

Ηλεκτρονική Υγεία

**ΣΥΣΤΗΜΑ ΠΡΟΒΛΕΨΗΣ ΚΑΡΔΙΟΛΟΓΙΚΩΝ ΠΑΘΗΣΕΩΝ ΜΕ ΤΗ  
ΧΡΗΣΗ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ**

**Κωνσταντίνος Δημητρίου**

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ**



**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**Απρίλιος 2021**

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**Σύστημα πρόβλεψης καρδιολογικών παθήσεων με τη χρήση μηχανικής  
μάθησης**

**Κωνσταντίνος Δημητρίου**

constandinosdemetriou@gmail.com

Απρίλιος 2021

# Περιεχόμενα

<b>Κεφάλαιο 1</b>	<b>Εισαγωγή.....</b>	<b>1</b>
<b>Κεφάλαιο 2</b>	<b>Σχετική εργασία.....</b>	<b>2</b>
2.1	Σύστημα πρόβλεψης καρδιακών παθήσεων με νευρωνικά δίκτυα.....	2
2.2	Μοντέλα πρόβλεψης καρδιακών παθήσεων για ασθενείς στην ανατολική Κίνα.....	2
2.3	Αλγόριθμοι μηχανικής μάθησης για ιατρικές διαγνώσεις.....	3
<b>Κεφάλαιο 3</b>	<b>Δεδομένα.....</b>	<b>4</b>
3.1	Επεξήγηση δεδομένων .....	4
3.2	Αναπαράσταση δεδομένων.....	6
3.3	Κανονικοποίηση δεδομένων.....	9
3.4	Συσχετίσεις δεδομένων .....	10
<b>Κεφάλαιο 4</b>	<b>Αλγόριθμοι μηχανικής μάθησης.....</b>	<b>12</b>
4.1	Logistic Regression .....	12
4.2	k-Nearest Neighbors (kNN).....	13
4.3	Multilayer Perceptron (MLP) .....	14
4.4	Decision Tree .....	15
4.5	Random Forest .....	16
4.6	Gaussian Naive Bayes .....	17
<b>Κεφάλαιο 5</b>	<b>Αξιολόγηση μοντέλων .....</b>	<b>19</b>
5.1	Accuracy .....	19
5.2	Διαχωρισμός δεδομένων σε train και test.....	19
5.3	Cross validation.....	20
5.4	Grid Search .....	21
5.5	ROC curve .....	23
<b>Κεφάλαιο 6</b>	<b>Πειραματικά αποτελέσματα.....</b>	<b>25</b>
6.1	Αποτελέσματα grid search.....	25
6.2	Αποτελέσματα cross validation .....	25
6.3	Ακρίβεια στα test δεδομένα.....	26
6.4	Area Under the curve .....	28
<b>Κεφάλαιο 7</b>	<b>Σύνδεση με την πανδημία COVID-19 .....</b>	<b>30</b>

<b>Κεφάλαιο 8 Συμπεράσματα.....</b>	<b>31</b>
<b>Βιβλιογραφία.....</b>	<b>32</b>
<b>ΠΑΡΑΡΤΗΜΑ Οδηγίες χρήσης συστήματος .....</b>	<b>33</b>

# Λίστα με τα σχήματα

Σχήμα 1: Ποσοστό ατόμων με και χωρίς καρδιακά προβλήματα.....	6
Σχήμα 2: Ποσοστό ανδρών και γυναικών στα δεδομένα .....	7
Σχήμα 3: Ιστόγραμμα ατόμων με και χωρίς καρδιακά προβλήματα ανά φύλο.....	7
Σχήμα 4: Ιστόγραμμα ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία.....	8
Σχήμα 5: Κατανομή ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία .....	8
Σχήμα 6: Κατανομή ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία ανά φύλο.....	8
Σχήμα 7: Ποσοστό ατόμων με και χωρίς καρδιακά προβλήματα ανά τύπο πόνου στο στήθος.....	9
Σχήμα 8: Correlation analysis μεταξύ όλων των features στα δεδομένα .....	10
Σχήμα 9: Απόλυτη τιμή του correlation analysis μεταξύ των features και του target .....	11
Σχήμα 10: Γραφική παράσταση sigmoid function.....	12
Σχήμα 11: Παράδειγμα εισαγωγής νέου query το οποίο θα ταξινομηθεί με τον αλγόριθμο 5-nearest neighborhoods.....	14
Σχήμα 12: Γραφική αναπαράσταση ενός multilayer perceptron νευρωνικό δίκτυο.....	15
Σχήμα 13: Παράδειγμα decision tree .....	16
Σχήμα 14: Τρόπος λειτουργίας random forest.....	17
Σχήμα 15: Διαχωρισμός των δεδομένων σε train και test .....	20
Σχήμα 16: Εικονική αναπαράσταση 5-fold cross validation .....	21
Σχήμα 17: True positive rate και False positive rate σε διαφορετικά classification thresholds ...	23
Σχήμα 18: Area under the ROC Curve.....	24
Σχήμα 19: Confusion matrixes που προέκυψαν για κάθε μοντέλο αξιολογώντας τα test δεδομένα .....	27
Σχήμα 20: ROC curve για τα test δεδομένα .....	29
Σχήμα 21: Διαπροσωπία του συστήματος.....	33
Σχήμα 22: Διαπροσωπία του συστήματος συμπληρωμένη με τα δεδομένα ασθενή χωρίς καρδιακή πάθηση .....	34
Σχήμα 23: Μήνυμα που παρουσιάζεται όταν το σύστημα προβλέψει ότι ο ασθενής δεν αντιμετωπίζει κάποια καρδιακή πάθηση.....	34
Σχήμα 24: Διαπροσωπία του συστήματος συμπληρωμένη με τα δεδομένα ασθενή με καρδιακή πάθηση.....	35
Σχήμα 25: Μήνυμα που παρουσιάζεται όταν το σύστημα προβλέψει ότι ο ασθενής αντιμετωπίζει κάποια καρδιακή πάθηση.....	35

# Λίστα με τους πίνακες

Πίνακας 1: Παράμετροι που διερευνήθηκαν στο grid search .....	22
Πίνακας 2: Οι βέλτιστες παράμετροι που επέλεξε το grid search .....	25
Πίνακας 3: Average accuracy και standard deviation για κάθε μοντέλο μηχανικής μάθησης.....	26
Πίνακας 4: Accuracy που προέκυψε για κάθε μοντέλο αξιολογώντας τα test δεδομένα .....	28
Πίνακας 5: ACU που προκύπτουν για κάθε μοντέλο .....	29

# Κεφάλαιο 1

## Εισαγωγή

Κατά την τελευταία δεκαετία, οι καρδιακές παθήσεις [1] γνωστές και ως καρδιαγγειακά νοσήματα παραμένουν η κύρια αιτία θανάτου παγκοσμίως. Μια εκτίμηση από τον Παγκόσμιο Οργανισμό Υγείας, είναι ότι συμβαίνουν περισσότεροι από 17,9 εκατομμύρια θάνατοι κάθε χρόνο παγκοσμίως λόγω των καρδιαγγειακών παθήσεων και από αυτούς τους θανάτους, το 80% οφείλεται στη στεφανιαία νόσο και σε εγκεφαλικό επεισόδιο. Πολλοί παράγοντες όπως προσωπικές και επαγγελματικές συνήθειες αλλά και η γενετική προδιάθεση οφείλονται σε καρδιακές παθήσεις. Συνήθης παράγοντες κινδύνου είναι το κάπνισμα, η υπερβολική κατανάλωση αλκοόλ και καφεΐνης, το άγχος και η έλλειψη σωματικής άσκησης μαζί με άλλους παράγοντες όπως η παχυσαρκία, η υπέρταση, η υψηλή χοληστερόλη στο αίμα αποτελούν παράγοντες προδιάθεσης για καρδιακές παθήσεις. Η αποτελεσματική και η έγκαιρη ιατρική διάγνωση καρδιακών παθήσεων παίζει καθοριστικό ρόλο για την πρόληψη του θανάτου.

Η μηχανική μάθηση είναι ένας από τους πιο γρήγορα εξελισσόμενους τομείς της τεχνητής νοημοσύνης. Οι αλγόριθμοι μηχανικής μάθησης [2] μπορούν να αναλύσουν τεράστιο όγκο δεδομένων από διάφορους τομείς εκ των οποίων ένας από τους σημαντικότερους είναι ο ιατρικός τομέας. Χρησιμοποιώντας τη μηχανική μάθηση μπορούμε να κατανοήσουμε πολύπλοκες και μη γραμμικές συσχετίσεις μεταξύ διαφόρων παραγόντων, μειώνοντας το σφάλμα στα προβλεπόμενα αποτελέσματα. Εξερευνώντας τεράστια σύνολα δεδομένων υπάρχει η δυνατότητα εξαγωγής κρυφών κρίσιμων πληροφοριών για τη λήψη αποφάσεων. Με τους αλγορίθμους μηχανικής μάθησης μπορούμε να κτίσουμε μοντέλα τα οποία εκπαιδεύουμε με ένα σύνολο από δεδομένα και αφού τα μοντέλα «μάθουν» μπορούμε να εκτελούμε προβλέψεις σε καινούργια δεδομένα.

Οι επαγγελματίες υγείας με τις γνώσεις τους αναλύουν αυτά τα δεδομένα για να καταλήξουν στις διαγνώσεις για τους ασθενείς τους. Ο ιατρικός τομέας περιλαμβάνει τεράστια δεδομένα ασθενών. Αυτά τα δεδομένα χρειάζονται ανάλυση από διάφορους αλγόριθμους μηχανικής μάθησης. Στην παρούσα εργασία, δοκιμάσαμε αλγόριθμους ταξινόμησης μηχανικής μάθησης για την πρόβλεψη καρδιακών παθήσεων σε ασθενείς.

# Κεφάλαιο 2

## Σχετική εργασία

### 2.1 Σύστημα πρόβλεψης καρδιακών παθήσεων με νευρωνικά δίκτυα

Οι Shah et al. (2018) στο επιστημονικό άρθρο με τίτλο “Effective heart disease prediction system using data mining techniques” [3] αναπτύξαν ένα αποτελεσματικό σύστημα πρόβλεψης καρδιακών παθήσεων χρησιμοποιώντας νευρωνικά δίκτυα για την πρόβλεψη του επιπέδου κινδύνου καρδιακών παθήσεων. Το σύστημα χρησιμοποιεί 15 ιατρικές παραμέτρους όπως ηλικία, φύλο, αρτηριακή πίεση, χοληστερόλη και παχυσαρκία για πρόβλεψη. Το σύστημα προβλέπει την πιθανότητα εμφάνισης καρδιακών παθήσεων σε ασθενείς. Υλοποίησαν μια αρχιτεκτονική νευρωνικού δικτύου πολλαπλών επιπέδων perceptron με αλγόριθμο εκπαίδευσης τον backpropagation. Το προτεινόμενο σύστημα μπορεί να εντοπίσει κρυφές γνώσεις, δηλαδή, μοτίβα και συσχετίσεις με καρδιακές παθήσεις από μια βάση δεδομένων με ιστορικά στοιχεία για ασθενείς με καρδιακές παθήσεις. Τα αποτελέσματα που προέκυψαν έδειξαν ότι το διαγνωστικό σύστημα που σχεδιάστηκε μπορεί να προβλέψει αποτελεσματικά το επίπεδο κινδύνου καρδιακών παθήσεων.

### 2.2 Μοντέλα πρόβλεψης καρδιακών παθήσεων για ασθενείς στην ανατολική Κίνα

Οι Yang et al. (2020) στο επιστημονικό άρθρο με τίτλο “Study of cardiovascular disease prediction model based on random forest in eastern China” [4] υλοποίησαν ένα μοντέλο πρόβλεψης καρδιακών παθήσεων. Η έρευνα βασίστηκε σε έναν μεγάλο μέρος του πληθυσμού με υψηλό κίνδυνο εμφάνισης καρδιακών παθήσεων στην ανατολική Κίνα χρησιμοποιώντας τον αλγόριθμο Random Forest. Σε αυτή τη μελέτη, 29930 άτομα με υψηλό κίνδυνο καρδιακών ασθενειών επιλέχθηκαν το 2014, τα οποία παρακολουθούνταν τακτικά χρησιμοποιώντας το ηλεκτρονικό σύστημα υγείας. Η ανάλυση έδειξε ότι σχεδόν 30 δείκτες συσχετίστηκαν με καρδιακά νοσήματα, συμπεριλαμβανομένων του φύλου, της ηλικίας, του οικογενειακού ιστορικού, του καπνίσματος, του αλκοόλ, της παχυσαρκίας, της υπερβολικής περιφέρειας της μέσης, της χοληστερόλης, της πυκνότητας λιποπρωτεΐνης, της γλυκόζης νηστείας στο αίμα και άλλων.



Χρησιμοποιήθηκαν αρκετές μέθοδοι για την κατασκευή του μοντέλου πρόβλεψης, όπως multivariate regression model, classification και regression tree (CART), Naïve Bayes, Bagged trees, Ada Boost και Random Forest. Τα αποτελέσματα έδειξαν ότι το Random Forest ήταν καλύτερο από τις άλλες μεθόδους με AUC 0,787.

