από την τιμή του χαρακτηριστικού αυτού περίπου ίση με 0.077, αντιστοιχεί η μέγιστη πιθανότητα να ανήκει ο όγκος στην κλάση της κακοήθειας.



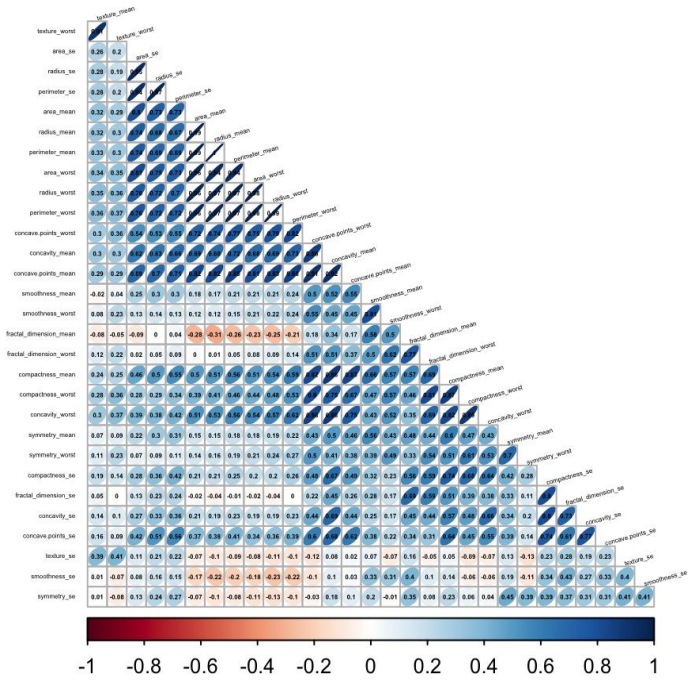
Εικόνα 3. 95 *Γράφημα πυκνότητας πιθανότητας της ανεξάρτητης μεταβλητής της μορφοκλασματικής διάστασης του κυτταρικού πυρήνα*

Συνοψίζοντας. Όσον αφορά τις συνεχείς μεταβλητές του συνόλου των δεδομένων που εξετάστηκαν, παρατηρείται ότι η μέση τιμή και η διάμεσος δεν παρουσιάζουν ιδιαίτερα μεγάλες αποκλίσεις. Τόσο για την κλάση των καλοηθών όγκων όσο και για την κλάση των κακοήθων. Επιπροσθέτως γίνεται αντιληπτό ότι οι όγκοι οι οποίοι ανήκουν στην κλάση της καλοήθειας παρουσιάζουν κατά κύριο λόγο μικρότερες τιμές στα χαρακτηριστικά του κυτταρικού πυρήνα σε αντίθεση με αυτές των όγκων που ανήκουν στην κλάση της κακοήθειας. Πιο συγκεκριμένα, οι κακοήθεις όγκοι παρουσιάζουν μεγαλύτερη μέση απόσταση του κυτταρικού πυρήνα σε σημεία της περιμέτρου του κατά μέσο όρο 5.313 pixels σε σχέση με τους καλοήθεις όγκους. Επίσης οι κακοήθεις όγκοι παρουσιάζουν μεγαλύτερη μέση τιμή υφής κατά μέσο 3.69 μονάδες της κλίμακας του γκρι έναντι των καλοηθών. Ακόμη κατά μέσο όρο οι κακοήθεις όγκοι παρουσιάζουν μέση τιμή περιμέτρου μεγαλύτερη κατά 37.29 pixels από τους καλοήθεις. Στη συνέχεια το μέσο μέγεθος του κυτταρικού πυρήνα των κακοήθων όγκων είναι κατά μέσο όρο μεγαλύτερο κατά 515.6 $pixels^2$ έναντι αυτού των καλοηθών, όπως επίσης και οι καλοήθεις όγκοι είναι κατά μέσο όρο 0.01 μονάδες πιο απαλοί από τους κακοήθεις. Επίσης οι κακοήθεις όγκοι είναι κατά μέσο όρο 0.07 μονάδες πιο συμπαγείς από τους καλοήθεις. Όσον αφορά την κοιλότητα, οι ασθενείς οι οποίοι έχουν διαγνωστεί με κακοήθη όγκο παρουσιάζουν μεγαλύτερη κοιλότητα κυτταρικού πυρήνα κατά μέσο όρο 0.11 μονάδες σε σχέση με αυτούς που έχουν διαγνωστεί με καλοήθεια, ενώ στους ασθενείς που έχουν διαγνωστεί με κακοήθεια παρατηρούνται κατά μέσο όρο 0.07 περισσότερα κοίλα τμήματα στον κυτταρικό τους πυρήνα. Έπειτα οι ευθείες δύο οριακών σημείων οι οποίες διασχίζουν το κέντρο του κυτταρικού πυρήνα είναι κατά μέσο όρο 0.02 pixels μεγαλύτερες στους κακοήθεις συγκριτικά με τους καλοήθεις ο όγκος. Τέλος ο κυτταρικός πυρήνας ενός καλοήθη όγκου

παρουσιάζει κατά μέσο όρο 0.0002 μονάδες παραπάνω όσον αφορά το χαρακτηριστικό της μορφοκλασματικής διάστασης του κακοήθη.

3.5 Συσχετίσεις μεταξύ των μεταβλητών

Στο συγκεκριμένο σκέλος της διπλωματικής εργασίας θα εξεταστούν οι σχέσεις που έχουν μεταξύ τους μεταβλητές για την καλύτερη κατανόηση τους. Ειδικότερα για την επίτευξη της καλύτερης κατανόησης είναι απαραίτητος ο υπολογισμός του συντελεστή γραμμικής συσχέτισης Pearson, ο οποίος στην ουσία αποτελεί μέτρο σε εντάσεις τις εξαρτήσεις μεταξύ δύο ποσοτικών μεταβλητών. Το πεδίο ορισμού του συντελεστή γραμμικής συσχέτισης αποτελείται από τις τιμές $[-1, 1]$. Όταν ο συντελεστής συσχέτισης λαμβάνει την τιμή -1 υπάρχει ένδειξη τελείας αρνητικής συσχετίσεις μεταξύ 2 ποσοτικών μεταβλητών. Στην περίπτωση όπου $r = 0$ οι δυο μεταβλητές προς εξέταση δεν σχετίζονται γραμμικά, ενώ όταν $r = +1$ τότε οι δύο μεταβλητές οι οποίες εξετάζονται υποδεικνύουν τέλεια θετική γραμμική συσχέτιση. Τα αποτελέσματα της παραμετρικής μεθόδου συσχέτισης Pearson που εφαρμόστηκε με σκοπό να εξεταστεί αν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών του συγκεκριμένου συνόλου δεδομένων εμφανίζεται στην **εικόνα 3.96**.



Εικόνα 3. 96 *Μέθοδος συσχέτισης με τη μέθοδο Pearson*

Αξίζει να σημειωθεί ότι στο παραπάνω σχήμα δεν αναγράφεται μόνο ο συντελεστής γραμμικής συσχέτισης αλλά αναπαρίσταται και γραφικά για κάθε ζεύγος μεταβλητών

Αρχικά, να επισημανθεί ότι παρόλο που υπάρχουν μεταβλητές οι οποίες έχουν αρνητική γραμμική συσχέτιση, δεν παρατηρούνται ακραίες τιμές. Έπειτα, εξετάζονται οι συσχετίσεις των μεταβλητών με δείκτη γραμμικής συσχέτισης μεγαλύτερο από 0.85. οι οποίες βρίσκονται σε υψηλά και κρίσιμα σημεία, η ανεξάρτητη μεταβλητή η οποία αφορά τη μέση απόσταση του κυτταρικού πυρήνα σε σημεία της περιμέτρου του παρουσιάζει υψηλή θετική γραμμική συσχέτιση με τις ανεξάρτητες μεταβλητές που αφορούν τη μέση περίμετρο του , το μέσο μεγέθους του, τη διακύμανση της απόστασης του πυρήνα σε σημεία της περιμέτρου του, τη διακυμάνσεις της περιμέτρου του και τέλος τη διακύμανση του μεγέθους του. Έπειτα, η ανεξάρτητη μεταβλητή που αφορά τη μέση υφή του κυτταρικού πυρήνα παρουσιάζει ισχυρή και θετική γραμμική συσχέτιση με την ανεξάρτητη μεταβλητή που αφορά τη διακύμανση της υφής του στην κλίμακα του γκρι. Επίσης, ανεξάρτητη μεταβλητή που αφορά τη μέση περίμετρο του κυτταρικού πυρήνα παρουσιάζει υψηλή και θετική γραμμική συσχέτιση με τις ανεξάρτητες μεταβλητές που αφορούν το μέσο μεγέθους του, τη διακύμανσης της απόστασης του πυρήνα σε σημεία της περιμέτρου του, τη διακύμανση της περιμέτρου του αλλά και τη διακύμανσης του μεγέθους του. Ακόμη, η ανεξάρτητη μεταβλητή που αφορά το μέσο μεγέθους του κυτταρικού πυρήνα παρουσιάζει υψηλή θετική γραμμική συσχέτιση με τις ανεξάρτητες μεταβλητές που αφορούν τη μέση περίμετρο του, τη διακύμανσης της ακτίνας από τον πυρήνα σε σημεία της περιμέτρου του, τη διακύμανση της περιμέτρου του και τέλος τη διακύμανση του μεγέθους του. Στη συνέχεια, η ανεξάρτητη μεταβλητή που αφορά τη μέση συμπαγότητα του κυτταρικού πυρήνα παρουσιάζει υψηλή θετική γραμμική συσχέτιση με την ανεξάρτητη μεταβλητή που αφορά τη μέση κοιλότητα του, το μέσο πλήθος των κοίλων τμημάτων του κυτταρικού πυρήνα και της διακύμανσης της συμπαγότητας του. Παρακάτω, η ανεξάρτητη μεταβλητή που αφορά τη μέση κοιλότητα του κυτταρικού πυρήνα παρουσιάζει υψηλή θετική γραμμική συσχέτιση με τις ανεξάρτητες μεταβλητές που αφορούν το μέσο πλήθος κοίλων σημείων του, την τυπική απόκλιση της κοιλότητας του, τη διακύμανση της συμπαγότητας του, τη διακύμανση της κοιλότητας του και τη διακύμανση του πλήθους των κοίλων τμημάτων του. Η ανεξάρτητη μεταβλητή του μέσου πλήθους των κοίλων τμημάτων του κυτταρικού πυρήνα παρουσιάζει υψηλή θετική γραμμική συσχέτιση μόνο με την ανεξάρτητη μεταβλητή που αφορά τη διακύμανση του πλήθους των κοίλων τμημάτων του κυτταρικού πυρήνα. Η τυπική απόκλιση της απόστασης του κυτταρικού πυρήνα σε σημεία της περιμέτρου του παρουσιάζει ισχυρή θετική γραμμική συσχέτιση με τις ανεξάρτητες μεταβλητές που αφορούν την τυπική απόκλισης περιμέτρου και του μεγέθους του κυτταρικού πυρήνα. Το χαρακτηριστικό του κυτταρικού πυρήνα που αφορά την τυπική απόκλιση της περιμέτρου του παρουσιάζει υψηλή

θετική γραμμική συσχέτιση με την ανεξάρτητη μεταβλητή της τυπικής απόκλισης του μεγέθους του. Η ανεξάρτητη μεταβλητή της τυπικής απόκλισης της συμπαγότητας του κυτταρικού πυρήνα παρουσιάζει υψηλή γραμμική θετική συσχέτιση με την ανεξάρτητη μεταβλητή της τυπικής απόκλισης της κοιλότητας του.

Λόγω της υψηλής γραμμικής συσχέτισης πολλών ανεξάρτητων μεταβλητών μεταξύ τους, εφαρμόστηκε μια μέθοδο κατά της οποίας επιλέγονται οι μεταβλητές με δείκτη γραμμικής συσχέτισης μικρότερο για τα σύνορα απόφασης 0.75 και στη συνέχεια εξετάστηκε ποιες μεταβλητες παρουσίασαν συντελεστή πληθωρισμού διακύμανσης (VIF: variance inflation factor) μικρότερο από 10. Στον **πίνακα 3.35**, παρουσιάζεται ο συντελεστής πληθωρισμού διακύμανσης των εναπομεινάντων μεταβλητών για σύνορο γραμμικής συσχέτισης ίσο με 0.75.

Μεταβλητή	V.I.F
radius_mean	3.167
texture_mean	2.547
smoothness_mean	4.480
compactness_mean	8.904
symmetry_mean	4.067
fractal_dimension_mean	10.134
radius_se	2.182
texture_se	2.280
smoothness_se	2.510
concave.points_se	3.250
symmetry_se	4.155
fractal_dimension_se	3.999
symmetry_worst	7.987

Πίνακας 3. 35 *Δείκτης συντελεστή πληθωρισμού διακύμανσης των μεταβλητών για σύνορο απόφασης γραμμικής συσχέτισης ίσο με 0.75*

Όπως προέκυψε και από τις μετρήσεις, παρατηρείται το φαινόμενο της πολυσυγγραμμικότητας κατά του οποίου παραπάνω από δύο μεταβλητές παρουσιάζουν υψηλή γραμμική συσχέτιση μεταξύ τους. Ως εκ τούτου η ανεξάρτητη μεταβλητή ‘’ fractal_dimension_mean’’ παρουσιάζει

VIF μεγαλύτερο από 10 γεγονός που υποδεικνύει την ύπαρξη επιβλαβούς πολυσυγγραμμικότητας. Για τη ‘’θεραπεία’’ του προβλήματος αυτού δεν αρκεί μόνο να παραμένουν οι ανεξάρτητες μεταβλητές με δείκτης γραμμικής συσχέτισης μικρότερο από 0.75 διότι μπορεί να υπάρχουν υψηλά γραμμικά συσχετιζόμενες μεταβλητές σε ζεύγη παραπάνω των δύο. Για τον λόγω αυτό επικαλείται ο δείκτης συνθήκης (CI: Condition index), ο οποίος αποτελεί μέτρο του κατά πόσο σχετίζονται μεταξύ τους δύο ή περισσότερες μεταβλητές σε ένα σύνολο δεδομένων. Όσο υψηλότερη είναι η τιμή του δείκτη συνθήκης, τόσο πιο σχετικές είναι οι μεταβλητές. Κρίσιμες τιμές για τον δείκτη συνθήκης θεωρούνται οι τιμές μεγαλύτερες από 10 έως 30. Όταν ο συντελεστής πληθωρισμού διακύμανσης ξεπερνά την τιμή 10 και ο αριθμός συνθήκης ξεπερνά τα διαστήματα 5 με 10 και 10 με 30 αντίστοιχα, τότε είναι βέβαιο πως οι ανεξάρτητες μεταβλητές του συνόλου δεδομένων πάσχουν από το φαινόμενο της πολυσυγγραμμικότητας. Ωστόσο, δεν είναι εφικτό να εντοπιστεί ποιες μεταβλητές ευθύνονται. Για τον προσδιορισμό των μεταβλητών οι οποίες προκαλούν επιβλαβή πολυσυγγραμμικότητα, χρησιμοποιείται ο δείκτης ‘’variance decomposition proportions’’ (VDP). Εάν οι τιμές του ‘’VDP’’ κυμαίνονται παραπάνω από 0,8 με 0,9 και αντιστοιχούν σε τιμές του δείκτη συνθήκης μεγαλύτερες του 10 έως 30, οι επεξηγηματικές μεταβλητές, οι οποίες σχετίζονται με τον ‘’variance decomposition proportion’’ είναι πολυσυγγραμμικές. Για τον λόγω αυτό αφαιρείται η μεταβλητή που αφορά το χαρακτηριστικό μέσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα (Hae, 2019). Στον **πίνακα 3.36**, αναγράφονται τα αποτελέσματα των διαγνωστικών μέτρων πολυσυγγραμμικότητας. Ειδικότερα, τα κελιά της στήλης ‘’condition index’’ με κόκκινο γέμισμα αφορούν τις τιμές του δείκτη οι οποίες ξεπερνούν τα κρίσιμα όρια, δηλαδή πάνω από 10. Επίσης, τα κελιά με κόκκινο γέμισμα της δεύτερης και της τρίτης στήλης αφορά τις τιμές του δείκτη VDP, οι οποίες ξεπερνούν τα κρίσιμα όρια. Πιο συγκεκριμένα, στη δεύτερη στήλη αναγράφονται οι τιμές του δείκτη ‘’VDP’’ της ελεύθερης μεταβλητής, ενώ για καθαρά χωροταξικούς λόγους η τρίτη στήλη εμπεριέχει τις τιμές του δείκτη ‘’VDP’’ της ανεξάρτητης μεταβλητής η οποία αφορά το χαρακτηριστικό της μέσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα καθώς είναι και η μόνη η οποία ξεπερνά τα κρίσιμα όρια.

Eigenvalue	Condition Index	INTERCEPT (Variance Decomposition Proportion)	fractal_dimension_mean
12.7117	1.00	0.00001	0.00001
0.3631	5.92	0.00024	0.00013
0.3113	6.39	0.00000	0.00003
0.2070	7.84	0.00004	0.00018
0.1052	10.99	0.00006	0.00001
0.0947	11.58	0.00020	0.00011
0.0749	13.03	0.00017	0.00035
0.0607	14.47	0.00124	0.00005
0.0295	20.74	0.00094	0.00287
0.0183	26.38	0.00125	0.00017
0.0126	31.72	0.00373	0.01856
0.0057	47.41	0.00090	0.00002
0.0046	52.52	0.04206	0.07288
0.0007	136.33	0.94916	0.90463

Πίνακας 3. 36 *Πίνακας αποτελεσμάτων διαγνωστικών πολυσυγγραμμικότητας*

Η πολυσυγγραμμικότητα αναφέρεται στην υψηλή συσχέτιση μεταξύ δύο ή περισσότερων προγνωστικών μεταβλητών σε ένα στατιστικό μοντέλο. Μπορεί να οδηγήσει σε ζητήματα όπως ασταθείς εκτιμήσεις των συντελεστών παλινδρόμησης και διογκωμένα τυπικά σφάλματα, τα οποία μπορεί να δυσκολέψουν την ερμηνεία των αποτελεσμάτων της ανάλυσης. Για τον λόγω αυτό η τεχνική του κεντραρίσματος ως προς τον μέσο όρο των μεταβλητών είναι μια κοινή τεχνική που χρησιμοποιείται στη στατιστική ανάλυση για τη μείωση των επιπτώσεων της πολυσυγγραμμικότητας, η οποία συμβαίνει όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό μεταξύ τους. Όταν μια μεταβλητή κεντράρεται ως προς τον μέσο όρο της, η μέση τιμή της μεταβλητής αφαιρείται από κάθε μεμονωμένη παρατήρηση, έτσι ώστε η μεταβλητή που προκύπτει να έχει μέση τιμή μηδέν. Αυτό έχει ως αποτέλεσμα την άρση τυχόν σταθερών διαφορών μεταξύ των παρατηρήσεων, διατηρώντας παράλληλα τις σχετικές διαφορές μεταξύ τους. Μέσω το κεντράρισμα των μεταβλητών, οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών μειώνονται, γεγονός που μπορεί να βοηθήσει στη ‘’θεραπεία’’ της πολυσυγγραμμικότητας. Αυτό συμβαίνει επειδή η πολυσυγγραμμικότητα προκύπτει όταν δύο

ή περισσότερες ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό, γεγονός που μπορεί να οδηγήσει σε ασταθείς εκτιμήσεις των συντελεστών παλινδρόμησης και διογκωμένα τυπικά σφάλματα. Το κεντράρισμα των μεταβλητών ως προς τη μέση τιμή τους μπορεί να είναι ιδιαίτερα χρήσιμο σε καταστάσεις όπου οι ανεξάρτητες μεταβλητές έχουν διαφορετικές μονάδες μέτρησης ή όταν η κλίμακα μιας μεταβλητής είναι πολύ μεγαλύτερη από τις άλλες. Αυτό οφείλεται στο ότι η τεχνική αυτή επιτρέπει μια πιο ουσιαστική ερμηνεία των συντελεστών παλινδρόμησης, καθώς αντιπροσωπεύουν το αποτέλεσμα μιας αλλαγής μιας μονάδας στην ανεξάρτητη μεταβλητή όταν όλες οι άλλες μεταβλητές διατηρούνται σταθερές στις μέσες τιμές τους.

Συνοπτικά, οι κεντραρίσματος μεταβλητές ως προς τη μέση τιμή της μπορούν να είναι μια χρήσιμη τεχνική για τη μείωση των επιπτώσεων της πολυσυγγραμμικότητας στην ανάλυση παλινδρόμησης. Μπορεί να βοηθήσει στη σταθεροποίηση των εκτιμήσεων των συντελεστών παλινδρόμησης και στη μείωση της πιθανότητας διογκωμένων τυπικών σφαλμάτων, ιδιαίτερα σε καταστάσεις όπου οι ανεξάρτητες μεταβλητές έχουν διαφορετικές μονάδες μέτρησης ή πολύ διαφορετικές κλίμακες.

Αφότου λοιπόν εφαρμόστηκε η τεχνική του κεντραρίσματος των ανεξάρτητων μεταβλητών υπολογίζονται πάλι οι δείκτες των διαγνωστικών της πολυσυγγραμμικότητας με αποτέλεσμα η ανεξάρτητη μεταβλητή η οποία αφορά το χαρακτηριστικό της μέσης μορφοκλασματικής διάστασης του κυτταρικού πυρήνα να ξεπερνά τα κρίσιμα όρια (0.8). Για τον λόγο αυτό η συγκεκριμένη ανεξάρτητη μεταβλητή αφαιρείται από το σύνολο των δεδομένων.

Eigenvalue	Condition Index	Intercept (Variance Decomposition Proportion)	fractal_dimension_mean
4.614	1.000	0.000	0.003
2.210	1.445	0.000	0.007
1.767	1.616	0.000	0.002
1.178	1.979	0.000	0.003
1.000	2.148	1.000	0.000
0.867	2.307	0.000	0.014
0.732	2.511	0.000	0.000
0.427	3.287	0.000	0.006
0.353	3.614	0.000	0.036
0.303	3.902	0.000	0.000
0.204	4.756	0.000	0.014
0.148	5.579	0.000	0.067
0.144	5.660	0.000	0.033
0.053	9.333	0.000	0.815

Πίνακας 3. 37 *Πίνακας αποτελεσμάτων διαγνωστικών πολυσυγγραμμικότητας μετά την εφαρμογή κεντραρίσματος των ανεξάρτητων μεταβλητών βάσει της μέσης τιμής τους*

Κεφάλαιο 4: Σύγκριση μεθόδων

4.1 Μεθοδολογία ανάλυσης

Στόχος του τετάρτου κεφαλαίου της εν λόγω διπλωματικής είναι αφενός η εκπαίδευση των αλγορίθμων μηχανικής μάθησης και η αξιολόγηση τους ως προς την προβλεπτική τους ικανότητα και αφετέρου η σύγκριση αυτών μεταξύ τους για την ανάδειξη του καλύτερου μοντέλου που ταξινομεί αποτελεσματικότερα τους όγκους στην σωστή κατηγορία στην οποία ανήκουν. Προτού εκπαιδευτούν οι αλγόριθμοι μηχανικής μάθησης, προηγείται η ορθολογική επιλογή των μεταβλητών που θα χρησιμοποιηθούν για την κατασκευή των υποδειγμάτων. Λόγω του προβλήματος που επιφέρει το φαινόμενο της πολυσυγγραμμικότητας, τα υποδείγματα θα χρησιμοποιήσουν τις μεταβλητές οι οποίες παρουσιάζουν συντελεστή πληθωρισμού διακύμανσης (VIF) μικρότερο από 10 καθώς και δείκτη variance decomposition proportion μικρότερο από 0.8. Οι μεταβλητές αυτές καθώς και ο συντελεστής πληθωρισμού διακύμανσης παρουσιάζονται στον παρακάτω **πίνακα (4.1)**

Μεταβλητή	VIF
radius_mean	2,75
texture_mean	2,36
smoothness_mean	4,20
compactness_mean	5,97
symmetry_mean	4,26
radius_se	2,79
texture_se	2,71
smoothness_se	2,54
concave.points_se	2,76
symmetry_se	4,51
fractal_dimension_se	4,95
symmetry_worst	9,62

Πίνακας 4.1 *Οι μεταβλητες με συντελεστή πληθωρισμού διακύμανσης (VIF) μικρότερο του 10.*

Αφού έχουν πλέον καθοριστεί οι 12 ανεξάρτητες μεταβλητές οι οποίες θα ληφθούν υπόψη για την πρόβλεψη της επικρατούσας κλάσης, γίνεται ο διαχωρισμός του συνόλου των δεδομένων σε ένα υποσύνολο εκπαίδευσης και σε ένα υποσύνολο ελέγχου. Το σύνολο εκπαίδευσης (training set) αποτελείται από το 70% του αρχικού δείγματος ενώ σύνολο ελέγχου αποτελείται

από το υπόλοιπο 30%. Το σύνολο ελέγχου (test set) χρησιμοποιείται για την εκπαίδευση και την εκμάθηση των αλγορίθμων ενώ το σύνολο ελέγχου για την αξιολόγηση της εκμάθησης των υποδειγμάτων, ενώ στη συνέχεια με τα μέτρα αξιολόγησης (accuracy, precision, recall, κ.ά.) ως εργαλεία αξιολογείται η απόδοση τους εφαρμόζοντας μια διαδικασία σύμφωνα με την οποία παραμετροποιούνται οι υπερπαράμετροι των υποδειγμάτων με διάφορες τιμές έτσι ώστε να βρεθεί ο βέλτιστος συνδυασμός κατά του οποίου μεγιστοποιείται η ικανότητα πρόβλεψης του κάθε μοντέλου.

4.2 Μοντέλο Λογιστικής Παλινδρόμησης

Αρχικά, κατασκευάζεται το υπόδειγμα του μοντέλου λογιστικής παλινδρόμησης εφόσον διασφαλιστεί ότι και στο σύνολο εκπαίδευσης αλλά και στο σύνολο ελέγχου διατηρείται το ποσοστό καλοηθών και κακοηθών όγκων. Για την επίτευξη αυτού χρησιμοποιείται η εντολή *rsample*, θέτοντας την τιμή *0,62* στο όρισμα *prop*, το οποίο ουσιαστικά αντιπροσωπεύει το ποσοστό των καλοηθών τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο ελέγχου. Με τη χρήση της συνάρτησης *glm()* και όρισμα *family = binomial*, επιτυγχάνεται η εκτέλεση της λογιστικής παλινδρόμησης έναντι κάποιου άλλου γενικευμένου γραμμικού μοντέλου. Παρακάτω στον πίνακα 4.2, παρουσιάζεται η σύνοψη των αποτελεσμάτων του υποδείγματος λογιστικής παλινδρόμησης.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.61	0.4417	-1.38	0.167439	
radius_mean	1.21	0.3388	3.581	0.000342	***
texture_mean	0.39	0.114	3.386	0.00071	***
smoothness_mean	88.67	52.076	1.703	0.088617	.
compactness_mean	25.90	21.1228	1.226	0.220089	
symmetry_mean	-56.14	25.864	-2.17	0.029972	*
radius_se	14.90	4.3981	3.387	0.000706	***
texture_se	1.49	1.0086	1.474	0.140435	
smoothness_se	136.81	242.3104	0.565	0.572346	
concave.points_se	103.54	127.9007	0.81	0.418206	
symmetry_se	-297.16	84.8068	-3.504	0.000458	***
fractal_dimension_se	-953.19	411.4208	-2.317	0.020513	*
symmetry_worst	69.64	15.81	4.41	0.00	***

Πίνακας 4.2 *Σήμανση σημαντικότητας: α = 0 '***', α = 0.001 '**', α = 0.01 '*', α = 0.05 '.', α = 0.1 ' '*

Στην πρώτη στήλη εμπεριέχονται οι ανεξάρτητες μεταβλητές ενώ η μεταβλητή intercept αντιπροσωπεύει την ελεύθερη μεταβλητή. Η στήλη *estimate*, αντιπροσωπεύει τους συντελεστές των ανεξάρτητων μεταβλητών. Ως παράδειγμα λαμβάνεται υπόψιν ο συντελεστής της μεταβλητής *radius_mean* που ισούται με 1.21. Ερμηνευτικά αυτό συμβολίζει πως για κάθε μοναδιαία αύξηση της μέσης απόστασης του όγκου από τα σημεία της περιμέτρου του, τότε η λογαριθμική απόδοση (log-odd) της μεταβλητής θα αυξηθεί κατά 1.21. Στην στήλη *Std. Error*, εμπεριέχονται οι τιμές του τυπικού σφάλματος. Στην λογιστική παλινδρόμηση, το τυπικό σφάλμα (*Standard error)* του εκτιμώμενου συντελεστή παλινδρόμησης υποδεικνύει την ακρίβεια της εκτίμησης της σχέσης μεταξύ μιας μεταβλητής πρόβλεψης και της λογαριθμικής απόδοσης. Η τιμή της *z-value,* ουσιαστικά αξιολογεί το αν χρειάζεται να απορριφθεί η μηδενική υπόθεση *(null hypothesis = H_0)*. Όσο μεγαλύτερη είναι η *z-value*, τόσο ισχυρότερη είναι η ένδειξη κατά της μηδενικής υπόθεσης. Όταν ισχύει η μηδενική υπόθεση σημαίνει πως δεν υπάρχει σχέση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής, η οποία είναι αντίθετη από την εναλλακτική υπόθεση *(alternative hypothesis = H_1)* σύμφωνα με την οποία υπάρχει κάποια σχέση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής. Στην στήλη *Pr(>|z|),* εμπεριέχονται οι συντελεστές της p-value η οποία ουσιαστικά είναι το μέτρο της σημαντικότητας μιας μεταβλητής. Έπειτα, εφαρμόζεται ο αλγόριθμος διαδικασία σταδιακής απόρριψης μεταβλητών (backward elimination). Αναλυτικότερα, στην εν λόγω διαδικασία εφόσον εισαχθούν όλες οι ανεξάρτητες μεταβλητές και σταδιακά σε κάθε διαδοχικό βήμα αφαιρείται η λιγότερο σημαντική. Η διαδικασία ολοκληρώνεται όταν απομείνουν μόνο οι σημαντικές μεταβλητές δηλαδή αυτές που συνεισφέρουν στο μοντέλο. Η σημαντικότητα κρίνεται με βάση την τιμή της p-value η οποία έχει ορισμένο επίπεδο σημαντικότητας ίσο με 0.05.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.66	0.4109	-1.601	0.109443	
radius_mean	1.37	0.3157	4.337	0.0000144	***
texture_mean	0.37	0.1094	3.401	0.000671	***
smoothness_mean	138.28	43.3255	3.192	0.001415	**
symmetry_mean	-47.55	23.6308	-2.012	0.044207	*
radius_se	14.57	4.2152	3.457	0.000546	***
texture_se	1.68	0.9112	1.846	0.064939	.
symmetry_se	-311.36	80.6923	-3.859	0.000114	***
fractal_dimension_se	-528.54	294.2412	-1.796	0.07245	.
symmetry_worst	71.88	15.1525	4.744	0.0000021	***

Πίνακας 4.3 *Σήμανση σημαντικότητας: α = 0 '***', α = 0.001 '**', α = 0.01 '*', α = 0.05 '.', α = 0.1 ' '*

Το τελικό μοντέλο λογιστικής παλινδρόμησης αποτελείται από συνολικά από εννέα ανεξάρτητες μεταβλητές οι οποίες έχουν ιδιαίτερη σημαντικότητα. Αναλυτικότερα, απαρτίζεται από την μέση απόσταση του πυρήνα σε σημεία της ακτίνας του όγκου, την μέση τιμή συμμετρίας του, την μέση τιμή απαλότητας, την μέση τιμή συμμετρίας, την τυπική απόκλιση της απόστασης του πυρήνα από σημεία της ακτίνας του, την τυπική απόκλιση της υφής του, την τυπική απόκλιση της συμμετρίας του όγκου, τη τυπική απόκλιση της μορφοκλασματική διάστασης αλλά και της διακύμανσης της συμμετρίας του. Στο **πίνακα 4.3** παρουσιάζονται οι συντελεστές της λογιστικής παλινδρόμησης και μέσα από αυτούς γίνεται αντιληπτό το πόσο ισχυρή επίδραση έχει η κάθε ανεξάρτητη μεταβλητή στην μεταβλητή απόκρισης.

Η πιο σημαντική και κατ' επέκτασιν η μεταβλητή με την μεγαλύτερη (αρνητική) επίδραση στο μοντέλο λογιστικής παλινδρόμησης είναι τυπική απόκλιση της μορφοκλασματικής διάστασης του όγκου με συντελεστή λογιστικής παλινδρόμησης -528.54, που σημαίνει ότι με μια μοναδιαία αύξηση, μειώνεται η λογαριθμική απόδοση *(log-odds)* κατά 528.54. Ακολουθεί η μεταβλητή που εκφράζει την τυπική απόκλιση της συμμετρίας με συντελεστή -311.36 και έπειτα η μέση τιμή απαλότητας με συντελεστή 138.28. Από την στιγμή που η συγκεκριμένη μεταβλητή έχει θετικό πρόσημο με μία μοναδιαία αύξηση η λογαριθμική απόδοση αυξάνεται κατά 138.28. Στη συνέχεια προηγείται η μεταβλητή που αντιπροσωπεύει τη μέση διακύμανση της συμμετρίας του όγκου με συντελεστή 71.88 και η μεταβλητή της μέσης τυπικής απόκλισης της απόστασης από τον πυρήνα σε σημεία του όγκου με συντελεστή 14.57. Έπειτα προπορεύονται οι μεταβλητές τις μέσης τυπικής απόκλισης της υφής στην κλίμακα του γκρι

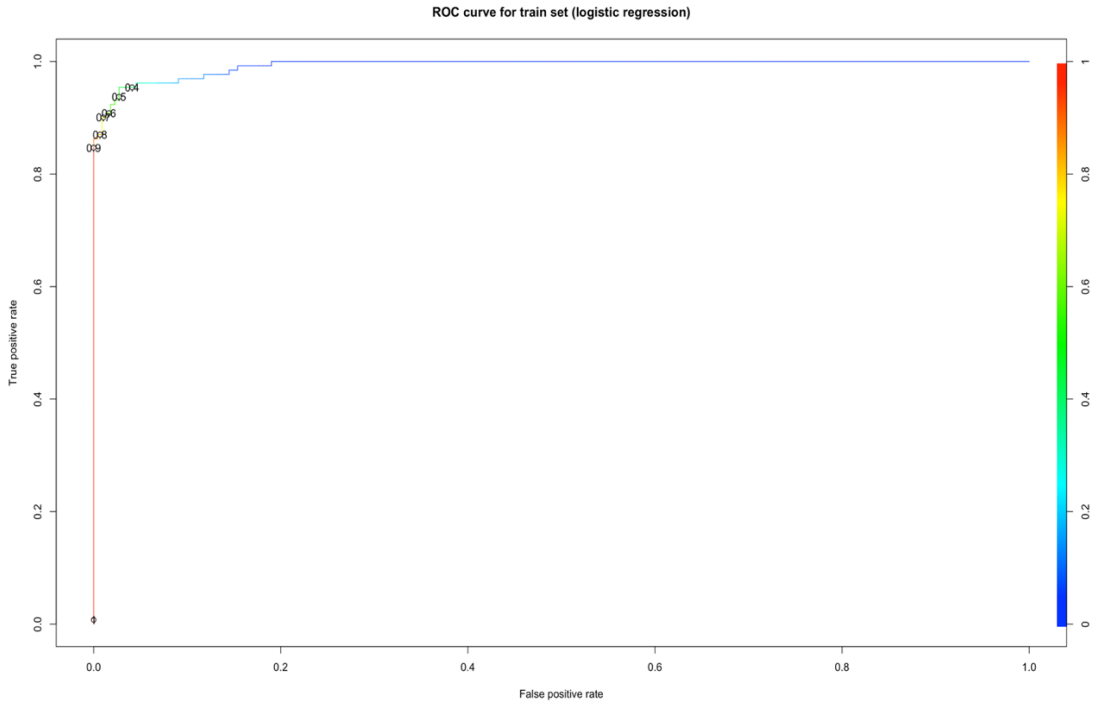
και της μέσης απόστασης του πυρήνα σε σημεία του όγκου με συντελεστές 1.68 και 1.37 αντίστοιχα. Τέλος, είναι η μεταβλητή της μέσης τιμής της υφής του όγκου στην κλίμακα του γκρι με συντελεστή λογιστικής παλινδρόμησης 0.37.

independed variable	odd ratio
radius_mean	3.36E+00
texture_mean	1.47E+00
smoothness_mean	3.23E+38
compactness_mean	1.78E+11
symmetry_mean	4.17E-25
radius_se	2.95E+06
texture_se	4.42E+00
smoothness_se	2.60E+59
concave.points_se	9.27E+44
symmetry_se	8.81E-130
fractal_dimension_se	0.00E+00
symmetry_worst	1.76E+30

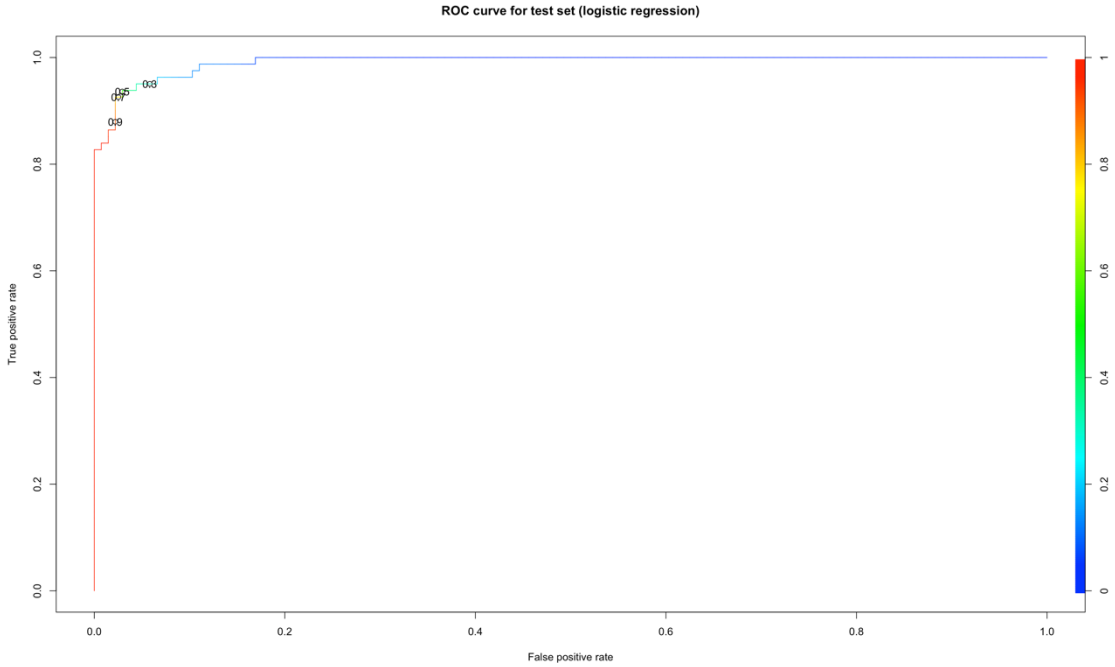
Πίνακας 4.4 *Τιμές των λόγων αποδόσεως (odd ratios) του υποδείγματος λογιστικής παλινδρόμησης*

