Ιόνιο Πανεπιστήμιο

Τμήμα Πληροφορικής

Αποθήκες και Εξόρυξη Δεδομένων

«Ανάπτυξη Μοντέλων Μηχανικής Μάθησης για την Πρόβλεψη του Καρκίνου του Μαστού»

Στοιχεία Φοιτητή:

Ονοματεπώνυμο	Νικόλαος Μπαλάτος
Αριθμός Μητρώου	Inf2021151
Email	inf2021151@ionio.gr / nbalatos@gmail.com

2025

Εισαγωγικό σημείωμα

Ο σκοπός της παρούσας εργασίας είναι η εφαρμογή διάφορων τεχνικών εξόρυξης γνώσης μέσω της εκτέλεσης αλγορίθμων μηχανικής μάθησης για την καλύτερη δυνατή πρόβλεψη του καρκίνου του μαστού. Για την επίτευξη του στόχου αξιοποιήθηκαν σχετικά δεδομένα από το Wisconsin Breast Cancer Dataset το οποίο περιλαμβάνει κλινικές και κυτταρολογικές μετρήσεις όγκων του μαστού οι οποίες επισημαίνονται ως καλοήθεις ή κακοήθεις. Στην συνέχεια αξιοποιήθηκαν αλγόριθμοι μηχανικής μάθησης για την εκπαίδευση μοντέλων πρόβλεψης καθώς και διάφορα στατιστικά στοιχεία για την αξιολόγηση των αποτελεσμάτων. Η εργασία εκπονήθηκε στα πλαίσια του μαθήματος «Αποθήκες και Εξόρυξη δεδομένων» του Ιονίου Πανεπιστημίου.

Περιεχόμενα

1. Εισαγωγή

Στην παρούσα εργασία, κύριος σκοπός είναι η εφαρμογή τεχνικών Εξόρυξης Δεδομένων μέσω της Μηχανικής Μάθησης για την ανάπτυξη μοντέλων πρόβλεψης του καρκίνου του μαστού. Για την διαδικασία της εκπαίδευσης των μοντέλων, σημαντικό παράγοντα αποτελούν τα δεδομένα που θα χρησιμοποιήσουμε καθώς από την ποιότητα τους και την προεπεξεργασία που θα υποστούν καθορίζεται σε μεγάλο βαθμό η ακρίβεια των μοντέλων πρόβλεψης. Στην εργασία αξιοποιήθηκαν τα δεδομένα του Wisconsin Breast Cancer Diagnostic Dataset, ένα ευρέως γνωστό σύνολο δεδομένων το οποίο περιλαμβάνει πολυάριθμες μετρήσεις από βιοψίες όγκων του μαστού για την ταξινόμηση τους ως καλοήθεις ή κακοήθεις, αξιοποιώντας 30 διακριτά χαρακτηριστικά σχετικά με τις φυσικές ιδιότητες των κυττάρων. Ο σκοπός της εργασίας αφορά σε δύο άξονες, αφενός την κατανόηση, την ανάλυση και την αξιολόγηση των δεδομένων μέσω τεχνικών προεπεξεργασίας και αφετέρου την αξιολόγηση διάφορων τεχνικών Μηχανικής Μάθησης με στόχο την εξαγωγή χρήσιμων συμπερασμάτων για την διάγνωση του καρκίνου του μαστού. Η παρούσα εργασία εντάσσεται στο πλαίσιο εξόρυξης γνώσης από ιατρικά δεδομένα, έναν τομέα με σημαντική ερευνητική και πρακτική αξία

1.1 Προσέγγιση της υλοποίησης

Για την επίλυση του προβλήματος ταξινόμησης που θέτει το σύνολο δεδομένων, αξιοποιήθηκαν βασικές προσεγγίσεις Μηχανικής Μάθησης, οι οποίες περιλαμβάνουν την χρήση αλγορίθμων όπως οι: **Logistic Regression, Random Forest Classifier, XGBoost.** Οι παραπάνω αλγόριθμοι εκτελέστηκαν σε συνδυασμό με τεχνικές προεπεξεργασίας του συνόλου δεδομένων όπως: Z-score outliers handling και Variance Threshold, με στόχο την επίτευξη της βέλτιστης αποδοτικότητας των μοντέλων. Η επιλογή των αλγορίθμων έγινε με βάση την μορφή του Dataset το οποίο περιέχει πολυάριθμα αριθμητικά χαρακτηριστικά καθώς και μια δυαδική κατηγορία ως τιμή εξόδου. Για την αξιολόγηση των αποτελεσμάτων επιλέχθηκε η μέθοδος 10-Fold Stratified Cross Validation σε συνδυασμό με τις μετρικές: Accuracy, Precision, Recall, F1-Score. Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για την εκπόνηση της εργασίας ήταν η Python σε συνδυασμό με της βιβλιοθήκες της, για διαχείριση δεδομένων, προεπεξεργασία, οπτικοποίηση και αλγορίθμων ταξινόμησης και αξιολόγησης.

1.2 Συνεισφορά

Η συνεισφορά της παρούσας εργασίας έγκειται στην ολοκληρωμένη ανάλυση του Breast Cancer Wisconsin Dataset με την Python, εστιάζοντας στην ανάλυση των δεδομένων και την προεπεξεργασία τους καθώς και στην συγκριτική αξιολόγηση διάφορων αλγορίθμων Μηχανικής Μάθησης. Εξετάζεται η επίδραση διάφορων τεχνικών ανάλυσης και προεπεξεργασίας καθώς και η αποδοτικότητα των μοντέλων που αναπτύχθηκαν μέσω διάφορων συγκρίσεων που σχετίζονται με τα αποτελέσματα. Παράλληλα η εργασία παρέχει και βιβλιογραφική επισκόπηση σχετικών έργων που θα

βοηθήσουν στην κατανόηση των πειραμάτων και του μικρόκοσμου του συνόλου δεδομένων που επιλέχθηκε.