## **2.3 Αλγόριθμοι μηχανικής μάθησης για ιατρικές διαγνώσεις**

Οι Fatima και Pasha (2017) στο επιστημονικό άρθρο με τίτλο “Survey of machine learning algorithms for disease diagnostic” [5] ανάλυσαν και σύγκριναν διαφορετικούς αλγορίθμους μηχανικής μάθησης για τη διάγνωση διαφόρων ασθενειών όπως καρδιακές παθήσεις, διαβήτης, ηπατική νόσο, δάγκειος πυρετός και ηπατίτιδα. Στη συγκεκριμένη εργασία υποστηρίζουν ότι στον τομέα της βιοϊατρικής, η αναγνώριση προτύπων και η μηχανική μάθηση υπόσχονται τη βελτιωμένη ακρίβεια στη διάγνωση της νόσου. Πολλοί αλγόριθμοι παρουσίασαν καλά αποτελέσματα. Παρατήρησαν ότι για την ανίχνευση καρδιακών παθήσεων, ο αλγόριθμος SVM παρέχει ακρίβεια 94,60%. Η νόσος του διαβήτη διαγιγνώσκεται με ακρίβεια 95% από τον αλγόριθμο Naive Bayes. Ο αλγόριθμος FT παρέχει το 97,10% της ορθότητας για τη διάγνωση της ηπατικής νόσου. Για την ανίχνευση του δάγκειου πυρετού, η ακρίβεια 100% επιτυγχάνεται με τη θεωρία RS. Τα feed forward νευρικά δίκτυα ταξινομούν σωστά την ηπατίτιδα καθώς παρέχει 98% ακρίβεια.

# Κεφάλαιο 3

## Δεδομένα

### 3.1 Επεξήγηση δεδομένων

Στη παρούσα εργασία χρησιμοποιήθηκαν δεδομένα από το “UCI Machine learning repository” και συγκεκριμένα οι βάσεις δεδομένων “Heart Disease Data Set” [6]. Οι δημιουργοί αυτών των βάσεων δεδομένων είναι:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Αυτή η βάση δεδομένων περιέχει 76 features, αλλά όλα τα δημοσιευμένα πειράματα αναφέρονται σε ένα υποσύνολο 14 από αυτά. Συγκεκριμένα η βάση δεδομένων του “cleveland” είναι η μόνη που έχει χρησιμοποιηθεί από ερευνητές μηχανικής μάθησης μέχρι σήμερα, γι’ αυτό αποφασίσαμε να χρησιμοποιήσουμε μόνο αυτήν. Η βάση δεδομένων του “cleveland” περιέχει 298 παραδείγματα με 14 features. Τα features περιέχουν βιοσήματα του ασθενή και ερμηνεύονται ως εξής:

1. **age**: ηλικία σε χρόνια
2. **sex**: φύλο
  - Τιμή 0: γυναίκα
  - Τιμή 1: άνδρας
3. **cp**: τύπος πόνου στο στήθος
  - Τιμή 0: τυπική στηθάγχη
  - Τιμή 1: άτυπη στηθάγχη
  - Τιμή 2: μη στηθαγχικό ή μη ισχαιμικό θωρακικό άλγος
  - Τιμή 3: ασυμφοματικός
4. **trestbps**: αρτηριακή πίεση σε ηρεμία σε mm Hg
5. **chol**: χοληστερόλη ολική σε mg/dl

6. **fbs**: γλυκόζη νηστείας

- Τιμή 0: γλυκόζη νηστείας  $\leq 120$  mg/dl
- Τιμή 1: γλυκόζη νηστείας  $> 120$  mg/dl

7. **restecg**: αποτελέσματα ηλεκτροκαρδιογραφήματος σε ηρεμία

- Τιμή 0: Φυσιολογικό
- Τιμή 1: ST-T wave abnormality (T wave inversions and/or ST elevation or depression of  $> 0.05$  mV)
- Τιμή 2: πιθανή ή συγκεκριμένη υπερτροφία της αριστερής κοιλίας με βάση τα κριτήρια του Estes

8. **thalach**: ο μέγιστος καρδιακός ρυθμός

9. **exang**: άσκηση που περιέχει στηθάγχη;

- Τιμή 0: η άσκηση δεν περιέχει στηθάγχη
- Τιμή 1: η άσκηση περιέχει στηθάγχη

10. **oldpeak**: ST depression induced by exercise relative to rest

11. **slope**: the slope of the peak exercise ST segment

- Τιμή 0: upsloping
- Τιμή 1: flat
- Τιμή 2: downsloping

12. **ca**: αριθμός κύριων αγγείων (0-3) που χρωματίζονται από φθοροσκόπηση

13. **thal**: μια διαταραχή του αίματος που ονομάζεται θαλασσαιμία

- Τιμή 0: Φυσιολογική
- Τιμή 1: Fixed Defect
- Τιμή 2: Reversible Defect

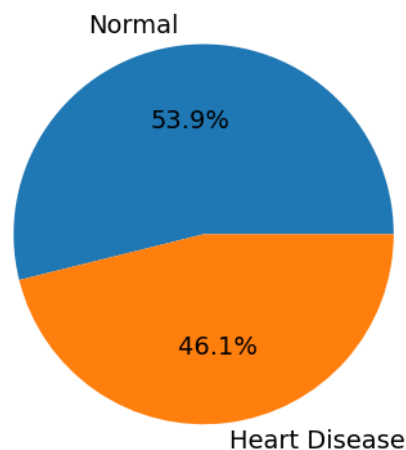
14. **target**: ο ασθενής έχει καρδιοπάθεια;

- Τιμή 0: ο ασθενής ΔΕΝ έχει καρδιοπάθεια
- Τιμή 1: ο ασθενής έχει καρδιοπάθεια

## 3.2 Αναπαράσταση δεδομένων

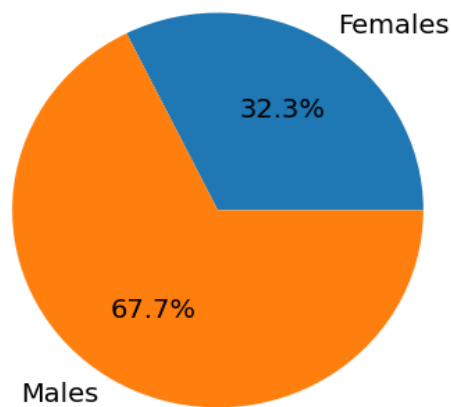
Πριν προχωρήσουμε στην επεξεργασία των δεδομένων μελετήσαμε τη μορφή τους έτσι ώστε να είμαστε βέβαιοι για τις μεθόδους που θα χρησιμοποιήσουμε. Δηλαδή έχοντας μια καλύτερη γνώση για τα υπάρχον δεδομένα θα κατευθύνουμε τη μελέτη μας προς κάποια κατεύθυνση.

Το Σχήμα 1 παρουσιάζει τη γραφική παράσταση με το ποσοστό ατόμων με και χωρίς καρδιακά προβλήματα στα δεδομένα. Παρατηρούμε ότι το 53.9% των ατόμων στα δεδομένα μας δεν αντιμετωπίζουν κάποιο καρδιακό νόσημα ενώ το 46.1% αντιμετωπίζει. Άρα τα δεδομένα μας ήταν σχεδόν ισοζυγισμένα και δε θα είχαμε πρόβλημα κατά την εκπαίδευση αλγορίθμων μηχανικής μάθησης.

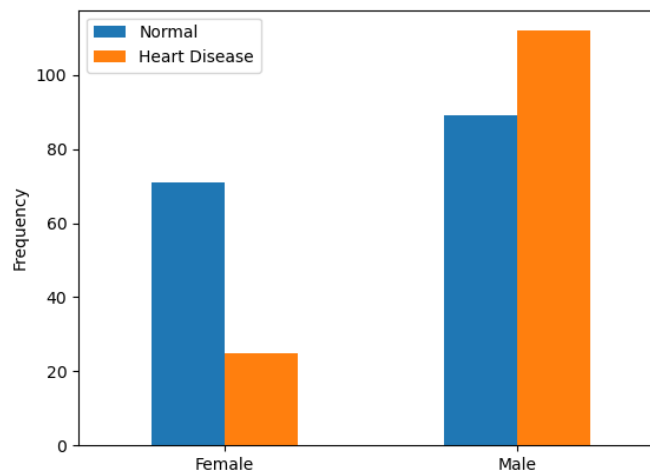


**Σχήμα 1:** Ποσοστό ατόμων με και χωρίς καρδιακά προβλήματα

Το Σχήμα 2 παρουσιάζει τη γραφική παράσταση με το ποσοστό των ανδρών και των γυναικών στα δεδομένα. Παρατηρούμε ότι το 67.7% των ατόμων στα δεδομένα μας είναι άνδρες ενώ το 32.3% είναι γυναίκες. Άρα τα δεδομένα που αφορούν τους άνδρες είναι περισσότερα από τα δεδομένα που αφορούν γυναίκες. Το Σχήμα 3 παρουσιάζει το ιστόγραμμα ανδρών και γυναικών στα δεδομένα. Παρατηρούμε ότι στα δεδομένα μας 71 γυναίκες δεν παρουσιάζουν κάποιο καρδιακό νόσημα ενώ 25 γυναίκες αντιμετωπίζουν κάποια καρδιακή πάθηση. Αντίθετα, 89 άνδρες στα δεδομένα μας δεν παρουσιάζουν κάποιο καρδιακό νόσημα ενώ 112 άνδρες αντιμετωπίζουν κάποια καρδιακή πάθηση.

