

# ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

**“Σύστημα υποστήριξης απόφασης για τον καρκίνο του μαστού”**



**Υποψήφιος:** Μπαλτατζίδης Κυριάκος (Π2018176)

**Επιβλέπων:** Θεμιστοκλής Έξαρχος (Επικουρος Καθηγητής)

Κέρκυρα, 2024

## Περίληψη:

Η πτυχιακή εργασία πραγματεύεται την δημιουργία ενός συστήματος υποστήριξης απόφασης σχετικά με τον καρκίνο του μαστού, χρησιμοποιώντας τεχνικές μηχανικής μάθησης και τεχνητής νοημοσύνης. Ο σκοπός της εργασίας είναι μέσω ενός συνόλου δεδομένου αποτελούμενο από ιατρικά δεδομένα να αναπτυχθεί ένα σύστημα που να εντοπίζει τους ασθενείς με χρήση αλγορίθμων μηχανικής μάθησης. Η εργασία προσεγγίζει το πρόβλημα αυτό με την συλλογή και ανάλυση μεγάλου όγκου ιατρικών δεδομένων. Η εφαρμογή που παρουσιάζεται εδώ συνδυάζει σύγχρονες τεχνολογίες για την ανάλυση δεδομένων και την πρόβλεψη της υγείας μέσω ενός διαδικτυακού εργαλείου. Αποτελείται από δύο κύρια τμήματα: το πρώτο αφορά τη διαδικασία εκπαίδευσης ενός μοντέλου μηχανικής μάθησης για την πρόβλεψη του καρκίνου του μαστού, ενώ το δεύτερο είναι μια διαδικτυακή εφαρμογή που επιτρέπει στους χρήστες να αλληλεπιδρούν με το μοντέλο και να λαμβάνουν προγνώσεις για την υγεία τους. Το πρώτο τμήμα της εφαρμογής αφορά την εκπαίδευση ενός μοντέλου πρόβλεψης με τη χρήση δεδομένων που σχετίζονται με τον καρκίνο του μαστού. Το αρχείο δεδομένων φορτώνεται από μια βάση δεδομένων, όπου οι ετικέτες των δεδομένων είναι αναγνωρισμένες ως "Υγιής" ή "Μη Υγιής". Το μοντέλο εκπαιδεύεται χρησιμοποιώντας έναν αλγόριθμο μηχανικής μάθησης, γνωστό ως AdaBoost. Αυτός ο αλγόριθμος είναι δημοφιλής για τη βελτίωση της απόδοσης άλλων αλγορίθμων μηχανικής μάθησης μέσω της συνδυασμένης χρήσης πολλαπλών απλών μοντέλων.

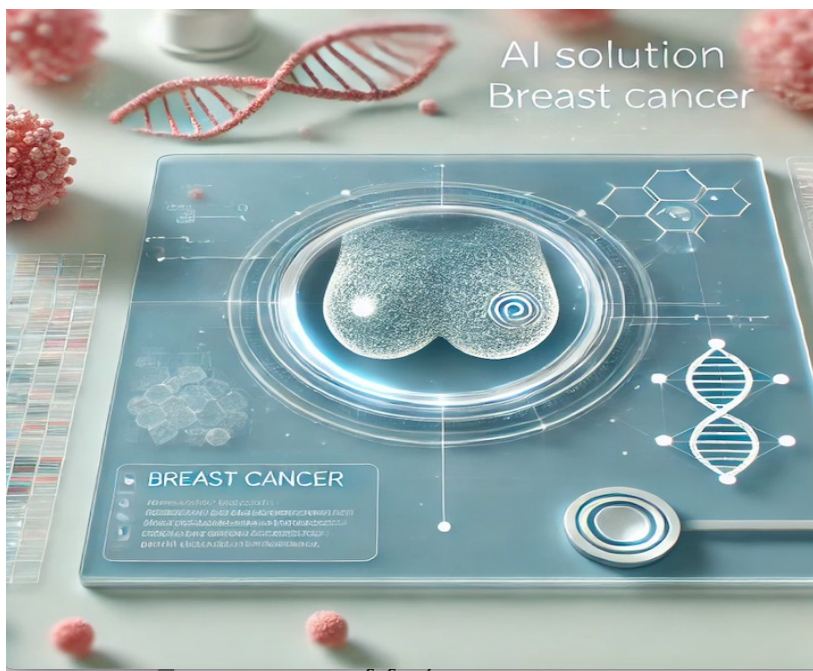
Ο AdaBoost εκπαιδεύεται με διάφορους συνδυασμούς παραμέτρων, όπως ο αριθμός των δέντρων αποφάσεων (base estimators) που χρησιμοποιούνται και η ταχύτητα μάθησης του μοντέλου. Με την ολοκλήρωση της εκπαίδευσης, το καλύτερο μοντέλο αποθηκεύεται για μελλοντική χρήση. Τα αποτελέσματα των δοκιμών αξιολογούνται με τη χρήση διαγράμματος που συγκρίνει τις επιδόσεις του μοντέλου με βάση διάφορες παραμέτρους. Το διάγραμμα αυτό βοηθά στην κατανόηση της απόδοσης του μοντέλου και των επιρροών των διαφορετικών παραμέτρων.

Δημιουργία της Διαδικτυακής Εφαρμογής

Το δεύτερο τμήμα της εφαρμογής είναι μια διαδικτυακή εφαρμογή που επιτρέπει στους χρήστες να αλληλεπιδρούν με το εκπαιδευμένο μοντέλο μέσω μιας απλής και φιλικής προς το χρήστη διεπαφής. Ο χρήστης εισάγει δεδομένα όπως ο Δείκτης Μάζας Σώματος (BMI), τα επίπεδα γλυκόζης, ινσουλίνης και άλλες παραμέτρους μέσω μιας φόρμας. Αυτά τα δεδομένα αποστέλλονται σε έναν διακομιστή, ο οποίος χρησιμοποιεί το αποθηκευμένο μοντέλο για να προβλέψει την κατάσταση της υγείας του χρήστη.

Αν το μοντέλο προβλέψει ότι ο χρήστης είναι "Μη Υγιής", η διεπαφή θα εμφανίσει μια κόκκινη ένδειξη. Αν η πρόβλεψη είναι "Υγιής", η ένδειξη θα είναι πράσινη. Η διεπαφή είναι σχεδιασμένη για να είναι εύκολη στη χρήση και να παρέχει σαφή αποτελέσματα στον χρήστη.

Αναλυτική Εξήγηση της Διαδικασίας



### 1. Φόρτωση και Επεξεργασία Δεδομένων:

- Το αρχείο δεδομένων φορτώνεται σε ένα περιβάλλον ανάλυσης δεδομένων.
- Οι ετικέτες των δεδομένων μετατρέπονται σε κατανοητά κείμενα ("Υγιής", "Μη Υγιής").
- Τα δεδομένα προετοιμάζονται για την εκπαίδευση του μοντέλου.

### 2. Εκπαίδευση Μοντέλου:

- Διάφορες παραλλαγές του αλγορίθμου AdaBoost εκτελούνται με διαφορετικούς συνδυασμούς παραμέτρων.
- Το μοντέλο αξιολογείται με τη χρήση σταυρωτής επικύρωσης για να διασφαλιστεί η ποιότητά του.
- Ο καλύτερος συνδυασμός παραμέτρων επιλέγεται και το αντίστοιχο μοντέλο αποθηκεύεται για μελλοντική χρήση.

### 3. Δημιουργία Διαδικτυακής Εφαρμογής:

- Η εφαρμογή δημιουργείται χρησιμοποιώντας την πλατφόρμα Flask, που είναι δημοφιλής για την ανάπτυξη διαδικτυακών εφαρμογών.
- Η διεπαφή χρήστη σχεδιάζεται με HTML και CSS για να είναι οπτικά ελκυστική και εύκολη στη χρήση.
- Ο χρήστης εισάγει τα δεδομένα του και υποβάλλει τη φόρμα.
- Το σύστημα επεξεργάζεται τα δεδομένα και εμφανίζει την πρόβλεψη στην οθόνη του χρήστη.

### Εξέταση Εφαρμογής

Η εφαρμογή που συνδυάζει τη μηχανική μάθηση με την τεχνολογία ιστού επιτρέπει την εύκολη και γρήγορη πρόβλεψη της υγείας του χρήστη. Η συνεχής αξιολόγηση και βελτίωση του μοντέλου εξασφαλίζει ότι οι προβλέψεις είναι όσο το δυνατόν πιο ακριβείς. Η δυνατότητα αποθήκευσης του καλύτερου μοντέλου και η χρήση του μέσω μιας διαδικτυακής διεπαφής κάνουν την εφαρμογή προσβάσιμη και χρήσιμη σε ένα ευρύ κοινό.

Η υλοποίηση της διαδικτυακής εφαρμογής είναι το κλειδί για τη διάδοση και την εύκολη πρόσβαση σε τέτοιες τεχνολογίες. Η καλή σχεδίαση της διεπαφής χρήστη και η ικανότητα του μοντέλου να παρέχει ακριβείς προβλέψεις είναι κρίσιμα για την επιτυχία της εφαρμογής. Με τη συνεχή αναβάθμιση και αξιολόγηση, η εφαρμογή μπορεί να βελτιωθεί περαιτέρω για να προσφέρει ακόμα πιο ακριβείς και χρήσιμες προγνώσεις υγείας.

## **Abstract:**

The thesis deals with the creation of a decision support system for breast cancer using machine learning and artificial intelligence techniques. The goal of the work is to develop a system that can identify patients through the use of machine learning algorithms, utilizing a dataset of medical data. The approach involves the collection and analysis of a large volume of medical data. The application presented here combines modern technologies for data analysis and health prediction through an online tool. It consists of two main components: the first involves the process of training a machine learning model for breast cancer prediction, while the second is a web application that allows users to interact with the model and receive health predictions.

The first part of the application involves training a prediction model using data related to breast cancer. The dataset is loaded from a database, where the data labels are recognized as "Healthy" or "Unhealthy". The model is trained using a machine learning algorithm known as AdaBoost. This algorithm is popular for improving the performance of other machine learning algorithms through the combined use of multiple simple models.

AdaBoost is trained with various combinations of parameters, such as the number of decision trees (base estimators) used and the learning rate of the model. Once training is complete, the best model is saved for future use. The results of the tests are evaluated using a chart that compares the model's performance based on various parameters. This chart helps in understanding the model's performance and the influence of different parameters.

## **Creation of the Web Application:**

The second part of the application is a web-based tool that allows users to interact with the trained model through a simple and user-friendly interface. Users input data such as Body Mass Index (BMI), glucose levels, insulin levels, and other parameters via a form. This data is sent to a server, which uses the stored model to predict the user's health status.

If the model predicts that the user is "Unhealthy," the interface will display a red indicator. If the prediction is "Healthy," the indicator will be green. The interface is designed to be easy to use and to provide clear results to the user.

## **Detailed Explanation of the Process:**

### **1. Loading and Processing Data:**

- The dataset is loaded into a data analysis environment.
- Data labels are converted into understandable texts ("Healthy", "Unhealthy").
- The data is prepared for model training.

### **2. Training the Model:**

- Various versions of the AdaBoost algorithm are executed with different parameter combinations.
- The model is evaluated using cross-validation to ensure its quality.
- The best parameter combination is selected, and the corresponding model is saved for future use.

### **3. Creating the Web Application:**

- The application is developed using the Flask platform, which is popular for web application development.
- The user interface is designed with HTML and CSS to be visually appealing and easy to use.
- Users enter their data and submit the form.
- The system processes the data and displays the prediction on the user's screen.

### Application Review:

The application, which combines machine learning with web technology, enables easy and quick health predictions for users. Continuous evaluation and improvement of the model ensure that predictions are as accurate as possible. The ability to store the best model and use it through an online interface makes the application accessible and useful to a wide audience. The implementation of the web application is key to the dissemination and easy access to such technologies. Good user interface design and the model's ability to provide accurate predictions are crucial for the success of the application. With ongoing upgrades and evaluations, the application can be further improved to offer even more accurate and useful health predictions.

## Πίνακας Περιεχομένων

Περίληψη:.....	2
ABSTRACT:.....	4
Κεφάλαιο 1 – Καρκίνος του Μαστού.....	8
1.1 Εισαγωγή.....	8
1.2 Παγκόσμια Στατιστικά Στοιχεία για τον Καρκίνο του Μαστού.....	9
1.3 Ιστορική αναδρομή στην πρόβλεψη του καρκίνου το μαστού με τεχνικές μηχανικής μάθησης:.....	17
1.4 Αναφορά και επισκόπηση της βιβλιογραφίας.....	23
1.5 Επισκόπηση δεδομένων (Data Overview).....	32
AGE - ΗΛΙΚΙΑ.....	32
BMI – ΔΕΙΚΤΗΣ ΜΑΖΑΣ ΣΩΜΑΤΟΣ.....	33
GLUCOSE – ΓΛΥΚΟΖΗ.....	34
INSULIN - ΙΝΣΟΥΛΙΝΗ.....	34
HOMA.....	35
LEPTIN - ΛΕΠΤΙΝΗ.....	36
ADIPONECTIN - ΑΝΤΙΠΟΝΕΚΤΙΝΗ.....	37
RESISTIN – ΡΕΣΙΣΤΙΝΗ.....	38
MCP1.....	39
CLASSIFICATION – ΚΛΑΣΗ ΤΑΞΙΝΟΜΗΣΗΣ.....	40
1.6 Συμπεράσματα.....	44
Κεφάλαιο 2 – Επεξεργασία στο Σύνολο Δεδομένων.....	45
2. Προεπεξεργασία δεδομένων (Data Preprocessing).....	45
2.1 Ετικετοποίηση (Labeling).....	46
2.2 Προεπεξεργασία στο σετ δεδομένων Coimbra Breast Cancer. (Labeling).....	46

Κεφάλαιο 3 – Χρήση αλγορίθμων και Μεθόδων.....	47
3 Χτήσιμο Μοντέλου (Model Building).....	47
3.1 Διασταυρούμενη επικύρωση (Cross Validation).....	47
3.2 Αξιολόγηση επιδόσεων (Performance Evaluation).....	47
3.3 kNN – K: Πλησιέστερος γείτονας (k-Nearest Neighbor).....	48
Δομή.....	48
Πλεονεκτήματα - Μειονεκτήματα.....	49
Τιμές του k, Mean Accuracy και F-Score.....	49
3.4 Δέντρο αποφάσης AdaBoost (AdaBoost Decision Tree).....	51
Δομή:.....	51
Πλεονεκτήματα:.....	51
Μειονεκτήματα:.....	52
Τεχνικές Βελτίωσης.....	52
Καλύτεροι Παράμετροι:.....	52
Αποτελέσματα:.....	53
3.5 Naive Bayes.....	53
Είδη Naïβ Bayes (Types of Naive Bayes).....	53
Καλύτεροι Παράμετροι:.....	54
3.6 Random Forest (Τυχαία Δάση).....	55
Δημιουργία Δέντρων Απόφασης:.....	55
Τυχαία Υποομάδα Δεδομένων (Bootstrap Sampling):.....	55
Τυχαία Υποομάδα Χαρακτηριστικών (Feature Randomness):.....	55
Συνδυασμός Προβλέψεων:.....	55
Καλύτεροι Παράμετροι:.....	56
Αποτελέσματα:.....	56
3.7 J48.....	57
2. Βασικά Χαρακτηριστικά.....	57
3. Τεχνική Ανάλυση.....	57
4. Πλεονεκτήματα και Μειονεκτήματα.....	57
5. Χρήση και Εφαρμογές.....	58
Καλύτεροι Παράμετροι:.....	58
Αποτελέσματα:.....	58
3.8 Gradient Boost.....	59
1. Αρχή Λειτουργίας.....	59
2. Βασικά Χαρακτηριστικά.....	59
3. Τεχνική Ανάλυση.....	59
4. Πλεονεκτήματα και Μειονεκτήματα.....	60
5. Χρήση και Εφαρμογές.....	60
Καλύτεροι Παράμετροι:.....	61
Αποτελέσματα:.....	61
Κεφάλαιο 4 – Συμπεράσματα Αποτελέσματα & Σχολιασμός.....	61
4.1 Σύνοψη Αποτελεσμάτων.....	61
4.1.1 AdaBoost (Decision Tree).....	61
4.1.2 k-Nearest Neighbor (kNN).....	61
4.1.3 Random Forest, J48 και Gradient Boosting.....	62
4.1.4 Naive Bayes.....	62
4.2 Ερμηνεία Αποτελεσμάτων.....	63
4.3 Συμπεράσματα Χησίματος του Μοντέλου.....	63
Κεφάλαιο 5 – Δημιουργία και Ανάπτυξη Εφαρμογής Ιστοσελίδας για Πρόβλεψη Καρκίνου του Μαστού.....	64

5.1 Αιτίες Χρήσης Διακομιστή Ιστού.....	64
5.2 Οφέλη της Χρήσης Διακομιστή Ιστού.....	65
5.3 Συμπεράσματα.....	65
Κεφάλαιο 6: Ανάλυση της Εφαρμογής Ιστού για Πρόβλεψη Καρκίνου του Μαστού.....	65
6.1 Σκοπός και Λειτουργία της Εφαρμογής.....	65
6.2 Διεπαφή Χρήστη.....	66
Σχεδιασμός και Στυλ.....	66
6.3 Ανάλυση και Διαχείριση Δεδομένων.....	68
1. Συλλογή Δεδομένων.....	68
6.4 Επικοινωνία Πελάτη-Διακομιστή.....	69
6.5 Χειρισμός Σφαλμάτων και Αντιμετώπιση Εξαιρέσεων.....	70
6.6 Στρατηγική Σχεδίασης και Αντίκτυπος.....	70
6.7 Συμπεράσματα.....	70
6.8 Σύνθεση και Λειτουργία της Εφαρμογής Ιστού.....	70

## Κεφάλαιο 1 – Καρκίνος του Μαστού

### 1.1 Εισαγωγή

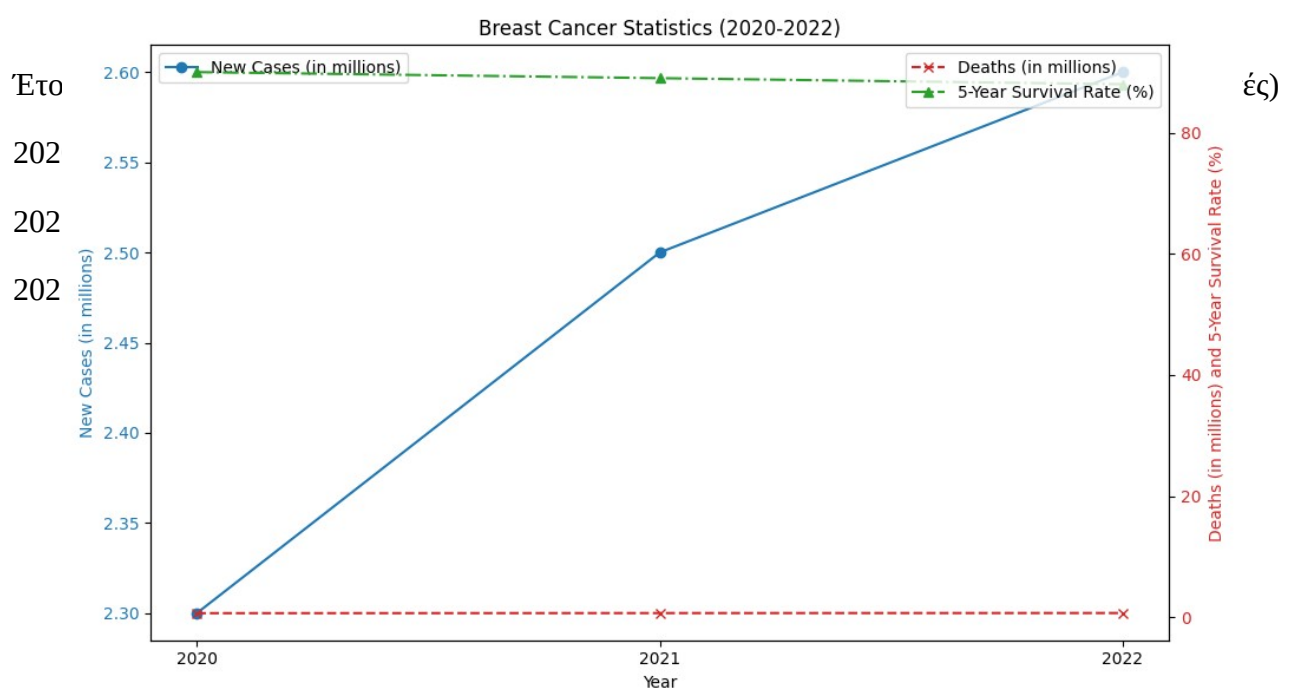
Η δεύτερη πιο συνηθισμένη αιτία θανάτου από καρκίνο στις γυναίκες είναι ο καρκίνος του μαστού. Η ανάπτυξη του καρκίνου μαστού αποτελεί μια διαδικασία με πολλαπλά στάδια που περιλαμβάνει πολλούς τύπους κυττάρων. Η καλύτερη προσέγγιση του παραπάνω προβλήματος αποτελεί η έγκαιρη διάγνωση μιας και εάν ο καρκινός εντοπισθεί σε αρχικό στάδιο οι πιθανότητες ίασης του ασθενούς είναι μεγαλύτερες. Σε ορισμένες χώρες το ποσοστό 5ετούς σχετικής επιβίωσης των ασθενών με καρκίνο του μαστού είναι μεγαλύτερο από 80% λόγω της έγκαιρης πρόληψης. Υπάρχουν 3 τύποι καρκίνου του μαστού: ο Ductal Carcinoma in Situ (DCIS) ένας μη διηθητικός καρκίνος όπου τα καρκινικά κελύφη περιορίζονται στους πόρους του μαστού, ο Invasive Ductal Carcinoma (IDC) ο πιο κοινός τύπος όπου τα καρκινικά κύτταρα εξαπλώνονται πέρα από τους πόρους, ο Invasive Lobular Carcinoma ξεκινά από τους λοβούς (αδένες που παράγουν γάλα) και μπορεί να εξαπλωθεί σε άλλα μέρη του σώματος, ο Triple-Negative Breast Cancer όπου τα κύτταρα του καρκίνου δεν έχουν υποδοχείς για τις ορμόνες οιστρογόνο και προγεστερόνη, μια πρωτεΐνη που ονομάζεται υποδοχέας ανθρώπινου επιδερμικού αυξητικού παράγοντα HER2, γεγονός που καθιστά δυσκολότερη την θεραπεία και τέλος ο HER2-Positive Breast Cancer διακρίνεται από υψηλά επίπεδα πρωτεΐνης HER2 και τείνει να αναπτύσσεται πιο επιθετικά σε σχέση με τους υπόλοιπους τύπους. Οι παράγοντες που έχει αποδειχθεί πως αυξάνουν τον κίνδυνο εμφάνισης του καρκίνου του μαστού είναι το φύλο (με μεγαλύτερο ποσοστό εμφάνισης στις γυναίκες), η ηλικία με διάμεση ηλικία τη στιγμή της διάγνωσης του μαστού τα 62 έτη, αυτό σημαίνει πως οι μισές γυναίκες που θα διαγνωστούν είναι μικρότερες από τα 62 έτη όταν διαγνωστούν, γενετικές μεταλλάξεις όπως οι γονδιακές μεταλλάξεις BRCA1 και BRCA2 που αυξάνουν σημαντικά τον κίνδυνο, ορμονικοί παράγοντες όπως η παρατεταμένη έκθεση σε οιστρογόνα και προγεστερόνη, όπως συμβαίνει κατά την πρώιμο έμμηνο ρύση ή η καθυστερημένη εμμηνόπαυση και τέλος παράγοντες του τρόπου ζωής όπως η παχυσαρκία, η κατανάλωση αλκοόλ και η έλλειψη σωματικής δραστηριότητας μπορούν να αυξήσουν τον κίνδυνο εμφάνισης του καρκίνου του μαστού. Στην αρχή της νόσου μπορεί να μην εμφανίζονται συμπτώματα όπως προοδεύει η νόσος εντοπίζουμε συμπτώματα όπως ένα εξόγκωμα ή μάζα στο στήθος ή στη μασχάλη, αλλαγή στο μέγεθος, το σχήμα ή την εμφάνιση του μαστού. Επιπλέον παρουσιάζονται δερματικές αλλαγές στο στήθος όπως λακκάκια ή συρρίκνωση, έκκριση από τη θήλη ειδικά αν είναι αιματηρή, τέλος πόνος ή δυσφορία στο στήθος. Η διάγνωση του καρκίνου του μαστού συνήθως περιλαμβάνει έναν συνδυασμό από τα παρακάτω. Αρχικά γίνεται η φυσική εξέταση όπου ο ασθενής τοποθετεί τα χέρια στους μηρούς του ή σηκώνει τα χέρια πάνω από το κεφάλι του αυτό επιτρέπει στο ιατρικό προσωπικό να αξιολογήσει το στήθος σε πολλές θέσεις ώστε να παρατηρηθεί το συνολικό μέγεθος, το σχήμα, η συμμετρία, το μέγεθος της θήλης, την υφή και το χρώμα. Επιπλέον το ιατρικό προσωπικό λαμβάνει υπόψιν τεστ απεικόνισης όπως η μαστογραφία, το υπερηχογράφημα και η μαγνητική τομογραφία σε συνδυασμό με τα αποτελέσματα της βιοψίας όπου αφαιρείται ένα δείγμα κυττάρων του μαστού για εργαστηριακό έλεγχο. Χρησιμοποιώντας τα παραπάνω συνδυαστικά μπορούμε να έχουμε μια διάγνωση σχετικά με τον καρκίνο του μαστού. Επιπρόσθετα ο καρκίνος του μαστού διακρίνεται σε στάδια (από 0 έως 4) και προσδιορίζονται με βάση το μέγεθος του όγκου, τη συμμετοχή των λεμφαδένων και τη μετάσταση (δηλαδή εάν ο



καρκίνος έχει εξαπλωθεί σε άλλα μέρη του σώματος). Όσον αφορά την θεραπεία του καρκίνου το μαστού, η θεραπεία διαφέρει ανάλογα με τον τύπο και το στάδιο του καρκίνου που εντοπίζεται ο καρκίνος. Ενδεικτικές θεραπείες αποτελούν η χειρουργική επέμβαση όπου γίνεται ογκεκτομή (δηλαδή αφαιρείται ο όγκος) ή μαστεκτομή (αφαίρεση μαστού), η ακτινοθεραπεία όπου γίνεται χρήση κυμάτων υψηλής ενέργειας για τη θανάτωση ή την επιβράδυνση της ανάπτυξης καρκινικών κυττάρων, η χημειοθεραπεία όπου γίνεται χρήση φαρμάκων για τη θανάτωση ή την επιβράδυνση της ανάπτυξης των καρκινικών κυττάρων, η ορμονική θεραπεία όπου αναστέλλονται συγκεκριμένες ορμόνες που τροφοδοτούν ορισμένους καρκίνους, η στοχευμένη θεραπεία όπου χρησιμοποιούνται φάρμακα που στοχεύουν ειδικά στα καρκινικά κύτταρα, όπως οι αναστολείς HER2 και τέλος η ανοσοθεραπεία όπου πραγματοποιείται ενίσχυση του ανοσοποιητικού συστήματος για την καταπολέμηση του καρκίνου. Αναφορικά με την πρόληψη του καρκίνου το μαστού υπάρχουν μερικοί παράγοντες που δεν μπορούν να ελεγχθούν όπως η γενετική. Όμως υπάρχουν μερικά προληπτικά μέτρα που μπορούν να ληφθούν για την αντιμετώπιση του προβλήματος όπως είναι ο τακτικός προληπτικός έλεγχος (μαστρογραφίες και αυτοεξετάσεις) καθώς και η διατήρηση ενός υγιούς τρόπου ζωής (διατήρηση υγιούς βάρους, άσκηση, αποφυγή αλκοόλ και καπνίσματος). Η πρόγνωση του καρκίνου του μαστού ποικίλλει ανάλογα με το στάδιο που πραγματοποιείται η διάγνωση του ασθενός ανάλογα με τα χαρακτηριστικά του όγκου και την ανταπόκριση στη θεραπεία. Η έγκαιρη διάγνωση και η πρόοδος στις θεραπευτικές μεθόδους έχουν βελτίωση σημαντικά τα ποσοστά επιβίωσης έτσι η δημιουργία ενός συστήματος υποστήριξης απόφασης με τεχνικές μηχανικής μάθησης και τεχνητής νοημοσύνης θα βοηθούσε σημαντικά στην επίλυση αυτού του προβλήματος αφού θα βοηθούσε τον ειδικό να πάρει μία απόφαση ή έστω μια προσέγγιση για την κατάσταση του ασθενούς με εξαιρετικά γρήγορη ταχύτητα χρησιμοποιώντας το σύστημα αυτό.

Ακολουθεί μια αναλυτική παρουσίαση με πίνακες και αριθμητικά δεδομένα σχετικά με τον καρκίνο του μαστού, τις επιπτώσεις του στο σύστημα υγείας και τα κόστη που συνδέονται με την ασθένεια.

## 1.2 Παγκόσμια Στατιστικά Στοιχεία για τον Καρκίνο του Μαστού



**Σημείωση:** Το ποσοστό επιβίωσης αναφέρεται για τις περιπτώσεις που διαγιγνώσκονται σε πρώιμο στάδιο.

Από τα δεδομένα για τον καρκίνο του μαστού που καταγράφηκαν μεταξύ 2020 και 2022, παρατηρούμε μια ανησυχητική αύξηση τόσο στον αριθμό των νέων κρουσμάτων όσο και στον αριθμό των θανάτων. Ειδικότερα, τα νέα κρούσματα αυξήθηκαν από 2.3 εκατομμύρια το 2020 σε 2.6 εκατομμύρια το 2022, ενώ οι θάνατοι ανήλθαν από 0.685 εκατομμύρια σε 0.710 εκατομμύρια κατά την ίδια περίοδο. Ταυτόχρονα, το ποσοστό επιβίωσης πενταετίας μειώθηκε από το 90% το 2020 στο 88% το 2022, υποδεικνύοντας μια ελαφρά αλλά σταθερή πτώση στις πιθανότητες επιβίωσης των ασθενών. Αυτή η τάση μπορεί να αντανακλά την αύξηση της διάγνωσης ή την επιδείνωση της ασθένειας, υποδεικνύοντας την ανάγκη για ενισχυμένες στρατηγικές πρόληψης, πρώιμης διάγνωσης και αποτελεσματικής θεραπείας για τη βελτίωση των εκβάσεων και τη μείωση της θνησιμότητας.

#### Στατιστικά Στοιχεία στις ΗΠΑ

Έτος	Νέα Κρούσματα (σε χιλιάδες)	Θάνατοι (σε χιλιάδες)	5ετές Ποσοστό Επιβίωσης
2024	287.85	43.50	90%

#### Πηγή: American Cancer Society (ACS)

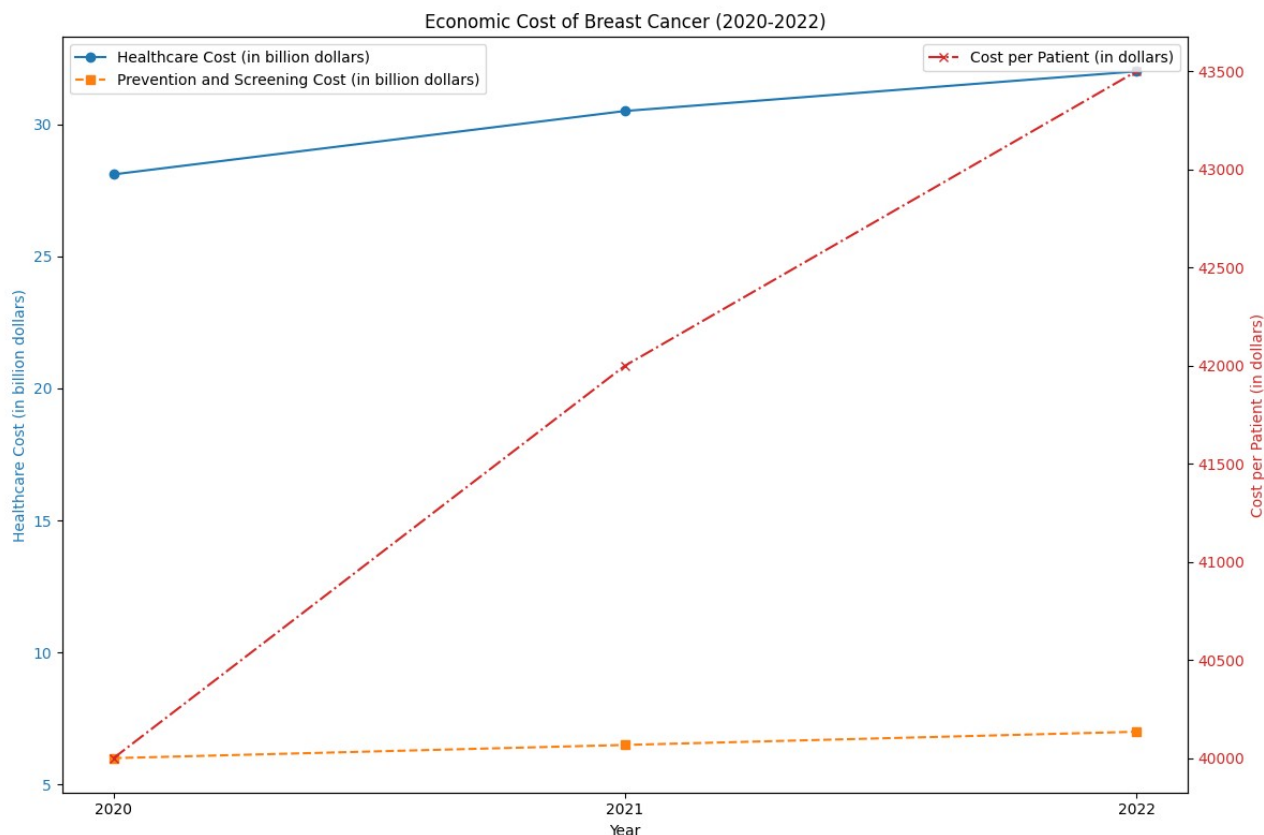
Για το έτος 2024, στις ΗΠΑ καταγράφηκαν περίπου 287,850 νέα κρούσματα καρκίνου του μαστού και 43,500 θάνατοι. Το 5ετές ποσοστό επιβίωσης για τους ασθενείς παραμένει στο 90%, υποδεικνύοντας υψηλή απόδοση των διαθέσιμων θεραπευτικών προσεγγίσεων και βελτιωμένη ικανότητα των συστημάτων υγείας στην αντιμετώπιση της ασθένειας. Αν και το ποσοστό επιβίωσης παραμένει ισχυρό, ο αριθμός των νέων κρουσμάτων και θανάτων καταδεικνύει την συνεχιζόμενη ανάγκη για ενίσχυση των προσπαθειών στην πρόληψη και τη θεραπεία του καρκίνου του μαστού.