2. Βιβλιογραφική Επισκόπηση

Το Wisconsin Diagnostic Breast Cancer Diagnostic Dataset (WDBC) είναι ένα από τα πιο δημοφιλή και ευρέως χρησιμοποιημένα datasets στην επιστημονική κοινότητα για την ταξινόμηση του καρκίνου του μαστού. Εξετάζοντας την βιβλιογραφία είναι πολυάριθμες οι έρευνες στις οποίες αξιοποιήθηκε το WDBC σε συνδυασμό με μεθόδους ταξινόμησης μέσω μηχανικής μάθησης. Είναι σημαντικό να σημειωθεί πως το dataset λόγω της φύσης του παρέχει στους ερευνητές την ευελιξία να το χρησιμοποιούν από τους απλούς αλγορίθμους ταξινόμησης όπως Logistic regression μέχρι τους πιο σύνθετους όπως Neural Networks και Deep Learning κάνοντας έτσι πολύ πιο εύκολη την σύγκριση της απόδοσης των διάφορων αλγορίθμων. Σε έρευνα του 2022 το WDBC αξιοποιήθηκε για την αξιολόγηση των αλγορίθμων μηχανικής μάθησης Random Forest, Gradient Boosting, SVM, Artificial Neural Networks και Multilayer Perceptron για την πρόβλεψη του καρκίνου του μαστού (Sanam Aamir et al. 2022). Στην εν λόγω εργασία αξιοποιήθηκαν και τεχνικές βελτίωσης της ταξινόμησης όπως Connection-Based feature selection και 5-fold cross- validation με τελική ακρίβεια των μοντέλων 99.12%. Ενώ σε διαφορετική έρευνα του 2024 δημοσιευμένη στο International Journal of Data Science and Big Data Analytics, εξετάστηκαν επτά αλγόριθμοι μηχανικής μάθησης (Logistic Regression, Random Forest, SVM, Decision Tree, XGBoost, Naive Bayes) και πιο αποδοτικός από όλους αναδείχθηκε ο SVM με συνολική ακρίβεια 97.13%. (Arjun Kumar & Akinul Islam, 2024). Τέλος σε έρευνα που αξιοποιήθηκαν και τεχνικές νευρωσικών δικτύων παρατηρήθηκε ακρίβεια 98% στις προβλέψεις του μοντέλου, υψηλότερη από κάθε άλλο αλγόριθμο ταξινόμησης μηχανικής μάθησης που εφαρμόστηκε στην έρευνα (Sirisha Yerraboina et al. 2024).

3. Μεθοδολογία

Η μεθοδολογία της παρούσας εργασίας περιλαμβάνει τα εξής στάδια: προεπεξεργασία των δεδομένων, επιλογή αλγορίθμων, εκπαίδευση και αξιολόγηση μοντέλων, και ανάλυση των αποτελεσμάτων.

Αρχικά, πραγματοποιήθηκε προεπεξεργασία των δεδομένων, η οποία περιλάμβανε τον καθαρισμό του dataset, τη διαχείριση των ελλειπών τιμών και τη μετατροπή των κατηγοριών σε αριθμητικά δεδομένα όπου ήταν απαραίτητο.

Στη συνέχεια, επιλέχθηκαν τρεις αλγόριθμοι μηχανικής μάθησης: Logistic Regression, Random Forest και XGBoost, οι οποίοι εφαρμόστηκαν για την ταξινόμηση των όγκων σε καλοήθεις και κακοήθεις.

Η εκπαίδευση των μοντέλων πραγματοποιήθηκε με την μέθοδο Train-Test Split (80%-20%), όπου το dataset χωρίστηκε σε 80% δεδομένα εκπαίδευσης και 20% δεδομένα δοκιμών, επιτρέποντας την αξιολόγηση της απόδοσης του κάθε αλγορίθμου σε άγνωστα δεδομένα.

Ενώ η εκπαίδευση αξιολογήθηκε με 10-Fold Stratified Cross-Validation, μια μέθοδο που επιτρέπει τη στατιστικά πιο ακριβή αξιολόγηση, καθώς το dataset χωρίζεται σε 10 ίσα τμήματα, με κάθε μοντέλο να εκπαιδεύεται 10 φορές, χρησιμοποιώντας διαφορετικά υποσύνολα για εκπαίδευση και αξιολόγηση σε κάθε επανάληψη.

Τέλος, τα αποτελέσματα των μοντέλων συγκρίθηκαν με βάση την ακρίβεια, το precision, το recall και το F1-score, προκειμένου να αναδειχθεί το μοντέλο με την καλύτερη απόδοση.

3.1 Περιγραφή των δεδομένων

Το σύνολο των δεδομένων που αξιοποιήθηκε στην παρούσα εργασία είναι το Wisconsin Diagnostic Breast Cancer Dataset, το οποίο στο παρελθόν, έχει χρησιμοποιηθεί εκτενώς στην έρευνα για τον καρκίνο του μαστού. Το dataset περιλαμβάνει πολυάριθμα δείγματα βιοψίας όγκων του μαστού και κύρια ασχολία του είναι η ταξινόμηση των όγκων σε καλοήθεις ή κακοήθεις. Τελικός στόχος της ανάλυσης των δεδομένων είναι η ανάπτυξη ενός μοντέλου το οποίο θα προβλέπει την διάγνωση ενός νέου ασθενούς με βάση τα χαρακτηριστικά του δείγματός του.