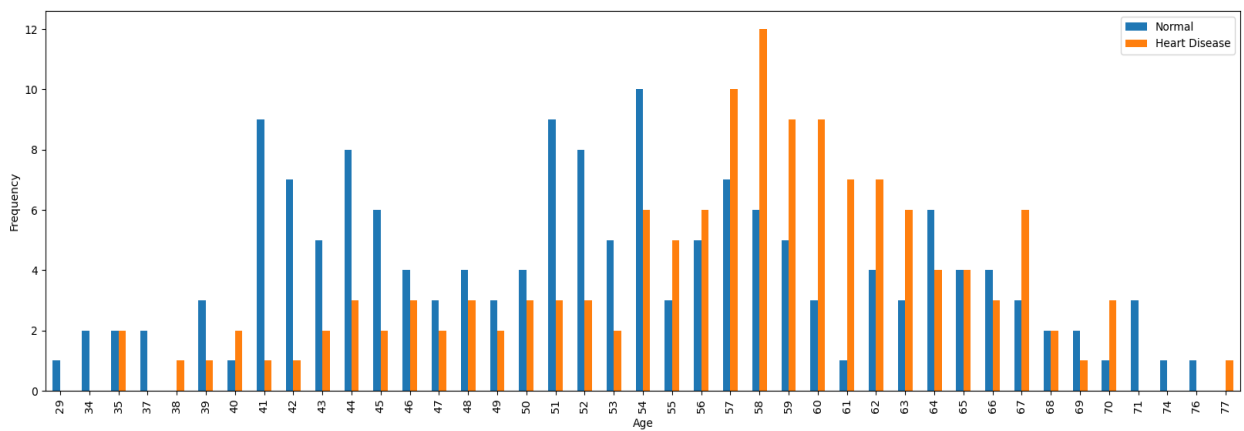


**Σχήμα 2:** Ποσοστό ανδρών και γυναικών στα δεδομένα

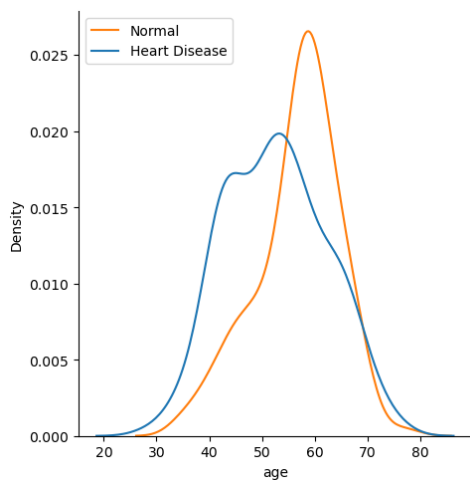


**Σχήμα 3:** Ιστόγραμμα ατόμων με και χωρίς καρδιακά προβλήματα ανά φύλο

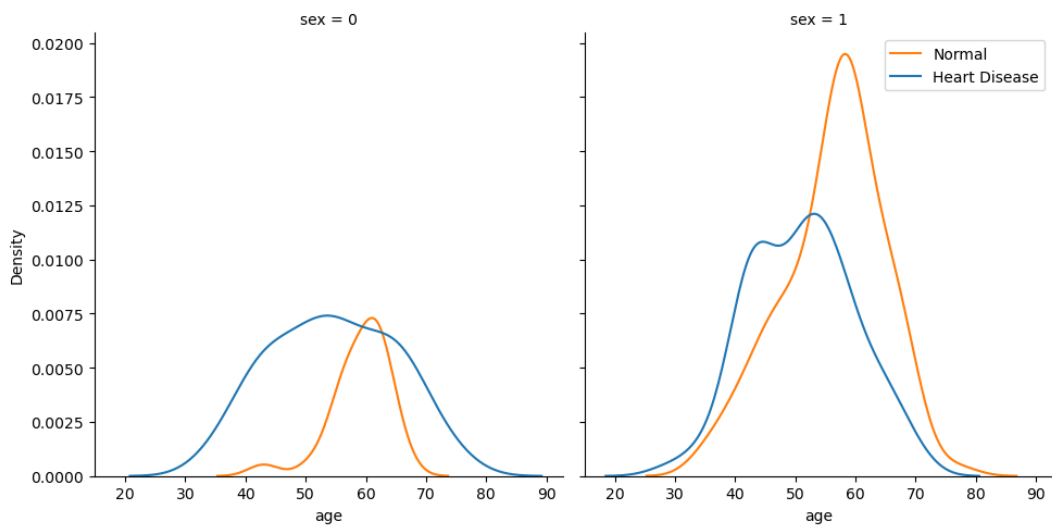
Το Σχήμα 4 απεικονίζει το ιστόγραμμα ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία. Παρατηρούμε ότι στα δεδομένα μας τα περισσότερα άτομα με καρδιακά προβλήματα βρίσκονται στην ηλικία των 58 ετών ενώ τα περισσότερα άτομα χωρίς καρδιακά προβλήματα βρίσκονται στην ηλικία των 54 ετών. Επίσης, παρατηρούμε η μικρότερη ηλικία για την οποία έχουμε δεδομένα είναι 29 ετών ενώ η μεγαλύτερη είναι 77 ετών. Το Σχήμα 5 παρουσιάζει τη γραφική παράσταση της κατανομή ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία. Όπως φαίνεται τα άτομα 60 – 70 ετών είναι που στα δεδομένα είναι εκείνα αντιμετωπίζουν περισσότερα καρδιακές παθήσεις ενώ τα άτομα 40 – 60 ετών είναι εκείνα που δεν αντιμετωπίζουν περισσότερο καρδιακές παθήσεις. Στο Σχήμα 6 φαίνεται η γραφική παράσταση της κατανομής των ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία ανά φύλο. Η κατανομές ανά φύλο έχουν περίπου την ίδια μορφή όπως και προηγούμενος.



**Σχήμα 4:** Ιστόγραμμα ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία

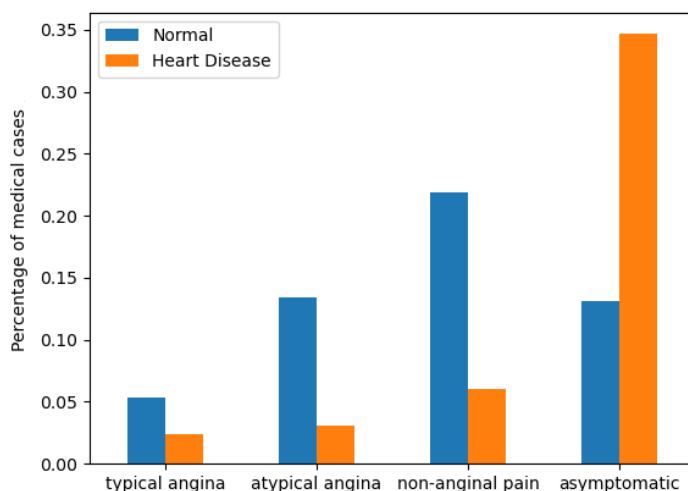


**Σχήμα 5:** Κατανομή ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία



**Σχήμα 6:** Κατανομή ατόμων με και χωρίς καρδιακά προβλήματα ανά ηλικία ανά φύλο

Το Σχήμα 7 απεικονίζει τη γραφική παράσταση με το ποσοστό ατόμων με και χωρίς καρδιακά προβλήματα ανά τύπο πόνου στο στήθος. Παρατηρούμε ότι τα περισσότερα άτομα στα δεδομένα που αντιμετωπίζουν κάποια καρδιακή ασθένεια δεν έχουν κάποιο πόνο στο στήθος. Αντίθετα, τα περισσότερα άτομα στα δεδομένα που δεν αντιμετωπίζουν κάποια καρδιακή ασθένεια υποφέρουν από μη στηθαγχικό ή μη ισχαιμικό θωρακικό άλγος.



**Σχήμα 7:** Ποσοστό ατόμων με και χωρίς καρδιακά προβλήματα ανά τύπο πόνου στο στήθος

### 3.3 Κανονικοποίηση δεδομένων

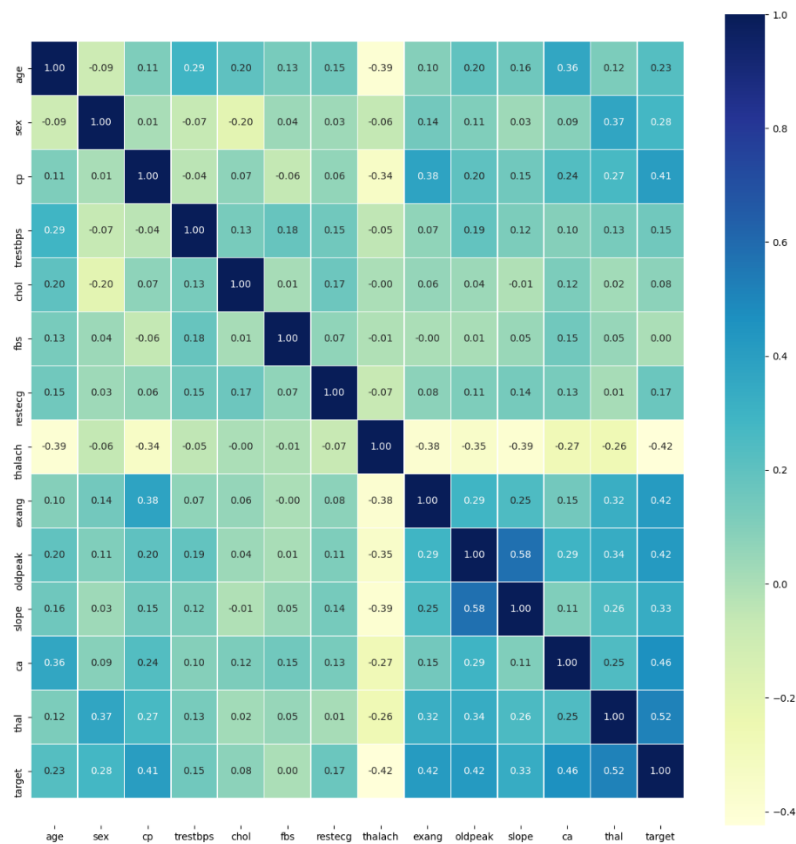
Για να βοηθήσουμε τα μοντέλα να μάθουν καλύτερα κανονικοποιήσαμε τα features αφαιρώντας τον μέσο όρο και διαιρώντας με την τυπική απόκλιση. Το κανονικοποιημένο score ενός δείγματος  $x$  υπολογίζεται ως εξής:

$$z = \frac{x - \mu}{s}$$

όπου  $\mu$  είναι ο μέσος όρος των δειγμάτων και  $s$  είναι η τυπική απόκλιση των δειγμάτων. Η κανονικοποίηση εκτελείται ανεξάρτητα σε κάθε feature υπολογίζοντας τα σχετικά στατιστικά στοιχεία για τα δείγματα. Στη συνέχεια αποθηκεύονται ο μέσος όρος και η τυπική απόκλιση για τα δεδομένα και χρησιμοποιούνται μεταγενέστερα στο μετασχηματισμό. Η κανονικοποίηση ενός συνόλου δεδομένων είναι μια κοινή απαίτηση για πολλά μοντέλα μηχανικής μάθησης. Ενδέχεται τα μοντέλα να συμπεριφέρονται άσχημα εάν τα μεμονωμένα features δεν είναι κανονικοποιημένα.

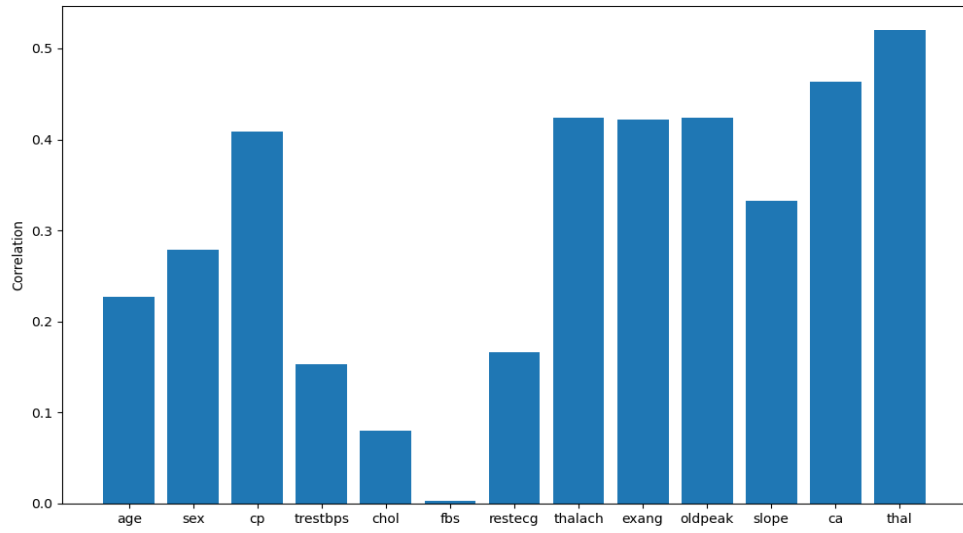
### 3.4 Συσχετίσεις δεδομένων

Το correlation analysis [7] είναι μια διμερής ανάλυση που μετρά τη δύναμη της συσχέτισης μεταξύ δύο μεταβλητών και της κατεύθυνσης της σχέσης. Όσον αφορά την ισχύ της σχέσης, η τιμή του συντελεστή συσχέτισης κυμαίνεται μεταξύ +1 και -1. Η τιμή  $\pm 1$  δείχνει τον τέλειο βαθμό συσχέτισης μεταξύ των δύο μεταβλητών. Καθώς η τιμή του συντελεστή συσχέτισης πηγαίνει προς το 0, η σχέση μεταξύ των δύο μεταβλητών είναι πιο αδύναμη. Η κατεύθυνση της σχέσης υποδεικνύεται από το σύμβολο του συντελεστή. Το σύμβολο + υποδηλώνει θετική σχέση δηλαδή ευθέως ανάλογη σχέση στην οποία με τη αύξηση της μια μεταβλητής αυξάνεται και η άλλη μεταβλητή. Αντίθετα το σύμβολο - υποδηλώνει αρνητική σχέση δηλαδή αντιστρόφως ανάλογη σχέση στην οποία με τη αύξηση της μια μεταβλητής μειώνεται η άλλη. Το Σχήμα 8 παρουσιάζει το correlation analysis μεταξύ όλων των features στα δεδομένα που χρησιμοποιήσαμε. Επίσης, το Σχήμα 9 παρουσιάζει την απόλυτη τιμή του correlation analysis μεταξύ των features και του target. Στην παρούσα εργασία δεν παρατηρήσαμε ξεκάθαρες ισχυρές συσχετίσεις κάποιου feature με το target γι' αυτό αποφασίσαμε να χρησιμοποιήσουμε αλγορίθμους μηχανικής μάθησης οι οποίοι θα χρησιμοποιούν όλα τα feature και θα κτίσουν τα μοντέλα πρόβλεψης αυτόματα.



Σχήμα 8: Correlation analysis μεταξύ όλων των features στα δεδομένα





**Σχήμα 9:** Απόλυτη τιμή του correlation analysis μεταξύ των features και του target

# Κεφάλαιο 4

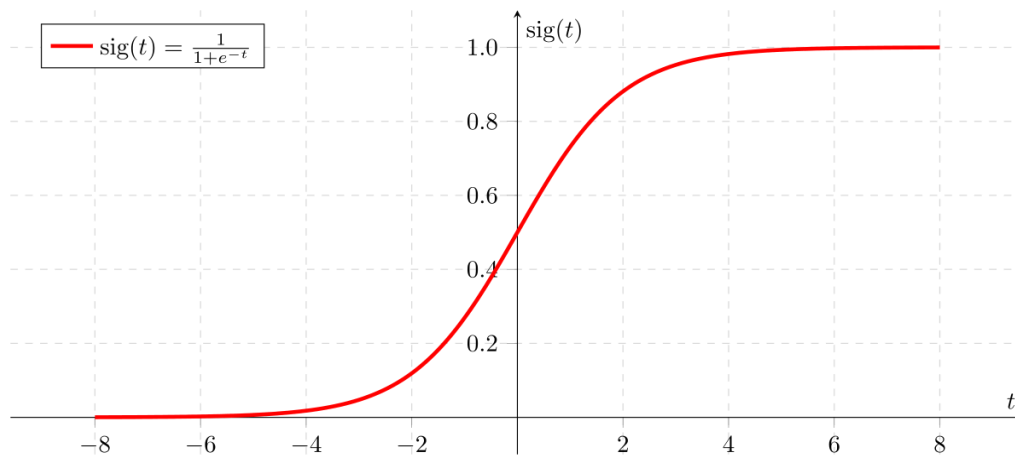
## Αλγόριθμοι μηχανικής μάθησης

### 4.1 Logistic Regression

Ο αλγόριθμος Logistic Regression [8] είναι ένας αλγόριθμος supervised learning που χρησιμοποιείται για να λύσει προβλήματα classification. Ο αλγόριθμος Logistic Regression χρησιμοποιεί μια sigmoid function η οποία δίνεται από τη σχέση:

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

Η υπόθεση του Logistic Regression περιορίζει τη συνάρτηση μεταξύ του 0 και του 1. Θα χρησιμοποιούμε το sigmoid function όπως φαίνεται στο Σχήμα 10 για να χαρτογραφήσουμε τις προβλέψεις στις πιθανότητες.



