ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

Ανάπτυξη κλινικού συστήματος υποστήριξης απόφασης για τη διάγνωση της καρδιακής στεφανιαίας νόσου

Κωνσταντίνος Πολυέζος
Αριθμός Μητρώου: 19017
mcse19017@uniwa.gr

Επιβλέπων Καθηγητής
Δρ. Πάρις Μαστοροκώστας

Αθήνα 2021

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η μεταπτυχιακή διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
	ΠΑΡΙΣ ΜΑΣΤΟΡΟΚΩΣΤΑΣ	ΚΑΘΗΓΗΤΗΣ	PARIS MASTOROKOSTAS Digitally signed by PARIS MASTOROKOSTAS Date: 2021.12.13 21:45:45 +02'00'
	ΓΕΩΡΓΙΟΣ ΠΡΕΖΕΡΑΚΟΣ	ΚΑΘΗΓΗΤΗΣ	Georgios Prezerakos Digitally signed by Georgios Prezerakos Date: 2021.12.18 15:42:57 +02'00'
	ΧΡΗΣΤΟΣ ΣΚΟΥΡΛΑΣ	ΟΜΟΤΙΜΟΣ ΚΑΘΗΓΗΤΗΣ	Christos Skourlas Digitally signed by Christos Skourlas Date: 2021.12.15 16:46:05 +02'00'

Δήλωση Συγγραφέα Μεταπτυχιακής Εργασίας

Ο κάτωθι υπογεγραμμένος Πολυέζος Κωνσταντίνος του Ηλία, με αριθμό μητρώου 19017, φοιτητής του Προγράμματος Μεταπτυχιακών Σπουδών «*Επιστήμη και Τεχνολογία της Πληροφορικής και των Υπολογιστών*» του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών

Αθήνα, Δευτέρα, 13 Δεκεμβρίου 2021

"Αλήθεια, τι δυνατή μηχανή που είναι ο άνθρωπος.
Τον ταΐζεις με ψωμί, κρασί και ψάρι και σου βγάζει νοήματα, γέλια και όνειρα"
Νίκου Καζαντζάκη, Βίος και πολιτεία του Αλέξη Ζορμπά

Ανάπτυξη κλινικού συστήματος υποστήριξης απόφασης για τη διάγνωση της καρδιακής στεφανιαίας νόσου

Σκοπός

Η αρτηριακή αποφρακτική νόσος, που συναντάται σε διάφορες μορφές, είναι στις μέρες μας η μεγαλύτερη αιτία θανάτου στις αναπτυγμένες χώρες. Η απόφραξη των αγγείων-αρτηριών οφείλεται σε μεγάλο ποσοστό στη δημιουργία της αθηρωματικής πλάκας. Ο προσδιορισμός και ο εντοπισμός της προδιάθεσης της αθηροσκλήρωσης αποτελεί πάρα πολύ σημαντικό βήμα για τη πρόληψη της αρτηριακής αποφρακτικής νόσου και συνεπώς την ελάττωση της νοσηρότητας και θνησιμότητας από αυτήν. Σκοπός της εργασίας είναι η ανάπτυξη ενός συστήματος υποβοήθησης διάγνωσης της αθηρωματικής νόσου, το οποίο βασίστηκε σε αλγόριθμους Μηχανικής Μάθησης.

Μεθοδολογία

Προς την κατεύθυνση αυτή, το σύνολο της μεταπτυχιακής εργασίας επικεντρώθηκε στην αξιολόγηση της συνεισφοράς πέντε τεχνικών Μηχανικής Μάθησης (Λογιστικής Παλινδρόμησης, Μηχανών Διανυσματικής Υποστήριξης, Απλού Μπέυζ ταξινομητή, Πολυεπίπεδου Νευρωνικού Δικτύου Perceptron-MLP, Δέντρων Απόφασης) για τη κατασκευή ενός αυτοματοποιημένου συστήματος πρόγνωσης της καρδιακής αρτηριακής αποφρακτικής νόσου. Για το σκοπό αυτό, με τη βοήθεια της γλώσσας προγραμματισμού ανοικτού κώδικα R, συγκρίθηκαν οι πέντε αλγόριθμοι με βάση το πλαίσιο δεδομένων Statlog (Heart) και αντιπαραβλήθηκαν τα αποτελέσματα των πέντε μεθόδων. Στη συνέχεια, στα πλαίσια της υλοποίησης ενός συστήματος υποστήριξης απόφασης, σχεδιάστηκε και η αντίστοιχη διαδραστική web εφαρμογή με τη βοήθεια της πλατφόρμας R Shiny.

Αποτελέσματα

Τα αποτελέσματα της επικύρωσης των μοντέλων, ανέδειξαν την υπεροχή του Νευρωνικού Δικτύου Perceptron με μέθοδο βελτιστοποίησης την αποσύνθεση βαρών (weight decay). Με αυτή τη πλέον πρόσφορη επιλογή αλγορίθμου, υλοποιήθηκε το κλινικό σύστημα υποστήριξης απόφασης για τη διάγνωση της καρδιακής ισχαιμικής νόσου. .

Συμπεράσματα

Στο τελευταίο μέρος της εργασίας συνοψίζονται τα αποτελέσματα της μελέτης και επιχειρείται η ερμηνεία τους.

Λέξεις κλειδιά: καρδιακές παθήσεις, μηχανική μάθηση, συστήματα υποστήριξης απόφασης, στεφανιαία νόσος,

Development of a clinical decision support system for the diagnosis of coronary heart disease

Purpose

Obstructive arterial disease, which exists in various forms, is now the leading cause of death in developed countries. The blockage of the arteries is due to a large percentage of the formation of atherosclerotic plaque. Determining and locating the predisposition to atherosclerosis is a very important step in the prevention of arterial occlusive disease and therefore the reduction of morbidity and mortality from it. The aim of this thesis is to develop a system to assist in the diagnosis of atherosclerotic disease, which was based on Machine Learning algorithms.

Methodology

Based on the above, this thesis focused on the evaluation of the contribution of five Machine Learning Techniques (Logistic regression, Support Vector Machines, Naive Bayes classificatior, Perceptron-MLP Multilevel Neural Network, Decision Trees) for the construction of an automated system for the prevention of coronary artery disease. For this purpose, with the help of the open source programming language R, the five algorithms were compared based on the Statlog (Heart) dataset and the results of the five methods were compared. Consequently, in the context of the implementation of a decision support system, the corresponding interactive web application was designed with the help of the R Shiny platform.

Results

The results of the validation of the models showed the superiority of the Perceptron Neural Network with weight decay regularization technique. With this most convenient algorithm option, the clinical decision support system for the diagnosis of ischemic heart disease was implemented.

Conclusions

In the last part of the thesis, the results of the study are summarized and their interpretation is attempted.