Το dataset αποτελείται από 569 δείγματα όπου κάθε δείγμα αντιπροσωπεύει έναν ασθενή, ενώ για κάθε δείγμα υπάρχουν 30 χαρακτηριστικά τα οποία περιγράφουν φυσικά και στατιστικά στοιχεία των κυττάρων που συλλέχθηκαν κατά την διαδικασία της βιοψίας. Παράλληλα το dataset περιέχει και την κατηγορία της διάγνωσης κάθε δείγματος η οποία ορίζεται είτε ως B (Benign - καλοήθης) είτε ως M (Malignant - κακοήθης).

Τα χαρακτηριστικά του συνόλου δεδομένων είναι αριθμητικά και περιγράφουν την εικόνα της κάθε βιοψίας. Μπορούμε να χωρίσουμε τα χαρακτηριστικά σε τρείς βασικές κατηγορίες.

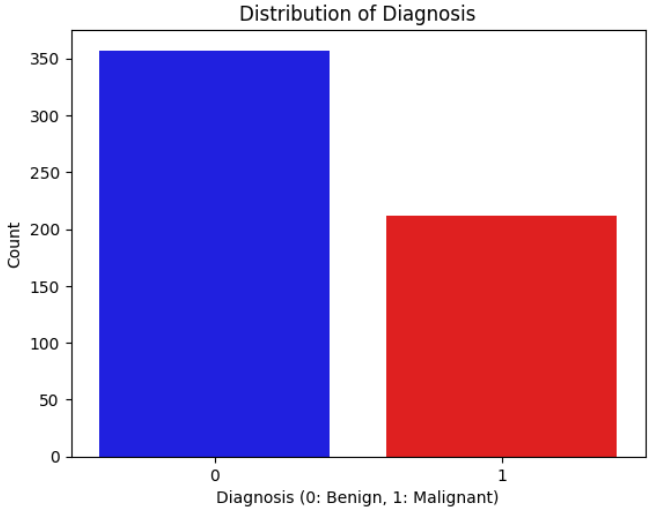
- **Μέση τιμή** χαρακτηριστικών (radius_mean, texture_mean) οι οποίες περιγράφουν τον μέσο όρο των μετρήσεων του κυττάρου.
- **Τυπική απόκλιση** χαρακτηριστικών (radius_se, texture_se) όπου περιγράφεται η μεταβλητότητα των μετρήσεων μεταξύ διαφορετικών κυττάρων.
- **Μέγιστη τιμή** χαρακτηριστικών (radius_worst, texture_worst) οι οποίες αναφέρονται στις πιο ακραίες τιμές που παρατηρήθηκαν στο σύνολο δεδομένων.

Είναι θετικό το γεγονός πως υπάρχουν ποικίλα χαρακτηριστικά όπως concavity_mean ή fractal_dimension_mean τα οποία παρέχουν διαφορετικές αλλά χρήσιμες πληροφορίες το κάθε ένα, που συμβάλλουν σημαντικά στην ανάπτυξη ενός μοντέλου πρόβλεψης.

3.2 Προεπεξεργασία των Δεδομένων

Στον τομέα της Μηχανικής Μάθησης, η προεπεξεργασία των δεδομένων, αποτελεί ένα από τα πιο κρίσιμα στάδια της διαδικασίας εκπαίδευσης καθώς μέσω αυτής διασφαλίζουμε την ποιότητα των δεδομένων και την βέλτιστη απόδοση των αλγορίθμων. Το dataset που αξιοποιήθηκε αρχικά περιείχε 569 δείγματα και 30 αριθμητικά χαρακτηριστικά τα οποία αφορούν στον καρκίνο του μαστού.

Στο πρώτο στάδιο της προεπεξεργασίας έγινε έλεγχος του συνόλου δεδομένων για ελλείπουσες τιμές και διπλότυπα, ενώ στην συνέχεια αφαιρέθηκε η στήλη ID καθώς δεν παρείχε κάποια χρήσιμη πληροφορία για την ανάπτυξη του μοντέλου πρόβλεψης. Στην συνέχεια εκτελέστηκε μετατροπή της στήλης διάγνωσης από B και M στην αριθμητική μορφή 0 και 1 όπου ως 0 χαρακτηρίζονται οι καλοήθεις όγκοι και ως 1 οι κακοήθεις, ώστε να μπορεί η στήλη να αξιοποιηθεί από τους αλγορίθμους μηχανικής μάθησης σε επόμενα στάδια.



Εικόνα #1 – Κατανομή της κλάσης «Διάγνωση» στις κατηγορίες Καλοήθης(0) & Κακοήθης(1)

Στο δεύτερο στάδιο της προεπεξεργασίας του συνόλου δεδομένων εκτελέστηκε έλεγχος ακραίων τιμών με την μέθοδο Z-score, με την οποία υπολογίστηκε πόσο απέχει κάθε τιμή από την μέση τιμή του εκάστοτε χαρακτηριστικού σε μονάδες τυπικής απόκλισης. Όσο μεγαλύτερη είναι η τυπική απόκλιση τόσο πιο διασκορπισμένες είναι οι τιμές του χαρακτηριστικού γύρω από την μέση τιμή τους. Για την εκτέλεση της εργασίας ορίστηκαν ως ακραίες, οι τιμές με Z-score μεγαλύτερο του 3. Η επιλογή αυτού του ορίου μας επιτρέπει να αποκλείσουμε μόνο τις πιο ακραίες τιμές που θα προσέθεταν θόρυβο. Δεν διαγράφηκαν ακραίες τιμές καθώς ήταν όλες σε επιτρεπτά όρια.