Παραπάνω, ο **πίνακα 4.4** περιέχει τους λόγους αποδόσεως (odds ratios) της κάθε ανεξάρτητης μεταβλητής. Αναλυτικότερα, στην λογιστική παλινδρόμηση ο λόγος απόδοσης είναι το μέτρο συσχέτισης μεταξύ της απόδοσης ενός γεγονότος, δηλαδή το κατά πόσο είναι πιθανό να συμβεί ανάλογα με τον συντελεστή της εκάστοτε ανεξάρτητης μεταβλητής, δηλαδή της επίδρασης της στην έκβαση του μοντέλου. Σε ένα υπόδειγμα λογιστικής παλινδρόμησης, ο συντελεστής παλινδρόμησης β_1 είναι η εκτιμώμενη αύξηση της λογαριθμικής απόδοσης του αποτελέσματος ανά μονάδα αύξησης του συντελεστή της ανεξάρτητης μεταβλητής. Με άλλα λόγια, η εκθετική συνάρτηση του συντελεστή παλινδρόμησης ($e^{\beta 1}$) είναι ο λόγος πιθανοτήτων που σχετίζεται με μια αύξηση κατά μία μονάδα στην έκθεση. Οι λόγοι αποδόσεων χρησιμοποιούνται για τη σύγκριση των σχετικών πιθανοτήτων εμφάνισης του αποτελέσματος που μας ενδιαφέρει (στη συγκεκριμένη περίπτωση εντοπισμός καλοήθειας ή κακοήθειας), δεδομένου του συντελεστή της εκάστοτε μεταβλητής ενδιαφέροντος (π.χ. συμμετρία όγκου). Ο λόγος αποδόσεων μπορούν επίσης να ερμηνεύσουν εάν ένας συντελεστής ανεξάρτητης μεταβλητής αποτελεί κίνδυνο για ένα συγκεκριμένο αποτέλεσμα και για να συγκρίνει το μέγεθος των διαφόρων παραγόντων

κινδύνου για αυτό το αποτέλεσμα. Όταν OD (odd ratio) ισούται με 1, ο συντελεστής ανεξάρτητης μεταβλητής δεν επηρεάζει την απόδοση του αποτελέσματος. Επίσης, όταν ο OD είναι μεγαλύτερος από 1, ο συντελεστής ανεξάρτητης μεταβλητής σχετίζεται με ισχυρή επιρροή της απόδοσης να συμβεί το εκάστοτε γεγονός ενώ αντιθέτως, όταν ο OD είναι μικρότερος από 1, τότε σχετίζεται με ανίσχυρη επιρροή. Σύμφωνα λοιπόν με τον **πίνακα 4.4**, την ισχυρότερη επιρροή σύμφωνα με τον λόγο αποδόσεως στην έκβαση της καλοήθειας την ασκεί αυτός της μεταβλητής της τυπικής απόκλισης του αριθμού των κοίλων σημείων ενός όγκου (concave.points_se) ίσος με 9.27 ενώ την μικρότερη αυτός της μεταβλητής που αφορά την τυπική απόκλιση της συμμετρίας του όγκου (symmetry_se). Πριν από οποιαδήποτε πρόβλεψη της επικρατούσας κλάσης, εφόσον έχει καθοριστεί το τελικό μοντέλο λογιστικής παλινδρόμησης, χρίζει επιτακτική ανάγκη η επιλογή του κατάλληλου συνόρου απόφασης *(threshold)* σύμφωνα με το οποίο θα πραγματοποιηθεί ο διαχωρισμός των παρατηρήσεων σε κάθε κλάση. Ο προεπιλεγμένος (default) συντελεστής διαχωρισμού αποφάσεως ισούται με 0.5, που σημαίνει ότι οι τιμές μικρότερες από 0.5 θα εκχωρούνται στην κλάση καλοήθης *(“B”)*, ενώ μεγαλύτερες από 0.5 θα εκχωρούνται στης κλάση κακοήθειας *(’’M’’)*. Για την βελτίωση του ταξινομιτή ρυθμίζεται ο συντελεστής συνόρου απόφασης έτσι ώστε να δίνεται βαρύτητα στην κλάση της καλοήθειας αφού αποτελούν την πλειοψηφία των παρατηρήσεων του συνολικού δείγματος. Με την καμπύλη *ROC* ως εργαλείο αξιολόγησης, είναι εύκολο να αξιολογηθεί η ικανότητα του μοντέλου να ταξινομεί ορθολογικά της παρατηρήσεις. Στον οριζόντιο άξονα μετριέται ο λόγος ή αλλιώς το ποσοστό των παρατηρήσεων που λανθασμένα έχουν ταξινομηθεί ως θετική κλάση άρα ανήκουν αρνητική κλάση. Ενώ στον κάθετο άξονα μετριέται το ποσοστό των παρατηρήσεων οι οποίες ανήκουν στην θετική κλάση και πράγματι ανήκουν στη θετική κλάση. Το ποια κλάση θεωρείται αρνητική ή θετική καθορίζεται από τον αναλυτή. Στην συγκεκριμένη διπλωματική εργασία καθορίζεται σε όλα τα μοντέλα ως θετική κλάση η καλοήθεια. Άρα για να γίνει καλύτερα κατανοητό, ο οριζόντιοw άξονας αντιπροσωπεύει το ποσοστό των ασθενών που διαγνώστηκαν με κακοήθεια και ταξινομήθηκαν ως καλοήθεια, ενώ ο κάθετος άξονας αντιπροσωπεύει το ποσοστό των ασθενών που διαγνώστηκαν με καλοήθεια και ταξινομήθηκαν σωστά ως καλοήθεις. Παρακάτω παρουσιάζεται η καμπύλη ROC (εικόνα 4.1) του συνόλου εκπαίδευσης αναγράφοντας κάποια από τα κομβικότερα σύνορα απόφασης καθώς και για το σύνολο ελέγχου (εικόνα 4.2) . Παρατηρούμε πως το προεπιλεγμένο σύνορο απόφασης το οποίο ισούται με 0.5 καθώς και το 0.62 (event rate) θεωρούνται αξιόλογα κριτήρια για την σωστή ταξινόμηση.



Εικόνα 4. 1 *Καμπύλη ROC συνόλου εκπαίδευσης του υποδείγματος λογιστικής παλινδρόμησης*

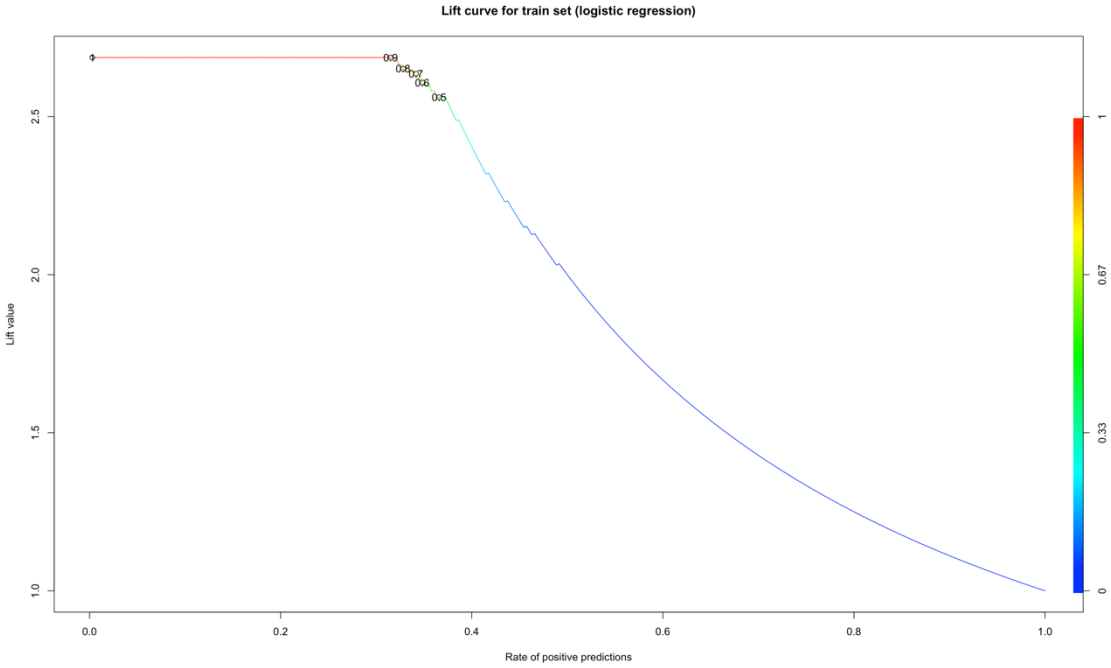


Εικόνα 4. 2 *Καμπύλη ROC συνόλου ελέγχου του υποδείγματος λογιστικής παλινδρόμησης*

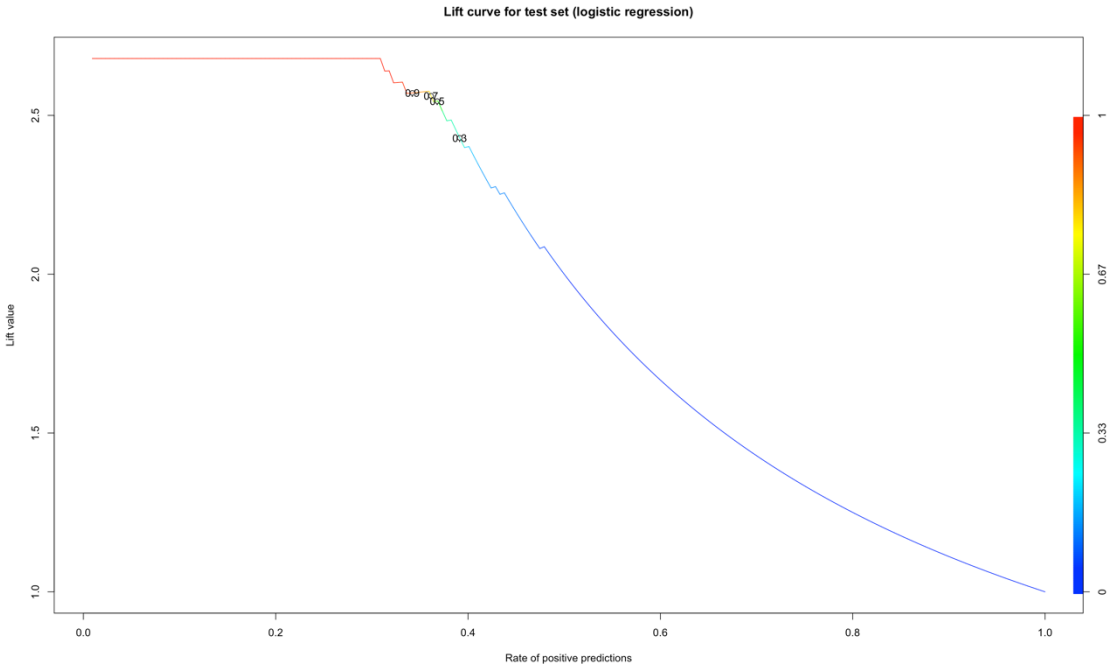
Ένα ακόμα γράφημα το οποίο αποτελεί χρήσιμο εργαλείο αξιολόγησης για ένα μοντέλο αλγορίθμου μηχανικής μάθησης είναι η καμπύλη lift βάσει της οποίας εκμαιεύονται πολύ σημαντικές πληροφορίες για την απόδοση του. Ειδικότερα, στον κάθετο άξονα μετράται η ανύψωση (lift) όπου εκφράζεται από τον μαθηματικό τύπο:

$$lift = \frac{predicted\ rate}{average\ rate} = \frac{ποσοστό\ θετικής\ κλάσης}{ποσοστό\ θετικής\ κλάσης\ στο\ σύνολο\ δεδομένων}$$

Μια υψηλή τιμή ανύψωσης σε ένα ορισμένο σημείο της καμπύλης δείχνει ότι το μοντέλο αποδίδει καλύτερα από ότι ένα με χαμηλότερο στον εντοπισμό θετικών περιπτώσεων σε αυτό το υποσύνολο του πληθυσμού. Η υψυλότερη τιμή ανύψωσης ονομάζεται σημείο μέγιστης ανύψωσης **(maximum lift point)**. Άρα, ένα σημείο στην καμπύλη ανύψωσης εκφράζει την ικανότητα πρόβλεψης της θετικής κλάσης (καλοήθεια) δεδομένου του ποσοστού της καλοήθειας στο σύνολο των δεδομένων για το εκάστοτε σύνορο απόφασης. Πιο απλά το διάγραμμα δείχνει πόσο καλά ‘’μαντεύει’’ ανάλογα με το πόσα γνωρίζει. Σύμφωνα με τις πληροφορίες που παρέχει το διάγραμμα, εύκολα αντλούνται οι πρόσθετες πληροφορίες ότι οι περισσότεροι υγιείς εξεταζόμενοι βρίσκονται αριστερά της καμπύλης ανύψωσης, ενώ όσο μεγαλύτερη είναι η ευθεία γραμμή πριν την πτώση της καμπύλης τόσο μεγαλύτερο είναι το ποσοστό που κατέχει την τιμή του μέγιστου σημείου ανύψησης. Παρατηρείται πως στην καμπύλη lift που αφορά το σύνολο εκπαίδευσης **(εικόνα 4.3),** τα σημεία που αντιστοιχούν στο σύνορο απόφασης, το οποίο ισούται με 0.6, βρίσκονται υψηλότερα έναντι αυτού που ισούται με 0.5. Αυτό σημαίνει πως το ποσοστό του δείγματος των εξεταζόμενων που αντιστοιχεί στο σύνορο απόφασης ίσο με 0.6 είναι πιο πιθανό να ανήκουν στην κλάση της καλοήθειας. Συνοψίζοντας, αν θεωρείσουμε το σημείο (0.4 , 2.5) ερμηνεύεται ως εξής. Το 40% του δείγματος των εξεταζόμενων είναι 2.5 φορές πιο πιθανό διαγνωστούν με καλοήθεια.



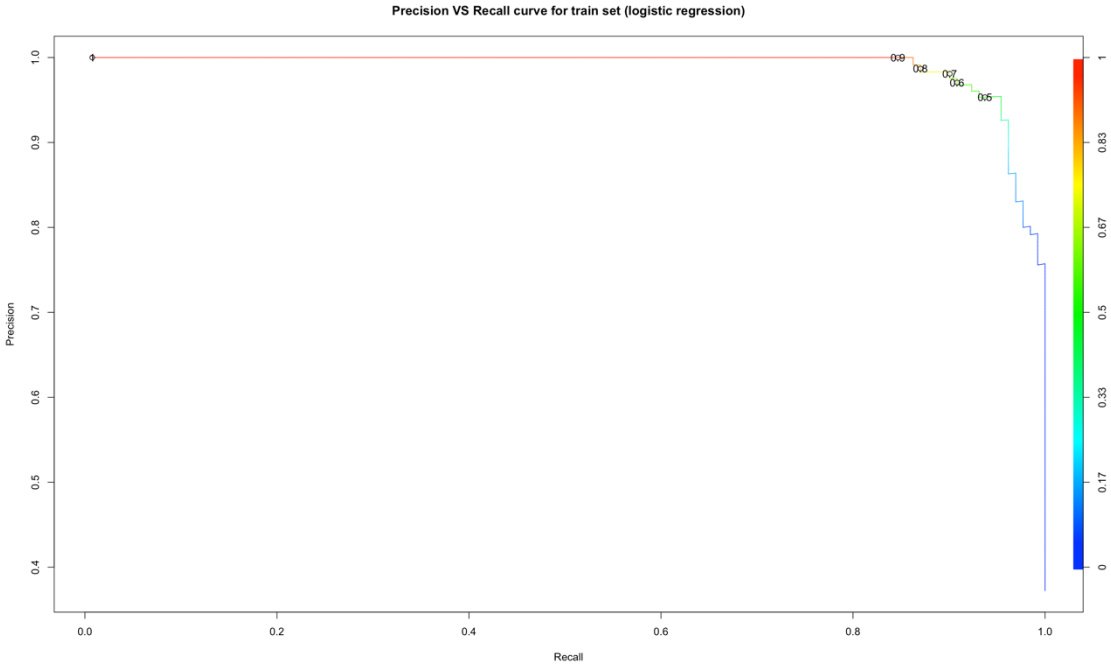
Εικόνα 4. 3 *Καμπύλη lift του συνόλου εκπαίδευσης για το υπόδειγμα λογιστικής παλινδρόμησης*



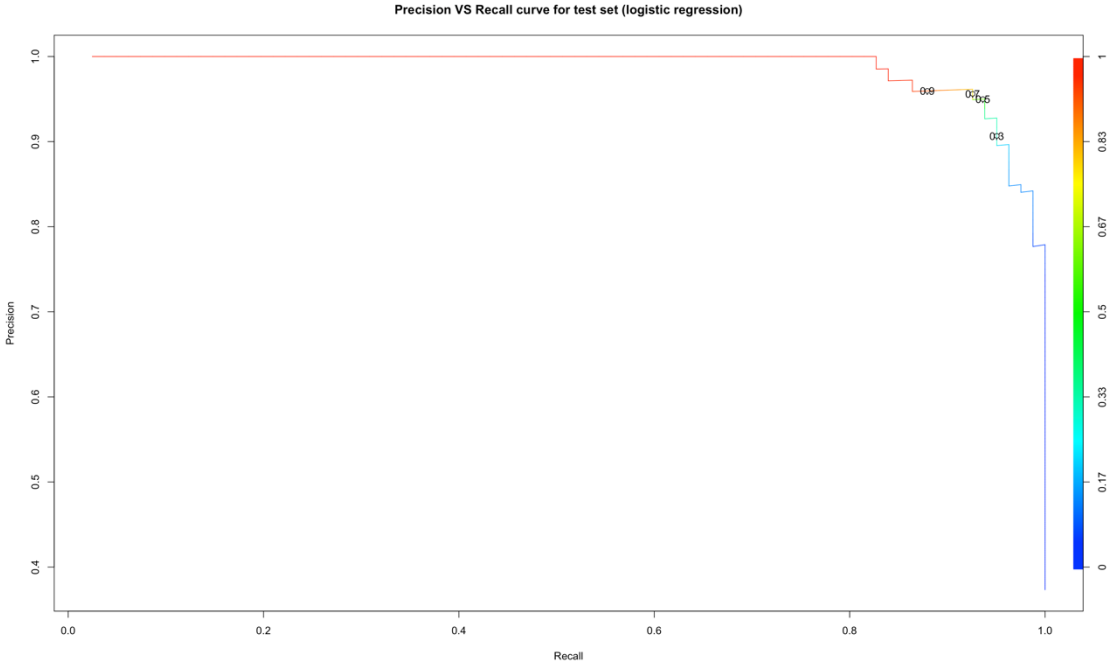
Εικόνα 4.4 *Καμπύλη lift του συνόλου ελέγχου για το υπόδειγμα λογιστικής παλινδρόμησης*

Τέλος, όσον αφορά τα γραφήματα που λαμβάνονται υπόψιν για την αξιολόγηση της απόδοσης του υποδείγματος υπάρχει το γράφημα precision vs recall. Στην **εικόνα 4.5** βρίσκεται το γράφημα precision vs recall για το σύνολο εκπαίδευσης ενώ στην **εικόνα 4.6** του συνόλου

ελέγχου. Ας υποθέσουμε ένα σημείο στην καμπύλη (0.9,0.9). Αυτό ερμηνεύεται ως ότι το 90% των ασθενών που ταξινομήθηκαν ως καλοήθεις είναι σωστά με πιθανότητα 90%.



Εικόνα 4. 5 *Καμπύλη precision vs recall συνόλου εκπαίδευσης για το υπόδειγμα λογιστικής παλινδρόμησης*



Εικόνα 4.6 *Καμπύλη precision vs recall συνόλου ελέγχου για το υπόδειγμα λογιστικής παλινδρόμησης*

Στη συνέχεια παρουσιάζεται ο **πίνακας 4.4** ο οποίος αντιπροσωπεύει τον confusion matrix του συνόλου εκπαίδευσης και για τα δύο σύνορα απόφασης. Όπου *n* ο αριθμός του δείγματος. Το άθροισμα των στηλών **B** του **πίνακα 4.5** αντιπροσωπεύει τους ασθενείς του συνόλου εκπαίδευσης που είναι πράγματι διαγνωσμένοι με καλοήθη όγκο, ενώ η στήλη **M** αντιπροσωπεύει τους ασθενείς που είναι διαγνωσμένοι με κακοήθη όγκο. Παραμένοντας στον **πίνακα 4.5**, το άθροισμα της γραμμής **predicted B**, αντιπροσωπεύει στο σύνολο εκπαίδευσης, το σύνολο των ασθενών που ταξινομήθηκαν ως καλοήθεις ενώ η γραμμή **predicted M**, το σύνολο των ασθενών που ταξινομήθηκαν ως κακοήθεις. Βάσει αυτών λοιπόν προκύπτει ότι 215 ασθενείς, για σύνορο απόφασης 0.5, με όγκο οι οποίοι ταξινομήθηκαν ως καλοήθεις ήταν πράγματι καλοήθεις ενώ 122 ασθενείς με όγκο ταξινομήθηκαν ως κακοήθεις και ήταν πράγματι κακοήθεις. Για το ίδιο σύνορο απόφασης, 6 ασθενείς με όγκο ταξινομήθηκαν ως κακοήθεις ενώ στην πραγματικότητα ήταν καλοήθεις και 9 ασθενείς με όγκο ταξινομήθηκαν ως καλοήθεις ενώ ήταν κακοήθεις.

Training set				
n = 352	Threshold = **0.5**		Threshold = **0.62**	
	B	M	B	M
Predicted B	215	9	218	12
Predicted M	6	122	3	119

Πίνακας 4. 5 *Confusion matrix του μοντέλου λογιστικής παλινδρόμησης συνόλου εκπαίδευσης του μοντέλου λογιστική παλινδρόμησης*

Μεταφέροντας στον **πίνακα 4.6**, παρουσιάζονται τα αντίστοιχα αποτελέσματα για το σύνολο ελέγχου με αριθμό δείγματος *n = 212*. Εφόσον η απόδοση και στο σύνολο εκπαίδευσης και ελέγχου είναι παρόμοια το μοντέλο δεν αντιμετωπίζει πρόβλημα υπερπροσαρμογής (overfitting).

Test set				
n = 217	Threshold = **0.5**		Threshold = **0.62**	
	B	M	B	M
Predicted B	132	6	133	6
Predicted M	4	75	3	75

Πίνακας 4. 6 *Confusion matrix του μοντέλου λογιστικής παλινδρόμησης συνόλου ελέγχου του μοντέλου λογιστική παλινδρόμησης*

Logistic Regression				
Evaluation metric	Threshold = **0.5**		Threshold = **0.62**	
	Training set	Test set	Training set	Test set
Accuracy	0.9573863	0.953917	0.9573862	0.958523
Precision	0.9598214	0.9565217	0.947826	0.956834
Recall (Sensitivity)	0.9728506	0.9705882	0.9864253	0.977941
Specificity	0.93129771	0.9259259	0.908396	0.9259259
F1 Score	0.966292	0.9635036	0.96674	0.9672723
AUC	0.99285	0.9917393	0.99285	0.9917393

Πίνακας 4. 7 *Μέτρα αξιολόγησης λογιστικής παλινδρόμησης για τα δυο σύνορα απόφασης*

Στον **πίνακα 4.7** παρουσιάζονται τα αποτελέσματα των μέτρων αξιολόγησης της λογιστικής παλινδρόμησης και για τα δύο σύνορα απόφασης. Ξεκινώντας με το σύνορο απόφασης ίσο με 0.62, παρουσιάζει ακρίβεια (accuracy) ίση με 95.85% και κατ' επέκτασιν **ποσοστό μη-ορθής ταξινόμησης** 4.15%, ποσοστό φαινομενικά καλό. Έπειτα, ο δείκτης precision, ίσος με 95.68%, υποδεικνύει ότι το συγκεκριμένο μοντέλο ταξινόμησε σωστά κατά 95.68% τους ασθενείς με καλοήθη όγκο. Επίσης, σύμφωνα με τα μέτρα recall και specificity, το υπόδειγμα πρόβλεψε σωστά την καλοήθεια που ήταν όντως καλοήθεια σε ποσοστό 97.79% και την κακοήθεια που ήταν όντως κακοήθεια με 92.59%. Να σημειωθεί ότι η ευαισθησία (sensitivity) παρουσιάζει καλυτέρα αποτελέσματα στο σύνορο απόφασης ίσο με 0.62 ενώ η ειδικότητα (specificity) σημειώνει τα ίδια αποτελέσματα και στα δύο σύνορα απόφασης. Ο δείκτης F1 Score ισούται με 96.72% ελάχιστα καλύτερο από το σύνορο 0.5 είναι σε υψηλά επίπεδα λόγω των υψηλών επιπέδων recall και precision. Τέλος, το μέτρο AUC (area under the curve), το οποίο ουσιαστικά είναι το εμβαδό την καμπύλης ROC, είναι και το πιο σημαντικό είναι ίσο και στα

δύο σύνορα απόφασης. Για 0.5 ισούται με 0.99285 και 0.9917393 για το σύνολο εκπαίδευσης και το σύνολο ελέγχου αντίστοιχα και για την τιμή 0.62 συνόρου αποφάσεως 0.99285 και 0.9917393. Όσο μεγαλύτερη είναι η τιμή του μέτρου AUC, τόσο καλύτερη επίδοση έχει και το μοντέλο. Οι τιμές για AUC παρουσιάζονται αρκετά υψηλές γεγονός που σημαίνει ότι το μοντέλο μας είναι πραγματικά αξιόπιστο.

4.3 Μοντέλο Decision Trees

Έπειτα, μετά το μοντέλο λογιστικής παλινδρόμησης ακολουθεί το υπόδειγμα δέντρων αποφάσεων (Decision Tree). Για υπόδειγμα του δέντρου απόφασης θα χρησιμοποιηθούν οι μεταβλητές του πίνακα 4.1. Για την εκπαίδευση του μοντέλου δέντρου απόφασης με το σύνολο εκπαίδευσης στο περιβάλλον της R, γίνεται χρήση της μεταβλητής ***rpart()***, ενώ για την οπτικοποίηση του μοντέλου με τη συνάρτηση ***rpart.plot()***, έχοντας ως σκοπό την πρόβλεψη της καλοήθειας.

n = 352	**Variables actually used in tree construction**			
	radius_mean	smoothness_mean	symmetry_worst	texture_mean
Root node error: 131/352 = 0.37216				

Πίνακας 4. 8 *Οι κύριες μεταβλητές βάσει των οποίων γίνεται ο διαχωρισμός των κόμβων*

Βάσει των αποτελεσμάτων που προκύπτουν στον **πίνακα 4.8**, για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν 4 κύριες μεταβλητές ως κριτήριο διαχωρισμού. Ειδικότερα, οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν για το κριτήριο διαχωρισμού είναι η μέση απόσταση του πυρήνα από σημεία της περιμέτρου του όγκου, η μέση τιμή απαλότητας του όγκου, η διακύμανση της συμμετρίας του όγκου και η μέση τιμή της υφής στην κλίμακα του γκρι. Είναι σημαντικό να σημειωθεί πως η συνάρτηση ***rpart()*** ενσωματώνει μια διαδικασία ‘’κλαδέματος’’ (pruning) ανάλογα με τον συντελεστή πολυπλοκότητας **(cp = complexity)**. Ό συντελεστής πολυπλοκότητας καθορίζει το βάθος (depth) του δέντρου, δηλαδή των αριθμό των διακλαδώσεων και το άθροισμα των καταληκτικών κόμβων. Όσο μεγαλύτερος είναι ο συντελεστής πολυπλοκότητας τόσο πιο ‘’κλαδεμένο’’ θα ναι και το δέντρο δηλαδή τόσο μικρότερο βάθος θα έχει.

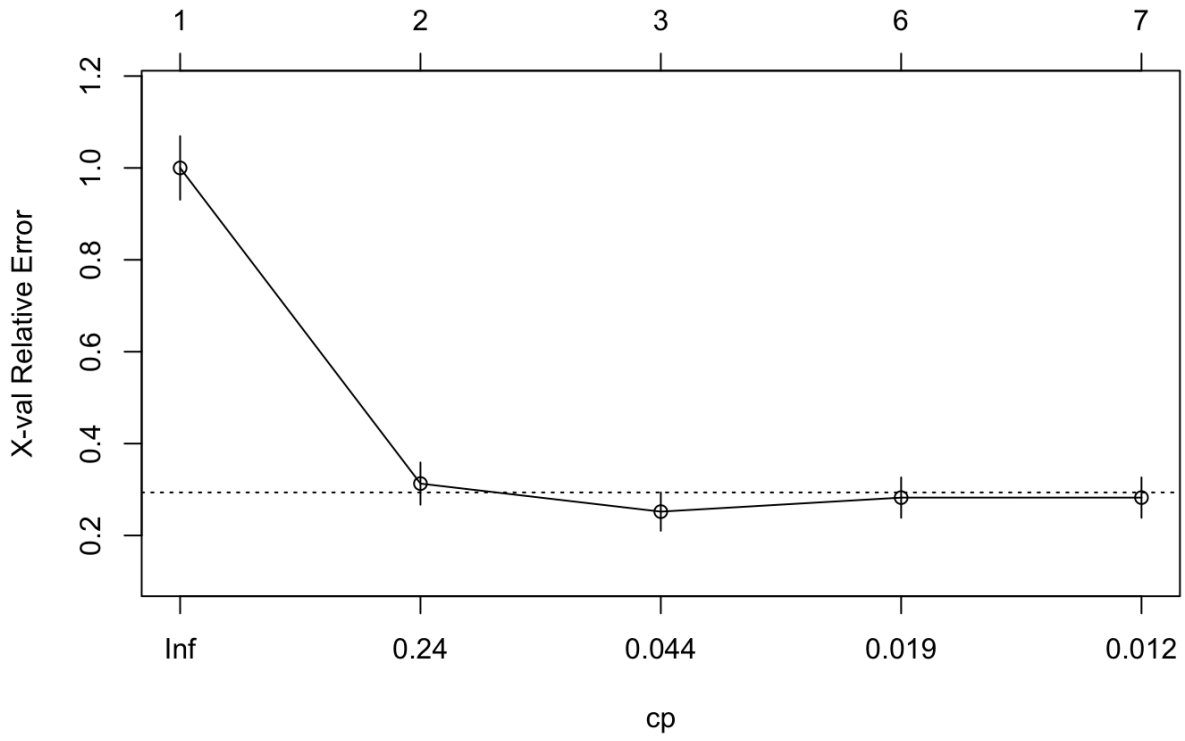
n = 352	CP	nsplit	rel error	xerror	xstd
1	0.70229	0	1	1	0.069229
2	0.083969	1	0.29771	0.31298	0.045944
3	0.022901	2	0.21374	0.25191	0.041746
4	0.015267	5	0.14504	0.28244	0.043925
5	0.01	6	0.12977	0.28244	0.043925

Πίνακας 4. 9 *Πίνακας αποτελεσμάτων διασταυρούμενης επικύρωσης*

Ερμηνεύοντας τον **πίνακα 4.9**, όπου **nsplit** (number of splits) ο αριθμός των διακλαδώσεων, **rel error** (relative error) το σχετικό σφάλμα το οποίο αφορά το σφάλμα για τις προβλέψεις των δεδομένων που χρησιμοποιήθηκαν για την εκτίμηση του μοντέλου**, xerror** το σφάλμα διασταυρωμένης επικύρωσης (cross validation). Κάθε σειρά στον **πίνακα 4.9** αντιπροσωπεύει διαφορετικό βάθος (depth) του δέντρου. Όσο μεγαλώνει ο αριθμός των διακλαδώσεων, ελαχιστοποιείται το σφάλμα ταξινόμησης στο σύνολο εκπαίδευσης με κόστος όμως να αυξάνεται ο κίνδυνος υπερπροσαρμογής (overfitting) του υποδείγματος. Το σφάλμα διασταυρωμένης επικύρωσης συνήθως αυξάνεται καθώς το δέντρο «μεγαλώνει» μετά το βέλτιστο επίπεδο. Ο εμπειρικός κανόνας είναι η επιλογή του χαμηλότερου επιπέδου βάσει της παρακάτω συνθήκης όπου:

$$rel\ error + xstd < xerror$$

Παρακάτω, στην **εικόνα 4.7** παρουσιάζεται το διάγραμμα σχετικού σφάλματος και συντελεστή πολυπλοκότητας απεικονίζοντας το πως επιδρά ο συντελεστής στο υπόδειγμα του δέντρου απόφασης.



Εικόνα 4. 7 *Διάγραμμα σχετικού σφάλματος προς συντελεστή πολυπλοκότητας*

Σύμφωνα λοιπόν με τα αποτελέσματα που εκμαιεύονται από τον **πίνακα 4.7** και την **εικόνα 4.7** η βέλτιστη επιλογή συντελεστή πολυπλοκότητας είναι ίση με 0.01 βάσει της οποίας ‘’κλαδεύεται’’ το υπόδειγμα. Η συνάρτηση ***plotcp()***, της βιβλιοθήκης ***rpart*** απεικονίζει γραφικά το πως προσαρμόζεται το υπόδειγμα δέντρου αποφάσεως με βάση τον συντελεστή πολυπλοκότητας. Δεν είναι αναγκαίο κάποιο άλλο σύνολο επικύρωσης όταν χρησιμοποιείται η συνάρτηση ***plotcp()***. Η ανάπτυξη του υποδείγματος σύμφωνα με την συνάρτηση ***rpart*** αρχικά γίνεται εφαρμόζοντας το αρχικό σύνολο δεδομένων με 57 καταληκτικούς κόμβους. Έπειτα, το δέντρο ‘’κλαδεύεται’’ στο βέλτιστο με το ελάχιστο ποσοστό λανθασμένης ταξινόμησης. Ως εκ τούτου, όταν γίνεται χρήση της συνάρτησης ***plotcp(),*** απεικονίζεται το σχετικό σφάλμα επικυρωμένης διασταύρωσης για κάθε υπό-δέντρο (subtree) από το μικρότερο στο μεγαλύτερο έτσι ώστε να συγκριθεί το ρίσκο για κάθε συντελεστή πολυπλοκότητας.

Αφότου, κλαδευτεί το αρχικό μοντέλο και προκύψει το νέο υπόδειγμα, αξιολογείται η επίδραση του κλαδέματος στο υπόδειγμα με βάση την ικανότητα πρόβλεψης του στα δεδομένα ελέγχου για τα δύο διαφορετικά σύνορα αποφάσεως (threshold). Σύμφωνα με τους πίνακες **4.10** και **4.11** προκύπτει πως το σύνορο απόφασης ίσο με 0.62 δεν έχει κάποια επίδραση και το μοντέλο αποδίδει το ίδιο.

Training set				
n = 352	Threshold = **0.5**		Threshold = **0.62**	
	Actual True	Actual False	Actual True	Actual False
Predicted True	217	13	217	13
Predicted False	4	118	4	118

Πίνακας 4. 10 *Confusion matrix μοντέλου δέντρου απόφασης συνόλου εκπαίδευσης για τον βέλτιστο συντελεστή πολυπλοκότητας ίσο με 0.01.*

Test set				
n = 217	Threshold = **0.5**		Threshold = **0.62**	
	Actual True	Actual False	Actual True	Actual False
Predicted True	128	9	128	9
Predicted False	8	72	8	72

Πίνακας 4. 11 *Confusion matrix μοντέλου δέντρου απόφασης συνόλου ελέγχου για τον βέλτιστο συντελεστή πολυπλοκότητας ίσο με 0.01.*

Ερμηνεύοντας τους πίνακες σύγχυσης (confusion matrix) και συγκεκριμένα του συνόλου εκπαίδευσης (**πίνακας 4.10)**, από τους συνολικά 221 όγκους που ανήκουν στη κλάση της καλοήθειας, οι 217 ταξινομήθηκαν σωστά ως καλοήθεις ενώ οι 4 ταξινομήθηκαν λανθασμένα ως κακοήθεις. Επίσης, από τους συνολικά 131 κακοήθεις όγκους ταξινομήθηκαν σωστά οι 118, ενώ οι 13 ταξινομήθηκαν λανθασμένα ως καλοήθεις. Για το σύνολο ελέγχου, από τους 136 συνολικά καλοήθεις όγκους ταξινομήθηκαν σωστά οι 128, ενώ οι 8 λανθασμένα ως κακοήθεις. Από τους 81 συνολικά κακοήθεις όγκους, οι 72 ταξινομήθηκαν σωστά ως κακοήθεις ενώ 9 ταξινομήθηκαν λανθασμένα ως καλοήθεις. Να σημειωθεί πως και για τα δύο σύνορα απόφασης προκύπτουν τα ίδια αποτελέσματα γεγονός που σημαίνει πως δεν επηρεάζεται η αλλαγή του συνόρου απόφασης.

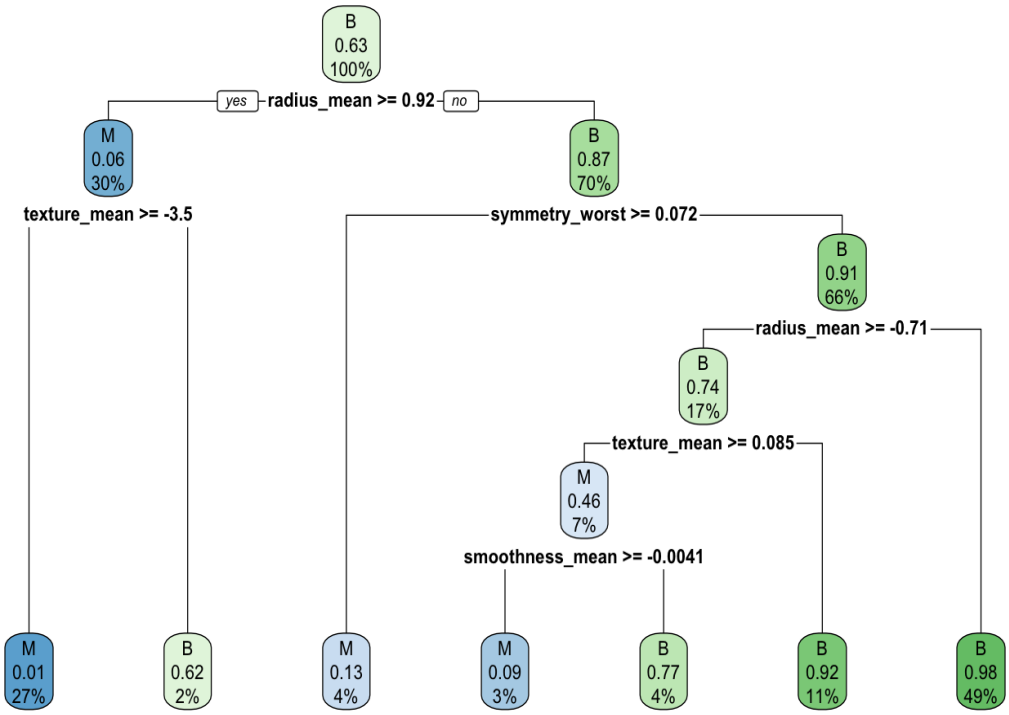
Έπειτα, στον **πίνακα 4.12**, παρουσιάζονται τα μέτρα αξιολόγησης του μοντέλου δέντρου απόφασης για τα δύο σύνορα απόφασης όπως επίσης και για τα δύο σύνολα δεδομένων, εκπαίδευσης και ελέγχου. Αναλυτικότερα, για το μέτρο accuracy στο σύνολο εκπαίδευσης το μοντέλο και για τα δύο σύνορα απόφασης ταξινομεί στη σωστή κλάση το 95.170% των παρατηρήσεων ενώ στο σύνολο ελέγχου το μοντέλο ταξινομεί το 92.170% των παρατηρήσεων.