Keywords: Cleveland heart disease dataset, Statlog (Heart) Data Set, coronary heart disease (CHD), clinical decision support system

Η εργασία αυτή αφιερώνεται στη σύζυγό μου Μαρία και στη κόρη μου Φωτεινή

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή ..... - 10 -
1. Βιβλιογραφική Έρευνα ..... - 11 -
1.1 Τα Κλινικά Συστήματα Υποβοήθησης Απόφασης ..... - 11 -
2. Θεωρητικό Υπόβαθρο ..... - 16 -
2.1 Καρδιακές Παθήσεις ..... - 16 -
2.2 Η Στεφανιαία νόσος ..... - 17 -
2.3 Αιτίες καρδιακών παθήσεων ..... - 18 -
2.4 Το πρότζεκτ Statlog ..... - 19 -
2.5 Το πλαίσιο δεδομένων Statlog (Heart) ..... - 20 -
2.6 Μηχανική μάθηση ..... - 22 -
2.7 Επικύρωση μοντέλου ..... - 23 -
2.8 Επιδόσεις ανά κλάση ..... - 25 -
2.9 Ανάλυση του πλαισίου δεδομένων ..... - 27 -
2.10 Λογιστική Παλινδρόμηση ..... - 30 -
2.10.1 Εισαγωγικά στοιχεία ..... - 30 -
2.10.2 Γραμμική παλινδρόμηση ..... - 31 -
2.10.3 Λογιστική παλινδρόμηση ..... - 32 -
2.10.4 Γενικευμένα γραμμικά μοντέλα ..... - 32 -
2.10.5 Μετρήσεις καλής προσαρμογής ..... - 33 -
2.11 Μηχανές Διανυσμάτων Υποστήριξης SVM ..... - 36 -
2.11.1 Γραμμικά διαχωρίσιμες κλάσεις σε επίπεδο δυο διαστάσεων ..... - 36 -
2.11.2 Γραμμική Κατηγοριοποίηση (Hard margin) ..... - 38 -
2.11.3 Μη Γραμμική Κατηγοριοποίηση (Soft margin) ..... - 39 -
2.11.4 Εσωτερικά γινόμενα ..... - 41 -
2.11.5 Συναρτήσεις Πυρήνα (Kernel Functions) ..... - 42 -
2.12 Νευρωνικό Δίκτυο ..... - 44 -
2.12.1 Ο βιολογικός νευρώνας ..... - 44 -
2.12.2 Ο τεχνητός νευρώνας (μοντέλο ΜακΚάλοχ-Πιτς) ..... - 45 -
2.12.3 Ο αισθητήρας perceptron (μοντέλο Ρόζενμπλατ) ..... - 46 -
2.12.4 Στοχαστική κατάβαση κλίσης ..... - 48 -
2.12.5 Πολυεπίπεδα δίκτυα perceptron ..... - 50 -
2.12.6 Εκπαίδευση πολυστρωματικού νευρωνικού δικτύου perceptron ..... - 52 -
2.13 Ταξινομητής Μπέϋζ ..... - 54 -
2.13.1 Το θεώρημα Μπέυζ ..... - 54 -
2.13.2 Υπό όρους ανεξαρτησία ..... - 56 -
2.13.3 Μπεϋζιανά δίκτυα ..... - 57 -
2.13.4 Ο απλός ταξινομητής Μπέυζ ..... - 58 -
2.14 Δέντρα απόφασης ..... - 60 -
2.14.1 Περιγραφή των Δέντρων Απόφασης ..... - 60 -
2.14.2 Ο αλγόριθμος CART ..... - 62 -
2.14.3 Αλγόριθμος C5.0 ..... - 65 -
3. Μεθοδολογία ..... - 67 -
3.1 Προετοιμασία Μοντέλου ..... - 68 -
3.2 Εκπαίδευση Πλαισίου Δεδομένων με Λογιστική Παλινδρόμηση ..... - 73 -
3.2.1 Εφαρμογή παλινδρόμησης LASSO στο μοντέλο ..... - 76 -
3.2.2 Απόδοση Συστήματος στη Λογιστική Παλινδρόμηση ..... - 80 -
3.3 Εκπαίδευση συστήματος με Μηχανές Διανυσμάτων Υποστήριξης ..... - 82 -
3.3.1 Μελέτη περίπτωσης με Γραμμικό Πυρήνα ..... - 82 -
3.3.2 Μελέτη περίπτωσης με Πυρήνα ακτινικής βάσης ..... - 86 -
3.3.3 Μελέτη περίπτωσης με Πολυωνυμικό πυρήνα ..... - 88 -
3.4 Εκπαίδευση Πλαισίου Δεδομένων με Νευρωνικό Δίκτυο ..... - 91 -
3.4.2 Μελέτη Περίπτωσης με Πολυεπίπεδο Νευρωνικό Δίκτυο Perceptron ..... - 91 -
3.5 Εκπαίδευση Πλαισίου Δεδομένων με Ταξινομητή Μπέϋζ .....

3.6 Εκπαίδευση Πλαισίου Δεδομένων με Δέντρα Απόφασης ..... - 98 -
3.6.1 C50 ..... - 98 -
3.6.2 CART ..... - 101 -
3.7 Υλοποίηση της εφαρμογής στο προγραμματιστικό περιβάλλον της R ..... - 104 -
4. Αποτελέσματα ..... - 114 -
5. Συζήτηση & Συμπεράσματα ..... - 117 -
Αναφορές - Πηγές ..... - 120 -