**Σχήμα 10:** Γραφική παράσταση sigmoid function

[Πηγή: <https://commons.wikimedia.org/wiki/File:Sigmoid-function-2.svg>]

Όταν χρησιμοποιούμε Linear Regression η φόρμουλα για την υπόθεση είναι:

$$h\theta(x) = \beta_0 + \beta_1 X$$

Στο Logistic Regression αναμένουμε ότι η υπόθεση μας θα μας δώσει τιμές μεταξύ του 0 και του 1 οπότε:

$$h\theta(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Το cost function αναπαριστά τον optimization στόχο δηλαδή θέλουμε να δημιουργούμε ένα cost function και το ελαχιστοποιούμε ώστε να μπορέσουμε να αναπτύξουμε ένα ακριβές μοντέλο με ελάχιστο error. Το cost function δίνεται από τη σχέση:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

όπου το  $y^{(i)}$  είναι το predicted output και  $h_{\theta}(x^{(i)})$  είναι η υπόθεση function. Χρησιμοποιούμε τη μέθοδο του gradient descent για να βρούμε την ελάχιστη τιμή του cost.

## 4.2 k-Nearest Neighbors (kNN)

Ο αλγόριθμος k-Nearest Neighbors [9] είναι ένας απλός αλγόριθμος supervised learning που χρησιμοποιείται για να λύσει προβλήματα classification ή regression. Ο αλγόριθμος kNN υποθέτει ότι παρόμοια πράγματα βρίσκονται κοντά το ένα στο άλλο. Το kNN χρησιμοποιεί την ιδέα του similarity (μερικές φορές ονομάζεται και distance, proximity, ή closeness) για να υπολογίσουμε την απόσταση μεταξύ σημείων σε ένα γράφημα. Υπάρχουν πολλές μέθοδοι υπολογισμού της απόστασης και ένας τρόπος μπορεί να είναι προτιμητέος από κάποιον άλλο τρόπο ανάλογα με το πρόβλημα που επιλύουμε. Τέτοιες μέθοδοι είναι euclidian, minkowski και manhattan distance. Για κάθε νέο query υπολογίζουμε το distance του με όλα τα vectors που υπάρχουν στο train set. Στη συνέχεια με βάση το distance επιλέγουμε τους k κοντινότερους γείτονες. Τέλος, κατατάσσουμε το query στην κατηγορία της πλειοψηφίας των γειτόνων. Το Σχήμα 11 παρουσιάζει την εισαγωγή ενός νέου query το οποίο θα ταξινομηθεί με βάση τον αλγόριθμο nearest neighborhoods με  $k = 5$ . Από τους 5 κοντινότερους γείτονες οι 3 ταξινομούνται στην κατηγορία A και οι άλλοι 2 στην κατηγορία B με αποτέλεσμα το query να ταξινομηθεί στην κατηγορία A στην οποία ανήκει και η πλειοψηφία των γειτόνων.



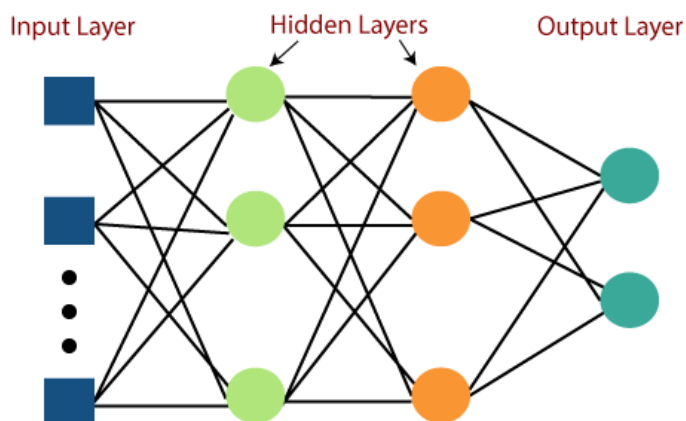
**Σχήμα 11:** Παράδειγμα εισαγωγής νέου query το οποίο θα ταξινομηθεί με τον αλγόριθμο 5-nearest neighborhoods

[Πηγή: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>]

### 4.3 Multilayer Perceptron (MLP)

Ο αλγόριθμος Multilayer Perceptron Classifier [10] είναι ένας αλγόριθμος supervised learning ο οποίος βασίζει τη λειτουργία του στα Τεχνητά Νευρωνικά Δίκτυα. Ο αλγόριθμος εκπαιδεύεται με ένα σύνολο εισόδων και τις επιθυμητές τους εξόδους και δυνητικά μπορεί να μάθει μια κατά προσέγγιση μη γραμμική συνάρτηση για classification ή regression. Η αρχιτεκτονική του δικτύου αποτελείται από ένα επίπεδο εισόδου (input layer), ένα ή περισσότερα μη γραμμικά επίπεδα τα οποία ονομάζονται κρυφά επίπεδα (hidden layer) και τέλος ένα επίπεδο εξόδου (output layer). Το επίπεδο εισόδου δεν είναι ενεργό καθώς απλά διοχετεύει τα χαρακτηριστικά (features) στους νευρώνες του επόμενου επιπέδου. Οι νευρώνες συνδέονται μεταξύ τους με τις συνάψεις. Κάθε σύναψη χαρακτηρίζεται από ένα βάρος (weight) η τιμή του οποίου μεταβάλλεται κατά τη διάρκεια της εκπαίδευσης. Κάθε νευρώνας σε κάθε ενεργό επίπεδο μετατρέπει τις τιμές που λαμβάνει από το προηγούμενο επίπεδο σε ένα weighted sum το οποίο διοχετεύεται σε μια μη γραμμική συνάρτηση ενεργοποίησης (πχ sigmoid function ή hyperbolic tan function). Για προβλήματα ταξινόμησης όπως είναι αυτά που έχουμε να λύσουμε σε αυτή την εργασία ο αλγόριθμος multilayer perceptron χρησιμοποιεί τη μέθοδο ανάστροφης μετάδοσης σφάλματος (backpropagation) κατά την εκπαίδευση. Αυτή η μέθοδος βασίζεται στην προσπάθεια να προσαρμόσουμε τα βάρη έτσι ώστε να ελαχιστοποιήσουμε τη τετραγωνική συνάρτηση σφάλματος. Αυτό επιτυγχάνεται με τη μέθοδο gradient descent οπότε και η αλλαγή των βαρών είναι ανάλογη ως προς το αρνητικό της κλίσης της συνάρτησης του σφάλματος σε ένα σημείο δηλαδή είναι ανάλογη με το αρνητικό της παραγώγου της συνάρτησης του σφάλματος ως προς τα

βάρη. Εν τέλη, ο αλγόριθμος δε χρειάζεται περισσότερα από τρία ενεργά επίπεδα για να λύσει οποιοδήποτε μη γραμμικά διαχωρίσιμο πρόβλημα εφόσον μπορεί να σχηματίσει οποιαδήποτε κυρτή επιφάνεια (με 1 κρυφό επίπεδο) ή οποιαδήποτε αυθαίρετη περιοχή (με 2 κρυφά επίπεδα). Το Σχήμα 12 παρουσιάζει μια γραφική αναπαράσταση ενός Multilayer Perceptron δίκτυο.



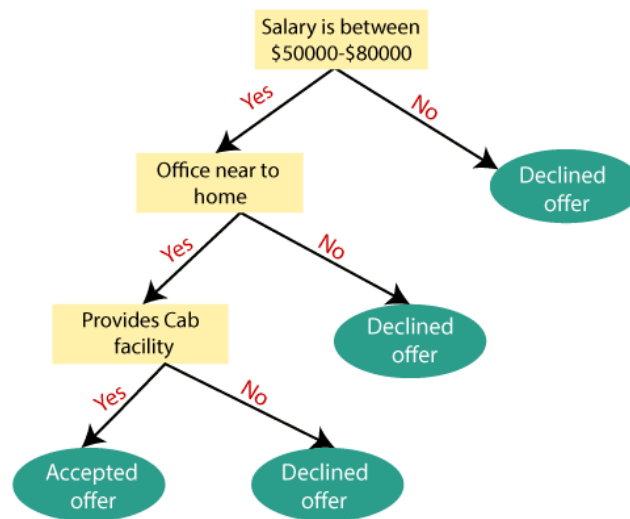
**Σχήμα 12:** Γραφική αναπαράσταση ενός multilayer perceptron νευρωνικό δίκτυο

[Πηγή: <https://github.com/Thanasis1101/MLP-from-scratch>]

## 4.4 Decision Tree

Ο αλγόριθμος Decision Tree Classifier [11] είναι ένας αλγόριθμος supervised learning που χρησιμοποιείται για να λύσει προβλήματα classification ή regression. Τα δέντρα αποφάσεων αναφέρονται σε ένα ιεραρχικό μοντέλο αποφάσεων καθώς και των συνέπειών τους. Ο στόχος, είναι να δημιουργηθούν μοντέλα τα οποία να προβλέπουν την τιμή μιας εισόδου δεδομένων με τη μάθηση απλών κανόνων απόφασης που προκύπτουν από τα χαρακτηριστικά (features) των δεδομένων έτσι ώστε σε άγνωστα δεδομένα να ακολουθήσει τη στρατηγική με την οποία έχει τη μεγαλύτερη πιθανότητα για να επιτύχει το στόχο του. Για προβλήματα ταξινόμησης όπως είναι αυτά που έχουμε να λύσουμε σε αυτή την εργασία ένα δέντρο απόφασης αναφέρεται ως ένα δέντρο ταξινόμησης. Το δέντρο αποφάσεων αποτελείται από κόμβους που σχηματίζουν ένα δέντρο με ρίζα. Σε κάθε φύλλο έχει αντιστοιχηθεί μία κατηγορία η οποία αναπαριστά την κατάλληλη έξοδο. Τα γεγονότα ταξινομούνται με βάση τη διαδρομή από τη ρίζα του δέντρου σε ένα φύλλο, σύμφωνα με τα αποτελέσματα των δοκιμών κατά μήκος της διαδρομής. Ο τρόπος με τον οποίο κτίζεται το δέντρο είναι: (1) ορίζεται ως ρίζα ο κόμβος με το καλύτερο χαρακτηριστικό από τα παραδείγματα (2) εάν για μια τιμή αυτού του χαρακτηριστικού για όλα τα παραδείγματα

που έχουν αυτή την τιμή η έξοδος είναι ίδια τότε δημιουργήσε φύλλο με αυτή την έξοδο (3) διαφορετικά δημιουργήσε ένα υποδέντρο αναδρομικά επιλέγοντας το αμέσως επόμενο καλύτερο χαρακτηριστικό. Το καλύτερο χαρακτηριστικό είναι εκείνο που με τη διάσπαση των παραδειγμάτων οδηγεί σε όσο το δυνατό μεγαλύτερη μείωση της εντροπίας. Το Σχήμα 13 παρουσιάζει ένα παράδειγμα decision tree που δημιουργήθηκε μετά την εκπαίδευση.