Στο σύνολο εκπαίδευσης, και για τα δύο σύνορα απόφασης, το ποσοστό των ασθενών οι οποίοι ανήκουν στην κλάση της καλοήθειας δεδομένου ότι έχει γίνει πρόβλεψη να ανήκουν στην κλάση της καλοήθειας ισούται με 94.35%. Όσον αφορά το σύνολο ελέγχου το ποσοστό αυτό ισούται με 93.43%

Το 98.190% των ασθενών του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως υγιείς και για τα δύο σύνορα απόφασης ενώ για το σύνολο ελέγχου το 94.120%. Όσον αφορά το μέτρο specificity, δεν υπάρχει διαφορά της απόδοσης του ανάμεσα στα δύο σύνορα απόφασης. Ειδικότερα, το 90.080% των κακοηθών όγκων του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ενώ για το σύνολο ελέγχου ταξινομήθηκαν σωστά σε ποσοστό 88.890%, ποσοστό όχι και τόσο ικανοποιητικό. Τέλος, το μοντέλο παρουσιάζει για το σύνολο εκπαίδευσης απόδοση για τα μέτρα F1-score και AUC ίση με 0.96232 και 0.97171 αντίστοιχα ενώ για το σύνολο ελέγχου 0.93773 και 0.94794.

Decision tree				
Evaluation metric	Threshold = **0.5**		Threshold = **0.62**	
	Training set	Test set	Training set	Test set
Accuracy	0.95170	0.92170	0.95170	0.92170
Precision	0.94350	0.93430	0.94350	0.93430
Recall (Sensitivity)	0.98190	0.94120	0.98190	0.94120
Specificity	0.90080	0.88890	0.90080	0.88890
F1 Score	0.96232	0.93773	0.96232	0.93774
AUC	0.97171	0.94794	0.97171	0.94794

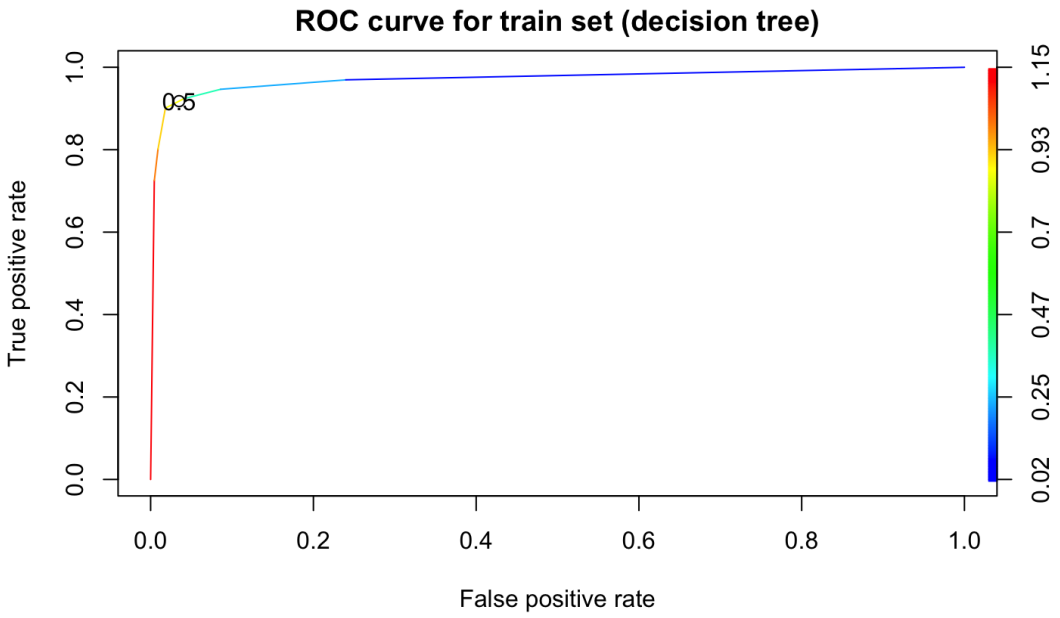
Πίνακας 4. 12 *Μέτρα αξιολόγησης του ''κλαδεμένου'' δέντρου απόφασης με συντελεστή πολυπλοκότητας ίσο με 0.01.*



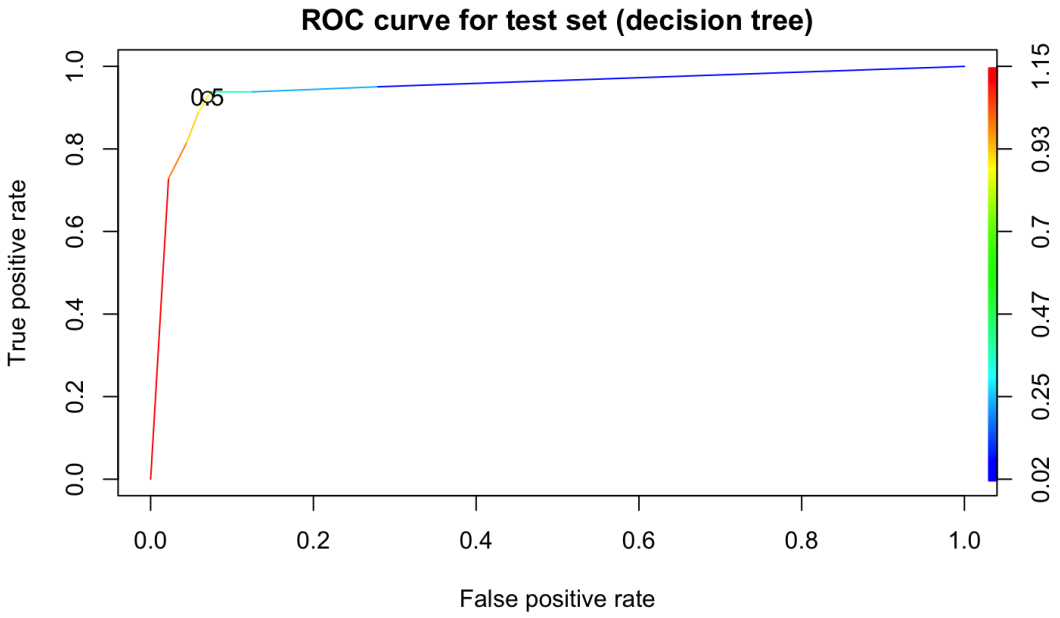
Εικόνα 4. 8 *Οπτικοποίηση του αλγορίθμου δέντρου απόφασης.*

Στην **εικόνα 4.8**, αναπαρίσταται το βέλτιστα ‘’κλαδεμένο’’ δέντρο απόφασης, το οποίο αποτελείται από 4 ανεξάρτητες μεταβλητές (πίνακας 4.8), 13 κόμβους και 6 διακλαδώσεις. Ξεκινώντας από τον πρώτο κόμβο, έχοντας το 100% του δείγματος του συνόλου εκπαίδευσης. Το δείγμα χωρίζεται βάσει της συνθήκης ότι η μέση απόσταση του πυρήνα από σημεία της περιμέτρου του όγκου να είναι μεγαλύτερη ή ίση από 0.92. Σε περίπτωση που αυτή η συνθήκη είναι αληθής, τότε το δείγμα χωρίζεται σε 70-30%, όπου το 30% του δείγματος επιβεβαιώνει την συνθήκη απόφασης με πιθανότητα 0.63. Δηλαδή κατά 63% ο ασθενής με μέση απόσταση του πυρήνα από σημεία του όγκου του μεγαλύτερη ή ίση από 0.92, ταξινομείται ως κακοήθης, ενώ με πιθανότητα κατά 8% ως καλοήθης.

Παρακάτω παρουσιάζεται η καμπύλη ROC για το τελικό, ‘’κλαδεμένο’’ υπόδειγμα από το οποίο επιβεβαιώνεται γραφικά ότι το σύνορο απόφασης 0.62 έναντι του 0.5 δεν παρουσιάζει παραπάνω απόδοση καθώς και την αναμενόμενη υπεροχή στην απόδοση της καμπύλης του συνόλου εκπαίδευσης έναντι του συνόλου ελέγχου.



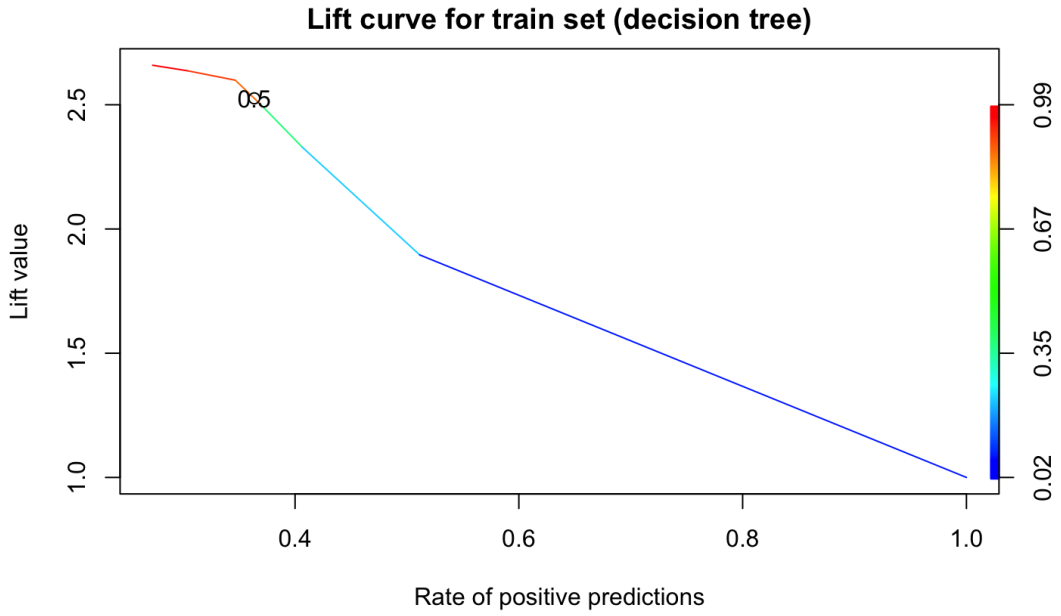
Εικόνα 4.9 *Καμπύλη ROC του συνόλου εκπαίδευσης για το υπόδειγμα δέντρου απόφασης.*



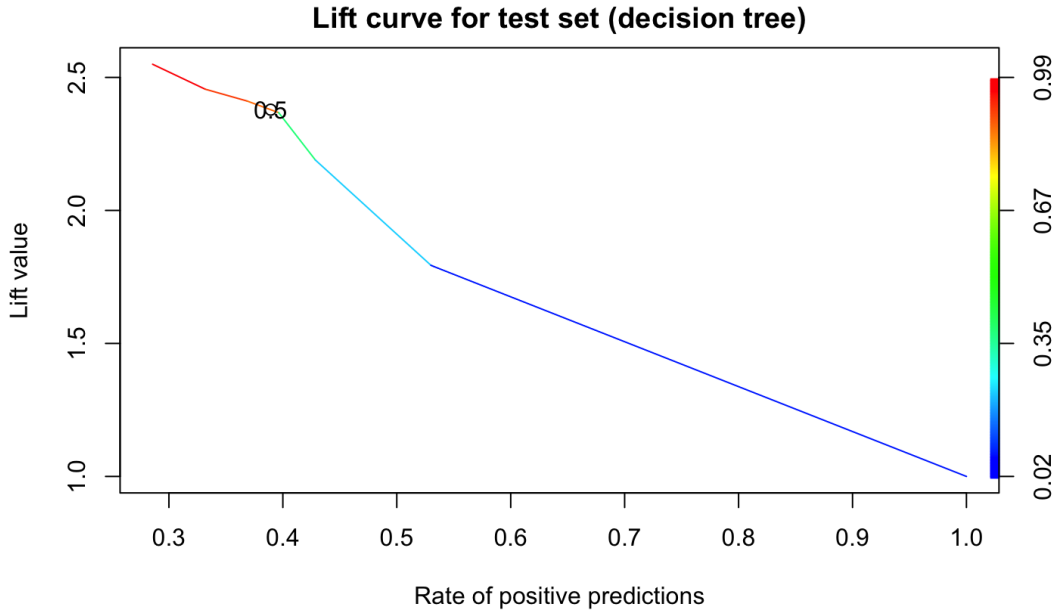
Εικόνα 4.10 *Καμπύλη ROC του συνόλου ελέγχου για το υπόδειγμα δέντρου απόφασης.*

Έπειτα, ακολουθούν οι καμπύλες lift για το σύνολο εκπαίδευσης (εικόνα 4.11) αλλά και για το σύνολο ελέγχου (εικόνα 4.12) καθώς και οι καμπύλες precision vs recall για τα δύο σύνολα

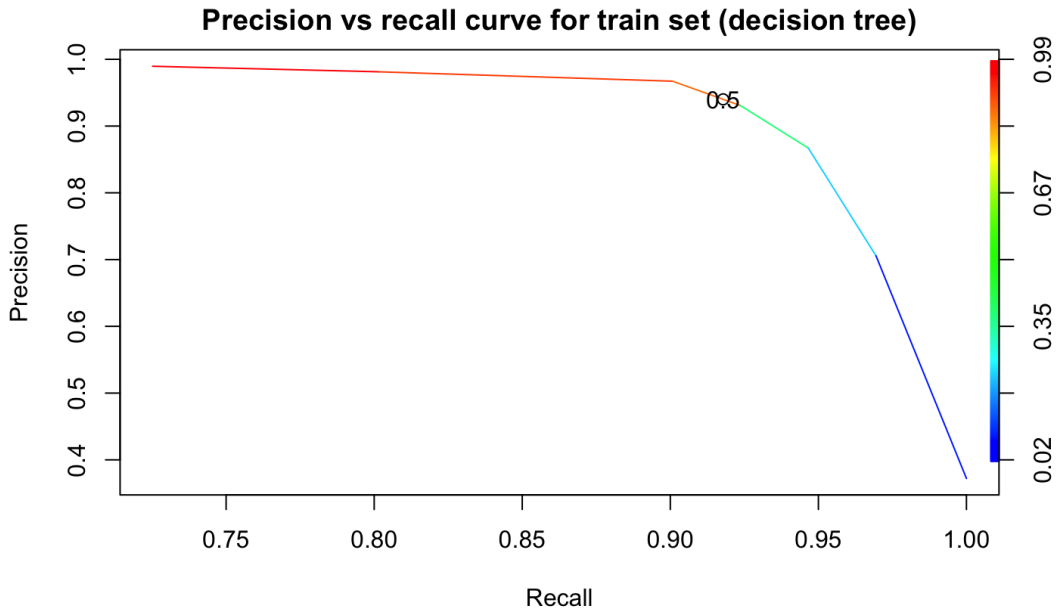
αντίστοιχα (εικόνα 4.13 και εικόνα 4.14). Διακρίνεται στην εικόνα 4.12 το ποσοστό του δείγματος το οποίο ανήκει στο σημείο μέγιστης ανύψωσης είναι ιδιαίτερα χαμηλό καθώς η πτώση είναι σχεδόν ακαριαία και ιδιαίτερα απότομη πράγμα που σημαίνει ότι όσο μειώνεται ο συντελεστής τόσο μικρότερη είναι η απόδοση να ανήκουν οι εξεταζόμενοι που ταξινομήθηκαν ως καλοήθεις στην κλάση της καλοήθειας.



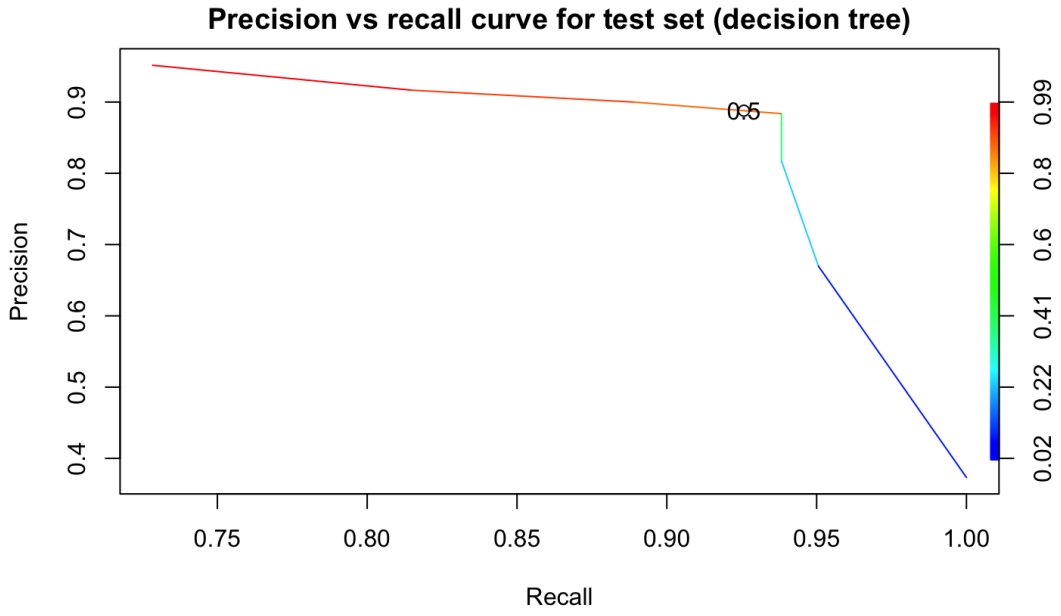
Εικόνα 4.11 *Καμπύλη Lift του συνόλου εκπαίδευσης για το υπόδειγμα δέντρου απόφασης*



Εικόνα 4.12 *Καμπύλη Lift του συνόλου ελέγχου για το υπόδειγμα δέντρου απόφασης*



Εικόνα 4.13 *Καμπύλη precision vs recall του συνόλου εκπαίδευσης για το υπόδειγμα δέντρου απόφασης*



Εικόνα 4.14 *Καμπύλη precision vs recall του συνόλου ελέγχου για το υπόδειγμα δέντρου απόφασης*

4.4 Μοντέλο Random Forest.

Επόμενος αλγόριθμος μηχανικής μάθησης ο οποίος θα εφαρμοστεί για την εκπαίδευση μοντέλου με σκοπό την πρόβλεψη της καλοήθειας είναι τα Τυχαία Δάση (Random Forest). Οι ανεξάρτητες μεταβλητές οι οποίες θα χρησιμοποιηθούν για αυτόν τον σκοπό είναι οι ίδιες που χρησιμοποιήθηκαν και στα προηγούμενα υποδείγματα (logistic regression και decision tree) και εμπεριέχονται στον **πίνακα 4.1**. Για την εκπαίδευση λοιπόν του μοντέλου αυτού στο περιβάλλον της R, γίνεται χρήση η συνάρτηση ***randomForest()*** στο σύνολο εκπαίδευσης. Το μοντέλο αυτό προκύπτει από την κατασκευή πολλαπλών δέντρων αποφάσεων σύμφωνα με κάποιες συγκεκριμένες υπερπαράμετροι οι οποίες καθορίζουν και με το τι κριτήρια θα πραγματοποιηθεί αυτή η διαδικασία. Για παράδειγμα η υπερπαράμετρος **ntree**, καθορίζει πόσα δέντρα θα κατασκευαστούν ενώ η **mtry** αντιπροσωπεύει τον αριθμό των δυνητικών ανεξάρτητων μεταβλητών από τις οποίες θα επιλεγεί η βέλτιστη από τον αλγόριθμο για τη δημιουργία κόμβου. Τα συνοπτικά αποτελέσματα του αλγορίθμου παρουσιάζονται παρακάτω στον **πίνακα 4.13**.

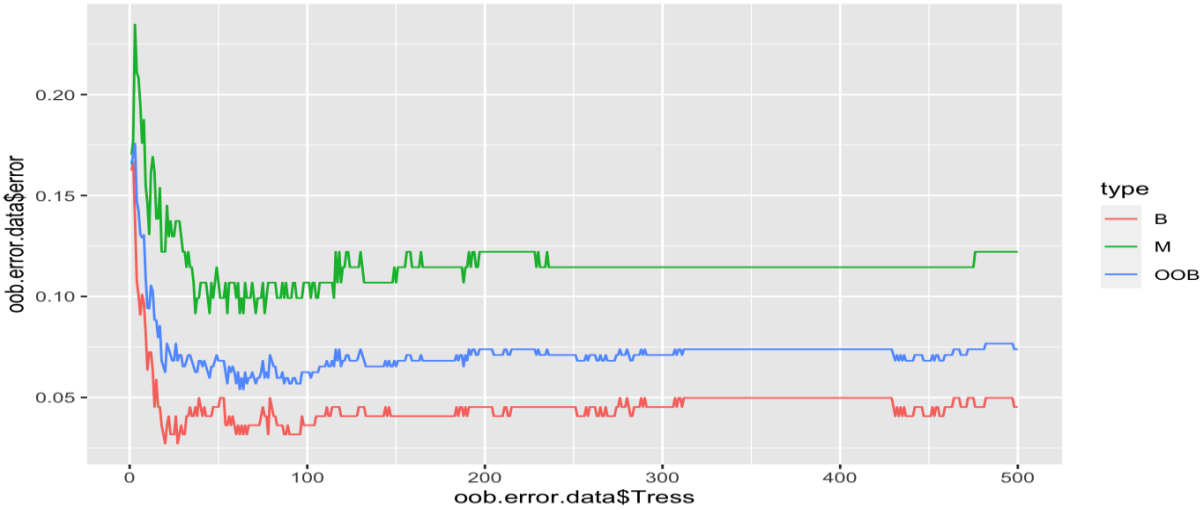
Type of random forest: Classification		
Number of trees	500	
No. of variables tried at each split:	3	
OOB estimate of error rate:	7.39%	
Confusion matrix:	Actual B	Actual M
Predicted B	212	17
Predicted M	9	114
Class.error	0.0407	0.1297

Πίνακας 4. 13 *Συνοπτικά αποτελέσματα του μοντέλου Random Forest για τις προεπιλεγμένες τιμές των υπερπαραμέτρων.*

Βάσει των αποτελεσμάτων που εκμαιεύονται από τον **πίνακα 4.13**, ο αριθμός των δέντρων που προέκυψε για την εφαρμογή του αλγορίθμου είναι ίσος με 500, καθώς και ο αριθμός των μεταβλητών σε κάθε κόμβο απόφασης για τον διαχωρισμό ισούται με 3. Το ‘’Out of bag error rate’’ η αλλιώς OOB estimate of error rate ισούται με 7.39%, ποσοστό σχετικά ικανοποιητικό για την απόδοσης του μοντέλου στα δεδομένα εκπαίδευσης. Το συγκεκριμένο ποσοστό αντιπροσωπεύει ουσιαστικά το με μέσο σφάλμα που προκύπτει από τα 500 δέντρα που

εκπαιδεύτηκαν. Ακόμη, παρατηρείται πως το μοντέλο τείνει να προβλέπει αποδοτικότερα την κλάση της καλοήθειας (true class) έναντι της κακοήθειας εφόσον το σφάλμα ταξινόμησης της θετικής κλάσης (B) είναι μικρότερο από αυτό της αρνητικής κλάσης (B). Η απόδοση του συγκεκριμένου μοντέλου ενδεχομένως να έχει πιθανά περιθώρια βελτίωσης καθώς τα αποτελέσματα τα οποία προέκυψαν στον **πίνακα 4.13**, αφορούν το μοντέλο με τις προεπιλεγμένες τιμές των υπερπαραμέτρων του. Για να εξεταστεί λοιπόν το ενδεχόμενο αυτό εφαρμόζεται μια διαδικασία αναζήτησης μέσω πλέγματος (Grid Search). Η αναζήτηση πλέγματος αποτελεί μια διαδικασία η οποία εφαρμόζεται κατά κόρον για την αναζήτηση των βέλτιστων τιμών των υπερπαραμέτρων. Στην ουσία αυτό που λαμβάνει χώρα στην εν λόγω διαδικασία είναι η αντικατάσταση τυχαίων μεταβλητών στο μοντέλο έτσι ώστε να συγκριθούν μεταξύ τους τα μοντέλα βάσει της απόδοσή τους και ως εκ τούτου να επιλεγεί το βέλτιστο. Το κριτήριο το οποίο θα αποτελέσει βασικό πυλώνα για την επιλογή του βέλτιστου μοντέλου είναι μια διαδικασία βαθμολόγησης των μέτρων απόδοσης. Αναλυτικότερα, για κάθε υπερπαράμετρο θα υπολογιστεί η απόδοση των μοντέλων για κάθε υποψήφια τιμή ως προς τον δείκτη AUC, και έπειτα θα επιλεγεί αυτή που μεγιστοποιεί τον δείκτη καθώς και οι γειτονικές της τιμές. Στη συνέχεια θα αναπτυχθεί μια διαδικασία βαθμολόγησης (Ranking) για το μοντέλο με την τιμή της υπερπαραμέτρου που μεγιστοποιεί τον δείκτη AUC, αλλά και αυτών με τις γειτονικές τιμές έτσι ώστε να συγκριθούν ως προς την απόδοση τους στα υπόλοιπα μέτρα αξιολόγησης.

Στο τέλος, η ίδια διαδικασία θα εφαρμοστεί και για τον συνδυασμό των εναπομεινάντων τιμών οι οποίες σύμφωνα με τα αποτελέσματα των παραπάνω διαδικασιών είναι πιθανότερο να βελτιστοποιούν το μοντέλο τυχαίου δάσους και ως εκ τούτου να καθοριστεί ο τελικός συνδυασμός. Παρακάτω, στην **εικόνα 4.15**, αναπαρίσταται το σφάλμα ταξινόμησης συναρτήσει του αριθμού που λαμβάνει η υπερπαράμετρος **ntree**. Λαμβάνοντας υπόψη ότι η προεπιλεγμένη τιμή για το πλήθος που χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου είναι 500, για την αναζήτηση του πλέγματος επιλέγονται οι τιμές.



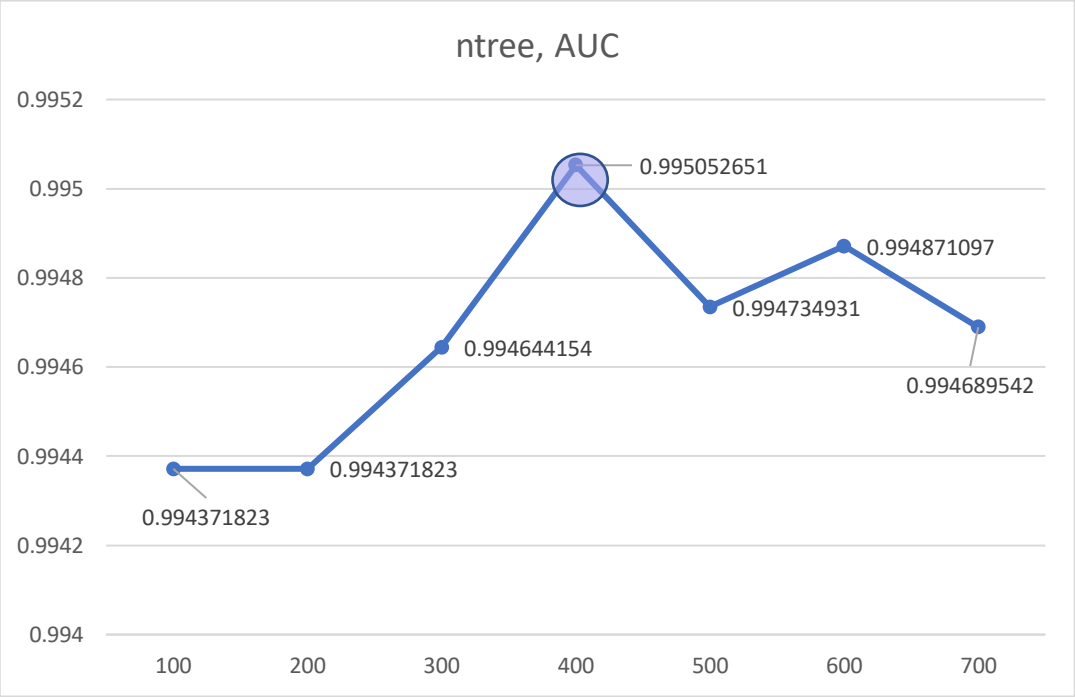
Εικόνα 4.15 *Γραφική αναπαράσταση των error rates συναρτήσει του πλήθους των δέντρων για το μοντέλο με τις προεπιλεγμένες τιμές των υπερπαραμέτρων του.*

Ξεκινώντας λοιπόν με την υπερπαράμετρο ntree οι οποία αντιπροσωπεύει τον αριθμό των δέντρων που κατασκευάζονται από τον αλγόριθμο. Στον **πίνακα 4.14** παρουσιάζονται τα αποτελέσματα των μέτρων αξιολόγησης από τη διαδικασία αναζήτησης πλέγματος για την εύρεση και επιλογή της βέλτιστης τιμής για την υπερπαραμέτρου ntree. Αποδεικνύεται ότι το μοντέλο τυχαίου δάσους παρουσιάζει την μέγιστη απόδοση για την τιμή ntree ίση με 400.

Hyperparameter: ntree						
ntree	AUC	accuracy	precision	sensitivity	specificity	F1-score
100	0.99437	0.95392	0.97015	0.95588	0.95062	0.96296
200	0.99437	0.94931	0.96992	0.95588	0.95062	0.95911
300	0.99464	0.95392	0.97727	0.94853	0.96296	0.96269
400	0.99505	0.95392	0.97727	0.94853	0.96296	0.96269
500	0.99473	0.95853	0.97744	0.94853	0.96296	0.96654
600	0.99487	0.96774	0.97778	0.94853	0.96296	0.97417
700	0.99469	0.96313	0.97761	0.94853	0.96296	0.97037
Best AUC performance = 0.99505, ntree = 400						

Πίνακας 4. 14 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο ntree.*

Η πορεία του μέτρου AUC φαίνεται παρακάτω στην **εικόνα 4.16** όπου αποδεικνύεται πως για ntree ίσο με 400 έχουμε την μέγιστη απόδοση του μέτρου.



Εικόνα 4.16 *Διάγραμμα απεικόνισης του μέτρου AUC συναρτήσει της τιμής ntree.*

Εκτός της τιμής 400 επιλέγουμε και τις γειτονικές δηλαδή τις τιμές 300, 500 και ως εκ τούτου πραγματοποιείται μια διαδικασία βαθμολόγησης της απόδοσης του μοντέλου για κάθε μέτρο αξιολόγησης αποτελέσματα της οποίας αναγράφονται στον πίνακα 4.15. Προκύπτει πως το μοντέλο random forest για ntree ίσο με 500 κατέχει την πρώτη θέση στη βαθμολογία. Αναλυτικότερα, βρίσκεται στη δεύτερη θέση στο μέτρο AUC, ενώ στα υπόλοιπα βρίσκεται στην πρώτη θέση σημειώνοντας μεγαλύτερη απόδοση σε σχέση με τα υπόλοιπα μοντέλα.

Hyperparameter: ntree						
mtry	auc	accuracy	precision	sensitivity	specificity	F1-score
300	0.99464	0.95392	0.97727	0.94853	0.96296	0.96269
400	0.99505	0.95392	0.97727	0.94853	0.96296	0.96269
500	0.99473	0.95853	0.97744	0.94853	0.96296	0.96654
Ranking						
mtry	auc	accuracy	precision	sensitivity	specificity	F1-score
300	3	2	2	1	1	2
400	1	2	2	1	1	2
500	2	1	1	1	1	1

Πίνακας 4. 15 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για το μοντέλο random forest για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο ntree*

Στον πίνακα 4.16 φαίνονται τα τελικά αποτελέσματα έχοντας το μοντέλο για ntree ίσο με 500 στην πρώτη θέση με μέση βαθμολογία 1.17, το μοντέλο για ntree ίσο με 400 με βαθμολογία 1.5 στη δεύτερη θέση και τέλος το μοντέλο για ntree ίσο με 300 και βαθμολογία 1.83 στην τρίτη θέση.

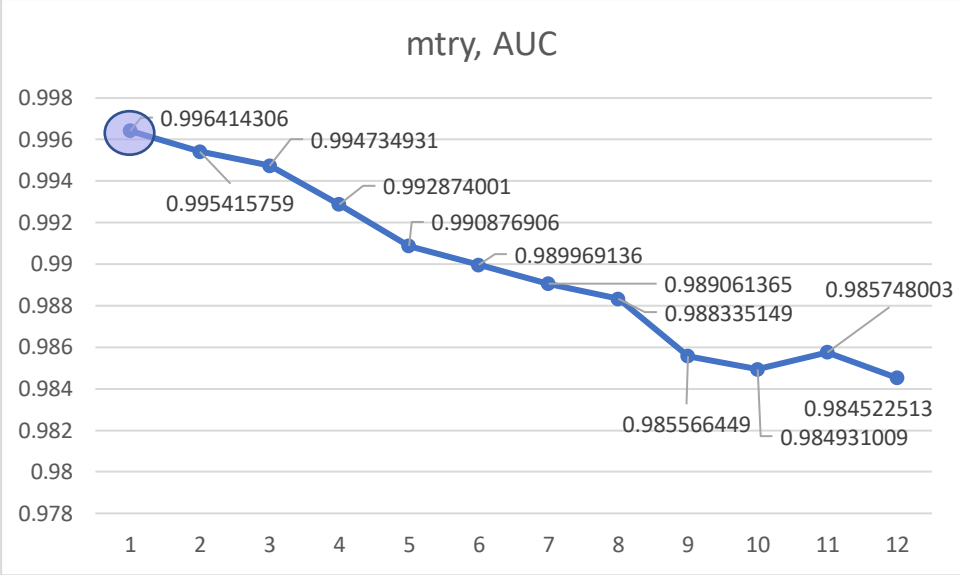
ntree	Average rank
300	1.83
400	1.50
500	1.17
Best rank = 1.17, ntree = 500	

Πίνακας 4. 16 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο ntree*

Στην συνέχεια πραγματοποιείται η ίδια διαδικασία για την υπερπαράμετρο mtry η οποία εκφράζει των αριθμό των ανεξάρτητων μεταβλητών που λαμβάνονται υπόψη από το μοντέλο τυχαίου δάσους ως κριτήριο διαχωρισμού σε κάθε κόμβο. Εξ ορισμού, το μοντέλο έχει ως προεπιλεγμένη τιμή για την υπερπαράμετρο mtry η οποία προκύπτει από τον μαθηματικό τύπο $mtry = \sqrt{p} = \sqrt{12} \approx 3$. Για τον λόγω του ότι η πορεία του δείκτη AUC διαγραμματικά είναι καθοδική (εικόνα 4.18), θεωρείται ορθολογική απόφαση να εξεταστεί ο συνολικός αριθμός των ανεξάρτητων μεταβλητών. Ο **πίνακα 4.17** περιέχει τα αποτελέσματα των μέτρων αξιολόγησης του αλγορίθμου για όλο το πλέγμα των τιμών της υπερπαραμέτρου mtry καθώς και η γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει της mtry στην εικόνα 4.17.

Hyperparameter: mtry						
mtry	auc	accuracy	precision	sensitivity	specificity	F1_score
1	0.996414	0.963134	0.977612	0.963235	0.962963	0.970370
2	0.995416	0.963134	0.977612	0.963235	0.962963	0.970370
3	0.994735	0.958525	0.977444	0.963235	0.962963	0.966543
4	0.992874	0.953917	0.970149	0.963235	0.950617	0.962963
5	0.990877	0.953917	0.970149	0.955882	0.950617	0.962963
6	0.989969	0.949309	0.962963	0.955882	0.938272	0.959410
7	0.989061	0.949309	0.962963	0.955882	0.938272	0.959410
8	0.988335	0.949309	0.962963	0.955882	0.938272	0.959410
9	0.985566	0.949309	0.962963	0.955882	0.938272	0.959410
10	0.984931	0.949309	0.962963	0.955882	0.938272	0.959410
11	0.985748	0.949309	0.962963	0.955882	0.938272	0.959410
12	0.984523	0.944700	0.955882	0.955882	0.925926	0.955882
Best AUC performance = 0.9964143 , mtry = 1						

Πίνακας 4. 17 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο mtry.*



Εικόνα 4.17 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει της mtry*

Βάσει των αποτελεσμάτων που προκύπτουν από την αναζήτηση πλέγματος για την υπερπαράμετρο mtry, προκύπτει πως για την τιμή ίση με 1, το μοντέλο παρουσιάζει την μεγαλύτερη απόδοση στο μέτρο AUC. Γι' αυτό λοιπόν επιλέγεται η τιμή 1 όπως επίσης και οι γειτονικές της 2 και 3 για να έτσι ώστε να συγκριθούν τα μοντέλα random forest μεταξύ τους βάσει την μέγιστη μέση απόδοση στα μέτρα αξιολόγησης. Τα αποτελέσματα της διαδικασίας αυτής παρουσιάζονται στον **πίνακα 4.18**.