Περιεχόμενα εικόνων

Εικόνα 1 Η Κίνα αναμένεται να παίξει σημαντικό ρόλο στα συστήματα υποβοήθησης απόφασης τις επόμενες δεκαετίες. ..... - 16 -
Εικόνα 2 Σημεία στο χώρο με καμπύλες ROC ..... - 26 -
Εικόνα 3 Σύγκριση κατηγοριοποιητών ..... - 27 -
Εικόνα 5 Γραμμική παλινδρόμηση ..... - 31 -
Εικόνα 6 Γραμμικά διαχωρίσιμες κλάσεις σε επίπεδο δυο διαστάσεων. ..... - 37 -
Εικόνα 7 Γραφική απεικόνιση περιθωρίου ..... - 40 -
Εικόνα 8 Γραμμικός ταξινομητής που έχει x μονάδες εισόδου, καθεμία από τις οποίες αντιστοιχεί στην τιμή ενός στοιχείου του διανύσματος εισόδου. Κάθε είσοδος (τιμή χαρακτηριστικού x_i πολλαπλασιάζεται με το αντίστοιχο βάρος της β_i ..... - 42 -
Εικόνα 9 Το ανθρώπινο νευρικό κύτταρο ..... - 44 -
Εικόνα 10 Το μοντέλο McCulloch-Pitts για τον τεχνητό νευρώνα ..... - 45 -
Εικόνα 11 Το μοντέλο Minsky-Papert για τον τεχνητό νευρώνα perceptron ..... - 47 -
Εικόνα 12 Ο τρόπος διαχωρισμού δυο κλάσεων με βάση τον νευρώνα perceptron ..... - 48 -
Εικόνα 13 Πολυστρωματικό Perceptron με δυο κρυμμένα επίπεδα ..... - 51 -
Εικόνα 14 Γραφική απεικόνιση παραδείγματος με τρεις ανεξάρτητες μεταβλητές ..... - 56 -
Εικόνα 15 Απλό Μπευζιανό δίκτυο με βάση το προηγούμενο παράδειγμα ..... - 57 -
Εικόνα 16 Αριστερά το αρχικό Μπευζιανό δίκτυο και δεξιά το απλοποιημένο ..... - 58 -
Εικόνα 17 Παράδειγμα ενός δικτύου Μπέυζ ..... - 59 -
Εικόνα 18 Ένα απλό παράδειγμα δέντρου απόφασης με δυο μεταβλητές X1 και X2 ..... - 61 -
Εικόνα 19 Απεικόνιση προηγούμενου παραδείγματος σε χώρο δυο διαστάσεων ..... - 62 -
Εικόνα 20 Κριτήριο διαχωρισμού για δυαδική ταξινόμηση. ..... - 66 -
Εικόνα 21 Συνεισφορά μεταβλητών του πλαισίου δεδομένων Statlog ..... - 71 -
Εικόνα 22 Μεταβλητές που αλληλεπιδρούν περισσότερο με την OUTPUT ..... - 72 -
Εικόνα 23 Γραφική παράσταση της μεταβλητής OUTPUT σε σχέση με την MAXHR ..... - 75 -
Εικόνα 24 Εφαρμογή παλινδρόμησης LASSO ..... - 79 -
Εικόνα 25 Καμπύλη ROC για το μοντέλο Λογιστικής Παλινδρόμησης ..... - 81 -
Εικόνα 26 Γραφική παράσταση της ακρίβειας σε σχέση με το βάρος του κόστους των λανθασμένων ταξινομήσεων για SVM με γραμμικό πυρήνα ..... - 85 -
Εικόνα 27 Γραφική παράσταση της ακρίβειας σε σχέση με το βάρος του κόστους των λανθασμένων ταξινομήσεων για SVM με Πυρήνα ακτινικής βάσης ..... - 87 -
Εικόνα 28 ακρίβεια σε σχέση την παράμετρο πολυπλοκότητας για διάφορες τιμές πολυωνύμων - 89 -
Εικόνα 29 Ο βέλτιστος αριθμός βαρών με το ελάχιστο μέσο τετράγωνο σφάλμα ..... - 93 -
Εικόνα 30 Το νευρωνικό δίκτυο με βάση το πλαίσιο δεδομένων Statlog (heart) ..... - 94 -
Εικόνα 31 Γραφική παράσταση της ακρίβειας σε σχέση με τη ρύθμιση του εύρους. Από αριστερά για την περίπτωση που δεν εφαρμόζεται λείανση Laplace και δεξιά για την περίπτωση που εφαρμόζεται αυτή η τεχνική ..... - 97 -
Εικόνα 32 Γραφική παράσταση ακρίβειας για 20 επαναληπτικές ωθήσεις ..... - 100 -
Εικόνα 33 Δέντρο απόφασης για αλγόριθμο C50 ..... - 100 -
Εικόνα 34 Γραφική παράσταση του συντελεστή πολυπλοκότητας σε σχέση με το συντελεστή RMSE ..... - 103 -

Εικόνα 35 Γραφική παράσταση της ακρίβειας του μοντέλου σε σχέση με την παράμετρο πολυπλοκότητας ..... - 103 -
Εικόνα 36 Δέντρο απόφασης για αλγόριθμο CART ..... - 104 -
Εικόνα 37 Αποτελέσματα από την εκτέλεση της εφαρμογής ..... - 105 -
Εικόνα 38 Παράμετροι εισόδου πρώτη περίπτωση ..... - 107 -
Εικόνα 39 Παράμετροι εισόδου δεύτερη περίπτωση ..... - 108 -
Εικόνα 40 Παράμετροι εισόδου τρίτη περίπτωση ..... - 109 -
Εικόνα 41 Παράμετροι εισόδου τέταρτη περίπτωση ..... - 110 -
Εικόνα 42 Φωτογραφίες από την εφαρμογή 1 ..... - 111 -
Εικόνα 43 Φωτογραφίες από την εφαρμογή 2 ..... - 112 -
Εικόνα 44 Φωτογραφίες από την εφαρμογή ..... - 113 -

Περιεχόμενα πινάκων

Πίνακας 1 Πλεονεκτήματα/Μειονεκτήματα Συστημάτων KBS ..... - 12 -
Πίνακας 2 Πλεονεκτήματα/Μειονεκτήματα Συστημάτων NonKBS ..... - 13 -
Πίνακας 3 Κατηγοριοποίηση των CDSS (Berlin A, Sorani M, Sim I. A, 2006) ..... - 13 -
Πίνακας 4 Σετ δεδομένων καρδιακών παθήσεων ..... - 21 -
Πίνακας 5 Συγκεντρωτική περιγραφή μεταβλητών πλαισίου δεδομένων statlog ..... - 30 -
Πίνακας 6 Συγκεντρωτικός πίνακας συχνοτήτων ..... - 114 -
Πίνακας 7 Στατιστικά μεταβλητών πλαισίου δεδομένων ..... - 115 -
Πίνακας 8 Συγκεντρωτικά αποτελέσματα για τη Λογιστική Παλινδρόμηση ..... - 116 -
Πίνακας 9 Συγκεντρωτικά αποτελέσματα για τις Μηχανές Διανυσμάτων Υποστήριξης ..... - 116 -
Πίνακας 10 Συγκεντρωτικά αποτελέσματα για τα Νευρωνικά Δίκτυα ..... - 116 -
Πίνακας 11 Συγκεντρωτικά αποτελέσματα για Μπεϋζιανό ταξινομητή ..... - 116 -
Πίνακας 12 Συγκεντρωτικά αποτελέσματα για τα Δέντρα Απόφασης ..... - 116 -
Πίνακας 13 Συγκεντρωτικά αποτελέσματα για όλους τους αλγόριθμους Μηχανικής Μάθησης της εργασίας ..... - 116 -

Εισαγωγή

Η στεφανιαία νόσος είναι από τα μεγαλύτερα προβλήματα υγείας, επειδή αποτελεί την πρώτη αιτία θανάτου στις σύγχρονες Δυτικές κοινωνίες. Το έτος 2001 ήταν η αιτία για το 1/3 των θανάτων στον κόσμο. Υπάρχει η εκτίμηση ότι λόγο της ραγδαίας αύξησης του αριθμού των ατόμων που πάσχουν από καρδιαγγειακά νοσήματα, το 2021 θα ευθύνονται για τον θάνατο σχεδόν 25 εκατομμυρίων ατόμων παγκοσμίως. Ειδικά οι φτωχές πληθυσμιακές ομάδες που δεν έχουν πρόσβαση στις διαγνωστικές εξετάσεις παρουσιάζουν μεγάλα ποσοστά καρδιαγγειακών παθήσεων [1].[2]