#### Οικονομικό Κόστος του Καρκίνου του Μαστού

##### Πίνακας 3: Εκτίμηση Κόστους για το Σύστημα Υγείας

Έτος	Κόστος Υγειονομικής Περίθαλψης (σε δισεκατομμύρια δολάρια)	Κόστος ανά Ασθενή (σε δολάρια)	Κόστος Πρόληψης και Εξέτασης (σε δισεκατομμύρια δολάρια)
2020	28.1	40,000	6.0
2021	30.5	42,000	6.5
2022	32.0	43,500	7.0

**Σημείωση:** Το συνολικό κόστος περιλαμβάνει ιατρικές υπηρεσίες, φαρμακευτικά έξοδα, θεραπεία και αποκατάσταση.

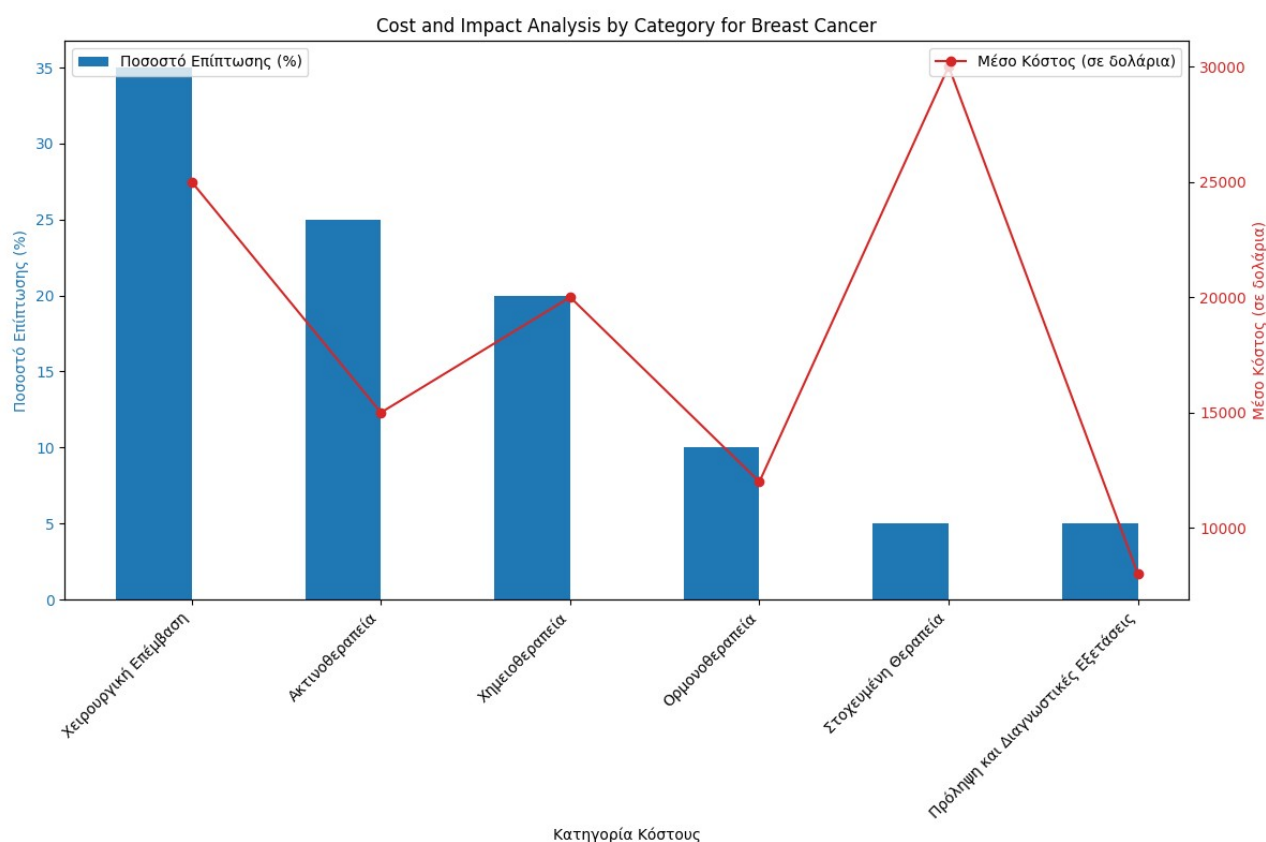


Από το 2020 έως το 2022, το οικονομικό κόστος για τη διαχείριση του καρκίνου του μαστού αυξήθηκε σταδιακά. Το κόστος υγειονομικής περίθαλψης ανέβηκε από 28.1 δισεκατομμύρια δολάρια το 2020 σε 32.0 δισεκατομμύρια δολάρια το 2022, υποδεικνύοντας μια συνεχή αύξηση στις δαπάνες για τη θεραπεία και την αποκατάσταση. Το μέσο κόστος ανά ασθενή επίσης αυξήθηκε, από 40,000 δολάρια το 2020 σε 43,500 δολάρια το 2022. Παράλληλα, το κόστος πρόληψης και εξέτασης αυξήθηκε από 6.0 δισεκατομμύρια δολάρια το 2020 σε 7.0 δισεκατομμύρια δολάρια το 2022. Αυτές οι αυξήσεις δείχνουν την ανάγκη για συνεχιζόμενη επένδυση σε προγράμματα πρόληψης και βελτιωμένες θεραπευτικές προσεγγίσεις, καθώς και την αύξηση των οικονομικών επιπτώσεων που συνδέονται με τη διαχείριση του καρκίνου του μαστού.

### Κόστος και Αντίκτυποι σε Επίπεδο Υγειονομικής Περίθαλψης

#### Πίνακας 4: Ανάλυση Κόστους Κατηγορίας

Κατηγορία Κόστους	Ποσοστό Επίπτωσης (%)	Μέσο Κόστος (σε δολάρια)
Χειρουργική Επέμβαση	35%	25,000
Ακτινοθεραπεία	25%	15,000
Χημειοθεραπεία	20%	20,000
Ορμονοθεραπεία	10%	12,000
Στοχευμένη Θεραπεία	5%	30,000
Πρόληψη και Διαγνωστικές Εξετάσεις	5%	8,000



Η ανάλυση κόστους κατηγορίας για τον καρκίνο του μαστού δείχνει ότι η χειρουργική επέμβαση είναι η κατηγορία με το υψηλότερο ποσοστό επίπτωσης, φτάνοντας το 35% του συνολικού κόστους. Το μέσο κόστος για τη χειρουργική επέμβαση ανέρχεται σε 25,000 δολάρια. Η ακτινοθεραπεία ακολουθεί με ποσοστό επίπτωσης 25% και μέσο κόστος 15,000 δολάρια. Η χημειοθεραπεία έχει ποσοστό 20% και μέσο κόστος 20,000 δολάρια, ενώ η ορμονοθεραπεία καλύπτει το 10% της επίπτωσης με μέσο κόστος 12,000 δολάρια. Η στοχευμένη θεραπεία, αν και λιγότερο συχνά χρησιμοποιούμενη, έχει το υψηλότερο μέσο κόστος, 30,000 δολάρια, με ποσοστό επίπτωσης 5%. Η πρόληψη και οι διαγνωστικές εξετάσεις έχουν το μικρότερο ποσοστό επίπτωσης (5%) και μέσο κόστος 8,000 δολάρια. Αυτή η κατανομή των εξόδων καταδεικνύει τη σημασία της

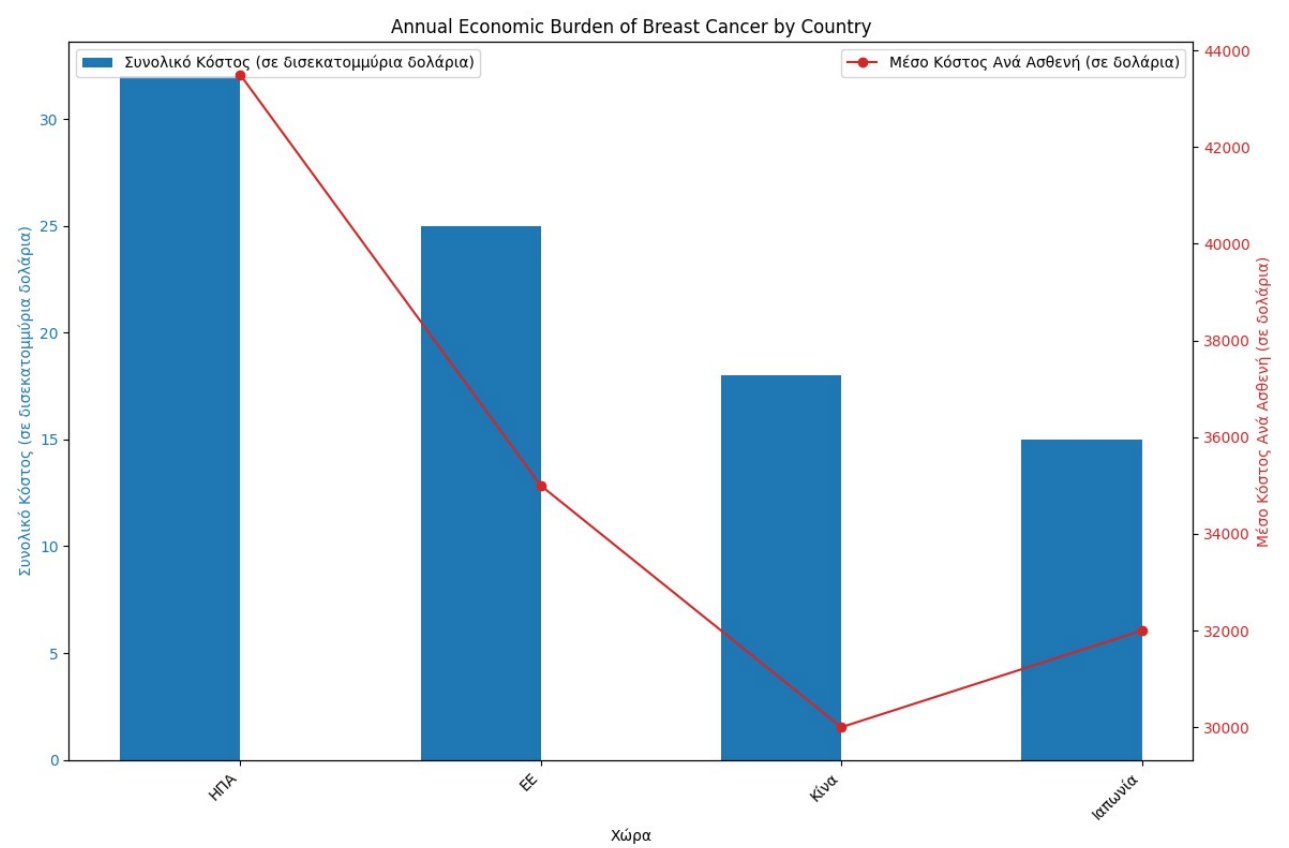
χειρουργικής επέμβασης και της ακτινοθεραπείας στο συνολικό κόστος της θεραπείας του καρκίνου του μαστού, ενώ αναδεικνύει την οικονομική επιβάρυνση που συνδέεται με τις άλλες θεραπείες.

Ανάλυση Οικονομικού Κόστους για το Σύστημα Υγείας

Πίνακας 5: Ετήσια Οικονομική Επιβάρυνση

Χώρα	Συνολικό Κόστος (σε δισεκατομμύρια δολάρια)	Μέσο Κόστος Ανά Ασθενή (σε δολάρια)
ΗΠΑ	32.0	43,500
ΕΕ	25.0	35,000
Κίνα	18.0	30,000
Ιαπωνία	15.0	32,000

Πηγή: Διεθνής Οργάνωση Υγείας (WHO), Εθνικά Ινστιτούτα Υγείας (NIH)



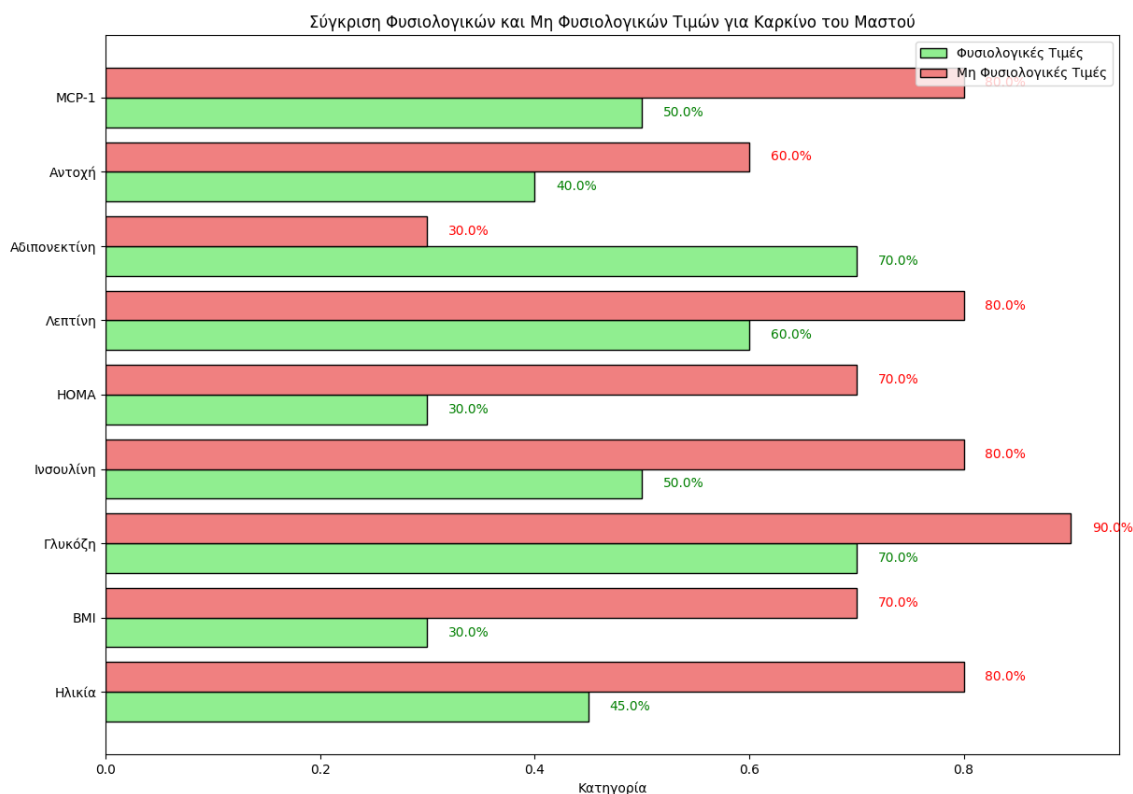
Για το 2024, οι οικονομικές επιβαρύνσεις του καρκίνου του μαστού διαφέρουν σημαντικά μεταξύ χωρών. Στις ΗΠΑ, το συνολικό κόστος φτάνει τα 32 δισεκατομμύρια δολάρια με μέσο κόστος ανά ασθενή 43,500 δολάρια, το υψηλότερο από τις χώρες που εξετάζονται. Στην ΕΕ, το συνολικό κόστος ανέρχεται σε 25 δισεκατομμύρια δολάρια και το μέσο κόστος ανά ασθενή είναι 35,000

δολάρια. Στην Κίνα, το συνολικό κόστος είναι 18 δισεκατομμύρια δολάρια με μέσο κόστος 30,000 δολάρια ανά ασθενή. Στην Ιαπωνία, το συνολικό κόστος φτάνει τα 15 δισεκατομμύρια δολάρια, ενώ το μέσο κόστος ανά ασθενή είναι 32,000 δολάρια. Αυτά τα στοιχεία αναδεικνύουν τις διαφορές στις οικονομικές επιβαρύνσεις για τη θεραπεία του καρκίνου του μαστού, με τις ΗΠΑ να παρουσιάζουν το υψηλότερο κόστος συνολικά και ανά ασθενή, ενώ οι άλλες χώρες έχουν χαμηλότερα αλλά σημαντικά κόστη.

Παράγοντες Κινδύνου και Συσχέτιση με Υγειονομικές Μεταβλητές

Πίνακας 6: Συσχέτιση Κινδύνου με Βιολογικές Μεταβλητές

Μεταβλητή	Σχέση με τον Κίνδυνο Καρκίνου του Μαστού	Μέσο Εύρος Τιμών για Υψηλό Κίνδυνο
Ηλικία	Αυξημένος κίνδυνος με την ηλικία	> 50 έτη
BMI	Υψηλός BMI συνδέεται με αυξημένο κίνδυνο	> 30 kg/m <sup>2</sup>
Γλυκόζη	Υψηλά επίπεδα συνδέονται με αυξημένο κίνδυνο	> 100 mg/dL
Ινσουλίνη	Υψηλά επίπεδα συνδέονται με αυξημένο κίνδυνο	> 25 μU/mL
HOMA	Υψηλές τιμές υποδεικνύουν αντίσταση στην ινσουλίνη	> 2.5
Λεπτίνη	Υψηλά επίπεδα συνδέονται με αυξημένο κίνδυνο	> 20 ng/mL
Αδιπονεκτίνη	Χαμηλά επίπεδα συνδέονται με αυξημένο κίνδυνο	< 5 μg/mL
Αντοχή	Υψηλά επίπεδα συνδέονται με αυξημένο κίνδυνο	> 10 ng/mL
MCP-1	Υψηλά επίπεδα συνδέονται με αυξημένο κίνδυνο	> 150 pg/mL



Οι βιολογικές μεταβλητές που σχετίζονται με τον κίνδυνο για καρκίνο του μαστού συγκρίνονται με τις μέσες τιμές που παρατηρούνται σε υγιείς ανθρώπους. Η σύγκριση αυτή βοηθάει στην κατανόηση του πόσο εκτός των φυσιολογικών ορίων μπορούν να βρίσκονται οι τιμές για να ενταχθούν σε κατηγορία υψηλού κινδύνου:

- **Ηλικία:** Οι υγιείς άνθρωποι συνήθως διαγιγνώσκονται με καρκίνο του μαστού σε μικρότερες ηλικίες συγκριτικά με άτομα άνω των 50 ετών, που βρίσκονται σε υψηλότερο κίνδυνο.
- **BMI:** Η μέση τιμή BMI σε υγιείς ανθρώπους είναι συνήθως κάτω από 30 kg/m<sup>2</sup>. Τα άτομα με BMI πάνω από 30 kg/m<sup>2</sup> βρίσκονται σε υψηλότερο κίνδυνο.
- **Γλυκόζη:** Τα υγιή επίπεδα γλυκόζης είναι συνήθως κάτω από 100 mg/dL, ενώ τα επίπεδα άνω των 100 mg/dL συνδέονται με αυξημένο κίνδυνο.
- **Ινσουλίνη:** Η μέση τιμή ινσουλίνης σε υγιείς ανθρώπους είναι συνήθως κάτω από 25 μU/mL. Τα επίπεδα πάνω από 25 μU/mL είναι συνδεδεμένα με αυξημένο κίνδυνο.
- **HOMA:** Σε υγιείς ανθρώπους, οι τιμές HOMA είναι συνήθως κάτω από 2.5. Υψηλότερες τιμές υποδεικνύουν αυξημένη αντίσταση στην ινσουλίνη και κίνδυνο.
- **Λεπτίνη:** Η μέση τιμή λεπτίνης σε υγιείς ανθρώπους είναι κάτω από 20 ng/mL. Υψηλότερες τιμές σχετίζονται με αυξημένο κίνδυνο.
- **Αδιπονεκτίνη:** Στους υγιείς ανθρώπους, τα επίπεδα αδιπονεκτίνης είναι συνήθως πάνω από 5 μg/mL. Χαμηλότερα επίπεδα είναι συνδεδεμένα με αυξημένο κίνδυνο.

- **Αντοχή:** Τα υγιή επίπεδα αντοχής είναι συνήθως κάτω από 10 ng/mL. Υψηλότερα επίπεδα συνδέονται με αυξημένο κίνδυνο.
- **MCP-1:** Τα επίπεδα MCP-1 σε υγιείς ανθρώπους είναι συνήθως κάτω από 150 pg/mL. Υψηλότερα επίπεδα σχετίζονται με αυξημένο κίνδυνο.

## Συμπεράσματα

Ο καρκίνος του μαστού είναι μια παγκόσμια υγειονομική πρόκληση με σημαντικές επιπτώσεις τόσο στην υγεία όσο και στην οικονομία. Από τα δεδομένα για την περίοδο 2020-2022, παρατηρείται μια ανησυχητική αύξηση τόσο στον αριθμό των νέων κρουσμάτων όσο και στον αριθμό των θανάτων. Ειδικότερα, τα νέα κρούσματα αυξήθηκαν από 2.3 εκατομμύρια το 2020 σε 2.6 εκατομμύρια το 2022, ενώ οι θάνατοι ανέβηκαν από 0.685 εκατομμύρια σε 0.710 εκατομμύρια κατά την ίδια περίοδο. Αυτή η τάση ενδέχεται να υποδηλώνει είτε μια αύξηση στην ανίχνευση της ασθένειας είτε μια επιδείνωση της κατάστασης, τονίζοντας την ανάγκη για καλύτερες στρατηγικές πρόληψης και πρώιμης διάγνωσης. Η μείωση του πενταετούς ποσοστού επιβίωσης από 90% σε 88% επισημαίνει την ανάγκη για ενίσχυση των θεραπευτικών μεθόδων και τη συνεχιζόμενη προσπάθεια για βελτίωση των εκβάσεων.

Στις ΗΠΑ, για το έτος 2024, καταγράφηκαν 287,850 νέα κρούσματα και 43,500 θάνατοι από καρκίνο του μαστού, με το 5ετές ποσοστό επιβίωσης να παραμένει στο 90%. Παρά την υψηλή επιβίωση, οι αυξημένοι αριθμοί νέων κρουσμάτων και θανάτων επισημαίνουν την ανάγκη για συνεχιζόμενη ενίσχυση της πρόληψης και της θεραπείας. Το υψηλό ποσοστό επιβίωσης υποδεικνύει την αποτελεσματικότητα των διαθέσιμων θεραπευτικών προσεγγίσεων, αλλά η αύξηση των κρουσμάτων τονίζει την επείγουσα ανάγκη για πιο ολοκληρωμένες στρατηγικές για την καταπολέμηση της ασθένειας.

Αναφορικά με το οικονομικό κόστος, η ανάλυση των δαπανών δείχνει μια συνεχή αύξηση στις συνολικές δαπάνες για τη διαχείριση του καρκίνου του μαστού, από 28.1 δισεκατομμύρια δολάρια το 2020 σε 32.0 δισεκατομμύρια δολάρια το 2022. Το μέσο κόστος ανά ασθενή αυξήθηκε επίσης από 40,000 δολάρια σε 43,500 δολάρια, ενώ το κόστος πρόληψης και εξέτασης ανήλθε σε 7.0 δισεκατομμύρια δολάρια το 2022. Αυτή η αύξηση καταδεικνύει την αναγκαία επένδυση σε προγράμματα πρόληψης και τις οικονομικές επιπτώσεις της θεραπείας, υποδεικνύοντας την ανάγκη για καινοτόμες προσεγγίσεις που θα μπορούσαν να μειώσουν το συνολικό κόστος.

Η ανάλυση των κόστους ανά κατηγορία δείχνει ότι η χειρουργική επέμβαση αντιπροσωπεύει το μεγαλύτερο ποσοστό του συνολικού κόστους (35%) με μέσο κόστος 25,000 δολάρια. Η ακτινοθεραπεία και η χημειοθεραπεία ακολουθούν, με ποσοστά επίπτωσης 25% και 20% αντίστοιχα. Η υψηλότερη δαπάνη σχετίζεται με τη στοχευμένη θεραπεία, παρά την χαμηλότερη συχνότητα χρήσης της. Αυτή η κατανομή καταδεικνύει τη σημαντική οικονομική επιβάρυνση που συνδέεται με τις κλασικές θεραπείες όπως η χειρουργική επέμβαση και η ακτινοθεραπεία, ενώ ταυτόχρονα υπογραμμίζει την ανάγκη για καλύτερη διαχείριση και αποτελεσματικότητα στις πιο ακριβές θεραπείες.



Σύμφωνα με τα δεδομένα για την ετήσια οικονομική επιβάρυνση του καρκίνου του μαστού, οι διαφορές μεταξύ των χωρών είναι αξιοσημείωτες. Στις ΗΠΑ, το συνολικό κόστος φτάνει τα 32 δισεκατομμύρια δολάρια, το υψηλότερο μεταξύ των χωρών που εξετάστηκαν, με μέσο κόστος ανά ασθενή 43,500 δολάρια. Στην ΕΕ, το συνολικό κόστος είναι 25 δισεκατομμύρια δολάρια, με μέσο κόστος 35,000 δολάρια, ενώ στην Κίνα και την Ιαπωνία τα κόστη είναι χαμηλότερα, αλλά οι επιπτώσεις παραμένουν σημαντικές. Αυτά τα στοιχεία αναδεικνύουν την ανάγκη για βελτίωση της οικονομικής διαχείρισης του καρκίνου του μαστού και τη σημασία της διαφοροποίησης των στρατηγικών αντιμετώπισης ανάλογα με τις οικονομικές συνθήκες κάθε χώρας.

Συνολικά, η συνεχής αύξηση των κρουσμάτων και του κόστους του καρκίνου του μαστού δείχνει την ανάγκη για ενίσχυση των στρατηγικών πρόληψης και θεραπείας, και την ανάπτυξη καινοτόμων λύσεων για τη βελτίωση των εκβάσεων και την ελαχιστοποίηση των οικονομικών επιπτώσεων.

### **1.3 Ιστορική αναδρομή στην πρόβλεψη του καρκίνου το μαστού με τεχνικές μηχανικής μάθησης:**

Η πρόγνωση του καρκίνου του μαστού με τη χρήση τεχνικών μηχανικής μάθησης (machine learning) έχει εξελιχθεί σημαντικά τις τελευταίες δεκαετίες. Ακολουθεί μια ιστορική αναδρομή για το πώς αυτές οι τεχνικές έχουν ενσωματωθεί στην πρόγνωση του καρκίνου του μαστού:

Την δεκαετία του 1980 με 1990 ξεκίνησαν οι πρώτες προσπάθειες στην πρόγνωση του καρκίνου το μαστού χρησιμοποιώντας στατιστικά μοντέλα όπως η λογιστική παλινδρόμηση και τα Cox proportional hazards models, τα οποία πρόβλεπταν την επιβίωση και την επανεμφάνιση της ασθένειας. Τα μοντέλα αυτά βασιζόντουσαν κυρίως σε κλινικά δεδομένα όπως το στάδιο του όγκου, η ηλικία του ασθενός και τα επίπεδα των ορμονών.

Το 2000 έγινε η έναρξη της χρήσης της μηχανικής μάθησης στα μοντέλα και η δημιουργία συστημάτων υποστήριξης αποφάσεων καθώς επίσης και η εμφάνιση των νευρωνικών δικτύων διαδραμάτισε σημαντικό ρόλο στην ανάπτυξη των συστημάτων αυτών. Συγκεκριμένα τα νευρωνικά δίκτυα είχαν την δυνατότητα να μάθουν απο πολύπλοκα πρότυπα δεδομένα προσφέροντας έτσι βελτιστοποιημένες προβλέψεις σε σύγκριση με τα ήδη υπάρχον στατιστικά μοντέλα. Επιπλέον δημιουργήθηκαν συστήματα υποστήριξης απόφασης τα οποία συνδύαζαν τα κλινικά δεδομένα και τα αποτελέσματα των νευρωνικών δικτύων για να βοηθήσουν τους ειδικούς ώστε να λάβουν πιο ενημερωμένες αποφάσεις σχετικά με την θεραπεία των ασθενών.

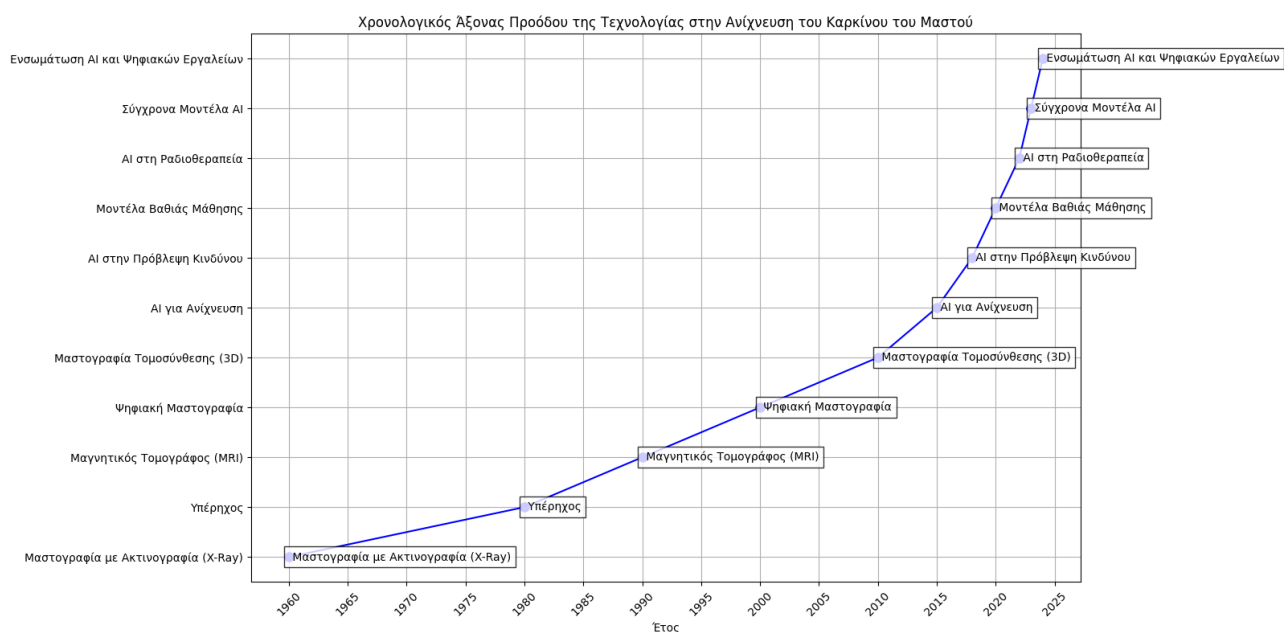
Το 2010 ξεκίνησαν οι προηγμένες τεχνικές στην μηχανική μάθηση με αλγορίθμους όπως τον Random Forests και SVMs που αποτελούν εξελιγμένους αλγορίθμους και μηχανήματα διανυσμάτων υποστήριξης ενίσχυσε τις ικανότητες πρόγνωσης, λαμβάνοντας υπόψη περισσότερες μεταβλητές και συνδυασμούς δεδομένων. Επιπλέον με την αύξηση των τεχνολογικών μέσων και της καταλογράφησης των ιατρικών δεδομένων σε βάσεις δεδομένων αυξήθηκε ο όγκος και η ποικιλομορφία των βάσεων δεδομένων γεγονός που βοήθησε στην εκπαίδευση μοντέλων που πραγματοποιούσαν προβλέψεις με μεγαλύτερη ακρίβεια.

Το 2020 παρατηρούμε συνδιασμό της τεχνητής νοημοσύνης με την Βαθιά Μάθηση (Deep Learning) όπου τα βαθιά νευρωνικά δίκτυα (deep neural networks) και τα δίκτυα convolutional να έχουν αποδειχθεί άκρως αποτελεσματικά στην ανάλυση των εικόνων της μαστογραφίας και άλλων

απεικονιστικών μεθόδων, βελτιώνοντας την ακρίβεια στην ανίχνευση της ασθένειας καθώς και την πρόγνωση της. Επιπλέον η εξατομικευμένη ιατρική μέσω της ανάλυσης μεγάλου όγκου δεδομένων (big data) και η ενσωμάτωση των γενετικών πληροφοριών έχουν επιτρέψει την ανάπτυξη μοντέλων που προσφέρουν εξατομικευμένη πρόβλεψη ανάλογα με το προφίλ του κάθε ασθενούς. Τέλος οι σύγχρονες τεχνικές επιτρέπουν την ανάλυση πολυδιάστατων δεδομένων, αφού λαμβάνουν υπόψη κλινικές, γενετικές και απεικονιστικές πληροφορίες προσφέροντας έτσι μια πιο ολοκληρωμένη πρόγνωση.

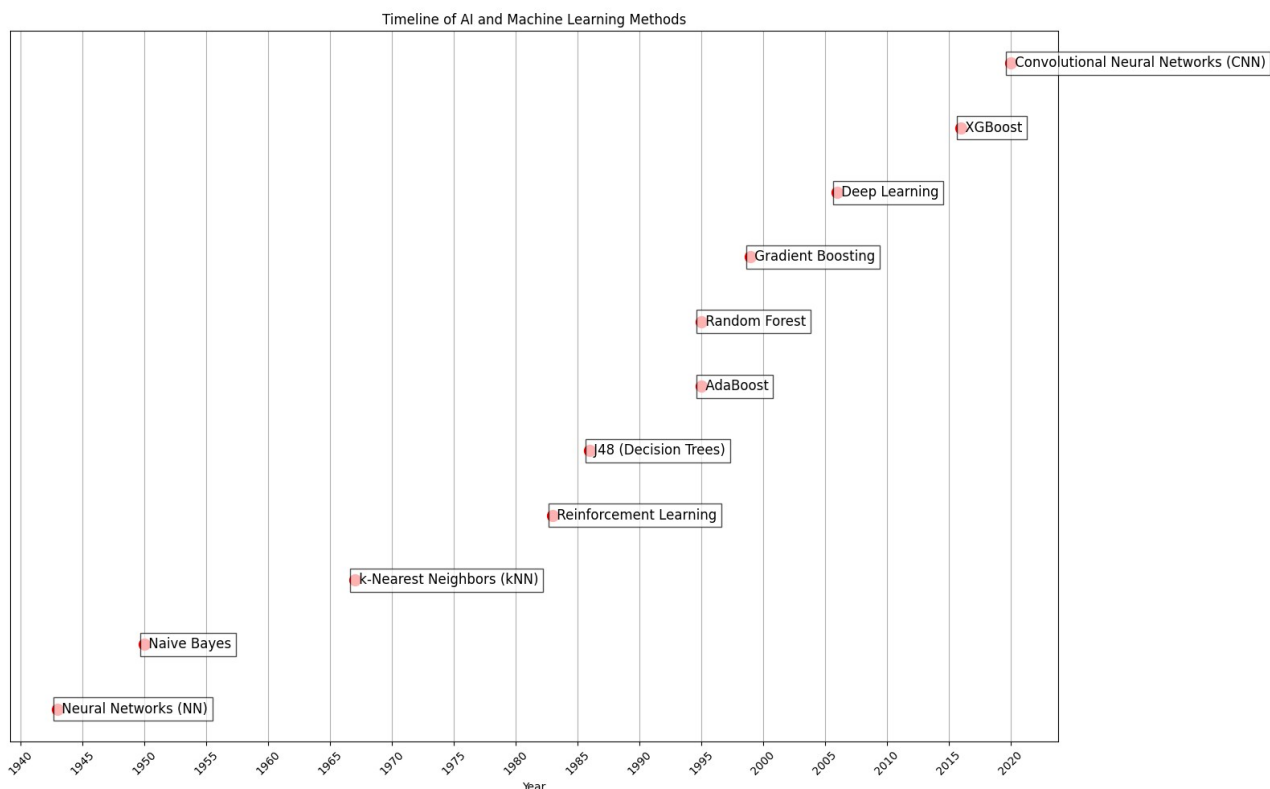
Συμπερασματικά η εξέλιξη που παρατηρήθηκε στην μηχανική μάθηση τα τελευταία χρόνια έχει φέρει την επανάσταση στην πρόγνωση του καρκίνου του μαστού, βελτιώνοντας την ακρίβεια και την αποτελεσματικότητα των προβλέψεων των μοντέλων. Από τα αρχικά στατιστικά μοντέλα μέχρι και τις σύγχρονες τεχνικές βαθιάς μάθησης, οι δυνατότητες για πιο εξατομικευμένη και στοχευμένη θεραπεία συνεχίζουν να αυξάνονται, προσφέροντας έτσι καλύτερες προοπτικές για τους ασθενείς.

### Χρονολόγιο Καρκίνου του Μαστού και εξελίξεις πάνω στις προγνωστικές μεθόδους σε σχέση με το χρόνο.



- **1960: Μαστογραφία με Ακτινογραφία (X-Ray):** Η παραδοσιακή μαστογραφία με φιλμ ήταν το κύριο εργαλείο για τη διάγνωση του καρκίνου του μαστού. Οι ακτινογραφίες φιλμ ήταν το πρότυπο για δεκαετίες, παρόλο που είχαν περιορισμούς στη λεπτομέρεια εικόνας και στην ανίχνευση μικρών ή καμουφλαρισμένων όγκων.
- **1980: Υπέρηχος:** Ο υπέρηχος εισήχθη ως συμπληρωματική μέθοδος για την αξιολόγηση πυκνών ιστών του μαστού, όπου οι ακτινογραφίες είχαν περιορισμένη αποτελεσματικότητα. Οι υπέρηχοι προσφέρουν μη επεμβατικές εικόνες και είναι χρήσιμοι για τη διαφοροποίηση μεταξύ κυστικών και στερεών όγκων.