Hyperparameter: mtry						
mtry	AUC	accuracy	precision	sensitivity	specificity	F1_score
1	0.996414	0.963134	0.977612	0.963235	0.962963	0.970370
2	0.995416	0.963134	0.977612	0.963235	0.962963	0.970370
3	0.994735	0.958525	0.977444	0.963235	0.962963	0.966543
Ranking						
mtry	AUC	accuracy	precision	sensitivity	specificity	F1_score
1	1	1	1	1	1	1
2	2	1	1	1	1	1
3	3	3	3	1	1	3

Πίνακας 4. 18 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για το μοντέλο random forest για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο mtry*

Αποδεικνύεται λοιπόν, βάσει του πίνακα 4.19, πως το μοντέλο το μοντέλο με mtry ίσο με 1 υπερέχει σε σχέση με τα υπόλοιπα, συγκεντρώνοντας βαθμολογία ίση με 1 κατέχοντας την πρώτη θέση σε όλα τα μέτρα αξιολόγησης. Έπειτα ακολουθεί το μοντέλο με mtry ίσο με 2 και βαθμολογία 1.17 και στην τελευταία θέση το μοντέλο με mtry ίσο με 3 και βαθμολογία 2.33.

mtry	Average rank
1	1.00
2	1.17
3	2.33
Best rank = 1 , mtry = 1	

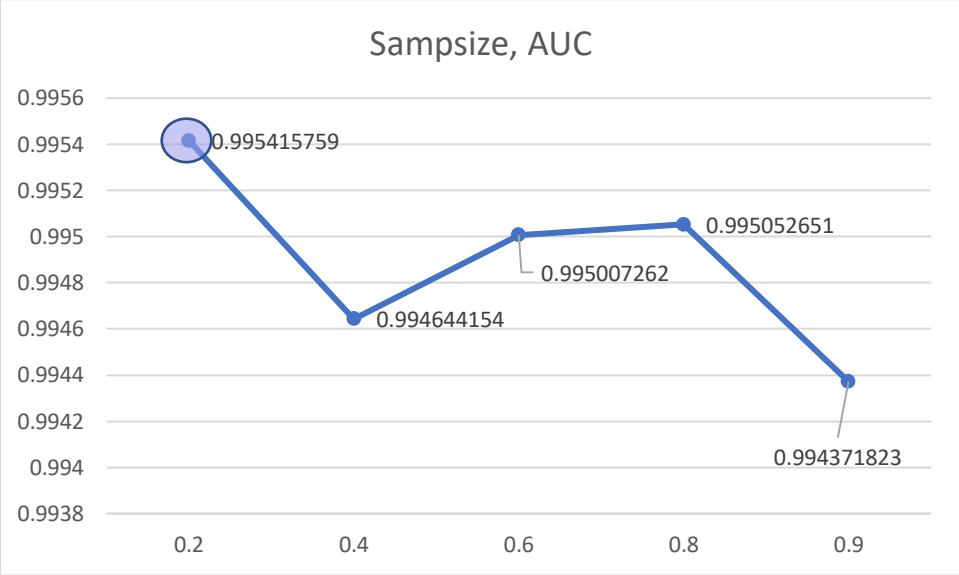
Πίνακας 4. 19 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο mtry*

Επόμενη υπερπαράμετρος στην οποία θα εφαρμοστεί η διαδικασία αναζήτησης μέσω πλέγματος είναι η **sampsize**, η οποία εκφράζει το ποσοστό του δείγματος το οποίο λαμβάνεται υπόψη για την ανάπτυξη του μοντέλου τυχαίου δάσους. Οι υποψήφιες τιμές της sampsize είναι 0.2 δηλαδή το 20% του δείγματος του συνόλου εκπαίδευσης, 0.4 (40%), 0.6 (60%), 0.8 (80%) και 0.9 (90%). Στον πίνακα 4.20 παρουσιάζονται τα μέτρα αξιολόγησης των μοντέλων για τιμές της **sampsize** καθώς και η γραφική απεικόνιση του μέτρου AUC συναρτήσει αυτής (εικόνα 4.18)

Hyperparameter: sampsize						
sampsize	AUC	accuracy	precision	sensitivity	specificity	F1-score
0.2	0.995416	0.963134	0.970588	0.970588	0.950617	0.970588
0.4	0.994644	0.963134	0.970588	0.970588	0.950617	0.970588
0.6	0.995007	0.967742	0.977778	0.970588	0.962963	0.974170
0.8	0.995053	0.958525	0.970370	0.970588	0.950617	0.966790
0.9	0.994372	0.963134	0.970588	0.970588	0.950617	0.970588
Best AUC performance = 0.9954157 , sampsize = 0.2						

Πίνακας 4. 20 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρος sampsize.*

Για το 20% του δείγματος το μοντέλο random forest σημειώνει την μέγιστη τιμή του μέτρου αξιολόγησης AUC ίσο με 0.9954157. Εκτός αυτής της τιμής του sampsize λαμβάνονται υπόψη και οι γειτονικές (0.4 και 0.6) έτσι ώστε να εφαρμοστεί η διαδικασία βαθμολόγησης.



Εικόνα 4.18 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει της sampsize.*

Τα αποτελέσματα παρουσιάζονται παρακάτω στον πίνακα 4.21 βάσει του οποίου προκύπτει πως το μοντέλο με τιμή sampsize ίση με 0.2 κατέχει την πρώτη θέση σε σχέση με τα υπόλοιπα μοντέλα στο μέτρο AUC, τη δεύτερη σε accuracy και precision, την πρώτη σε ευαισθισία και τη δεύτερη σε specificity και F1-score. Παρομοίως ακολουθεί και η διαδικασία και για τα υπόλοιπα μοντέλα.

Hyperparameter: sampsize						
sampsize	AUC	accuracy	precision	sensitivity	specificity	F1-score
0.2	0.995416	0.963134	0.970588	0.970588	0.950617	0.970588
0.4	0.994644	0.963134	0.970588	0.970588	0.950617	0.970588
0.6	0.995007	0.967742	0.977778	0.970588	0.962963	0.974170
Ranking						
sampsize	AUC	accuracy	precision	sensitivity	specificity	F1-score
0.2	1	2	2	1	2	2
0.4	3	2	2	1	2	2
0.6	2	1	1	1	1	1

Πίνακας 4. 21 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για το μοντέλο random forest για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο sampsize.*

Εν τέλει το μοντέλο που σημειώνει τη μέγιστη μέση απόδοση στα μέτρα αξιολόγησης είναι αυτό με που χρησιμοποιεί το 60% του συνόλου εκπαίδευσης, δηλαδή για sampsize ίσο με 0.6 με βαθμολογία 1.16 **(πίνακας 4.22)**.

sampsize	Average rank
0.2	1.667
0.4	2.000
0.6	1.167
Best rank = 1.16 , Sampsize = 0.6	

Πίνακας 4. 22 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο sampsize*

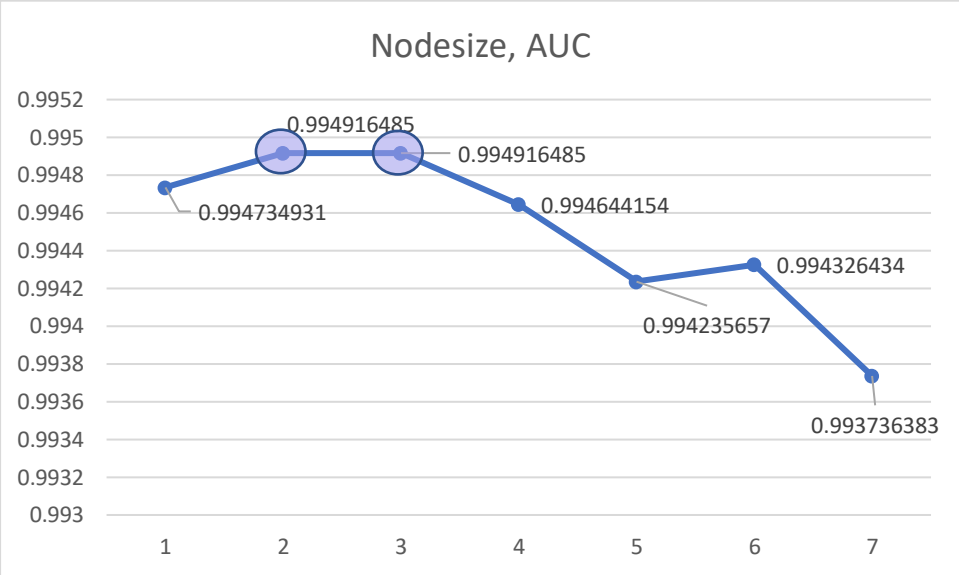
Προτελευταία υπερπαράμετρος που θα εξεταστεί είναι η **nodesize**. Αναφορικά, η nodesize εκφράζει των αριθμό των παρατηρήσεων σε κάθε καταληκτικό κόμβο. Στον **πίνακα 4.23**, παρουσιάζονται τα μέτρα αξιολόγησης για όλες τις υποψήφιες τιμές της υπερπαραμέτρου. Σύμφωνα λοιπόν με τον **πίνακα 4.23** προκύπτει πως για τις τιμές της nodesize ίσες με 2 και 3 σημειώνεται η μέγιστη τιμή του μέτρου AUC ίση με 0.9947349.

Hyperparameter: nodesize						
nodesize	auc	accuracy	precision	sensitivity	specificity	F1-score
1	0.9947349	0.9585253	0.9774436	0.9558824	0.9629630	0.9665428
2	0.9949165	0.9539171	0.9701493	0.9558824	0.9506173	0.9629630
3	0.9949165	0.9677419	0.9777778	0.9558824	0.9629630	0.9741697
4	0.9946442	0.9585253	0.9703704	0.9558824	0.9506173	0.9667897
5	0.9942357	0.9539171	0.9701493	0.9705882	0.9506173	0.9629630
6	0.9943264	0.9631336	0.9705882	0.9705882	0.9506173	0.9705882
7	0.9937364	0.9631336	0.9705882	0.9632353	0.9506173	0.9705882
Best AUC performance = 0.9949165, nodesize = 1						

Πίνακας 4. 23 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο nodesize.*

Επιλέγονται λοιπόν οι τιμές 2 και 3 καθώς και οι γειτονικές τους 1 και 4 για την διαδικασία βαθμολόγησης που πραγματοποιείται στον πίνακα 4.24. Με τιμή για nodesize ίση με 3, το

μοντέλο κατέχει την πρώτη θέση σε όλα τα μέτρα αξιολόγησης. Ακολουθεί το μοντέλο με τιμή 1, στην προτελευταία θέση το μοντέλο με τιμή 4 και τέλος το μοντέλο με τιμή 4.



Εικόνα 4.19 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει της nodesize.*

Hyperparameter: nodesize						
nodesize	AUC	accuracy	precision	sensitivity	specificity	F1-score
1	0.9947349	0.9585253	0.9774436	0.9558824	0.9629630	0.9665428
2	0.9949165	0.9539171	0.9701493	0.9558824	0.9506173	0.9629630
3	0.9949165	0.9677419	0.9777778	0.9558824	0.9629630	0.9741697
4	0.9946442	0.9585253	0.9703704	0.9558824	0.9506173	0.9667897
Ranking						
nodesize	AUC	accuracy	precision	sensitivity	specificity	F1-score
1	3	2	2	1	1	3
2	1	4	4	1	3	4
3	1	1	1	1	1	1
4	4	2	3	1	3	2

Πίνακας 4. 24 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για το μοντέλο random forest για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο nodesize*

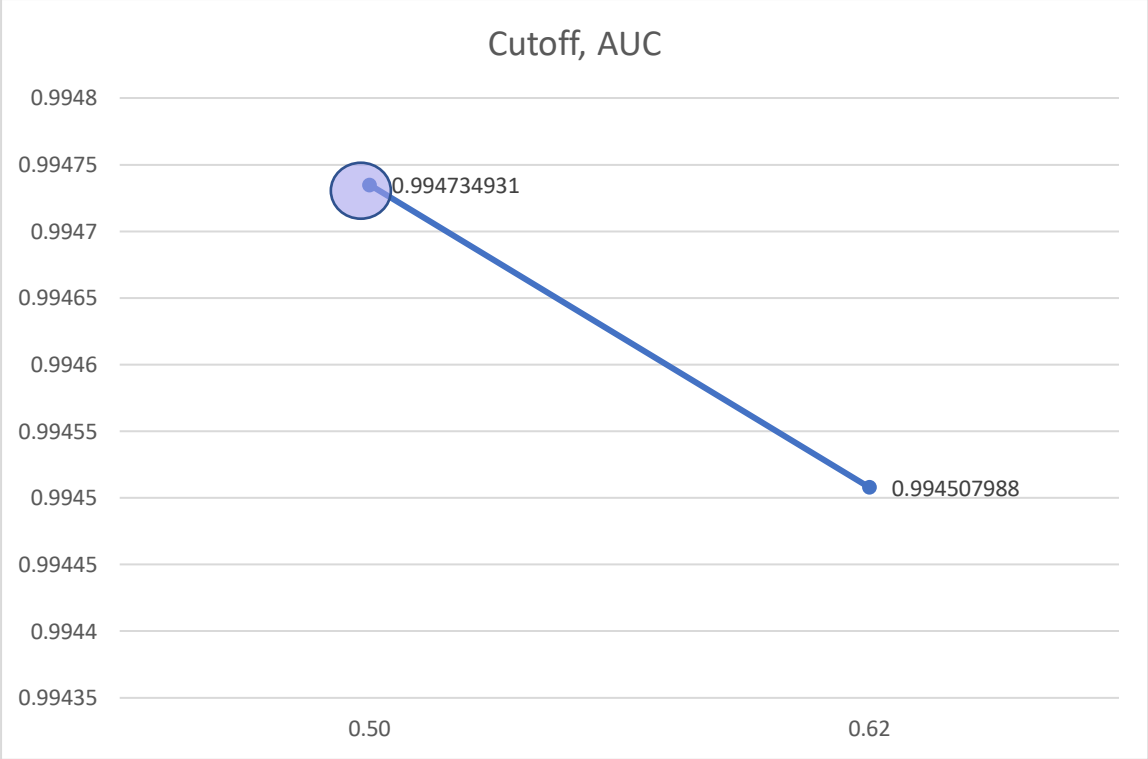
sampsize	Average rank
1	2
2	2.83
3	1
4	2.5
Best rank = 1, Sampsize = 3	

Πίνακας 4. 25 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο nodesize.*

Τελευταία υπερμεταβλητή είναι η cutoff όπου στην πραγματικότητα ερμηνεύεται ως το σύνορο απόφασης. Γι ’αυτό λοιπόν θα εξεταστούν δύο τιμές, 0.5 και 0.62, όπως αλώστε εξετάστηκαν και τα προηγούμενα μοντέλα. Είναι ευδιάκριτο λοιπόν πως για σύνορο απόφασης ίσο με 0.5 το μοντέλο αποδίδει καλύτερα στο μέτρο αξιολόγησης AUC με τιμή 0.9947349 **(πίνακας 4.26)**.

Hyperparameter: Cutoff						
nodesize	auc	accuracy	precision	sensitivity	specificity	F1-score
0.50	0.9947349	0.9585253	0.9774436	0.9558824	0.9629630	0.9665428
0.62	0.9945080	0.9585253	0.9847328	0.9558824	0.9753086	0.9662921
Best AUC performance = 0.9947349 , Cutoff = 0.5						

Πίνακας 4. 26 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο cutoff.*



Εικόνα 4.20 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει της cutoff.*

Ανεξαρτήτως της υπεροχής που σημειώνει το μοντέλο με τιμή για cutoff ίση με 0.5, αποδεικνύεται εν τέλη ότι το μοντέλο με τιμή ίση με 0.62 αποδίδει καλύτερα κατά μέσο όρο στα μέτρα αξιολόγησης.

Hyperparameter: Cutoff						
cutoff	auc	accuracy	precision	sensitivity	specificity	F1-score
0.50	0.9947349	0.9585253	0.9774436	0.9558824	0.9629630	0.9665428
0.62	0.9945080	0.9585253	0.9847328	0.9558824	0.9753086	0.9662921
Best AUC performance = 0.9947349 , Cutoff = 0.5						
0.50	1	1	2	1	2	1
0.62	2	1	1	1	1	2

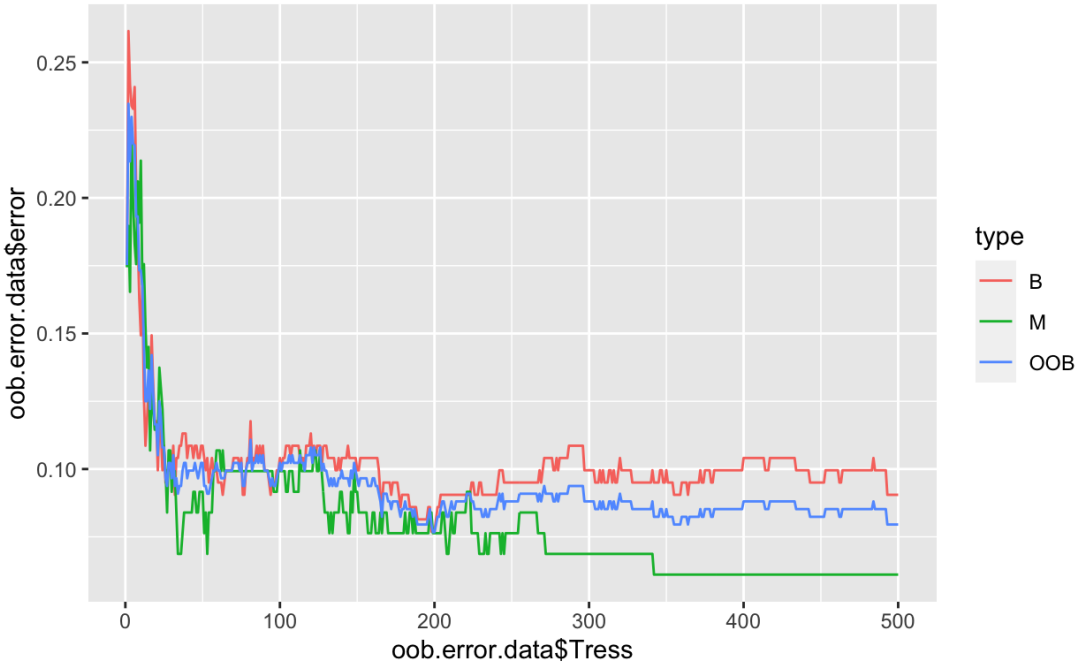
Πίνακας 4. 27 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για το μοντέλο random forest για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο cutoff.*

Cutoff	Average rank
0.5	1.83
0.62	1
Best rank = 1 , Cutoff = 0.62	

Πίνακας 4. 28 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο cutoff.*

Για να βρεθεί λοιπόν ο βέλτιστος συνδυασμός τιμών των υπερμεταβλητών μετά από αυτή τη διαδικασία, πραγματοποιείται ξανά αλλά αυτή τη φορά με τις τιμές των πινάκων 4.16, 4.19, 4.22 και 4.28

Συνοψίζοντας, μετά την διαδικασία αναζήτησης πλέγματος το μοντέλο random forest αποτελείται από 500 δέντρα, 1 ανεξάρτητη μεταβλητή που λαμβάνεται υπόψη ως κριτήριο για τη δημιουργία κόμβου, 40% του δείγματος του συνόλου εκπαίδευσης, κάθε κόμβος αποτελείται από 2 παρατηρήσεις ενώ το βέλτιστο σύνορο απόφασης ισούται με 0.62. Έπειτα παρουσιάζονται η γραφική αναπαράσταση των error rates συναρτήσει του πλήθους των δέντρων **(εικόνα 4.21)** για το μοντέλο με τις βέλτιστες τιμές των υπερπαραμέτρων του όπως και τα συνοπτικά αποτελέσματα του βέλτιστου μοντέλου με την εντολή ***summary()*** **(πίνακας 4.30)** στο περιβάλλον της R. Παρατηρείται πως η διαφορά μεταξύ των μοντέλων ως προς τον δείκτη OOB error rate είναι σχεδόν αμελητέα. Ειδικότερα, το μοντέλο με τις προεπιλεγμένες τιμές των υπερπαραμέτρων παρουσιάζει OOB error rate ίσο με 7.39%, ενώ αυτό με τις βέλτιστες ίσο με 7.95%. Στο βέλτιστο μοντέλο δηλαδή παρουσιάστηκε αύξηση στο σφάλμα της τάξης του 0.56%. Επίσης, το βέλτιστο μοντέλο παρουσιάζει class error για την κλάση της κακοήθειας ίσο με 0.06106, δηλαδή το 6.1% των ασθενών με κακοήθεια δεν ταξινομήθηκαν σωστά, ποσοστό το οποίο μειώθηκε συγκριτικά με το αρχικό μοντέλο το οποίο παρουσίασε class error για την κλάση της κακοήθειας ίσο με 0.1297, δηλαδή 12.9%. Αυτή η βελτίωση εξομάλυνε την μεροληψία που παρουσίαζε το μοντέλο απέναντι στις δύο κλάσεις, προβλέποντας την θετική κλάση σε μεγαλύτερο ποσοστό σε σχέση με την αρνητική. Όσον αφορά το class error της θετικής κλάσης, ανέβηκε κατά 5% σε σχέση με το αρχικό μοντέλο. Ανεξαρτήτως αυτής της αύξησης ως τελικό μοντέλο τυχαίου δάσους θα επιλεγεί αυτό που υπέστη παραμετροποίηση διότι παρουσιάσει αντικειμενικά την μέγιστη μέση απόδοση σε όλα τα μέτρα αξιολόγησης σε σχέση με τα υπόλοιπα μοντέλα και προβλέπει σωστά σε παρόμοια ποσοστά τις δύο κλάσεις.



Εικόνα 4.21 *Γραφική αναπαράσταση των error rates συναρτήσει του πλήθους των δέντρων για το μοντέλο με τις βέλτιστες τιμές των υπερπαραμέτρων του.*

Type of random forest: Classification		
Number of trees	500	
No. of variables tried at each split:	1	
OOB estimate of error rate:	7.95%	
Confusion matrix:	Actual B	Actual M
Predicted B	201	8
Predicted M	20	123
Class.error	0.09049	0.06106

Πίνακας 4. 29 *Συνοπτικά αποτελέσματα του μοντέλου Random Forest για τις βέλτιστες τιμές των υπερπαραμέτρων.*

Οι πίνακες σύγχυσης (confusion matrix) του συνόλου εκπαίδευσης παρουσιάζονται στον **πίνακα 4.31** για τα δύο σύνορα απόφασης ενώ του συνόλου ελέγχου στον **πίνακα 4.32** όπως και τα μέτρα αξιολόγησης του συνόλου εκπαίδευσης και ελέγχου για τα σύνορα απόφασης 0.5 αλλά και 0.62 του βέλτιστου μοντέλου random forest παρουσιάζονται στον **πίνακα 4.33**. Είναι ξεκάθαρο πως το μοντέλο με σύνορο απόφασης ίσο με 0.62 υπερέχει έναντι αυτού με 0.5 με

τη μόνη εξαίρεση του μέτρου της ευαισθησίας (sensitivity) όπου το μοντέλο με cutoff ίσο με 0.5 υπερέχει.

Confusion Matrix training set				
n = 352	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	221	7	219	0
Predicted M	0	124	2	131

Πίνακας 4. 30 *Πίνακας σύγχυσης (confusion matrix) του συνόλου εκπαίδευσης για τα δύο σύνορα απόφασης.*

Confusion Matrix test set				
n = 217	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	133	3	129	1
Predicted M	3	78	7	80

Πίνακας 4. 31 *Πίνακας σύγχυσης (confusion matrix) του συνόλου ελέγχου για τα δύο σύνορα απόφασης.*

Έπειτα, στον **πίνακα 4.33**, παρουσιάζονται τα μέτρα αξιολόγησης του μοντέλου τυχαίου δάσους για τα δύο σύνορα απόφασης όπως επίσης και για τα δύο σύνολα δεδομένων, εκπαίδευσης και ελέγχου. Αναλυτικότερα. για το μέτρο accuracy στο σύνολο εκπαίδευσης το μοντέλο με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 98.01% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 99.43%, ενώ στο σύνολο ελέγχου με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 97.24% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 96.31%.

Στο σύνολο εκπαίδευσης, για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 96.93%, ενώ για 0.62 ισούται με 100%. Στο σύνολο ελέγχου για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 0.9779, ενώ για 0.62 ισούται με 0.9923.

Το 100% των ασθενών του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως υγιείς για το σύνορο απόφασης 0.5 ενώ για 0.62 το 0.991%. Αντίστοιχα, για το σύνολο ελέγχου το 0. 9779% για 0.5 και 0.9485% για 0.62. Επίσης, το 9466% των όγκων του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως κακοήθης για το σύνορο απόφασης 0.5 ενώ για 0.62 το 100%. Αντίστοιχα, για το σύνολο ελέγχου το 96.30% για 0.5 και 98.77% για 0.62. Τέλος, για τα μέτρα F1-score και AUC σημειώθηκαν αποδόσεις ίσες με 98.4409% και 99.989% για σύνορο απόφασης ίσο με 0.5 του συνόλου εκπαίδευσης ενώ για 0.62 σημειώθηκε 99.5454% και 99.9723% αντίστοιχα, ενώ στο σύνολο ελέγχου για σύνορο απόφασης 0.5 σημειώθηκε απόδοση στα μέτρα F1-score και AUC, ίση με 97.79% και 99.646%, όπως επίσης για 0.62 σημειώθηκε 97.76% και 99.68% αντίστοιχα.

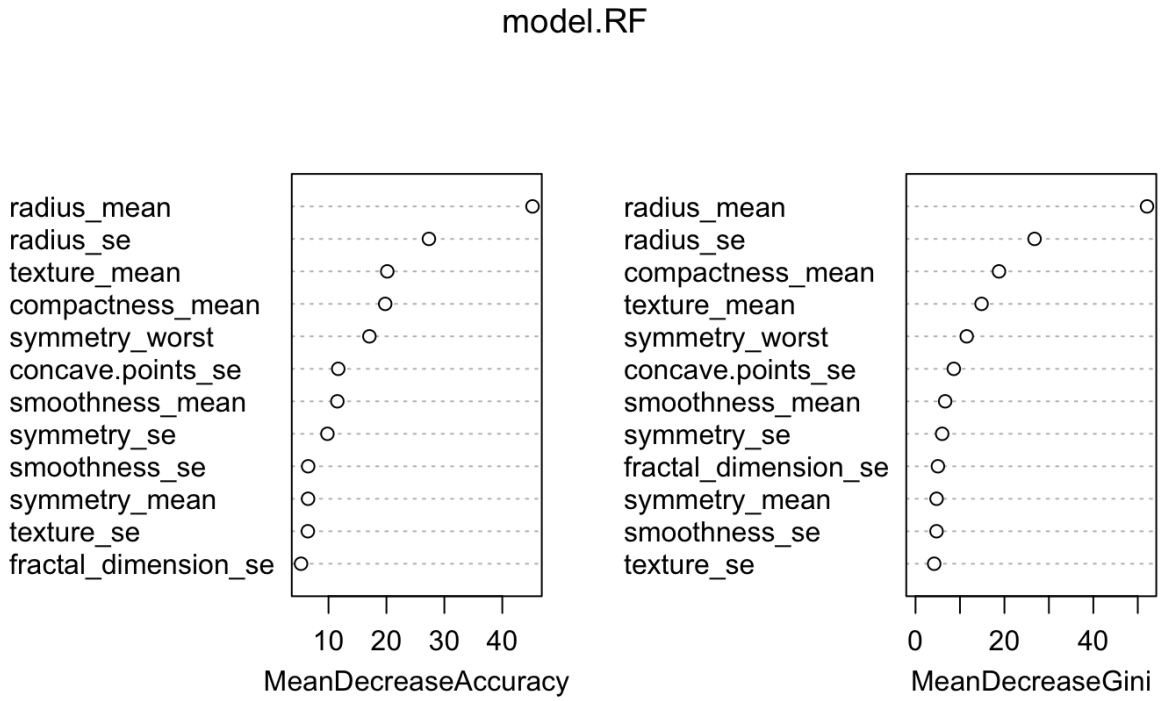
Evaluation Metric	Threshold = 0.5		Threshold = 0.62	
	Training set	Test set	Training set	Test set
Accuracy	0.9801	0.9724	0.9943	0.9631
Precision	0.9693	0.9779	1	0.9923
Recall (Sensitivity)	1	0.9779	0.991	0.9485
Specificity	0.9466	0.9630	1	0.9877
F1-Score	0.984409	0.9779	0.995454	0.9776
AUC	0.99989	0.99646	0.999723	0.9968

Πίνακας 4. 32 *Μέτρα αξιολόγησης του βέλτιστου μοντέλου random forest, για τα δύο σύνορα απόφασης αλλά και για τα δύο σύνολα (εκπαίδευσης και ελέγχου)*

	B	M	MeanDecreaseAccuracy	MeanDecreaseGini
radius_mean	0.065141936	0.100149613	0.078101644	11.428173
texture_mean	0.023473147	0.039796092	0.029516651	5.959998
smoothness_mean	0.006139608	0.025879861	0.013460376	4.60615
compactness_mean	0.02822893	0.049264734	0.036006099	6.544982
symmetry_mean	0.003652843	0.015830157	0.008183294	3.35211
radius_se	0.03713232	0.056212639	0.044156078	8.160703
texture_se	0.004616935	0.002555712	0.003850657	2.719164
smoothness_se	0.006157382	0.00351241	0.005156213	2.978116
concave.points_se	0.016249021	0.031882233	0.021945489	5.841454
symmetry_se	0.008806269	0.009588549	0.009057881	3.358962
fractal_dimension_se	0.002411874	0.007887676	0.00442673	2.982203
symmetry_worst	0.012633694	0.0229091	0.016495722	4.593356

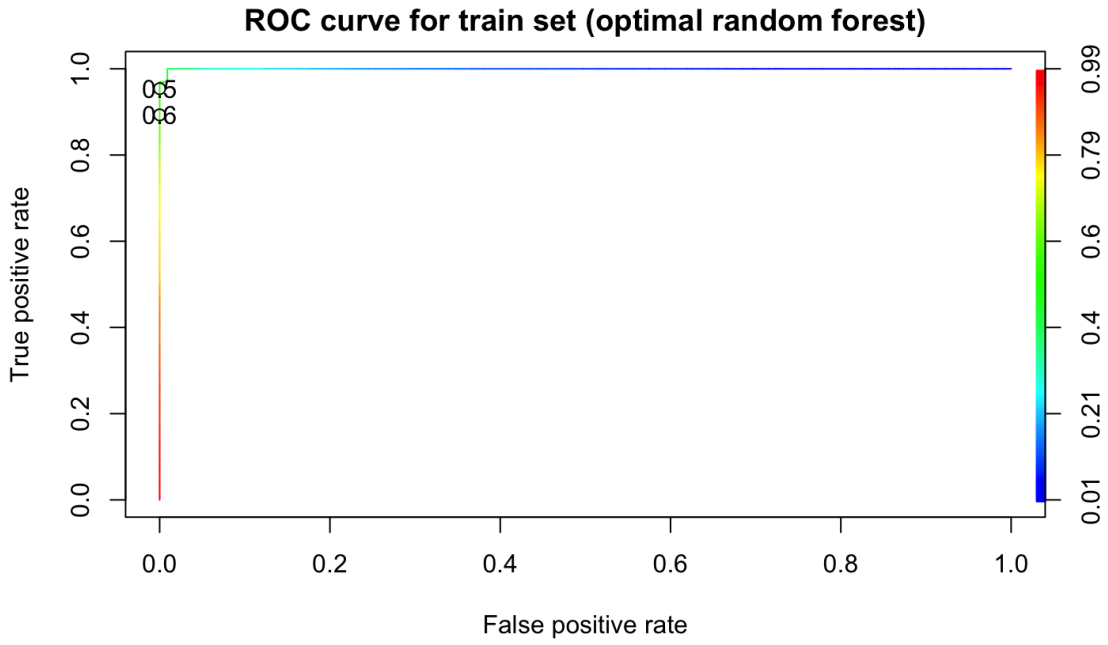
Πίνακας 4. 33 *Συνοπτικά αποτελέσματα του μοντέλου Random Forest για τις βέλτιστες τιμές των υπερπαραμέτρων του.*

Στην εικόνα 4.20 φαίνονται τα μέτρα των δεικτών αυτών για κάθε μεταβλητή́ κατά́ φθίνουσα σειρά́, βάσει των οποίων μπορούμε να λάβουμε απόφαση για το ποιοι είναι οι βασικοί́ προγνωστικοί́ παράγοντες του μοντέλου Random Forest. Είναι λοιπόν εμφανές πως ό́τι οι σημαντικότερες μεταβλητες που εξασφαλίζουν καλύτερη πρόβλεψη για τις δυο κλά́σεις είναι η μέση απόσταση του όγκου από τον πυρήνα σε σημεία της ακτίνας του καθώς και τυπική απόκλιση της απόστασης του όγκου από τον πυρήνα σε σημεία της ακτίνας του. Τα υπόλοιπα στατιστικά́ στοιχειά ενός όγκου από δεδομένα εικόνας είναι σε μικρότερο βαθμό́ σημαντικά́ για την πρόβλεψη της επικρατούσας κλάσης.

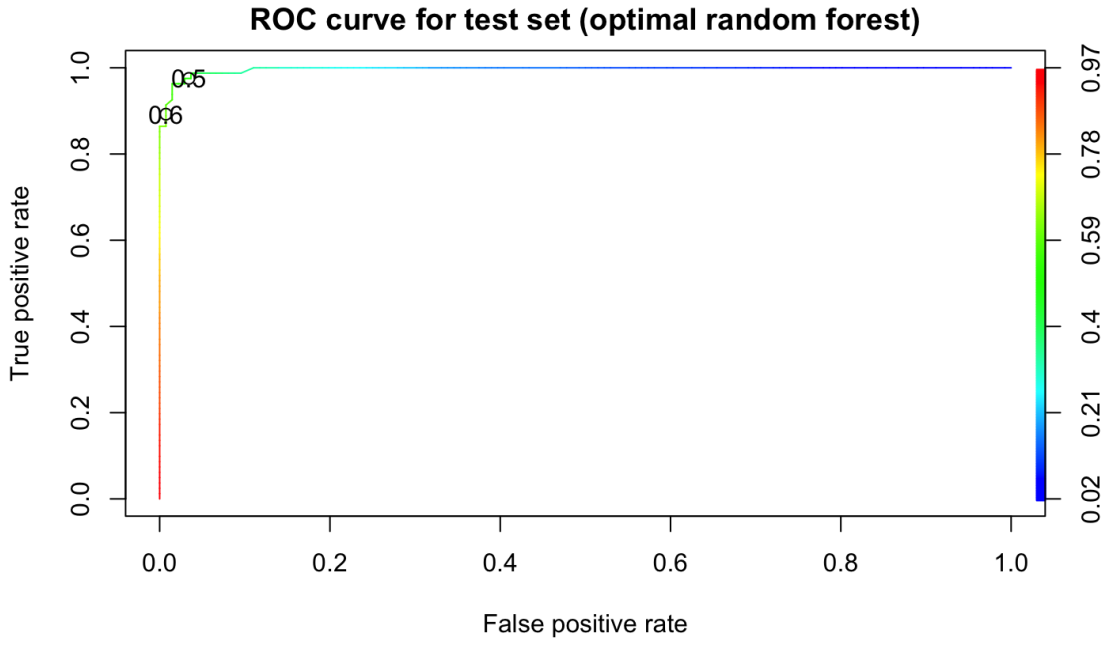


Εικόνα 4.22 *Διαγραμματική απεικόνιση σημαντικότητας μεταβλητών βέλτιστου μοντέλου random forest.*

Τέλος, ακολουθούν οι καμπύλες ROC, lift αλλά και precision vs recall για το βέλτιστο μοντέλο random forest για το σύνολο απόφασης αλλά και για το σύνολο ελέγχου. Στην **εικόνα 4.23** φαίνεται η καμπύλη ROC του συνόλου εκπαίδευσης ενώ στην **εικόνα 4.24** του συνόλου ελέγχου. Είναι ευδιάκριτο πως για το σύνορο απόφασης 0.5 το μοντέλο αποδίδει καλύτερα και στα δύο σύνολα όσον αφορά το μέτρο sensitivity (true positive rate), όπου μετριέται στον κάθετο άξονα. Παρατηρείται ότι το βέλτιστα παραμετροποιημένο μοντέλο για σύνορο απόφασης 0.5 προβλέπει μεγαλύτερο ποσοστό της θετικής κλάσης λόγω του ότι είναι πιο ψηλά στην καμπύλη έναντι αυτού που ισούται με 0.62. Το μοντέλο παρουσιάζει δείκτη AUC ίσο με 0.997413 **(πίνακας 4.30)**, δηλαδή το μοντέλο μπορεί να διακρίνει το 99.7% των παρατηρήσεων μεταξύ και των δύο κλάσεων.