Προς την κατεύθυνση αυτή, η παρούσα μεταπτυχιακή διπλωματική εργασία αρχικά ασχολείται με την δημιουργία ενός προγνωστικού μοντέλου της αθηρωματικής νόσου με αλγόριθμους Μηχανικής Μάθησης. Η βασική επιδίωξη της διπλωματικής εργασίας είναι η σχεδίαση και η επιτυχής ανάλυση ενός Κλινικού Συστήματος Υποστήριξης Απόφασης. Η εφαρμογή των γνώσεων που αποκτήθηκαν για την καλύτερη κατανόηση των θεμάτων που αφορούν τον τομέα της Μηχανικής Μάθησης είναι επίσης ζητούμενο σε αυτήν την τελευταία φάση του ΠΜΣ. Επομένως, η διπλωματική εργασία αποτελεί πρωτίστως ευκαιρία για μια εφαρμοσμένη μελέτη που ίσως αποτελέσει και εφαλτήριο για μια μελλοντική σταδιοδρομία.

Συγκεκριμένα, στο Κεφάλαιο 1 πραγματοποιείται μια εκτενής βιβλιογραφική ανασκόπηση των συστημάτων υποστήριξης απόφασης. Σε μεγάλο μέρος της, η βιβλιογραφική προσέγγιση βασίζεται στην επιστημονική δημοσίευση των Berlin και Sorani [19], σχετικά με τη ταξινόμηση-κατάταξη των κλινικών συστημάτων βασιζόμενων στη γνώση τα οποία υποστηρίζουν δραστηριότητες λήψης αποφάσεων.

Στο Κεφάλαιο 2 παρουσιάζεται η θεωρητική πτυχή της μεταπτυχιακής διατριβής. Αναλύεται θεματικά η καρδιακή ισχαιμική νόσος και γίνεται ενδελεχής ανάλυση των προσδιοριστών του πλαισίου δεδομένων statlog(heart) που βρίσκεται στο διαδικτυακό αποθετήριο του UCI (https://archive.ics.uci.edu/ml/index.php). Επίσης, σκιαγραφείται η επιστημονική περιοχή της Μηχανικής Μάθησης (Machine Learning) και των αλγορίθμων αυτής, που χρησιμοποιούνται στην εν λόγω εργασία. Σε αυτό το στάδιο, η επίλυση προβλημάτων με ευφυείς προσεγγίσεις που βασίζεται στη δυνατότητα των συστημάτων να μαθαίνουν αποκτά διάσταση στην πρόβλεψη των καρδιακών παθήσεων. Στο ίδιο κεφάλαιο, γίνεται λόγος για την ακρίβεια ενός μοντέλου που πρέπει να εκτιμάται έναντι παρατηρήσεων, οι οποίες δεν ανήκουν σε ένα σύνολο εκπαίδευσης. Παρουσιάζονται η μέθοδος holdout, η διασταυρούμενη επικύρωση 10 τμημάτων, η μέθοδος «άφησε ένα έξω» και η μέθοδος bootstrap. Για την εκτίμηση και παρουσίαση της ικανότητας των μοντέλων να προβλέπουν συγκεκριμένη τιμή κλάσης έχουν προταθεί ειδικές τεχνικές που αναφέρονται ως πίνακας σύγχυσης (confusion matrix) και οι καμπύλες ROC (Receiver Operating Characteristics).

Στο Κεφάλαιο 3 ακολουθούν τα βήματα της μεθοδολογίας. Με τη βοήθεια της γλώσσας προγραμματισμού R, πραγματοποιείται η προετοιμασία του μοντέλου (Παραγ. 3.1), προκειμένου να είναι έτοιμο για την μετέπειτα προγραμματιστική επεξεργασία. Ακολούθως παρουσιάζονται, εκπαιδεύονται και αξιολογούνται αλγόριθμοι όπως, η Λογιστική (ή Λογαριθμική) Παλινδρόμηση (Παραγ. 3.2), οι Μηχανές Διανυσμάτων Υποστήριξης (Παραγ. 3.3), τα Νευρωνικά Δίκτυα (Παραγ. 3.4), ο απλός ταξινομητής Μπέυζ και το δίκτυο Μπέυζ (Παραγ.

3.5) και τα Δένδρα Απόφασης (Παραγ. 3.6). Στη συνέχεια με βάση τη μέθοδο που παρουσιάζει την υψηλότερη ακρίβεια μοντέλου, δομείται εφαρμογή λογισμικού συστήματος υποστήριξης της καρδιακής ισχαιμικής νόσου με τη βοήθεια της πλατφόρμας R και του αντίστοιχου πακέτου R Shiny (Παραγ. 3.7) .

Στο Κεφάλαιο 4 παρουσιάζονται συγκριτικά τα αποτελέσματα, όπου διαφαίνεται ότι το πολυεπίπεδο Νευρωνικό Δίκτυο Perceptron έχει τη μεγαλύτερη απόδοση με βάση το συγκεκριμένο βελτιστοποιημένο πλαίσιο δεδομένων, όπως αυτό προέκυψε από την προεργασία του Κεφαλαίου 3. Στο Κεφάλαιο 5 συνοψίζονται τα αποτελέσματα της μελέτης και επιχειρείται η ερμηνεία τους.

1. Βιβλιογραφική Έρευνα

1.1 Τα Κλινικά Συστήματα Υποβοήθησης Απόφασης

Η λήψη ιατρικών αποφάσεων από τους γιατρούς, στη καθημερινή διαγνωστική διαδικασία, απαιτεί γνώσεις και εμπειρία. Η διάγνωση στηρίζεται κατά κύριο λόγο στα συμπτώματα και σε διάφορα διαγνωστικά κριτήρια τους ασθενούς. Για την αντιμετώπιση των περιπτώσεων αυτών πρέπει αρκετές φορές να εισαχθούν μέθοδοι υποβοήθησης της λήψης αποφάσεων, που θα στηρίζονται σε συγκεκριμένες διεργασίες και θα εκτελούνται με τη βοήθεια της τεχνολογίας [16]. Η μεγάλη ανάπτυξη των σύγχρονων τεχνολογιών πληροφορικής έχουν αναδείξει νέες τάσεις στην ανάπτυξη των υπολογιστικών συστημάτων, με στόχο την υποβοήθηση κλινικών αποφάσεων διαγνωστικού χαρακτήρα με βάση εξατομικευμένα δεδομένα του ασθενούς. Σύμφωνα με τον ορισμό του Dinevski [17], τα Κλινικά Συστήματα Υποβοήθησης Λήψης Αποφάσεων (CDSSs) είναι εφαρμογές Η/Υ οι οποίες υποβοηθούν τον γιατρό στην λήψη απόφασης για συγκεκριμένα περιστατικά (case-specific). Τα συστήματα υποστήριξης κλινικών αποφάσεων συμβάλλουν στη βελτίωση των παρεχόμενων υπηρεσιών υγείας και ταυτόχρονα μειώνουν το κόστος (π.χ. ελαχιστοποίησης ιατρικών λαθών ή της ελαχιστοποίησης του αριθμού επανεισαγωγών μέσω της χρήσης φορητών συσκευών από τους ασθενείς). Ένα σύστημα υποστήριξης κλινικών αποφάσεων είναι σχεδιασμένο ώστε να χρησιμοποιεί μαθηματικά μοντέλα προσομοίωσης, γνωσιακές βάσεις δεδομένων, μεθόδους αναγνώρισης προτύπων και την Τεχνητή Νοημοσύνη για την κωδικοποίηση της υπάρχουσας γνώσης και για την επίλυση των προβλημάτων που προκύπτουν στην κλινική πράξη [15].