Στην συνέχεια πραγματοποιήθηκε επιλογή Χαρακτηριστικών (Feature Selection) στο dataset. Για τον σκοπό αυτό αξιοποιήθηκε η τεχνική Variance Threshold, κατά την οποία γίνεται εντοπισμός και αφαίρεση όλων των χαρακτηριστικών με πολύ χαμηλή

διακύμανση. Όταν κάποιο χαρακτηριστικό έχει πολύ χαμηλή διακύμανση δεν μας παρέχει σημαντική πληροφορία, καθώς δεν υπάρχει διαφοροποίηση μεταξύ των δειγμάτων και για τον λόγω αυτό δεν μας βοηθάει στην ανάπτυξη μοντέλου πρόβλεψης. Για την εργασία χρησιμοποιήθηκε Variance Threshold = 0.1, δηλαδή απομακρύνθηκαν όλα τα χαρακτηρίστηκα με διακύμανση μικρότερη από 0.1. Τα χαρακτηριστικά που διατηρήθηκαν μετά την εφαρμογή του Variance Threshold (0.01) είναι: Radius1, Texture1, Perimeter1, Area1, Radius2, Texture2, Perimeter2, Area2, Radius3, Texture3, Perimeter3, Area3, Compactness3, Concavity3.

Τέλος σχετικά με την μικρή ανισορροπία που εντοπίζουμε στο dataset, σαν μεθοδολογία, δεν εκτελούμε καμία ενέργεια καθώς πρώτα θα εκτελέσουμε του αλγορίθμους μηχανικής μάθησης και εφόσον εντοπίσουμε biased μοντέλα θα γυρίσουμε σε προηγούμενα βήματα για να προσθέσουμε τεχνικές εξισορρόπησης. Τελικά στην περίπτωση μας οι αλγόριθμοι προσαρμόστηκαν και δεν χρειάστηκε εξισορρόπηση των κλάσεων

Τα στάδια προεπεξεργασίας που εφαρμόστηκαν είναι κρίσιμα για την επίτευξη της βέλτιστης αποδοτικότητας των μοντέλων μηχανικής μάθησης καθώς με της παραπάνω τεχνικές απομακρύνθηκαν δεδομένα που προσέθεταν θόρυβο και στα δεδομένα που διατηρήθηκαν διασφαλίστηκε η ομοιόμορφη κατανομή τους.

3.3 Εκπαίδευση των Μοντέλων Μηχανικής Μάθησης

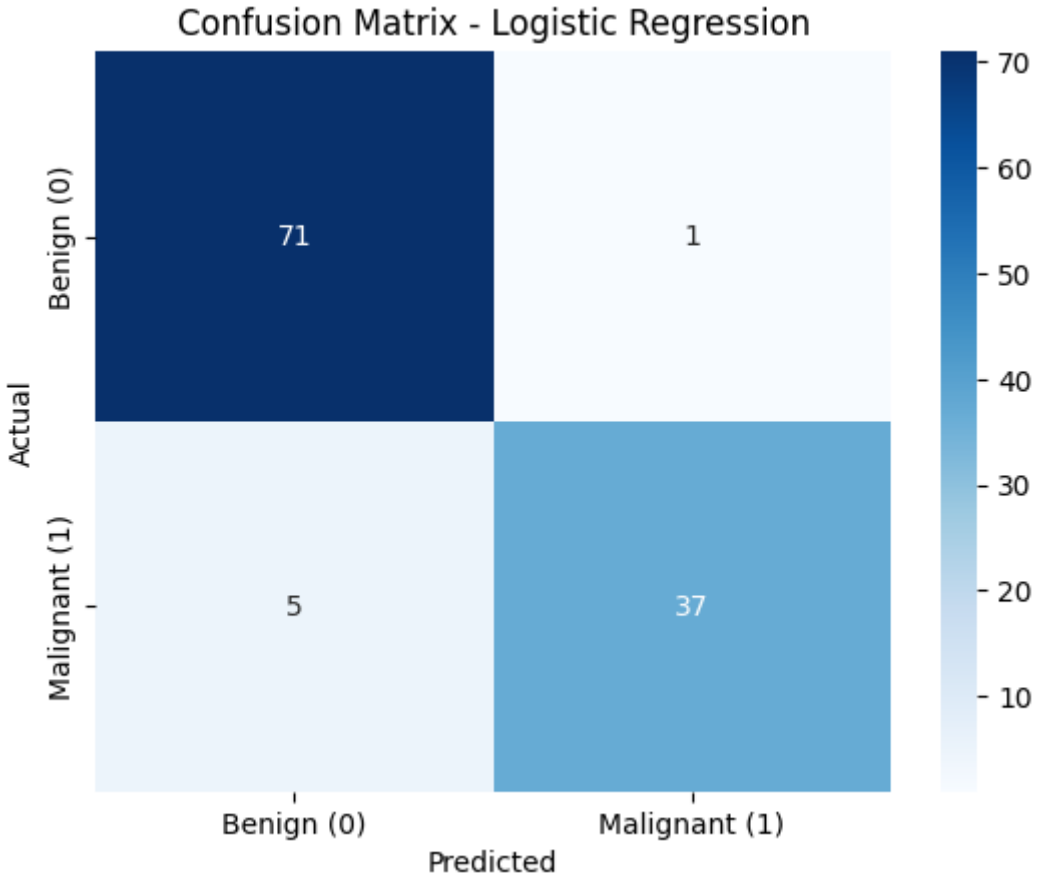
Αφού ολοκληρώθηκε το στάδιο της προεπεξεργασίας των δεδομένων ακολουθήσε η ανάπτυξη μοντέλων Μηχανικής Μάθησης για τα οποία αξιοποιήθηκαν οι αλγόριθμοι: Logistic Regression, Random Forest Classifier, XGBoost, με τελικό στόχο την ακριβή ταξινόμηση των όγκων των δειγμάτων. Για την εκπαίδευση όλων των μοντέλων έγινε χρήση Train Test Split της τάξεως 80%-20%