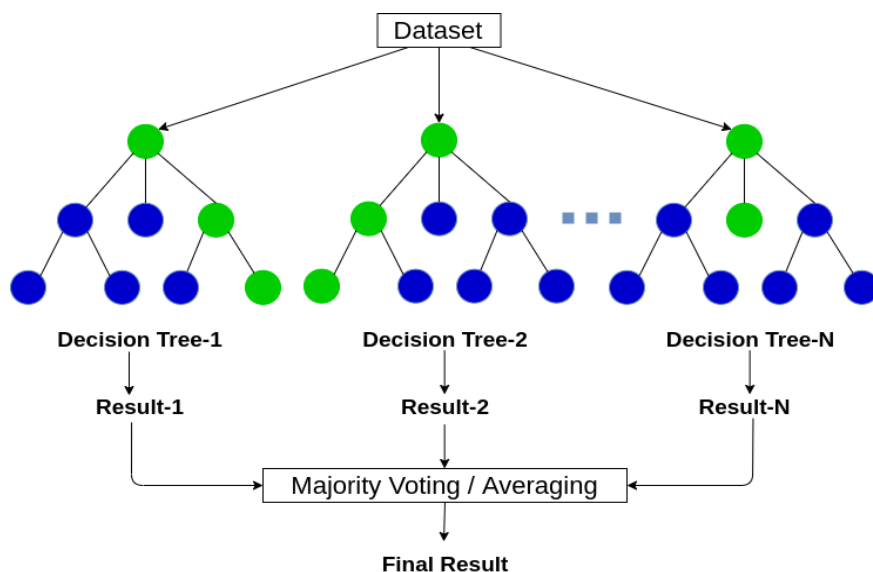


**Σχήμα 13:** Παράδειγμα decision tree

[Πηγή: <https://www.mygreatlearning.com/blog/decision-tree-algorithm/>]

## 4.5 Random Forest

Τα Random Forest [12] είναι ένας αλγόριθμος supervised learning που χρησιμοποιείται για να λύσει προβλήματα classification ή regression. Όπως υποδηλώνει το όνομά του, αποτελούνται από ένα μεγάλο αριθμό μεμονωμένων decision trees που λειτουργούν ως σύνολο. Κάθε μεμονωμένο δέντρο στο Random Forest παράγει μια πρόβλεψη για ένα class και το class με τις περισσότερες ψήφους γίνεται η πρόβλεψη του μοντέλου μας. Ο λόγος για τον οποίο τα Random Forest λειτουργούν τόσο καλά είναι ότι τα δέντρα προστατεύουν το ένα το άλλο από τα ατομικά τους λάθη (αρκεί να μην κάνουν λάθος όλα στην ίδια κατεύθυνση). Το Random Forest προσθέτει επιπλέον τυχαιότητα στο μοντέλο, ενώ κτίζει τα δέντρα. Αντί να αναζητά το πιο σημαντικό χαρακτηριστικό όταν διαχωρίζει έναν κόμβο, αναζητά το καλύτερο χαρακτηριστικό σε ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτό οδηγεί σε μια μεγάλη ποικιλία που γενικά οδηγεί σε ένα καλύτερο μοντέλο. Το Σχήμα 14 παρουσιάζει τον τρόπο λειτουργίας των Random Forest.



**Σχήμα 14:** Τρόπος λειτουργίας random forest

[Πηγή: <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/>]

## 4.6 Gaussian Naive Bayes

Το Naive Bayes [13] είναι ένας αλγόριθμος supervised learning που χρησιμοποιείται για να λύσει προβλήματα classification. Ο αλγόριθμος βασίζεται στη θεωρία των πιθανοτήτων. Ας ξεκινήσουμε με το εξής ερώτημα «Δεδομένου ενός σημείου  $x$ , ποια είναι η πιθανότητα του  $x$  να ανήκει σε μια κλάση  $c$ ;». Ο Naive Bayes classifier προσπαθεί να υπολογίσει αυτές τις πιθανότητες απευθείας. Επομένως, δεδομένου ενός σημείου  $x$ , θέλουμε να υπολογίσουμε τη  $p(c | x)$  για όλες τις κλάσεις  $c$  και η έξοδος είναι το  $c$  με τη μεγαλύτερη πιθανότητα. Αυτό μπορεί να γραφτεί ως εξής:

$$prediction(x) = \arg \max p(c | x)$$

όπου το  $\max p(c | x)$  επιστρέφει τη μέγιστη πιθανότητα ενώ το  $\arg \max p(c | x)$  επιστέφει το  $c$  με τη ψηλότερη πιθανότητα. Μπορούμε να υπολογίσουμε το  $p(c | x)$ , με το θεώρημα του Bayes:

$$p(c | x) = \frac{p(x | c) \times p(c)}{p(x)} = \frac{p(x | c) \times p(c)}{\sum_c p(x | c) \times p(c)}$$

Πως υπολογίζουμε το  $p(x | c)$  και  $p(c)$ ; Αυτό είναι το θέμα της εκπαίδευσης του Bayes classifier. Ο απλούστερος τρόπος για υπολογίσουμε το  $p(c)$  είναι να υπολογίσουμε τις σχετικές συχνότητες των κλάσεων και να τις χρησιμοποιήσουμε σαν πιθανότητες. Για να υπολογίσουμε τη πιθανότητα

$p(x / c)$  θα χρειαστεί να κάνουμε τη naïve υπόθεση ότι τα features  $x_1, x_2$  είναι στοχαστικά ανεξάρτητα δεδομένου του  $c$ .

$$p(x_1, x_2 | c) = p(x_1 | c) \times p(x_2 | c)$$

Από αυτό το μέρος προέρχεται η naïve προσέγγιση του Bayes επειδή αυτή η εξίσωση δεν ισχύει γενικά. Για να υπολογίσουμε τη πιθανότητα  $p(x / c)$  θα χρησιμοποιήσουμε τη Gaussian κατανομή η οποία δίνεται από τη σχέση:

$$p(x_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} \times e^{-\frac{1}{2}\left(\frac{x_i - \mu_{i,j}}{\sigma_{i,j}}\right)^2} \text{ για } i = 1, 2 \text{ και } j = 1, 2, 3$$

όπου  $\mu_{i,j}$  είναι ο μέσος και  $\sigma_{i,j}$  είναι το standard deviation.



# Κεφάλαιο 5

## Αξιολόγηση μοντέλων

### 5.1 Accuracy

Για την αξιολόγηση της επίδοσης των αλγορίθμων μηχανικής μάθησης χρησιμοποιήθηκε η μετρική του *accuracy* σε προβλέψεις με άγνωστα δεδομένα δηλαδή δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση.

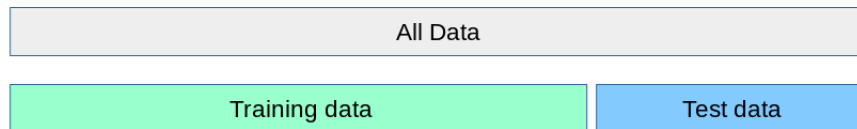
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ένα *true positive (TP)* είναι το αποτέλεσμα όπου το μοντέλο προβλέπει σωστά ότι ένας ασθενής έχει καρδιακή πάθηση. Ομοίως, ένα *true negative (TN)* είναι ένα αποτέλεσμα όπου το μοντέλο προβλέπει σωστά ότι ένας ασθενής δεν έχει καρδιακή πάθηση. Αντίθετα, ένα *false positive (FP)* είναι ένα αποτέλεσμα όπου το μοντέλο προβλέπει εσφαλμένα ότι ένας ασθενής έχει καρδιακή πάθηση. Αντίστοιχα, ένα *false negative (FN)* είναι το αποτέλεσμα όπου το μοντέλο προβλέπει εσφαλμένα ότι ένας ασθενής δεν έχει καρδιακή πάθηση. Οι πιο πάνω πληροφορίες συνήθως βρίσκονται σε ένα *confusion matrix* το οποίο έχει τη μορφή:

$$confusion\ matrix = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

### 5.2 Διαχωρισμός δεδομένων σε train και test

Η διαδικασία διαχωρισμού των δεδομένων σε train-test χρησιμοποιείται για την εκτίμηση της απόδοσης των αλγορίθμων μηχανικής μάθησης. Εάν τεστάρουμε το μοντέλο που αναπτύξαμε με τα ίδια δεδομένα που το εκπαιδεύσαμε θα έχουμε ψηλό score αλλά ενδέχεται το σύστημα να αποτυγχάνει όταν κάνει προβλέψεις σε άγνωστα δεδομένα. Αυτή η κατάσταση ονομάζεται *overfitting*. Για να αποφύγουμε αυτή την κατάσταση μια συνήθης πρακτική είναι να κρατάμε ένα μέρος των δεδομένων σαν test. Στη περίπτωση μας χωρίσαμε τα δεδομένα σε 80% για train και 20% για test. Το Σχήμα 15 παρουσιάζει το διαχωρισμό των δεδομένων σε train και test.



**Σχήμα 15:** Διαχωρισμός των δεδομένων σε train και test

[Πηγή: [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)]

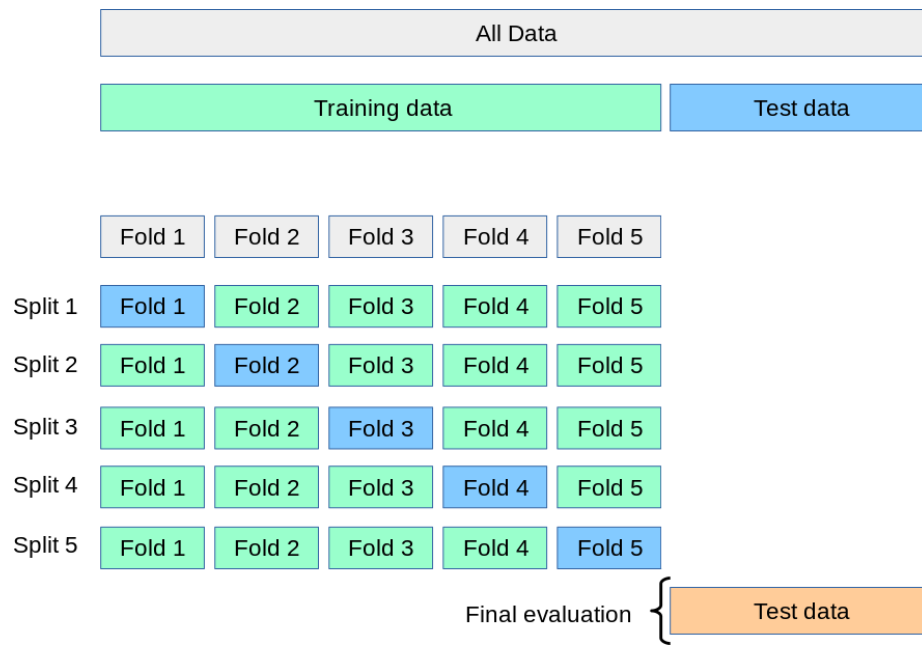
## 5.3 Cross validation

Το cross validation [14] είναι μια μέθοδος που χρησιμοποιείται για την αξιολόγηση της αποτελεσματικότητας των μοντέλων μηχανικής μάθησης. Συνήθως χρησιμοποιείται για σύγκριση και αξιολόγηση διάφορων μοντέλων μηχανικής μάθησης και για επιλογή του καλύτερου. Είναι επίσης μια διαδικασία επανα-δειγματοληψίας που χρησιμοποιείται για την αξιολόγηση ενός μοντέλου εάν έχουμε περιορισμένα δεδομένα, βοηθώντας μας να βεβαιωθούμε ότι το μοντέλο δεν θα πάθει overfitting. Έχει μια μόνο παράμετρο, το  $k$ , που αναφέρεται στον αριθμό ομάδων (folds) στις οποίες θα διαιρεθεί ένα σύνολο δεδομένων.

Η διαδικασία αυτή συχνά ονομάζεται  $k$ -fold cross-validation. Η διαδικασία που ακολουθεί είναι η εξής:

1. Σπάζουμε τα δεδομένα για το train σε  $k$  folds (ομάδες).
2. Επιλέγουμε ένα fold και το χρησιμοποιούμε σαν test set.
3. Χρησιμοποιούμε τα υπόλοιπα  $k-1$  folds σαν για να εκπαιδεύουμε το μοντέλο.
4. Υπολογίζουμε το accuracy του εκπαιδευμένου μοντέλου στο test set.
5. Επαναλαμβάνουμε αυτήν τη διαδικασία έως ότου κάθε  $k$ -fold χρησιμοποιηθεί σαν test set.
6. Τέλος, υπολογίζουμε το average accuracy (μέση ακρίβεια) και το standard deviation (τυπική απόκλιση). Αυτές θα είναι οι μετρήσεις για την απόδοση του μοντέλου.

Αυτό σημαίνει ότι κάθε ομάδα έχει την ευκαιρία να χρησιμοποιηθεί σαν test set 1 φορά και να χρησιμοποιηθεί για την εκπαίδευση του μοντέλου  $k-1$  φορές. Το Σχήμα 16 παρουσιάζει τη διαδικασία του διαχωρισμού των δεδομένων σε train και test και το διαχωρισμό των δεδομένων στα train δεδομένα που θα γίνει στην περίπτωση 5-fold cross validation.



**Σχήμα 16:** Εικονική αναπαράσταση 5-fold cross validation

[Πηγή: [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)]

## 5.4 Grid Search

Ένα μοντέλο μηχανικής μάθησης έχει πολλές παραμέτρους που δεν εκπαιδεύονται από τα train δεδομένα. Αυτές οι παράμετροι είναι πολύ σημαντικές εφόσον ελέγχουν την ακρίβεια του μοντέλου. Για παράδειγμα, ο ρυθμός μάθησης (learning rate) ενός νευρικού δικτύου είναι παράμετρος επειδή ορίζεται ρητά πριν την εκπαίδευση. Από την άλλη, τα βάρη (weights) ενός νευρικού δικτύου δεν είναι παράμετρος επειδή εκπαιδεύονται από το train set. Το grid search [15] είναι μια τεχνική που επιχειρεί να βρει τις βέλτιστες τιμές των παραμέτρων.