Οι Akerkar και Sajja [18] προσδιόρισαν τους βασικούς παράγοντες που οδηγούν στην ταχεία εφαρμογή και ανάπτυξη των CDSSs :

- Η μονιμότητα της γνώσης
- Η διαθεσιμότητα
- Η αποδοτικότητα (ταχύτητα, ακρίβεια, έλεγχος και αποθήκευση)
- Η συνέπεια και η αξιοπιστία
- Η αιτιολόγηση του "συλλογισμού" λήψης απόφασης
- Η ικανότητα αυτό-εκπαίδευσης και αναπροσαρμογής

Ενώ οι Berlin, Sorani και Sim [19] ανέδειξαν τις κατηγορίες και επιχείρησαν έναν οντολογικό προσδιορισμό των CDSSs. Υπάρχουν δυο βασικές κατηγορίες που σχετίζονται με τη λειτουργία των CDSSs: Τα *Forward Chain Systems* και τα *Backward Chain Systems*. Στην πρώτη περίπτωση, παρουσιάζονται οι συνθήκες του προβλήματος στο σύστημα και το σύστημα επεξεργάζεται και συστήνει την ανάλογη δράση (έξοδος). Στην δεύτερη περίπτωση, παρουσιάζονται μια ή παραπάνω υποθέσεις στο σύστημα και το σύστημα ανακτά τα σχετικά με την υπόθεση δεδομένα, τα επεξεργάζεται και επιβεβαιώνει ή απορρίπτει την υπόθεση.

Τα CDSS επίσης, διακρίνονται και ως προς την αρχιτεκτονική σε δυο κατηγορίες: Συστήματα Βασισμένα Στην Εξαγωγή Γνώσης (*Knowledge Base Systems/KBS*) και Συστήματα που Δεν Βασίζονται στην Εξαγωγή Γνώσης. (*Non-Knowledge Base Systems/NKBS*) 0. Στα KBS η βάση γνώσης των συστημάτων περιλαμβάνει δεδομένα και επεξεργασμένες πληροφορίες, κανόνες και πρωτόκολλα κωδικοποιημένα, ευρετικούς και εμπειρικούς κανόνες. Σε αυτό το στάδιο, η δημιουργία της KB απαιτεί την βοήθεια "Ειδικών" στο εκάστοτε κλινικό αντικείμενο, καθώς και ευρύτερη γνώση στο πεδίο (βιβλιογραφική γνώση, μαθηματικά μοντέλα, κλινικά πρωτόκολλα κλπ). Η εξαγωγή των συμπερασμάτων (μηχανή εξαγωγής συμπερασμάτων) γίνεται με επιλογές των κατάλληλων "κανόνων" από την KB. Τα NonKBS συστήματα παίρνουν την πληροφορία (είσοδοι) και την επεξεργάζονται για την παραγωγή "συμβουλής". Η βάση γνώσης δεν υπάρχει ως ξεχωριστή οντότητα αλλά είναι εμπεδωμένη στην μηχανή εξαγωγής συμπερασμάτων κατά τη διάρκεια ανάπτυξης του συστήματος. Τα συστήματα αυτά απαιτούν μεγάλο αριθμό δεδομένων για να αναπτυχθεί η δομή τους. Οι Ειδικοί έχουν περιορισμένο ρόλο στην ανάπτυξη του συστήματος.

Ένα σύστημα KBS δέχεται μια σειρά από δεδομένα όπως μετρήσεις φυσιολογίας, δημογραφικά κ.α. Χρησιμοποιεί μια βάση γνώσεων και μια μηχανή εξαγωγής συμπερασμάτων (κανόνες, δέντρα αποφάσεων, πίνακες αληθείας, δίκτυα αποφάσεων) και παράγει μια έξοδο (σύσταση, διάγνωση, σύγκριση αρχικών υποθέσεων κ.α.). Στα συστήματα NonKBS η Βάση Γνώσεων είναι εμπεδωμένη στην μηχανή εξαγωγής συμπερασμάτων. Αν για παράδειγμα, η μηχανή εξαγωγής συμπερασμάτων είναι ένα σύστημα ασαφούς λογικής, η γνώση είναι εμπεδωμένη στους κανόνες και στις γλωσσικές μεταβλητές (ασαφή σύνολα) 0. Αν η μηχανή εξαγωγής συμπερασμάτων είναι ένα εκπαιδευόμενο Νευρωνικό Δίκτυο, τότε τα βάρη των συνδέσεων είναι η αποκομισμένη γνώση από την εκπαίδευση, καθώς και η αρχιτεκτονική, εμπεριέχονται στο δίκτυο. Αντίστοιχα, σε ένα σύστημα Μπεϋσιανών δικτύων, η γνώση είναι αποθηκευμένη στην δομή και στις πιθανότητες.

Πίνακας 1 Πλεονεκτήματα/Μειονεκτήματα Συστημάτων KBS

Πλεονεκτήματα	***Μειονεκτήματα***
Εξαγωγή γνώσης από ειδικούς	Εξαγωγή γνώσης από ειδικούς
Εύκολη εξαγωγή συμπερασματων	Μεγάλες Βάσεις Δεδομένων
Μονιμότητα της γνώσης	Δεν έχουν ικανότητα αυτοεκπαίδευσης
Αποδοτικότητα	Η KB είναι η γνώση του μηχανικού
Συνέπεια στις απόφασεις	Συνέπεια στις αποφάσεις

Πίνακας 2 Πλεονεκτήματα/Μειονεκτήματα Συστημάτων NonKBS

Πλεονεκτήματα	*Μειονεκτήματα*
Δεν απαιτείται η εξαγωγή γνώσης από ειδικούς	Η ποιότητα του συστήματος εξαρτάται από τον όγκο και την ποιότητα των δεδομένων
Μονιμότητα της γνώσης	Καλή απόδοση σε δεδομένα του παρελθόντος. Δεν είναι γνωστό πως ανταποκρίνονται σε νέες συνθήκες.
Αποδοτικότητα	Κάποια από τα συστήματα (NN) είναι "Μαύρα Κουτιά", δεν είναι για τον τελικό χρήστη διαυγής η σχέση αιτίου-αιτιατού.
Μπορούν να ξαναεκπαιδευτούν όταν έχουμε διαθέσιμα νέα δεδομένα	