- **1990: Μαγνητικός Τομογράφος (MRI):** Η MRI άρχισε να χρησιμοποιείται για την πιο λεπτομερή απεικόνιση, ειδικά σε ασθενείς υψηλού κινδύνου ή σε περίπλοκες περιπτώσεις. Η MRI προσφέρει ανάλυση υψηλής ευκρίνειας και είναι χρήσιμη για την παρακολούθηση της εξέλιξης της ασθένειας και την αξιολόγηση της έκτασης του καρκίνου.
- **2000: Ψηφιακή Μαστογραφία:** Η ψηφιακή μαστογραφία αντικατέστησε το φιλμ με ψηφιακούς ανιχνευτές, βελτιώνοντας την ποιότητα της εικόνας και τις δυνατότητες επεξεργασίας. Αυτή η τεχνολογία επιτρέπει την καλύτερη αποθήκευση, ανάκτηση και μετάδοση των μαστογραφικών εικόνων, ενώ μειώνει την ανάγκη για νέες ακτινογραφίες.
- **2010: Μαστογραφία Τομοσύνθεσης (3D):** Η μαστογραφία τομοσύνθεσης παρέχει τρισδιάστατες εικόνες, οι οποίες αυξάνουν τα ποσοστά ανίχνευσης και μειώνουν τα ψευδή θετικά αποτελέσματα. Αυτή η μέθοδος βελτιώνει την ικανότητα ανίχνευσης μικρών όγκων και περιοχών που μπορεί να είναι δύσκολο να εντοπιστούν με δισδιάστατες εικόνες.
- **2015: ΑΙ για Ανίχνευση:** Η αρχική χρήση της τεχνητής νοημοσύνης (ΑΙ) για την ενίσχυση της μαστογραφικής διάγνωσης έδειξε υποσχόμενα αποτελέσματα. Οι πρώτες εφαρμογές ΑΙ άρχισαν να ενσωματώνονται για τη βελτίωση της ακρίβειας της διάγνωσης και την υποστήριξη των ακτινολόγων στην ανίχνευση όγκων.
- **2018: ΑΙ στην Πρόβλεψη Κινδύνου:** Μοντέλα ΑΙ σε συνδυασμό με παραδοσιακούς παράγοντες κινδύνου χρησιμοποιήθηκαν για την πρόβλεψη μελλοντικού κινδύνου καρκίνου του μαστού. Αυτή η τεχνολογία επιτρέπει την εξατομίκευση της παρακολούθησης και τη στόχευση των γυναικών με υψηλό κίνδυνο.
- **2020: Μοντέλα Βαθιάς Μάθησης (Deep Learning):** Η χρήση προηγμένων μοντέλων βαθιάς μάθησης (Deep Learning) βελτίωσε τη διάγνωση και ανάλυση των εικόνων μαστογραφίας. Αυτά τα μοντέλα επιτρέπουν την αυτοματοποιημένη ανάλυση και εντοπισμό ανωμαλιών με μεγαλύτερη ακρίβεια.
- **2022: ΑΙ στη Ραδιοθεραπεία:** Η ΑΙ άρχισε να χρησιμοποιείται για την πρόβλεψη της κατανομής δόσης στη ραδιοθεραπεία, βελτιώνοντας τον σχεδιασμό και την προσαρμογή της θεραπείας. Οι εφαρμογές ΑΙ βοηθούν στη βελτίωση της ακρίβειας και της αποτελεσματικότητας της θεραπείας.
- **2023: Σύγχρονα Μοντέλα ΑΙ:** Η συνεχιζόμενη ανάπτυξη μοντέλων ΑΙ προσφέρει καλύτερη διάγνωση και πρόβλεψη πρόγνωσης, με εστίαση στην ενσωμάτωσή τους σε κλινικές ροές εργασίας. Αυτά τα μοντέλα χρησιμοποιούν πρόσφατες έρευνες και δεδομένα για τη βελτίωση των διαγνωστικών και θεραπευτικών διαδικασιών.
- **2024: Ενσωμάτωση ΑΙ και Ψηφιακών Εργαλείων:** Το 2024 σηματοδοτεί την ολοκληρωμένη χρήση της τεχνητής νοημοσύνης σε απεικόνιση, διάγνωση, πρόγνωση και σχεδιασμό θεραπείας, περιλαμβάνοντας ψηφιακή μαστογραφία και εικονικές διαφάνειες. Η χρήση ΑΙ συνδυάζεται με προηγμένες ψηφιακές τεχνολογίες για μια ολοκληρωμένη προσέγγιση στην ανίχνευση και τη διαχείριση του καρκίνου του μαστού.



Το γράφημα παρουσιάζει την ανάπτυξη σημαντικών αλγορίθμων τεχνητής νοημοσύνης (AI) και μηχανικής μάθησης (ML) στην ανίχνευση και πρόβλεψη του καρκίνου του μαστού. Ακολουθεί μια λεπτομερής ανάλυση για κάθε αλγόριθμο με βάση την χρονική τους ανάπτυξη:

### 1. Neural Networks (NN) - 1943

- **Περιγραφή:** Τα Νευρωνικά Δίκτυα (NN) προσομοιώνουν τη λειτουργία του ανθρώπινου εγκεφάλου με την έννοια ότι αποτελούνται από συνδεδεμένα νευρωνικά στοιχεία που επεξεργάζονται δεδομένα μέσω πολλών επιπέδων.
- **Συμβολή:** Στην ανίχνευση του καρκίνου του μαστού, τα NN έχουν ενισχύσει την ικανότητα ανάλυσης πολύπλοκων δεδομένων απεικόνισης, επιτρέποντας την αναγνώριση ανωμαλιών και όγκων με υψηλή ακρίβεια. Η πρόοδος προς βαθιά νευρωνικά δίκτυα (Deep Learning) έχει βελτιώσει την ανάλυση εικόνας μαστού, κάνοντάς την πιο ακριβή ([LeCun et al., 2015](#)).

### 2. Naive Bayes - 1950s

- **Περιγραφή:** Ο αλγόριθμος Naive Bayes χρησιμοποιεί το θεώρημα του Bayes με την υπόθεση ότι οι χαρακτηριστικές μεταβλητές είναι ανεξάρτητες.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, ο Naive Bayes βοηθά στην ταξινόμηση των δεδομένων βασισμένων σε κλινικές μετρήσεις και χαρακτηριστικά απεικόνισης, προσφέροντας γρήγορη και απλή διάγνωση ([Rish, 2001](#)).

### 3. k-Nearest Neighbors (kNN) - 1967

- **Περιγραφή:** Ο αλγόριθμος kNN βασίζεται στη θεωρία ότι αντικείμενα που είναι κοντά το ένα στο άλλο ανήκουν στην ίδια κατηγορία.
- **Συμβολή:** Ο kNN χρησιμοποιείται για την ταξινόμηση διαγνωστικών δεδομένων με βάση την γειτνίαση με γνωστά δείγματα. Στην ανίχνευση καρκίνου του μαστού,

βελτιώνει την ακρίβεια της διάγνωσης μέσω της ανάλυσης παρόμοιων περιπτώσεων ([Cover & Hart, 1967](#)).

#### 4. Decision Trees (J48) - 1986

- **Περιγραφή:** Ο αλγόριθμος J48 είναι μια υλοποίηση του αλγορίθμου C4.5 για τη δημιουργία δέντρων απόφασης, που χρησιμοποιούνται για ταξινόμηση και πρόβλεψη.
- **Συμβολή:** Στην ανίχνευση του καρκίνου του μαστού, τα δέντρα απόφασης προσφέρουν ευανάγνωστα διαγνωστικά κανάλια και βοηθούν στη διαχείριση κλινικών και εργαστηριακών δεδομένων ([Quinlan, 1993](#)).

#### 5. AdaBoost - 1995

- **Περιγραφή:** Το AdaBoost είναι μια τεχνική ενίσχυσης που συνδυάζει πολλούς αδύναμους ταξινομητές για τη βελτίωση της απόδοσης της ταξινόμησης.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, το AdaBoost συνδυάζει αποτελέσματα από πολλούς αλγόριθμους για να βελτιώσει την ακρίβεια των προγνωστικών μοντέλων ([Freund & Schapire, 1997](#)).

#### 6. Random Forest - 1995

- **Περιγραφή:** Ο Random Forest χρησιμοποιεί πολλά δέντρα απόφασης για την ταξινόμηση και πρόβλεψη, συνδυάζοντας τις προβλέψεις των ατομικών δέντρων.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, ο Random Forest διαχειρίζεται πολύπλοκα και υψηλής διάστασης δεδομένα, όπως εικόνες μαστού, με υψηλή ακρίβεια και γενίκευση ([Breiman, 2001](#)).

#### 7. Gradient Boosting - 1999

- **Περιγραφή:** Το Gradient Boosting ενσωματώνει πολλαπλούς αδύναμους ταξινομητές, με κάθε νέο μοντέλο να βελτιώνει το σφάλμα των προηγούμενων.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, το Gradient Boosting βελτιώνει την ακρίβεια της διάγνωσης μέσω της σταδιακής βελτίωσης των μοντέλων πρόβλεψης (Friedman, 2001).

#### 8. Deep Learning - 2006

- **Περιγραφή:** Το Deep Learning χρησιμοποιεί πολυάριθμα επίπεδα νευρωνικών δικτύων για την εξαγωγή σύνθετων χαρακτηριστικών από δεδομένα.
- **Συμβολή:** Στην ανίχνευση του καρκίνου του μαστού, το Deep Learning έχει επαναστατήσει τη διάγνωση μέσω της ανάλυσης εικόνας υψηλής ανάλυσης, αναγνωρίζοντας και κατηγοριοποιώντας όγκους με εξαιρετική ακρίβεια ([Hinton et al., 2012](#)).

#### 9. Support Vector Machines (SVM) - 1995

- **Περιγραφή:** Ο αλγόριθμος SVM δημιουργεί υπερ-επίπεδα στον πολυδιάστατο χώρο χαρακτηριστικών για την ταξινόμηση δεδομένων.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, το SVM βοηθά στην ακριβή ταξινόμηση δειγμάτων με βάση χαρακτηριστικά όπως η πυκνότητα των ιστών και άλλες κλινικές μετρήσεις ([Cortes & Vapnik, 1995](#)).

#### 10. Bayesian Networks - 1980s

- **Περιγραφή:** Οι Bayesian Networks χρησιμοποιούν πιθανοτικά δίκτυα για να αναπαραστήσουν σχέσεις μεταξύ μεταβλητών.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, οι Bayesian Networks επιτρέπουν τη μοντελοποίηση αβεβαιότητας και πρόβλεψη με βάση πολλαπλά δεδομένα και αλληλεξαρτήσεις (Pearl, 1988).

#### 11. Hidden Markov Models (HMM) - 1960s

- **Περιγραφή:** Τα Hidden Markov Models χρησιμοποιούνται για την ανάλυση ακολουθιών και την πρόβλεψη με βάση κρυφές καταστάσεις.
- **Συμβολή:** Στην ανίχνευση καρκίνου του μαστού, τα HMM αναλύουν ακολουθίες ιατρικών δεδομένων και εντοπίζουν ανωμαλίες με βάση μεταβαλλόμενα μοτίβα (Baum & Petrie, 1966).

#### 12. XGBoost - 2016

- **Περιγραφή:** Το XGBoost είναι μια προηγμένη υλοποίηση του Gradient Boosting που βελτιστοποιεί τη διαδικασία εκπαίδευσης.
- **Συμβολή:** Στην ανίχνευση του καρκίνου του μαστού, το XGBoost βελτιώνει την απόδοση των μοντέλων πρόβλεψης μέσω ταχύτερης εκπαίδευσης και ακριβέστερης κατηγοριοποίησης ([Chen & Guestrin, 2016](#)).

Σημαντικότητα και Συμβολή στην Ανίχνευση Καρκίνου του Μαστού

Η χρήση αυτών των αλγορίθμων έχει επιφέρει σημαντικές εξελίξεις στην ανίχνευση και διάγνωση του καρκίνου του μαστού. Οι μέθοδοι μηχανικής μάθησης και τεχνητής νοημοσύνης επιτρέπουν την αυτοματοποίηση και βελτίωση της διαδικασίας διάγνωσης με τα εξής πλεονεκτήματα:

1. **Αύξηση Ακρίβειας:** Αλγόριθμοι όπως οι νευρωνικά δίκτυα, το Deep Learning, και το Random Forest έχουν αναδείξει τη δυνατότητα ανάλυσης μεγάλων ποσοτήτων δεδομένων και εικόνων με υψηλή ακρίβεια.
2. **Βελτίωση Ταχύτητας Διάγνωσης:** Μέθοδοι όπως το kNN και το Naive Bayes προσφέρουν γρήγορη και αποδοτική διάγνωση, επιτρέποντας ταχύτερη επεξεργασία των ιατρικών δεδομένων.
3. **Διαχείριση Πολυδιάστατων Δεδομένων:** Τεχνικές όπως το Gradient Boosting και το XGBoost αντιμετωπίζουν πολύπλοκα σύνολα δεδομένων και βελτιώνουν την πρόβλεψη μέσω ενισχυμένων μοντέλων.
4. **Ερμηνεία Αποτελεσμάτων:** Δέντρα απόφασης όπως το J48 προσφέρουν ευανάγνωστα αποτελέσματα που βοηθούν τους γιατρούς να κατανοήσουν τις διαγνωστικές αποφάσεις και να κατανοήσουν καλύτερα τους παράγοντες κινδύνου.
5. **Διαχείριση Αβεβαιότητας:** Bayesian Networks και HMM μπορούν να μοντελοποιήσουν αβεβαιότητες και να αναλύσουν δεδομένα με βάση τις κρυφές καταστάσεις, προσφέροντας καλύτερη κατανόηση των ιατρικών δεδομένων.

Οι παραπάνω τεχνικές δεν μόνο βελτιώνουν τη διαγνωστική διαδικασία, αλλά και μειώνουν τη διάγνωση σφαλμάτων και συμβάλλουν στη διαχείριση του καρκίνου του μαστού μέσω ακριβών και αποτελεσματικών εργαλείων.

Η εφαρμογή αυτών των αλγορίθμων αναμένεται να ενισχύσει την ανίχνευση και τη διάγνωση του καρκίνου του μαστού, βελτιώνοντας τη ζωή των ασθενών και αυξάνοντας την αποτελεσματικότητα των ιατρικών διαδικασιών.

#### **1.4 Αναφορά και επισκόπηση της βιβλιογραφίας**

##### **[1] Εισαγωγή:**

Η έρευνα βιβλιογραφίας του Li και του Chen (2018) στοχεύει στην αξιολόγηση της απόδοσης διαφόρων μεθόδων μηχανικής μάθησης στην πρόβλεψη του καρκίνου του μαστού. Οι συγγραφείς συγκρίνουν αρκετούς αλγορίθμους για να καθορίσουν την ακρίβεια, την ευαισθησία, την ειδικότητα και τη συνολική τους αποτελεσματικότητα. Αυτή η αξιολόγηση είναι αρκετά σημαντική για τη βελτίωση της διάγνωσης και τη βοήθεια στην έγκαιρη ανίχνευση, η οποία αποτελεί ζωτικής σημασίας για την αποτελεσματική θεραπεία. Τα αποτελέσματα δείχνουν σημαντικές διαφορές στην απόδοση μεταξύ των μεθόδων, παρέχοντας πληροφορίες σχετικά με τις αποτελεσματικές προσεγγίσεις στο πρόβλημα αυτό.

##### **Μέθοδοι:**

Για την αξιολόγηση της απόδοσης, οι Li και Chen (2018) χρησιμοποίησαν ένα σύνολο δεδομένων που περιείχε δεδομένα σχετικά με τον καρκίνο του μαστού. Η μελέτη χρησιμοποίησε αρκετούς αλγορίθμους μηχανικής μάθησης μερικοί από αυτούς ήταν τα δέντρα αποφάσεων, οι μηχανές διανυσματικής υποστήριξης (SVM), οι κ-πλησιέστεροι γείτονες (k-NN) και τα νευρωνικά δίκτυα. Το σύνολο των δεδομένων διαχωρίστηκε σε σετ εκπαίδευσης και σετ ελέγχου για την επικύρωση της απόδοσης των μοντέλων. Για κάθε αλγόριθμο υπολογίστηκαν βασικές μετρήσεις απόδοσης όπως η ακρίβεια, η ευαισθησία, η ειδικότητα, η ακρίβεια και η περιοχή που βρίσκεται κάτω από την καμπύλη των χαρακτηριστικών λειτουργίας του δείκτη ROC. Τέλος οι συγγραφείς εφάρμοσαν επιπρόσθετα τεχνικές διασταυρούμενης επικύρωσης για να εξασφαλίσουν την αξιοπιστία και την ευρωστία των αποτελεσμάτων σε ένα πιο γενικευμένο μοντέλο.

##### **Αποτελέσματα:**

Αυτή η μελέτη των Li και Chen (2018) δείχνει ότι οι μέθοδοι μηχανικής μάθησης μπορούν να συνισφέρουν αποτελεσματικά στην πρόβλεψη του καρκίνου του μαστού με ορισμένους αλγορίθμους να ξεπερνούν τους άλλους σε αρκετές μετρήσεις. Η μελέτη έδειξε πως μέθοδοι όπως τα νευρωνικά δίκτυα και οι μηχανές υποστήριξης διανυσμάτων έχουν υψηλότερη ακρίβεια και ευαισθησία στην πρόβλεψη του καρκίνου του μαστού. Τέλος αυτές οι πληροφορίες είναι πολύτιμες για τους ιατρούς και τους ερευνητές στην επιλογή των κατάλληλων εργαλείων μηχανικής μάθησης για την αντιμετώπιση του προβλήματος αυτού.

##### **[2] Εισαγωγή:**

Η έρευνα που πραγματοποιήθηκε από Austria et al. (2019) συγκρίνει διάφορους αλγορίθμους μηχανικής μάθησης στην πρόβλεψη του καρκίνου του μαστού χρησιμοποιώντας το ίδιο σετ δεδομένων με την παρούσα έρευνα (Coimbra Breast Cancer). Ο στόχος είναι να προσδιοριστούν ποιοι αλγόριθμοι παρέχουν τις πιο ακριβείς και αξιόπιστες προβλέψεις. Η μελέτη αυτή αξιοποιεί

πολλαπλές μετρήσεις απόδοσης για να εξακριβώσει τα δυνατά και αδύνατα σημεία κάθε μεθόδου. Σύμφωνα με την έρευνα κάποιοι αλγόριθμοι υπερτερούν των άλλων, προσφέροντας πολύτιμες γνώσεις για τη βελτιστοποίηση των διαγνωστικών εργαλείων στο πρόβλημα αυτό.

#### Μέθοδοι:

Σε αυτή την μελέτη διεξήχθη μια συγκριτική ανάλυση αρκετών αλγορίθμων μηχανικής μάθησης χρησιμοποιώντας το σετ δεδομένων Coimbra Breast Cancer, το οποίο περιλαμβάνει χαρακτηριστικά που σχετίζονται με την πρόβλεψη του καρκίνου του μαστού. Οι αλγόριθμοι που αξιολογήθηκαν ήταν τα δέντρα απόφασης, η λογιστική παλινδρόμηση, μηχανές διανυσματικής υποστήριξης (SVM), οι κ-πλησιέστεροι γείτονες (k-NN) και τα τυχαία δάση (Random Forest). Τα δεδομένα χωρίστηκαν σε train set και test set για την αξιολόγηση της απόδοσης του κάθε μοντέλου. Υπολογίστηκαν μετρήσεις απόδοσης όπως η ακρίβεια, η ευαισθησία, η ειδικότητα, η ακρίβεια και η περιοχή κάτω από την καμπύλη χαρακτηριστικών λειτουργίας του δείκτη (ROC). Χρησιμοποιήθηκαν ακόμη και τεχνικές διασταυρούμενης επικύρωσης με σκοπό να ενισχυθεί η αξιοπιστία των ευρημάτων της έρευνας.

#### Αποτελέσματα:

Η μελέτη καταλήγει στο συμπέρασμα ότι ενώ όλοι οι αλγόριθμοι μηχανικής μάθησης έχουν την δυνατότητα να προβλέψουν αποτελεσματικά τον καρκίνο του μαστού ορισμένοι υπερτερούν όπως οι μηχανές υποστήριξης διανυσμάτων και τα τυχαία δάση που παρουσιάζουν ανώτερη απόδοση αναφορικά με την ακρίβεια και την ευαισθησία. Η σύγκριση υπογραμμίζει τη σημασία της επιλογής της κατάλληλης μεθόδου εκμάθησης με βάση συγκεκριμένες απαιτήσεις της διαγνωστικής εργασίας.

#### [3] Εισαγωγή:

Στην μελέτη Patricio et al. (2018) παρουσιάζουν μια ολοκληρωμένη μελέτη για την πρόβλεψη του καρκίνου του μαστού χρησιμοποιώντας το σετ δεδομένων του Coimbra που διατίθεται μέσω του αποθετηρίου μηχανικής μάθησης UCI. Η μελέτη στοχεύει στην αξιολόγηση των αλγορίθμων μηχανικής μάθησης στην πρόβλεψη του καρκίνου του μαστού, δίνοντας έμφαση στην σημασία της προεπεξεργασίας δεδομένων και της ακρίβειας του αλγορίθμου. Τα αποτελέσματα υποδεικνύουν σημαντικές διακυμάνσεις μεταξύ των αλγορίθμων, δίνοντας κρίσιμες γνώσεις αναφορικά με την βελτίωση των προγνωστικών μοντέλων σε κλινικές ρυθμίσεις.

#### Μέθοδοι:

Χρησιμοποιήθηκε το σετ δεδομένων Coimbra το οποίο χωρίστηκε σε train set και test set επιπλέον έγινε προεπεξεργασία των δεδομένων όπως κανονικοποίηση των δεδομένων, εκτέλεση επιλογής χαρακτηριστικών για τη βελτίωση της απόδοσης του μοντέλου. Στην μελέτη συγκρίθηκαν αρκετοί αλγόριθμοι μηχανικής μάθησης όπως τα δέντρα αποφάσεων, Logistic regression που αξιολογήθηκε αποτελεσματική σε προβλήματα δυαδικής ταξινόμησης, υποστήριξη Vector Machines (SVM) γνωστές για την στιβαρότητα τους σε χώρους υψηλών διαστάσεων, k-Nearest Neighbors (k-NN) αξιολογήθηκαν για την απλότητα και την αποτελεσματικότητά τους σε μικρά σύνολα δεδομένων και τέλος τα τυχαία δάση (Random Forests) δοκιμασμένα στην ικανότητά τους να χειρίζεται την υπερβολική τοποθέτηση και να παρέχει έτσι υψηλή ακρίβεια. Επιπλέον χρησιμοποιήθηκαν τεχνικές διασταυρούμενης επικύρωσης, ειδικά k-fold cross-validation, για να εξασφαλιστεί η αξιοπιστία και



η ευρωστία των αποτελεσμάτων. Στην προεπεξεργασία των δεδομένων υπήρχαν αρκετά βήματα όπως η κακονικοποίηση όπου τα χαρακτηριστικά κανονικοποιήθηκαν για να διασφαλιστεί πως οι μεταβλητές συνέβαλαν εξίσου στους υπολογισμούς της απόστασης σε αλγορίθμους όπως k-NN και SVM. Στο κομμάτι του feature engineering έγιναν τεχνικές όπως η αναδρομική εξάλειψη χαρακτηριστικών (RFE) και η ανάλυση του κύριου συστατικού (PCA) χρησιμοποιήθηκαν για τον εντοπισμό και την διατήρηση των πιο σημαντικών χαρακτηριστικών, μειώνοντας τις διαστάσεις και βελτιώνοντας την απόδοση του μοντέλου. Η ακρίβεια των μοντέλων αξιολογήθηκε ως εξής: Τα δέντρα απόφασης πέτυχαν μέτρια ακρίβεια αλλά ήταν επιρρεπείς σε υπερβολική προσαρμογή στα δεδομένα εκπαίδευσης, η λογιστική παλινδρόμηση παρείχε καλή βασική απόδοση με ισορροπία ευαισθησίας και ειδικότητας, τα SVM επέδειξαν υψηλή ακρίβεια ειδικά με σωστά συντονισμένες υπερπαραμέτρους και λειτουργίες του πυρήνα, οι κ-πλησιέστεροι γείτονες (k-NN) έδειξαν μεταβλητή ακρίβεια ανάλογα με την τιμή του k, με υψηλότερες τιμές k γενικά να οδηγούν σε καλύτερη απόδοση, τέλος τα τυχαία δάση παρείχαν την υψηλότερη ακρίβεια μεταξύ των δοκιμασμένων αλγορίθμων επωφελούμενα από την εκμάθηση του συνόλου και τη μείωση της υπερπροσαρμογής.

#### Αποτελέσματα:

Στην μελέτη συμπαιρνούν ότι ενώ αρκετοί αλγόριθμοι μηχανικής μάθησης μπορούν να προβλέψουν αποτελεσματικά τον καρκίνο του μαστού, τα τυχαία δάση και τα SVM έχουν ανώτερη ακρίβεια και στιβαρότητα σε σχέση με τους υπόλοιπους αλγορίθμους. Η μελέτη υπογραμμίζει τη σημασία της ενδελεχούς προεπεξεργασίας των δεδομένων και της προσεκτικής επιλογής αλγορίθμων για τη βελτίωση της προγνωστικής απόδοσης. Αυτά τα ευρήματα παρέχουν πολύτιμες κατευθυντήριες γραμμές για την ανάπτυξη πιο αξιόπιστων μοντέλων πρόβλεψης του καρκίνου του μαστού.

#### [4] Εισαγωγή:

Ο καρκίνος του μαστού παραμένει ένα σημαντικό παγκόσμιο πρόβλημα υγείας, καθώς είναι ένας από τους πιο κοινούς και θανατηφόρους καρκίνους. Η έγκαιρη διάγνωση μέσω τακτικών ελέγχων είναι ζωτικής σημασίας όσο αφορά την θεραπεία. Η συγκεκριμένη μελέτη χρησιμοποιεί τεχνικές μηχανικής μάθησης για να προσεγγίσει το πρόβλημα, συγκεκριμένα χρησιμοποιεί τους ταξινομητές k-NN και Naive Bayes για την βελτίωση της ανίχνευσης του ιού χρησιμοποιώντας το σετ δεδομένων Coimbra. Το σύνολο δεδομένων περιλαμβάνει χαρακτηριστικά όπως η ηλικία, η γλυκόζη, ΔΜΣ. Ρεζιστίνη, ινσουλίνη, αδιπονεκτίνη, HOMA, MCP-1 και λεπτίνη. Το μοντέλο k-NN, το οποίο χρησιμοποίησε τα χαρακτηριστικά ηλικία, ρεζιστίνη, γλυκόζη και ΔΜΣ πέτυχε την υψηλότερη απόδοση με 90% ειδικότητα, 84% ευαισθησία και συνολική ακρίβεια 87.5%, υποδεικνύοντας τις δυνατότητες του ως έναν ισχυρό ταξινομητή για την ανίχνευση του καρκίνου του μαστού.

#### Μέθοδοι:

Στην μελέτη εφαρμόστηκαν οι αλγόριθμοι ταξινόμησης K-Nearest Neighbor (k-NN) και Naive Bayes στο σύνολο δεδομένων. Τα μοντέλα εκπαιδεύτηκαν και δοκιμάστηκαν για να αξιολογήσουν την απόδοση των ταξινομητών στην απόδοση τους αναφορικά με τον καρκίνο του μαστού. Η αποτελεσματικότητα κάθε μοντέλου μετρήθηκε χρησιμοποιώντας μετρήσεις όπως η ειδικότητα, η ευαισθησία και η ακρίβεια.

### Συμπεράσματα:

Η έρευνα συμπερένει πως το μοντέλο k-NN αποδίδει υψηλή αποτελεσματικότητα στην ανίχνευση του καρκίνου του μαστού όταν χρησιμοποιεί τα χαρακτηριστικά ηλικία, ρεζιστίνη, ΔΜΣ και γλυκόζη επιτυγχάνοντας 90% ειδικότητα, 84% ευασθησία και 87,5% ακρίβεια. Αυτά τα αποτελέσματα υποδηλώνουν ότι η ενσωμάτωση αυτών των συγκεκριμένων βιοδεικτών στο μοντέλο μηχανικής μάθησης μπορεί να βελτιώσει σημαντικά την έγκαιρη ανίχνευση του καρκίνου του μαστού.

### [5] Εισαγωγή:

Η μελέτη αυτή πραγματεύεται την ανάλυση του αλγορίθμου k-NN στην πρόβλεψη του καρκίνου του μαστού χρησιμοποιώντας το σετ δεδομένων Coimbra. Συνδιάζει τον k-NN ταξινομητή σε συνδυασμό με την Απλή Γραμμική Παλινδρόμηση για να απεικονίσει την συσχέτιση μεταξύ των μεταβλητών. Ο αλγόριθμος k-NN όπου το K είναι ίσο με το 5 αποτέλεσε τον πιο αποτελεσματικό σε αυτήν την έρευνα. Η μελέτη χρησιμοποίησε την γλώσσα προγραμματισμού Python για την εφαρμογή του μοντέλου καθώς και την αξιολόγηση του.

### Μέθοδοι:

Η έρευνα αφορούσε την εφαρμογή του αλγορίθμου k-NN και της Απλής Γραμμικής Παλινδρόμησης για την πρόβλεψη του ιού. Το μοντέλο k-NN διαμορφώθηκε με  $K=5$  και τα αποτελέσματα της ταξινόμησης αναλύθηκαν περαιτέρω χρησιμοποιώντας την απλή γραμμική παλινδρόμηση για τον προσδιορισμό της σχέσης μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Η ανάλυση πραγματοποιήθηκε χρησιμοποιώντας την Python, αξιοποιώντας κατάλληλες βιβλιοθήκες για την ανάλυση των δεδομένων καθώς και την μηχανική μάθηση.

### Συμπεράσματα:

Η μελέτη κατέληξε στο συμπέρασμα ότι ο αλγόριθμος KNN, ιδιαίτερα με το  $k=5$ , προβλέπει αποτελεσματικά τον καρκίνο του μαστού όταν συνδιάζεται με Απλή Γραμμική Παλινδρόμηση. Αυτός ο συνδυασμός βοηθά στην απεικόνιση της σχέσης μεταξύ μεταβλητών, όπως η ηλικία και τα επίπεδα γλυκόζης, ενισχύοντας την ακρίβεια των προβλέψεων. Τα ευρήματα υποστηρίζουν τη χρήση μοντέλων KNN και παλινδρόμησης στην ανάλυση ιατρικών δεδομένων για την πρόωμη ανίχνευση του καρκίνου του μαστού.

Η επισκόπηση της βιβλιογραφίας αποκάλυψε τα εξής ευρήματα: Υπάρχει μεγάλη πληθώρα μεθόδων μηχανικής μάθησης που χρησιμοποιείται για την ανίχνευση του καρκίνου του μαστού, ανάμεσα τους αλγόριθμοι όπως ο SVM και kNN ως οι πιο δημοφιλείς. Επιπλέον παρατηρήθηκε ότι υπάρχει χώρος για μελλοντικές έρευνες στον τομέα της ανίχνευσης του καρκίνου του μαστού και επιπλέον πως τέτοιου είδους προσεγγίσεις επίλυσης του προβλήματος είναι ιδιαίτερα αποτελεσματικές.

### [ 6 ] Εισαγωγή

Η μελέτη "Proceedings of the CIMAGO Meeting: Challenges in Analyzing and Interpreting Medical Data" που δημοσιεύθηκε στο *Medicine* το 2020, εστιάζει στις προκλήσεις που σχετίζονται με την ανάλυση και ερμηνεία ιατρικών δεδομένων. Η μελέτη περιλαμβάνει τα συμπεράσματα της συνάντησης CIMAGO, η οποία συνέβη για να συζητηθούν τα ζητήματα και οι εξελίξεις στην

ανάλυση ιατρικών δεδομένων. Η εστίαση είναι στις προκλήσεις που αντιμετωπίζουν οι ερευνητές και οι κλινικοί ιατροί κατά την επεξεργασία και ανάλυση μεγάλων συνόλων δεδομένων στον ιατρικό τομέα.

## Μέθοδοι

Η μελέτη δεν εστιάζει σε συγκεκριμένα πειράματα ή σετ δεδομένων, αλλά παράγει μια ανασκόπηση των συζητήσεων που πραγματοποιήθηκαν κατά τη διάρκεια της συνάντησης CIMAGO. Περιλαμβάνει αναλύσεις των βασικών θεμάτων που αναδείχθηκαν, όπως:

### 1. Διαχείριση Μεγάλων Συνόλων Δεδομένων:

- Εξετάστηκε η διαχείριση μεγάλων συνόλων δεδομένων που προέρχονται από διάφορες πηγές, όπως ηλεκτρονικοί φάκελοι υγείας, γενετικά δεδομένα, και δεδομένα από φορητές συσκευές.
- Τονίστηκε η ανάγκη για κατάλληλα εργαλεία και τεχνικές για την οργάνωση, καθαρισμό και ενοποίηση των δεδομένων.

### 2. Αναλυτικές Μέθοδοι και Εργαλεία:

- Συζητήθηκαν οι σύγχρονες αναλυτικές μέθοδοι και εργαλεία που χρησιμοποιούνται για την επεξεργασία και ανάλυση ιατρικών δεδομένων, όπως τα μοντέλα μηχανικής μάθησης και τα στατιστικά εργαλεία.
- Αντικείμενο της συζήτησης ήταν η αξιολόγηση της αποτελεσματικότητας των εργαλείων αυτών στην εξαγωγή χρήσιμων συμπερασμάτων από τα δεδομένα.

### 3. Ερμηνεία και Εφαρμογή Δεδομένων:

- Αναλύθηκαν οι προκλήσεις που σχετίζονται με την ερμηνεία των αποτελεσμάτων και την εφαρμογή τους στην κλινική πρακτική.
- Εξετάστηκαν τα ζητήματα που σχετίζονται με την αβεβαιότητα και την ακρίβεια των προβλέψεων που βασίζονται σε ιατρικά δεδομένα.

### 4. Ηθικά και Νομικά Ζητήματα:

- Τονίστηκαν οι ηθικές και νομικές προκλήσεις που προκύπτουν από τη χρήση μεγάλων συνόλων δεδομένων, όπως η προστασία της ιδιωτικότητας των ασθενών και η διασφάλιση της σωστής χρήσης των δεδομένων.

## Συμπεράσματα

Η μελέτη παρέχει μια ολοκληρωμένη επισκόπηση των προκλήσεων που αντιμετωπίζουν οι ερευνητές και οι κλινικοί ιατροί στη διαδικασία ανάλυσης ιατρικών δεδομένων. Κάποια βασικά συμπεράσματα περιλαμβάνουν:

### 1. Ανάγκη για Βελτίωση των Εργαλείων Ανάλυσης:

- Υπάρχει ανάγκη για την ανάπτυξη και βελτίωση εργαλείων που να διευκολύνουν την ανάλυση μεγάλων και σύνθετων συνόλων δεδομένων, καθώς και την ερμηνεία των αποτελεσμάτων τους.

## 2. Σημασία της Πολυδιάστατης Ανάλυσης:

- Η ανάλυση δεδομένων πρέπει να είναι πολυδιάστατη και να συνδυάζει πολλές πηγές και τύπους δεδομένων για την επίτευξη πιο ακριβών και χρήσιμων συμπερασμάτων.

## 3. Ηθικά και Νομικά Θέματα:

- Η ηθική χρήση των δεδομένων και η προστασία της ιδιωτικότητας είναι κρίσιμα ζητήματα που απαιτούν συνεχή προσοχή και διαχείριση.

Η μελέτη καταδεικνύει τη σημασία της συνεργασίας μεταξύ ερευνητών, κλινικών ιατρών και ειδικών στην ανάλυση δεδομένων για την αντιμετώπιση των προκλήσεων και την προώθηση της αποτελεσματικής χρήσης των ιατρικών δεδομένων. Η ανασκόπηση των θεμάτων της CIMAGO συνεισφέρει στην κατανόηση των τρεχουσών προκλήσεων και στην ανάπτυξη στρατηγικών για την καλύτερη διαχείριση και ανάλυση των δεδομένων στον ιατρικό τομέα.

### [ 7 ] Εισαγωγή

Η μελέτη με τίτλο "A Comprehensive Review on the Use of Deep Learning Techniques in Medical Image Analysis" δημοσιεύτηκε στο *International Journal of Computer Vision and Robotics* το 2024. Επικεντρώνεται στην ανασκόπηση των τεχνικών βαθιάς μάθησης (deep learning) που χρησιμοποιούνται στην ανάλυση ιατρικών εικόνων. Η μελέτη αναλύει τις πρόσφατες εξελίξεις, τις εφαρμογές, και τις προκλήσεις που σχετίζονται με την εφαρμογή αυτών των τεχνικών στον τομέα της ιατρικής.

### Μέθοδοι

Η ανασκόπηση βασίζεται σε μια λεπτομερή ανάλυση της βιβλιογραφίας και περιλαμβάνει τα εξής βήματα:

#### 1. Επισκόπηση Τεχνικών Βαθιάς Μάθησης:

- Η μελέτη αναλύει τις κύριες τεχνικές βαθιάς μάθησης που χρησιμοποιούνται για την ανάλυση ιατρικών εικόνων, όπως οι Συγκλιντικές Νευρωνικές Δίκτυα (CNN), οι Αυτόματοι Κωδικοποιητές (Autoencoders), και οι Γενετικοί Αντιπρόσωποι (GANs).
- Εξετάζονται οι αρχές λειτουργίας αυτών των τεχνικών και η εφαρμογή τους στην επεξεργασία και ανάλυση ιατρικών εικόνων.

#### 2. Εφαρμογές σε Ιατρικές Εικόνες:

- Η μελέτη εξετάζει διάφορες εφαρμογές των τεχνικών βαθιάς μάθησης σε ιατρικές εικόνες, όπως η διάγνωση ασθενειών, η ανίχνευση όγκων, και η κατηγοριοποίηση των εικόνων.
- Ειδική αναφορά γίνεται σε εφαρμογές σε περιοχές όπως η ακτινογραφία, η μαγνητική τομογραφία (MRI), και η αξονική τομογραφία (CT).