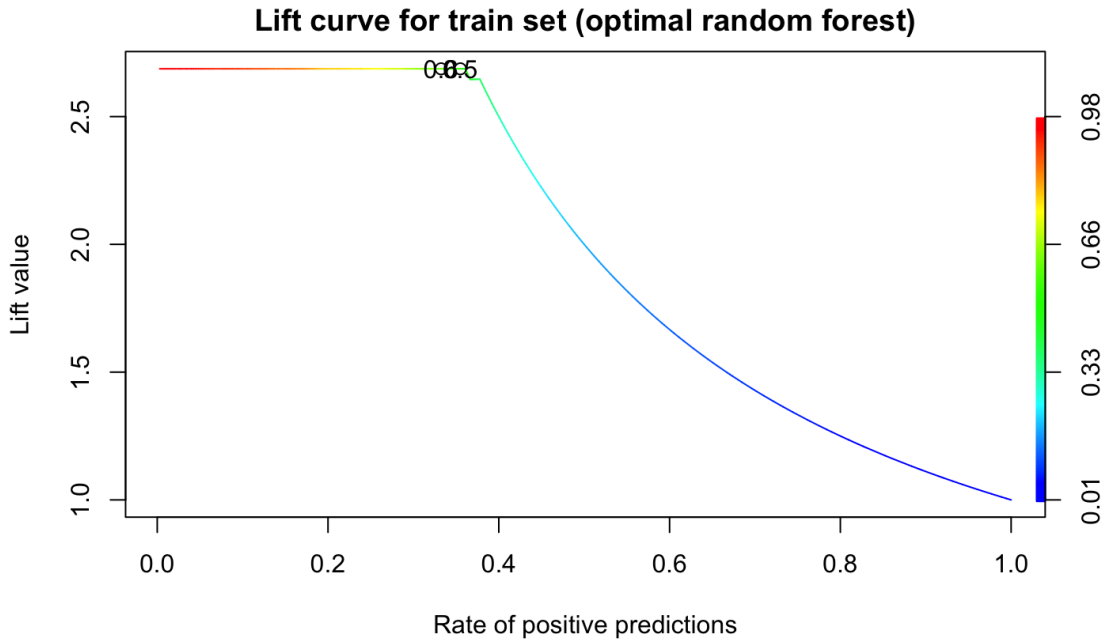
Εικόνα 4.23 *Καμπύλη ROC για το σύνολο εκπαίδευσης του βέλτιστου μοντέλου random forest*



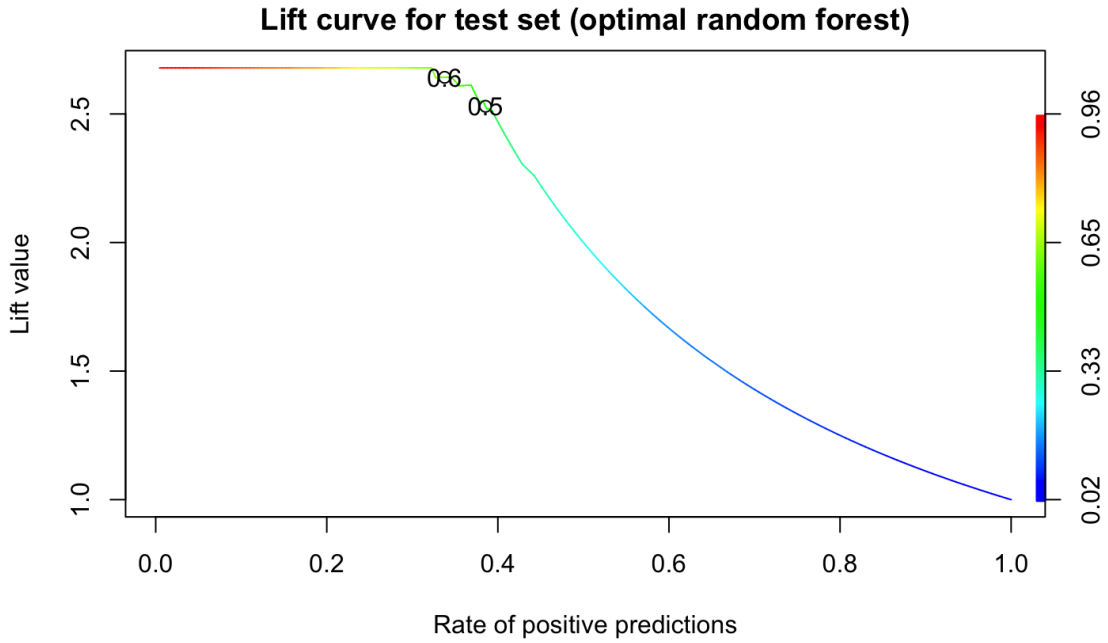
Εικόνα 4. 24 *Καμπύλη ROC για το σύνολο ελέγχου του βέλτιστου μοντέλου random forest.*

Ακολουθούν οι καμπύλες lift για το σύνολο εκπαίδευσης (εικόνα 4.25) και για το σύνολο ελέγχου (εικόνα 4.26). Είναι ευδιάκριτο πως το μοντέλο με σύνορο απόφασης ίσο με 0.62 υπερέχει έναντι αυτού με 0.5. Ερμηνεύοντας την καμπύλη lift εκμαιεύεται η πληροφορία ότι για σύνορο απόφασης ίσο με 0.62 είναι μεγαλύτερη η απόδοση η παρατήρηση η οποία έχει

ταξινομηθεί ως θετική (καλοήθεια) να ναι όντως θετική. Στην καμπύλη για το σύνολο εκπαίδευσης (εικόνα 4.23) η απόδοση αυτή ισούται με την μέγιστη τιμή ανύψωσης, ενώ για το σύνολο ελέγχου (εικόνα 4.24) η τιμή ανήκει μετά το σημείο της πτώσης. Παρ' αυτά το μοντέλο με σύνορο απόφασης ίσο με 0.62 εξακολουθεί να παρουσιάζει υπεροχή ενώ αυτό με 0.5 αποδεικνύεται ιδιαίτερα πιο αδύναμο.

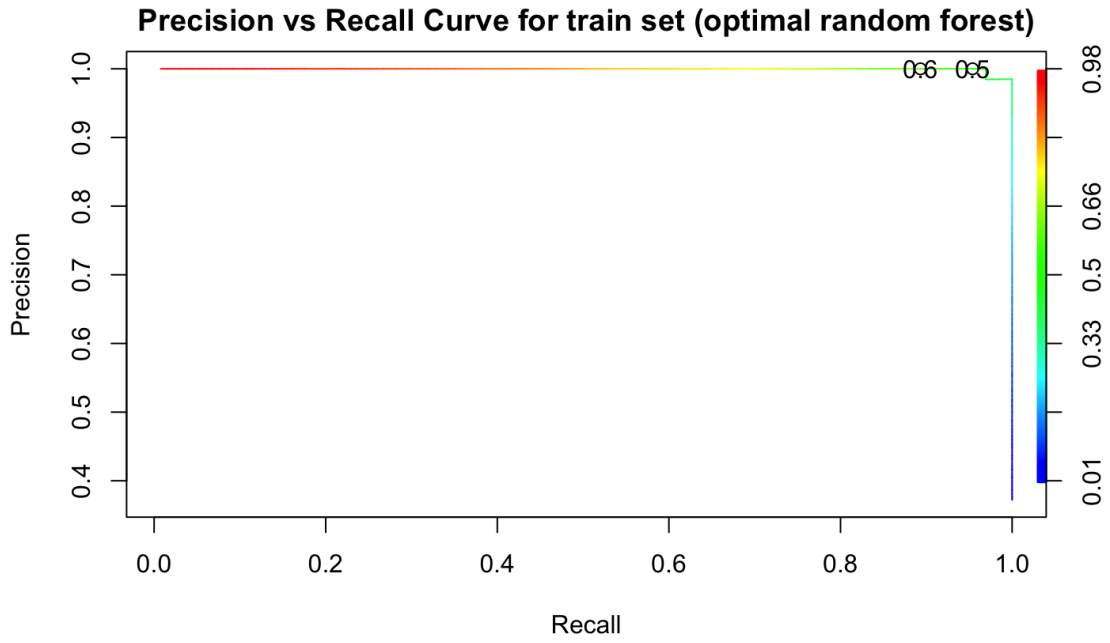


Εικόνα 4.25 *Καμπύλη lift για το σύνολο εκπαίδευσης του βέλτιστου μοντέλου random forest.*

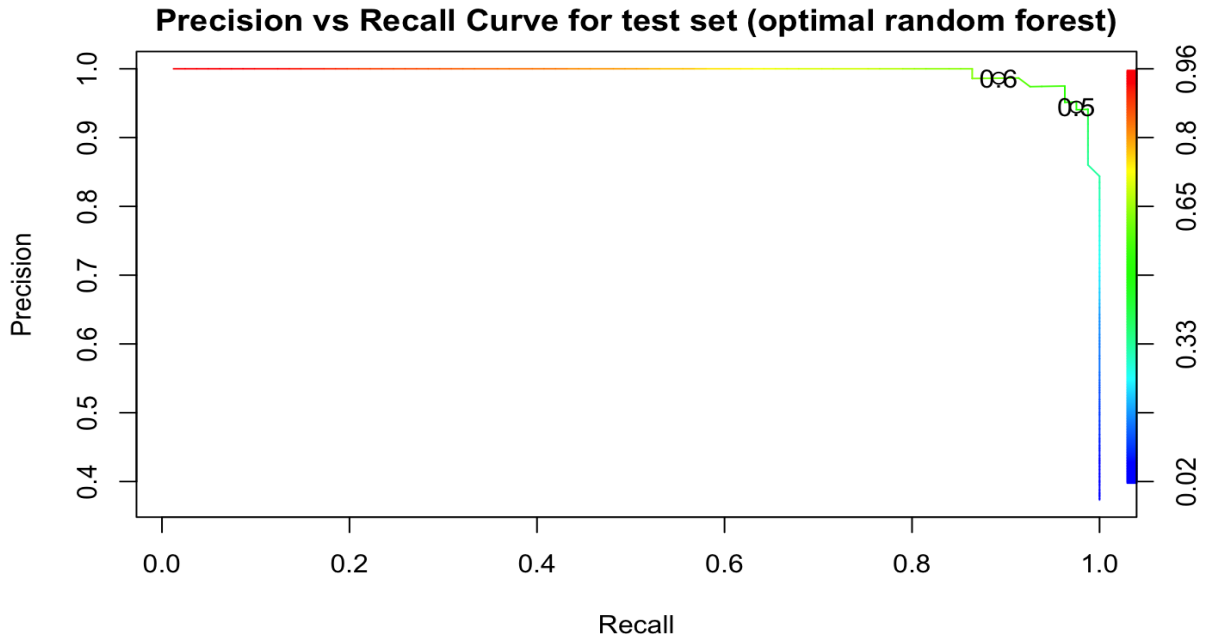


Εικόνα 4.26 *Καμπύλη lift για το σύνολο ελέγχου του βέλτιστου μοντέλου random forest*

Έπειτα, τελευταίο διάγραμμα που αφορά το μοντέλο τυχαίου δάσους είναι αυτό του precision vs recall. Στις εικόνες 4.27 και 4.28 όπου αντιστοιχούν τα γραφήματα precision vs recall για το σύνολο εκπαίδευσης και ελέγχου, αναπαρίστανται γραφικά η απόδοση του μέτρου precision (κάθετος άξονας) και recall (οριζόντιος άξονας). Στο σύνολο ελέγχου όπως και στο σύνολο εκπαίδευσης το σύνορο απόφασης ίσο με 0.6 φαίνεται να αποδίδει καλύτερα από αυτό που ισούται με 0.5 λόγω της μεγαλύτερης ποσοστιαίας υπεροχής του στο μέτρο αξιολόγησης precision έναντι του μέτρου recall.



Εικόνα 4.27 *Καμπύλη precision vs recall για το σύνολο εκπαίδευσης του βέλτιστου μοντέλου random forest*



Εικόνα 4.28 *Καμπύλη precision vs recall για το σύνολο εκπαίδευσης του βέλτιστου μοντέλου random forest.*

4.5 Μοντέλο Support Vector Machine (SVM).

Ο τελευταίος αλγόριθμος που θα χρησιμοποιηθεί για την ανάλυση του συνόλου δεδομένων που αφορά τα χαρακτηρίστηκα κυτταρικού πυρήνα από δεδομένα εικόνας είναι αυτός των Μηχανών Διανυσμάτων Υποστήριξης (support vector machine). Κατά τον συγκεκριμένο αλγόριθμο γίνεται ο διαχωρισμός της κλάσης του συνόλου δεδομένων μέσω ενός υπερεπίπεδου. Αυτός ο διαχωρισμός γίνεται με βάση κάποιων μαθηματικών συναρτήσεων πυρήνα (kernels), οι οποίες διαχωρίζονται σε γραμμικές (linear kernel), πολυωνυμική (polynomial kernel) και γκαουσιανή (radial kernel). Στη συγκεκριμένη διπλωματική εργασία θα εξεταστούν και οι τρεις περιπτώσεις στο περιβάλλον της R.

Οι μεταβλητές βάσει των οποίων θα εκπαιδευτεί το υπόδειγμα είναι οι ίδιες που χρησιμοποιήθηκαν και στους υπολοίπους αλγορίθμους και αναγράφονται στον πίνακα 4.1. Βέβαια, εκτός από την ευελιξία της επιλογής της μαθηματικής συνάρτησης πυρήνα για την επίτευξη των βέλτιστων αποτελεσμάτων, υπάρχουν και κάποιες υπερπαράμετροι που μπορούν να μεγιστοποιήσουν την απόδοση του μοντέλου όπως ακριβώς και στο υπόδειγμα του τυχαίου δάσους που εξετάστηκε παραπάνω. Ο συντονισμός του αλγορίθμου στοχεύει να βρει τις καλύτερες παραμέτρους για την ταξινόμηση του συνόλου δεδομένων εκπαίδευσης και την

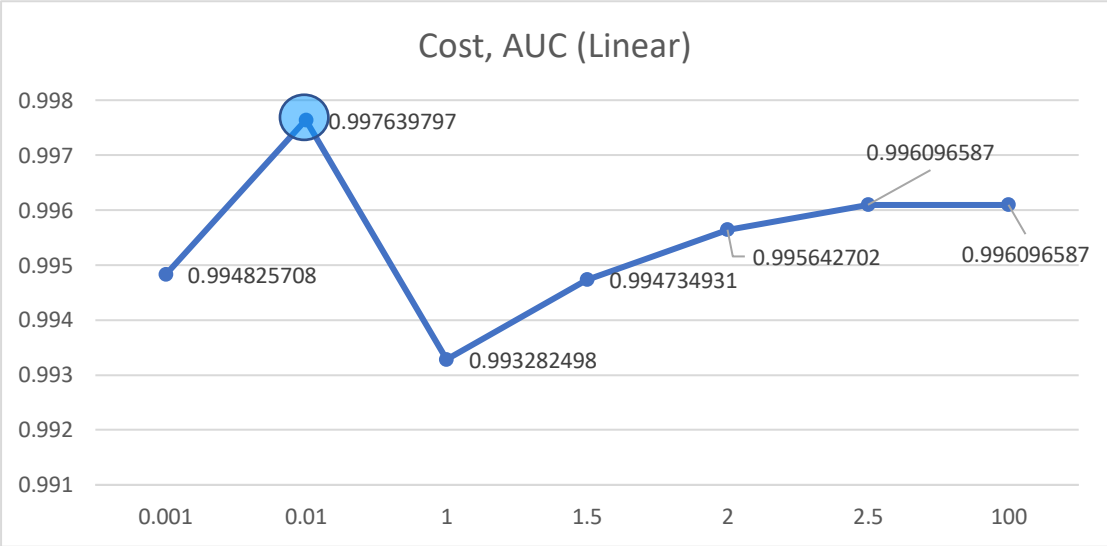
αποφυγή εσφαλμένης ταξινόμησης των παρατηρήσεων εκτελώντας διασταυρουμένη επικύρωση σε ένα σύνολο μοντέλων ενδιαφέροντος.

Οι βασικοί λοιπόν παράμετροι που πρόκειται να εξεταστούν για την επίτευξη της μέγιστης ακρίβειας της πρόβλεψης είναι αυτή του κόστους (cost), gamma (μόνο για την γκαουσιανή συνάρτηση) και του βαθμού πολυωνύμου (degree) της πολυωνυμικής συνάρτησης πυρήνα. Οι παράμετρος cost είναι απαραίτητη για όλα τα μοντέλα ανεξαρτήτως της επιλογής συνάρτησης πυρήνα και ουσιαστικά αντιπροσωπεύει το κατά πόσο το μοντέλο ‘’συγχωρρεί’’ μια λανθασμένη ταξινόμηση. Όσο μεγαλύτερος είναι ο βαθμός του κόστους τόσο μεγαλύτερη είναι και η ‘’ανοχή’’ του μοντέλου απέναντι στις λάθος ταξινομήσεις στο σύνολο των δεδομένων εκπαίδευσης. Όσο για την υπερπαράμετρο gamma, η οποία αφορά μόνο την γκαουσιανή συνάρτηση πυρήνα, έχει ως στόχο να ομαδοποιήσει όσο το δυνατών περισσότερα δεδομένα που ανήκουν στην ίδια κλάση και αντιπροσωπεύει το πόση καμπυλότητα είναι επιθυμητή στο όριο απόφασης. Τέλος, η υπερπαράμετρος degree της πολυωνυμικής συνάρτησης πυρήνα ουσιαστικά είναι ο βαθμός του πολυωνύμου που χρησιμοποιείται από τον αλγόριθμο για την εύρεση του πολυεπιπέδου βάσει του οποίου γίνεται η εύρεση του υπερεπίπεδου (hyperplane).

Ξεκινώντας από την γραμμική συνάρτηση πυρήνα (linear kernel), εφαρμόζεται η διαδικασία βελτιστοποίησης της αποδοτικότητας του μοντέλου, δηλαδή στην επίτευξη της μέγιστης απόδοσης των μέτρων αξιολόγησης με την ίδια διαδικασία αναζήτησης πλέγματος που εφαρμόστηκε ομοίως και στο μοντέλο random forest για τον προσδιορισμό της βέλτιστης τιμής του κόστους. Στον **πίνακα 4.32**, παρουσιάζονται τα μέτρα αξιολόγησης που προκύπτουν για κάθε τιμή της υπερπαραμέτρου cost έτσι ώστε να εντοπιστεί αυτή για την οποία βελτιστοποιείται η απόδοση του μοντέλου ως προς το μέτρο αξιολόγησης AUC. Το μοντέλο λαμβάνει τις τιμές (0.001, 0.01, 1, 1.5, 2, 2.5, 100). Προκύπτει λοιπόν ότι το μοντέλο με τιμή κόστους 0.01 αποδίδει στο μέγιστο όσον αφορά το μέτρο αξιολόγησης AUC, σε σχέση με τις υπόλοιπες του συνόλου δυνητικών τιμών που έχουν οριστεί για την διαδικασία μέσω πλέγματος.

Kernel: Linear						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
0.001	0.99483	0.94931	0.97710	0.94118	0.96296	0.95880
0.01	0.99764	0.96313	0.98485	0.95588	0.97531	0.97015
1	0.99328	0.97235	0.97794	0.97794	0.96296	0.97794
1.5	0.99473	0.96774	0.97778	0.97059	0.96296	0.97417
2	0.99564	0.96774	0.97778	0.97059	0.96296	0.97417
2.5	0.99610	0.96774	0.97778	0.97059	0.96296	0.97417
100	0.99610	0.97235	0.98507	0.97059	0.97531	0.97778
Best AUC performance =0.99764, Cost = 0.01						

Πίνακας 4. 34 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο cost του μοντέλου γραμμικής συνάρτησης πυρήνα.*



Εικόνα 4.29 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει του κόστους του μοντέλου με γραμμική συνάρτηση πυρήνα.*

Έπειτα, λαμβάνονται υπόψιν οι γειτονικές τιμές του κόστους (0.001 και 1) έτσι ώστε να εφαρμοστεί η διαδικασία βαθμολόγησης των μέτρων αξιολόγησης για τα μοντέλα με τις

προαναφερθείσες τιμές κόστους. Η διαδικασία παρουσιάζεται αναλυτικά στον **πίνακα 4.33**, καθώς και τα αποτελέσματα αυτής στον **πίνακα 4.34**.

Hyperparameter: Cost						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
0.001	0.99483	0.94931	0.97710	0.94118	0.96296	0.95880
0.01	0.99764	0.96313	0.98485	0.95588	0.97531	0.97015
1	0.99328	0.97235	0.97794	0.97794	0.96296	0.97794
Ranking						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
0.001	2	3	3	3	2	3
0.01	1	2	1	2	1	2
1	3	1	2	1	2	1

Πίνακας 4. 35 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο cost του μοντέλου SVM με γραμμική συνάρτηση πυρήνα.*

Cost	Average rank
0.001	2.67
0.01	1.50
1	1.67
Best rank = 2.17, Cost = 0.01	

Πίνακας 4. 36 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο cost του μοντέλου SVM με γραμμική συνάρτηση πυρήνα.*

Στην πρώτη θέση σημειώνοντας την μέγιστη μέση απόδοση στα μέτρα ξιολόγησης ίση με 1.5, βρίσκεται το μοντέλο με τιμή κόστους ίσο με 0.01. Έπειτα ακολουθεί το μοντέλο με τιμή κόστους ίση με 1 και βαθμολογία ίση με 1.67, ενώ στην τελευταία θέση βρίσκεται το μοντέλο με τιμή κόστους ίση με 0.001 σημειώνοντας την χαμηλότερη μέση βαθμολογία ίση με 2.67. Ως εκ τούτου από τη στιγμή που το μοντέλου SVM με γραμμική συνάρτηση πυρήνα έχει προκύψει και κατ' επέκτασιν εκπαιδευτεί αυτό με τιμή κόστους το 0.01 ακολουθεί, ακολουθεί η ταξινόμηση δεδομένων για το σύνολο εκπαίδευσης αλλά και ελέγχου για κάθε μία από τις δύο

κλάσεις. Στους **πίνακες 4.35** και **4.36**, απεικονίζονται οι πίνακες σύγχυσης του μοντέλου για τα δύο σύνολα δεδομένων όχι μόνο για το προεπιλεγμένο σύνορο απόφασης (0.5) αλλά και γι' αυτό το οποίο αντιπροσωπεύει την συχνότητα εμφάνισης της θετικής κλάσης.

Kernel: Linear				
n = 352	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	215	11	209	11
Predicted M	6	120	12	120

Πίνακας 4. 37 *Confusion matrix για τα δύο σύνορα απόφασης 0.5 και 0.62 του συνόλου εκπαίδευσης για το μοντέλο SVM με γραμμική συνάρτηση πυρήνα και τιμή κόστους 0.01.*

Kernel: Linear				
n = 217	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	130	2	130	2
Predicted M	6	79	6	79

Πίνακας 4. 38 *Confusion matrix για τα δύο σύνορα απόφασης 0.5 και 0.62 του συνόλου ελέγχου για το μοντέλο SVM με γραμμική συνάρτηση πυρήνα και τιμή κόστους 0.01.*

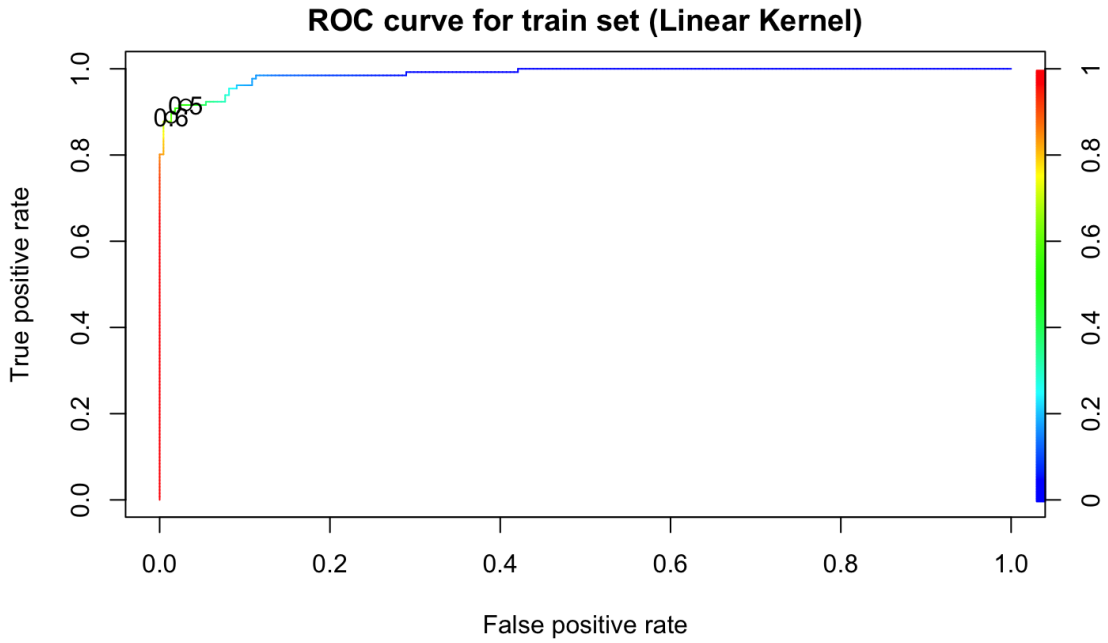
Ερμηνεύοντας λοιπόν το μοντέλο σύμφωνα με τα μέτρα αξιολόγησης που παρουσιάζονται στον **πίνακα 4.37**, για το μέτρο accuracy στο σύνολο εκπαίδευσης το μοντέλο με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 95.170% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 93.470%, ενώ στο σύνολο ελέγχου με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 96.313% των παρατηρήσεων όπως επίσης και για το σύνορο 0.62 το οποίο ταξινομεί το 96.313%. Στο σύνολο εκπαίδευσης, για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 95.130%, ενώ για 0.62 ισούται με 95%. Στο σύνολο ελέγχου για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 98.485, όπως επίσης και για σύνορο απόφασης 0.62. Το 97.290% των ασθενών του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως υγιείς για το σύνορο απόφασης 0.5 ενώ για 0.62 το 94.570%. Αντίστοιχα, για το σύνολο ελέγχου το 95.588% για 0.5 αλλά και για σύνορο 0.62. Επίσης, το 91.600% των όγκων του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως κακοήθης και για τα δύο σύνορα απόφασης, ενώ για το σύνολο ελέγχου το 97.531% για 0.5 και για 0.62. Τέλος, για τα μέτρα F1-score και AUC

σημειώθηκαν αποδόσεις ίσες με 0.96197% και 0.98757% για σύνορο απόφασης ίσο με 0.5 του συνόλου εκπαίδευσης ενώ για 0.62 σημειώθηκε 94.785% και 98.757% αντίστοιχα, ενώ στο σύνολο ελέγχου για σύνορο απόφασης 0.5 σημειώθηκε απόδοση στα μέτρα F1-score και AUC, ίση με 97.015% και 99.764%, όπως επίσης για 0.62 σημειώθηκε 97.015% και 99.764% αντίστοιχα.

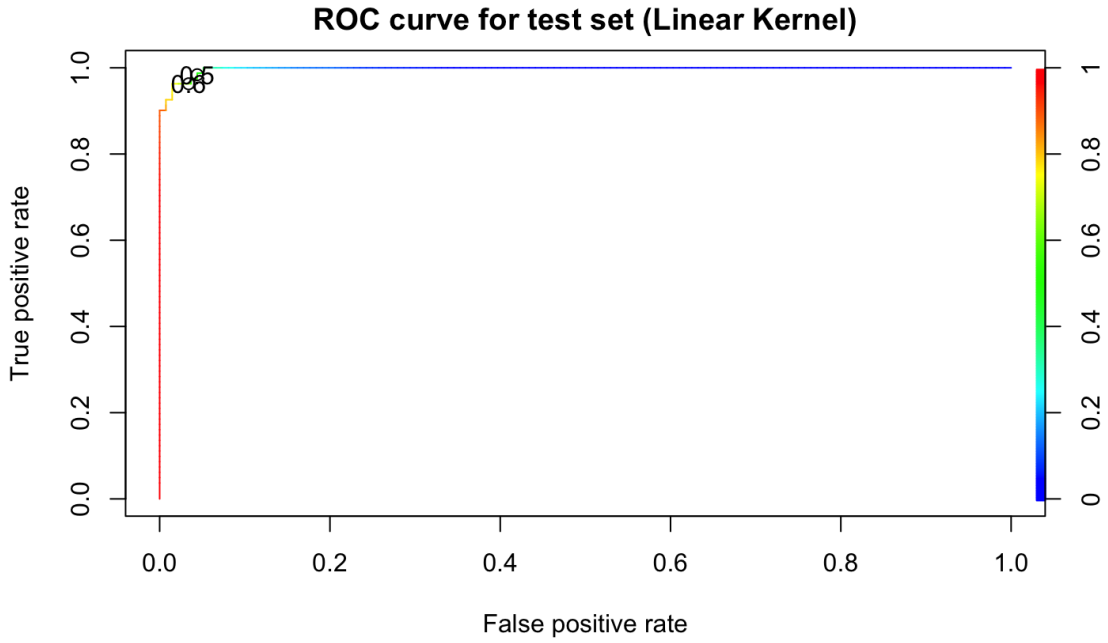
Kernel: Linear				
Evaluation Metric	Threshold = 0.5		Threshold = 0.62	
	Training set	Test set	Training set	Test set
Accuracy	0.95170	0.96313	0.93470	0.96313
Precision	0.95130	0.98485	0.95000	0.98485
Recall (Sensitivity)	0.97290	0.95588	0.94570	0.95588
Specificity	0.91600	0.97531	0.91600	0.97531
F1-Score	0.96197	0.97015	0.94785	0.97015
AUC	0.98757	0.99764	0.98757	0.99764

Πίνακας 4. 39 *Μέτρα αξιολόγησης του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για τα δύο σύνορα απόφασης του συνόλου εκπαίδευσης και ελέγχου.*

Παρακάτω στην **εικόνα 4.30**, όπου απεικονίζεται η καμπύλη ROC, φαίνεται υπεροχή του μοντέλου με σύνορο απόφασης όσον αφορά τα μέτρα sensitivity και specificity στο σύνολο εκπαίδευσης, ενώ στην **εικόνα 4.31**, η αποδόσεις των μοντέλων με τα δύο σύνορα απόφασης σχεδόν εφάπτονται με αυτό του 0.5 να σημειώνει καλύτερη απόδοση.

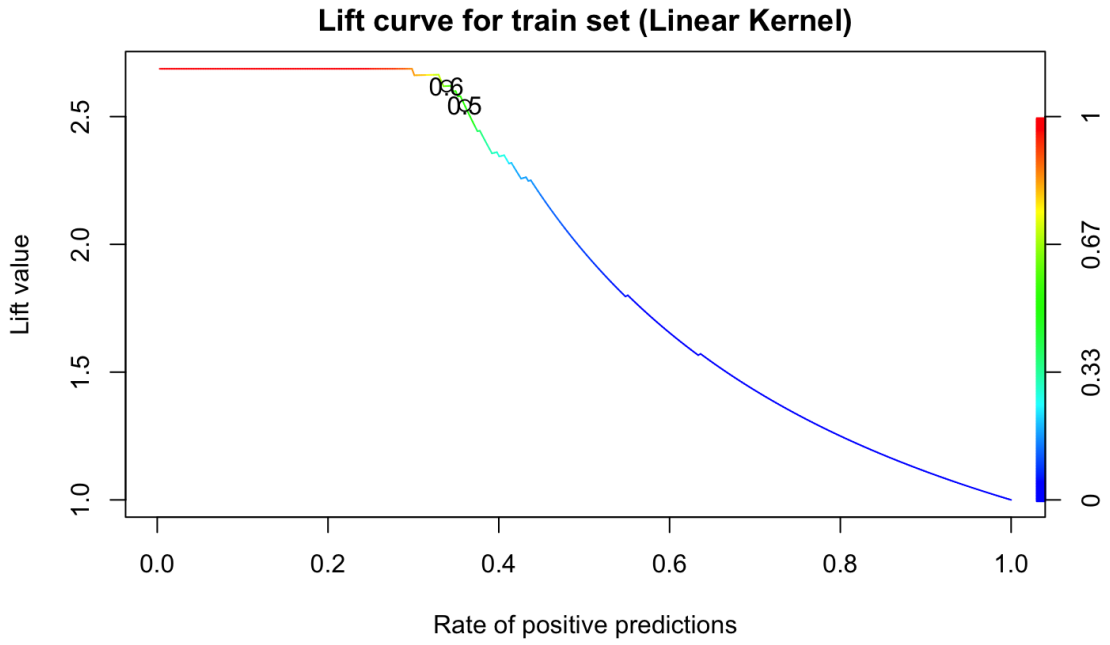


Εικόνα 4.30 *Καμπύλη ROC του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*

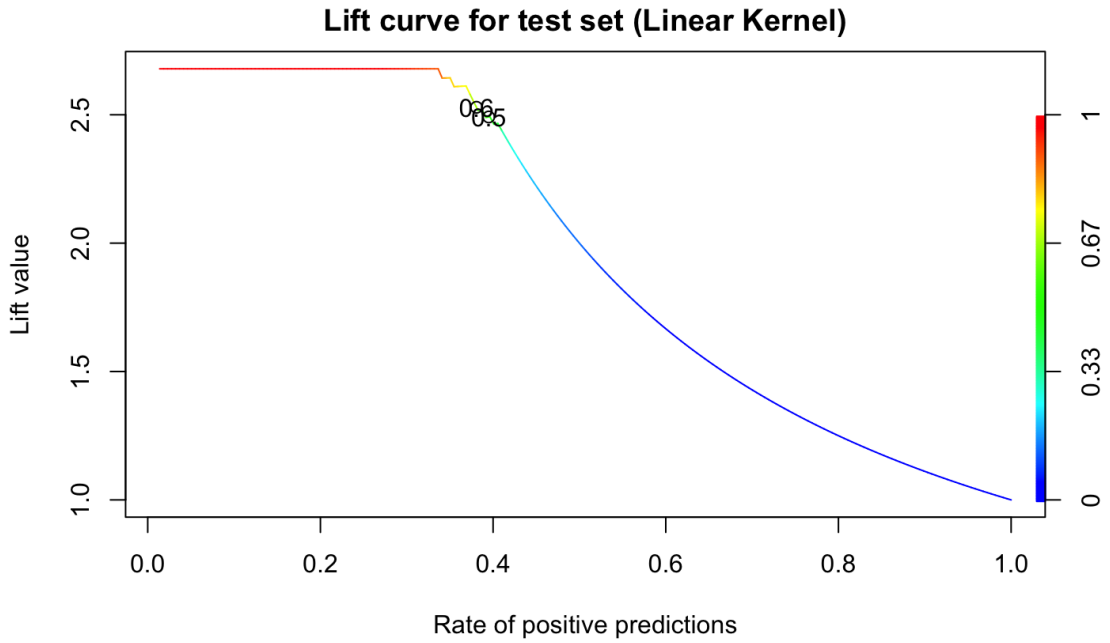


Εικόνα 4.31 *Καμπύλη ROC του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Στην συνέχεια στις εικόνες 4.32 και 33 βρίσκονται οι καμπύλες lift για τα σύνολα εκπαίδευσης και ελέγχου αντίστοιχα, με το μοντέλο με σύνορο απόφασης ίσο με 0.62 να βρίσκεται ψηλότερα, δηλαδή να κατέχει υψηλότερη τιμή ανύψωσης. Αυτό σημαίνει πως για το ποσοστό που των ασθενών που ανήκουν στη θετική κλάση έχουν την απόδοση να ανήκουν όντως στην θετική κλάση ίση με την τιμή ανύψωσης. Για το σύνολο ελέγχουν παρουσιάζεται πάλι η υπεροχή του μοντέλου με σύνορο απόφασης ίσο με 0.62 αλλά είναι ιδιαίτερα μικρή.

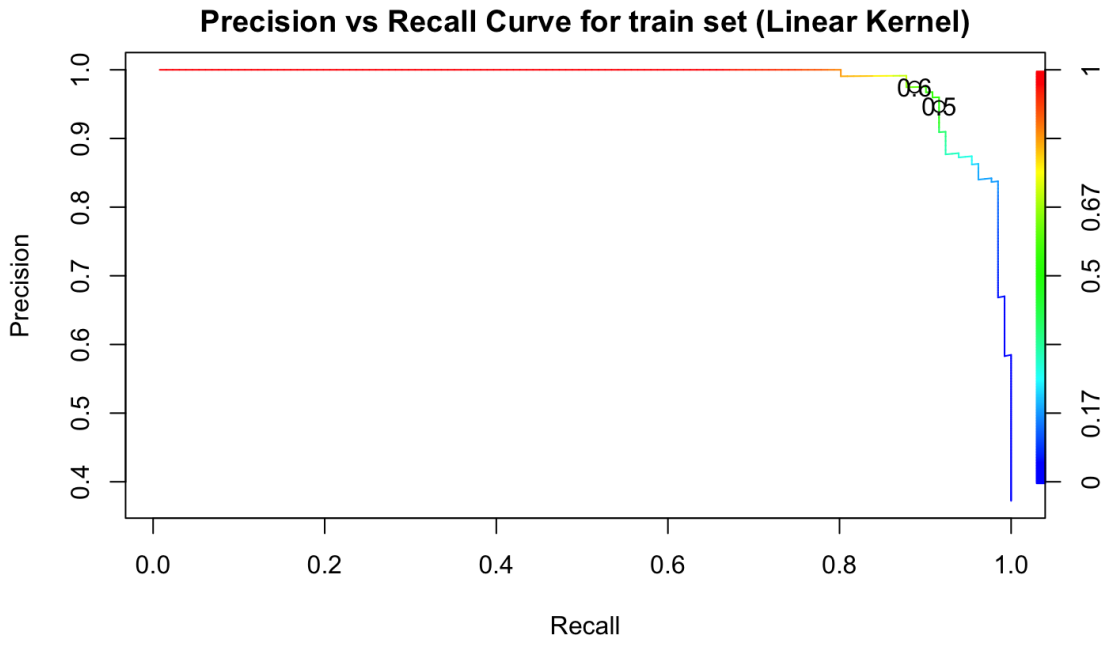


Εικόνα 4.32 *Καμπύλη Lift του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*

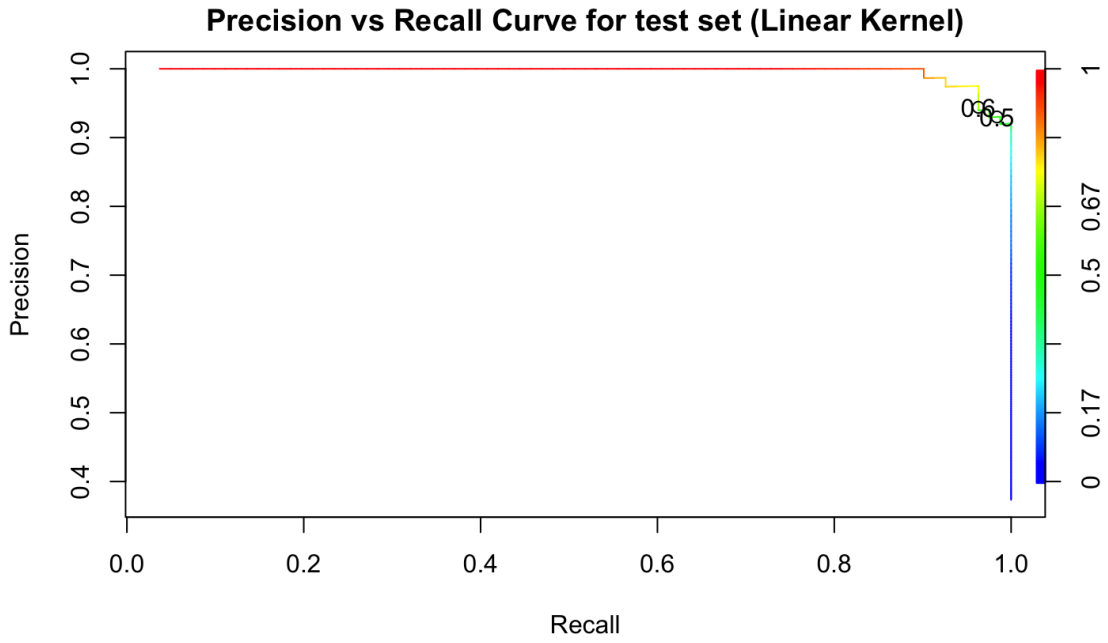


Εικόνα 4.33 *Καμπύλη Lift του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Τέλος, στις εικόνες 4.34 και 35 το μοντέλο για σύνορο απόφασης ίσο με 0.62 υπερέχει έναντι του 0.5 όπως αποδεικνύεται από το διάγραμμα precision vs recall του συνόλου εκπαίδευσης και ελέγχου όσον αφορά το μέτρο precision δίοτι βρίσκεται πιο ψηλά στην καμπύλη ενώ για το μέτρο recall το μοντέλο με το σύνορο απόφασης ίσο με 0.5 διότι βρίσκεται πιο δεξιά.