3.3.1 Logistic Regression

Η Λογιστική Παλινδρόμηση είναι ένας γραμμικός αλγόριθμος μηχανικής μάθησης αρκετά απλός αλλά πολύ αποτελεσματικός στην δυαδική ταξινόμηση, σε σχέση με τα διάφορα χαρακτηριστικά του εκάστοτε δείγματος. Η Λογιστική Παλινδρόμηση είναι ιδιαίτερα χρήσιμη όταν τα δεδομένα έχουν γραμμική σχέση με την έξοδο, ενώ είναι ερμηνεύσιμη και εύκολη στην εφαρμογή.

Στην περίπτωση της παρούσας εργασίας με βάση τα δεδομένα του dataset τα οποία χρησιμοποιήθηκαν ως input ο αλγόριθμος εκτέλεσε προβλέψεις ταξινομώντας διάφορα δείγματα στις κατηγορίες 0 (Καλοήθης) ή 1 (Κακοήθης) με μεγάλη ακρίβεια και αποτελεσματικότητα. Είναι σημαντικό να σημειωθεί πως κατά την εκτέλεση του αλγορίθμου ορίστηκε η μεταβλητή stratify = y ώστε να εξασφαλίσουμε την ίδια κατανομή των κατηγοριών διάγνωσης και στο train και στο test split. Στην εικόνα που ακολουθεί απεικονίζονται τα αποτελέσματα της εκτέλεσης του αλγορίθμου Logistic Regression στον οποίο είχαμε συνολικά: **37 True Positives** και **71 True Negatives** δηλαδή 108 συνολικά σωστά ταξινομημένες προβλέψεις και μόλις **1 False Positives** και **5 False Negatives**, δηλαδή 6 λανθασμένα ταξινομημένες προβλέψεις.

Συνολική ακρίβεια του μοντέλου και υπόλοιπες μετρικές θα υπολογιστούν στην συνέχεια. Από τον πίνακα σύγχυσης του μοντέλου όμως έχουμε πολύ ικανοποιητικές προβλέψεις