Είναι μια εξαντλητική αναζήτηση που πραγματοποιείται σε συγκεκριμένες τιμές παραμέτρων ενός μοντέλου. Πολλές φορές το grid search χρησιμοποιεί k-fold cross validation για να βρει τις βέλτιστες παραμέτρους. Ο Πίνακας 1 παρουσιάζει τις τιμές των παραμέτρων που εξερευνήθηκαν κατά τη διάρκεια του Grid Search για κάθε μοντέλο μηχανικής μάθησης.

Μοντέλο	Παράμετρος	Τιμές
Logistic Regression	C	1, 2, 3, 4
	penalty	l1, l2
	solver	newton-cg, lbfgs, liblinear
	max_iter	100, 300, 500, 600, 700, 800, 900, 1000
k-Nearest Neighbors	n_neighbors	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
	weights	uniform, distance
	algorithm	auto, ball_tree, kd_tree, brute
	metric	euclidean, minkowski, manhattan
Multilayer Perceptron	hidden_layer_sizes	(50, 50, 50), (50, 100, 50), (100,)
	activation	identity, logistic, tanh, relu
	solver	lbfgs, sgd, adam
	alpha	0.0001, 0.05
	learning_rate	constant, invscaling, adaptive
	max_iter	200, 300, 500, 800, 1000, 1500, 2000, 2500, 3000
Decision Tree	criterion	gini, entropy
	splitter	best, random
	max_features	auto, sqrt, log2
	max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
	min_samples_split	1, 2, 3, 4, 5, 6, 7, 8, 9
	min_samples_leaf	1, 2, 3, 4, 5, 6, 7, 8, 9
Random Forest	n_estimators	100, 500, 1000, 1500
	criterion	gini, entropy
	max_depth	None, 4, 50, 100, 200, 300
	max_features	auto, sqrt, log2
Gaussian Naive Bayes		

**Πίνακας 1:** Παράμετροι που διερευνήθηκαν στο grid search

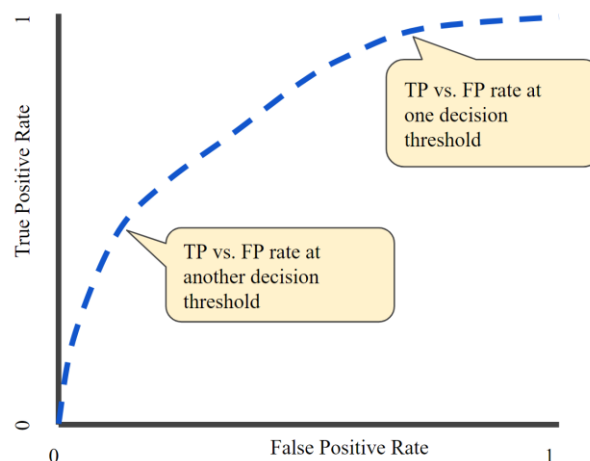
## 5.5 ROC curve

Μια ROC curve (receiver operating characteristic curve) [16] είναι μια γραφική παράσταση που απεικονίζει τη διαγνωστική ικανότητα ενός συστήματος δυαδικού classifier σε διάφορα thresholds. Η ROC είναι μια καμπύλη πιθανότητας και το AUC (Area Under the Curve) αναπαριστά την ικανότητα του μοντέλου να διαχωρίζει τις κλάσεις. Όσο πιο ψηλό είναι το AUC τόσο καλύτερα το μοντέλο διαχωρίζει τις κλάσεις. Η καμπύλη ROC αναπαρίσταται με το *true positive rate* να βρίσκεται στον y-άξονα και το *false positive rate* στο x-άξονα.

$$\text{true positive rate} = \frac{TP}{TP + FN}$$

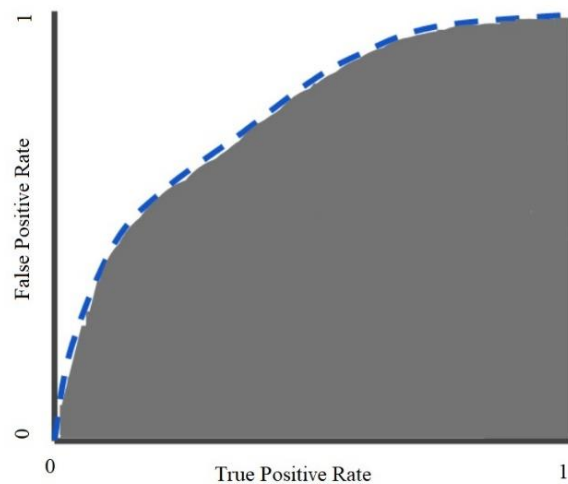
$$\text{false positive rate} = \frac{FP}{FP + TN}$$

Μια καμπύλη ROC απεικονίζει το *true positive rate* εναντίον του *false positive rate* εναντίον σε διαφορετικά classification thresholds. Χαμηλότερα classification thresholds κατατάσσουν περισσότερα στοιχεία ως θετικά, αυξάνοντας έτσι τόσο τα *false positive* όσο και τα *true positive*. Το παρακάτω Σχήμα 17 δείχνει μια τυπική καμπύλη ROC.



**Σχήμα 17:** True positive rate και False positive rate σε διαφορετικά classification thresholds  
[Πηγή: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>]

Το AUC μετρά το εμβαδόν κάτω από την καμπύλη ROC από το (0,0) μέχρι το (1,1). Το Η AUC παρέχει ένα συνολικό μέτρο απόδοσης σε όλα τα πιθανά classification thresholds. Όσο μεγαλύτερο είναι το AUC τόσο πιο ψηλή είναι η απόδοση του μοντέλου με μέγιστη τιμή το 100% που σημαίνει ότι το μοντέλο ήταν ορθό σε όλες τις προβλέψεις του. Αντίστοιχα, Όσο μικρότερο είναι το AUC τόσο πιο χαμηλή είναι η απόδοση του μοντέλου με ελάχιστη τιμή το 0% που σημαίνει ότι το μοντέλο ήταν λανθασμένο σε όλες τις προβλέψεις του. Το Σχήμα 18 παρουσιάζει με γκρίζο χρώμα το AUC.



**Σχήμα 18:** Area under the ROC Curve

[Πηγή: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>]

# Κεφάλαιο 6

## Πειραματικά αποτελέσματα

### 6.1 Αποτελέσματα grid search

Το πρώτο βήμα στην πειραματική αξιολόγηση των μοντέλων μας ήταν να εντοπίσουμε τις βέλτιστες παραμέτρους για τα μοντέλα. Όπως ήδη έχουμε αναφέρει το grid search είναι μια τεχνική που επιχειρεί να βρει τις βέλτιστες τιμές των παραμέτρων. Ο Πίνακας 2 παρουσιάζει τις βέλτιστες παραμέτρους που επέλεξε το grid search για το κάθε μοντέλο.

Model	Best Parameters
Logistic Regression	'C': 1, 'max_iter': 300, 'penalty': 'l2', 'solver': 'lbfgs'
k-Nearest Neighbors	'algorithm': 'auto', 'metric': 'manhattan', 'n_neighbors': 5, 'weights': 'uniform'
Multilayer Perceptron	'activation': 'identity', 'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'learning_rate': 'invscaling', 'max_iter': 1500, 'solver': 'adam'
Decision Tree	'criterion': 'entropy', 'max_depth': 7, 'max_features': 'sqrt', 'min_samples_leaf': 8, 'min_samples_split': 8, 'splitter': 'best'
Random Forest	'criterion': 'gini', 'max_depth': 4, 'max_features': 'log2', 'n_estimators': 100
Gaussian Naive Bayes	

Πίνακας 2: Οι βέλτιστες παράμετροι που επέλεξε το grid search

### 6.2 Αποτελέσματα cross validation

Στη συνέχεια χρησιμοποιήσαμε τις βέλτιστες παραμέτρους που προέκυψαν από το grid search για να εκπαιδύσουμε τα μοντέλα και να αξιολογήσουμε το accuracy και το standard deviation που προκύπτουν από τη διαδικασία του 5-fold cross validation.

Model	Average Accuracy	Standard Deviation
Logistic Regression	80.59 %	1.65 %
k-Nearest Neighbors	70.46 %	3.81 %
Multilayer Perceptron	79.34 %	4.19 %
Decision Tree	72.56 %	6.49 %
Random Forest	81.46 %	4.74 %
Gaussian Naive Bayes	83.12 %	1.89 %

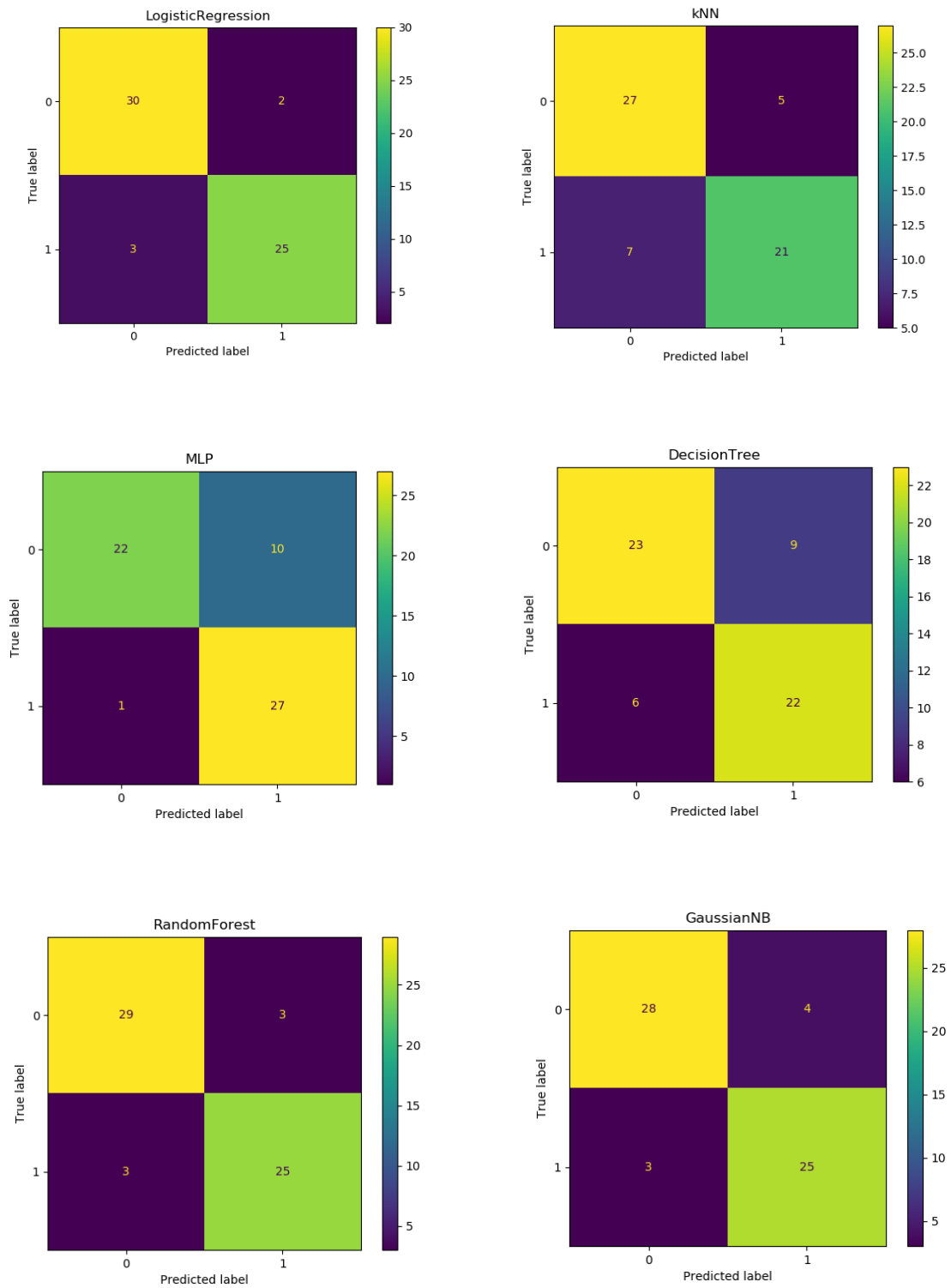
**Πίνακας 3:** Average accuracy και standard deviation για κάθε μοντέλο μηχανικής μάθησης

Ο Πίνακας 3 παρουσιάζει το average accuracy και το standard deviation για κάθε μοντέλο μηχανικής μάθησης. Παρατηρούμε ότι το average accuracy κυμαίνεται από 70.46% έως 83.12% ανάλογα με τον αλγόριθμο που εκπαιδεύσαμε το μοντέλο. Ένα άλλο σημείο που θα πρέπει να λάβουμε υπόψιν μας είναι το standard deviation που προκύπτει από το 5-fold cross validation. Ιδανικά θέλουμε να έχουμε μικρό standard deviation διότι έτσι μπορούμε να πούμε ότι δεν υπάρχουν μεγάλες αποκλίσεις μεταξύ των accuracy στα διάφορα folds δεδομένων που εξετάζονται. Στην περίπτωση μας όλα τα standard deviation που προκύπτουν κυμαίνονται από 1.65% έως 6.49%.