Πίνακας 3 Κατηγοριοποίηση των CDSS (Berlin A, Sorani M, Sim I. A, 2006)

Με βάση το ευρύτερο πλαίσιο (Context)	*Περιγραφή*
Κλινική περιοχή	Νοσηλευτικές μονάδες / εξωτερικά ιατρεία
Κλινικό έργο	Πρόγνωση / διάγνωση / θεραπεία
Περιοχή βελτιστοποίησης	Κλινικά αποτελέσματα/ αποτελέσματα του συστήματος
Τρόπος εφαρμογής	Ανεξάρτητα-απόμακρα από τον ασθενή / ενώπιον του ασθενή/ σε συνεργασία με τον ασθενή
Βαθμός υποχρεωτικής εφαρμογής	Οι συστάσεις είναι υποχρεωτικές / ή συμβουλευτικές με βάση την πολιτική του οργανισμού
Πιθανοί φραγμοί στην εφαρμογή	Οικονομικοί / κοινωνικοί / ηθικοί φραγμοί / νομικοί

Με βάση τη Βάση Γνώσης και τις πηγές δεδομένων	*Περιγραφή*
Πηγή Γνώσης	Ειδικοί, πρωτόκολλα, μαθηματικά μοντέλα, αποθηκευμένα δεδομένα και πληροφορίες
Πηγή δεδομένων	Φάκελος (χαρτί) ασθενούς, Ηλεκτρονικός Φάκελος, Αυτόματη απαγωγή
Κωδικοποίηση δεδομένων	Κείμενο, αριθμητικές τιμές, πιθανότητες
Βαθμός προσαρμογής	Βαθμός στον οποίο οι αποφάσεις προσαρμόζονται στον συγκεκριμένο ασθενή, κλινικά δεδομένα και ιστορικό
Μηχανισμός ενημέρωσης	Μηχανισμός αναπροσαρμογής συστήματος σε νέα «δεδομένα», αυτόματος, χειροκίνητος

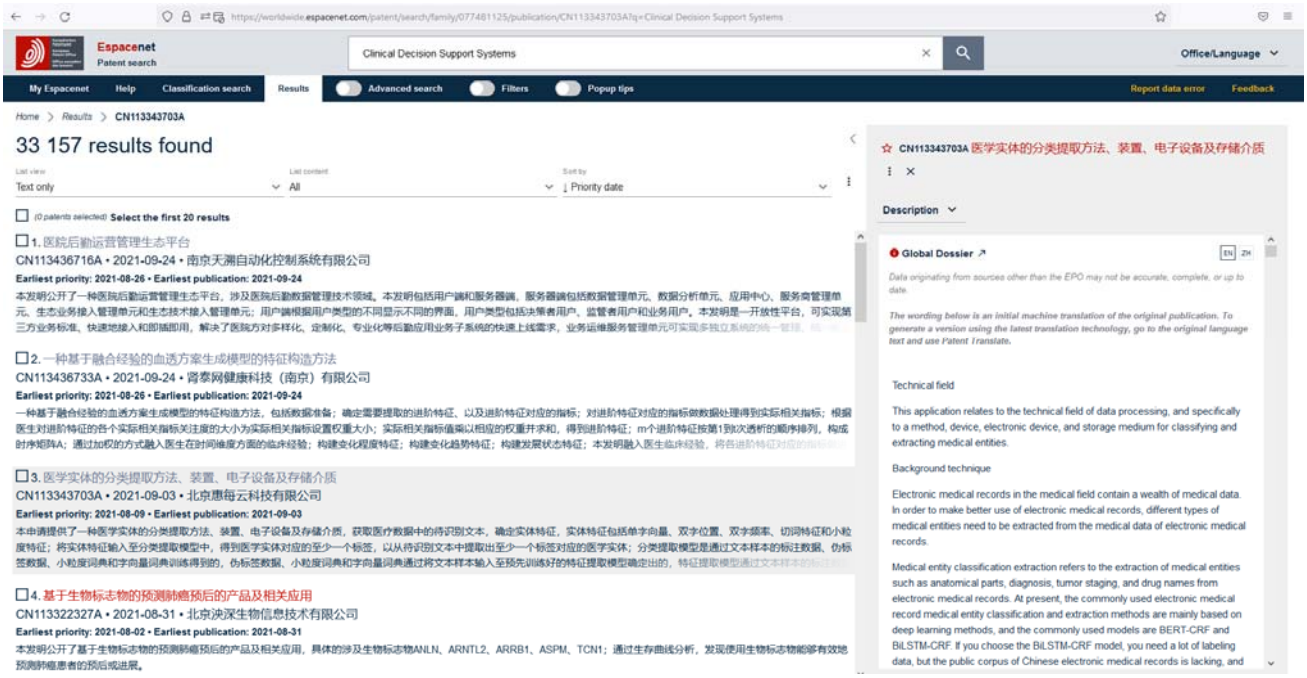
Με βάση την Απόφαση	*Περιγραφή*
Μέθοδος Εξαγωγής Συμπερασμάτων	Κανόνες, Bayesian δίκτυα, Fuzzy, Neural Networks…
Βαθμός επείγοντος εφαρμογής απόφασης	Η σύσταση πρέπει να εφαρμοστεί άμεσα ή σε εύθετο χρόνο
Βαθμός κατηγορηματικότητας	Ο βαθμός στον οποίο οι συστάσεις είναι κατηγορηματικές και δεν υπεισέρχεται βαθμός αβεβαιότητας
Βαθμός περιπλοκότητας	Βαθμός περιπλοκότητας διαδικασίας λήψης απόφασης
Βαθμός ανάδρασης	Κατά πόσο ο γιατρός πρέπει να ενημερώσει το σύστημα για τις προθέσεις του να εφαρμόσει τις συστάσεις

Με βάση την μέθοδο «παροχής» της σύστασης	***Περιγραφή***
Τρόπος απεικόνισης συμβουλής	Δικτυακά, οθόνη λογισμικού, εκτύπωση
Τρόπος συμβουλής	Βαθμός «αυθαιρεσίας» αποφάσεων
Ενσωμάτωση περαιτέρω βημάτων	Εάν το σύστημα προτείνει κλινικά βήματα ταυτόχρονα με τη συμβουλή
Διαθεσιμότητα επεξήγησης	Εάν το σύστημα επεξηγεί την λογική της απόφασης
Δια-δραστικότητα	Εάν το σύστημα είναι δια-δραστικό προς τον χρήστη