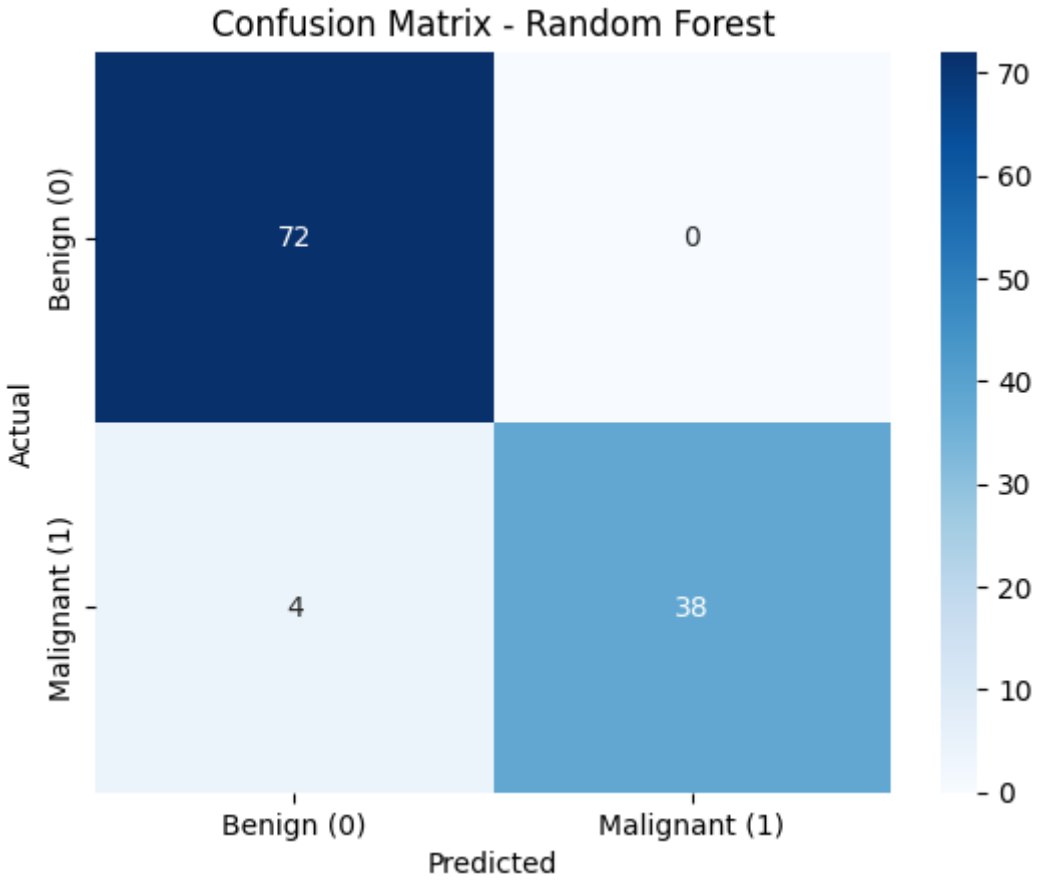


Εικόνα #2 – Πίνακας Σύγχυσης για το μοντέλο Λογιστικής Παλινδρόμησης

3.3.2 Random Forest Classifier

Ο αλγόριθμος Random Forest Classifier είναι ένας αρκετά ισχυρός αλγόριθμος ταξινόμησής ο οποίος βασίζεται στην μέθοδο δημιουργίας δένδρων απόφασης και της ενσωματωμένης μηχανικής μάθησης. Δημιουργεί έναν μεγάλο αριθμό δέντρων απόφασης κατά τη διαδικασία εκπαίδευσης και συνδυάζει τις προβλέψεις τους για να παράγει ένα πιο ακριβές και σταθερό αποτέλεσμα. Το κύριο πλεονέκτημά του είναι ότι μειώνει το πρόβλημα του overfitting, το οποίο συχνά επηρεάζει τα μεμονωμένα δέντρα απόφασης, εξασφαλίζοντας καλύτερη γενίκευση στα δεδομένα. Ο αλγόριθμος αυτός είναι ιδιαίτερα χρήσιμος σε προβλήματα με πολλαπλές μεταβλητές εισόδου και δεδομένα με μεγάλη διακύμανση

Στην εικόνα που ακολουθεί απεικονίζονται τα αποτελέσματα της εκτέλεσης του Random Forest Classifier από τον οποίο λάβαμε συνολικά: **38 True Positives** και **72 True Negatives** δηλαδή 110 συνολικά σωστά ταξινομημένες προβλέψεις και μόλις **0 False Positives** και **4 False Negatives**, δηλαδή 4 λανθασμένα ταξινομημένες προβλέψεις.



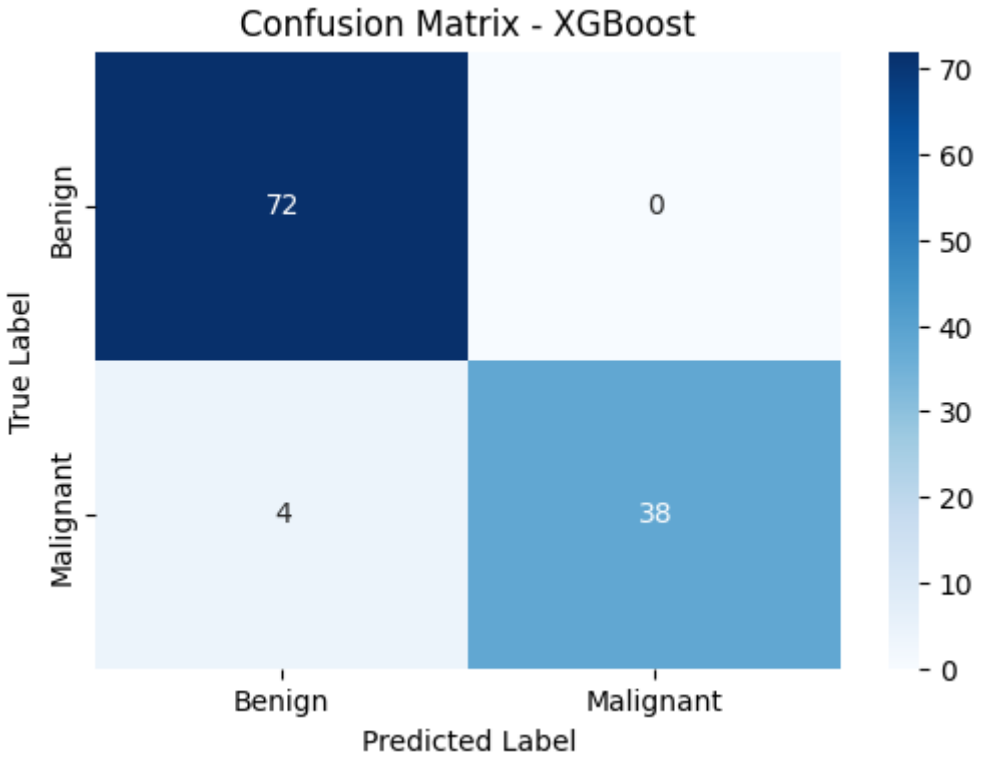
Εικόνα #3 – Πίνακας Σύγχυσης για το μοντέλο Random Forest Classifier

Συνολική ακρίβεια του μοντέλου και υπόλοιπες μετρικές θα υπολογιστούν στην συνέχεια. Από τον πίνακα σύγχυσης του μοντέλου όμως έχουμε πολύ ικανοποιητικές προβλέψεις

3.3.3 XGBoost

Ο αλγόριθμος XGBoost (Extreme Gradient Boosting) είναι ένας από τους πιο αποδοτικούς και σύγχρονους αλγορίθμους ταξινόμησης. Βασίζεται σε τεχνικές ενίσχυσης (boosting), όπου διαδοχικά μοντέλα μαθαίνουν από τα λάθη των προηγούμενων, δημιουργώντας έναν τελικό ισχυρό ταξινομητή. Στην παρούσα εργασία εφαρμόστηκε με χρήση παραμέτρων που εξασφάλιζαν σταθερότητα και αποφυγή υπερεκπαίδευσης (όπως use_label_encoder=False, eval_metric='logloss'). Ο XGBoost διαχειρίστηκε αποτελεσματικά τα δεδομένα, παρουσιάζοντας αξιόπιστες προβλέψεις και ανταγωνιστικά αποτελέσματα σε σχέση με τους άλλους αλγορίθμους.

Στην εικόνα που ακολουθεί απεικονίζονται τα αποτελέσματα της εκτέλεσης του Random Forest Classifier από τον οποίο λάβαμε συνολικά: **38 True Positives** και **72 True Negatives** δηλαδή 110 συνολικά σωστά ταξινομημένες προβλέψεις και μόλις **0 False Positives** και **4 False Negatives**, δηλαδή 4 λανθασμένα ταξινομημένες προβλέψεις.