### 6.3 Ακρίβεια στα test δεδομένα

Όπως ήδη έχουμε αναφέρει χωρίσαμε τα δεδομένα σε 80%% για train και 20% για test. Η τελική αξιολόγηση του συστήματος έγινε στα test δεδομένα δηλαδή άγνωστα για το σύστημα δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση του. Το Σχήμα 19 παρουσιάζει τα confusion matrixes που προέκυψαν για κάθε μοντέλο αξιολογώντας τα test δεδομένα. Ιδανικά θέλουμε στα μοντέλα μας να έχουμε όσο το δυνατόν λιγότερα false positives και false negatives για να έχουμε συστήματα με όσο το δυνατόν ψηλότερη ακρίβεια.





**Σχήμα 19:** Confusion matrixes που προέκυψαν για κάθε μοντέλο αξιολογώντας τα test δεδομένα

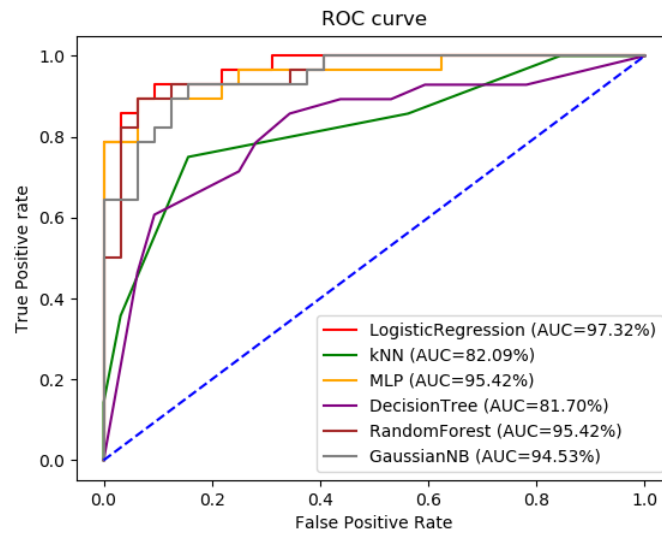
Ο Πίνακας 4 παρουσιάζει το accuracy που προέκυψε για κάθε μοντέλο αξιολογώντας τα test δεδομένα. Παρατηρούμε ότι το accuracy κυμαίνεται από 80% έως 91.67%. Το γεγονός αυτό μας αφήνει απόλυτα ικανοποιημένους εφόσον μπορούμε να πούμε σε άγνωστα δεδομένα μπορούμε να προβλέψουμε εάν ένα άτομο αντιμετωπίζει κάποια καρδιακή πάθηση με μεγάλη ακρίβεια (91.67% σε περίπτωση που χρησιμοποιήσουμε τον αλγόριθμο Logistic Regression).

Model	Accuracy
Logistic Regression	<b>91.67 %</b>
k-Nearest Neighbors	80 %
Multilayer Perceptron	81.67 %
Decision Tree	75 %
Random Forest	90 %
Gaussian Naive Bayes	88.33 %

**Πίνακας 4:** Accuracy που προέκυψε για κάθε μοντέλο αξιολογώντας τα test δεδομένα

## 6.4 Area Under the curve

Μια τελευταία τεχνική που αποφασίσαμε να χρησιμοποιήσουμε για να αξιολογήσουμε τα μοντέλα μας ήταν το AUC (Area Under the Curve) στις ROC γραφικές παραστάσεις. Όπως ήδη αναφέραμε μια ROC curve είναι μια γραφική παράσταση που απεικονίζει τη διαγνωστική ικανότητα ενός συστήματος δυαδικού classifier. Όσο μεγαλύτερο είναι το εμβαδόν που σχηματίζεται κάτω από την καμπύλη τόσο καλύτερα το μοντέλο διαχωρίζει τους ασθενείς που αντιμετωπίζουν κάποια καρδιακή ασθένεια από τους ασθενείς που δεν έχουν κάποιο πρόβλημα καρδιάς. Το Σχήμα 20 παρουσιάζει τα ROC curves. Όσο μεγαλύτερο είναι το εμβαδόν που σχηματίζει μια καμπύλη τόσο μεγαλύτερη είναι η ακρίβεια του αλγορίθμου. Όπως φαίνεται και από τον Πίνακα 5 τα AUC κυμαίνονται από 81.7% έως 97.32%. Έτσι μπορούμε να συμπεράνουμε ότι οι τεχνικές που αναπτύξαμε δουλεύουν καλά εφόσον το AUC είναι αρκετά ψηλό.



**Σχήμα 20:** ROC curve για τα test δεδομένα

Model	Accuracy
Logistic Regression	<b>97.32 %</b>
k-Nearest Neighbors	82.09 %
Multilayer Perceptron	95.42 %
Decision Tree	81.7 %
Random Forest	95.42 %
Gaussian Naive Bayes	94.53 %

**Πίνακας 5:** ACU που προκύπτουν για κάθε μοντέλο

# Κεφάλαιο 7

## Σύνδεση με την πανδημία COVID-19

Δυστυχώς τον τελευταίο χρόνο ολοκλήρως ο πλανήτης βρίσκεται αντιμέτωπος με την πανδημία του νέου κορονοϊού COVID-19. Η πανδημία άλλαξε την καθημερινότητα μας και απειλεί τη δημόσια υγεία. Η κάθε χώρα ανακοίνωσε τα δικά της περιοριστικά μέτρα και εφάρμοσε τα δικά της πρωτόκολλα με στόχο τον περιορισμό της εξάπλωσης του ιού. Μερικά από τα σημαντικότερα μέτρα που εφάρμοσε η πολιτεία είναι η καραντίνα για τα άτομα που ήρθαν σε επαφή με ύποπτο ή επιβεβαιωμένο κρούσμα και η κοινωνική αποστασιοποίηση.

Έχει παρατηρηθεί ότι αρκετοί γιατροί δεν εξετάζουν πλέον με φυσική παρουσία τους ασθενείς τους φοβούμενοι την εξάπλωση του ιού. Επίσης, πολλοί ασθενείς αποφεύγουν να επισκεφθούν τους γιατρούς τους προσπαθώντας να περιορίσουν τις μετακινήσεις τους είτε επειδή βρίσκονται σε καραντίνα. Αυτό όμως έχει σαν αποτέλεσμα την καθυστέρηση της διάγνωσης σε περίπτωση κάποιος πάθησης και κατά συνέπεια καθυστέρηση στη θεραπεία της ασθένειας. Έτσι εγκυμονούν σοβαροί κίνδυνοι για την υγεία του ασθενή ενώ πολλές φορές απειλείται ακόμα και η ίδια η ζωή του.

Έχει γίνει λοιπόν αδήριτη η ανάγκη παροχής ιατρικής φροντίδας και η ανταλλαγή ιατρικής γνώσης μεταξύ απομακρυσμένων περιοχών με τη χρήση τηλεπικοινωνιακών μέσων. Στην παρούσα εργασία υλοποιήσαμε ένα σύστημα το οποίο με βάση κάποιες πληροφορίες και βιοσήματα για τον ασθενή μπορεί να προβλέψει με ψηλή ακρίβεια εάν κάποιος ασθενής αντιμετωπίζει κάποια καρδιακή πάθηση ή όχι. Όπως ήδη αναφέραμε την τελευταία δεκαετία οι καρδιακές παθήσεις είναι η κύρια αιτία θανάτου παγκοσμίως. Με το σύστημα μας δυνητικά ένας ασθενής θα μπορεί να καταχωρήσει τα ιατρικά δεδομένα του μέσω μια απλής διαπροσωπίας και το σύστημα θα προβλέψει εάν αντιμετωπίζει κάποιο καρδιακό νόσημα. Έτσι ένας ασθενής μπορεί να έχει μια πρώτη διάγνωση για την κατάσταση της καρδιάς του χωρίς να χρειάζεται να επισκεφθεί με φυσική παρουσία το γιατρό του. Όπως θα εξηγήσουμε και στη συνέχεια τέτοια συστήματα δεν αντικαθιστούν σε καμία περίπτωση τον γιατρό.

# Κεφάλαιο 8

## Συμπεράσματα

Σε αυτή την εργασία υλοποιήσαμε ένα σύστημα το οποίο με βάση κάποια βιοσήματα που αφορούν τον ασθενή προβλέπει εάν ο ασθενής αντιμετωπίζει κάποια καρδιακή πάθηση. Για να το πετύχουμε αυτό δοκιμάσαμε διάφορους αλγόριθμους μηχανικής μάθησης για να εκπαιδεύσουμε μοντέλα με ιατρικά δεδομένα και να εντοπίσουμε εκείνο με τη μεγαλύτερη ακρίβεια σε άγνωστα δεδομένα. Η έρευνα μας έδειξε ότι ο αλγόριθμος Logistic Regression είναι εκείνος που πετυχαίνει υψηλότερη ακρίβεια με 91.67% στις προβλέψεις στα test δεδομένα και υψηλότερο AUC με 97.32%. Γι' αυτό το λόγο αποθηκεύσαμε τη γνώση που παράχθηκε από το συγκεκριμένο μοντέλο και κάθε φορά που χρησιμοποιούμε το σύστημα απλά τη φορτώνουμε και εκτελούμε τις προβλέψεις μας στα νέα δεδομένα που εισάγονται στο σύστημα.

Θέλουμε να τονίσουμε το σύστημα που αναπτύξαμε σε καμία περίπτωση δεν αντικαθιστούν τον ιατρό και την ιατρική εξέταση. Ο γιατρός μέσω της φυσικής εξέτασης και της επικοινωνίας με τον ασθενή μπορεί να λάβει υπόψιν παραμέτρους στη διάγνωση του που το σύστημα αγνοεί. Παρόλα αυτά το σύστημα μας παραμένει μια καλή προσέγγιση εάν ένας ασθενής επιθυμεί να λάβει μια πρώτη διάγνωση για την κατάσταση της καρδιάς του χωρίς να χρειάζεται να επισκεφθεί με φυσική παρουσία το γιατρό του. Επίσης, το σύστημα μπορεί να χρησιμοποιηθεί και τον ίδιο τον γιατρό σε περίπτωση που δεν είναι σίγουρος για τη διάγνωση του και επιθυμεί να λάβει ακόμα μια άποψη.

Όπως ήδη έχουμε αναφέρει το σύστημα χρησιμοποιεί 14 features για να εκτελέσει τις προβλέψεις του. Αυτό ίσως αποτελεί ένα μειονέκτημα του συστήματος εφόσον ένας απλός χρήστης να μην έχει τόσες πολλές πληροφορίες έτσι ώστε να είναι σε θέση να συμπληρώσει όλα τα πεδία που απαιτούνται για να εκτελέσει το σύστημα την πρόβλεψη του.

Τέλος, εικάζεται ότι μέχρι το 2045 η τεχνητή νοημοσύνη θα έχει φτάσει τις ικανότητες της βιολογικής νοημοσύνης. Έτσι, είναι πιθανόν μέχρι τότε τέτοιου είδους συστήματα να είναι ευρέως διαδεδομένα τα επόμενα χρόνια και να χρησιμοποιούνται από τον καθένα μας.

# Βιβλιογραφία

- [1] M. Seckeler and T. Hoke, "The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease," *Clinical epidemiology*, vol. 3, p. 67, 2011.
- [2] S. Weng, J. Reps, J. Kai, J. Garibaldi and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PloS one*, vol. 12, p. e0174944, 2017.
- [3] P. Singh, S. Singh and G. Pandi-Jain, "Effective heart disease prediction system using data mining techniques," *International journal of nanomedicine*, vol. 13, p. 121, 2018.
- [4] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu and J. Yan, "Study of cardiovascular disease prediction model based on random forest in eastern China," *Scientific reports*, vol. 10, pp. 1--8, 2020.
- [5] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, p. 1, 2017.
- [6] A. Janosi, W. Steinbrunn, M. Pfisterer and R. Detrano, "Machine Learning Repository," UCI, [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>. [Accessed April 2021].
- [7] N. Gogtay and U. Thatte, "Principles of correlation analysis," *Journal of the Association of Physicians of India*, vol. 65, pp. 78--81, 2017.
- [8] J. Peng, L. Lee and M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, pp. 3--14, 2002.
- [9] P. Cunningham and S. Delany, "k-Nearest neighbour classifiers," *Mult Classif Syst*, 2007.
- [10] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, pp. 183-197, 1991.
- [11] H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 74--78, 2018.
- [12] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5--32, 2001.
- [13] H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," *Artificial Intelligence and Signal Processing Conference (AISP)*, pp. 209--212, 2017.
- [14] M. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, pp. 108--132, 2000.
- [15] S. LaValle, M. Branicky and S. Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *The International Journal of Robotics Research*, vol. 23, pp. 673--692, 2004.
- [16] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, pp. 1145--1159, 1997.