Με βάση την ενσωμάτωση στην ροή εργασιών (workflow)	***Περιγραφή***
Εξακρίβωση χρήστη	Χρήση με ιδιότητα (ασθενής, γιατρός, μη κλινικό προσωπικό)
Στοχευμένη παροχή συμβουλών	Ιδιότητα-ταυτοποίηση προσώπου που πρέπει να ενημερωθεί για τις συμβουλές
Ταυτοποίηση ενδιάμεσου	Ταυτοποίηση του προσώπου που εισάγει τα στοιχεία
Ταυτοποίηση αποδέκτη	Ταυτοποίηση του προσώπου αποδέκτη της συμβουλής
Ενσωμάτωση στην ροή	Εάν το σύστημα απαιτεί εργασίες οι οποίες δεν είναι τμήμα της καθημερινής εργασίας του προσωπικού, εάν οι λοιπές διαδικασίες παγώνουν ως αποτέλεσμα των συστάσεων

Οι Ledley και Lusted το 1959 [33] ήταν οι πρώτοι που δημοσίευσαν άρθρο σχετικό με την υποβοήθηση λήψης κλινικής απόφασης. Στο άρθρο, γιατροί έλαβαν οδηγίες πώς να δημιουργήσουν διαγνωστικές βάσεις δεδομένων χρησιμοποιώντας κάρτες με εγκοπές (edge-notched cards). Ο Horrocks [34] χρησιμοποίησε Μπαεϋσιανά Δίκτυα (Πιθανότητες) για να διαγνώσει την πιο πιθανή πάθηση με συμπτώματα ισχυρού πόνου στην κοιλιακή χώρα. Αρκετά πρωτοποριακό για την εποχή του διαδραστικό πρόγραμμα υπολογιστών υλοποιήθηκε από τον Shortliffe το 1974 και ονομάστηκε MYCIN [35]. Το πρόγραμμα αυτό βασίστηκε σε μεγάλο βαθμό σε τεχνικές τεχνητής νοημοσύνης (AI), στις οποίες χρησιμοποιήθηκαν κανόνες λήψης αποφάσεων για να συμβουλεύσουν τους γιατρούς σχετικά με την κατάλληλη επιλογή θεραπείας μολυσματικών ασθενειών. Ο Bates το 1998 [36] δημιούργησε το πληροφοριακό σύστημα CPOE που αφορά την ορθή συνταγογράφηση φαρμάκων σε ασθενείς με χρήση λογισμικού υποστήριξης αποφάσεων προκειμένου να αποφεύγονται οι ανεπιθύμητες ενέργειες φάρμακων (Adverse Drug Events). Με μαζική αποστολή μηνυμάτων ηλεκτρονικού ταχυδρομείου υπενθύμισης εμβολιασμού της γρίπης, οι Baker και άλλοι [37], κατάφεραν ποσοστό αύξησης εμβολιασμού 45.2% σύμφωνα με την έρευνα. Σχετική δημοσίευση των Bogusevicius, Maleckas κ.α. σχετικά με την υποβοήθηση διάγνωσης με τη βοήθεια υπολογιστή της απόφραξης του λεπτού εντέρου σε σύγκριση με τις κλασσικές απεικονιστικές τεχνικές [38] ανέδειξε τη δυνατότητα πρόβλεψης των CDSS σε πολύ μικρότερο χρονικό διάστημα.

Ένα αρκετά επιτυχημένο CDSS που βρίσκεται ακόμη σε λειτουργία και ξεκίνησε να αναπτύσσεται το 1983, δημιουργήθηκε από τον Pryor του πανεπιστήμιο της Γιούτα και εφαρμόζεται στο νοσοκομείο LDS των ΗΠΑ [39]. Το σύστημα HELP όπως ονομάστηκε, είναι πλέον απαραίτητο εργαλείο για την κάλυψη των διοικητικών, κλινικών, διδακτικών και ερευνητικών αναγκών του νοσοκομείου, καθώς και για την παροχή της δυνατότητας λήψης αποφάσεων. Ο Manotti [40] ανέπτυξε ένα πρόγραμμα δοσολογίας υπολογιστικής νοημοσύνης για παρακολούθηση και καθοδήγηση της από του στόματος αντιπηκτικής θεραπείας σε εξωτερικούς ασθενείς. Οι Boukhors κ.α. υλοποίησαν ένα υπολογιστικό σύστημα δοσολογίας ινσουλίνης για ασθενείς με σακχαρώδη διαβήτη τύπου 1 [41]. Το 2002 δημιουργείται το πρότζεκτ CROPS: Ένα σύστημα ηλεκτρονικής αναφοράς παθολογίας καρκίνου με δομημένη εισαγωγή δεδομένων [42]. Το λογισμικό CaseWalker που αφορούσε την εισαγωγή δεδομένων με καθοδήγηση από υπολογιστή, επέδειξε πολύ μεγαλύτερη αποτελεσματικότητα σε σχέση με τις γραπτές υπενθυμίσεις, στη βελτίωση της τήρησης μιας κατευθυντήριας γραμμής κλινικής πρακτικής για τη διάγνωση και θεραπεία της κατάθλιψης [43]. Οι Lutz και άλλοι [44] μελέτησαν την αποτελεσματικότητα των ενημερωτικών δελτίων προερχόμενων από υπολογιστή στη βελτίωση του αριθμού και της ποικιλίας φρούτων και λαχανικών που καταναλώνονται από ενήλικες. Οι Goodey και Brickley [45] μελέτησαν σύστημα υπολογιστικής ευφυΐας βασισμένο σε νευρωνικά δίκτυα για την καθοδήγηση απόφασης παραπομπής σε στοματικό χειρουργό για αφαίρεση τρίτου κάτω γομφίου (φρονιμίτη). Οι Bakar και άλλοι [46] ανέπτυξαν διαδικτυακή εφαρμογή υποβοήθησης διάγνωσης της καρδιακής ισχαιμικής νόσου με το όνομα HDP. Το πλαίσιο δεδομένων ήταν το statlogHeart και ο αλγόριθμος εκπαίδευσης Νευρωνικό Δίκτυο. Οι Oluoch κ.α. [50] μελέτησαν την αποτελεσματικότητα των κλινικών συστημάτων υποστήριξης αποφάσεων σε ανθρώπους που είχαν μολυνθεί με HIV και ξεκίνησαν αντιρετροϊκή θεραπεία (ART) σε χώρες με περιορισμένους πόρους. Η έρευνα βασίστηκε σε άρθρα από τις βάσεις δεδομένων MEDLINE, EMBASE, CINAHL και Global Health Library μέχρι και τον Ιανουάριο του 2012. Οι μελέτες, αρκετές από τις οποίες πραγματοποιήθηκαν στην υποσαχάρια Αφρική, ανέδειξαν ως ιδανική λύση για την φροντίδα ασθενών με HIV, τα κλινικά συστήματα υποστήριξης με βάση τα ηλεκτρονικά αρχεία ασθενούς (EMR). Στα αποτελέσματα της έρευνας υπήρξαν αναφορές για μείωση των ιατρικών λαθών, μείωση των χαμένων ραντεβού, ελαχιστοποίηση σφαλμάτων των αποτελεσμάτων στους προγνωστικούς δείκτες CD4 (Τ βοηθητικά κύτταρα) και μείωση του χρόνου αναμονής των ασθενών. Αυτό που αξίζει να αναφερθεί είναι ότι μετά από σχετική έρευνα στη βάση δεδομένων του «Ευρωπαϊκού Γραφείου Διπλωμάτων Ευρεσιτεχνίας» (Esp@cenet) σχετικά με τις αναδυόμενες τάσεις στις τεχνολογίες των κλινικών συστημάτων απόφασης, αναδείχθηκε ότι τα συστήματα υποβοήθησης με αναγνώριση προτύπων (pattern recognition) έχουν κυρίαρχο ρόλο. Επίσης σημειώνεται εντυπωσιακή άνοδος των διπλωμάτων ευρεσιτεχνίας που προέρχονται από την Κίνα, (τέλη 2012 –σήμερα), κάτι που ίσως δείχνει ότι στο εγγύς μέλλον θα υπερισχύσει ως τεχνολογικός κολοσσός.