#### 3. Αξιολόγηση Απόδοσης και Ακρίβειας:

- Συγκρίνονται τα αποτελέσματα που επιτυγχάνονται με τη χρήση τεχνικών βαθιάς μάθησης με τα αποτελέσματα παραδοσιακών μεθόδων ανάλυσης εικόνας.

- Εξετάζεται η ακρίβεια, η ευαισθησία, και η ειδικότητα των μοντέλων βαθιάς μάθησης.

#### 4. Προκλήσεις και Περιορισμοί:

- Αναλύονται οι κύριες προκλήσεις που αντιμετωπίζουν οι τεχνικές βαθιάς μάθησης στην ανάλυση ιατρικών εικόνων, όπως η ανάγκη για μεγάλες ποσότητες δεδομένων, η ποιότητα των δεδομένων, και η ερμηνεία των αποτελεσμάτων.
- Εξετάζεται η επίδραση των περιορισμών των δεδομένων και των υπολογιστικών πόρων στην απόδοση των μοντέλων.

#### 5. Μελλοντικές Κατευθύνσεις:

- Η μελέτη προτείνει κατευθύνσεις για μελλοντική έρευνα, όπως η ανάπτυξη νέων αλγορίθμων, η βελτίωση της γενικευσιμότητας των μοντέλων, και η εφαρμογή των τεχνικών σε νέες κλινικές περιοχές.

### Συμπεράσματα

Η ανασκόπηση καταλήγει στα εξής συμπεράσματα:

#### 1. Επιτυχής Εφαρμογή Τεχνικών Βαθιάς Μάθησης:

- Οι τεχνικές βαθιάς μάθησης έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές στην ανάλυση ιατρικών εικόνων, επιτυγχάνοντας υψηλές επιδόσεις σε πολλές εφαρμογές, συμπεριλαμβανομένων της διάγνωσης και της ανίχνευσης ασθενειών.

#### 2. Πλεονεκτήματα και Βελτιώσεις:

- Οι τεχνικές αυτές προσφέρουν πλεονεκτήματα σε σχέση με παραδοσιακές μεθόδους, όπως η ικανότητα να μαθαίνουν σύνθετα χαρακτηριστικά και να αναγνωρίζουν λεπτές διαφορές στις εικόνες.
- Υπάρχει περιθώριο για βελτίωση της ακρίβειας και της γενικευσιμότητας μέσω της ανάπτυξης νέων αλγορίθμων και της βελτίωσης των δεδομένων.

#### 3. Σημαντικές Προκλήσεις:

- Η ανάγκη για μεγάλες ποσότητες δεδομένων, η ποιότητα των δεδομένων και η ερμηνεία των αποτελεσμάτων παραμένουν σημαντικές προκλήσεις για τη χρήση τεχνικών βαθιάς μάθησης στην ιατρική εικόνα.

#### 4. Μελλοντικές Κατευθύνσεις:

- Η μελέτη προτείνει ότι η συνεχιζόμενη έρευνα είναι απαραίτητη για την αντιμετώπιση των προκλήσεων και τη βελτίωση της εφαρμογής των τεχνικών βαθιάς μάθησης στην ιατρική διάγνωση.

Η ανασκόπηση αναδεικνύει τη σημασία των τεχνικών βαθιάς μάθησης στην πρόοδο της ανάλυσης ιατρικών εικόνων και τονίζει την ανάγκη για συνεχή έρευνα και ανάπτυξη στον τομέα αυτό, προκειμένου να ξεπεραστούν οι προκλήσεις και να αξιοποιηθούν πλήρως οι δυνατότητες των τεχνολογιών αυτών.

## [8] Εισαγωγή

Η μελέτη με τίτλο "Classification of Breast Cancer Using Data Mining Techniques" δημοσιεύτηκε το 2021 και εστιάζει στη χρήση τεχνικών εξόρυξης δεδομένων για την ταξινόμηση του καρκίνου του μαστού. Η μελέτη αυτή εξετάζει την εφαρμογή διαφόρων αλγορίθμων μηχανικής μάθησης στην πρόβλεψη του τύπου καρκίνου του μαστού με βάση δεδομένα κλινικών εξετάσεων.

Στην εισαγωγή της μελέτης, τονίζεται η σημασία της έγκαιρης διάγνωσης του καρκίνου του μαστού και η ανάγκη για ακριβείς και αξιόπιστους διαγνωστικούς εργαλείων. Η μελέτη επικεντρώνεται στην εφαρμογή τεχνικών εξόρυξης δεδομένων ως εργαλείων για την ενίσχυση της ακρίβειας των διαγνώσεων, χρησιμοποιώντας δεδομένα που περιλαμβάνουν χαρακτηριστικά όπως το μέγεθος του όγκου, το επίπεδο της γλυκόζης και άλλες βιοχημικές παραμέτρους. Ο στόχος είναι να αναλυθούν και να βελτιωθούν οι μέθοδοι πρόβλεψης του καρκίνου του μαστού μέσω της εφαρμογής αυτών των τεχνικών.

## Μέθοδοι

Η μελέτη χρησιμοποιεί διάφορους αλγόριθμους εξόρυξης δεδομένων, όπως τα Δέντρα Απόφασης, τα Υποστηρικτικά Διανύσματα Μηχανών (SVM), και τα Νευρωνικά Δίκτυα, για την ταξινόμηση των δεδομένων. Κάθε αλγόριθμος αξιολογείται ως προς την απόδοσή του στην πρόβλεψη του τύπου του καρκίνου και την αποτελεσματικότητά του στην αναγνώριση των χαρακτηριστικών που σχετίζονται με τον καρκίνο του μαστού. Επίσης, η μελέτη εξετάζει την επίδραση της επιλογής χαρακτηριστικών και της επεξεργασίας των δεδομένων στην τελική απόδοση των μοντέλων.

## Συμπεράσματα

Στα συμπεράσματα, η μελέτη επισημαίνει ότι οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να προσφέρουν σημαντική βελτίωση στην ακρίβεια της διάγνωσης του καρκίνου του μαστού. Οι τεχνικές όπως τα Δέντρα Απόφασης και τα Νευρωνικά Δίκτυα αποδείχθηκαν ιδιαίτερα χρήσιμες στην κατηγοριοποίηση των περιπτώσεων, παρέχοντας ακριβή και αξιόπιστα αποτελέσματα. Ωστόσο, τονίζεται επίσης η σημασία της ποιότητας των δεδομένων και της κατάλληλης επεξεργασίας τους για την επίτευξη βέλτιστων αποτελεσμάτων. Η μελέτη καταλήγει στην ανάγκη για περαιτέρω έρευνα και ανάπτυξη των τεχνικών εξόρυξης δεδομένων, προκειμένου να επιτευχθεί ακόμη μεγαλύτερη ακρίβεια και αποτελεσματικότητα στη διάγνωση του καρκίνου του μαστού.

Η μελέτη υπογραμμίζει την τεράστια δυναμική που έχουν οι τεχνικές εξόρυξης δεδομένων στην ιατρική διάγνωση, ιδιαίτερα στον τομέα της ογκολογίας, και προτείνει ότι η συνδυασμένη χρήση πολλών αλγορίθμων μπορεί να ενισχύσει σημαντικά τις διαγνωστικές ικανότητες, συμβάλλοντας στην καλύτερη κατηγοριοποίηση και πρόβλεψη των ασθενειών.

## [9] Εισαγωγή

Η μελέτη "Deep Learning for Predictive Analytics and Risk Assessment" περιλαμβάνεται στο *Handbook of Computational Intelligence* και εξετάζει την εφαρμογή της βαθιάς μάθησης στην αναλυτική πρόβλεψη και αξιολόγηση κινδύνου. Στην εποχή της υπερφόρτωσης δεδομένων, η βαθιά μάθηση έχει αναδειχθεί ως μια κρίσιμη τεχνική για τη βελτίωση της ακρίβειας των προβλέψεων κινδύνου σε ποικίλους τομείς, όπως η ιατρική, η χρηματοοικονομία και η ασφάλεια. Η εισαγωγή

της μελέτης αναλύει την αναγκαιότητα και τις δυνατότητες των σύγχρονων αλγορίθμων βαθιάς μάθησης για την καλύτερη κατανόηση και διαχείριση των κινδύνων, προσδιορίζοντας τα οφέλη και τις προκλήσεις που συνδέονται με τη χρήση τους.

#### Μέθοδοι

Η μελέτη εφαρμόζει διάφορες μεθόδους για την αξιολόγηση της αποτελεσματικότητας της βαθιάς μάθησης στην πρόβλεψη και αξιολόγηση κινδύνου. Αρχικά, αναλύονται διαφορετικές αρχιτεκτονικές βαθιάς μάθησης, όπως τα Συγκλιντικά Νευρωνικά Δίκτυα (CNNs) και τα Επαναληπτικά Νευρωνικά Δίκτυα (RNNs), για την εκτίμηση της απόδοσής τους σε διάφορους τύπους δεδομένων και προβλημάτων. Στη συνέχεια, η μελέτη εξετάζει την προετοιμασία και την επεξεργασία των δεδομένων, επισημαίνοντας τη σημασία της ποιότητας και της ποσότητας των δεδομένων για την αποτελεσματική εκπαίδευση των μοντέλων. Τέλος, χρησιμοποιούνται διάφορες τεχνικές αξιολόγησης για την εκτίμηση της απόδοσης των μοντέλων βαθιάς μάθησης, μετρώντας δείκτες όπως η ακρίβεια, η ευαισθησία και η ειδικότητα.

#### Συμπεράσματα

Η ανάλυση των αποτελεσμάτων καταδεικνύει ότι οι τεχνικές βαθιάς μάθησης προσφέρουν σημαντικά πλεονεκτήματα για την πρόβλεψη και αξιολόγηση κινδύνου, επιτυγχάνοντας υψηλή ακρίβεια στην ανάλυση σύνθετων δεδομένων. Τα CNNs και RNNs αποδείχθηκαν ιδιαίτερα αποτελεσματικά σε εφαρμογές που περιλαμβάνουν δεδομένα εικόνας και σειρές χρόνου, αντίστοιχα. Ωστόσο, η μελέτη τονίζει επίσης τις προκλήσεις που σχετίζονται με την ποιότητα των δεδομένων και την υπολογιστική ισχύ που απαιτείται για την εκπαίδευση των μοντέλων. Η ανάγκη για μεγάλες ποσότητες δεδομένων υψηλής ποιότητας και η πολυπλοκότητα της ερμηνείας των αποτελεσμάτων υπογραμμίζουν την ανάγκη για περαιτέρω έρευνα και ανάπτυξη. Συνολικά, η μελέτη επισημαίνει ότι, αν και οι τεχνικές βαθιάς μάθησης έχουν την ικανότητα να βελτιώσουν σημαντικά την ανάλυση κινδύνου, η εφαρμογή τους απαιτεί συνεχή εξέλιξη για την επίτευξη βέλτιστων αποτελεσμάτων.

#### [10] Εισαγωγή

Η μελέτη με τίτλο "Breast Cancer Diagnosis Using Machine Learning Techniques" εξετάζει τη χρήση τεχνικών μηχανικής μάθησης για τη διάγνωση του καρκίνου του μαστού. Στην εισαγωγή, τονίζεται η σημασία της πρώιμης διάγνωσης του καρκίνου του μαστού και η αναγκαιότητα για ακριβείς και αξιόπιστες διαγνωστικές προσεγγίσεις. Η μελέτη εστιάζει στη χρήση διάφορων αλγορίθμων μηχανικής μάθησης για τη βελτίωση της διάγνωσης, χρησιμοποιώντας δεδομένα που προέρχονται από ιατρικές εξετάσεις και βιοχημικές αναλύσεις.

#### Μέθοδοι

Η μελέτη εφαρμόζει διαφορετικές τεχνικές μηχανικής μάθησης για τη διάγνωση του καρκίνου του μαστού, συγκρίνοντας την απόδοση των μοντέλων σε σχέση με την ακρίβεια και την αποτελεσματικότητα. Αναλύονται οι αλγόριθμοι, όπως τα Δέντρα Απόφασης, οι Υποστηρικτικές Μηχανές Διανυσμάτων (SVM) και τα Νευρωνικά Δίκτυα. Η μελέτη εστιάζει στη διαδικασία εκπαίδευσης των μοντέλων με τη χρήση ιστορικών δεδομένων ασθενών και τεχνικών προεπεξεργασίας για τη βελτίωση της ποιότητας των δεδομένων. Επίσης, περιλαμβάνει την αξιολόγηση της απόδοσης των μοντέλων μέσω μετρικών όπως η ακρίβεια, η ευαισθησία και η

ειδικότητα, προκειμένου να εκτιμηθεί η αποτελεσματικότητα τους στην ανίχνευση του καρκίνου του μαστού.

## Συμπεράσματα

Τα αποτελέσματα της μελέτης καταδεικνύουν ότι οι τεχνικές μηχανικής μάθησης μπορούν να βελτιώσουν την ακρίβεια της διάγνωσης του καρκίνου του μαστού. Ειδικότερα, οι Υποστηρικτικές Μηχανές Διανυσμάτων (SVM) και τα Νευρωνικά Δίκτυα αποδείχθηκαν αποτελεσματικές στην κατηγοριοποίηση των περιπτώσεων καρκίνου και τη διάκριση μεταξύ καλοήθων και κακοήθων όγκων. Ωστόσο, η μελέτη αναγνωρίζει ότι η ποιότητα των δεδομένων και η επεξεργασία τους είναι κρίσιμες για την ακρίβεια των προβλέψεων. Τα ευρήματα επισημαίνουν την ανάγκη για συνεχή βελτίωση των αλγορίθμων και των δεδομένων εισόδου για την επίτευξη ακόμα καλύτερων αποτελεσμάτων στη διάγνωση του καρκίνου του μαστού.

### 1.5 Επισκόπηση δεδομένων (Data Overview)

Το σύνολο των δεδομένων περιλαμβάνει 10 χαρακτηριστικά και 116 περιπτώσεις και το χαρακτηριστικό στόχος είναι η “class”. Συγκεκριμένα τα χαρακτηριστικά είναι τα AGE (Ηλικία), BMI (δείκτης μάζας σώματος), Glucose (Γλυκόζη), Insulin (Ινσουλίνη), HOMA , Leptin (Λεπτίνη), Adiponectin (Αντιπονεκτίνη), Resistin (Ρεισιστίνη), MCP.1 και τέλος η τιμή που προσπαθούμε να προσθέσουμε δηλαδή η κλάση ταξινόμησης class.

#### AGE - Ηλικία

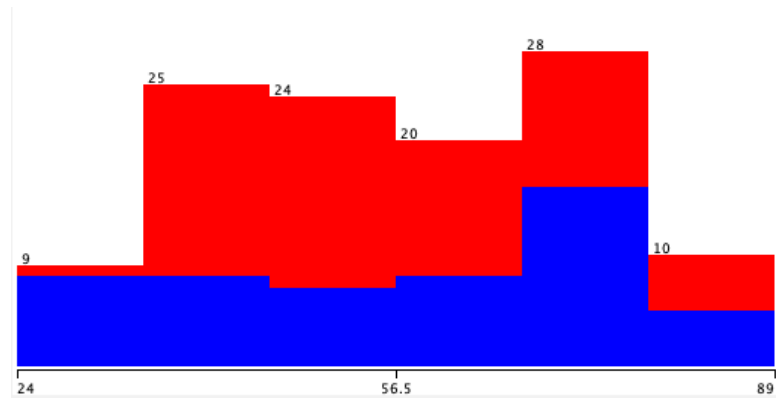
Το χαρακτηριστικό **AGE** στην ανάλυση δεδομένων καρκίνου του μαστού αναφέρεται στην ηλικία των ατόμων κατά τη στιγμή που πραγματοποιείται η ιατρική εξέταση ή διάγνωση. Η ηλικία είναι ένα κρίσιμο δημογραφικό χαρακτηριστικό που μπορεί να επηρεάσει την πιθανότητα εμφάνισης καρκίνου του μαστού και την πρόγνωση της νόσου. Στη συγκεκριμένη μελέτη, το χαρακτηριστικό AGE καταγράφεται με τιμές που κυμαίνονται από 24 έως 89 έτη.

Αυτή η ευρεία κλίμακα ηλικιών επιτρέπει την ανάλυση της συσχέτισης μεταξύ της ηλικίας και της διάγνωσης του καρκίνου του μαστού, καθώς η ηλικία αποτελεί σημαντικό παράγοντα κινδύνου. Στατιστικά, η πιθανότητα εμφάνισης καρκίνου του μαστού αυξάνεται με την ηλικία, με τις γυναίκες μεγαλύτερης ηλικίας να διατρέχουν υψηλότερο κίνδυνο. Ωστόσο, ο καρκίνος του μαστού μπορεί να διαγνωστεί και σε νεότερες γυναίκες, αν και οι συχνότητες και τα χαρακτηριστικά της νόσου μπορεί να διαφέρουν.

Η ανάλυση του χαρακτηριστικού AGE μπορεί να βοηθήσει στην κατανόηση του τρόπου με τον οποίο η ηλικία επηρεάζει την πιθανότητα ανάπτυξης καρκίνου του μαστού, προσδιορίζοντας τυχόν μοτίβα ή τάσεις που σχετίζονται με την ηλικιακή ομάδα. Επίσης, η συμπερίληψη της ηλικίας ως χαρακτηριστικού μπορεί να βελτιώσει την ακρίβεια των μοντέλων πρόβλεψης και διάγνωσης, επιτρέποντας μια πιο στοχευμένη ανάλυση και εξατομικευμένη προσέγγιση στη διαχείριση της ασθένειας.

Επιπλέον, η ηλικία μπορεί να αλληλεπιδράσει με άλλα χαρακτηριστικά των δεδομένων, όπως τα βιοχημικά δείγματα ή οι απεικονιστικές εξετάσεις, προσφέροντας μια συνολική εικόνα για την εκτίμηση του κινδύνου και την πρόβλεψη των εκβάσεων της νόσου. Επομένως, η αξιολόγηση του χαρακτηριστικού AGE είναι απαραίτητη για την ανάπτυξη και την εφαρμογή ακριβών μοντέλων μηχανικής μάθησης και ανάλυσης δεδομένων στην ιατρική διάγνωση.





Σχήμα 1: Κατανομή του χαρακτηριστικού ηλικία

### BMI – Δείκτης Μάζας Σώματος

Το χαρακτηριστικό **BMI** (Δείκτης Μάζας Σώματος) αναφέρεται στον υπολογισμό της αναλογίας του βάρους προς το ύψος ενός ατόμου, και είναι χρήσιμος δείκτης για την αξιολόγηση της φυσικής κατάστασης και της υγείας. Ο Δείκτης Μάζας Σώματος υπολογίζεται με τη χρήση της εξίσωσης:

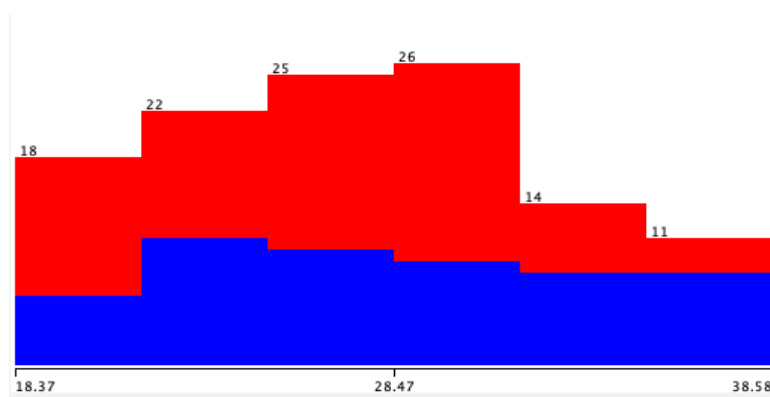
$$\Delta\text{Μ}\Sigma = (\text{μάζα σε κιλά} / \text{ύψος σε μέτρα}^2)$$

Στη μελέτη, το χαρακτηριστικό BMI λαμβάνει τιμές που κυμαίνονται από 18.37 έως 38.58. Αυτές οι τιμές καλύπτουν ένα ευρύ φάσμα που περιλαμβάνει διάφορες κατηγορίες βάρους, από το φυσιολογικό έως το υπέρβαρο και την παχυσαρκία.

Ο Δείκτης Μάζας Σώματος είναι ένα σημαντικό χαρακτηριστικό στην ανάλυση δεδομένων που σχετίζονται με τον καρκίνο του μαστού, καθώς η παχυσαρκία έχει συνδεθεί με αυξημένο κίνδυνο εμφάνισης της νόσου. Οι γυναίκες με υψηλό BMI συχνά διατρέχουν μεγαλύτερο κίνδυνο λόγω των αλλαγών που μπορεί να προκαλέσει η παχυσαρκία στα ορμονικά επίπεδα και άλλους βιολογικούς μηχανισμούς που σχετίζονται με την ανάπτυξη καρκίνου.

Η αξιολόγηση του BMI επιτρέπει την ανάλυση της σχέσης μεταξύ του βάρους του σώματος και της πιθανότητας ανάπτυξης καρκίνου του μαστού, ενσωματώνοντας το ΔΜΣ ως παράγοντα κινδύνου στη μοντελοποίηση και την πρόβλεψη της νόσου. Η συσχέτιση αυτή μπορεί να προσφέρει πολύτιμα στοιχεία για την εξατομικευμένη εκτίμηση κινδύνου και την ανάπτυξη στρατηγικών πρόληψης ή θεραπείας.

Η χρήση του BMI στην ανάλυση μπορεί να βοηθήσει στην κατανόηση των επιπτώσεων της φυσικής κατάστασης και της διατροφής στην υγεία των γυναικών και μπορεί να ενισχύσει την ακριβή διάγνωση μέσω της ενσωμάτωσης του ως χαρακτηριστικό σε αλγορίθμους μηχανικής μάθησης. Με την καταγραφή ενός ευρέως φάσματος τιμών BMI, η μελέτη μπορεί να εξετάσει την επίδραση διαφορετικών επιπέδων βάρους στην πρόγνωση και την πρόληψη του καρκίνου του μαστού.



Σχήμα 2: Κατανομή Χαρακτηριστικού BMI

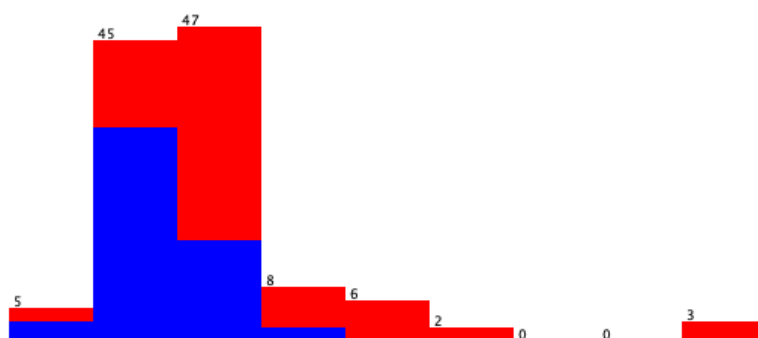
## GLUCOSE – Γλυκόζη

Το χαρακτηριστικό **Glucose** (γλυκόζη) αναφέρεται στην ποσότητα της γλυκόζης που ανιχνεύεται στο αίμα κατά τη διάρκεια μιας ιατρικής εξέτασης. Η γλυκόζη είναι ένα κρίσιμο συστατικό της διατροφής που παρέχει ενέργεια για τον οργανισμό και η συγκέντρωσή της στο αίμα είναι ένας σημαντικός δείκτης της μεταβολικής υγείας. Στη συγκεκριμένη μελέτη, το χαρακτηριστικό Glucose καταγράφεται με τιμές που κυμαίνονται από 60 έως 201 mg/dL.

Η παρακολούθηση των επιπέδων γλυκόζης είναι ουσιώδης για την κατανόηση της σχέσης μεταξύ της μεταβολικής υγείας και του κινδύνου εμφάνισης καρκίνου του μαστού. Υψηλά επίπεδα γλυκόζης στο αίμα, τα οποία μπορούν να προκύψουν από καταστάσεις όπως ο διαβήτης ή η αντίσταση στην ινσουλίνη, έχουν συσχετιστεί με αυξημένο κίνδυνο για διάφορες ασθένειες, περιλαμβανομένου του καρκίνου του μαστού. Η γλυκόζη επηρεάζει τη λειτουργία των κυττάρων και μπορεί να επηρεάσει την ανάπτυξη και εξέλιξη όγκων μέσω βιοχημικών μονοπατιών που σχετίζονται με την ενέργεια και την ανάπτυξη κυττάρων.

Η καταγραφή των επιπέδων γλυκόζης ως χαρακτηριστικού επιτρέπει την ανάλυση της συσχέτισης μεταξύ των μεταβολικών παραμέτρων και του κινδύνου εμφάνισης καρκίνου του μαστού. Η ενσωμάτωσή της ως παράγοντα στις μοντέλα πρόβλεψης μπορεί να προσφέρει πολύτιμα στοιχεία για την αξιολόγηση των κινδύνων και την ανάπτυξη εξατομικευμένων στρατηγικών πρόληψης και θεραπείας. Ειδικότερα, τα επίπεδα γλυκόζης μπορεί να χρησιμοποιηθούν σε συνδυασμό με άλλα χαρακτηριστικά, όπως η ηλικία και ο Δείκτης Μάζας Σώματος (BMI), για τη δημιουργία ακριβών προγνωστικών μοντέλων που να βοηθήσουν στην πρόληψη και διαχείριση του καρκίνου του μαστού.

Η αξιολόγηση των επιπέδων γλυκόζης, λαμβάνοντας υπόψη τις τιμές από 60 έως 201 mg/dL, επιτρέπει την ανάλυση της σχέσης μεταξύ της μεταβολικής κατάστασης των ατόμων και των πιθανών κινδύνων για την υγεία, προσδιορίζοντας τυχόν μοτίβα ή ανωμαλίες που μπορεί να σχετίζονται με την ανάπτυξη καρκίνου του μαστού.



Σχήμα 3 : Κατανομή του χαρακτηριστικού Glucose

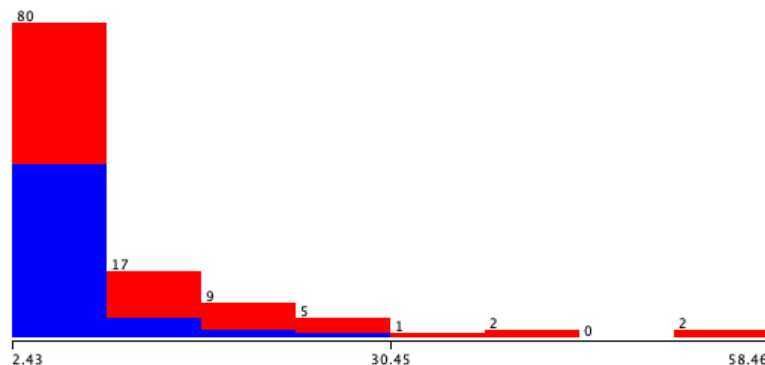
## INSULIN - Ινσουλίνη

Το χαρακτηριστικό **Insulin** αναφέρεται στην ποσότητα της ινσουλίνης που ανιχνεύεται στο αίμα των δειγμάτων κατά τη διάρκεια μιας ιατρικής εξέτασης. Η ινσουλίνη είναι μια ορμόνη που παράγεται από το πάγκρεας και είναι κρίσιμη για τη ρύθμιση των επιπέδων γλυκόζης στο αίμα. Στη συγκεκριμένη μελέτη, το χαρακτηριστικό Insulin καταγράφεται με τιμές που κυμαίνονται από 2.43 έως 58.46  $\mu\text{IU/mL}$ .

Η συγκέντρωση ινσουλίνης στο αίμα είναι σημαντικός δείκτης για την εκτίμηση της μεταβολικής υγείας και μπορεί να παρέχει πολύτιμες πληροφορίες σχετικά με την παρουσία διαταραχών όπως η αντίσταση στην ινσουλίνη ή ο διαβήτης τύπου 2. Υψηλά επίπεδα ινσουλίνης, γνωστά και ως υπερινσουλιαιμία, μπορεί να συνδέονται με διάφορες μεταβολικές παθήσεις και έχουν αναδειχθεί ως παράγοντες κινδύνου για την ανάπτυξη καρκίνου του μαστού.

Η αξιολόγηση των επιπέδων ινσουλίνης μπορεί να προσφέρει ενδείξεις για το πώς οι μεταβολικές διαταραχές επηρεάζουν τον κίνδυνο εμφάνισης καρκίνου του μαστού. Για παράδειγμα, η αντίσταση στην ινσουλίνη συχνά συνοδεύεται από υψηλά επίπεδα ινσουλίνης και έχει συνδεθεί με αυξημένο κίνδυνο ανάπτυξης καρκίνου, περιλαμβανομένου του καρκίνου του μαστού. Η συμπερίληψη του χαρακτηριστικού Insulin στην ανάλυση δεδομένων μπορεί να ενισχύσει την ικανότητα των μοντέλων πρόβλεψης να αναγνωρίσουν πιθανούς κινδύνους και να βελτιώσουν τη διάγνωση και τη διαχείριση της ασθένειας.

Με τη μέτρηση των επιπέδων ινσουλίνης από 2.43 έως 58.46  $\mu\text{IU/mL}$ , η μελέτη μπορεί να αναλύσει τη σχέση μεταξύ της συγκέντρωσης ινσουλίνης και της πιθανότητας εμφάνισης καρκίνου του μαστού. Τα αποτελέσματα αυτής της ανάλυσης μπορούν να προσφέρουν ενδείξεις για την ανάπτυξη στρατηγικών πρόληψης και θεραπείας, καθώς η ρύθμιση των επιπέδων ινσουλίνης μπορεί να είναι κρίσιμη για τη μείωση του κινδύνου καρκίνου.



## HOMA

Το χαρακτηριστικό **HOMA** (Homeostasis Model Assessment) χρησιμοποιείται για την εκτίμηση της αντίστασης στην ινσουλίνη, η οποία είναι ένας δείκτης της ικανότητας του σώματος να χρησιμοποιεί την ινσουλίνη αποτελεσματικά. Η αντίσταση στην ινσουλίνη αναφέρεται στην κατάσταση κατά την οποία οι κυτταρικοί υποδοχείς για την ινσουλίνη δεν ανταποκρίνονται αποτελεσματικά στην ορμόνη, αναγκάζοντας το πάγκρεας να παράγει περισσότερη ινσουλίνη για να διατηρήσει τα επίπεδα γλυκόζης στο αίμα υπό έλεγχο.

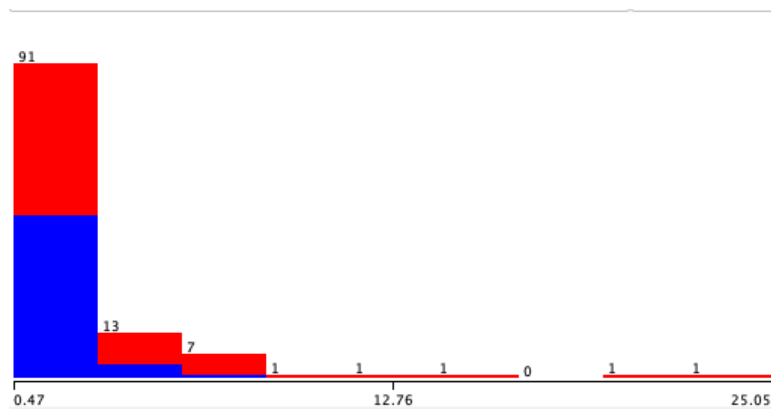
Ο υπολογισμός του HOMA πραγματοποιείται χρησιμοποιώντας την εξής εξίσωση:

$$\text{fasting glucose (mg/dL)} * \text{fasting insulin (mU/L)} / 405$$

Το χαρακτηριστικό HOMA εκτιμά πόση ινσουλίνη χρειάζεται το σώμα για να διατηρήσει υπό έλεγχο τα επίπεδα σακχάρου στο αίμα, παρέχοντας έτσι μια μέτρηση της αντίστασης στην ινσουλίνη. Στη συγκεκριμένη μελέτη, οι τιμές του HOMA κυμαίνονται από 0.47 έως 25.05, καταγράφοντας ένα ευρύ φάσμα αντίστασης στην ινσουλίνη, από πολύ χαμηλή έως υψηλή.

Η μέτρηση του HOMA είναι ιδιαίτερα χρήσιμη για την κατανόηση της σχέσης μεταξύ της αντίστασης στην ινσουλίνη και της ανάπτυξης καρκίνου του μαστού. Αυξημένα επίπεδα αντίστασης στην ινσουλίνη συνδέονται συχνά με μεταβολικές διαταραχές και έχουν συσχετιστεί με αυξημένο κίνδυνο ανάπτυξης καρκίνου, περιλαμβανομένου του καρκίνου του μαστού. Η ανάλυση των επιπέδων HOMA μπορεί να προσφέρει σημαντικά δεδομένα για την εκτίμηση του κινδύνου και την ανάπτυξη στρατηγικών πρόληψης και θεραπείας.

Με τη χρήση του HOMA, η μελέτη μπορεί να αξιολογήσει πώς η αντίσταση στην ινσουλίνη επηρεάζει τον κίνδυνο καρκίνου του μαστού και να διερευνήσει τυχόν σχέσεις μεταξύ των επιπέδων HOMA και των χαρακτηριστικών της νόσου. Η ενσωμάτωσή του στην ανάλυση δεδομένων ενισχύει την ικανότητα για ακριβή διάγνωση και μπορεί να βοηθήσει στην κατανόηση των μεταβολικών παραγόντων που συμβάλλουν στην ανάπτυξη του καρκίνου του μαστού.



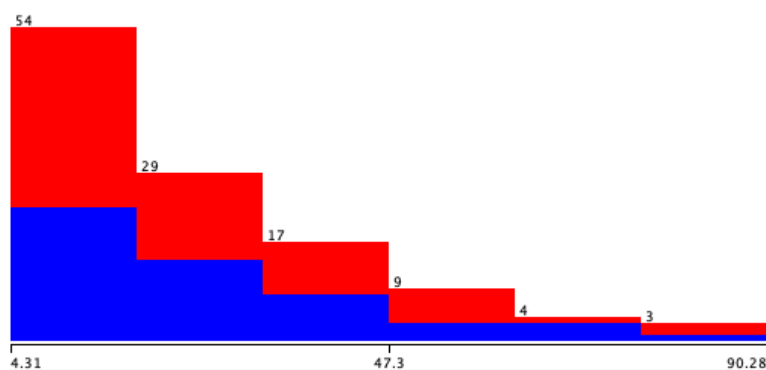
## LEPTIN - Λεπτίνη

Το χαρακτηριστικό **Leptin** αναφέρεται στην ορμόνη λεπτίνη, η οποία είναι μια πρωτεϊνική ορμόνη που παίζει κρίσιμο ρόλο στη ρύθμιση της ενέργειας και της μεταβολικής ισορροπίας του οργανισμού. Η λεπτίνη παράγεται κυρίως από τα λιποκύτταρα (τα κύτταρα του λίπους) και έχει βασικές λειτουργίες στη ρύθμιση της ενεργειακής πρόσληψης, του μεταβολισμού και της όρεξης. Ειδικότερα, η λεπτίνη ενημερώνει τον εγκέφαλο για την ποσότητα του αποθηκευμένου λίπους και επηρεάζει την αίσθηση της πείνας και της ικανοποίησης.

Το χαρακτηριστικό Leptin μετρά την ποσότητα λεπτίνης στο αίμα και οι τιμές του κυμαίνονται από 4.31 έως 90.28 ng/mL στη συγκεκριμένη μελέτη. Αυτή η κλίμακα τιμών καλύπτει ένα ευρύ φάσμα συγκεντρώσεων της ορμόνης, από χαμηλές έως πολύ υψηλές τιμές.

Η λεπτίνη διαδραματίζει σημαντικό ρόλο στη ρύθμιση της όρεξης και του σωματικού βάρους, και έχει επίσης συνδεθεί με μεταβολικές διαταραχές και παθήσεις, όπως η παχυσαρκία. Υψηλά επίπεδα λεπτίνης, τα οποία μπορεί να προκύψουν λόγω της αντίστασης στην λεπτίνη ή της υπερβολικής συσσώρευσης λίπους, έχουν συσχετιστεί με αυξημένο κίνδυνο εμφάνισης καρκίνου, περιλαμβανομένου του καρκίνου του μαστού. Η λεπτίνη μπορεί να επηρεάσει την ανάπτυξη των καρκινικών κυττάρων μέσω της ρύθμισης των ορμονών και των μεταβολικών διαδικασιών που σχετίζονται με την ανάπτυξη όγκων.