# ΠΑΡΑΡΤΗΜΑ

## Οδηγίες χρήσης συστήματος

Το Σχήμα 21 παρουσιάζει τη διαπροσωπία του συστήματος που υλοποιήσαμε. Ο ασθενής καλείται να συμπληρώσει τα 13 πεδία που υπάρχουν στη φόρμα που αφορούν τα ιατρικά του δεδομένα. Αυτά τα δεδομένα περιγράφονται αναλυτικά και στο Κεφάλαιο 3.1. Πατώντας το κουμπί “Submit” το σύστημα χρησιμοποιεί το εκπαιδευμένο μοντέλο μηχανικής μάθησης, που είχε τη ψηλότερη ακρίβεια, για να κάνει την πρόβλεψη. Η πρόβλεψη παρουσιάζεται σε παράθυρο μηνύματος. Πατώντας το κουμπί “Clear” τα πεδία καθαρίζονται και επαναφέρονται στην αρχική τους κατάσταση.

♥ Prediction of heart disease

*Prediction of heart disease in a patient*

Age

Sex

Chest pain type

Resting blood pressure (mm Hg)

Serum cholestoral (mg/dl)

Fasting blood sugar (mg/dl)

Resting electrocardiographic results

Maximum heart rate achieved

Exercise induced angina

ST depression induced by exercise relative to rest

The slope of the peak exercise ST segment

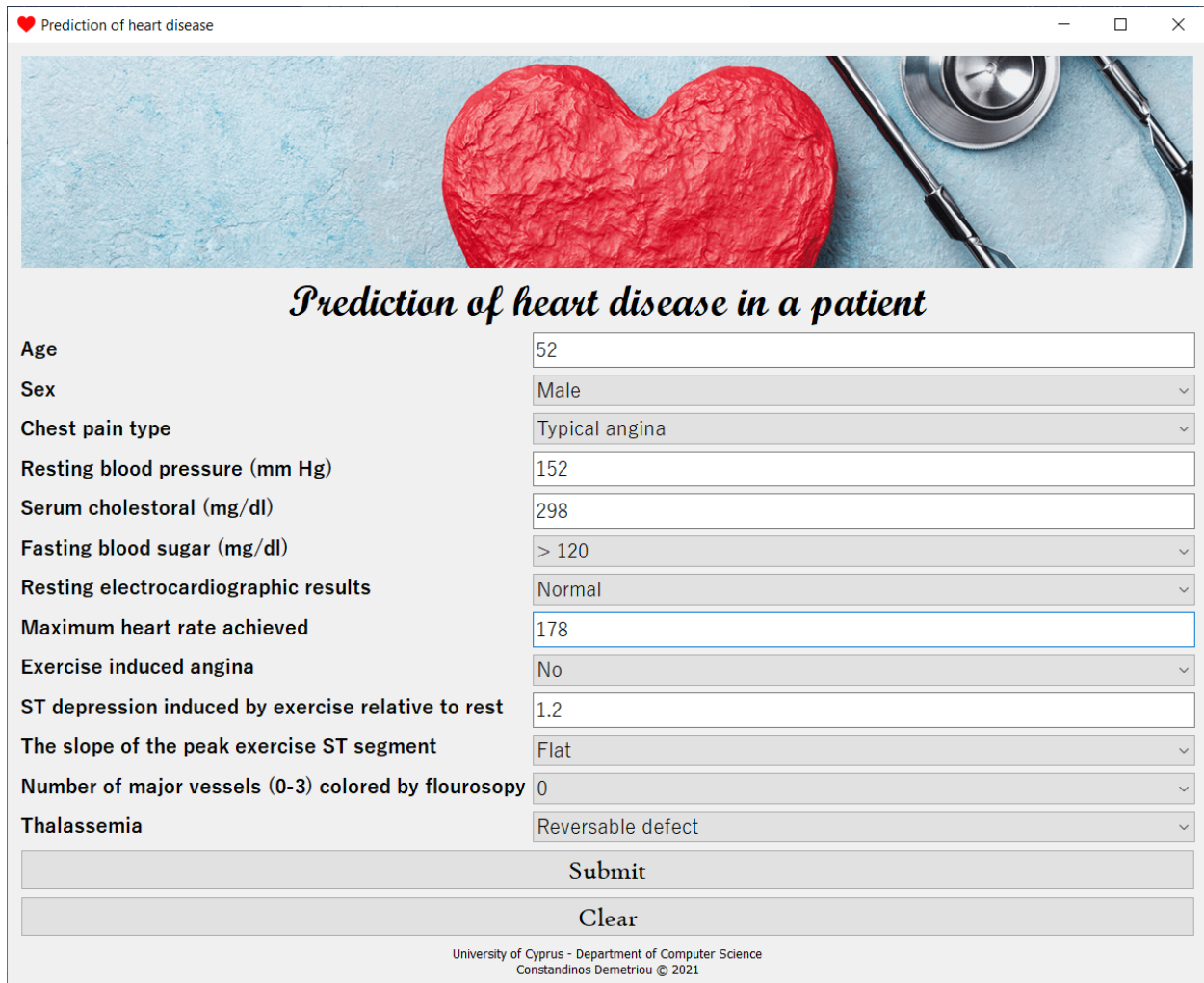
Number of major vessels (0-3) colored by flourosopy

Thalassemia

University of Cyprus - Department of Computer Science  
Constandinos Demetriou © 2021

Σχήμα 21: Διαπροσωπία του συστήματος

Το Σχήμα 22 παρουσιάζει τη διαπροσωπία του συστήματος συμπληρωμένη με τα δεδομένα ασθενή χωρίς καρδιακή πάθηση. Όταν ο χρήστης πατήσει το κουμπί “Submit” το σύστημα θα εκτελέσει την πρόβλεψη του και θα παρουσιαστεί το μήνυμα που φαίνεται στο Σχήμα 23.



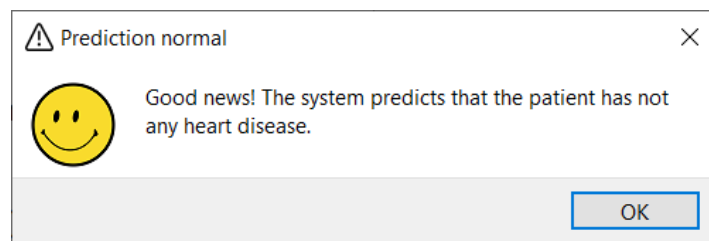
Field	Value
Age	52
Sex	Male
Chest pain type	Typical angina
Resting blood pressure (mm Hg)	152
Serum cholestoral (mg/dl)	298
Fasting blood sugar (mg/dl)	> 120
Resting electrocardiographic results	Normal
Maximum heart rate achieved	178
Exercise induced angina	No
ST depression induced by exercise relative to rest	1.2
The slope of the peak exercise ST segment	Flat
Number of major vessels (0-3) colored by flourosopy	0
Thalassemia	Reversable defect

Submit

Clear

University of Cyprus - Department of Computer Science  
Constandinos Demetriou © 2021

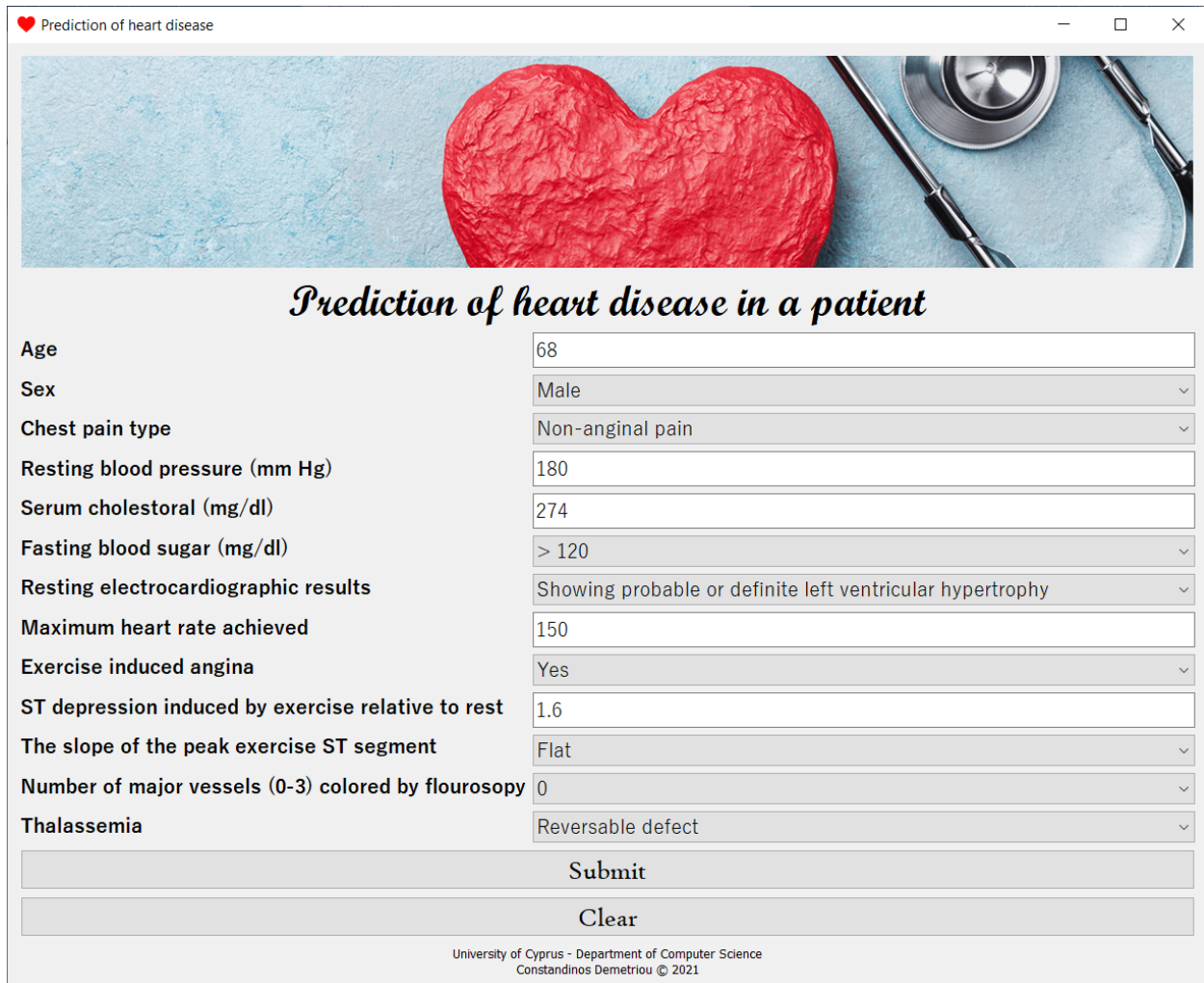
**Σχήμα 22:** Διαπροσωπία του συστήματος συμπληρωμένη με τα δεδομένα ασθενή χωρίς καρδιακή πάθηση



**Σχήμα 23:** Μήνυμα που παρουσιάζεται όταν το σύστημα προβλέψει ότι ο ασθενής δεν αντιμετωπίζει κάποια καρδιακή πάθηση



Το Σχήμα 24 παρουσιάζει τη διαπροσωπία του συστήματος συμπληρωμένη με τα δεδομένα ασθενή με καρδιακή πάθηση. Όταν ο χρήστης πατήσει το κουμπί “Submit” το σύστημα θα εκτελέσει την πρόβλεψη του και θα παρουσιαστεί το μήνυμα που φαίνεται στο Σχήμα 25.



Age 68

Sex Male

Chest pain type Non-anginal pain

Resting blood pressure (mm Hg) 180

Serum cholestoral (mg/dl) 274

Fasting blood sugar (mg/dl) > 120

Resting electrocardiographic results Showing probable or definite left ventricular hypertrophy

Maximum heart rate achieved 150

Exercise induced angina Yes

ST depression induced by exercise relative to rest 1.6

The slope of the peak exercise ST segment Flat

Number of major vessels (0-3) colored by flourosopy 0

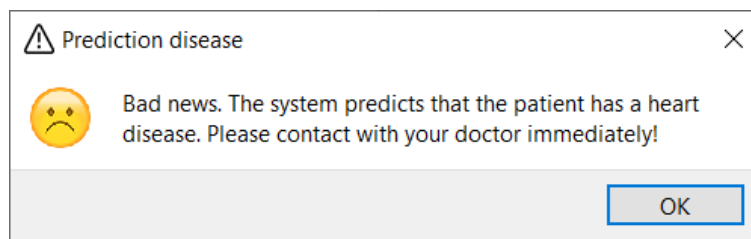
Thalassemia Reversable defect

Submit

Clear

University of Cyprus - Department of Computer Science  
Constandinos Demetriou © 2021

**Σχήμα 24:** Διαπροσωπία του συστήματος συμπληρωμένη με τα δεδομένα ασθενή με καρδιακή πάθηση



**Σχήμα 25:** Μήνυμα που παρουσιάζεται όταν το σύστημα προβλέψει ότι ο ασθενής αντιμετωπίζει κάποια καρδιακή πάθηση