Εικόνα 4.34 *Καμπύλη Precision vs Recall του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*

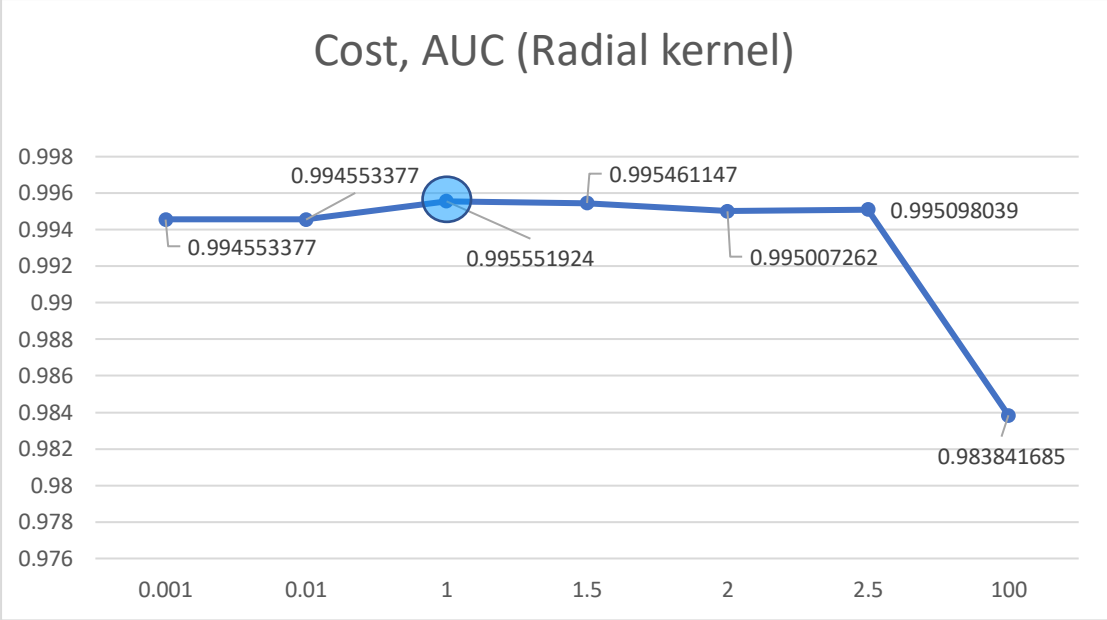


Εικόνα 4.35 *Καμπύλη Precision vs Recall του μοντέλου SVM με γραμμική συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Μετά την γραμμική συνάρτηση πυρήνα, σειρά έχει η γκαουσιανή όπου το σύνολο των δυνητικών τιμών αποτελείται από (0.001, 0.01, 1, 1.5, 2, 2.5, 100). Ακολουθεί η ίδια διαδικασία, υπολογίζοντας τα μέτρα αξιολόγησης των μοντέλων για κάθε υποψήφια τιμή. Στον **πίνακα 4.38** παρουσιάζονται τα μέτρα αξιολόγησης της διαδικασίας αναζήτησης της βέλτιστης τιμής του μοντέλου της γκαουσιανή συνάρτηση πυρήνα, όπως επίσης και η γραφική αναπαράσταση του μέτρου AUC συναρτήσει του κόστους (**εικόνα 4.36)** με την τιμή κόστους ίση με 1 να σημειώνει την υψηλότερη τιμή.

Hyperparameter: Cost (radial kernel)						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
0.001	0.99455	0.94009	0.99200	0.91176	0.98765	0.95019
0.01	0.99455	0.93548	0.99194	0.90441	0.98765	0.94615
1	0.99555	0.97235	0.99242	0.96324	0.98765	0.97761
1.5	0.99546	0.97696	0.99248	0.97059	0.98765	0.98141
2	0.99501	0.97235	0.98507	0.97059	0.97531	0.97778
2.5	0.99510	0.96313	0.97059	0.97059	0.95062	0.97059
100	0.98384	0.92166	0.92806	0.94853	0.87654	0.93818
Best AUC performance = 0.99555 , Cost = 1						

Πίνακας 4. 40 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο cost του μοντέλου γκαουσιανής συνάρτησης πυρήνα.*



Εικόνα 4.36 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει του κόστους του μοντέλου με γκαουσιανή συνάρτηση πυρήνα.*

Σύμφωνα λοιπόν με την διαδικασία βαθμολόγησης (**πίνακας 4.39)** τα αποτελέσματα που προκύπτουν από αυτή στον **πίνακα 4.40**, υποδεικνύουν πως το μοντέλο SVM με γκαουσιανή συνάρτηση πυρήνα ίση με 1.5, σημειώνει την καλύτερη βαθμολογία στα μέτρα αξιολόγησης σε σχέση με τις γειτονικές της και ως εκ τούτου αποτελεί την βέλτιστη τιμή.

Hyperparameter: Cost (radial kernel)						
Cost	auc	accuracy	precision	sensitivity	specificity	F1-score
0.01	0.99455	0.93548	0.99194	0.90441	0.98765	0.94615
1	0.99555	0.97235	0.99242	0.96324	0.98765	0.97761
1.5	0.99546	0.97696	0.99248	0.97059	0.98765	0.98141
Ranking						
Cost	auc	accuracy	precision	sensitivity	specificity	F1-score
0.01	3	3	3	3	1	3
1	1	2	2	2	1	2
1.5	2	1	1	1	1	1

Πίνακας 4. 41 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο cost του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα.*

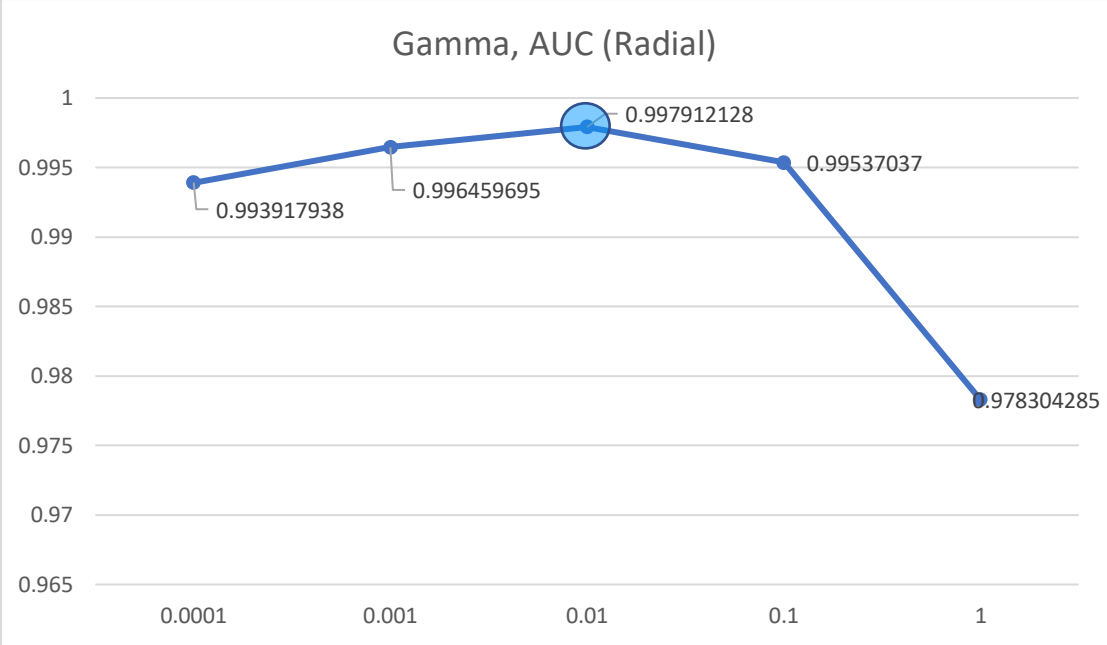
Cost	Average rank
0.01	2.67
1	1.67
1.5	1.17
Best rank = 1.17, Cost = 1.5	

Πίνακας 4. 42 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο Cost του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα*

Η γκαουσιανή συνάρτηση πυρήνα του μοντέλου SVM, εκτός από την υπερπαράμετρο του κόστους (cost), λαμβάνει και την gamma. Το σύνολο δυνητικών τιμών (0.0001, 0.001, 0.01, 0.1, 1, 2) που λαμβάνονται υπόψη για την επιλογή της βέλτιστης καθώς και τα μέτρα αξιολόγησης των μοντέλων για κάθε τιμή αυτού παρουσιάζονται στον **πίνακα 4.41**. Είναι εμφανές πως η τιμή της υπερπαραμέτρου gamma ίση 0.01 σημειώνει την υψηλότερη απόδοση για το μέτρο AUC ίσο με 0.997912, όπως άλλωστε φαίνεται και στην γραφική αναπαράσταση στην **εικόνα 4.37**.

Hyperparameter: Gamma (radial kernel)						
Gamma	auc	accuracy	precision	sensitivity	specificity	F1_score
0.0001	0.99392	0.94009	0.97674	0.92647	0.96296	0.95094
0.001	0.99646	0.95853	0.97744	0.95588	0.96296	0.96654
0.01	0.99791	0.96774	0.99237	0.95588	0.98765	0.97378
0.1	0.99537	0.97235	0.99242	0.96324	0.98765	0.97761
1	0.97830	0.90323	0.88591	0.97059	0.79012	0.92632
Best AUC performance = 0.99791 , Gamma = 0.01						

Πίνακας 4. 43 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο Gamma του μοντέλου γκαουσιανής συνάρτησης πυρήνα.*



Εικόνα 4. 37 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει του gamma του μοντέλου με γκαουσιανή συνάρτηση πυρήνα.*

Όπως λοιπόν πραγματοποιείται η διαδικασία βαθμολόγησης μέχρι στιγμής, λαμβάνοντας υπόψη και τις γειτονικές τιμές αυτής που το μοντέλο σημείωσε την υψηλότερη τιμή για τον δείκτη AUC, η ίδια διαδικασία πραγματοποιείται και εδώ. Ως εκ τούτου στον **πίνακα 4.42** παρουσιάζονται η διαδικασίας βαθμολόγησης ενώ στον **πίνακα 4.31** τα αποτελέσματα αυτής. Βέλτιστη τιμή της υπερπαραμέτρου gamma είναι η 0.1 η οποία βρίσκετε στην πρώτη θέση σε σχέση με τις υπόλοιπες βάσει της καθολικής απόδοσης στα μέτρα αξιολόγησης.

Hyperparameter: Gamma (radial kernel)						
Gamma	auc	accuracy	precision	sensitivity	specificity	F1-score
0.001	0.99646	0.95853	0.97744	0.95588	0.96296	0.96654
0.01	0.99791	0.96774	0.99237	0.95588	0.98765	0.97378
0.1	0.99537	0.97235	0.99242	0.96324	0.98765	0.97761
Ranking						
Gamma	auc	accuracy	precision	sensitivity	specificity	F1-score
0.001	2	3	3	2	3	3
0.01	1	2	2	2	1	2
0.1	3	1	1	1	1	1

Πίνακας 4. 44 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο gamma του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα.*

Gamma	Average rank
0.001	2.67
0.01	1.67
0.1	1.33
Best rank = 1.33 , Gamma = 0.1	

Πίνακας 4. 45 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο gamma του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα.*

Σύμφωνα με τον **πίνακα 4.44**, αποδεικνύεται πως για σύνορο απόφασης ίσο με 0.62, το μοντέλο SVM με γκαουσιανή συνάρτηση πυρήνα, αποδίδει καλύτερα στο σύνολο εκπαίδευσης ενώ για το σύνολο ελέγχου η απόδοση του μοντέλου για σύνορο απόφασης ίσο με 0.62 αποδίδει καλύτερα έναντι του 0.5.

Kernel: Radial				
n = 352	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	217	12	212	11
Predicted M	4	119	9	120

Πίνακας 4. 46 *Confusion matrix για τα δύο σύνορα απόφασης 0.5 και 0.62 του συνόλου εκπαίδευσης για το μοντέλο SVM με γκαουσιανή συνάρτηση πυρήνα και τιμή κόστους 1.5 και gamma ίση με 0.1.*

Kernel: Radial				
n = 217	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	130	1	129	1
Predicted M	6	80	7	80

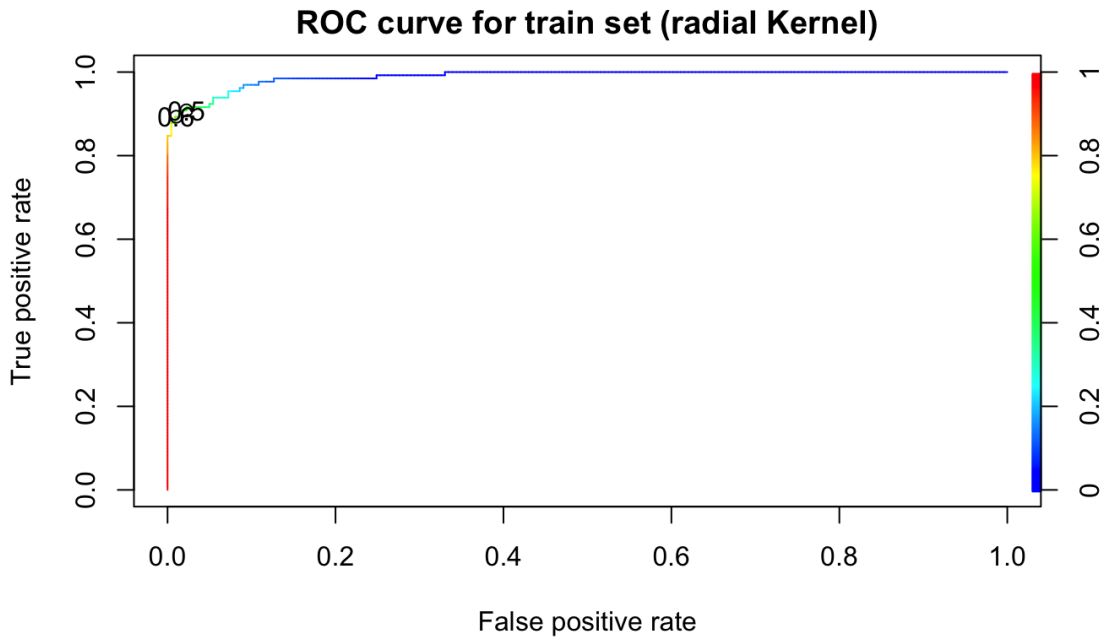
Πίνακας 4. 47 *Confusion matrix για τα δύο σύνορα απόφασης 0.5 και 0.62 του συνόλου εκπαίδευσης για το μοντέλο SVM με γκαουσιανή συνάρτηση πυρήνα και τιμή κόστους 1.5 και gamma ίση με 0.1.*

Έπειτα, στον **πίνακα 4.46**, παρουσιάζονται τα μέτρα αξιολόγησης του μοντέλου για τα δύο σύνορα απόφασης όπως επίσης και για τα δύο σύνολα δεδομένων, εκπαίδευσης και ελέγχου.

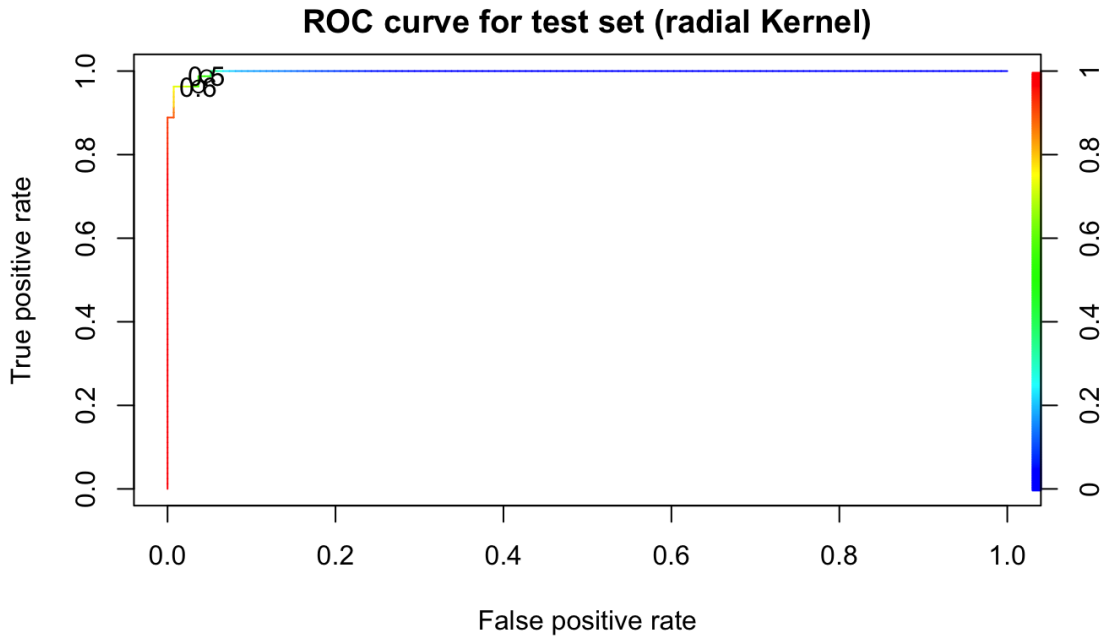
Για το μέτρο accuracy στο σύνολο εκπαίδευσης το μοντέλο με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 95.45% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 94.32%, ενώ στο σύνολο ελέγχου με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 96.774% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 96.31%. Στο σύνολο εκπαίδευσης, για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 94.760%, ενώ για 0.62 ισούται με 95.07%. Στο σύνολο ελέγχου για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 99.236%, όπως επίσης και για σύνορο απόφασης 0.62. Το 98.19% των ασθενών του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως υγιείς για το σύνορο απόφασης 0.5 ενώ για 0.62 το 95.93%. Αντίστοιχα, για το σύνολο ελέγχου το 95.55% για 0.5 και 94.85% για 0.62. Επίσης, το 90.84% των όγκων του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως κακοήθης για το σύνορο απόφασης 0.5 ενώ για 0.62 το 91.600%. Αντίστοιχα, για το σύνολο ελέγχου το 98.7654% για 0.5 και 98.770% για 0.62. Τέλος, για τα μέτρα F1-score και AUC σημειώθηκαν αποδόσεις ίσες με 96.44% και 98.947% για σύνορο απόφασης ίσο με 0.5 του συνόλου εκπαίδευσης ενώ για 0.62 σημειώθηκε 95.495% και 98.947% αντίστοιχα, ενώ στο σύνολο ελέγχου για σύνορο απόφασης 0.5 σημειώθηκε απόδοση στα μέτρα F1-score και AUC, ίση με 973783% και 997912%, όπως επίσης για 0.62 σημειώθηκε 96.9920% και 99.7910% αντίστοιχα.

Kernel: Radial				
Evaluation Metric	Threshold = 0.5		Threshold = 0.62	
	Training set	Test set	Training set	Test set
Accuracy	0.95450	0.9677419	0.94320	0.963100
Precision	0.94760	0.9923664	0.9507	0.992300
Recall (Sensitivity)	0.98190	0.9558823	0.95930	0.948500
Specificity	0.90840	0.9876543	0.91600	0.987700
F1-Score	0.96444	0.973783	0.95495	0.969920
AUC	0.98947	0.997912	0.98947	0.997910

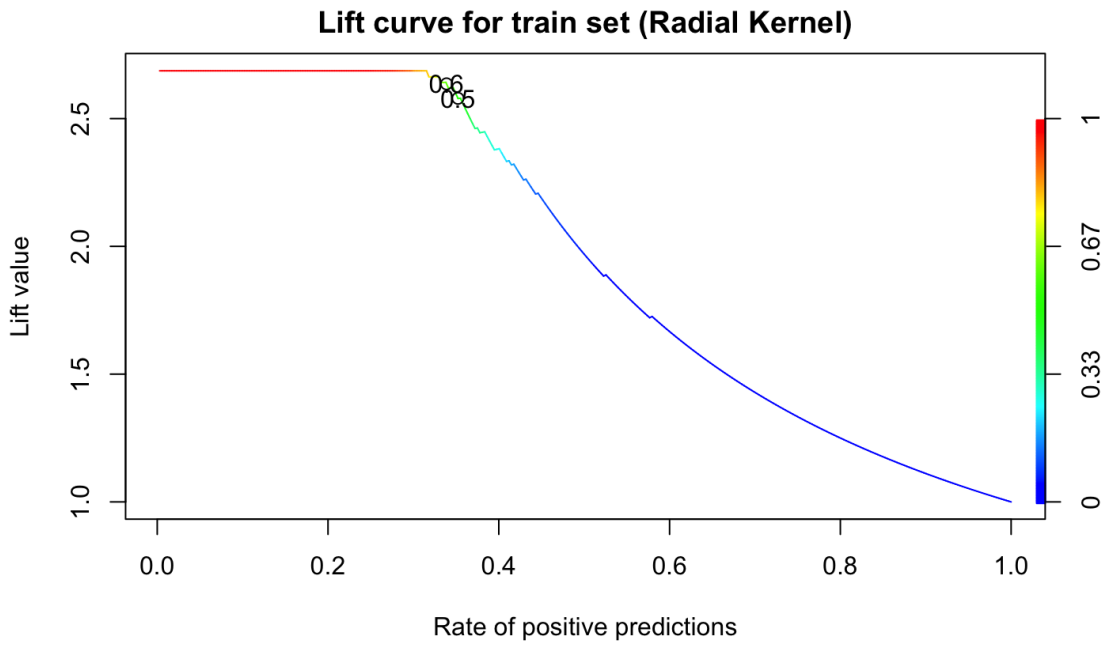
Πίνακας 4. 48 *Μέτρα αξιολόγησης του βέλτιστου μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για τα δύο σύνορα απόφασης του συνόλου εκπαίδευσης και ελέγχου.*



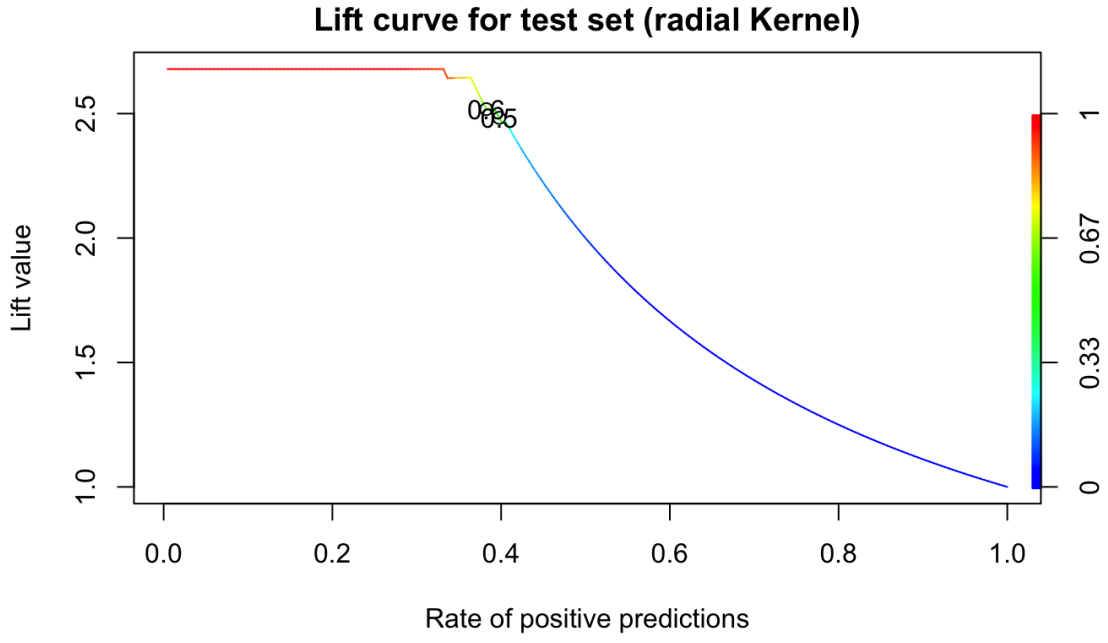
Εικόνα 4.38 *Καμπύλη ROC του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*



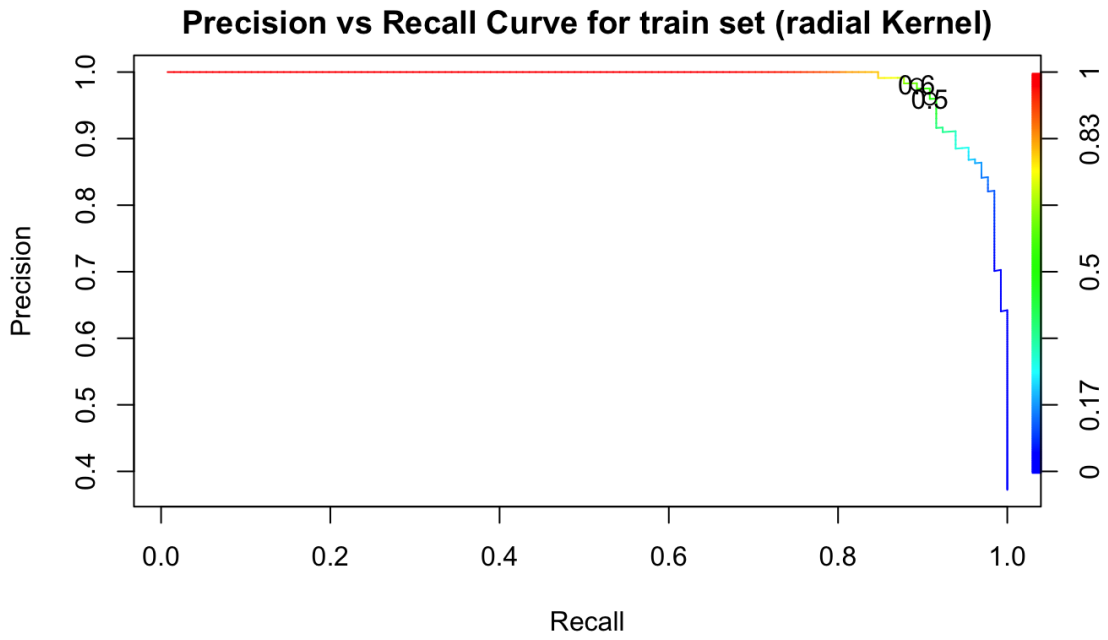
Εικόνα 4.39 *Καμπύλη ROC του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για το σύνολο ελέγχου.*



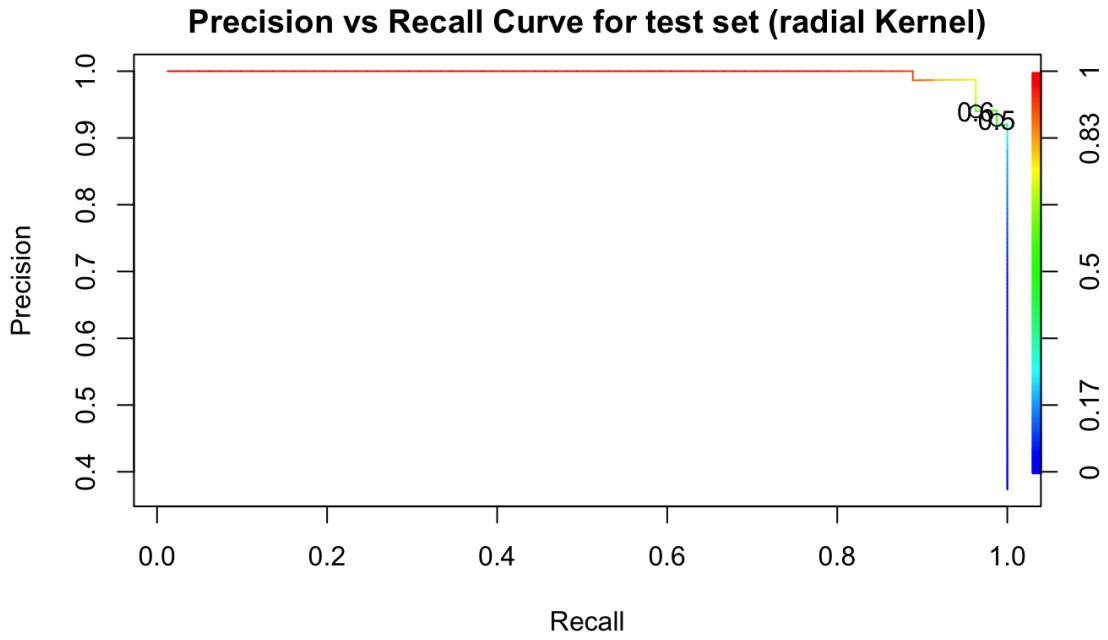
Εικόνα 4.40 *Καμπύλη Lift του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*



Εικόνα 4. 41 *Καμπύλη Lift του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για το σύνολο ελέγχου.*



Εικόνα 4.42 *Καμπύλη Precision vs Recall του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*



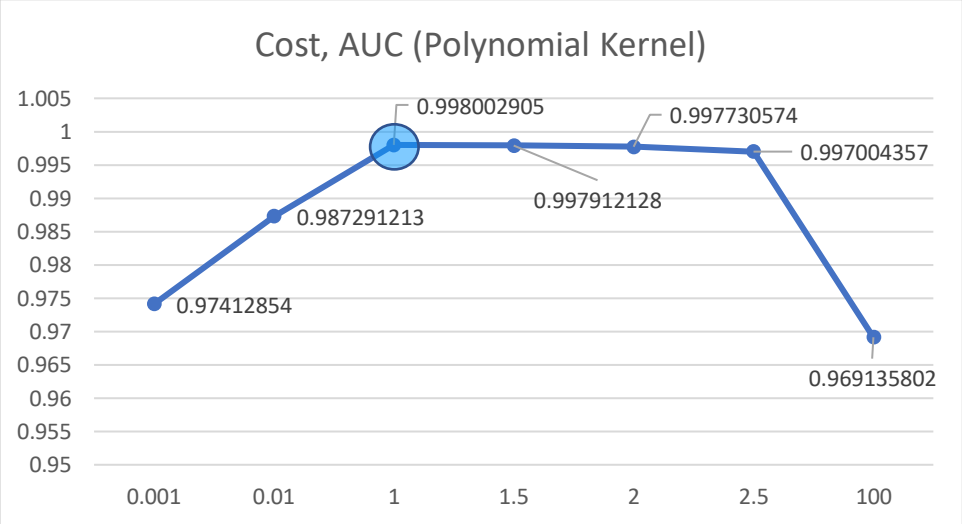
Εικόνα 4. 43 *Καμπύλη Precision vs Recall του μοντέλου SVM με γκαουσιανή συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Τέλος, σειρά έχει η πολυωνυμική συνάρτηση πυρήνα βάσει της οποίας εξετάζονται οι υπερμεταβλητές του κόστους (cost) και ο βαθμός πολυωνύμου (degree). Αρχικά εξετάζεται το ποιά τιμή για την υπερπαράμετρο του κόστους σημειώνει μεγαλύτερη απόδοση στο μέτρο AUC, έναντι των υπόλοιπων υποψήφιων με σύνολο τιμών (0.001, 0.01, 1, 1.5, 2, 2.5, 100).

Για την τιμή κόστους λοιπόν ίση με 1, το μοντέλο SVM με πολυωνυμική συνάρτηση πυρήνα σημειώνει την μέγιστη απόδοση στο μέτρο AUC ίση με 0.998 (**πίνακας 4.47)**.

Hyperparameter: Cost (polynomial kernel)						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
0.001	0.97413	0.89862	0.88514	0.96324	0.79012	0.92254
0.01	0.98729	0.94470	0.94286	0.97059	0.90123	0.95652
1	0.99800	0.96313	0.97761	0.96324	0.96296	0.97037
1.5	0.99791	0.96313	0.97761	0.96324	0.96296	0.97037
2	0.99773	0.96774	0.97778	0.97059	0.96296	0.97417
2.5	0.99700	0.96774	0.97778	0.97059	0.96296	0.97417
100	0.96914	0.94470	0.94286	0.97059	0.90123	0.95652
Best AUC performance = 0.998 , Cost = 1						

Πίνακας 4. 49 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο cost μοντέλου πολυωνυμικής συνάρτησης πυρήνα.*



Εικόνα 4.44 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει του κόστους του μοντέλου με πολυωνυμική συνάρτηση πυρήνα.*

Για την διαδικασία βαθμολόγησης λαμβάνονται υπόψιν και οι γειτονικές τιμές κόστους (0.02 και 1.5) **(πίνακας 4.48)** αποτελέσματα της οποίας φαίνονται στον **πίνακα 4.49** με καταλήγουσα τιμή ως βέλτιστη ίση με 1 με μέγιστη μέση βαθμολογία ίση με 1.17.

Hyperparameter: Degree (polynomial kernel)						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
2	0.82725	0.79724	0.78049	0.94118	0.55556	0.85333
3	0.99800	0.96313	0.97761	0.96324	0.96296	0.97037
4	0.83878	0.69585	0.67677	0.98529	0.20988	0.80240
Ranking						
degree	auc	accuracy	precision	sensitivity	specificity	F1-score
2	3	2	2	3	2	2
3	1	1	1	2	1	1
4	2	3	3	1	3	3

Πίνακας 4. 50 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο κόστους του μοντέλου SVM με πολυωνυμικής συνάρτηση πυρήνα.*

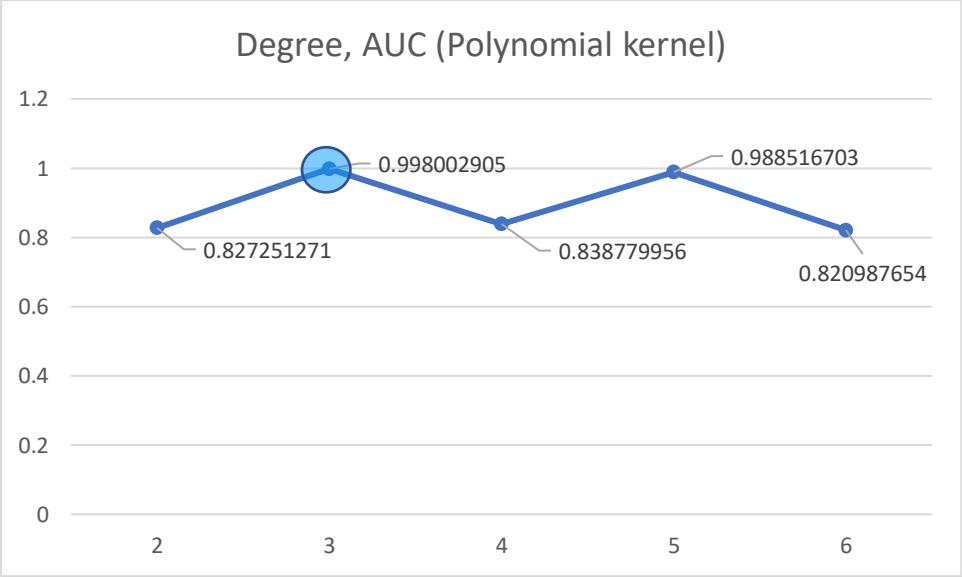
Cost	Average rank
0.01	2.67
1	1.17
1.5	1.33
Best rank = 1.17 , Cost = 1	

Πίνακας 4. 51 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο cost του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα.*

Τέλος, τελευταία υπερπαράμετρο η οποία χρήζει βελτιστοποίησης είναι αυτή του βαθμού πολυωνύμου του μοντέλου SVM. Το σύνολο των υποψήφιων τιμών αποτελείται από τις τιμές (2, 3, 4, 5, 6). Να σημειωθεί πως ο μοναδιαίος βαθμός πολυωνύμου δεν λαμβάνεται υπόψη δίοτι ουσιαστικά έχει εξεταστεί στη γραμμική συνάρτηση πυρήνα. Από τον **πίνακα 4.50**, αποδεικνύεται πως ο βαθμός πολυωνύμου ίσος με 3 σημειώνει τη μέγιστη απόδοση για το μέτρο AUC ίσο με 0.998, ενώ στην **εικόνα 4.45** φαίνεται η γραφική απεικόνιση των τιμών που λαμβάνει το μέτρο AUC, συναρτήσει του βαθμού πολυωνύμου.

Hyperparameter: Degree (polynomial kernel)						
degree	auc	accuracy	precision	sensitivity	specificity	F1_score
2	0.82725	0.79724	0.78049	0.94118	0.55556	0.85333
3	0.99800	0.96313	0.97761	0.96324	0.96296	0.97037
4	0.83878	0.69585	0.67677	0.98529	0.20988	0.80240
5	0.98852	0.92166	0.89404	0.99265	0.80247	0.94077
6	0.82099	0.63594	0.63256	1.00000	0.02469	0.77493
Best AUC performance = 0.998, degree = 3						

Πίνακας 4. 52 *Μέτρα αξιολόγησης για τις υποψήφιες τιμές για την υπερπαράμετρο degree μοντέλου πολυωνυμικής συνάρτησης πυρήνα.*



Εικόνα 4.45 *Γραφική απεικόνιση των τιμών του μέτρου AUC συναρτήσει του βαθμού πολυωνύμου του μοντέλου με πολυωνυμική συνάρτηση πυρήνα.*

Έπειτα από την διαδικασία βαθμολόγησης η οποία λαμβάνει χώρα στον **πίνακα 4.50**, τα αποτελέσματα της οποίας φαίνονται στον **πίνακα 4.51**. Για βαθμό πολυωνύμου λοιπόν ίσο με 3, το μοντέλο SVM με πολυωνυμική συνάρτηση πυρήνα ως βέλτιστη τιμή πολυωνύμου σημειώνει την μέση μέγιστη απόδοση στα μέτρα αξιολόγησης.