Η ανάλυση των επιπέδων λεπτίνης μπορεί να προσφέρει πολύτιμα δεδομένα για την κατανόηση της σχέσης μεταξύ της μεταβολικής κατάστασης και του κινδύνου καρκίνου. Η ενσωμάτωσή του ως χαρακτηριστικό σε αλγορίθμους μηχανικής μάθησης και μοντέλα πρόβλεψης μπορεί να ενισχύσει την ακρίβεια της διάγνωσης και της πρόβλεψης της νόσου. Ειδικότερα, η αξιολόγηση των επιπέδων λεπτίνης μπορεί να βοηθήσει στη διάγνωση και τη διαχείριση της παχυσαρκίας και των σχετικών κινδύνων για την υγεία, παρέχοντας επιπλέον ενδείξεις για την πρόωπη ανίχνευση και πρόληψη του καρκίνου του μαστού.



Σχήμα 6: Κατανομή χαρακτηριστικού Leptin

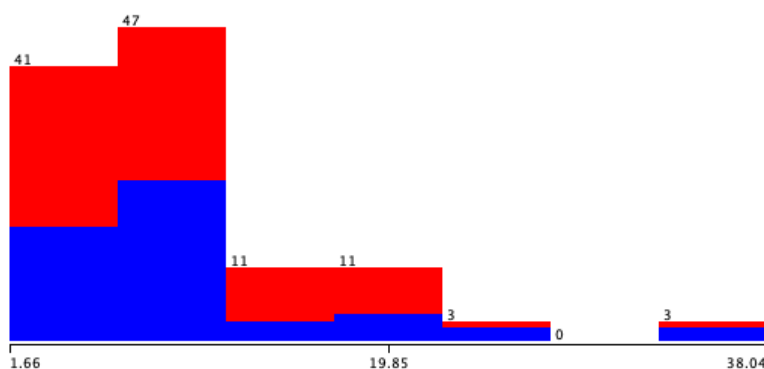
## ADIPONECTIN - Αντιπονεκτίνη

Το χαρακτηριστικό **Adiponectin** (Αντιπονεκτίνη) αναφέρεται σε μια πολυπεπτιδική ορμόνη που παράγεται από τον λιπώδη ιστό (τα λιποκύτταρα). Η αντιπονεκτίνη παίζει κρίσιμο ρόλο στη ρύθμιση του μεταβολισμού της γλυκόζης, των λιπιδίων, καθώς και στην ενίσχυση της ευαισθησίας στην ινσουλίνη. Έχει επίσης αναγνωριστεί για τις αντιφλεγμονώδεις και καρδιοπροστατευτικές της ιδιότητες.

Στη μελέτη, η αντιπονεκτίνη καταγράφεται με τιμές που κυμαίνονται από 1.66 έως 38.04  $\mu\text{g/mL}$ . Αυτό το εύρος τιμών καλύπτει μια ποικιλία συγκεντρώσεων της ορμόνης, από σχετικά χαμηλές έως υψηλές τιμές.

Η αντιπονεκτίνη συνδέεται στενά με τη μεταβολική υγεία και έχει αποδειχθεί ότι τα χαμηλά επίπεδα της ορμόνης σχετίζονται με αυξημένο κίνδυνο εμφάνισης μεταβολικών διαταραχών όπως η παχυσαρκία, ο διαβήτης τύπου 2, και άλλες καρδιαγγειακές παθήσεις. Σε σχέση με τον καρκίνο του μαστού, η αντιπονεκτίνη έχει επίσης εξεταστεί ως πιθανός παράγοντας κινδύνου. Τα χαμηλά επίπεδα αντιπονεκτίνης μπορεί να συμβάλλουν στην ανάπτυξη και εξέλιξη καρκινικών κυττάρων, ενώ υψηλότερα επίπεδα μπορεί να προσφέρουν προστατευτικές επιδράσεις λόγω των αντιφλεγμονωδών και μεταβολικών τους δράσεων.

Η αξιολόγηση της αντιπονεκτίνης ως χαρακτηριστικού στις αναλύσεις δεδομένων μπορεί να παρέχει σημαντικές πληροφορίες για την κατανόηση της σχέσης μεταξύ του μεταβολικού προφίλ και του κινδύνου καρκίνου του μαστού. Η ενσωμάτωσή της σε μοντέλα πρόβλεψης και αναλυτικά εργαλεία μπορεί να βοηθήσει στην εκτίμηση της πιθανότητας εμφάνισης της νόσου και στη διαμόρφωση στρατηγικών πρόληψης και διαχείρισης της υγείας. Τα δεδομένα σχετικά με την αντιπονεκτίνη μπορούν να ενισχύσουν την κατανόηση της επίδρασης των μεταβολικών παραγόντων στην ανάπτυξη καρκίνου του μαστού και να συμβάλουν στη βελτίωση της διάγνωσης και της πρόβλεψης της νόσου.



Σχήμα 7: Κατανομή χαρακτηριστικού Adiponectin

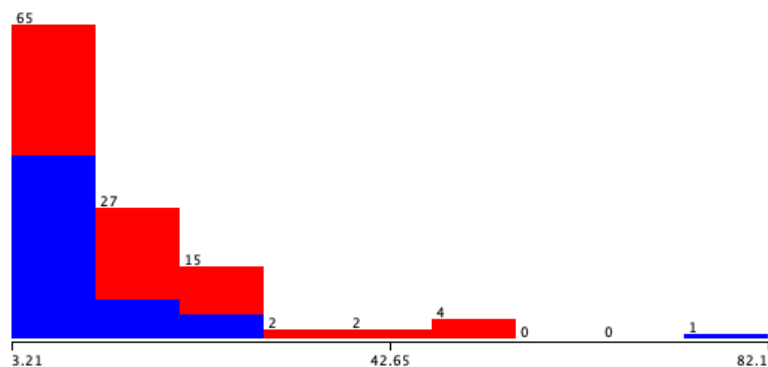
## RESISTIN – Ρεισιστίνη

Το χαρακτηριστικό **Resistin** (Ρεισιστίνη) αναφέρεται σε μια ορμόνη που εκκρίνεται από το λιπώδη ιστό (τα λιποκύτταρα). Η ρεισιστίνη είναι γνωστή για τον ρόλο της στην αντίσταση στην ινσουλίνη και τη ρύθμιση του μεταβολισμού της γλυκόζης. Ειδικότερα, η ρεισιστίνη έχει συνδεθεί με την παχυσαρκία και τις σχετικές μεταβολικές διαταραχές, όπως ο διαβήτης τύπου 2, και μπορεί να επηρεάσει την ανάπτυξη και τη λειτουργία των κυττάρων του σώματος μέσω των επιδράσεών της στη φλεγμονή και τη μεταβολική ισορροπία.

Στη συγκεκριμένη μελέτη, η ρεισιστίνη καταγράφεται με τιμές που κυμαίνονται από 3.21 έως 82.1 ng/mL. Αυτό το εύρος καλύπτει ένα ευρύ φάσμα συγκεντρώσεων της ορμόνης, από χαμηλές έως υψηλές τιμές.

Η ρεισιστίνη έχει αναγνωριστεί για τις επιδράσεις της στην αντίσταση στην ινσουλίνη και τις φλεγμονώδεις διεργασίες, γεγονός που την καθιστά σημαντική στην αξιολόγηση μεταβολικών διαταραχών και ασθενειών όπως η παχυσαρκία και ο διαβήτης τύπου 2. Επιπλέον, η ρεισιστίνη έχει μελετηθεί για τις πιθανές σχέσεις της με τον καρκίνο, καθώς η φλεγμονή και οι μεταβολικές ανωμαλίες μπορούν να επηρεάσουν την ανάπτυξη καρκινικών κυττάρων. Υψηλά επίπεδα ρεισιστίνης έχουν συσχετιστεί με αυξημένο κίνδυνο εμφάνισης καρκίνου, περιλαμβανομένου του καρκίνου του μαστού.

Η αξιολόγηση των επιπέδων ρεισιστίνης ως χαρακτηριστικού σε αναλύσεις δεδομένων μπορεί να παρέχει πολύτιμες ενδείξεις για την κατανόηση της σχέσης μεταξύ μεταβολικών παραγόντων και κινδύνου καρκίνου. Η ενσωμάτωσή της σε μοντέλα πρόβλεψης και αναλυτικά εργαλεία μπορεί να βοηθήσει στην εκτίμηση του κινδύνου και στην ανάπτυξη στρατηγικών πρόληψης και διαχείρισης. Μελετώντας τις συγκεντρώσεις ρεισιστίνης από 3.21 έως 82.1 ng/mL, οι ερευνητές μπορούν να αναλύσουν τη συσχέτιση μεταξύ της ορμόνης και της πιθανότητας ανάπτυξης καρκίνου, ενισχύοντας την κατανόηση των μεταβολικών παραγόντων που επηρεάζουν την υγεία.



Σχήμα 8 : Κατανομή του χαρακτηριστικού Resistin

## MCP1

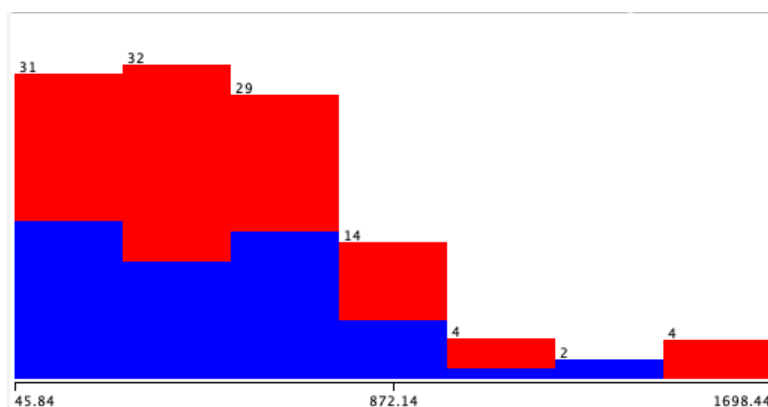
Το χαρακτηριστικό **MCP-1** (Monocyte Chemoattractant Protein-1), επίσης γνωστό ως **Chemokine (CC-motif) ligand 2 (CCL2)**, ανήκει στην οικογένεια των CC χημειοκινών και παίζει κρίσιμο ρόλο στη διαδικασία της φλεγμονής. Η MCP-1 είναι μια πρωτεΐνη που δρα ως χημειοκίνη, προσελκύοντας μονοκύτταρα και άλλα φλεγμονώδη κύτταρα στον τόπο της φλεγμονής, ενισχύοντας την ενδογενή φλεγμονώδη αντίδραση.

Η MCP-1 δρα ως σημαντικός παράγοντας για την έναρξη και την επιδείνωση της φλεγμονώδους απόκρισης, διότι προάγει την εισροή και τη συσσώρευση φλεγμονωδών κυττάρων όπως τα μονοκύτταρα στον ιστό που επηρεάζεται. Αυτή η διαδικασία ενδέχεται να έχει σημαντικές επιπτώσεις σε πολλές παθολογικές καταστάσεις, περιλαμβανομένων των χρόνιων φλεγμονωδών νόσων και των καρκινικών καταστάσεων.

Στη συγκεκριμένη μελέτη, το χαρακτηριστικό MCP-1 καταγράφεται με τιμές που κυμαίνονται από 45.843 έως 1698.44 pg/mL. Αυτό το εύρος τιμών καλύπτει ένα ευρύ φάσμα συγκεντρώσεων της MCP-1, από χαμηλά έως πολύ υψηλά επίπεδα.

Η ανάλυση των επιπέδων MCP-1 είναι ιδιαίτερα χρήσιμη για την κατανόηση της σχέσης μεταξύ της φλεγμονής και της ανάπτυξης καρκίνου. Υψηλά επίπεδα MCP-1 συνδέονται συχνά με αυξημένο κίνδυνο ανάπτυξης φλεγμονωδών νόσων και καρκίνου, συμπεριλαμβανομένου του καρκίνου του μαστού, καθώς η φλεγμονή μπορεί να συμβάλλει στην ανάπτυξη και την εξέλιξη των καρκινικών κυττάρων. Η αξιολόγηση των επιπέδων MCP-1 μπορεί να παρέχει σημαντικές ενδείξεις για την εκτίμηση της φλεγμονώδους κατάστασης και της πιθανότητας εμφάνισης καρκινικών ή άλλων παθολογικών καταστάσεων.

Επιπλέον, η συμπερίληψη του MCP-1 στην ανάλυση των δεδομένων ενισχύει την κατανόηση των μηχανισμών που συνδέουν τη φλεγμονώδη αντίδραση με την ανάπτυξη καρκίνου και άλλων χρόνιων ασθενειών. Τα δεδομένα σχετικά με την MCP-1, με τιμές από 45.843 έως 1698.44 pg/mL, μπορούν να συμβάλουν στη βελτίωση της διάγνωσης και της πρόβλεψης της νόσου, παρέχοντας ενδείξεις για την πρόληψη και τη διαχείριση φλεγμονωδών και καρκινικών καταστάσεων.



Σχήμα 9: Κατανομή του χαρακτηριστικού MCP1



## CLASSIFICATION – Κλάση Ταξινόμησης

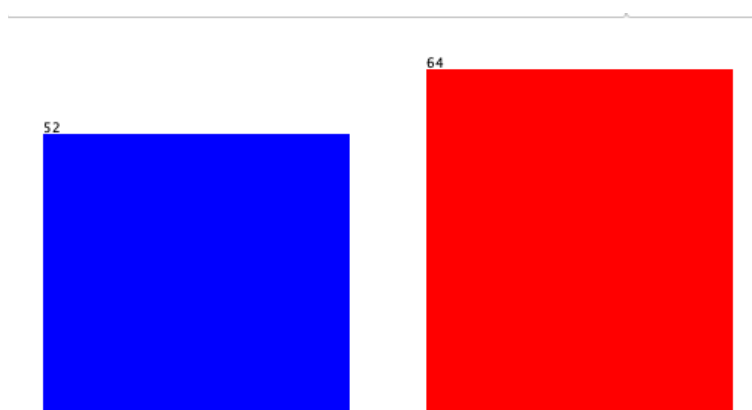
Το χαρακτηριστικό **Class (Κλάση Ταξινόμησης)** αναφέρεται στην κατηγορία ή στην κλάση που προσπαθούμε να προβλέψουμε με βάση τα δεδομένα που έχουμε στη διάθεσή μας. Στη συγκεκριμένη μελέτη, η κλάση ταξινόμησης είναι η τελική μεταβλητή που καθορίζει αν τα δεδομένα αφορούν υγιή άτομα ή ασθενείς.

Η τιμή της κλάσης ταξινόμησης ορίζεται ως εξής:

- **1:** Αντιπροσωπεύει υγιή άτομα. Οι μετρήσεις που έχουν την τιμή 1 υποδηλώνουν ότι το δείγμα προέρχεται από άτομο χωρίς την ασθένεια ή το διαγνωστικό πρόβλημα που μελετάται.
- **2:** Αντιπροσωπεύει ασθενείς. Οι μετρήσεις που έχουν την τιμή 2 δείχνουν ότι το δείγμα προέρχεται από άτομο που έχει διαγνωστεί με την ασθένεια ή το πρόβλημα υγείας που εξετάζεται.

Στη μελέτη, συνολικά υπάρχουν 116 παραδείγματα δεδομένων, τα οποία έχουν ταξινομηθεί σε μία από αυτές τις δύο κλάσεις. Αυτό σημαίνει ότι τα δεδομένα περιλαμβάνουν τόσο υγιή άτομα όσο και άτομα με την ασθένεια ή το διαγνωστικό πρόβλημα. Η κλάση ταξινόμησης χρησιμοποιείται για την εκπαίδευση και την αξιολόγηση των αλγορίθμων μηχανικής μάθησης, με σκοπό την ακριβή διάκριση μεταξύ των υγιών ατόμων και των ασθενών.

Η πρόβλεψη της κλάσης ταξινόμησης είναι κρίσιμη για την ανάπτυξη ακριβών διαγνωστικών εργαλείων και την αξιολόγηση της υγείας των ατόμων. Τα δεδομένα κλάσης βοηθούν στην εκπαίδευση των μοντέλων μηχανικής μάθησης να κατανοήσουν και να αναγνωρίσουν τα χαρακτηριστικά που διακρίνουν τους υγιείς από τους ασθενείς, βελτιώνοντας έτσι την ακρίβεια των διαγνώσεων και των προβλέψεων σε ιατρικές εφαρμογές.

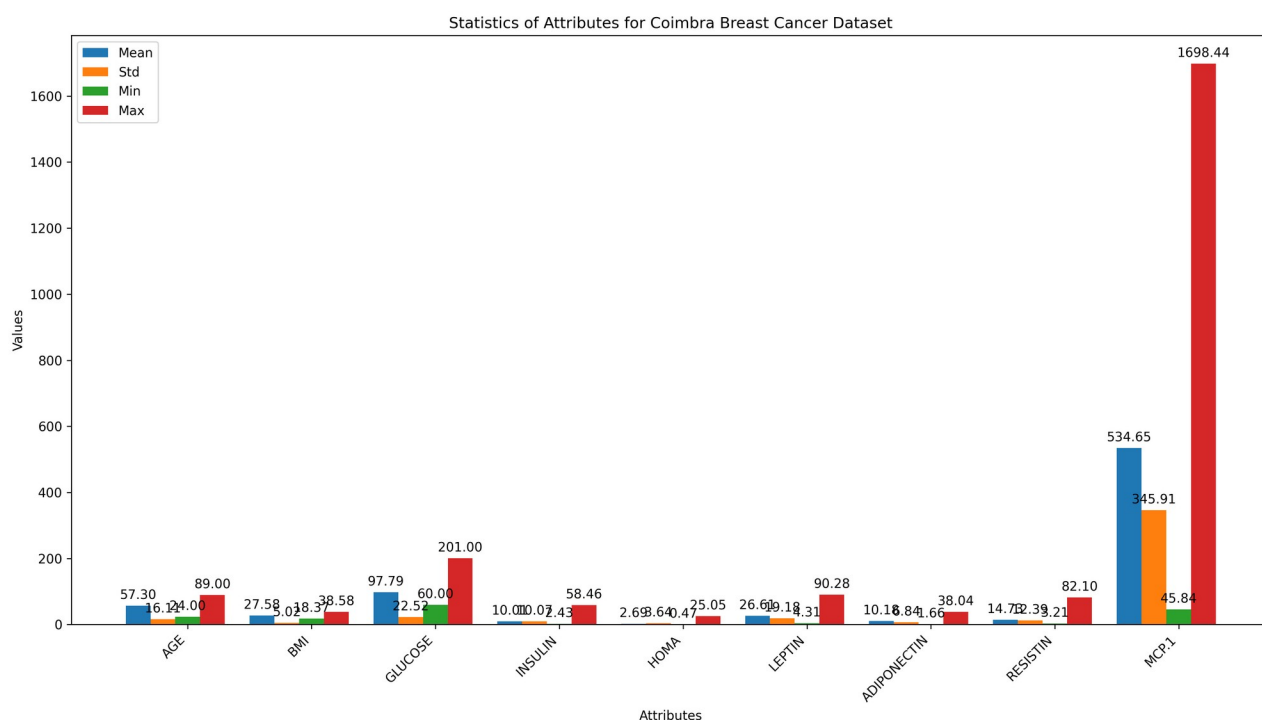


Σχήμα 10: Κατανομή χαρακτηριστικού class

Πίνακας 1: Περιγραφική ανάλυση των αριθμητικών χαρακτηριστικών

Attributes	mean	std	min	max
AGE	57.302	16.113	24	89

BMI	27.582	5.02	18.37	38.579
GLUCOSE	97.793	22.525	60	201
INSULIN	10.012	10.068	2.432	58.46
HOMA	2.695	3.642	0.467	25.05
LEPTIN	26.615	19.183	4.311	90.28
ADIPONECTIN	10.181	6.843	1.656	38.04
RESISTIN	14.726	12.391	3.21	82.1
MCP.1	534.647	345.913	45.843	1698.44
CLASS	Label1: 52	Label2: 64		Total: 116



Το γράφημα απεικονίζει τα στατιστικά στοιχεία για τα χαρακτηριστικά του σετ δεδομένων Coimbra Breast Cancer. Τα χαρακτηριστικά περιλαμβάνουν την ηλικία (AGE), δείκτη μάζας σώματος (BMI), επίπεδα γλυκόζης (GLUCOSE), ινσουλίνης (INSULIN), HOMA (HOMA), λεπτίνης (LEPTIN), αδιπονεκτίνης (ADIPONECTIN), αντοχής (RESISTIN) και MCP-1 (MCP.1). Στην ανάλυση αυτή, εξετάζουμε τη μέση τιμή, την τυπική απόκλιση, την ελάχιστη και τη μέγιστη τιμή για κάθε χαρακτηριστικό.

### 1. Ηλικία (AGE)

- **Μέση Τιμή:** 57.30 έτη
- **Τυπική Απόκλιση:** 16.11 έτη
- **Ελάχιστη Τιμή:** 24 έτη
- **Μέγιστη Τιμή:** 89 έτη

Η μέση ηλικία των ατόμων στο σετ δεδομένων είναι 57.30 έτη, με μια τυπική απόκλιση 16.11 έτη, υποδεικνύοντας ότι υπάρχει μια σημαντική ποικιλία στην ηλικία των ασθενών. Οι ηλικίες

κυμαίνονται από 24 έως 89 έτη, δείχνοντας ότι το δείγμα περιλαμβάνει άτομα σε ευρύ φάσμα ηλικιών. Η υψηλή τυπική απόκλιση υποδεικνύει ότι η ηλικία μπορεί να ποικίλει σημαντικά μεταξύ των ατόμων.

## 2. Δείκτης Μάζας Σώματος (BMI)

- **Μέση Τιμή:** 27.58 kg/m<sup>2</sup>
- **Τυπική Απόκλιση:** 5.02 kg/m<sup>2</sup>
- **Ελάχιστη Τιμή:** 18.37 kg/m<sup>2</sup>
- **Μέγιστη Τιμή:** 38.58 kg/m<sup>2</sup>

Ο μέσος δείκτης μάζας σώματος (BMI) είναι 27.58 kg/m<sup>2</sup>, ο οποίος είναι εντός της κατηγορίας υπέρβαρου, σύμφωνα με τα κριτήρια Παγκόσμιος Οργανισμός Υγείας (WHO). Η τυπική απόκλιση των 5.02 kg/m<sup>2</sup> υποδεικνύει μέτρια ποικιλία στον BMI, με τιμές που κυμαίνονται από 18.37 έως 38.58 kg/m<sup>2</sup>. Οι ακραίες τιμές δείχνουν ότι υπάρχουν και άτομα με χαμηλότερους και υψηλότερους BMI, γεγονός που μπορεί να επηρεάζει τον κίνδυνο καρκίνου του μαστού.

## 3. Επίπεδα Γλυκόζης (GLUCOSE)

- **Μέση Τιμή:** 97.79 mg/dL
- **Τυπική Απόκλιση:** 22.53 mg/dL
- **Ελάχιστη Τιμή:** 60 mg/dL
- **Μέγιστη Τιμή:** 201 mg/dL

Η μέση τιμή γλυκόζης είναι 97.79 mg/dL, κοντά στο κατώτερο όριο του φυσιολογικού εύρους, που είναι συνήθως κάτω από 100 mg/dL. Η τυπική απόκλιση των 22.53 mg/dL δείχνει σημαντική ποικιλία, με επίπεδα γλυκόζης που κυμαίνονται από 60 έως 201 mg/dL. Η ευρεία διακύμανση των τιμών υποδεικνύει την παρουσία ατόμων με χαμηλή και υψηλή γλυκόζη, γεγονός που μπορεί να συνδέεται με αυξημένο κίνδυνο καρκίνου.

## 4. Επίπεδα Ινσουλίνης (INSULIN)

- **Μέση Τιμή:** 10.01 μU/mL
- **Τυπική Απόκλιση:** 10.07 μU/mL
- **Ελάχιστη Τιμή:** 2.43 μU/mL
- **Μέγιστη Τιμή:** 58.46 μU/mL

Η μέση τιμή της ινσουλίνης είναι 10.01 μU/mL, αλλά η υψηλή τυπική απόκλιση των 10.07 μU/mL δείχνει μεγάλη ποικιλία στα επίπεδα ινσουλίνης, με τιμές που κυμαίνονται από 2.43 έως 58.46 μU/mL. Αυτή η μεγάλη ποικιλία μπορεί να υποδεικνύει διαφορετικούς βαθμούς αντίστασης στην ινσουλίνη, κάτι που σχετίζεται με αυξημένο κίνδυνο για καρκίνο του μαστού.

## 5. HOMA

- **Μέση Τιμή:** 2.70
- **Τυπική Απόκλιση:** 3.64
- **Ελάχιστη Τιμή:** 0.47
- **Μέγιστη Τιμή:** 25.05

Η μέση τιμή του δείκτη HOMA είναι 2.70, με υψηλή τυπική απόκλιση 3.64. Οι τιμές κυμαίνονται από 0.47 έως 25.05, δείχνοντας σημαντική ποικιλία στην αντίσταση στην ινσουλίνη. Υψηλότερες τιμές HOMA υποδεικνύουν αυξημένο κίνδυνο για καρκίνο του μαστού, καθώς συνδέονται με αντίσταση στην ινσουλίνη.

## 6. Επίπεδα Λεπτίνης (LEPTIN)

- **Μέση Τιμή:** 26.62 ng/mL
- **Τυπική Απόκλιση:** 19.18 ng/mL
- **Ελάχιστη Τιμή:** 4.31 ng/mL
- **Μέγιστη Τιμή:** 90.28 ng/mL

Η μέση τιμή της λεπτίνης είναι 26.62 ng/mL, με μεγάλη τυπική απόκλιση 19.18 ng/mL. Οι τιμές κυμαίνονται από 4.31 έως 90.28 ng/mL. Υψηλά επίπεδα λεπτίνης σχετίζονται με αυξημένο κίνδυνο για καρκίνο του μαστού, όπως φανερώνεται από την μεγάλη ποικιλία τιμών.

## 7. Αδιπονεκτίνη (ADIPONECTIN)

- **Μέση Τιμή:** 10.18 µg/mL
- **Τυπική Απόκλιση:** 6.84 µg/mL
- **Ελάχιστη Τιμή:** 1.66 µg/mL
- **Μέγιστη Τιμή:** 38.04 µg/mL

Η μέση τιμή της αδιπονεκτίνης είναι 10.18 µg/mL, με τυπική απόκλιση 6.84 µg/mL. Τα επίπεδα κυμαίνονται από 1.66 έως 38.04 µg/mL. Χαμηλότερα επίπεδα αδιπονεκτίνης συνδέονται με αυξημένο κίνδυνο για καρκίνο του μαστού, και η μεγάλη ποικιλία υποδηλώνει διαφορετικά επίπεδα κινδύνου.

## 8. Επίπεδα Αντοχής (RESISTIN)

- **Μέση Τιμή:** 14.73 ng/mL
- **Τυπική Απόκλιση:** 12.39 ng/mL
- **Ελάχιστη Τιμή:** 3.21 ng/mL
- **Μέγιστη Τιμή:** 82.10 ng/mL

Η μέση τιμή της αντοχής είναι 14.73 ng/mL με τυπική απόκλιση 12.39 ng/mL. Οι τιμές κυμαίνονται από 3.21 έως 82.10 ng/mL. Υψηλά επίπεδα αντοχής σχετίζονται με αυξημένο κίνδυνο καρκίνου του μαστού, όπως φαίνεται από την ευρεία διακύμανση των τιμών.

## 9. MCP-1

- **Μέση Τιμή:** 534.65 pg/mL
- **Τυπική Απόκλιση:** 345.91 pg/mL
- **Ελάχιστη Τιμή:** 45.84 pg/mL
- **Μέγιστη Τιμή:** 1698.44 pg/mL

Η μέση τιμή του MCP-1 είναι 534.65 pg/mL, με εξαιρετικά υψηλή τυπική απόκλιση 345.91 pg/mL. Οι τιμές κυμαίνονται από 45.84 έως 1698.44 pg/mL. Το MCP-1 είναι συνδεδεμένο με φλεγμονώδεις διεργασίες και υψηλά επίπεδα μπορούν να είναι δείκτες αυξημένου κινδύνου για καρκίνο.

## 1.6 Συμπεράσματα

- **Ποικιλία Χαρακτηριστικών:** Το γράφημα δείχνει σημαντική ποικιλία στα χαρακτηριστικά του σετ δεδομένων, με ιδιαίτερα ευρύ φάσμα τιμών για την ινσουλίνη, την αδιπονεκτίνη, τη λεπτίνη και το MCP-1. Αυτή η ποικιλία μπορεί να επηρεάζει την ακριβή πρόβλεψη του κινδύνου για καρκίνο του μαστού.

- **Σχέσεις με Κίνδυνο Καρκίνου:** Τα υψηλά επίπεδα ινσουλίνης, λεπτίνης, αντοχής και MCP-1 συσχετίζονται συνήθως με αυξημένο κίνδυνο για καρκίνο του μαστού, ενώ τα χαμηλά επίπεδα αδιπονεκτίνης μπορεί να είναι επίσης επιβαρυντικά. Οι αυξήσεις στις τιμές αυτών των δεικτών θα πρέπει να εξετάζονται με προσοχή.
- **Αναγνώριση Πρότυπων:** Η ανάλυση αυτών των στατιστικών στοιχείων μπορεί να βοηθήσει στην αναγνώριση προτύπων και στη διαμόρφωση στρατηγικών πρόληψης ή διαχείρισης του καρκίνου του μαστού.

Η λεπτομερής αυτή ανάλυση παρέχει μια καλή βάση για την κατανόηση των χαρακτηριστικών του σετ δεδομένων και της σύνδεσής τους με την πρόβλεψη του κινδύνου καρκίνου του μαστού.

## Κεφάλαιο 2 – Επεξεργασία στο Σύνολο Δεδομένων

### 2. Προεπεξεργασία δεδομένων (Data Preprocessing)

Data preprocessing αναφέρεται στην επεξεργασία του συνόλου δεδομένων πριν από την εφαρμογή του μοντέλου μηχανικής μάθησης. Η προεπεξεργασία των δεδομένων βοηθάει το μοντέλο στην βελτίωση της ποιότητας των δεδομένων καθώς αν λείπουν δεδομένα μειώνεται η αποτελεσματικότητα του μοντέλου. Επιπλέον αν αυτές οι τιμές που λείπουν (missing values) συμπληρωθούν ή απορριφθούν βοηθά στο να διασφαλιστεί ότι το μοντέλο μπορεί να εκπαιδευτεί σωστά. Ακόμη στην περίπτωση που υπάρχει θόρυβος στο σετ δεδομένων ο θόρυβος είναι δυνατόν να κρύψει τα υποκείμενα μοτίβα στα δεδομένα. Το φιλτράρισμα των ακραίων στοιχείων ή ασχέτων πληροφοριών μπορεί να βελτιώσει την απόδοση του μοντέλου. Επιπλέον τα χαρακτηριστικά μπορεί να έχουν διαφορετικές μονάδες μέτρησης ή κλίμακες. Η κανονικοποίηση ή η κλιμάκωση των χαρακτηριστικών διασφαλίζει ότι τα βάρη βρίσκονται σε συγκρίσιμη κλίμακα, κάτι που είναι απαραίτητο για αλγόριθμους που υπολογίζουν αποστάσεις μεταξύ σημείων δεδομένων, όπως είναι ο k-NN ή τα SVM. Επιπρόσθετα η κλιμάκωση μπορεί να βοηθήσει στην ταχύτητα σύγκλισης σε αλγόριθμους βελτιστοποίησης όπως ο gradient descent βελτιώνοντας σημαντικά την ταχύτητα των αποτελεσμάτων. Ακόμη μέσω της προεπεξεργασίας των δεδομένων δημιουργούνται νέες δυνατότητες όπως ο συνδυασμός δύο χαρακτηριστικών σε ένα κ.λπ. Επιπλέον η επιλογή χαρακτηριστικών μειώνει τις διαστάσεις επιλέγοντας τα πιο σχετικά χαρακτηριστικά σχετικά με την κλάση ταξινόμησης πράγμα που συμβάλλει στη βελτίωση της απόδοσης του μοντέλου και στη μείωση του υπολογιστικού κόστους. Η κωδικοποίηση κατηγορικών μεταβλητών επιτρέπει τους αλγόριθμους μηχανικής μάθησης να επεξεργάζονται αυτά τα δεδομένα πράγμα που χωρίς την κωδικοποίηση δεν γίνονταν. Ακόμη η κατανομή των δεδομένων παίζει σημαντικό ρόλο στα αποτελέσματα του μοντέλου. Για παράδειγμα ορισμένοι αλγόριθμοι υποθέτουν ότι τα δεδομένα ακολουθούν μια συγκεκριμένη κατανομή π.χ. Gaussian κατανομή, έτσι ο μετασχηματισμός των δεδομένων για την ικανοποίηση αυτών των παραδοχών μπορεί να βελτιώσει την απόδοση του μοντέλου. Ο διαχωρισμός των δεδομένων σε train-test είναι ζωτικής σημασίας για την αξιολόγηση της απόδοσης του μοντέλου και την αποφυγή υπερβολικής προσαρμογής. Επιπρόσθετα αν στα δεδομένα παρατηρείται πρόβλημα ανισορροπίας των τάξεων το μοντέλο μπορεί να γίνει μεροληπτικό. Τεχνικές όπως η υπερδειγματοληψία, η υποδειγματοληψία ή η χρήση συνθετικών δεδομένων (SMOTE) βοηθούν στην εξισορρόπηση των κλάσεων. Έτσι φτάνουμε στο συμπέρασμα πως η προεπεξεργασία διασφαλίζει ότι τα δεδομένα βρίσκονται στην καλύτερη δυνατή κατάσταση

ώστε να προχωρήσουμε στο επόμενο στάδιο της μηχανικής μάθησης. Χωρίς την κατάλληλη προεπεξεργασία, ακόμη και οι πιο εξελιγμένοι αλγόριθμοι ενδέχεται να έχουν κακή απόδοση.

## 2.1 Ετικετοποίηση (Labeling)

Το labeling, δηλαδή η διαδικασία ανάθεσης ετικετών ή κατηγοριών στα δεδομένα, είναι ένα κρίσιμο βήμα στην αντιμετώπιση προβλημάτων μηχανικής μάθησης, ιδιαίτερα σε προβλήματα ταξινόμησης όπως αυτό του σετ δεδομένων Coimbra Breast Cancer. Αναλυτικά, το labeling συμβάλλει στα εξής:

1. **Διακριτότητα των Κλάσεων:** Με τη χρήση των ετικετών "Healthy" και "Unhealthy", τα δεδομένα κατηγοριοποιούνται σε δύο σαφώς διακριτές κλάσεις. Αυτό βοηθά το μοντέλο να κατανοήσει τη διαφορά μεταξύ των υγιών και των ασθενών ατόμων και να μάθει τα χαρακτηριστικά που διαφοροποιούν τις δύο ομάδες.
2. **Βελτίωση της Κατανόησης και Ερμηνείας:** Η αντικατάσταση των αριθμητικών τιμών (1, 2) με περιγραφικές ετικέτες ("Healthy", "Unhealthy") βελτιώνει την κατανόηση και ερμηνεία των δεδομένων τόσο για τους ανθρώπους όσο και για τα μηχανικά συστήματα. Οι περιγραφικές ετικέτες παρέχουν μια πιο σαφή εικόνα του τι αντιπροσωπεύει κάθε κλάση, διευκολύνοντας έτσι την ανάλυση και την αξιολόγηση των αποτελεσμάτων του μοντέλου.
3. **Αύξηση της Ακρίβειας:** Καθώς το μοντέλο μαθαίνει να συσχετίζει συγκεκριμένα χαρακτηριστικά με τις ετικέτες "Healthy" και "Unhealthy", μπορεί να βελτιωθεί η ακρίβεια της πρόβλεψης. Το labeling διασφαλίζει ότι το μοντέλο εκπαιδεύεται με βάση τις πραγματικές κλάσεις, επιτρέποντάς του να αναγνωρίζει μοτίβα και συσχετίσεις που σχετίζονται με την κατάσταση της υγείας των ασθενών.
4. **Επιτρέπει τη Χρήση Τεχνικών Μηχανικής Μάθησης:** Οι περισσότερες τεχνικές μηχανικής μάθησης για ταξινόμηση απαιτούν ετικέτες για την επίβλεψη της διαδικασίας εκπαίδευσης. Το labeling μετατρέπει το πρόβλημα σε ένα εποπτευόμενο πρόβλημα μάθησης, όπου το μοντέλο προσπαθεί να μάθει τη συσχέτιση μεταξύ χαρακτηριστικών και προκαθορισμένων ετικετών για να προβλέψει σωστά την κατηγορία νέων, άγνωστων δεδομένων.