Εικόνα #4 – Πίνακας Σύγχυσης για τον αλγόριθμο XGBoost

Συνολική ακρίβεια του μοντέλου και υπόλοιπες μετρικές θα υπολογιστούν στην συνέχεια. Από τον πίνακα σύγχυσης του μοντέλου όμως έχουμε πολύ ικανοποιητικές προβλέψεις

3.4 Ανάλυση και Αξιολόγηση των Αποτελεσμάτων

Για την αξιολόγηση της απόδοσης των μοντέλων, εφαρμόστηκε η μέθοδος 10-Fold Stratified Cross-Validation πάνω σε ολόκληρο το σύνολο δεδομένων. Η τεχνική αυτή επιτρέπει την πιο αντικειμενική αποτίμηση της γενικής ικανότητας των μοντέλων, καθώς όλα τα δεδομένα χρησιμοποιούνται τόσο για εκπαίδευση όσο και για έλεγχο, σε εναλλασσόμενα υποσύνολα.

Σύμφωνα με τα αποτελέσματα, το μοντέλο **Random Forest** σημείωσε την υψηλότερη ακρίβεια με **95.79%**, ακολουθούμενο από το **Logistic Regression** με **95.26%** και τον **XGBoost** με **94.73%**. Η διαφορά είναι μικρή, και δείχνει ότι και τα τρία μοντέλα είχαν εξαιρετική απόδοση.

Αναλύοντας τις υπόλοιπες μετρικές, ο Random Forest κατέγραψε:

1. Recall 93.33%
2. Precision 95.67%
3. F1-score 94.21%
4. ROC AUC 98.66%.

Το Logistic Regression εμφάνισε

1. Recall 91.45%
2. Precision 95.87%
3. F1-score 93.31%
4. ROC AUC 99.19%

επιτυγχάνοντας την υψηλότερη τιμή στην ευαισθησία καλοηθών περιπτώσεων (Recall_Benign = 97.46%).

Ο XGBoost είχε επίσης σταθερές επιδόσεις με

1. Recall 92.38%
2. Precision 93.70%
3. F1-score 92.66%
4. ROC AUC 98.97%,

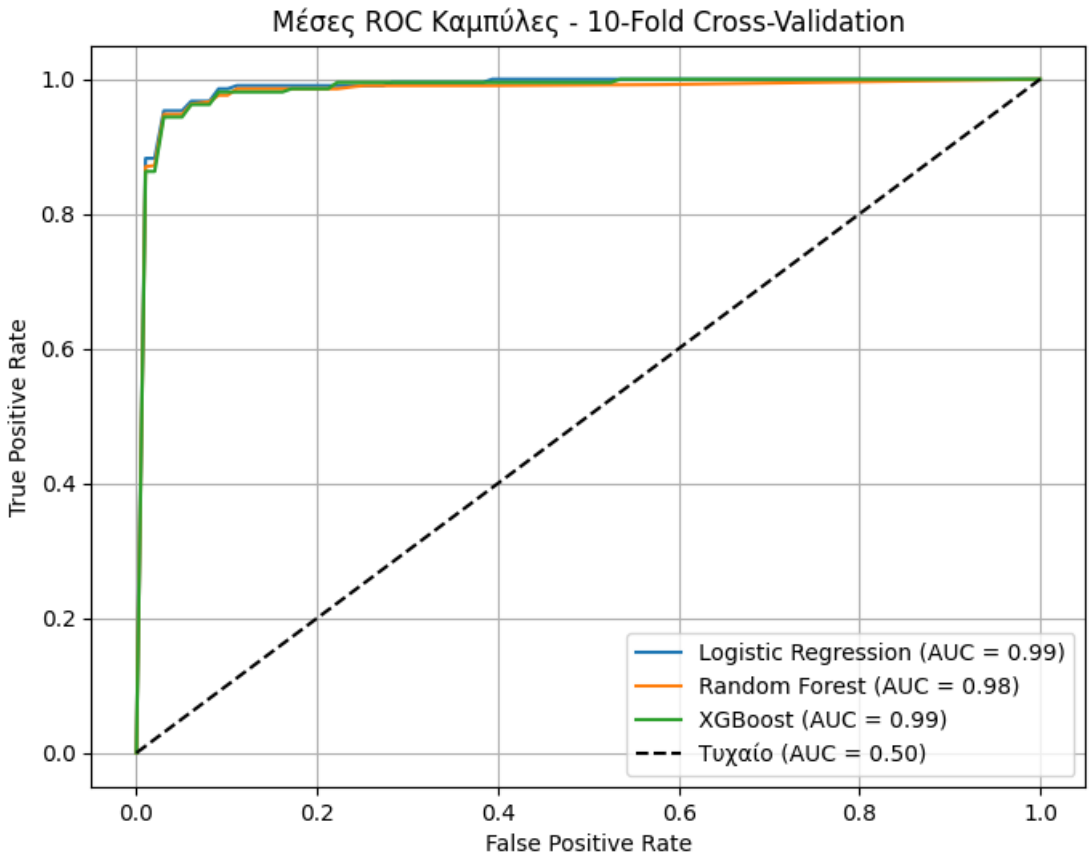
διατηρώντας την καλύτερη ισορροπία ανάμεσα στις δύο κατηγορίες με Recall_Benign = 96.07% και Recall_Malignant = 92.38%.