Hyperparameter: Degree (polynomial kernel)						
Cost	auc	accuracy	precision	sensitivity	specificity	F1_score
2	0.82725	0.79724	0.78049	0.94118	0.55556	0.85333
3	0.99800	0.96313	0.97761	0.96324	0.96296	0.97037
4	0.83878	0.69585	0.67677	0.98529	0.20988	0.80240
Ranking						
degree	auc	accuracy	precision	sensitivity	specificity	F1-score
2	3	2	2	3	2	2
3	1	1	1	2	1	1
4	2	3	3	1	3	3

Πίνακας 4. 53 *Διαδικασία βαθμολόγησης των μέτρων αξιόλογης για την εύρεση της βέλτιστης τιμής για την υπερπαράμετρο degree του μοντέλου SVM με πολυωνυμικής συνάρτηση πυρήνα.*

degree	Average rank
2	2.33
3	1.17
4	2.50
Best rank = 1.17 , degree = 3	

Πίνακας 4. 54 *Αποτελέσματα διαδικασίας βαθμολόγησης για την υπερπαράμετρο degree του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα.*

Kernel: Polynomial				
n = 352	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	219	18	218	9
Predicted M	2	113	3	122

Πίνακας 4. 55 *Confusion matrix για τα δύο σύνορα απόφασης 0.5 και 0.62 του συνόλου εκπαίδευσης για το μοντέλο SVM με πολυωνυμική συνάρτηση πυρήνα, τιμή κόστους 1.5 και gamma ίση με 1.*

Kernel: Polynomial				
n = 217	Threshold = 0.5		Threshold = 0.62	
	Actual B	Actual M	Actual B	Actual M
Predicted B	131	3	129	1
Predicted M	5	78	7	80

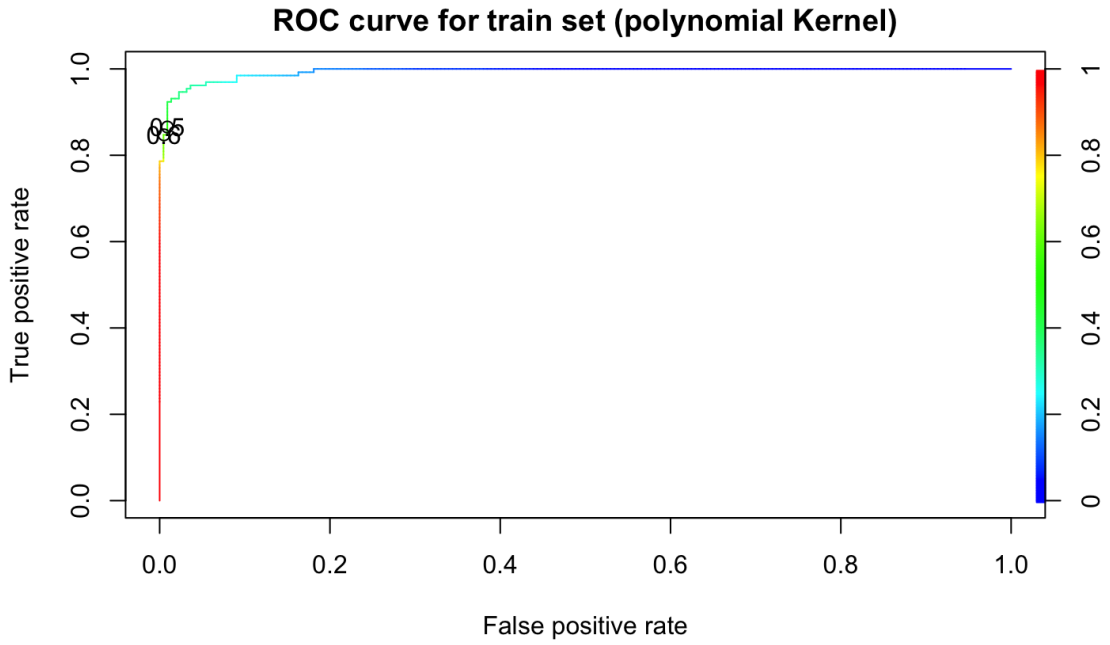
Πίνακας 4. 56 *Confusion matrix για τα δύο σύνορα απόφασης 0.5 και 0.62 του συνόλου ελέγχου για το μοντέλο SVM με πολυωνυμική συνάρτηση πυρήνα, τιμή κόστους 1.5 και gamma ίση με 1.*

Έπειτα, στον **πίνακα 4.54**, παρουσιάζονται τα μέτρα αξιολόγησης του μοντέλου με πολυωνυμική συνάρτηση πυρήνα για τα δύο σύνορα απόφασης όπως επίσης και για τα δύο σύνολα δεδομένων, εκπαίδευσης και ελέγχου. Για το μέτρο accuracy στο σύνολο εκπαίδευσης το μοντέλο με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 94.32% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 96.590%, ενώ στο σύνολο ελέγχου με σύνορο απόφασης 0.5 ταξινομεί στη σωστή κλάση το 96.313% των παρατηρήσεων έναντι του συνόρου 0.62 το οποίο ταξινομεί το 96.310%. Στο σύνολο εκπαίδευσης, για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 92.41%, ενώ για 0.62 ισούται με 96.04%. Στο σύνολο ελέγχου για σύνορο απόφασης 0.5, το ποσοστό των ασθενών οι οποίοι ταξινομήθηκαν σωστά στη θετική κλάση και είναι όντως υγιής, ισούται με ποσοστό 97.761%, ενώ για το σύνορο απόφασης 0.62 ισούται με 99.23%. Το 99.10% των ασθενών του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως υγιείς για το σύνορο απόφασης 0.5 ενώ για 0.62 το 98.64%. Αντίστοιχα, για το σύνολο ελέγχου το 96.324% για 0.5 και 94.85% για 0.62. Επίσης, το 86.260% των όγκων του συνόλου εκπαίδευσης ταξινομήθηκαν σωστά ως κακοήθης για το σύνορο απόφασης 0.5 ενώ για 0.62 το 93.130%. Αντίστοιχα, για το σύνολο ελέγχου το 96.296% για 0.5 και 98.770% για 0.62. Τέλος, για τα μέτρα F1-score και AUC σημειώθηκαν αποδόσεις ίσες με 95.633% και 99.364% για σύνορο απόφασης ίσο με 0.5 του συνόλου εκπαίδευσης ενώ για 0.62 σημειώθηκε 97.321% και 98.947% αντίστοιχα, ενώ στο σύνολο ελέγχου για σύνορο απόφασης 0.5 σημειώθηκε απόδοση στα μέτρα F1-score και AUC, ίση με 97.037% και 99.8%, όπως επίσης για 0.62 σημειώθηκε 96.9920% και 99.8% αντίστοιχα.

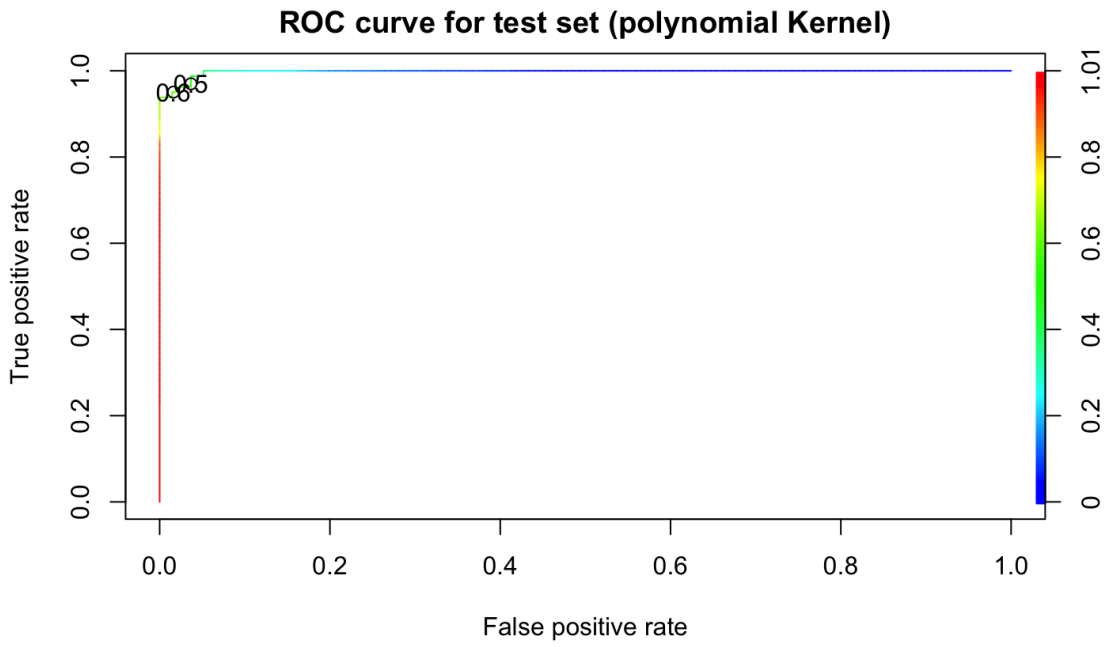
Kernel: Polynomial				
Evaluation Metric	Threshold = 0.5		Threshold = 0.62	
	Training set	Test set	Training set	Test set
Accuracy	0.94320	0.96313	0.96590	0.96310
Precision	0.92410	0.97761	0.96040	0.99230
Recall (Sensitivity)	0.99100	0.96324	0.9864	0.94850
Specificity	0.86260	0.96296	0.93130	0.98770
F1-Score	0.95633	0.97037	0.97321	0.96992
AUC	0.99364	0.99800	0.99364	0.99800

Πίνακας 4. 57 *Μέτρα αξιολόγησης του βέλτιστου μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για τα δύο σύνορα απόφασης του συνόλου εκπαίδευσης και ελέγχου.*

Οι καμπύλες ROC για το σύνολο εκπαίδευσης και ελέγχου παρουσιάζονται παρακάτω στις **εικόνες 4.46** και **4.47** αντίστοιχα βάσει των οποίων προκύπτουν τα μέτρα AUC όντας το εμβαδό κάτω από τις καμπύλες. Είναι ευδιάκριτη η υπεροχή του μοντέλου με σύνορο απόφασης ίσο με 0.5 έναντι του 0.62.

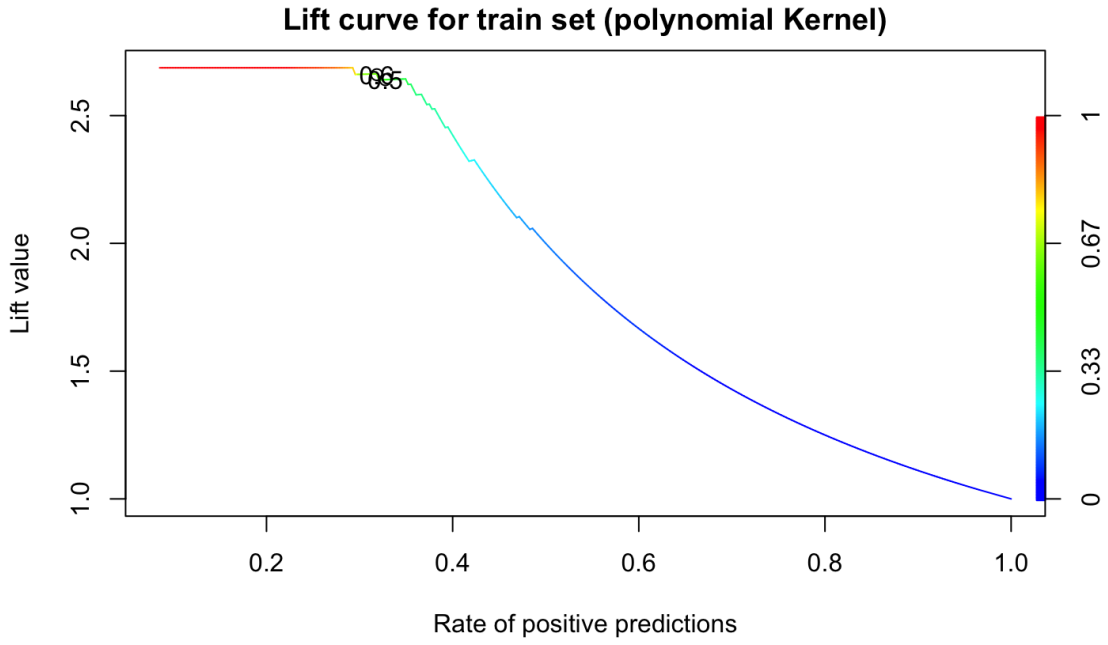


Εικόνα 4.46 *Καμπύλη ROC του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*

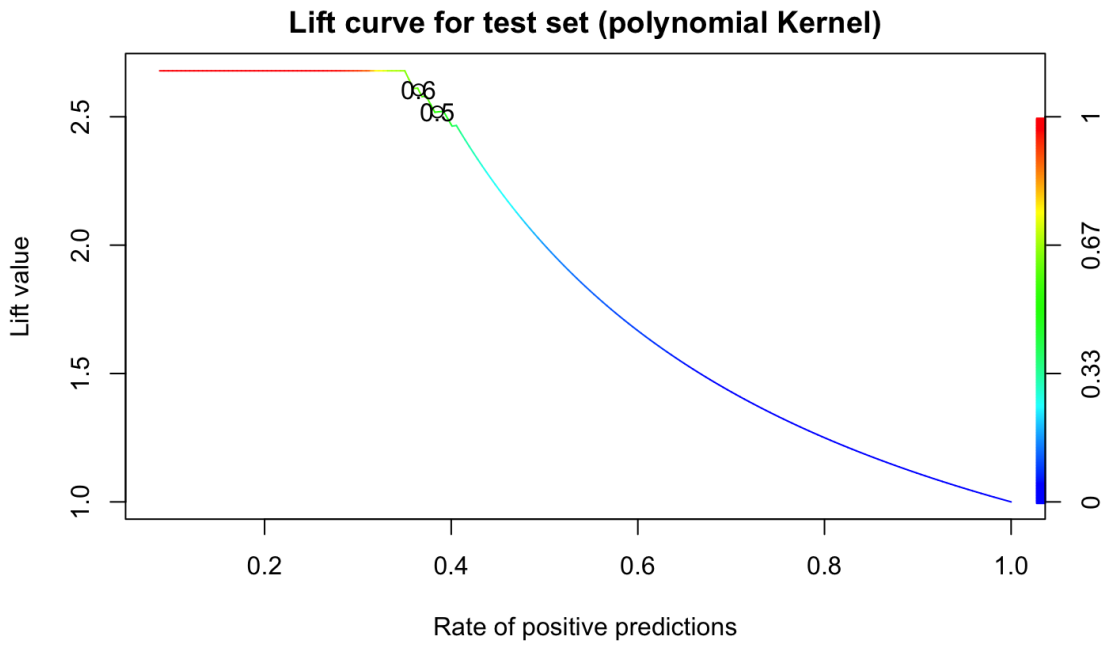


Εικόνα 4.47 *Καμπύλη ROC του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Επίσης, στις **εικόνες 4.48** και **4.49** οι καμπύλες lift για τα δύο σύνολα δεδομένων, με το σύνορο απόφασης ίσο με 0.62 να κατέχει μεγαλύτερη τιμή ανύψωσης έναντι του 0.5 τόσο στο σύνολο εκπαίδευσης αλλά και στο σύνολο ελέγχου.

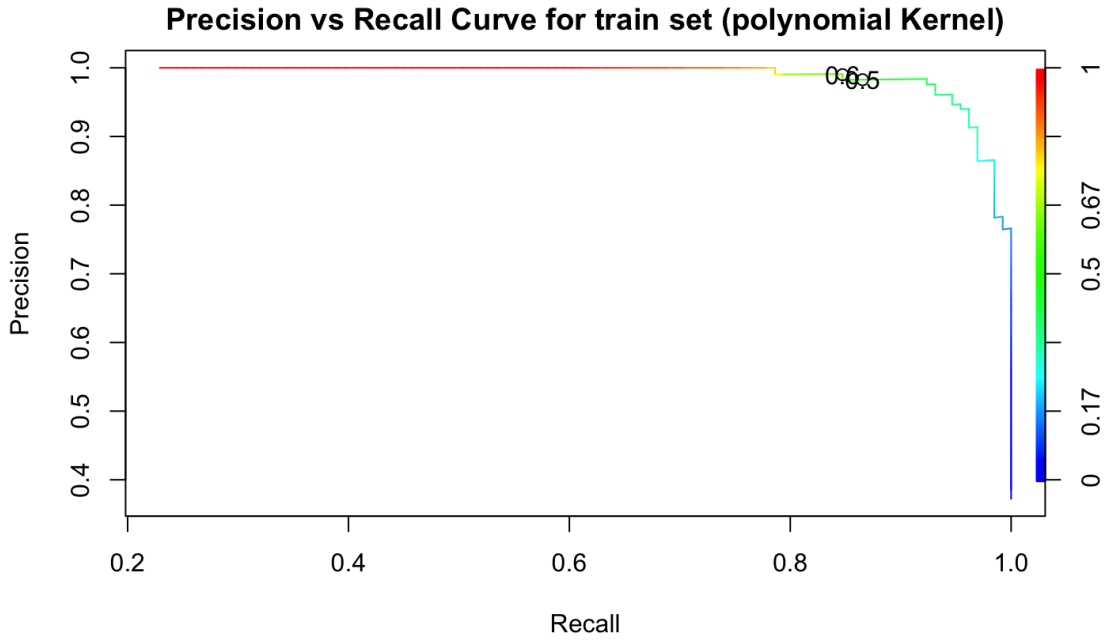


Εικόνα 4.48 *Καμπύλη Lift του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*

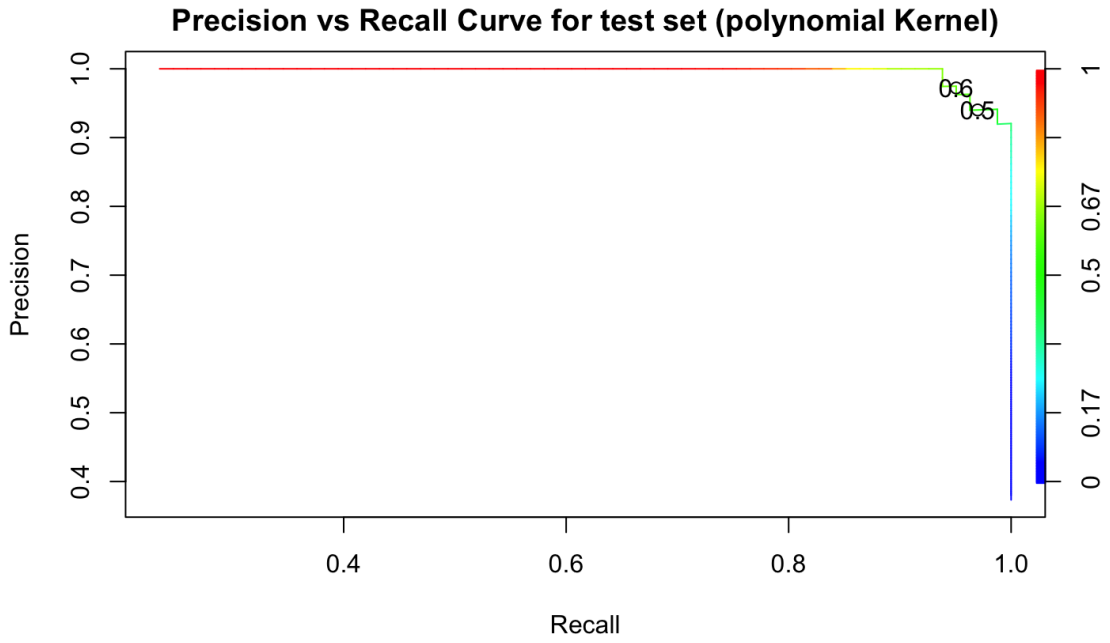


Εικόνα 4.49 *Καμπύλη Lift του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Τέλος, στις **εικόνες 4.49** και **4.50** βρίσκονται τα διαγράμματα precision vs recall του συνόλου εκπαίδευσης και ελέγχου.



Εικόνα 4.50 *Καμπύλη Precision vs Recall του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για το σύνολο εκπαίδευσης.*



Εικόνα 4.50 *Καμπύλη Precision vs Recall του μοντέλου SVM με πολυωνυμική συνάρτηση πυρήνα για το σύνολο ελέγχου.*

Συνοψίζοντας, οι τιμές των υπερπαραμέτρων για τα μοντέλα SVM βάσει των οποίων σημειώνουν τη μέγιστη μέση απόδοση στα μέτρα αξιολόγησης είναι για γραμμική συνάρτηση πυρήνα κόστος ίσο με 0.01, για γκαουσιανή κόστος ίσο με 1.5 και gamma ίσο με 0.1 και τέλος για πολυωνυμική κόστος ίσο με 1 και βαθμό πολυωνύμου ίσο με 3.

4.6 Σύγκριση μεθόδων

Σε αυτό το κομμάτι της διπλωματικής εργασίας θα λάβει χώρα η συγκέντρωση αλλά και η σύγκριση των μέτρων αξιόλογης όλων των αλγορίθμων μηχανικής μάθησης που έχουν εκπαιδευτεί έτσι ώστε να αναδειχτεί αυτός ο οποίος σημείωσε τη μέγιστη μέση απόδοση. Παρακάτω στον **πίνακα 4.50** παρουσιάζονται όλα τα μοντέλα καθώς και το σύνορο απόφασης το οποίο υπερέχει και βελτιστοποιεί το κάθε μοντέλο. Έπειτα, με εφαρμόζοντας τη διαδικασία βαθμολόγησης διακρίνεται η θέση την οποία κατέχει ο κάθε αλγόριθμος στο κάθε μέτρο αξιολόγησης. Αναλυτικότερα, ξεκινώντας με το υπόδειγμα της λογιστικής παλινδρόμησης το οποίο αποδίδει καλύτερα για σύνορο απόφασης ίσο με 0.62 έναντι του 0.5, κατέχει την προ τελευταία θέση , ενώ βάσει του μέτρου AUC, βρίσκεται στην τέταρτη θέση. Βάσει αυτών των αποτελεσμάτων το υπόδειγμα λογιστικής παλινδρόμησης βρίσκεται στην προ τελευταία θέση γεγονός πως σημαίνει ότι το υπόδειγμα λογιστικής παλινδρόμησης έχει μικρή προβλεπτική ικανότητα. Στη συνέχεια ακολουθεί το υπόδειγμα των δέντρων απόφασης με σύνορο απόφασης

ίσο με 0.5 το οποίο βρίσκεται στην τελευταία θέση σε όλα τα μέτρα αξιολόγησης και κατ' επέκτασιν στην ιεραρχία, σημειώνοντας την πιο φτωχή προβλεπτική ικανότητα. Έπειτα ο αλγόριθμος μηχανικής μάθησης τυχαίου δάσους βρίσκεται στην δεύτερη θέση της βαθμολογίας όσον αφορά την μέση απόδοση του στα μέτρα αξιολόγησης. Πιο αναλυτικά κατέχει την πρώτη θέση στα μέτρα specificity και F1-score, ενώ κατέχει την πέμπτη θέση στο μέτρο recall, την τέταρτη στα μέτρα accuracy και AUC ενώ την δεύτερη στο μέτρο precision. Τέλος, ο αλγόριθμος SVM εξετάζεται ξεχωριστά ανάλογα με την συνάρτηση πυρήνα. Ως εκ τούτου, το μοντέλο με γραμμική συνάρτηση κατέχει την τέταρτη θέση στην συνολική βαθμολογία, τη δεύτερη το μοντέλο με πολυωνυμική συνάρτηση και την πρώτη θέση το μοντέλο με γκαουσιανή συνάρτηση πυρήνα, παρουσιάζοντας τη μέγιστη μέση απόδοση έναντι των υπολοίπων αλγορίθμων μηχανικής μάθησης.

Threshold	**Algorithm**	**Accuracy**	**Precision**	**Recall (Sensitivity)**	**Specificity**	**F1-Score**	**AUC**
0.62	Logistic Regression	0.95852	0.95683	0.97794	0.92593	0.96727	0.99174
0.5	Decision Trees	0.92170	0.93430	0.94120	0.88890	0.93773	0.94794
0.62	Random Forest	0.96310	0.99230	0.94850	0.98770	0.97760	0.99680
0.5	SVM Linear	0.96313	0.98485	0.95588	0.97531	0.97015	0.99764
0.5	SVM Radial	0.96774	0.99237	0.95588	0.98765	0.97378	0.99791
0.50	SVM Polynomial	0.96313	0.97761	0.96324	0.96296	0.97037	0.99800
Ranking							
Threshold	**Algorithm**	**Accuracy**	**Precision**	**Recall (Sensitivity)**	**Specificity**	**F1-Score**	**AUC**
0.62	Logistic Regression	5	5	1	5	5	5
0.5	Decision Trees	6	6	6	6	6	6
0.62	Random Forest	4	2	5	1	1	4
0.5	SVM Linear	2	3	3	3	4	3
0.5	SVM Radial	1	1	3	2	2	2
0.50	SVM Polynomial	2	4	2	4	3	1

Πίνακας 4. 58 *Πίνακας συγκεντρωτικών αποτελεσμάτων των μοντέλων ταξινόμησης*

Threshold	Algorithm	Average rank
0.62	Logistic Regression	4.33
0.50	Decision Trees	6.00
0.62	Random Forest	2.83
0.50	SVM Linear	3.00
0.50	SVM Radial	1.83
0.50	SVM Polynomial	2.67

Πίνακας 4. 59 *Αποτελέσματα βαθμολογίας του συγκεντρωτικού πίνακα αποτελεσμάτων*

Συνεπώς, προκύπτει πως ο καταλληλότερος αλγόριθμος με τη μέγιστη μέση προβλεπτική ικανότητα είναι αυτός των μηχανών διανυσμάτων υποστήριξης με γκαουσιανή συνάρτηση πυρήνα.

Κεφάλαιο 5: Συμπέρασμα

Στο πέμπτο και τελευταίο κεφάλαιο της διπλωματικής εργασίας παραθέτονται τα αποτελέσματα που προέκυψαν από τα πειράματα του εργαστηριακού σκέλους στο κεφάλαιο 4, έτσι ώστε να προσδιοριστούν τα σημαντικότερα στατιστικά στοιχεία από την μοντελοποίηση

χαρακτηριστικών κυτταρικού πυρήνα από δεδομένα εικόνας για τη διάγνωση του όγκου του μαστού.

Αρχικά στην εισαγωγή έγινε αναφορά στη σύναψη της τεχνολογίας με τις θετικές επιστήμες για την δημιουργία κατάλληλων ‘’εργαλείων’’ και τα προβλήματα της σύγχρονης κοινωνίας και κατ’ επέκτασιν στη σύγχρονη αγορά τα οποία επιλύονται μέσω αυτών αλλά με σκοπό αλλά και τα και επιλύονται Στο κεφάλαιο 1, έγινε εμβάθυνση στο πεδίο έρευνας πάνω στο οποίο βασίζεται η διπλωματική εργασία. Στο κεφάλαιο 2 έγινε αναφορά στη μηχανική μάθηση και στους αλγορίθμους εποπτευόμενης μάθησης. Έγινε αναλυτική αναφορά στους αλγορίθμους λογιστικής παλινδρόμησης (logistic regression), δέντρων απόφασης (decision trees), τυχαίου δάσους (Random Forest) και μηχανών διανυσμάτων υποστήριξης (SVM – Support Vector Machine), έγινε σχολιασμός ο τρόπος λειτουργίας βάσει του οποίου καταλήγει η επιλογή της επικρατούσας κλάσης σε προβλήματα ταξινόμησης όπως επίσης και των μέτρων αξιολόγησης.

Στο κεφάλαιο 3, έγινε μια αναλυτική περιγραφή των στοιχείων του συνόλου δεδομένων που χρησιμοποιήθηκαν για το εργαστηριακό σκέλος μαζί με τα περιγραφικά στατιστικά τους χαρακτηριστικά οπτικοποιώντας το συγκεκριμένο σύνολο δεδομένων με τη χρήση γραφημάτων για την πλήρη κατανόηση και αντίληψη των στατιστικών στοιχείων που χρησιμοποιούνται για τη μοντελοποίηση χαρακτηριστικών κυτταρικού πυρήνα από δεδομένα εικόνας.

Στο κεφάλαιο 4 έγινε χρήση των αλγορίθμων μηχανικής μάθησης με σκοπό την εκπαίδευση και δημιουργία υποδειγμάτων θέτοντας ως στόχο την αξιολόγηση της προβλεπτικής τους ικανότητας και κατ’ επέκτασιν την πρόβλεψη της καλοήθειας ενός όγκου του μαστού και στον εντοπισμό των καταλυτικών παραγόντων που διαχωρίζουν την καλοήθεια από την κακοήθεια ενός όγκου του μαστού. Για την επίλυση του προβλήματος χρησιμοποιήθηκαν μοντέλα ταξινόμησης τα οποία εκπαιδεύτηκαν στο περιβάλλον της R. Ειδικότερα, χρησιμοποιήθηκαν οι αλγόριθμοι λογιστικής παλινδρόμησης (logistic regression), δέντρων απόφασης (decision trees), τυχαίου δάσους (Random Forest) και μηχανών διανυσμάτων υποστήριξης (SVM – Support Vector Machine) με γραμμική (Linear), γκαουσιανή (Radial) και πολυωνυμική (Polynomial) συνάρτηση πυρήνα με στόχο τη δημιουργία υποδειγμάτων για την ταξινόμηση των δεδομένων σε κάθε κλάση. Επιπροσθέτως, χρησιμοποιήθηκαν διάφορες τεχνικές βελτιστοποίησης για τον προσδιορισμό των παραμέτρων του κάθε παραγόμενου μοντέλου λαμβάνοντας υπόψιν ότι το ποσοστό των καλοηθών όγκων μαστού στο συγκεκριμένο σύνολο

είναι 62%. Έπειτα αξιολογήθηκαν τα μοντέλα που δημιουργήθηκαν από τους παραπάνω αλγορίθμους ενώ στο τέλος αξιολογήθηκε η προβλεπτική τους ικανότητα με τη χρήση των μέτρων αξιολόγησης.

Τέλος, κάποιες ιδέες για περαιτέρω οι οποίες θα μπορούσαν να εμπλουτίσουν την έρευνα θα ήταν:

1. η αύξηση του μεγέθους του δείγματος δεδομένων σε τέτοιο βαθμό έτσι ώστε να εκπαιδευτούν σε περισσότερα δεδομένα οι αλγόριθμοι χωρίς να προκληθεί το φαινόμενο της υπέρ-προσαρμογής.
2. Εκπαίδευση περισσοτέρων αλγορίθμων μηχανικής αλλά και βαθιάς μάθησης όπως:
    2.1.1 LDA/QDA
    2.1.2 Gradient Boosting Methods (GBM, XGBoost & LightGBM)
3. Περισσότερες επεξηγηματικές μεταβλητές
4. Πιο εξαντλητικό grid search και ranking όπως:
    4.1 Μεγαλύτερο πλέγμα υποψήφιων τιμών για κάθε υπέρ-παράμετρο
    4.2 Προσθήκη επιπλέων μέτρων αξιολόγησης (π.χ. negative predictive value, balanced accuracy)

Aiolli, F., & Sperduti, A. (2005). *Multiclass Classification with Multi-Prototype Support Vector Machines.* Ανάκτηση 12 30, 2022, από https://www.jmlr.org/papers/volume6/aiolli05a/aiolli05a.pdf

Akinsola, J. (2017). *Supervised Machine Learning Algorithms: Classification and Comparison.* Ανάκτηση 12 30, 2022, από https://www.researchgate.net/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison

Bagchi, T. (2022, 1 1). *Support Vector Machines--An Overview.* Ανάκτηση 12 30, 2022, από https://www.researchgate.net/publication/358021073_Support_Vector_Machines--An_Overview

Bottaci, L., Drew, P., Hartley, J., Hadfield, M., Farouk, R., & Lee, P. (1997). *Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions.* Ανάκτηση 11 11, 2022, από https://www.sciencedirect.com/science/article/pii/S014067369611196X/pdfft?md5=a3edf7b5fa9e23ae7bdfb89d5bf48064&pid=1-s2.0-S014067369611196X-main.pdf

Cicchetti, D. (1992). *Neural networks and diagnosis in the clinical laboratory: state of the art.* Ανάκτηση 11 11, 2022, από https://watermark.silverchair.com/clinchem0009.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAtIwggLOBgkqhkiG9w0BBwagggK_MIICuwIBADCCArQGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMAqT57EOoU3YqA6LbAgEQgIICha0fE2ME196XLBjvzqjyhszVzlUc_wyHwtUZBfIGbS

Elkington, L., Adhikari, P., & Pradhan, P. (2022). *Fractal Dimension Analysis to Detect the Progress of Cancer Using Transmission Optical Microscopy*. Retrieved from mdip: https://www.mdpi.com/2673-4125/2/1/5

Exarchos, K., Goletsis, Y., & Fotiadis, D. (2012). *Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence.* Ανάκτηση 11 11, 2022, από https://www.researchgate.net/profile/Dimitrios-Fotiadis/publication/51588389_Multiparametric_Decision_Support_System_f

or_the_Prediction_of_Oral_Cancer_Reoccurrence/links/56531b6a08aefe619b190340/Multiparametric-Decision-Support-System-for-the-Prediction-o

Frost, J. (2022). *The Difference between Linear and Nonlinear Regression Models.* Ανάκτηση 12 30, 2022, από https://statisticsbyjim.com/regression/difference-between-linear-nonlinear-regression-models/

Hae, K. J. (2019). *Multicollinearity and misleading statistical results.* Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/

Homer, M. J. (1996). *Mammographic Interpretation: A Practical Approach 2nd Edition.* McGraw-Hill Professional.

Kadam, V., Kanhere, S., & Mahindrakar, S. (2020, 10). *Regression Techniques in Machine Learning &Applications: A Review.* Ανάκτηση 12 30, 2022, από https://d1wqtxts1xzle7.cloudfront.net/64866432/32019-libre.pdf?1604664314=&response-content-disposition=inline%3B+filename%3DRegression_Techniques_in_Machine_Learnin.pdf&Expires=1676031211&Signature=cCagt1iyZIbedVu4Ut8THatpiQskHBinmiAdivKfwyUUay88e8GubE~p

Karimi, Z. (2021, 10 1). *Confusion Matrix.* Ανάκτηση 12 30, 2022, από https://www.researchgate.net/publication/355096788_Confusion_Matrix

Kim, J. H. (2019). *Multicollinearity and misleading statistical results*. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/

Kim, J. H. (2019). *Multicollinearity and misleading statistical results*. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *National Library of Medecine*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). *Machine learning: a review of classification and combining techniques.* Ανάκτηση 12 30, 2022, από https://www.researchgate.net/profile/P-

Pintelas/publication/226525180_Machine_learning_A_review_of_classification_and_combining_techniques/links/0fcfd5119227feb83c000000/Machine-learning-A-review-of-classification-and-combining-techniques.pdf

Kumar, A. (2021, 10 19). *Difference between Parametric vs Non-Parametric Models.* Ανάκτηση 12 30, 2022, από https://vitalflux.com/difference-between-parametric-vs-non-parametric-models/?utm_content=cmp-true

Kumar, R., & Srivastava, S. (2017, 2). *Machine Learning: A Review on Binary Classification.* Ανάκτηση 12 30, 2022, από https://www.researchgate.net/profile/Saurabh-Srivastava-8/publication/313779520_Machine_Learning_A_Review_on_Binary_Classification/links/5a140771aca27240e30848cf/Machine-Learning-A-Review-on-Binary-Classification.pdf

Loh, W.-Y. (2011, 1 1). *Classification and regression trees.* Ανάκτηση 1 20, 2019, από stat.wisc.edu: https://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

Louridas, P. (2020). *Algorithms.* Ανάκτηση 12 30, 2022, από https://sciarium.com/file/509924/#:~:text=Louridas%20explains%20not%20just%20what%20algorithms%20are%20but,of%20choosing%20the%20best%20algorithm%20for%20particular%20tasks.