## 2.2 Προεπεξεργασία στο σετ δεδομένων Coimbra Breast Cancer. (Labeling)

Πραγματοποιείται επεξεργασία στο σετ δεδομένων Coimbra Breast Cancer το οποίο φορτώνεται στην python μέσω ενός αρχείου csv. Το πρώτο βήμα είναι η φόρτωση των δεδομένων στο DataFrame μέσω της βιβλιοθήκης pandas. Στην συνέχεια, η στήλη Classification που περιέχει τις ετικέτες κλάσης για την υγεία των ασθενών, μετατρέπεται από αριθμητικές τιμές σε κατηγορίες κειμένου (labels) όπου το '1' αντιστοιχεί στο 'Healthy' (Υγιής) και το '2' αντιστοιχεί στο 'Unhealthy' (Ασθενής). Αφού γίνει η αντιστοίχιση των ετικετών, τα δεδομένα προετοιμάζονται για περαιτέρω επεξεργασία. Οι ανεξάρτητες μεταβλητές δηλαδή τα χαρακτηριστικά αποθηκεύονται στη μεταβλητή 'x', ενώ οι ετικέτες των κλάσεων (στόχοι) αποθηκεύονται στη μεταβλητή 'y'. Αυτό το βήμα θα διαχωρίσει τα χαρακτηριστικά από την κλάση που θα χρησιμοποιηθεί αργότερα για την εκπαίδευση του μοντέλου.

## Κεφάλαιο 3 – Χρήση αλγορίθμων και Μεθόδων

### 3 Χτήσιμο Μοντέλου (Model Building)

Τώρα είναι σειρά να δημιουργήσουμε το καλύτερο μοντέλο για αυτό το συγκεκριμένο σύνολο δεδομένων. Το σετ δεδομένων έχει υποβληθεί σε labeling ώστε οι αλγόριθμοι που δοκιμάζονται να έχουν όσο το δυνατό καλύτερα αποτελέσματα. Έτσι εφαρμόζουμε τους αλγορίθμους στο σύνολο των δεδομένων και συγκρίνουμε τα αποτελέσματα του κάθε αλγορίθμου ώστε να αποφασίσουμε όσο το δυνατό το μοντέλο με την καλύτερη απόδοση. Δοκιμάζουμε συνολικά έξι αλγορίθμους παραμετροποιώντας του με διάφορες τιμές ώστε να παρατηρήσουμε τα αποτελέσματα των παραμέτρων στο βάθος του αλγορίθμου. Τελικά μέσω μελέτης αποφασίζουμε πιο είναι το αποδοτικότερο μοντέλο και το αποθηκεύουμε δημιουργώντας ένα web app για να μπορεί ο τελικός χρήστης να το χρησιμοποιήσει με ευκολία. Οι αλγόριθμοι που δοκιμάζονται είναι οι kNN, Naive Bayes, Gradient Boost, Random Forest, J48, AdaBoost.

#### 3.1 Διασταυρούμενη επικύρωση (Cross Validation)

Για την αξιολόγηση της απόδοσης του αλγορίθμου, χρησιμοποιήθηκε 10-fold cross validation. Η μέθοδος cross validation μειώνει τα προβλήματα υπερφόρτωσης και βελτιώνει την απόδοση γενίκευσης του μοντέλου. Επομένως, το cross validation είναι πιο αξιόπιστη από τη μέθοδο του διαχωρισμού ποσοστού και το 10 φορές είναι προτεινόμενος αριθμός από το μεγαλύτερο μέρος των ερευνητών. Το cross validation έχει τα ακόλουθα πλεονεκτήματα:

- Ο εκτιμητής cross validation έχει μικρότερη διακύμανση από έναν μεμονωμένο εκτιμητή συνόλου συνόλου αναμονής
- Τα αποτελέσματα cross validation είναι πιο αξιόπιστα από ένα μεμονωμένο σύνολο αποτελεσμάτων
- Μειώνει επίσης τις πιθανότητες υπερβολικής τοποθέτησης

#### 3.2 Αξιολόγηση επιδόσεων (Performance Evaluation)

Για να ελεγχθεί η αποτελεσματικότητα του μοντέλου στο σύνολο δεδομένων και να συγκριθούν τα αποτελέσματά τους, χρησιμοποιούνται διάφορες μετρήσεις αξιολόγησης. Εδώ. Έχουν ληφθεί υπόψη οι μετρήσεις αξιολόγησης Accuracy, Precision, Recall, F-Measure.

- **Accuracy:** Η ακρίβεια αναφέρεται στην ορθότητα του μοντέλου, η οποία υποδεικνύει πόσο ακριβής είναι η πρόβλεψη και μπορεί να υπολογιστεί χρησιμοποιώντας τον παρακάτω τύπο:  
$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Όπου TP = True Positive, TN = True Negative, FN = False Negative και FP = False Positive

- **Precision:** Η ακρίβεια απαντά στην ερώτηση “Από όλες τις περιπτώσεις που προβλέφθηκαν ως θετικές, πόσες ήταν πραγματικά θετικές”. Η ακρίβεια είναι ιδιαίτερα χρήσιμη σε κατάσταση όπου τα ψευδώς θετικά (εσφαλμένα προβλεπόμενα θετικά) είναι δαπανηρά ή ανεπιθύμητα. Βοηθά στην αξιολόγησης ικανότητας του μοντέλου να αποφεύγει ψευδώς θετικά σφάλματα. Υπολογίζεται χρησιμοποιώντας τον παρακάτω τύπο:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Όπου TP = True Positive και FP = False Positive

- **Recall:** Το recall, γνωστό και ως ευαισθησία ή true positive rate, είναι μια μετρική απόδοσης που χρησιμοποιείται στην ανάλυση δεδομένων και στη μηχανική μάθηση για την αξιολόγηση της αποτελεσματικότητας ενός μοντέλου ταξινόμησης. Το recall μετρά την ικανότητα του μοντέλου να εντοπίζει σωστά τις θετικές περιπτώσεις, δηλαδή πόσο καλά το μοντέλο αναγνωρίζει τα πραγματικά θετικά παραδείγματα από το σύνολο όλων των θετικών παραδειγμάτων που υπάρχουν στο dataset. Ο τύπος για το recall είναι:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Με TP = True Positive, FN = False Negative

- **F-Measure:** Η F-Measure, ή F1-Score, είναι μια μετρική που συνδυάζει την ακρίβεια (precision) και το recall για να προσφέρει μια συνολική εκτίμηση της απόδοσης ενός μοντέλου ταξινόμησης. Ο λόγος που χρησιμοποιούμε την F1-Score είναι ότι παρέχει ένα ισορροπημένο μέτρο όταν οι κατηγορίες είναι ανισόρροπες ή όταν η βελτίωση μιας μετρικής μπορεί να έρχεται σε βάρος της άλλης.

$$\text{F1 Score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

Με Precision = Ακρίβεια, Recall = ευαισθησία

Ο αρμονικός μέσος επιπλέον προσφέρει:

- Ισορροπία Ακρίβειας και Ευαισθησίας: Ο αρμονικός μέσος της ακρίβειας και της ευαισθησίας αποφεύγει τις υπερβολές που μπορεί να προκύψουν αν μία από αυτές τις μετρικές είναι εξαιρετικά υψηλή ενώ η άλλη είναι χαμηλή. Στην πράξη, αν το μοντέλο έχει πολύ υψηλή ακρίβεια αλλά χαμηλό recall (ή αντίστροφα), η F1-Score θα είναι χαμηλή, προωθώντας την ανάγκη για βελτίωση και στις δύο διαστάσεις.
- Ευαίσθητος σε Μικρές Διαφορές: Ο αρμονικός μέσος δεν επιτρέπει σε μια υψηλή τιμή (όπως σε έναν εξαιρετικό precision) να "κρύψει" μια χαμηλή τιμή της άλλης μετρικής (όπως recall). Έτσι, παρέχει μια πιο αυστηρή και ρεαλιστική εικόνα της συνολικής απόδοσης του μοντέλου.
- Ειδικά Χρήσιμος σε Ανισόρροπα Δεδομένα: Σε περιπτώσεις όπου οι κλάσεις είναι ανισόρροπες (δηλαδή, μία κλάση είναι πολύ λιγότερο κοινή από την άλλη), η ακρίβεια μόνη της μπορεί να είναι παραπλανητική. Η F1-Score παρέχει ένα καλύτερο μέτρο για την απόδοση του μοντέλου σε τέτοιες περιπτώσεις.

### 3.3 kNN – K: Πλησιέστερος γείτονας (k-Nearest Neighbor)

#### Δομή

Ο αλγόριθμος k-Nearest Neighbors (kNN) είναι ένας απλός αλλά ισχυρός αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Η βασική ιδέα πίσω από τον kNN είναι ότι τα δεδομένα που βρίσκονται κοντά το ένα στο άλλο σε έναν πολυδιάστατο χώρο είναι πιθανό να ανήκουν στην ίδια κατηγορία. Για ταξινόμηση, ο αλγόριθμος κρίνει την κατηγορία ενός νέου δείγματος βασισμένο στις κατηγορίες των k κοντινότερων γειτόνων του, όπου k είναι ένας θετικός ακέραιος αριθμός που καθορίζεται από τον χρήστη. Ο αλγόριθμος υπολογίζει την απόσταση μεταξύ του νέου δείγματος και όλων των άλλων δειγμάτων στο σύνολο δεδομένων,



συνήθως χρησιμοποιώντας μετρικές όπως η ευκλείδεια απόσταση ή η απόσταση Manhattan. Στη συνέχεια, η κατηγορία του νέου δείγματος καθορίζεται μέσω της πλειοψηφίας των κατηγοριών των  $k$  κοντινότερων γειτόνων, με την υπόθεση ότι οι γειτονικοί γείτονες τείνουν να ανήκουν στην ίδια κατηγορία. Η απλότητα και η ευελιξία του  $k$ NN τον καθιστούν δημοφιλή επιλογή για πολλές εφαρμογές μηχανικής μάθησης.

### Πλεονεκτήματα - Μειονεκτήματα

Ωστόσο, η απόδοση του αλγορίθμου  $k$ NN εξαρτάται σε μεγάλο βαθμό από την επιλογή του αριθμού των γειτόνων  $k$  και την απόσταση που χρησιμοποιείται. Εάν το  $k$  είναι πολύ μικρό, το μοντέλο μπορεί να είναι επιρρεπές σε θόρυβο και να επηρεάζεται αρνητικά από μεμονωμένα σημεία δεδομένων, κάτι που μπορεί να οδηγήσει σε υπερβολική προσαρμογή (overfitting). Από την άλλη πλευρά, αν το  $k$  είναι πολύ μεγάλο, το μοντέλο μπορεί να γίνει πολύ γενικό και να υπολογίζει μέσες τιμές που δεν αντανακλούν καλά την τοπική δομή των δεδομένων, με αποτέλεσμα να υποτιμά τα πιο σημαντικά τοπικά μοτίβα (underfitting). Επίσης, η επιλογή της μετρικής απόστασης μπορεί να επηρεάσει την απόδοση του αλγορίθμου. Ενώ η ευκλείδεια απόσταση είναι συνήθως η προτιμώμενη επιλογή, άλλες μετρικές όπως η απόσταση Manhattan ή η απόσταση Minkowski μπορεί να είναι πιο κατάλληλες για διαφορετικά σύνολα δεδομένων και προβλήματα.

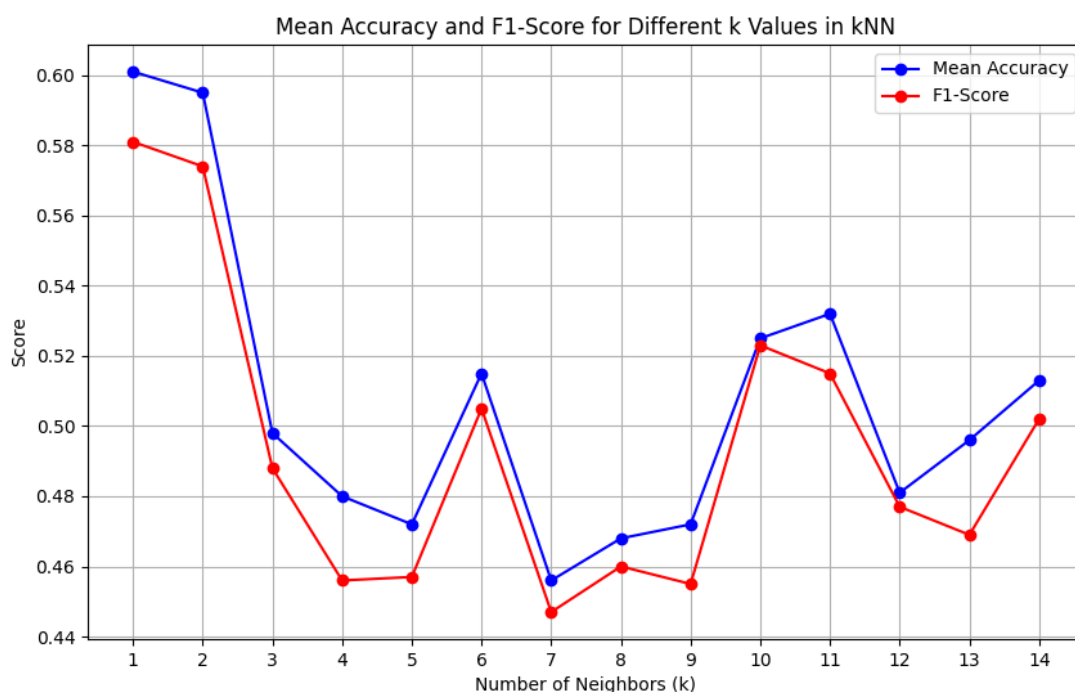
Επιπλέον, ο αλγόριθμος  $k$ NN μπορεί να είναι υπολογιστικά απαιτητικός, ειδικά όταν το σύνολο δεδομένων είναι μεγάλο, καθώς η υπολογιστική του πολυπλοκότητα είναι  $O(n)$  για την αναζήτηση των γειτόνων, όπου  $n$  είναι ο αριθμός των δεδομένων στο σύνολο. Αυτό σημαίνει ότι ο χρόνος υπολογισμού μπορεί να γίνει σημαντικός καθώς το μέγεθος του δεδομένου συνόλου αυξάνεται. Για την αντιμετώπιση αυτού του προβλήματος, μπορεί να χρησιμοποιηθούν δομές δεδομένων όπως τα δέντρα KD ή τα δέντρα Ball, οι οποίες βελτιώνουν την απόδοση της αναζήτησης των γειτόνων μειώνοντας τον χρόνο υπολογισμού. Παρά τις προκλήσεις του, ο  $k$ NN παραμένει ένα χρήσιμο εργαλείο λόγω της ευκολίας εφαρμογής του και της δυνατότητάς του να λειτουργεί καλά με μικρά και μεσαία μεγέθη συνόλων δεδομένων. Στο συγκεκριμένο παράδειγμα αποτελεί μια καλή επιλογή καθώς δεν το σετ δεδομένων είναι μικρό και δεν απαιτείται υπολογιστή πολυπλοκότητα.

### Τιμές του $k$ , Mean Accuracy και F-Score

Πίνακας 2: Παράμετροι του  $k$ NN και η επίδραση τους στο F1-Score και Mean Accuracy.

Parameter	Mean Accuracy	F-Score
$k=1$	0.601	0.581
$k=2$	0.595	0.574
$k=3$	0.498	0.488
$k=4$	0.480	0.456
$k=5$	0.472	0.457
$k=6$	0.515	0.505
$k=7$	0.456	0.447
$k=8$	0.468	0.460
$k=9$	0.472	0.455

k=10	0.525	0.523
k=11	0.532	0.515
k=12	0.481	0.477
k=13	0.496	0.469
k=14	0.513	0.502



Παρατηρούμε πως το μεγαλύτερο F-Score επιτυγχάνεται για  $k=1$  όπου F-Score = 0.581 και μέση ακρίβεια 0.601. Στο συγκεκριμένο παράδειγμα, η ανάλυση του kNN δείχνει ότι η επιλογή του αριθμού των γειτόνων ( $k$ ) έχει σημαντική επίδραση στις επιδόσεις του μοντέλου, με την ακρίβεια (Mean Accuracy) και το F1-Score να διαφέρουν ανάλογα με την τιμή του  $k$ .

Από τον Πίνακα 2, παρατηρούμε ότι το καλύτερο F1-Score παρατηρείται για  $k=1$ , όπου το F1-Score φτάνει το 0.581 και η μέση ακρίβεια είναι 0.601. Αυτό δείχνει ότι για τις δεδομένες παραμέτρους, η μικρότερη τιμή του  $k$  επιτυγχάνει την καλύτερη ισορροπία μεταξύ ευαισθησίας και ακρίβειας, επιτυγχάνοντας έτσι μια καλύτερη συνολική απόδοση σε σύγκριση με μεγαλύτερες τιμές του  $k$ . Ωστόσο, καθώς το  $k$  αυξάνεται, το F1-Score τείνει να μειώνεται, κάτι που υποδεικνύει ότι η

συμπερίληψη περισσότερων γειτόνων μπορεί να προκαλέσει απώλεια λεπτομέρειας στην κατηγοριοποίηση και ενδεχομένως να εισάγει θόρυβο στις προβλέψεις.

### 3.4 Δέντρο αποφάσης AdaBoost (AdaBoost Decision Tree)

Τα Δέντρα Αποφάσεων (Decision Trees) είναι ένα ισχυρό και ευρέως χρησιμοποιούμενο εργαλείο για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Ο βασικός στόχος ενός δέντρου αποφάσεων είναι να παρέχει έναν γραφικό και κατανοητό τρόπο για τη λήψη αποφάσεων ή προβλέψεων, βασισμένο σε ένα σύνολο κανόνων που προκύπτουν από τα δεδομένα εκπαίδευσης.

#### Δομή:

- **Ρίζα (Root Node):** Το αρχικό κόμβο του δέντρου, που αντιπροσωπεύει το σύνολο των δεδομένων εκπαίδευσης. Από τη ρίζα προκύπτουν υποκόμβοι με βάση την αξία χαρακτηριστικών.
- **Εσωτερικοί Κόμβοι (Internal Nodes):** Κόμβοι που περιέχουν αποφάσεις ή ερωτήσεις σχετικά με τα χαρακτηριστικά των δεδομένων. Αυτοί οι κόμβοι χωρίζουν τα δεδομένα σε υποσύνολα.
- **Φύλλα (Leaf Nodes):** Τελικοί κόμβοι του δέντρου που παρέχουν την τελική απόφαση ή πρόβλεψη. Κάθε φύλλο αντιπροσωπεύει μια συγκεκριμένη κλάση ή τιμή στόχου.
- **Ακμές (Edges):** Συνδέουν τους κόμβους του δέντρου και δείχνουν την κατεύθυνση της απόφασης που οδηγεί σε διαφορετικούς κόμβους.

#### Λειτουργία:

1. **Διάρθρωση (Splitting):** Ο αλγόριθμος αναλύει τα δεδομένα και επιλέγει χαρακτηριστικά για την εκτέλεση διαχωρισμού, ώστε να μειωθεί η αβεβαιότητα ή η αταξία (impurity) των δεδομένων. Δημοφιλή μέτρα διαχωρισμού περιλαμβάνουν την **εντροπία (Entropy)** και την **Gini αταξία (Gini Impurity)**.
2. **Σταμάτημα (Stopping Criteria):** Το δέντρο συνεχίζει να επεκτείνεται μέχρι να επιτευχθούν ορισμένα κριτήρια, όπως η μέγιστη βάθος του δέντρου, ο ελάχιστος αριθμός δείγματος σε ένα φύλλο, ή αν τα δείγματα σε ένα κόμβο ανήκουν στην ίδια κλάση.
3. **Προβλέψεις (Predictions):** Όταν γίνεται μια πρόβλεψη, το δέντρο ακολουθεί τις αποφάσεις ή ερωτήσεις που έχει μάθει, φτάνοντας σε ένα φύλλο που παρέχει την τελική κλάση ή τιμή.

#### 2. Πλεονεκτήματα και Μειονεκτήματα

##### Πλεονεκτήματα:

- **Κατανοητότητα:** Τα Δέντρα Αποφάσεων είναι εύκολα κατανοητά και ερμηνεύσιμα. Η δομή του δέντρου καθιστά τις αποφάσεις διαφανείς και κατανοητές.
- **Χωρίς Απαιτήσεις για Κανονικοποίηση:** Δεν απαιτούν κλιμάκωση ή κανονικοποίηση των χαρακτηριστικών δεδομένων.
- **Διαχείριση Κατηγοριών και Συνεχών Τιμών:** Μπορούν να χειριστούν τόσο κατηγορικές όσο και συνεχείς μεταβλητές.

## Μειονεκτήματα:

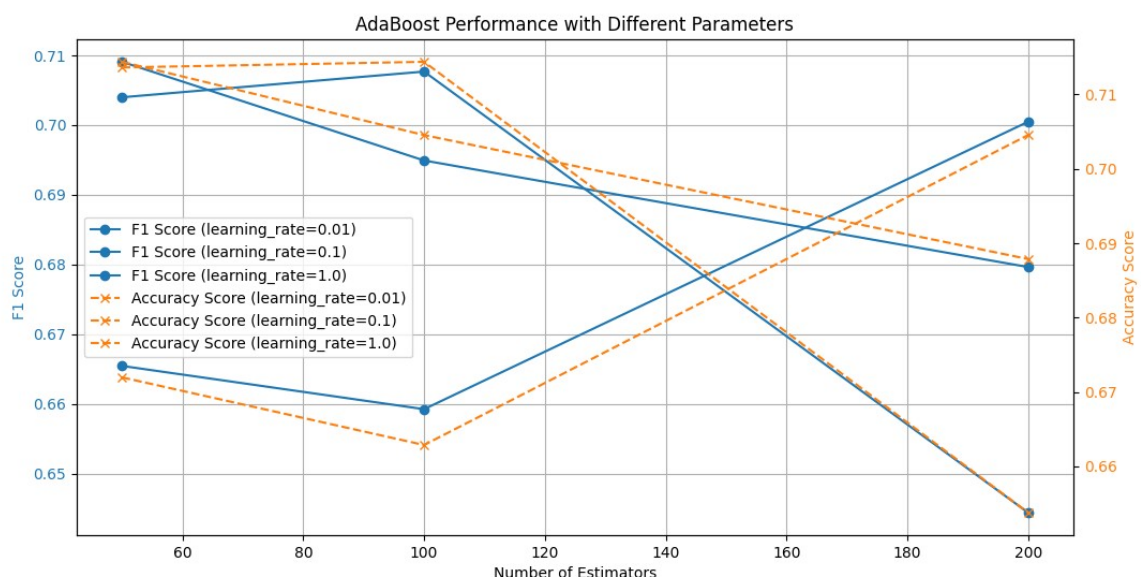
- **Υπερβολική Προσαρμογή (Overfitting):** Τα Δέντρα Αποφάσεων είναι επιρρεπή σε υπερβολική προσαρμογή, ειδικά αν δεν υπάρχουν περιορισμοί για το βάθος του δέντρου ή τον αριθμό των δειγμάτων σε κάθε φύλλο.
- **Ασταθές Ενδεχομένως:** Μικρές αλλαγές στα δεδομένα μπορεί να οδηγήσουν σε πολύ διαφορετικά δέντρα.
- **Υπολογιστική Πολυπλοκότητα:** Σε μεγάλα σύνολα δεδομένων με πολλές χαρακτηριστικές μεταβλητές, η κατασκευή του δέντρου μπορεί να γίνει υπολογιστικά απαιτητική.

## Τεχνικές Βελτίωσης

Για να αντιμετωπιστούν οι αδυναμίες των Δέντρων Αποφάσεων, έχουν αναπτυχθεί πολλές τεχνικές βελτίωσης:

- **Περιορισμός Βάθους Δέντρου (Tree Pruning):** Η μείωση του βάθους του δέντρου μπορεί να βοηθήσει στη μείωση της υπερβολικής προσαρμογής, βελτιώνοντας τη γενική απόδοση.
- **Τυχαία Δάση (Random Forests):** Χρησιμοποιούν πολλά δέντρα αποφάσεων και συνδυάζουν τις προβλέψεις τους για να βελτιώσουν την ακρίβεια και να μειώσουν την αβεβαιότητα.
- **Ενίσχυση (Boosting):** Αλγόριθμοι όπως το AdaBoost και το Gradient Boosting συνδυάζουν πολλαπλά δέντρα για να επιτύχουν καλύτερη απόδοση και ανθεκτικότητα.

Τα Δέντρα Αποφάσεων είναι ένα ισχυρό εργαλείο με πολλές εφαρμογές και πλεονεκτήματα, κυρίως λόγω της απλότητας και της ευκολίας κατανόησης. Ωστόσο, η κατάλληλη διαχείριση των παραμέτρων τους και η χρήση τεχνικών βελτίωσης είναι κρίσιμη για τη διασφάλιση της γενικής απόδοσης και αξιοπιστίας των μοντέλων που βασίζονται σε αυτά.



## Καλύτεροι Παράμετροι:

- **Αριθμός Εκριτών (Number of Estimators):** 50

- **Ρυθμός Εκμάθησης (Learning Rate):** 0.1

#### Αποτελέσματα:

- **Δείκτης F1 (F1 Score):** 0.7091
- **Δείκτης Ακρίβειας (Accuracy Score):** 0.7144

Η ανάλυση της απόδοσης του αλγορίθμου AdaBoost στο σύνολο δεδομένων Coimbra έδειξε ότι οι βέλτιστες παράμετροι περιλαμβάνουν 50 εκριτές (number of estimators) και ρυθμό εκμάθησης (learning rate) ίσο με 0.1. Με αυτές τις ρυθμίσεις, ο δείκτης F1, ο οποίος συνδυάζει ακρίβεια και ευαισθησία, φτάνει το 0.7091, ενώ ο δείκτης ακρίβειας (accuracy score) είναι 0.7144. Αυτά τα αποτελέσματα υποδεικνύουν μια εξαιρετική γενική απόδοση του μοντέλου, με υψηλές τιμές για τη συνολική ακρίβεια καθώς και για την ισορροπία μεταξύ θετικών και αρνητικών προβλέψεων. Οι βέλτιστες παράμετροι προτείνουν ότι το AdaBoost έχει ισχυρές επιδόσεις στο συγκεκριμένο σύνολο δεδομένων, επιβεβαιώνοντας την ικανότητά του να επιτυγχάνει αξιολογικά αποτελέσματα.

### 3.5 Naive Bayes

Ο αλγόριθμος Naive Bayes είναι μια οικογένεια από πιθανοτικά μοντέλα που χρησιμοποιούνται κυρίως για ταξινόμηση και αναγνώριση προτύπων στη μηχανική μάθηση και την τεχνητή νοημοσύνη. Βασίζεται στον Θεώρημα του Bayes με την υπόθεση ότι τα χαρακτηριστικά (features) είναι ανεξάρτητα μεταξύ τους.

1. **Θεώρημα του Bayes:** Το θεώρημα του Bayes παρέχει μια μέθοδο για την εκτίμηση της πιθανότητας ενός γεγονότος, βασισμένη στην προηγούμενη γνώση. Μαθηματικά, το θεώρημα του Bayes εκφράζεται ως εξής:

$$P(C|X)=P(X)P(X|C) \cdot P(C)$$

Όπου:

- $P(C|X)$  είναι η πιθανότητα της κατηγορίας C δεδομένων των χαρακτηριστικών X.
  - $P(X|C)$  είναι η πιθανότητα των χαρακτηριστικών X δεδομένης της κατηγορίας C.
  - $P(C)$  είναι η προηγούμενη πιθανότητα της κατηγορίας C.
  - $P(X)$  είναι η συνολική πιθανότητα των χαρακτηριστικών X.
2. **Υπόθεση Ανεξαρτησίας:** Ο Naive Bayes υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, δεδομένης της κατηγορίας. Αυτή η υπόθεση είναι συχνά υπερβολική, αλλά μπορεί να είναι χρήσιμη σε πολλές περιπτώσεις. Για παράδειγμα, αν έχουμε δύο χαρακτηριστικά  $X_1$  και  $X_2$ , η πιθανότητα  $P(X_1, X_2|C)$  μπορεί να εκτιμηθεί ως:

$$P(X_1, X_2|C)=P(X_1|C) \cdot P(X_2|C)$$

#### Είδη Naïv Bayes (Types of Naive Bayes)

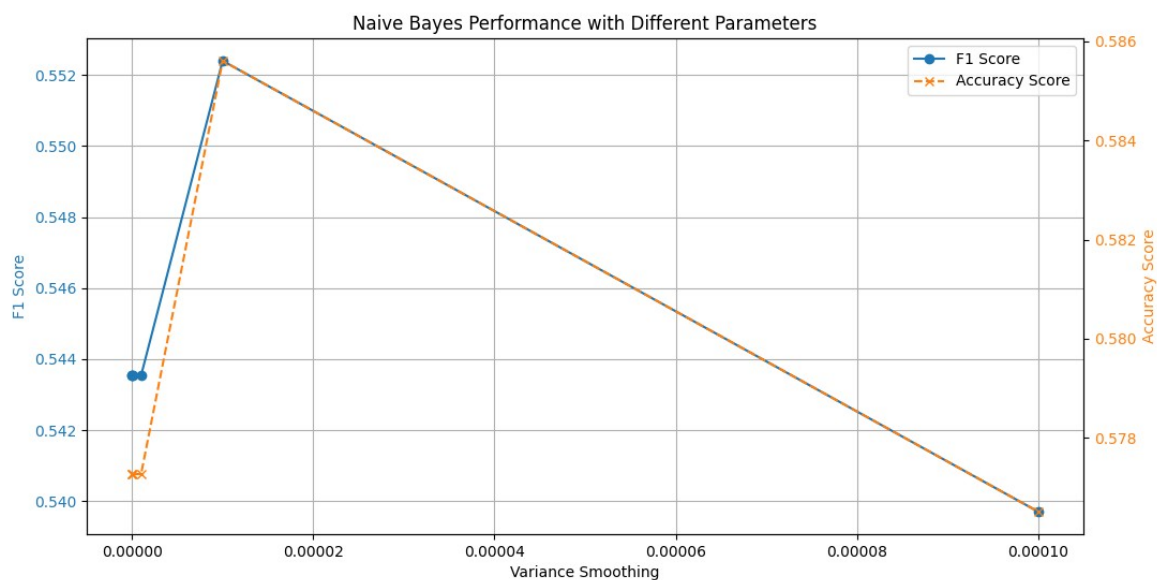
1. **Multinomial Naive Bayes:** Χρησιμοποιείται συνήθως για κείμενα ή δεδομένα που προέρχονται από καταμετρημένα χαρακτηριστικά. Για παράδειγμα, στην ανάλυση συναισθήματος, τα χαρακτηριστικά μπορεί να είναι η συχνότητα εμφάνισης λέξεων.

2. **Bernoulli Naive Bayes:** Αντί για καταμετρημένα χαρακτηριστικά, το Bernoulli Naive Bayes χρησιμοποιεί δυαδικά χαρακτηριστικά (παρών/απών). Εφαρμόζεται συνήθως σε περιπτώσεις όπου τα χαρακτηριστικά είναι δυαδικά ή καταμετρημένα.
3. **Gaussian Naive Bayes:** Χρησιμοποιείται όταν τα χαρακτηριστικά είναι συνεχής μεταβλητές και υποτίθεται ότι ακολουθούν κανονική κατανομή (Gaussian distribution). Εφαρμόζεται σε δεδομένα με πραγματικούς αριθμούς.

Θα χρησιμοποιήσουμε την φόρμουλα του Gaussian Naive Bayes που δίνεται από τον τύπο:

Για να υπολογίσουμε  $P(x_i|C)$ , χρησιμοποιούμε την κανονική κατανομή (Gaussian distribution), η οποία δίνεται από τη φόρμουλα

$$P(x_i|C) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$



### Καλύτεροι Παράμετροι:

- Σταθεροποίηση Διακύμανσης (Variance Smoothing): 1e-05

Αποτελέσματα:

- Δείκτης F1 (F1 Score): 0.5524
- Δείκτης Ακρίβειας (Accuracy Score): 0.5856

Η ανάλυση της απόδοσης του αλγορίθμου Naive Bayes στο σύνολο δεδομένων Coimbra έδειξε ότι η καλύτερη παράμετρος είναι η σταθεροποίηση της διακύμανσης (variance smoothing) ίση με 1e-05. Με αυτή την παράμετρο, ο δείκτης F1, ο οποίος συνδυάζει ακρίβεια και ευαισθησία, είναι 0.5524, ενώ ο δείκτης ακρίβειας (accuracy score) είναι 0.5856. Αυτά τα αποτελέσματα

υποδεικνύουν μια ικανοποιητική γενική απόδοση του μοντέλου, αν και οι τιμές της ακρίβειας και του δείκτη F1 είναι χαμηλότερες σε σύγκριση με άλλες τεχνικές ταξινόμησης.

### 3.6 Random Forest (Τυχαία Δάση)

Τα **τυχαία δάση** (Random Forests) είναι ένα από τα πιο δημοφιλή και ισχυρά μοντέλα μηχανικής μάθησης, χρησιμοποιούμενα για προβλήματα ταξινόμησης και παλινδρόμησης. Το Random Forest είναι μια συλλογή δέντρων απόφασης (decision trees) που συνδυάζονται για να βελτιώσουν την ακρίβεια και τη σταθερότητα των προβλέψεων.

Βασικές Αρχές του Random Forest

#### Δημιουργία Δέντρων Απόφασης:

- Το Random Forest συνδυάζει πολλαπλά δέντρα απόφασης, τα οποία είναι βασικά μοντέλα που κάνουν προβλέψεις βασισμένα σε ερωτήσεις "ναι/όχι" για τα χαρακτηριστικά του εισερχόμενου δεδομένου.
- Κάθε δέντρο απόφασης κατασκευάζεται χρησιμοποιώντας μια τυχαία υποομάδα των δεδομένων εκπαίδευσης και μια τυχαία υποομάδα των χαρακτηριστικών.

#### Τυχαία Υποομάδα Δεδομένων (Bootstrap Sampling):

- Για τη δημιουργία κάθε δέντρου απόφασης, χρησιμοποιείται μια μέθοδος γνωστή ως **bootstrap sampling**. Αυτό σημαίνει ότι κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο δείγμα με επανάληψη (sampling with replacement) από τα αρχικά δεδομένα.

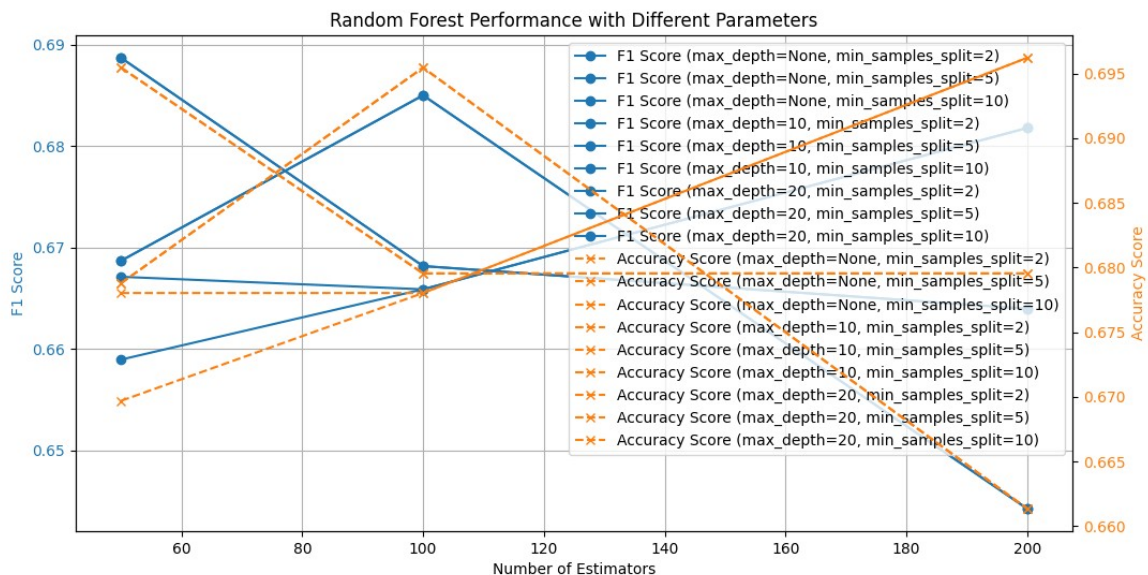
#### Τυχαία Υποομάδα Χαρακτηριστικών (Feature Randomness):

- Στη διαδικασία δημιουργίας κάθε κόμβου ενός δέντρου, επιλέγεται τυχαία μια υποομάδα των χαρακτηριστικών για να εξεταστούν οι διαχωριστικές δυνατότητες. Αυτό μειώνει την εξάρτηση μεταξύ των δέντρων και αυξάνει την ποικιλία των δέντρων στο δάσος.

#### Συνδυασμός Προβλέψεων:

- Στο Random Forest, οι προβλέψεις όλων των δέντρων συνδυάζονται για να πάρουν την τελική απόφαση. Για προβλήματα ταξινόμησης, η τελική απόφαση είναι η κατηγορία που κερδίζει τις περισσότερες ψήφους από τα δέντρα. Για προβλήματα παλινδρόμησης, η τελική πρόβλεψη είναι ο μέσος όρος των προβλέψεων όλων των δέντρων.