Τα παραπάνω επιβεβαιώνουν ότι όλα τα μοντέλα είναι ικανά για ταξινόμηση όγκων με υψηλή ακρίβεια, ενώ το Random Forest εμφανίζεται ελαφρώς ανώτερο στις συνολικές μετρικές. Ωστόσο, το Logistic Regression παρουσιάζει την καλύτερη ικανότητα διάκρισης μεταξύ των καλοήθων περιπτώσεων, και ο XGBoost παρέχει σταθερές και ισορροπημένες προβλέψεις.

Πίνακες αποτελεσμάτων 10-Fold Stratified Cross-Validation

	Accuracy	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.9526	0.9587	0.9145	0.9331
Random Forest	0.9578	0.9567	0.9333	0.9421
XGBoost	0.947	0.9370	0.9238	0.9266

	Recall Benign	**Recall Malignant**
Logistic Regression	0.9746	0.9145
Random Forest	0.9717	0.9333
XGBoost	0.9607	0.9238



Εικόνα #6 – Roc-Auc Curves των μοντέλων που αξιοποιήθηκαν

3.5 Συμπεράσματα και προτάσεις για μελλοντικές βελτιώσεις

Η εργασία ανέδειξε την αποτελεσματικότητα διαφόρων αλγορίθμων μηχανικής μάθησης στην πρόβλεψη του καρκίνου του μαστού, αξιοποιώντας το Wisconsin Diagnostic Breast Cancer Dataset. Μέσα από προσεκτική προεπεξεργασία των δεδομένων και εφαρμογή 10-Fold Cross Validation, διασφαλίστηκε η αξιόπιστη αξιολόγηση των μοντέλων.

Τα αποτελέσματα ανέδειξαν τον Random Forest ως το πιο αποδοτικό μοντέλο με βάση την συνολική του απόδοση. Ο Logistic Regression κατέγραψε ιδιαίτερα υψηλή ακρίβεια και ήταν εξαιρετικός στην ανίχνευση καλοήθων περιπτώσεων, ενώ ο XGBoost παρείχε ιδιαίτερα σταθερές και ισορροπημένες προβλέψεις μεταξύ των κατηγοριών. Η συνολική ακρίβεια και των τριών μοντέλων υπερέβη το 94%, κάτι που επιβεβαιώνει τη χρησιμότητά τους σε προβλήματα ιατρικής διάγνωσης.

Στο μέλλον, η ενσωμάτωση πιο σύνθετων αλγορίθμων όπως τα τεχνητά νευρωνικά δίκτυα (Neural Networks) ή συνδυαστικά μοντέλα (ensemble learning) θα μπορούσε να ενισχύσει περαιτέρω την απόδοση. Επιπλέον, η εφαρμογή τεχνικών εξισορρόπησης κλάσεων και αυτοματοποιημένη επιλογή χαρακτηριστικών θα μπορούσαν να βελτιστοποιήσουν ακόμη περισσότερο την αποτελεσματικότητα των προβλέψεων.

Παράρτημα Κώδικα των Μοντέλων

```python
df = pd.read_csv("C:/Users/nbala/Desktop/BreCanPred/Dataset/wdbc_cleaned.csv")
X = df.drop(columns=['Diagnosis'])
y = df['Diagnosis']

# διαχωρισμός σε training και test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y

# εκτέλεση της Logistic Regression
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

# προβλέψεις
y_pred = model.predict(X_test)

# Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", xticklabels=["Benign (0)", "Malignant (1
            yticklabels=["Benign (0)", "Malignant (1)"])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix - Logistic Regression")
plt.show()
```

✓ 0.3s Python

```python
# εκτέλεση του Random Forest
model = RandomForestClassifier(n_estimators=100, random_state=42, bootstrap=True, max_depth=10)
model.fit(X_train, y_train)

# προβλέψεις
y_pred = model.predict(X_test)

# Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", xticklabels=["Benign (0)", "Malignant (1
            yticklabels=["Benign (0)", "Malignant (1)"])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix - Random Forest")
plt.show()
```

✓ 0.2s Python

```python
# Διαχωρισμός χαρακτηριστικών και ετικετών
X = df.drop(columns=['Diagnosis'])
y = df['Diagnosis']

# Διαχωρισμός σε training και test set (80-20)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, stratify=y, random_state=4

# Ορισμός και εκπαίδευση του XGBoost μοντέλου
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)
xgb_model.fit(X_train, y_train)

# Πρόβλεψη στο test set
y_pred = xgb_model.predict(X_test)

# Υπολογισμός και εκτύπωση του Confusion Matrix
cm = confusion_matrix(y_test, y_pred)

# Οπτικοποίηση του Confusion Matrix
plt.figure(figsize=(6,4))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=["Benign", "Malignant"], yticklabels=
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.title("Confusion Matrix - XGBoost")
plt.show()
```

✓ 0.4s Python

Βιβλιογραφία

[1] Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., Alhaisoni, M., Khan, M. A., Khan, K., & Ahmad, J. (2022). Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. Computational and mathematical methods in medicine, 2022, 5869529.
https://doi.org/10.1155/2022/5869529

[2] Arjun Kumar Bose Arnob and Akinul Islam Jony / Int.J.Data.Sci. & Big Data Anal. (2024). Comparing Machine Learning Algorithms for Breast Cancer Diagnosis: Wisconsin Diagnostic Dataset Analysis
https://doi.org/10.51483/IJDSBDA.4.2.2024.1-11

[3] Sirisha Yerraboina, Harikrishna Bommala and Vineela Madireddy (2024). Breast cancer classification and prediction methods by employing machine and deep learning approaches-A survey
MATEC Web Conf., 392 (2024) 01137
https://doi.org/10.1051/matecconf/202439201137