Mechanistic Interpretation of Machine Learning Inference: A Fuzzy Feature Importance Fusion Approach. (2021, 10 22). Ανάκτηση 1 30, 2023, από https://arxiv.org/pdf/2110.11713v1.pdf

Mullainathan, S., & Spiess, J. (2017). *Machine Learning: An Applied Econometric Approach.* Ανάκτηση 12 30, 2022, από https://scholar.harvard.edu/files/sendhil/files/jep.31.2.87.pdf

Nasteski, V. (2017). *An overview of the supervised machine learning methods.* Ανάκτηση 12 30, 2022, από https://www.researchgate.net/publication/328146111_An_overview_of_the_supervised_machine_learning_methods

Ng, A. (2022). *Support Vector Machines.* Ανάκτηση 12 30, 2022, από https://see.stanford.edu/materials/aimlcs229/cs229-notes3.pdf

Parka, K., Alib, A., Kimc, D., Ana, Y., Kimb, M., & Shin, H. (2013). *Robust predictive model for evaluating breast cancer survivability* . Ανάκτηση 11 11, 2022, από https://www.sciencedirect.com/science/article/abs/pii/S0952197613001140

Philip, A. T., Pal, S., & Verma, A. (2020). *A study of cancer prediction using neural network.* Ανάκτηση 11 11, 2022, από https://www.irjet.net/archives/V7/i11/IRJET-V7I1184.pdf

Shalizi, C. (2012). *Logistic Regression.* Ανάκτηση 12 30, 2022, από https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

Singh, H. (2015). *Practical Machine Learning and Image Processing.* Ανάκτηση 12 30, 2022, από http://repo.darmajaya.ac.id/4349/1/Practical%20Machine%20Learning%20and%20Image%20Processing_%20For%20Facial%20Recognition%2C%20Object%20Detection%2C%20and%20Pattern%20Recognition%20Using%20Python%20%28%20PDFDrive%20%29.pdf

Soofi, A. A., & Awan, A. (2017). *Classification Techniques in Machine Learning: Applications and Issues.* Ανάκτηση 12 30, 2022, από https://pdfs.semanticscholar.org/2678/e213cec548d278879ceaf01582ee8913cc3f.pdf

Yang, A. Y. (2022). *Random Forests.* Ανάκτηση 12 30, 2022, από https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf

Yufeng, G. (2017, 8 31). *The 7 Steps of Machine Learning.* Ανάκτηση 12 30, 2022, από https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

```
#Observations' proportion per class

table(nBreastCancerData$diagnosis)

prop.table(table(nBreastCancerData$diagnosis))*100


#Descriptive Staristics per quarter

by(new_BreastCancerData, BreastCancerData$diagnosis, summary)

stat.des = by(BreastCancerData, BreastCancerData$diagnosis,
describe)


stat.des.B = stat.des$B

stat.des.M = stat.des$M


#Barplots

ggplot(BreastCancerData, aes(x= diagnosis, fill=as.factor(diagnosis)
)) +

  geom_bar( ) +

  scale_fill_hue(c = 60) +

  theme(legend.position="none")


#Violin plot

ggplot(BreastCancerData, aes(x = diagnosis, y =, fill = diagnosis)) +

  geom_violin(trim = FALSE) +

  geom_boxplot(width = 0.07)

  labs(x = "", y = "Diagnosis", title = "My Violin Plot")
```

```
cols <- c("#7400b8", "#80ffdb")

#Histogram Per Class (transparent)

ggplot(new_BreastCancerData, aes(x=, fill=diagnosis)) +
geom_histogram(alpha=0.5, position="identity")

#Histogram per class density

cols <- c("#7400b8", "#80ffdb") #coloring

ggplot(new_BreastCancerData, aes(x =, fill = diagnosis)) +
  geom_density(alpha = 0.6) +
  scale_fill_manual(values = cols)

#Pearson Correlation

#Correlation

w.corr<-cor(BreastCancerData[,c(3:32)],method="pearson")

corrplot(w.corr, order='hclust', method='ellipse',addCoef.col =
'black',type='lower', number.cex = 0.25,tl.cex = 0.25, diag=F,tl.col
= 'black',tl.srt=15)

#Centering Data

DF = scale(BreastCancerData[, -c(1, 2)], scale = FALSE)

MeanCenteredDataSet = data.frame(DF)

NewCenteredDataSet = cbind( new_BreastCancerData[, c(1,2)],
MeanCenteredDataSet)
```

```
new_BreastCancerData[, -33]

View(NewCenteredDataSet)


#COLLINEARITY DIAGNOSTICS

model = lm(formula = diagnosis ~

                  radius_mean

                + texture_mean

                + smoothness_mean

                + compactness_mean

                + symmetry_mean

                + fractal_dimension_mean

                + radius_se

                + texture_se

                + smoothness_se

                + concave.points_se

                + symmetry_se

                + fractal_dimension_se

                + symmetry_worst, data = new_BreastCancerData)


#VIFs and Tolerance

VIFs_and_Tolerance = ols_vif_tol(lm.fits)


#Eigenvalues and condition indices

Eigenvalues_and_condition_indices = ols_eigen_cindex(lm.fits)


# collinearity diagnostics
```

```
collinearity_diagnostics = ols_coll_diag(lm.fits)


#LOGISTIC REGRESSION
train <- read.csv("~/Desktop/diplomatiki/01 Thesis Labs/train.csv")
test <- read.csv("~/Desktop/diplomatiki/01 Thesis Labs/test.csv")


str(train$diagnosis)
summary(test$diagnosis)
train$diagnosis = as.factor(train$diagnosis)
class(train$diagnosis)



str(test$diagnosis)
summary(test$diagnosis)
test$diagnosis = as.factor(test$diagnosis)
class(test$diagnosis)




train$diagnosis <- relevel(train$diagnosis, ref = "B")
test$diagnosis <- relevel(test$diagnosis, ref = "B")


attach(train)
set.seed(111)
glm.fits = glm(diagnosis ~
                 radius_mean
```

```
                + texture_mean
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst,
                family=binomial(logit), data = train)

vifs = vif(glm.fits)

odds.ratio = exp(glm.fits$coefficients[-1])
print(odds.ratio)
set.seed(111)
glm.fits.step = stepAIC(glm.fits, direction = "backward")

#Evaluation metrics for train and test [50-50 threshold]
#For train

glm.probs.50.train = predict(glm.fits.step, type = "response", newdata
= train)
```

```
glm.probs.50.train.labels = as.factor(ifelse(glm.probs.50.train > 0.5, "B", "M"))

table.50.train = confusionMatrix(table(glm.probs.50.train.labels, train$diagnosis), positive = "B")

print(table.50.train)


##### for test

glm.probs.50.test = predict(glm.fits.step, type = "response", newdata = test)

glm.probs.50.test.labels = as.factor(ifelse(glm.probs.50.test > 0.5, "B", "M"))



table.50.test = confusionMatrix(table(glm.probs.50.test.labels, test$diagnosis), positive = "B")


print(table.50.test)

#Evaluation metrics for train and test [63-37 threshold]

#For train

glm.probs.62.train = predict(glm.fits.step, type = "response", newdata = train)

glm.probs.62.train.labels = as.factor(ifelse(glm.probs.62.train > 0.62, "B", "M"))


table.62.train = confusionMatrix(table(glm.probs.62.train.labels, train$diagnosis), positive = "B" )


print(table.62.train)
```

```
##### for test

glm.probs.62.test = predict(glm.fits.step, type = "response", newdata
= test)

glm.probs.62.test.labels = as.factor(ifelse(glm.probs.62.test > 0.62,
"B", "M"))


table.62.test    =    confusionMatrix(table(glm.probs.62.test.labels,
test$diagnosis), positive = "B")


print(table.62.test)


#Roc curve for train set

ROCR_pred_train = prediction(glm.probs.50.train, train$diagnosis)

ROCR_perf_train = performance(ROCR_pred_train, 'tpr', 'fpr')

plot(ROCR_perf_train,  main="ROC  curve  for  train  set  (logistic
regression)", print.cutoffs.at = seq(0.4, by = 0.1), colorize = TRUE)


auc.train <- performance(ROCR_pred_train, "auc")

auc.train@y.values[[1]]


#Roc curve for test set

ROCR_pred_test = prediction(glm.probs.50.test, test$diagnosis)

ROCR_perf_test = performance(ROCR_pred_test, 'tpr', 'fpr')

plot(ROCR_perf_test,  main="ROC  curve  for  test  set  (logistic
regression)", print.cutoffs.at = seq(0.3, by = 0.2), colorize = TRUE)
```

```
auc.test <- performance(ROCR_pred_test, "auc")

auc.test@y.values[[1]]


#Lift curve for Train set

perf = performance(ROCR_pred_train,"lift","rpp")

plot(perf, main="Lift curve for train set (logistic regression)",
print.cutoffs.at = seq(0.5, by = 0.1),  colorize=T)


#lift curve for test set

perf = performance(ROCR_pred_test,"lift","rpp")

plot(perf, main="Lift curve for test set (logistic regression)",
print.cutoffs.at = seq(0.3, by = 0.2), colorize=T)


#Precision VS Recall curve for Train Set

precision.vs.recall.train    =    performance(ROCR_pred_train,"prec",
"rec")

plot(precision.vs.recall.train, main = "Precision VS Recall curve for
train set (logistic regression)",

     print.cutoffs.at = seq(0.5, by = 0.1), colorize=T)


#Precision vs recall curve for test set

precision.vs.recall.test = performance(ROCR_pred_test,"prec", "rec")

plot(precision.vs.recall.test, main =  "Precision VS Recall curve for
test set (logistic regression)",

     print.cutoffs.at = seq(0.3, by = 0.2), colorize=T)
```

```
#DECISION TREES
set.seed(111)
Decision.Trees.fit = rpart(diagnosis ~
                  radius_mean
                + texture_mean
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst, method = "class", data = train )
summary(Decision.Trees.fit)

rpart.plot(Decision.Trees.fit, fallen.leaves = TRUE)
plotcp(Decision.Trees.fit)
printcp(Decision.Trees.fit)

set.seed(111)
pruned_model <- prune(Decision.Trees.fit, cp = 0.01)
summary(pruned_model)

rpart.plot(pruned_model, fallen.leaves = TRUE)
```

```
plotcp(pruned_model)

printcp(pruned_model)


#train 50-50

diagnosis.preds.train = predict(pruned_model, type = "prob", newdata
= train)

diagnosis.preds.train = diagnosis.preds.train[, 1]


ypred.train.50 =  as.factor(ifelse(diagnosis.preds.train > 0.5, "M",
"B"))

ypred.train.50  = relevel(ypred.train.50 , ref = "M")


table.50.train     =     caret::confusionMatrix(table(ypred.train.50,
train$diagnosis), positive = "B" )

print(table.50.train)


diagnosis.preds.train.50  =  predict(pruned_model,  type  =  "prob",
newdata = train)

diagnosis.preds.train.50 = diagnosis.preds.train.50[, 1]

auc.train.50    =    performance(prediction(diagnosis.preds.train.50,
train$diagnosis), "auc")@y.values[[1]]


#train 62-38

ypred.train.62 = as.factor(ifelse(diagnosis.preds.train > 0.62, "M",
"B"))

ypred.train.62  = relevel(ypred.train.62 , ref = "M")
```

```
table.62.train     =     caret::confusionMatrix(table(ypred.train.62,
train$diagnosis), positive = "B")

print(table.62.train)


diagnosis.preds.train.62  =  predict(pruned_model,  type  =  "prob",
newdata = train)

diagnosis.preds.train.62 = diagnosis.preds.train.62[, 1]

auc.train.62    =    performance(prediction(diagnosis.preds.train.62,
train$diagnosis), "auc")@y.values[[1]]




#test 50-50

diagnosis.preds.test = predict(pruned_model, type = "prob", newdata =
test)

diagnosis.preds.test = diagnosis.preds.test[, 1]


ypred.test.50  =   as.factor(ifelse(diagnosis.preds.test  >  0.5,  "M",
"B"))

ypred.test.50  = relevel(ypred.test.50 , ref = "M")


table.50.test      =      caret::confusionMatrix(table(ypred.test.50,
test$diagnosis), positive = "B" )

print(table.50.test)


diagnosis.preds.test.50 = predict(pruned_model, type = "prob", newdata
= test)

diagnosis.preds.test.50 = diagnosis.preds.test.50[, 1]
```

```
auc.test.50     =     performance(prediction(diagnosis.preds.test.50, test$diagnosis), "auc")@y.values[[1]]




#test 62-38

ypred.test.62  =  as.factor(ifelse(diagnosis.preds.test  >  0.62,  "M", "B"))

ypred.test.62  = relevel(ypred.test.62 , ref = "M")


table.62.test      =      caret::confusionMatrix(table(ypred.test.62, test$diagnosis), positive = "B")

print(table.62.test)


diagnosis.preds.test.62 = predict(pruned_model, type = "prob", newdata = test)

diagnosis.preds.test.62 = diagnosis.preds.test.62[, 1]

auc.test.62     =     performance(prediction(diagnosis.preds.test.62, test$diagnosis), "auc")@y.values[[1]]


#FOR TRAIN

#predictions.train  =  predict(pruned_model,  newdata  =  train,  type  = "prob")

pred.obj.train = prediction(diagnosis.preds.train, train$diagnosis)

perf.obj.train = performance(pred.obj.train, "tpr", "fpr")

plot(perf.obj.train, colorize = TRUE, main="ROC curve for train set (decision tree)", print.cutoffs.at = 0.5)
```

```
#FOR TEST

#predictions.test = predict(pruned_model, newdata = test, type = "prob")

pred.obj.test = prediction(diagnosis.preds.test, test$diagnosis)

perf.obj.test = performance(pred.obj.test, "tpr", "fpr")

plot(perf.obj.test, colorize = TRUE, main="ROC curve for test set (decision tree)", print.cutoffs.at = 0.5)

#prob.val.train = predict(pruned_model, train, type = "prob")

pred.val.train = prediction(diagnosis.preds.train, train$diagnosis)

perf.val.train = performance(pred.val.train, "lift")


plot(performance(pred.val.train, measure="lift", x.measure="rpp"), main="Lift curve for train set (decision tree)", print.cutoffs.at = 0.5, colorize=TRUE)


#Lift Curve for test set

#prob.val.test = predict(pruned_model, test, type = "prob")

pred.val.test = prediction(diagnosis.preds.test, test$diagnosis)

perf.val.test = performance(pred.val.test, "lift")


plot(performance(pred.val.test, measure="lift", x.measure="rpp"), main="Lift curve for test set (decision tree)", print.cutoffs.at = 0.5, colorize=TRUE)


#Precision VS Recall Curves

#For train set
```

```
plot(performance(pred.val.train, measure="prec", x.measure="rec"),
main="Precision vs recall curve for train set (decision tree)",
print.cutoffs.at = 0.5, colorize=TRUE)


#For Test Set

plot(performance(pred.val.test, measure="prec", x.measure="rec"),
main="Precision vs recall curve for test set (decision tree)",
print.cutoffs.at = 0.5, colorize=TRUE)
```

#Random Forest

```
set.seed(123)

model.RF = randomForest (diagnosis ~

                              radius_mean

               + texture_mean

               + smoothness_mean

               + compactness_mean

               + symmetry_mean

               + radius_se

               + texture_se

               + smoothness_se

               + concave.points_se

               + symmetry_se

               + fractal_dimension_se

               + symmetry_worst, importance = TRUE, proximity = TRUE,
data = train)

attach(train)

# Define the hyperparameter values to be searched
```

```
ntree_vals = c(300, 400, 500)

mtry_vals = c(1, 2, 3)

sampsize_vals = c(0.2, 0.4, 0.6)

nodesize_vals = c(1, 2, 3, 4)

cutoff_vals = c(0.5, 0.62)


# Initialize a list to store the results

results = list()


# Initialize lists to store the results

results_ntree = numeric()

results_mtry = numeric()

results_sampsize = numeric()

results_cutoff = numeric()

results_nodesize = numeric()


results_auc = numeric()

results_accuracy = numeric()

results_precision = numeric()

results_sensitivity = numeric()

results_specificity = numeric()

results_F1_score = numeric()


# Loop through all combinations of hyperparameters

for (i in ntree_vals) {

  for (j in mtry_vals) {
```

```
for (k in sampsize_vals) {
  for (l in cutoff_vals) {
    for (m in nodesize_vals) {

      # Train the random forests model
      set.seed(123)
      model = randomForest(diagnosis ~
                           radius_mean
           + texture_mean
           + smoothness_mean
           + compactness_mean
           + symmetry_mean
           + radius_se
           + texture_se
           + smoothness_se
           + concave.points_se
           + symmetry_se
           + fractal_dimension_se
           + symmetry_worst,
                           ntree = i,
                           mtry = j,
                           sampsize = floor(nrow(train) * k),
                           cutoff = c(l, abs(1 - l)),
                           nodesize = m,
                         data = train)
```

```
          # Make predictions on the test data and confusion matrix
          predictions = predict(model, newdata = test)
          table = confusionMatrix(table(predictions, test$diagnosis),
positive = "B")

          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model
          predictions = predict(model, newdata = test, type = "prob")
          predictions = predictions[, 2]
          auc = performance(prediction(predictions, test$diagnosis),
"auc")@y.values[[1]]
          accuracy = table[["overall"]][["Accuracy"]]
          precision = table[["byClass"]][["Pos Pred Value"]]
          sensitivity= table[["byClass"]][["Sensitivity"]]
          specificity = table[["byClass"]][["Specificity"]]
          F1_score = table[["byClass"]][["F1"]]

          # Store the results
          results_ntree = c(results_ntree, i)
          results_mtry = c(results_mtry, j)
          results_sampsize = c(results_sampsize, k)
          results_cutoff = c(results_cutoff, l)
          results_nodesize = c(results_nodesize, m)
          results_auc = c(results_auc, auc)
          results_accuracy= c(results_accuracy, accuracy)
          results_precision = c(results_precision, precision)
```

```
          results_sensitivity = c(results_sensitivity, sensitivity)
          results_specificity = c(results_specificity, specificity)
          results_F1_score = c(results_F1_score, F1_score)
        }
      }
    }
  }
}

#Combine the results into a single data frame
results = data.frame(ntree = results_ntree,
                     mtry = results_mtry,
                     sampsize = results_sampsize,
                     cutoff = results_cutoff,
                     nodesize = results_nodesize,
                     auc = results_auc,
                     accuracy = results_accuracy,
                     precision = results_precision,
                     sensitivity = results_sensitivity,
                     specificity = results_specificity,
                     F1_score = results_F1_score)

set.seed(123)
model.RF = randomForest (diagnosis ~
                           radius_mean
               + texture_mean
```

```
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst, importance = TRUE, proximity = TRUE,
                            ntree = 500 ,
                            mtry = 1 ,
                            sampsize = floor(nrow(train) * 0.4),
                            nodesize = 2,
                           cutoff = c(0.62, 0.38),
                          data = train)

ooberror = mean(model.RF$err.rate[,1])

oob.error.data = data.frame(
  Tress = rep(1:nrow(model.RF$err.rate), times = 3),
  type = rep(c("OOB", "B", "M"), each = nrow(model.RF$err.rate)),
  error = c(model.RF$err.rate[,"OOB"],
            model.RF$err.rate[, "B"],
            model.RF$err.rate[,"M"]))
#PLOT
```

```
ggplot(data = oob.error.data, aes(x = oob.error.data$Tress, y = oob.error.data$error)) +
  geom_line(aes(color = type))

#varImpPlot(model.RF)

print(model.RF$importance)
varImpPlot(model.RF)

#train
predictions.optimal.train = predict(model.RF, newdata = train)

          table.optimal.train = confusionMatrix(table(predictions.optimal.train, train$diagnosis), positive = "B")


predictions.optimal.train = predict(model.RF, newdata = train, type = "prob")
          predictions.optimal.train = predictions.optimal.train[, 2]
          auc.optimal.train = performance(prediction(predictions.optimal.train, train$diagnosis), "auc")@y.values[[1]]

          print(auc.optimal.train)
          print(table.optimal.train)
```

```
#test

predictions.optimal.test = predict(model.RF, newdata = test)

          table.optimal.test                                          =
confusionMatrix(table(predictions.optimal.test,     test$diagnosis),
positive = "B")



predictions.optimal.test = predict(model.RF, newdata = test, type =
"prob")

          predictions.optimal.test = predictions.optimal.test[, 2]

          auc.optimal.test                                            =
performance(prediction(predictions.optimal.test,    test$diagnosis),
"auc")@y.values[[1]]


          print(auc.optimal.test)

          print(table.optimal.test)

          contrasts(train$diagnosis)


#Roc curve for train set

ROCR_pred_train        =        prediction(predictions.optimal.train,
train$diagnosis)

ROCR_perf_train = performance(ROCR_pred_train, 'tpr', 'fpr')

plot(ROCR_perf_train, main="ROC curve for train set (optimal random
forest)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


auc.train = performance(ROCR_pred_train, "auc")

auc.train = auc.train@y.values[[1]]
```

```
#Roc curve for test set

ROCR_pred_test = prediction(predictions.optimal.test, test$diagnosis)

ROCR_perf_test = performance(ROCR_pred_test, 'tpr', 'fpr')

plot(ROCR_perf_test, main="ROC curve for test set (optimal random forest)", print.cutoffs.at =c(0.5, 0.6), colorize = TRUE)


auc.test = performance(ROCR_pred_test, "auc")

auc.test = auc.test@y.values[[1]]


perf.pred.train = prediction(predictions.optimal.train, train$diagnosis)

lift.train = performance(perf.pred.train, "lift", "rpp")

plot(lift.train, main="Lift curve for train set (optimal random forest)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)



prec.rec.train <- performance(perf.pred.train, "prec", "rec")

plot(prec.rec.train, main = "Precision vs Recall Curve for train set (optimal random forest)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)



perf.pred.test = prediction(predictions.optimal.test, test$diagnosis)

lift.test = performance(perf.pred.test, "lift", "rpp")

plot(lift.test, main="Lift curve for test set (optimal random forest)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)
```

```
prec.rec.test <- performance(perf.pred.test, "prec", "rec")

plot(prec.rec.test, main = "Precision vs Recall Curve for test set (optimal random forest)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)
```

#SUPPORT VECTOR MACHINE

```
#LINEAR KERNEL


# Define the hyperparameter values to be searched

cost_vals = c(0.001, 0.01, 1, 1.5, 2, 2.5, 100)


# Initialize a list to store the results

results = list()


# Initialize lists to store the results

results_cost = numeric()


results_auc = numeric()

results_accuracy = numeric()

results_precision = numeric()

results_sensitivity = numeric()

results_specificity = numeric()

results_F1_score = numeric()


# Loop through all combinations of hyperparameters

for (i in cost_vals) {
```

```
  set.seed(111)
svm.model = svm(diagnosis ~
                              radius_mean
                + texture_mean
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst,
                            cost = i,
                  type  =  "C-classification",  kernel  =  "linear",
probability= TRUE, data = train)


          # Make predictions on the test data and confusion matrix
           predictions = predict(svm.model, newdata = test, probability
= TRUE)
           probs = attr(predictions, "probabilities")
           probs = probs[, 1]
           pred.labels = as.factor(ifelse(probs > 0.5, "B", "M"))
           table = confusionMatrix(table(pred.labels, test$diagnosis),
positive = "B")
```

```
          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model
          probs = predict(svm.model, newdata = test, probability =
TRUE)
          probs = attr(predictions, "probabilities")
          probs = probs[, 2]
          auc   =   performance(prediction(probs,   test$diagnosis),
"auc")@y.values[[1]]
          accuracy = table[["overall"]][["Accuracy"]]
          precision = table[["byClass"]][["Pos Pred Value"]]
          sensitivity= table[["byClass"]][["Sensitivity"]]
          specificity = table[["byClass"]][["Specificity"]]
          F1_score = table[["byClass"]][["F1"]]

          # Store the results
          results_cost = c(results_cost, i)

          results_auc = c(results_auc, auc)
          results_accuracy= c(results_accuracy, accuracy)
          results_precision = c(results_precision, precision)
          results_sensitivity = c(results_sensitivity, sensitivity)
          results_specificity = c(results_specificity, specificity)
          results_F1_score = c(results_F1_score, F1_score)
}
```

```
          #Combine the results into a single data frame
results = data.frame(cost = results_cost,
                     auc = results_auc,
                     accuracy = results_accuracy,
                     precision = results_precision,
                     sensitivity = results_sensitivity,
                     specificity = results_specificity,
                     F1_score = results_F1_score)

#RADIAL KERNEL

# Define the hyperparameter values to be searched
cost_vals = c(0.01, 1, 1.5)
gamma_vals = c(0.001, 0.01, 1)

# Initialize a list to store the results
results = list()

# Initialize lists to store the results
results_cost = numeric()
results_gamma = numeric()

results_auc = numeric()
results_accuracy = numeric()
results_precision = numeric()
results_sensitivity = numeric()
```

```
results_specificity = numeric()
results_F1_score = numeric()


for (i in cost_vals) {
 for (j in gamma_vals) {
set.seed(1111)
  svm.model = svm(diagnosis ~
                              radius_mean
                + texture_mean
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst,
                            cost = i,
                            gamma = j,
                  type  =  "C-classification",  kernel  =  "radial",
probability= TRUE, data = train)


 # Make predictions on the test data and confusion matrix
```

```
          predictions = predict(svm.model, newdata = test, probability
= TRUE)

          probs = attr(predictions, "probabilities")

          probs = probs[, 1]

          pred.labels = as.factor(ifelse(probs > 0.5, "B", "M"))

          table = confusionMatrix(table(pred.labels, test$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs = predict(svm.model, newdata = test, probability =
TRUE)

          probs = attr(predictions, "probabilities")

          probs = probs[, 2]

          auc   =   performance(prediction(probs,   test$diagnosis),
"auc")@y.values[[1]]

          accuracy = table[["overall"]][["Accuracy"]]

          precision = table[["byClass"]][["Pos Pred Value"]]

          sensitivity= table[["byClass"]][["Sensitivity"]]

          specificity = table[["byClass"]][["Specificity"]]

          F1_score = table[["byClass"]][["F1"]]


          # Store the results

          results_cost = c(results_cost, i)

          results_gamma = c(results_gamma, j)


          results_auc = c(results_auc, auc)
```

```
          results_accuracy= c(results_accuracy, accuracy)
          results_precision = c(results_precision, precision)
          results_sensitivity = c(results_sensitivity, sensitivity)
          results_specificity = c(results_specificity, specificity)
          results_F1_score = c(results_F1_score, F1_score)
}
}

          #Combine the results into a single data frame
results = data.frame(cost = results_cost,
                     gamma = results_gamma,
                     auc = results_auc,
                     accuracy = results_accuracy,
                     precision = results_precision,
                     sensitivity = results_sensitivity,
                     specificity = results_specificity,
                     F1_score = results_F1_score)

#POLYNOMIAL KERNEL

# Define the hyperparameter values to be searched
cost_vals = c(0.01, 1, 1.5)
degree_vals = c(2, 3, 4)

# Initialize a list to store the results
results = list()
```

```
# Initialize lists to store the results
results_cost = numeric()
results_degree = numeric()

results_auc = numeric()
results_accuracy = numeric()
results_precision = numeric()
results_sensitivity = numeric()
results_specificity = numeric()
results_F1_score = numeric()

# Loop through all combinations of hyperparameters
for (i in cost_vals) {
  for (j in degree_vals) {
set.seed(2)
svm.model = svm(diagnosis ~
                                radius_mean
                + texture_mean
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
```

```
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst,
                            cost = i,
                            degree = j,
                  type = "C-classification", kernel = "polynomial",
probability= TRUE, data = train)


 # Make predictions on the test data and confusion matrix
          predictions = predict(svm.model, newdata = test, probability
= TRUE)
          probs = attr(predictions, "probabilities")
          probs = probs[, 1]
          pred.labels = as.factor(ifelse(probs > 0.5, "B", "M"))
          table = confusionMatrix(table(pred.labels, test$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model
          probs = predict(svm.model, newdata = test, probability =
TRUE)
          probs = attr(predictions, "probabilities")
          probs = probs[, 2]
          auc   =   performance(prediction(probs,   test$diagnosis),
"auc")@y.values[[1]]
          accuracy = table[["overall"]][["Accuracy"]]
          precision = table[["byClass"]][["Pos Pred Value"]]
```

```
            sensitivity= table[["byClass"]][["Sensitivity"]]
            specificity = table[["byClass"]][["Specificity"]]
            F1_score = table[["byClass"]][["F1"]]

            # Store the results
            results_cost = c(results_cost, i)
            results_degree = c(results_degree, j)

            results_auc = c(results_auc, auc)
            results_accuracy= c(results_accuracy, accuracy)
            results_precision = c(results_precision, precision)
            results_sensitivity = c(results_sensitivity, sensitivity)
            results_specificity = c(results_specificity, specificity)
            results_F1_score = c(results_F1_score, F1_score)
 }
}
            #Combine the results into a single data frame
results = data.frame(cost = results_cost,
                     degree = results_degree,

                     auc = results_auc,
                     accuracy = results_accuracy,
                     precision = results_precision,
                     sensitivity = results_sensitivity,
                     specificity = results_specificity,
                     F1_score = results_F1_score)
```

```
set.seed(111)
svm.model.linear = svm(diagnosis ~
                            radius_mean
               + texture_mean
               + smoothness_mean
               + compactness_mean
               + symmetry_mean
               + radius_se
               + texture_se
               + smoothness_se
               + concave.points_se
               + symmetry_se
               + fractal_dimension_se
               + symmetry_worst,
                 type = "C-classification", kernel = "linear", cost =
0.01, probability= TRUE, data = train)

 # Make predictions on the test data and confusion matrix

#Train
          predictions.linear.train    =    predict(svm.model.linear,
newdata = train, probability = TRUE)
          probs.linear.train     =     attr(predictions.linear.train,
"probabilities")
          probs.linear.train = probs.linear.train[, 1]
```

```
          pred.labels.linear.train                                    =
as.factor(ifelse(probs.linear.train > 0.5, "B", "M"))

          table.linear.train                                          =
confusionMatrix(table(pred.labels.linear.train,    train$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.linear.train  =  predict(svm.model.linear,  newdata  =
train, probability = TRUE)

          probs.linear.train    =    attr(predictions.linear.train,
"probabilities")

          probs.linear.train = probs.linear.train[, 2]

          auc.linear.train                                            =
performance(prediction(probs.linear.train,         train$diagnosis),
"auc")@y.values[[1]]

          accuracy.linear.train                                       =
table.linear.train[["overall"]][["Accuracy"]]

          precision.linear.train                                      =
table.linear.train[["byClass"]][["Pos Pred Value"]]

          sensitivity.linear.train                                    =
table.linear.train[["byClass"]][["Sensitivity"]]

          specificity.linear.train                                    =
table.linear.train[["byClass"]][["Specificity"]]

          F1_score.linear.train                                       =
table.linear.train[["byClass"]][["F1"]]


#test

          predictions.linear.test = predict(svm.model.linear, newdata
= test, probability = TRUE)
```

```
          probs.linear.test      =      attr(predictions.linear.test,
"probabilities")

          probs.linear.test = probs.linear.test[, 1]

          pred.labels.linear.test                                    =
as.factor(ifelse(probs.linear.test > 0.5, "B", "M"))

          table.linear.test                                          =
confusionMatrix(table(pred.labels.linear.test,      test$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.linear.test  =  predict(svm.model.linear,  newdata  =
test, probability = TRUE)

          probs.linear.test      =      attr(predictions.linear.test,
"probabilities")

          probs.linear.test = probs.linear.test[, 2]

          auc.linear.test = performance(prediction(probs.linear.test,
test$diagnosis), "auc")@y.values[[1]]

          accuracy.linear.test                                       =
table.linear.test[["overall"]][["Accuracy"]]

          precision.linear.test                                      =
table.linear.test[["byClass"]][["Pos Pred Value"]]

          sensitivity.linear.test                                    =
table.linear.test[["byClass"]][["Sensitivity"]]

          specificity.linear.test                                    =
table.linear.test[["byClass"]][["Specificity"]]

          F1_score.linear.test                                       =
table.linear.test[["byClass"]][["F1"]]
```

```
#lINEAR KERNEL

#Train 62-38

          predictions.linear.train.62   =   predict(svm.model.linear,
newdata = train, probability = TRUE)

          probs.linear.train.62  =  attr(predictions.linear.train.62,
"probabilities")

          probs.linear.train.62 = probs.linear.train.62[, 1]

          pred.labels.linear.train.62                                  =
as.factor(ifelse(probs.linear.train.62 > 0.62, "B", "M"))

          table.linear.train.62                                        =
confusionMatrix(table(pred.labels.linear.train.62,  train$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.linear.train.62 = predict(svm.model.linear, newdata =
train, probability = TRUE)

          probs.linear.train.62  =  attr(predictions.linear.train.62,
"probabilities")

          probs.linear.train.62 = probs.linear.train.62[, 2]

          auc.linear.train.62                                          =
performance(prediction(probs.linear.train.62,      train$diagnosis),
"auc")@y.values[[1]]

          accuracy.linear.train.62                                     =
table.linear.train.62[["overall"]][["Accuracy"]]

          precision.linear.train.62                                    =
table.linear.train.62[["byClass"]][["Pos Pred Value"]]

          sensitivity.linear.train.62                                  =
table.linear.train.62[["byClass"]][["Sensitivity"]]
```

```
          specificity.linear.train.62                                    =
table.linear.train.62[["byClass"]][["Specificity"]]

          F1_train.linear.train.62                                       =
table.linear.train.62[["byClass"]][["F1"]]


#test 62-38

          predictions.linear.test.62   =   predict(svm.model.linear,
newdata = test, probability = TRUE)

          probs.linear.test.62   =   attr(predictions.linear.test.62,
"probabilities")

          probs.linear.test.62 = probs.linear.test.62[, 1]

          pred.labels.linear.test.62                                     =
as.factor(ifelse(probs.linear.test.62 > 0.5, "B", "M"))

          table.linear.test.62                                           =
confusionMatrix(table(pred.labels.linear.test.62,   test$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.linear.test.62 = predict(svm.model.linear, newdata =
test, probability = TRUE)

          probs.linear.test.62   =   attr(predictions.linear.test.62,
"probabilities")

          probs.linear.test.62 = probs.linear.test.62[, 2]

          auc.linear.test.62                                             =
performance(prediction(probs.linear.test.62,        test$diagnosis),
"auc")@y.values[[1]]

          accuracy.linear.test.62                                        =
table.linear.test.62[["overall"]][["Accuracy"]]
```

```
          precision.linear.test.62                                    =
table.linear.test.62[["byClass"]][["Pos Pred Value"]]

          sensitivity.linear.test.62                                  =
table.linear.test.62[["byClass"]][["Sensitivity"]]

          specificity.linear.test.62                                  =
table.linear.test.62[["byClass"]][["Specificity"]]

          F1_score.linear.test.62                                     =
table.linear.test.62[["byClass"]][["F1"]]



#Plots for linear kernel

#ROC Curve for Train set

ROCR_pred_linear_train       =       prediction(probs.linear.train,
train$diagnosis)

ROCR_perf_linear_train  =  performance(ROCR_pred_linear_train,  'tpr',
'fpr')

plot(ROCR_perf_linear_train,  main="ROC  curve  for  train  set  (Linear
Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)



#Roc curve for test set

ROCR_pred_linear_test = prediction(probs.linear.test, test$diagnosis)

ROCR_perf_linear_test  =  performance(ROCR_pred_linear_test,  'tpr',
'fpr')

plot(ROCR_perf_linear_test,  main="ROC  curve  for  test  set  (Linear
Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)
```

```
#Lift curve for train set

perf.pred.train.linear = prediction(probs.linear.train, train$diagnosis)

lift.train.linear = performance(perf.pred.train.linear, "lift", "rpp")

plot(lift.train.linear, main="Lift curve for train set (Linear Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


#Lift curve for test set

perf.pred.test.linear = prediction(probs.linear.test, test$diagnosis)

lift.test.linear = performance(perf.pred.test.linear, "lift", "rpp")

plot(lift.test.linear, main="Lift curve for test set (Linear Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


#Pecision VS Recall curve

#For train set

prec.rec.train.linear <- performance(perf.pred.train.linear, "prec", "rec")

plot(prec.rec.train.linear , main="Precision vs Recall Curve for train set (Linear Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


#for test set

prec.rec.test.linear <- performance(perf.pred.test.linear, "prec", "rec")

plot(prec.rec.test.linear, main="Precision vs Recall Curve for test set (Linear Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


set.seed(1111)
```

```
svm.model.radial = svm(diagnosis ~
                               radius_mean
                + texture_mean
                + smoothness_mean
                + compactness_mean
                + symmetry_mean
                + radius_se
                + texture_se
                + smoothness_se
                + concave.points_se
                + symmetry_se
                + fractal_dimension_se
                + symmetry_worst,
                            type = "C-classification", cost = 1, gamma
= 0.01, kernel = "radial", probability= TRUE, data = train)


#Train
          predictions.radial.train    =    predict(svm.model.radial,
newdata = train, probability = TRUE)
          probs.radial.train     =     attr(predictions.radial.train,
"probabilities")
          probs.radial.train = probs.radial.train[, 1]
          pred.labels.radial.train                                    =
as.factor(ifelse(probs.radial.train > 0.5, "B", "M"))
          table.radial.train                                          =
confusionMatrix(table(pred.labels.radial.train,    train$diagnosis),
positive = "B")
```

```
          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.radial.train  =  predict(svm.model.radial,  newdata  =
train, probability = TRUE)

          probs.radial.train     =     attr(predictions.radial.train,
"probabilities")

          probs.radial.train = probs.radial.train[, 2]

          auc.radial.train                                          =
performance(prediction(probs.radial.train,          train$diagnosis),
"auc")@y.values[[1]]

          accuracy.radial.train                                     =
table.radial.train[["overall"]][["Accuracy"]]

          precision.radial.train                                    =
table.radial.train[["byClass"]][["Pos Pred Value"]]

          sensitivity.radial.train                                  =
table.radial.train[["byClass"]][["Sensitivity"]]

          specificity.radial.train                                  =
table.radial.train[["byClass"]][["Specificity"]]

          F1_score.radial.train                                     =
table.radial.train[["byClass"]][["F1"]]


#test

          predictions.radial.test = predict(svm.model.radial, newdata
= test, probability = TRUE)

          probs.radial.test      =      attr(predictions.radial.test,
"probabilities")

          probs.radial.test = probs.radial.test[, 1]
```

```
          pred.labels.radial.test                                    =
as.factor(ifelse(probs.radial.test > 0.5, "B", "M"))

          table.radial.test                                          =
confusionMatrix(table(pred.labels.radial.test,      test$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.radial.test  =  predict(svm.model.radial,  newdata  =
test, probability = TRUE)

          probs.radial.test      =      attr(predictions.radial.test,
"probabilities")

          probs.radial.test = probs.radial.test[, 2]

          auc.radial.test = performance(prediction(probs.radial.test,
test$diagnosis), "auc")@y.values[[1]]

          accuracy.radial.test                                       =
table.radial.test[["overall"]][["Accuracy"]]

          precision.radial.test                                      =
table.radial.test[["byClass"]][["Pos Pred Value"]]

          sensitivity.radial.test                                    =
table.radial.test[["byClass"]][["Sensitivity"]]

          specificity.radial.test                                    =
table.radial.test[["byClass"]][["Specificity"]]

          F1_score.radial.test                                       =
table.radial.test[["byClass"]][["F1"]]


#Radial Kernel

#Train 62-38
```

```
          predictions.radial.train.62   =   predict(svm.model.radial,
newdata = train, probability = TRUE)

          probs.radial.train.62  =  attr(predictions.radial.train.62,
"probabilities")

          probs.radial.train.62 = probs.radial.train.62[, 1]

          pred.labels.radial.train.62                                  =
as.factor(ifelse(probs.radial.train.62 > 0.62, "B", "M"))

          table.radial.train.62                                        =
confusionMatrix(table(pred.labels.radial.train.62,  train$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.radial.train.62 = predict(svm.model.radial, newdata =
train, probability = TRUE)

          probs.radial.train.62  =  attr(predictions.radial.train.62,
"probabilities")

          probs.radial.train.62 = probs.radial.train.62[, 2]

          auc.radial.train.62                                          =
performance(prediction(probs.radial.train.62,      train$diagnosis),
"auc")@y.values[[1]]

          accuracy.radial.train.62                                     =
table.radial.train.62[["overall"]][["Accuracy"]]

          precision.radial.train.62                                    =
table.radial.train.62[["byClass"]][["Pos Pred Value"]]

          sensitivity.radial.train.62                                  =
table.radial.train.62[["byClass"]][["Sensitivity"]]

          specificity.radial.train.62                                  =
table.radial.train.62[["byClass"]][["Specificity"]]
```

```
          F1_score.radial.train.62                                    =
table.radial.train.62[["byClass"]][["F1"]]


#test 62-38

          predictions.radial.test.62   =   predict(svm.model.radial,
newdata = test, probability = TRUE)

          probs.radial.test.62   =   attr(predictions.radial.test.62,
"probabilities")

          probs.radial.test.62 = probs.radial.test.62[, 1]

          pred.labels.radial.test.62                                  =
as.factor(ifelse(probs.radial.test.62 > 0.62, "B", "M"))

          table.radial.test.62                                        =
confusionMatrix(table(pred.labels.radial.test.62,   test$diagnosis),
positive = "B")


          # Evaluate the AUC, accuracy, specificity & sensitivity of
the model

          probs.radial.test.62 = predict(svm.model.radial, newdata =
test, probability = TRUE)

          probs.radial.test.62   =   attr(predictions.radial.test.62,
"probabilities")

          probs.radial.test.62 = probs.radial.test.62[, 2]

          auc.radial.test.62                                          =
performance(prediction(probs.radial.test.62,        test$diagnosis),
"auc")@y.values[[1]]

          accuracy.radial.test.62                                     =
table.radial.test.62[["overall"]][["Accuracy"]]

          precision.radial.test.62                                    =
table.radial.test.62[["byClass"]][["Pos Pred Value"]]
```

```
          sensitivity.radial.test.62                                    =
table.radial.test.62[["byClass"]][["Sensitivity"]]

          specificity.radial.test.62                                    =
table.radial.test.62[["byClass"]][["Specificity"]]

          F1_score.radial.test.62                                       =
table.radial.test.62[["byClass"]][["F1"]]



#Plots for radial kernel

#ROC Curve for Train set

ROCR_pred_radial_train       =       prediction(probs.radial.train,
train$diagnosis)

ROCR_perf_radial_train  =  performance(ROCR_pred_radial_train,  'tpr',
'fpr')

plot(ROCR_perf_radial_train,  main="ROC  curve  for  train  set  (radial
Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)



#Roc curve for test set

ROCR_pred_radial_test = prediction(probs.radial.test, test$diagnosis)

ROCR_perf_radial_test  =  performance(ROCR_pred_radial_test,  'tpr',
'fpr')

plot(ROCR_perf_radial_test,  main="ROC  curve  for  test  set  (radial
Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)



#Lift curve for train set
```

```
perf.pred.train.radial = prediction(probs.radial.train, train$diagnosis)

lift.train.radial = performance(perf.pred.train.radial, "lift", "rpp")

plot(lift.train.radial, main="Lift curve for train set (Radial Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


#Lift curve for test set

perf.pred.test.radial = prediction(probs.radial.test, test$diagnosis)

lift.test.radial = performance(perf.pred.test.radial, "lift", "rpp")

plot(lift.test.radial, main="Lift curve for test set (radial Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


#Pecision VS Recall curve

#For train set

prec.rec.train.radial = performance(perf.pred.train.radial, "prec", "rec")

plot(prec.rec.train.radial , main="Precision vs Recall Curve for train set (radial Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)


#for test set

prec.rec.test.radial = performance(perf.pred.test.radial, "prec", "rec")

plot(prec.rec.test.radial, main="Precision vs Recall Curve for test set (radial Kernel)", print.cutoffs.at = c(0.5, 0.6), colorize = TRUE)
```