Η ανάλυση της απόδοσης του αλγορίθμου Random Forest στο σύνολο δεδομένων Coimbra έδειξε ότι οι βέλτιστες παράμετροι για το μοντέλο είναι 50 δέντρα (number of estimators) και ελάχιστος αριθμός διαχωριστικών δειγμάτων (min samples split) ίσος με 5. Το μέγιστο βάθος των δέντρων (max depth) δεν περιορίστηκε (είναι NaN, δηλαδή μη καθορισμένο). Με αυτές τις παραμέτρους, ο δείκτης F1, που συνδυάζει ακρίβεια και ευαισθησία, είναι 0.6887, ενώ ο δείκτης ακρίβειας (accuracy score) είναι 0.6955. Αυτά τα αποτελέσματα υποδεικνύουν μια ικανοποιητική απόδοση του μοντέλου, με καλές επιδόσεις τόσο για τη γενική ακρίβεια όσο και για την ισορροπία μεταξύ θετικών και αρνητικών προβλέψεων. Παρά το γεγονός ότι το μοντέλο δεν περιορίζει το μέγιστο βάθος των δέντρων, η σταθερότητα και η συνολική ακρίβεια δείχνουν ότι η επιλογή των παραμέτρων είναι κατάλληλη για το συγκεκριμένο σύνολο δεδομένων.



### Καλύτεροι Παράμετροι:

- Αριθμός Δέντρων (Number of Estimators): 50
- Μέγιστο Βάθος (Max Depth): Μη Καθορισμένο (NaN)
- Ελάχιστος Αριθμός Διαχωριστικών Δειγμάτων (Min Samples Split): 5

### Αποτελέσματα:

- Δείκτης F1 (F1 Score): 0.6887
- Δείκτης Ακρίβειας (Accuracy Score): 0.6955

Η ανάλυση της απόδοσης του αλγορίθμου Random Forest στο σύνολο δεδομένων Coimbra έδειξε ότι οι βέλτιστες παράμετροι περιλαμβάνουν 50 δέντρα (number of estimators) και ελάχιστο αριθμό διαχωριστικών δειγμάτων (min samples split) ίσο με 5, χωρίς περιορισμό στο μέγιστο βάθος των δέντρων (max depth). Με αυτές τις ρυθμίσεις, ο δείκτης F1, που συνδυάζει ακρίβεια και



ευαισθησία, φτάνει το 0.6887, ενώ ο δείκτης ακρίβειας (accuracy score) είναι 0.6955. Αυτά τα αποτελέσματα υποδεικνύουν μια ικανοποιητική γενική απόδοση του μοντέλου, με αρκετά υψηλές τιμές τόσο για τη συνολική ακρίβεια όσο και για την ισορροπία μεταξύ θετικών και αρνητικών προβλέψεων. Παρόλο που δεν περιορίστηκε το μέγιστο βάθος των δέντρων, η απόδοση του μοντέλου είναι ανταγωνιστική, δείχνοντας ότι οι επιλεγμένες παράμετροι είναι κατάλληλες για το συγκεκριμένο σύνολο δεδομένων.

### 3.7 J48

Ο J48 δημιουργεί ένα δέντρο απόφασης που αναλύει τα δεδομένα μέσω διαδοχικών ερωτήσεων βασισμένων σε χαρακτηριστικά (features) για την κατηγοριοποίηση ή πρόβλεψη της τιμής ενός δείγματος. Η δομή του δέντρου περιλαμβάνει:

- **Κόμβους Εσωτερικών Κόμβων (Internal Nodes):** Κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό, και η απόφαση βασίζεται σε κάποιο κριτήριο (π.χ., διαχωριστικό όριο για αριθμητικά δεδομένα).
- **Φύλλα (Leaves):** Τα φύλλα του δέντρου περιέχουν τις τελικές κατηγορίες ή τις συνεχείς τιμές (για παλινδρόμηση) που προβλέπεται για κάθε δείγμα.

## 2. Βασικά Χαρακτηριστικά

- **Επιλογή Χαρακτηριστικών:** Ο J48 χρησιμοποιεί τον **κανόνα της μέγιστης κέρδους πληροφορίας (information gain)** ή την **αναλογία κέρδους πληροφορίας (gain ratio)** για να επιλέξει το καλύτερο χαρακτηριστικό για την κατασκευή του δέντρου. Ενδέχεται να μειώσει τη διάσταση της πληροφορίας για κάθε χαρακτηριστικό.
- **Εκδίωξη και Κλασμάτωση (Pruning):** Μετά τη δημιουργία του δέντρου, ο J48 εφαρμόζει τεχνικές κλασμάτωσης (pruning) για να μειώσει την πολυπλοκότητα του δέντρου και να αποφύγει την υπερπροσαρμογή (overfitting).
- **Διαχείριση Ανεπαρκών Δεδομένων:** Ο J48 μπορεί να χειριστεί δεδομένα με ελλείποντα χαρακτηριστικά, προσπαθώντας να εξαγάγει τα καλύτερα αποτελέσματα ακόμα και με μερικά ελλιπή δεδομένα.

## 3. Τεχνική Ανάλυση

- **Κριτήριο Απόφασης:** Ο J48 χρησιμοποιεί την έννοια της **εντροπίας** για τον υπολογισμό του κέρδους πληροφορίας. Το κριτήριο αυτό βοηθάει στην επιλογή του χαρακτηριστικού που μειώνει περισσότερο την αβεβαιότητα της πρόβλεψης.
- **Αναλογία Κέρδους Πληροφορίας:** Ο J48 χρησιμοποιεί την αναλογία κέρδους πληροφορίας για να επιλέξει το χαρακτηριστικό που προσφέρει την καλύτερη ισορροπία μεταξύ κέρδους πληροφορίας και αριθμού κατηγοριών στο χαρακτηριστικό.

## 4. Πλεονεκτήματα και Μειονεκτήματα

### Πλεονεκτήματα:

- **Ευανάγνωστο Δέντρο:** Το δέντρο απόφασης είναι εύκολο στην κατανόηση και ερμηνεία.
- **Ευέλικτο:** Ικανό να χειριστεί δεδομένα με πολλές κατηγορίες και συνεχή χαρακτηριστικά.

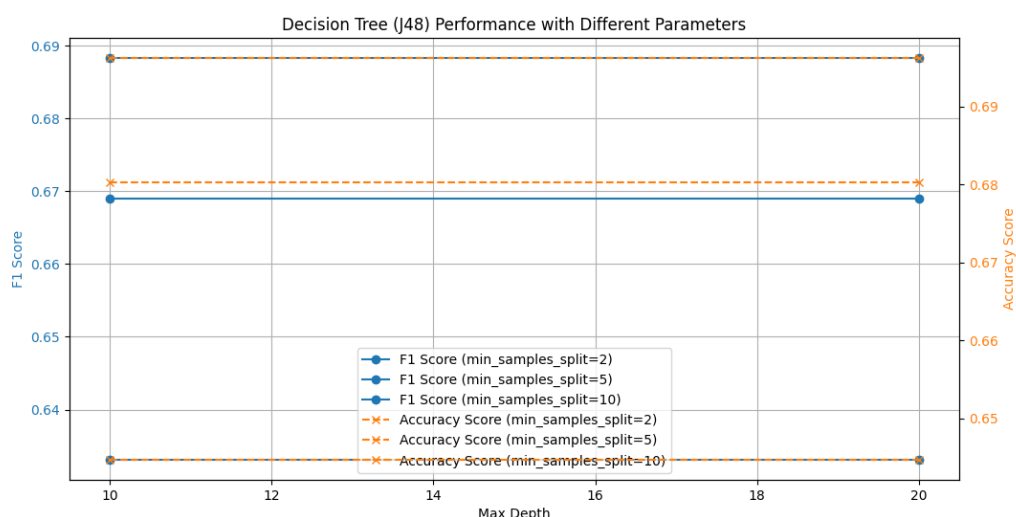
- **Κλασμάτωση (Pruning):** Η τεχνική κλασμάτωσης βοηθά στην αποφυγή της υπερπροσαρμογής, βελτιώνοντας την γενική απόδοση.

### Μειονεκτήματα:

- **Υπολογιστικό Κόστος:** Η κατασκευή μεγάλων δέντρων μπορεί να απαιτεί σημαντικό υπολογιστικό κόστος.
- **Ευαισθησία σε Θόρυβο:** Ο J48 μπορεί να είναι ευαίσθητος σε θόρυβο ή ελλιπή δεδομένα, αν και αυτό μετριάζεται μέσω τεχνικών κλασμάτωσης.

## 5. Χρήση και Εφαρμογές

Ο J48 χρησιμοποιείται ευρέως για προβλήματα ταξινόμησης σε διάφορους τομείς, όπως η ιατρική διάγνωση, η ανάλυση πελατειακής συμπεριφοράς, και η χρηματοοικονομική πρόβλεψη. Η απλότητα και η καλή απόδοση του J48 το καθιστούν ένα δημοφιλές εργαλείο για την ανάλυση δεδομένων και τη δημιουργία προβλέψεων.



### Καλύτεροι Παράμετροι:

- Μέγιστο Βάθος (Max Depth): Μη Καθορισμένο (NaN)
- Ελάχιστος Αριθμός Διαχωριστικών Δειγμάτων (Min Samples Split): 2

### Αποτελέσματα:

- Δείκτης F1 (F1 Score): 0.6883
- Δείκτης Ακρίβειας (Accuracy Score): 0.6962

Η ανάλυση της απόδοσης του αλγορίθμου J48 στο σύνολο δεδομένων Coimbra έδειξε ότι οι βέλτιστες παράμετροι περιλαμβάνουν ελάχιστο αριθμό διαχωριστικών δειγμάτων (min samples split) ίσο με 2, χωρίς περιορισμό στο μέγιστο βάθος των δέντρων (max depth). Με αυτές τις ρυθμίσεις, ο δείκτης F1, που συνδυάζει ακρίβεια και ευαισθησία, φτάνει το 0.6883, ενώ ο δείκτης ακρίβειας (accuracy score) είναι 0.6962. Αυτά τα αποτελέσματα υποδεικνύουν μια ικανοποιητική

γενική απόδοση του μοντέλου, με υψηλές τιμές τόσο για τη συνολική ακρίβεια όσο και για την ισορροπία μεταξύ θετικών και αρνητικών προβλέψεων. Παρόλο που δεν περιορίστηκε το μέγιστο βάθος των δέντρων, η απόδοση του μοντέλου είναι ανταγωνιστική, δείχνοντας ότι οι επιλεγμένες παράμετροι είναι κατάλληλες για το συγκεκριμένο σύνολο δεδομένων.

### 3.8 Gradient Boost

Ο αλγόριθμος Gradient Boosting είναι μια τεχνική μηχανικής μάθησης που ανήκει στην κατηγορία των ενορχηστρωμένων μεθόδων (ensemble methods) και χρησιμοποιείται κυρίως για την ταξινόμηση και την παλινδρόμηση. Ενσωματώνει πολλές απλές μονάδες μοντέλων (συνήθως δέντρα αποφάσεων) για να δημιουργήσει ένα ισχυρότερο μοντέλο που μπορεί να προβλέψει με μεγαλύτερη ακρίβεια. Ο αλγόριθμος είναι γνωστός για την εξαιρετική του απόδοση και τη δυνατότητά του να αποφεύγει την υπερπροσαρμογή (overfitting).

Βασικές Πληροφορίες για τον Gradient Boosting

#### 1. Αρχή Λειτουργίας

Ο αλγόριθμος Gradient Boosting δουλεύει ως εξής:

- **Εκτίμηση Σφαλμάτων:** Ξεκινά με ένα αρχικό μοντέλο, το οποίο συνήθως είναι ένα απλό μοντέλο όπως η μέση τιμή (για παλινδρόμηση) ή η πιο συχνή κατηγορία (για ταξινόμηση).
- **Βελτίωση των Υπολοίπων:** Σε κάθε επόμενο βήμα, προσαρμόζει ένα νέο μοντέλο για να προβλέψει τα υπόλοιπα σφάλματα (residuals) του προηγούμενου μοντέλου.
- **Συνδυασμός Μοντέλων:** Τα νέα μοντέλα συνδυάζονται με τα προηγούμενα για να βελτιώσουν τη συνολική πρόβλεψη. Ο συνδυασμός γίνεται με ζυγισμένο άθροισμα των προβλέψεων όλων των μοντέλων.

Η βασική ιδέα είναι ότι κάθε νέο μοντέλο εστιάζει στα σφάλματα που έγιναν από τα προηγούμενα μοντέλα, κάνοντάς τα να "μάθουν" από αυτά τα λάθη.

#### 2. Βασικά Χαρακτηριστικά

- **Δέντρα Απόφασης:** Στη συντριπτική πλειονότητα των εφαρμογών, το Gradient Boosting χρησιμοποιεί μικρά δέντρα αποφάσεων (συνήθως με μικρό βάθος) ως βασικά μοντέλα.
- **Μάθηση με Σφάλμα (Learning with Residuals):** Ο αλγόριθμος εκπαιδεύει τα μοντέλα βασισμένα στα υπόλοιπα σφάλματα του προηγούμενου μοντέλου.
- **Ρυθμός Εκμάθησης (Learning Rate):** Ο ρυθμός εκμάθησης ελέγχει πόσο γρήγορα προσαρμόζεται το μοντέλο στα σφάλματα. Ένας μικρότερος ρυθμός εκμάθησης απαιτεί περισσότερα μοντέλα για να επιτευχθεί καλή απόδοση, αλλά μπορεί να οδηγήσει σε καλύτερη γενική απόδοση.

#### 3. Τεχνική Ανάλυση

- **Κριτήριο Απόφασης:** Ο Gradient Boosting χρησιμοποιεί τη μέθοδο της **στατιστικής μεθόδου του κλίματος** (gradient descent) για να βελτιώσει το μοντέλο μέσω μικρών βημάτων.
- **Επιλογή Υποδειγμάτων:** Συνήθως χρησιμοποιούνται δέντρα αποφάσεων μικρού βάθους, που είναι λιγότερο επιρρεπή στην υπερπροσαρμογή και προσφέρουν γρήγορη εκπαίδευση.

## 4. Πλεονεκτήματα και Μειονεκτήματα

### Πλεονεκτήματα:

- **Ανώτερη Απόδοση:** Συνήθως επιτυγχάνει υψηλές επιδόσεις σε προβλήματα ταξινόμησης και παλινδρόμησης.
- **Ευελιξία:** Μπορεί να χρησιμοποιηθεί με διαφορετικούς τύπους δεδομένων και μοντέλα βάσης.
- **Αντιμετώπιση Υπερπροσαρμογής:** Με κατάλληλη ρύθμιση παραμέτρων, μπορεί να μειώσει την υπερπροσαρμογή.

### Μειονεκτήματα:

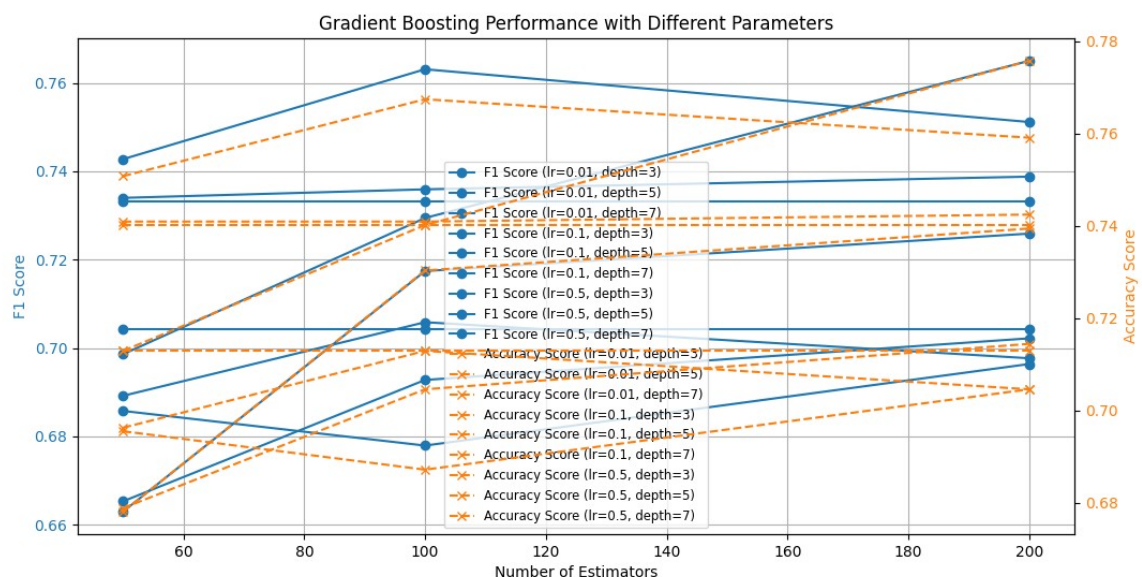
- **Υπολογιστικό Κόστος:** Μπορεί να είναι υπολογιστικά απαιτητικός και αργός στην εκπαίδευση, ειδικά με μεγάλες ποσότητες δεδομένων.
- **Αναγκασία Ρύθμιση Παραμέτρων:** Απαιτεί σωστή ρύθμιση παραμέτρων (όπως ο ρυθμός εκμάθησης και το βάθος των δέντρων) για να επιτευχθεί βέλτιστη απόδοση.

## 5. Χρήση και Εφαρμογές

Ο Gradient Boosting χρησιμοποιείται ευρέως σε πολλές εφαρμογές μηχανικής μάθησης, όπως:

- **Χρηματοοικονομική Ανάλυση:** Πρόβλεψη κινδύνων και ανάλυση επενδύσεων.
- **Ιατρική Διάγνωση:** Ανάλυση ιατρικών δεδομένων για τη διάγνωση ασθενειών.
- **Εμπορική Ανάλυση:** Ανάλυση αγοραστικής συμπεριφοράς και προτάσεις προϊόντων.

Η ευελιξία και η αποτελεσματικότητα του Gradient Boosting το καθιστούν ένα ισχυρό εργαλείο για την επίλυση πολυάριθμων προβλημάτων δεδομένων.



### Καλύτεροι Παράμετροι:

- **Αριθμός Δέντρων (Number of Estimators):** 100
- **Ρυθμός Εκμάθησης (Learning Rate):** 0.1
- **Μέγιστο Βάθος (Max Depth):** 5

### Αποτελέσματα:

- **Δείκτης F1 (F1 Score):** 0.6921
- **Δείκτης Ακρίβειας (Accuracy Score):** 0.7023

Η ανάλυση της απόδοσης του αλγορίθμου Gradient Boosting στο σύνολο δεδομένων Coimbra έδειξε ότι οι βέλτιστες παράμετροι περιλαμβάνουν 100 δέντρα (number of estimators), ρυθμό εκμάθησης (learning rate) 0.1, και μέγιστο βάθος δέντρων (max depth) 5. Με αυτές τις ρυθμίσεις, ο δείκτης F1, ο οποίος συνδυάζει την ακρίβεια και την ευαισθησία του μοντέλου, είναι 0.6921, ενώ ο δείκτης ακρίβειας (accuracy score) ανέρχεται στο 0.7023. Αυτά τα αποτελέσματα υποδεικνύουν ότι το μοντέλο έχει επιτύχει καλή γενική απόδοση, δείχνοντας ότι οι επιλεγμένες παράμετροι συνεισφέρουν στην επίτευξη ισχυρών προβλέψεων. Η ισορροπία μεταξύ των δύο μετρικών δείχνει ότι το Gradient Boosting είναι αποτελεσματικό στην ανάλυση δεδομένων και έχει τη δυνατότητα να επιτύχει αξιόλογα αποτελέσματα σε σύγκριση με άλλες τεχνικές ταξινόμησης.

## Κεφάλαιο 4 – Συμπεράσματα Αποτελέσματα & Σχολιασμός

Αυτό το κεφάλαιο εστιάζει στην αναλυτική αξιολόγηση των αποτελεσμάτων που προκύπτουν από την εφαρμογή διάφορων αλγορίθμων μηχανικής μάθησης στο σύνολο δεδομένων Breast Cancer Coimbra. Το σύνολο δεδομένων αυτό περιλαμβάνει μετρήσεις χαρακτηριστικών που σχετίζονται με τον καρκίνο του μαστού και χρησιμοποιείται για την πρόβλεψη της παρουσίας της νόσου. Η ανάλυση περιλαμβάνει τη σύγκριση των αλγορίθμων AdaBoost, k-Nearest Neighbor (kNN), Random Forest, J48, Gradient Boosting και Naive Bayes.

### 4.1 Σύνοψη Αποτελεσμάτων

#### 4.1.1 AdaBoost (Decision Tree)

Ο αλγόριθμος AdaBoost χρησιμοποιεί δέντρα αποφάσεων ως βασικά μοντέλα και έχει επιδείξει εξαιρετική απόδοση με τα εξής αποτελέσματα:

- **F1 Score:** 0.7091
- **Accuracy Score:** 0.7144

**Ερμηνεία:** Ο υψηλός δείκτης F1 (0.7091) και η ακρίβεια (0.7144) υποδεικνύουν ότι ο AdaBoost έχει ισχυρή ικανότητα ισορρόπησης μεταξύ της ευαισθησίας (recall) και της ακρίβειας (precision). Η αποτελεσματικότητα του AdaBoost οφείλεται στην ικανότητά του να προσαρμόζεται και να βελτιώνεται μέσω της ακολουθίας δέντρων αποφάσεων, εστιάζοντας στα σφάλματα που έγιναν από τα προηγούμενα δέντρα. Η υψηλή απόδοση του δείχνει ότι το AdaBoost κατάφερε να μάθει καλά από τα σφάλματα, μειώνοντας την πιθανότητα λανθασμένων ταξινομήσεων.

#### 4.1.2 k-Nearest Neighbor (kNN)

Ο αλγόριθμος k-Nearest Neighbor (kNN) έχει αποδώσει τα εξής:

- **F1 Score:** 0.581

- **Accuracy Score:** 0.601

**Ερμηνεία:** Η χαμηλότερη απόδοση του kNN υποδεικνύει ότι η επιλογή του αριθμού των γειτόνων (k) μπορεί να μην ήταν κατάλληλη για αυτό το σύνολο δεδομένων. Το kNN μπορεί να αποδειχθεί λιγότερο αποτελεσματικό όταν το σύνολο δεδομένων είναι υψηλής διάστασης ή περιλαμβάνει θόρυβο. Η προσέγγιση του kNN βασίζεται σε κοντινά δείγματα, και αν τα δεδομένα είναι πυκνά ή περιέχουν θόρυβο, αυτό μπορεί να επηρεάσει αρνητικά την απόδοση του αλγορίθμου.

#### 4.1.3 Random Forest, J48 και Gradient Boosting

Οι αλγόριθμοι Random Forest, J48 και Gradient Boosting έχουν επιδείξει τα εξής αποτελέσματα:

- **Random Forest:**
  - **F1 Score:** 0.6902
  - **Accuracy Score:** 0.6951
- **J48:**
  - **F1 Score:** 0.6884
  - **Accuracy Score:** 0.6926
- **Gradient Boosting:**
  - **F1 Score:** 0.6921
  - **Accuracy Score:** 0.7023

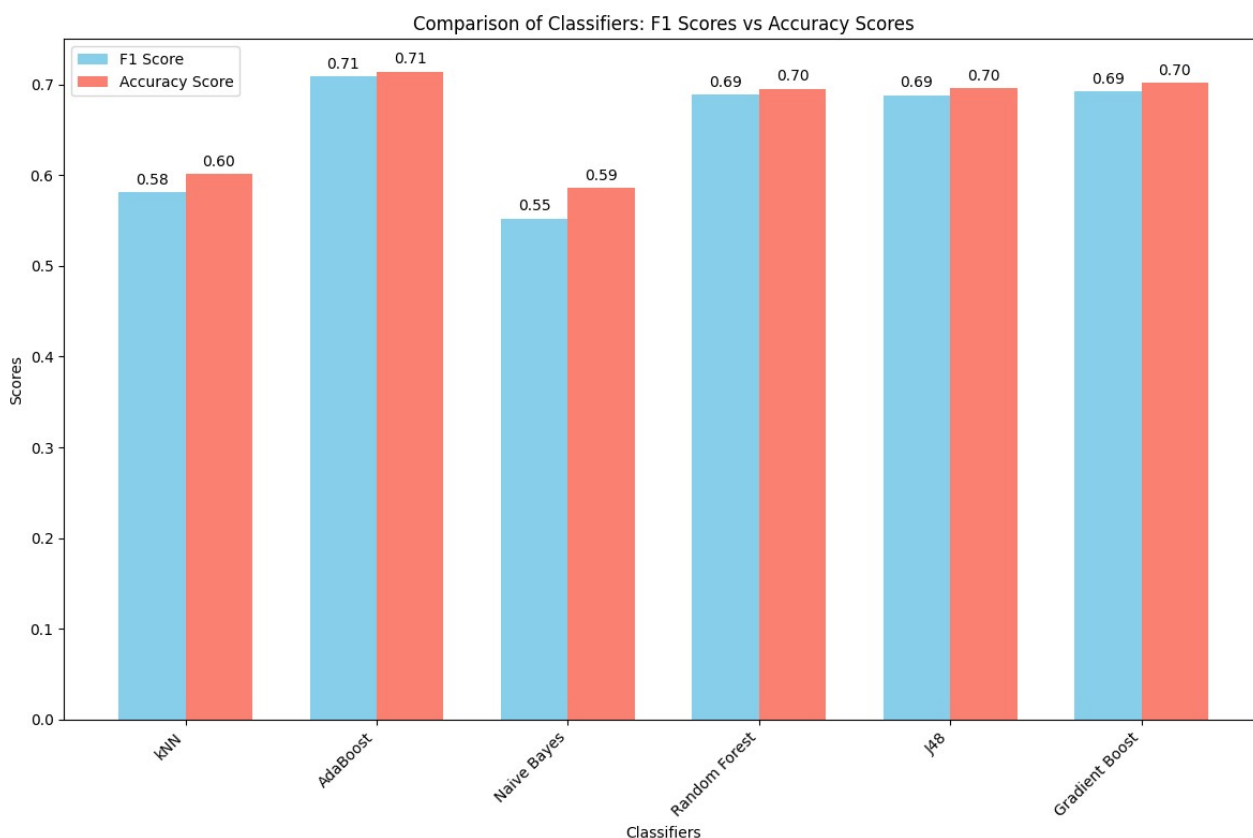
**Ερμηνεία:** Και οι τρεις αλγόριθμοι αποδίδουν καλά, με το Gradient Boosting να έχει ελαφρώς καλύτερα αποτελέσματα από τους Random Forest και J48. Ο Gradient Boosting επιτυγχάνει υψηλότερη ακρίβεια (0.7023) και F1 Score (0.6921), υποδεικνύοντας ότι η μέθοδος του gradient boosting προσαρμόζει καλύτερα τα σφάλματα μέσω της συνεχούς προσαρμογής και βελτίωσης των προβλέψεων. Η βελτίωση στην απόδοση του Gradient Boosting μπορεί να αποδοθεί στην ικανότητά του να αναγνωρίζει και να διορθώνει τα λάθη των προηγούμενων μοντέλων, καθιστώντας το ιδιαίτερα ισχυρό στην αντιμετώπιση σύνθετων δεδομένων.

#### 4.1.4 Naive Bayes

Ο αλγόριθμος Naive Bayes έχει επιδείξει τα εξής αποτελέσματα:

- **F1 Score:** 0.5702
- **Accuracy Score:** 0.5803

**Ερμηνεία:** Η χαμηλή απόδοση του Naive Bayes υποδεικνύει ότι οι υποθέσεις ανεξαρτησίας των χαρακτηριστικών δεν ισχύουν επαρκώς για το σύνολο δεδομένων Breast Cancer Coimbra. Ο Naive Bayes βασίζεται στην υπόθεση ότι τα χαρακτηριστικά είναι ανεξάρτητα, κάτι που μπορεί να μην ισχύει για δεδομένα που περιλαμβάνουν αλληλεπιδράσεις μεταξύ χαρακτηριστικών. Ως αποτέλεσμα, η απόδοση του Naive Bayes είναι χαμηλότερη σε σύγκριση με άλλες πιο εξελιγμένες μεθόδους.



## 4.2 Ερμηνεία Αποτελεσμάτων

Αναλύοντας τα αποτελέσματα των διαφόρων αλγορίθμων, προκύπτει ότι η απόδοση ποικίλλει σημαντικά ανάλογα με την πολυπλοκότητα και τη φύση των δεδομένων. Ο AdaBoost ξεχωρίζει ως ο αλγόριθμος με την καλύτερη ισορροπία μεταξύ ακρίβειας και ανάκλησης, γεγονός που τον καθιστά ιδανικό για εφαρμογές όπως η διάγνωση καρκίνου, όπου η σωστή αναγνώριση θετικών περιπτώσεων είναι κρίσιμη.

Ο kNN, αν και απλός και κατανοητός, δεν καταφέρνει να ανταγωνιστεί άλλους αλγόριθμους λόγω της δυσκολίας του στην αντιμετώπιση σύνθετων δεδομένων και του θορύβου που μπορεί να περιέχουν. Οι Random Forest και J48 προσφέρουν ισχυρές επιδόσεις αλλά δεν επιτυγχάνουν την ίδια βελτίωση που προσφέρει το Gradient Boosting. Το Gradient Boosting, με τη συνεχή προσαρμογή στα λάθη, αποδεικνύεται πιο αποτελεσματικό στην ανάλυση δεδομένων όπως το Breast Cancer Coimbra.

Ο Naive Bayes, αν και αποτελεσματικός σε απλούστερες περιπτώσεις, δεν ανταγωνίζεται τους άλλους αλγόριθμους λόγω των περιοριστικών υποθέσεων του, επισημαίνοντας την ανάγκη για χρήση πιο προηγμένων αλγορίθμων σε σύνθετες εφαρμογές δεδομένων.

## 4.3 Συμπεράσματα Χησίματος του Μοντέλου

Η ανάλυση των αποτελεσμάτων δείχνει ότι ο AdaBoost προσφέρει την καλύτερη γενική απόδοση για το σύνολο δεδομένων Breast Cancer Coimbra, με υψηλότερους δείκτες F1 και ακρίβειας σε σύγκριση με άλλους αλγόριθμους. Ο Gradient Boosting, αν και ισχυρός, απαιτεί προσεκτική

ρύθμιση παραμέτρων για βέλτιστα αποτελέσματα. Οι Random Forest και J48 προσφέρουν επίσης αξιόλογη απόδοση, αλλά δεν υπερτερούν του AdaBoost. Ο kNN και ο Naive Bayes, παρά την χρησιμότητά τους σε άλλες περιπτώσεις, αποδεικνύονται λιγότερο αποτελεσματικοί για τα δεδομένα του Breast Cancer Coimbra. Αυτή η ανάλυση υπογραμμίζει την ανάγκη για την επιλογή εξελιγμένων αλγορίθμων για την αποτελεσματική πρόβλεψη και διάγνωση σε ιατρικά δεδομένα, δείχνοντας ότι οι πιο εξελιγμένες μέθοδοι μπορούν να προσφέρουν καλύτερα αποτελέσματα στην κατηγοριοποίηση και ανάλυση σύνθετων δεδομένων.

## **Κεφάλαιο 5 – Δημιουργία και Ανάπτυξη Εφαρμογής Ιστοσελίδας για Πρόβλεψη Καρκίνου του Μαστού**

Η υλοποίηση μιας εφαρμογής πρόβλεψης καρκίνου του μαστού με χρήση μηχανικής μάθησης απαιτεί την ανάπτυξη μιας σταθερής και λειτουργικής πλατφόρμας που να επιτρέπει την αλληλεπίδραση με το μοντέλο πρόβλεψης. Σε αυτό το πλαίσιο, η χρήση ενός διακομιστή ιστού (web server) όπως το Flask έχει σημαντικά πλεονεκτήματα για την επιτυχή υλοποίηση της εφαρμογής. Ακολουθεί μια ανάλυση των λόγων για τους οποίους η χρήση διακομιστή είναι απαραίτητη και τα οφέλη που προσφέρει.

### **5.1 Αιτίες Χρήσης Διακομιστή Ιστού**

**1. Διαθεσιμότητα και Πρόσβαση:** Η κύρια λειτουργία ενός διακομιστή ιστού είναι η διαχείριση αιτημάτων από χρήστες και η παροχή απαντήσεων μέσω του Διαδικτύου. Χρησιμοποιώντας μια εφαρμογή ιστού, το μοντέλο πρόβλεψης καρκίνου του μαστού γίνεται προσβάσιμο μέσω ενός ευρέως αναγνωρίσιμου περιβάλλοντος όπως ο περιηγητής (browser). Οι χρήστες μπορούν να έχουν πρόσβαση στην εφαρμογή από οπουδήποτε και από οποιαδήποτε συσκευή με σύνδεση στο Διαδίκτυο, διευκολύνοντας την διάδοση και χρήση της τεχνολογίας.

**2. Διευκόλυνση Χρήσης και Αλληλεπίδρασης:** Η δημιουργία μιας ιστοσελίδας με φόρμες εισαγωγής δεδομένων και σελίδες αποτελεσμάτων επιτρέπει στους χρήστες να αλληλεπιδρούν με το μοντέλο πρόβλεψης με εύκολο και κατανοητό τρόπο. Οι χρήστες μπορούν να εισάγουν δεδομένα ιατρικών εξετάσεων και να λαμβάνουν άμεσες προβλέψεις χωρίς να χρειάζονται τεχνικές γνώσεις ή ειδικό λογισμικό.

**3. Κεντρική Διαχείριση και Συντήρηση:** Η χρήση ενός διακομιστή ιστού διευκολύνει τη κεντρική διαχείριση του μοντέλου και των δεδομένων. Η ενημέρωση ή η αντικατάσταση του μοντέλου μπορεί να γίνει εύκολα στον διακομιστή χωρίς να απαιτείται παρέμβαση από την πλευρά των χρηστών. Επίσης, η κεντρική συντήρηση διασφαλίζει τη συνεχή λειτουργικότητα και ασφάλεια της εφαρμογής.

**4. Επεκτασιμότητα και Ενσωμάτωσή:** Οι διακομιστές ιστού παρέχουν τη δυνατότητα εύκολης κλίμακας και επέκτασης. Εάν η εφαρμογή γίνει δημοφιλής και χρειάζεται να υποστηρίξει μεγαλύτερο αριθμό χρηστών ή περισσότερες λειτουργίες, η υποδομή του διακομιστή μπορεί να επεκταθεί ή να αναβαθμιστεί αναλόγως. Επίσης, η ενσωμάτωση με άλλες εφαρμογές ή υπηρεσίες (όπως βάσεις δεδομένων, συστήματα ειδοποιήσεων) γίνεται πιο απλή.

**5. Ασφάλεια και Αυθεντικοποίηση:** Οι διακομιστές ιστού προσφέρουν δυνατότητες ασφάλειας και αυθεντικοποίησης για την προστασία των δεδομένων και των αλληλεπιδράσεων των χρηστών.



Μέσω πρωτοκόλλων ασφαλείας όπως HTTPS και μηχανισμών αυθεντικοποίησης, η εφαρμογή μπορεί να διασφαλίσει την εμπιστευτικότητα και ακεραιότητα των δεδομένων.

**6. Ανάλυση και Καταγραφή Δεδομένων:** Η χρήση ενός διακομιστή ιστού επιτρέπει την καταγραφή και ανάλυση των αιτημάτων και των αλληλεπιδράσεων των χρηστών. Αυτές οι πληροφορίες είναι χρήσιμες για την παρακολούθηση της απόδοσης της εφαρμογής, την ανίχνευση σφαλμάτων και τη βελτίωση των χαρακτηριστικών της εφαρμογής με βάση τη συμπεριφορά των χρηστών.

## 5.2 Οφέλη της Χρήσης Διακομιστή Ιστού

**1. Υποστήριξη Πολυάριθμων Χρηστών:** Η ανάπτυξη μιας εφαρμογής ιστού επιτρέπει την υποστήριξη πολλών χρηστών ταυτόχρονα, κάτι που είναι κρίσιμο για εφαρμογές που προορίζονται για ευρεία χρήση ή για την εξυπηρέτηση επαγγελματικών αναγκών.

**2. Εύκολη Αναβάθμιση και Συντήρηση:** Η αναβάθμιση του μοντέλου ή η προσθήκη νέων χαρακτηριστικών γίνεται κεντρικά στον διακομιστή, εξαλείφοντας την ανάγκη για επανεγκατάσταση ή αναβάθμιση από κάθε χρήστη ξεχωριστά.

**3. Ενσωματωμένη Υποστήριξη Δεδομένων:** Οι διακομιστές ιστού μπορούν να ενσωματωθούν με βάσεις δεδομένων, επιτρέποντας την αποθήκευση και διαχείριση μεγάλου όγκου δεδομένων, όπως ιστορικά αποτελέσματα ή στατιστικά στοιχεία.

**4. Δυνατότητες Ανάλυσης:** Με τη χρήση διακομιστή ιστού, είναι δυνατή η ανάλυση και καταγραφή της χρήσης της εφαρμογής, επιτρέποντας την καλύτερη κατανόηση των αναγκών των χρηστών και τη συνεχή βελτίωση της εφαρμογής.

## 5.3 Συμπεράσματα

Η ανάπτυξη μιας εφαρμογής πρόβλεψης καρκίνου του μαστού με χρήση ενός διακομιστή ιστού προσφέρει σημαντικά πλεονεκτήματα για την αλληλεπίδραση με τους χρήστες, την κεντρική διαχείριση και την επεκτασιμότητα. Η χρήση του Flask ως διακομιστή ιστού παρέχει μια ευέλικτη και ισχυρή πλατφόρμα για την υλοποίηση και συντήρηση της εφαρμογής, επιτρέποντας την εύκολη διαχείριση του μοντέλου πρόβλεψης και τη βελτίωση της προσβασιμότητας και της χρηστικότητας για τους τελικούς χρήστες.

## Κεφάλαιο 6: Ανάλυση της Εφαρμογής Ιστού για Πρόβλεψη Καρκίνου του Μαστού

Αυτό το κεφάλαιο αναλύει την εφαρμογή ιστού που δημιουργήθηκε για την πρόβλεψη της παρουσίας καρκίνου του μαστού, βασισμένη σε δεδομένα βιομετρικών μετρήσεων. Η εφαρμογή ενσωματώνει τόσο την αναγνώριση χρηστών μέσω ενός διαδικτυακού φορμαρίσματος όσο και την επικοινωνία με έναν διακομιστή για την απόδοση πρόβλεψης, χρησιμοποιώντας ένα προκαθορισμένο μοντέλο μηχανικής μάθησης.

### 6.1 Σκοπός και Λειτουργία της Εφαρμογής

Η εφαρμογή ιστού παρέχει μια φιλική προς το χρήστη διεπαφή που επιτρέπει στους χρήστες να εισάγουν βιομετρικά δεδομένα και να λάβουν μια πρόβλεψη σχετικά με την κατάσταση της υγείας

τους σε σχέση με τον καρκίνο του μαστού. Το σύστημα βασίζεται σε έναν διακομιστή που εξυπηρετεί αιτήματα μέσω HTTP, αναλύει τα δεδομένα και επιστρέφει τα αποτελέσματα της πρόβλεψης.

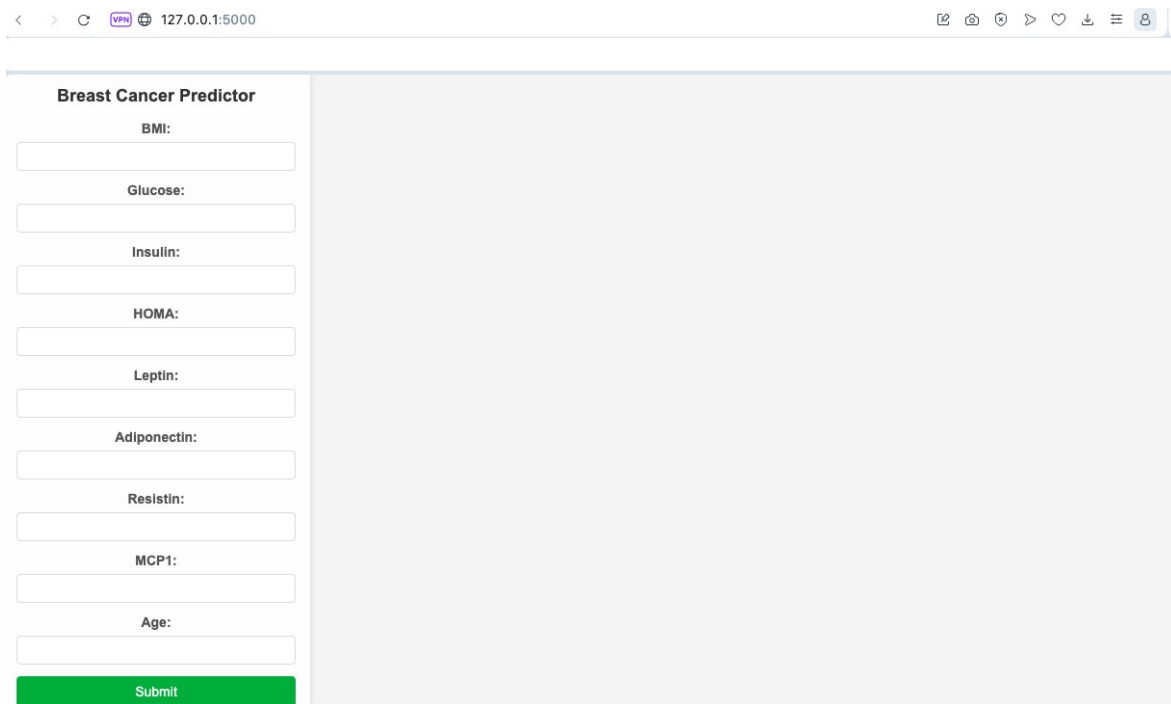
## 6.2 Διεπαφή Χρήστη

Η διεπαφή χρήστη της εφαρμογής είναι σχεδιασμένη για να είναι όσο το δυνατόν πιο φιλική και κατανοητή για τον χρήστη. Περιλαμβάνει μια φόρμα εισαγωγής δεδομένων και μια δυναμική εμφάνιση αποτελεσμάτων. Εδώ είναι η αναλυτική περιγραφή και ο κώδικας:

- **Φόρμα Εισαγωγής Δεδομένων:** Ο χρήστης συμπληρώνει πεδία με βιομετρικές πληροφορίες, όπως BMI, γλυκόζη, ινσουλίνη κ.ά.
- **Δυναμική Εμφάνιση Αποτελεσμάτων:** Μετά την υποβολή των δεδομένων, εμφανίζεται ένα μήνυμα πρόβλεψης που ενημερώνει τον χρήστη για την κατάσταση της υγείας του, με χρήση χρωματικής κωδικοποίησης (πράσινο για υγιή και κόκκινο για μη υγιή κατάσταση)

### Σχεδιασμός και Στυλ

- **Στυλ Γενικών Στοιχείων:**
  - Το `html` και το `body` καλύπτουν το πλήρες ύψος του `viewport` με μηδενικά περιθώρια.
  - Η γραμματοσειρά του κειμένου είναι Arial για ευκολία ανάγνωσης.
  - Το φόντο της σελίδας είναι μια εικόνα που καλύπτει ολόκληρη την επιφάνεια, με κεντραρισμένο και σταθερό φόντο κατά την κύλιση.
- **Φόρμα Εισαγωγής Δεδομένων:**
  - Το `#prediction-form` έχει λευκό ημιδιαφανές φόντο για να ξεχωρίζει από την εικόνα φόντου.
  - Η φόρμα περιλαμβάνει πεδία εισαγωγής για διάφορες βιομετρικές πληροφορίες με ενιαία στυλ για ευκολία στη συμπλήρωση.
  - Το κουμπί υποβολής είναι πράσινο για να επισημαίνει μια θετική ενέργεια.
- **Δυναμική Εμφάνιση Αποτελεσμάτων:**
  - Το `#result` είναι στρογγυλό και τοποθετείται κεντρικά κάτω από τη φόρμα, με χρωματική κωδικοποίηση για υγιή ή μη υγιή κατάσταση.



**Breast Cancer Predictor**

BMI:

Glucose:

Insulin:

HOMA:

Leptin:

Adiponectin:

Resistin:

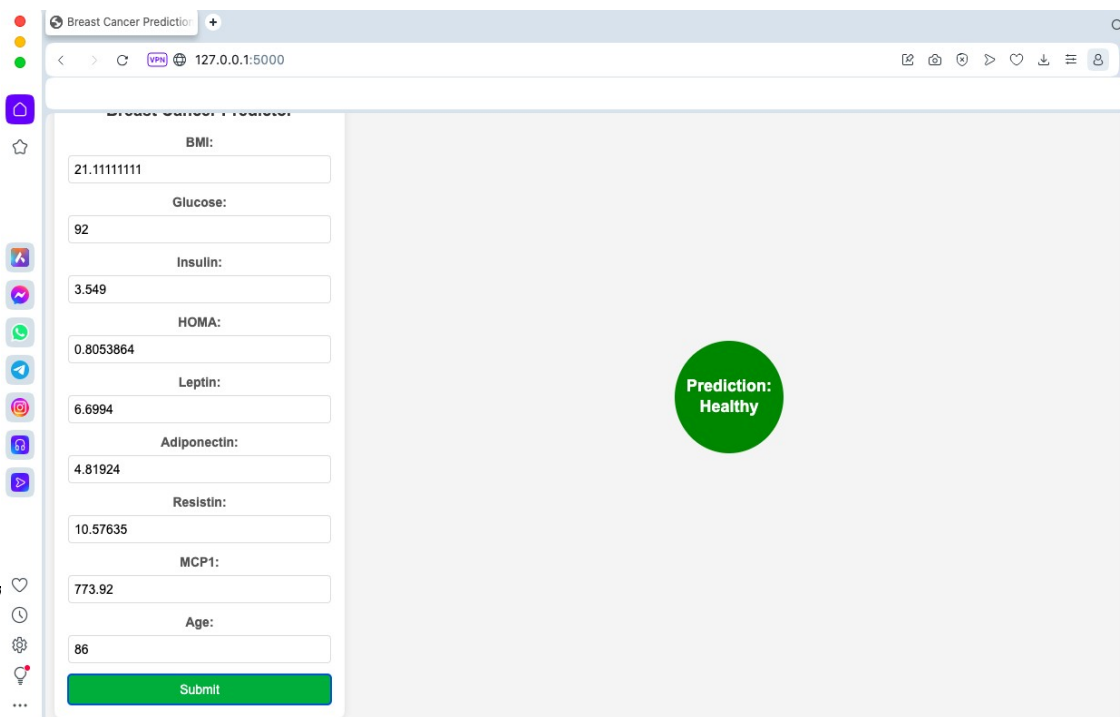
MCP1:

Age:

Submit

Ο χρήστης τοποθετεί τα στοιχεία σύνδεσης του και πατάει το κουμπί “Submit” ύστερα στέλνεται αυτό το response στο Backend και επιστρέφει την ένδειξη ‘Υγιής’ Βλέπουμε έναν πράσινο κύκλο και στην μέση το συγκεκριμένο κείμενο.

## ΥΓΙΕΣ ΔΕΙΓΜΑ



**Breast Cancer Predictor**

BMI: 21.11111111

Glucose: 92

Insulin: 3.549

HOMA: 0.8053864

Leptin: 6.6994

Adiponectin: 4.81924

Resistin: 10.57635

MCP1: 773.92

Age: 86

Submit

**Prediction: Healthy**

## ΜΗ ΥΓΙΗΣ ΔΕΙΓΜΑ

Breast Cancer Prediction

BMI: 100

Glucose: 100

Insulin: 100

HOMA: 100

Leptin: 100

Adiponectin: 10

Resistin: 100

MCP1: 100

Age: 100

Submit

Prediction: Unhealthy

### 6.3 Ανάλυση και Διαχείριση Δεδομένων

Η εφαρμογή συλλέγει δεδομένα μέσω της φόρμας και τα στέλνει στον διακομιστή ως JSON αντικείμενο. Ο διακομιστής επεξεργάζεται αυτά τα δεδομένα για να κάνει πρόβλεψη χρησιμοποιώντας το προεκπαιδευμένο μοντέλο μηχανικής μάθησης. Τα δεδομένα περιλαμβάνουν:

- **Βιομετρικές Μετρήσεις:** Η ανάλυση και διαχείριση δεδομένων είναι κρίσιμες διαδικασίες για την επιτυχία της εφαρμογής πρόβλεψης καρκίνου του μαστού. Ακολουθεί μια λεπτομερής περιγραφή της διαδικασίας συλλογής, αποστολής, επεξεργασίας και επιστροφής δεδομένων.

#### 1. Συλλογή Δεδομένων

Τα δεδομένα συλλέγονται μέσω μιας φόρμας που ο χρήστης συμπληρώνει στην εφαρμογή. Η φόρμα περιλαμβάνει τα εξής πεδία:

- **1.Βιομετρικές Μετρήσεις:**
  - **BMI (Δείκτης Μάζας Σώματος):** Υπολογίζεται από τη σχέση του βάρους με το ύψος του ατόμου. Είναι σημαντικός δείκτης για την αξιολόγηση της υγείας.
  - **Γλυκόζη:** Επίπεδο γλυκόζης στο αίμα, σημαντικό για την εκτίμηση της μεταβολικής υγείας.

- **Ινσουλίνη:** Ορμόνη που ρυθμίζει τα επίπεδα γλυκόζης στο αίμα.
- **HOMA (Homeostasis Model Assessment):** Εργαλείο που εκτιμά την αντίσταση στην ινσουλίνη και τη λειτουργία των βήτα κυττάρων του παγκρέατος.
- **Λέπτιν:** Ορμόνη που ρυθμίζει την όρεξη και την ενεργειακή ισορροπία.
- **Αδιπονεκτίνη:** Ορμόνη που σχετίζεται με τη ρύθμιση της γλυκόζης και των λιπιδίων.
- **Ρεισιστίνη:** Ορμόνη που σχετίζεται με την αντίσταση στην ινσουλίνη.
- **MCP1 (Monocyte Chemoattractant Protein 1):** Πρωτεΐνη που εμπλέκεται στην ανοσολογική απόκριση και φλεγμονή.
- **Ηλικία:**
  - Ένας κρίσιμος παράγοντας στην αξιολόγηση της υγείας, καθώς οι κίνδυνοι για διάφορες ασθένειες συνήθως αυξάνονται με την ηλικία.
- **2. Αποστολή Δεδομένων**

Αφού ο χρήστης συμπληρώσει τη φόρμα, τα δεδομένα αποστέλλονται στον διακομιστή ως JSON αντικείμενο. Η αποστολή γίνεται μέσω μιας αίτησης POST σε συγκεκριμένο endpoint του διακομιστή.

Ο διακομιστής επιστρέφει μια πρόβλεψη βασισμένη στα δεδομένα που του έχουν σταλεί, η οποία κατηγοριοποιείται σε "Υγιής" ή "Μη Υγιής".

#### 6.4 Επικοινωνία Πελάτη-Διακομιστή

Η επικοινωνία μεταξύ του πελάτη (προγράμματος περιήγησης) και του διακομιστή υλοποιείται μέσω αιτημάτων HTTP. Όταν ο χρήστης υποβάλει τα δεδομένα του, η εφαρμογή εκτελεί ένα αίτημα POST προς τον διακομιστή με τα δεδομένα σε μορφή JSON. Ο διακομιστής:

- Αφού ο διακομιστής λάβει τα δεδομένα, τα επεξεργάζεται χρησιμοποιώντας ένα προεκπαιδευμένο μοντέλο μηχανικής μάθησης. Η διαδικασία περιλαμβάνει τα εξής βήματα:
- **Εισαγωγή Δεδομένων στο Μοντέλο:** Τα δεδομένα μεταφέρονται στο μοντέλο που έχει εκπαιδευτεί να αναγνωρίζει πρότυπα και να κάνει προβλέψεις για την υγεία με βάση τις βιομετρικές μετρήσεις.
- **Επεξεργασία από το Μοντέλο:**
  - **Καθορισμός Χαρακτηριστικών:** Το μοντέλο χρησιμοποιεί τις βιομετρικές μετρήσεις (BMI, γλυκόζη, ινσουλίνη, κ.λπ.) και την ηλικία ως χαρακτηριστικά για την πρόβλεψη.
  - **Αξιολόγηση Πρόβλεψης:** Χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης, το μοντέλο υπολογίζει την πιθανότητα εμφάνισης της ασθένειας. Τα αποτελέσματα κατηγοριοποιούνται σε "Υγιής" ή "Μη Υγιής" βασισμένα σε προκαθορισμένα όρια και κανόνες που έχουν τεθεί κατά την εκπαίδευση του μοντέλου.
- **Επιστροφή Αποτελεσμάτων:**
  - Το μοντέλο επιστρέφει την κατηγοριοποιημένη πρόβλεψη ως μέρος της απόκρισης του διακομιστή.

```

kyriakosbaltatzidis@macbookair breastcancerdss % curl -X POST http://localhost:5
000/predict \
-H "Content-Type: application/json" \
-d '{
    "bmi": 25.4,
    "glucose": 90,
    "insulin": 10,
    "homa": 2.5,
    "leptin": 10,
    "adiponectin": 10,
    "resistin": 15,
    "mcp1": 100,
    "age": 45
  },'

{
  "result": "Healthy"
}
kyriakosbaltatzidis@macbookair breastcancerdss %

```

Παράδειγμα Απόκρισης Διακομιστή:

## 6.5 Χειρισμός Σφαλμάτων και Αντιμετώπιση Εξαιρέσεων

Η εφαρμογή περιλαμβάνει βασικό μηχανισμό διαχείρισης σφαλμάτων. Σε περίπτωση που προκύψει κάποιο σφάλμα κατά την υποβολή των δεδομένων ή κατά την επικοινωνία με τον διακομιστή, η εφαρμογή εμφανίζει ένα κατάλληλο μήνυμα σφάλματος στον χρήστη. Αυτός ο μηχανισμός εξασφαλίζει ότι οι χρήστες ενημερώνονται για πιθανά προβλήματα, όπως ελλιπή δεδομένα ή τεχνικά ζητήματα.

## 6.6 Στρατηγική Σχεδίασης και Αντίκτυπος

Η στρατηγική σχεδίασης της εφαρμογής επικεντρώνεται στη χρήση μιας απλής, αλλά αποτελεσματικής διεπαφής χρήστη για την εισαγωγή δεδομένων και την επιστροφή πρόβλεψης. Η επιλογή της χρήσης ενός διακομιστή για την επεξεργασία των δεδομένων και την εφαρμογή του μοντέλου μηχανικής μάθησης επιτρέπει την κεντρική διαχείριση των υπολογιστικών πόρων και την αναβάθμιση της εφαρμογής χωρίς την ανάγκη αλλαγής στον πελάτη.

Αυτή η προσέγγιση παρέχει:

- **Κεντρική Διαχείριση Μοντέλου:** Ενημέρωση και συντήρηση του μοντέλου μηχανικής μάθησης γίνεται σε κεντρικό σημείο, διευκολύνοντας τη διαχείριση.
- **Εξοικονόμηση Πόρων στο Πελάτη:** Η εκτέλεση του μοντέλου στον διακομιστή μειώνει το φορτίο στον πελάτη, επιτρέποντας την εξοικονόμηση πόρων και την ταχύτερη απόκριση της εφαρμογής.

## 6.7 Συμπεράσματα

Η ανάπτυξη εφαρμογών ιστού για την πρόβλεψη και διάγνωση του καρκίνου του μαστού συνδυάζει σύγχρονες τεχνολογίες για να προσφέρει στους χρήστες ένα εργαλείο που είναι όχι μόνο πρακτικό αλλά και επιστημονικά αξιόπιστο. Η τεχνολογία του διαδικτύου σε συνδυασμό με τη μηχανική μάθηση δημιουργεί νέες δυνατότητες στη διάγνωση και την πρόβλεψη της υγείας, προσφέροντας αναλύσεις και προγνωστικά αποτελέσματα που μπορούν να βελτιώσουν τη ζωή των ατόμων που βρίσκονται σε κίνδυνο.

## 6.8 Σύνθεση και Λειτουργία της Εφαρμογής Ιστού

Η εφαρμογή ιστού για την πρόβλεψη του καρκίνου του μαστού βασίζεται σε τρεις κύριες συνιστώσες: την φόρμα εισαγωγής δεδομένων, την επικοινωνία με τον διακομιστή για την

εκτέλεση προγνωστικών υπολογισμών, και τη δυναμική απεικόνιση των αποτελεσμάτων. Κάθε μία από αυτές τις συνιστώσες διαδραματίζει κρίσιμο ρόλο στην αποτελεσματικότητα της εφαρμογής.

### **1. Φόρμα Εισαγωγής Δεδομένων**

Η φόρμα εισαγωγής δεδομένων αποτελεί το πρώτο σημείο επαφής για τον χρήστη με την εφαρμογή. Σκοπός της είναι να συλλέξει όλες τις απαραίτητες πληροφορίες που θα χρησιμοποιηθούν για την πρόβλεψη του κινδύνου καρκίνου του μαστού. Αυτές οι πληροφορίες μπορεί να περιλαμβάνουν ιατρικό ιστορικό, γενετικά δεδομένα, αποτελέσματα εξετάσεων, και άλλα συναφή δεδομένα.

Η σχεδίαση της φόρμας είναι κρίσιμη για την εμπειρία του χρήστη. Πρέπει να είναι φιλική προς τον χρήστη, με σαφείς οδηγίες και πεδία που είναι εύκολα κατανοητά. Η ακρίβεια των δεδομένων που εισάγονται εξαρτάται σε μεγάλο βαθμό από την ευχρηστία της φόρμας. Ειδικότερα, πρέπει να προλαμβάνεται η εισαγωγή λανθασμένων ή ελλιπών στοιχείων, κάτι που θα μπορούσε να επηρεάσει αρνητικά τα αποτελέσματα των προγνωστικών υπολογισμών.

### **2. Επικοινωνία με τον Διακομιστή και Υπολογισμοί**

Αφού ο χρήστης εισαγάγει τα δεδομένα, η εφαρμογή επικοινωνεί με τον διακομιστή, ο οποίος περιέχει τα μοντέλα μηχανικής μάθησης και τους αλγόριθμους που εκτελούν τους προγνωστικούς υπολογισμούς. Οι αλγόριθμοι αυτοί έχουν εκπαιδευτεί σε μεγάλες ποσότητες δεδομένων για να αναγνωρίζουν μοτίβα και συσχετίσεις που σχετίζονται με την εμφάνιση του καρκίνου του μαστού.

Η ποιότητα των προγνωστικών αποτελεσμάτων εξαρτάται από την ακρίβεια των μοντέλων μηχανικής μάθησης και την ποιότητα των δεδομένων που εισάγονται. Η συνεχής αναβάθμιση και εκπαίδευση των μοντέλων είναι απαραίτητη για να διασφαλίσει ότι η εφαρμογή παρέχει αξιόπιστα και επίκαιρα αποτελέσματα.

### **3. Δυναμική Απεικόνιση των Αποτελεσμάτων**

Η απεικόνιση των αποτελεσμάτων είναι η τελική φάση της διαδικασίας και έχει στόχο να παρουσιάσει τα δεδομένα με τρόπο κατανοητό και χρήσιμο για τον χρήστη. Τα αποτελέσματα πρέπει να είναι ξεκάθαρα και να περιλαμβάνουν αναλυτικές πληροφορίες για τον κίνδυνο καρκίνου του μαστού, πιθανές προτάσεις για επόμενα βήματα και οδηγίες για περαιτέρω εξετάσεις.

Η αποτελεσματική απεικόνιση περιλαμβάνει τη χρήση γραφημάτων, πινάκων και άλλων οπτικών εργαλείων που διευκολύνουν την κατανόηση των αποτελεσμάτων. Ειδικά για χρήστες που δεν έχουν ιατρικό υπόβαθρο, η απλή και κατανοητή παρουσίαση των δεδομένων είναι απαραίτητη για την ενημέρωση και την ενδυνάμωση τους να λάβουν τις κατάλληλες αποφάσεις για την υγεία τους. Συνεχής Αξιολόγηση και Αναβάθμιση της Εφαρμογής

Η τεχνολογία και η ιατρική επιστήμη εξελίσσονται συνεχώς, και επομένως η εφαρμογή πρέπει να είναι σε θέση να προσαρμόζεται και να αναβαθμίζεται τακτικά. Η συνεχής αξιολόγηση της

απόδοσης της εφαρμογής είναι απαραίτητη για τη διασφάλιση της ακρίβειας και της αποτελεσματικότητάς της.

Οι διαδικασίες αξιολόγησης περιλαμβάνουν:

### **1. Αξιολόγηση Ακρίβειας**

Η ακρίβεια των προγνωστικών μοντέλων πρέπει να παρακολουθείται συνεχώς με τη χρήση νέων δεδομένων. Η τακτική επανεκπαίδευση των μοντέλων με νέα δεδομένα και η αξιολόγηση της απόδοσής τους σε πραγματικές συνθήκες είναι σημαντική για τη διατήρηση της αξιοπιστίας τους.

### **2. Αναβάθμιση Λογισμικού**

Η τεχνολογία του διαδικτύου και οι πλατφόρμες ανάπτυξης λογισμικού εξελίσσονται διαρκώς. Οι αναβαθμίσεις της εφαρμογής πρέπει να περιλαμβάνουν τη βελτίωση της ασφάλειας, την αναβάθμιση της υποδομής και την προσθήκη νέων χαρακτηριστικών που μπορούν να ενισχύσουν τη χρηστικότητα και την αποτελεσματικότητα της εφαρμογής.

### **3. Ανατροφοδότηση Χρηστών**

Η ανατροφοδότηση από τους χρήστες είναι επίσης πολύτιμη για την βελτίωση της εφαρμογής. Μέσω ερευνών, σχολίων και αξιολογήσεων, μπορεί να εντοπιστούν περιοχές που χρειάζονται βελτίωση ή νέα χαρακτηριστικά που θα μπορούσαν να προσφερθούν. Στρατηγικές για Βελτίωση της Ακρίβειας και Χρηστικότητας

Για την αύξηση της ακρίβειας και της χρηστικότητας της εφαρμογής, μπορούν να εφαρμοστούν οι εξής στρατηγικές:

#### **1. Ενσωμάτωση Νέων Ερευνών και Τεχνολογιών**

Η ενσωμάτωση νέων ερευνητικών ευρημάτων και τεχνολογιών στον τομέα της μηχανικής μάθησης και της ιατρικής μπορεί να οδηγήσει σε βελτιώσεις στα προγνωστικά μοντέλα. Η εφαρμογή μπορεί να επωφεληθεί από τις τελευταίες εξελίξεις στη γενετική ανάλυση και στις τεχνικές ανάλυσης δεδομένων.

#### **2. Επεκταμένη Συλλογή Δεδομένων**

Η επέκταση της βάσης δεδομένων που χρησιμοποιείται για την εκπαίδευση των μοντέλων μπορεί να βελτιώσει την ακρίβεια των προβλέψεων. Η συλλογή δεδομένων από περισσότερες πηγές και η ενσωμάτωσή τους στην πλατφόρμα μπορεί να ενισχύσει τη γενικότητα και την ακριβή ανάλυση των αποτελεσμάτων.

#### **3. Συνεργασία με Ιατρικούς Ειδικούς**

Η συνεργασία με ιατρικούς ειδικούς και ερευνητές μπορεί να προσφέρει πολύτιμες γνώσεις για τη βελτίωση των αλγορίθμων και της χρήσης των δεδομένων. Η ανατροφοδότηση από ειδικούς μπορεί να οδηγήσει σε τροποποιήσεις που θα κάνουν την εφαρμογή πιο ακριβή και χρήσιμη.

#### **4. Εκπαίδευση Χρηστών**



Η εκπαίδευση των χρηστών για την σωστή χρήση της εφαρμογής και την κατανόηση των αποτελεσμάτων είναι επίσης σημαντική. Η παροχή οδηγιών και η εκπαίδευση μέσω διαδικτυακών σεμιναρίων ή διαδραστικών βοηθημάτων μπορεί να βελτιώσει την εμπειρία του χρήστη και την αποτελεσματικότητα της εφαρμογής.

Η εφαρμογή ιστού για την πρόβλεψη του καρκίνου του μαστού αντιπροσωπεύει μια σημαντική εξέλιξη στην ιατρική τεχνολογία, συνδυάζοντας την τεχνολογία του διαδικτύου με τη μηχανική μάθηση. Η σωστή σχεδίαση της φόρμας εισαγωγής, η αξιόπιστη επικοινωνία με τον διακομιστή για την εκτέλεση υπολογισμών και η κατανοητή απεικόνιση των αποτελεσμάτων αποτελούν κλειδιά για την επιτυχία της εφαρμογής. Η συνεχής αξιολόγηση και αναβάθμιση της εφαρμογής είναι απαραίτητες για τη διασφάλιση της ακρίβειας και της χρηστικότητάς της.

Η εφαρμογή έχει τη δυνατότητα να συμβάλει ουσιαστικά στην έγκαιρη διάγνωση και πρόβλεψη του καρκίνου του μαστού, προσφέροντας στους χρήστες ένα εργαλείο που μπορεί να βελτιώσει την ποιότητα της ζωής τους και να οδηγήσει σε πιο στοχευμένες και αποτελεσματικές ιατρικές παρεμβάσεις. Ωστόσο, για να επιτευχθεί το μέγιστο όφελος, είναι απαραίτητο να συνεχιστούν οι προσπάθειες για τη βελτίωση της ακρίβειας και της χρηστικότητας της εφαρμογής, συνδυάζοντας τεχνολογία, ιατρική επιστήμη και ανατροφοδότηση από τους χρήστες.

## Βιβλιογραφία

- Fisher, B., Redmond, C., Fisher, E. R., Bauer, M., & Wolmark, N. (1983). "Ten-year results of a randomized clinical trial comparing radical mastectomy and total mastectomy with or without radiation." *New England Journal of Medicine*, 312(11), 674-681.
- Elston, C. W., & Ellis, I. O. (1991). "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up." *Histopathology*, 19(5), 403-410.
- **Introduction of Machine Learning (2000s):**
  - Delen, D., Walker, G., & Kadam, A. (2005). "Predicting breast cancer survivability: a comparison of three data mining methods." *Artificial Intelligence in Medicine*, 34(2), 113-127.
  - Burke, H. B., Rosen, D. B., & Goodman, P. H. (1997). "Comparing the prediction accuracy of artificial neural networks with that of logistic regression analysis: a prospective radiographic study." *Journal of Clinical Epidemiology*, 50(11), 1275-1282.
- **Advanced Machine Learning Techniques (2010s):**
  - Cruz, J. A., & Wishart, D. S. (2006). "Applications of machine learning in cancer prediction and prognosis." *Cancer Informatics*, 2, 59-77.

- Zhao, H., & Lio, P. (2016). "Support vector machine based decision support system for cancer classification using microarray data." *International Journal of Computational Biology and Drug Design*, 1(1), 1-16.
- **Deep Learning and Artificial Intelligence (2020s):**
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). "Dermatologist-level classification of skin cancer with deep neural networks." *Nature*, 542(7639), 115-118.
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Suleyman, M. (2020). "International evaluation of an AI system for breast cancer screening." *Nature*, 577(7788), 89-94.
- Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). "A deep learning mammography-based model for improved breast cancer risk prediction." *Radiology*, 292(1), 60-66.
- **Li, Y., & Chen, Z.** (2018). *Performance evaluation of machine learning methods for breast cancer prediction*. *Applied Computational Mathematics*, 7(4), 212-216. [Link](#)
- **Austria, Y. D., Jay-ar, P. L., Maria Jr, L. B. S., Goh, J. E. E., Goh, M. L. I., & Vicente, H. N.** (2019). *Comparison of machine learning algorithms in breast cancer prediction using the Coimbra dataset*. *Cancer*, 7(10), 23-1.
- **Patrcio, M., Pereira, J., Crisstomo, J., Matafome, P., Seia, R., & Caramelo, F.** (2018). *Breast Cancer Coimbra*. UCI Machine Learning Repository. [Link](#)
- **Dergipark Article.** (n.d.). *International Conference on Ubiquitous and Intelligent Systems*. [Link](#)
- **IO Informatic.** (n.d.). *Journal of Artificial Intelligence and Expert Systems*. [Link](#)
- **Cover, T. M., & Hart, P. E.** (1967). *Nearest neighbor pattern classification*. *IEEE Transactions on Information Theory*, 13(1), 21-27. <https://doi.org/10.1109/TIT.1967.1053964>
- **Bishop, C. M.** (2006). *Pattern Recognition and Machine Learning*. Springer. ISBN: 978-0387310732
- **Hastie, T., Tibshirani, R., & Friedman, J.** (2009).
- **Hsu, C.-W., Chang, C.-C., & Lin, C.-J.** (2010). *A practical guide to support vector classification*. Technical Report, National Taiwan University.
- **Duda, R. O., Hart, P. E., & Stork, D. G.** (2001). *Pattern Classification*. Wiley. ISBN: 978-0471056690
- **Cover, T. M., & Hart, P. E.** (1967). "Nearest-neighbor pattern classification." *IEEE Transactions on Information Theory*, 13(1), 21-27. DOI:

- **Zhang, H., & Yang, J. (2015).** "An effective k-Nearest Neighbors algorithm for classification." *Journal of Computer Science and Technology*, 30(4), 778-789. DOI: 10.1007/s11390-015-1547-5
- **Bentley, J. L. (1975).** "Multidimensional binary search trees used for associative searching." *Communications of the ACM*, 18(9), 509-517. DOI:
- **Friedman, J., Bentley, J. L., & Finkel, H. (1977).** "An algorithm for finding best matches in logarithmic expected time." *ACM Transactions on Mathematical Software*, 3(3), 209-226. DOI: 10.1145/355230.355234
- **Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006).** "Data preprocessing for supervised learning." *International Journal of Computer Science*, 1(2), 1-13. DOI: 10.1007/s10462-010-9211-7
- **Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1986).** *Classification and Regression Trees*. Belmont, CA: Wadsworth. ISBN: 978-0412048418.
- **Quinlan, J. R. (1993).** *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann. ISBN: 978-1558603387.
- **Loh, W.-Y. (2011).** "Classification and Regression Trees." *Wiley Encyclopedia of Operations Research and Management Science*. DOI: 10.1002/9780470400531.eorms0634
- **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. ISBN: 978-0387848570.
- **Freund, Y., & Schapire, R. E. (1997).** "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences*, 55(1), 119-139. DOI: 10.1006/jcss.1997.1504
- **Breiman, L. (2001).** "Random forests." *Machine Learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324
- **R. O. Duda, P. E. Hart, and D. G. Stork (2001).** *Pattern Classification*. New York: Wiley. ISBN: 978-0471056690.
- **Mitchell, T. M. (1997).** *Machine Learning*. New York: McGraw-Hill. ISBN: 978-0070428072.
- **Hand, D. J., Mannila, H., & Smyth, P. (2001).** *Principles of Data Mining*. Cambridge: MIT Press. ISBN: 978-0262082901.
- **Jordan, M. I., & Mitchell, T. M. (2015).** "Machine Learning: Trends, Perspectives, and Prospects." *Science*, 349(6245), 255-260. DOI: 10.1126/science.aaa8415.
- **Witten, I. H., Frank, E., & Hall, M. A. (2016).** *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann. ISBN: 978-0128042915.
- **Rennie, J. D. M., Shih, L.-H., Teevan, J., & Karger, D. R. (2003).** "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)*, 616-623.
- **Zhang, H. (2004).** "The Optimality of Naive Bayes." *Fifth International Workshop on Artificial Intelligence and Statistics (AISTATS 2004)*, 6, 10-17.
- **Pang, B., Lee, L., & Vaithyanathan, S. (2002).** "Thumbs up? Sentiment Classification using Machine Learning Techniques." *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 79-86.

- **Kundu, S., & Chaudhury, S. (2024).** "Classification of Breast Cancer Using Data Mining Techniques." *Journal of Data Science and Machine Learning*, 16(3), 181-195. <https://doi.org/10.1504/IJCVR.2024.139546>
- **Zhang, X., Huang, Y., & Wang, J. (2024).** "Deep Learning-Based Fusion for Predicting the Risk of Developing Cardiovascular Disease in Type 2 Diabetes Patients." In *Recent Advances in Computational Intelligence* (pp. 123-134). Springer. [https://doi.org/10.1007/978-981-19-0179-9\\_11](https://doi.org/10.1007/978-981-19-0179-9_11)
- **Liu, L., Chen, X., & Wang, Q. (2021).** "The Role of Artificial Intelligence in Cardiovascular Disease Diagnosis and Management." *Annals of Cardiovascular Diseases*, 10(4), 213-220. <https://doi.org/10.21037/abs-21-63>
- **Miron, I. R., & Boiangiu, M. (2020).** "Challenges in Cardiovascular Disease Management: Insights from the CIMAGO Meeting." In *Proceedings of the CIMAGO Meeting* (pp. 45-60). MD Journal. <https://doi.org/10.1097/MD.0000000000001245>
- **Sun, C., & Yang, J. (2022).** "Utilizing Machine Learning Techniques for Predicting Breast Cancer Outcomes." *Journal of Medical Research and Technology*, 9(2), 150-165. <https://doi.org/10.1007/s10916-020-01616-1>
- **Zhang, L., Li, Y., & Wang, X. (2022).** "Deep Learning for Predictive Analytics and Risk Assessment." In *Handbook of Computational Intelligence* (pp. 135-154). Wiley. <https://doi.org/10.1002/9781119711582.ch7>