Εικόνα 1 Η Κίνα αναμένεται να παίξει σημαντικό ρόλο στα συστήματα υποβοήθησης απόφασης τις επόμενες δεκαετίες.

2. Θεωρητικό Υπόβαθρο

Σε αυτή την ενότητα γίνεται η παρουσίαση του θεωρητικού πλαισίου της εργασίας. Περιλαμβάνει μια ανασκόπηση των καρδιακών παθήσεων με έμφαση τη στεφανιαία νόσο και μια αναφορά στο πρότζεκτ statlog που βασίζεται το σετ δεδομένων της εργασίας. Επιπρόσθετα αναλύονται οι 14 παράμετροι που σχετίζονται με το σετ δεδομένων Statlog Heart που αφορά την πρόβλεψη της ισχαιμικής νόσου. Ακολουθεί η θεωρητική προσέγγιση των πέντε μοντέλων μηχανικής μάθησης που θα χρησιμοποιηθούν για την εκπαίδευση του προβλήματος ταξινόμησης.

2.1 Καρδιακές Παθήσεις

Σύμφωνα με τον Παγκόσμιο οργανισμό Υγείας[1] (WHO), με τον όρο "*καρδιακές παθήσεις*" εννοούμε ένα σύνολο ασθενειών που προσβάλουν την καρδιά και τα αιμοφόρα αγγεία. Σε αυτές περιλαμβάνονται:

1.Η στεφανιαία νόσος, που ονομάζεται και ισχαιμική πάθηση.
2.Η νόσος των αρτηριών που τροφοδοτούν τον εγκέφαλο με αίμα και ευθύνεται για τα εγκεφαλικά επεισόδια.
3.Η περιφερειακή αρτηριακή νόσος που επηρεάζει τις αρτηρίες που τροφοδοτούν με αίμα τα άκρα, δηλαδή τα πόδια και τα χέρια.
4.Η ρευματική καρδίτιδα, μια ασθένεια που οφείλεται στη μόλυνση από το βακτηρίδιο του στρεπτόκοκκου και προκαλεί ασθένεια των μυών και των βαλβίδων της καρδιάς.
5.Μια πολύπλοκη κατηγορία εκ γενετής παθήσεων λόγω δυσγενεσιών των δομών της καρδιάς και των αγγείων του κυκλοφορικού συστήματος. Αυτή η κατηγορία είναι γνωστή και ως "Συγγενείς καρδιοπάθειες".
6.Οι βαθιές θρομβώσεις και πνευμονικές εμβολές που βρίσκουν αίτιο στους σχηματισμούς θρόμβων στις

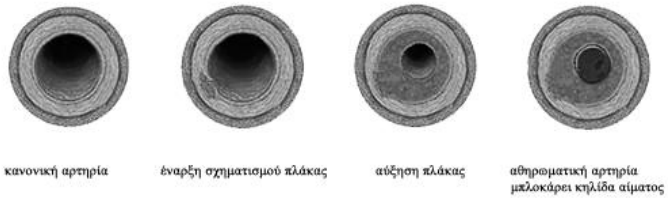
φλέβες των κάτω άκρων και που υπάρχει η πιθανότητα να διασπαστούν, να αποκολληθούν και να μεταφερθούν στην καρδιά και στους πνεύμονες.

Στην έκθεση της οργάνωσης για την υγεία WHO υπάρχουν αρκετά αξιοσημείωτες και ανησυχητικές αναφορές. Οι καρδιακές παθήσεις αποτελούν τον πρώτο παράγοντα θανάτων παγκοσμίως. Το έτος 2012 πέθαναν από κάποια καρδιακή νόσο περίπου 17.5 εκατομμύρια άνθρωποι, δηλαδή το 31% των αιτιών θανάτου σε παγκόσμιο επίπεδο. Προσδιορίζεται ότι 7.4 εκατομμύρια άνθρωποι έφυγαν από τη ζωή λόγω της στεφανιαίας νόσου, ενώ 6.7 εκατομμύρια πληθυσμού υπέστησαν εγκεφαλικά επεισόδια. Ποσοστό του 1/3 εμφανίζεται σε χώρες με χαμηλά και μεσαία εισοδήματα. Στις ηλικίες κάτω των 70 ετών, στην κατηγορία θανάτων προερχομένων από μη-μεταδοτικές ασθένειες, το 37% είναι καρδιακές παθήσεις.

Στον Ευρωπαϊκό χώρο, τα καρδιαγγειακά νοσήματα προκαλούν κάθε χρόνο 4 εκατομμύρια θανάτους. Η θνησιμότητα που παρουσιάζει η στεφανιαία νόσος στον ελληνικό πληθυσμό, όπως έχει αναφερθεί από την Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ), είναι 110 θάνατοι ανά 100.000 άτομα.

2.2 Η Στεφανιαία νόσος

Οι καρδιακές παθήσεις, και ειδικά η στεφανιαία νόσος[1], είναι αρκετά επικίνδυνες γιατί αν ο ασθενής αγνοήσει τα πρώιμα συμπτώματα, υπάρχει πιθανότητα να οδηγηθεί σε μια σοβαρή υποκείμενη καρδιαγγειακή διαταραχή που θα προκαλέσει ακαριαίο θάνατο. Ο θωρακικός πόνος, οι ενοχλήσεις στα κάτω άκρα, η βραχύπνοια, το αίσθημα δυσφορίας που συνοδεύεται από ιδρώτα, ζαλάδα ή χλωμάδα, δυσκολίες ομιλίας και κατανόησης του λόγου και το συχνότερο σύμπτωμα της αποπληξίας είναι συμπτώματα καρδιακής νόσου και που σε τέτοιες περιπτώσεις, θα πρέπει να ζητείται άμεσα η ιατρική συμβουλή. Η στεφανιαία νόσος[2] δημιουργείται όταν οι αθηρωματικες πλάκες (χοληστερόλη, λιπώδη στοιχεία, ινώδης ιστός και εναποθέσεις ασβεστίου), γεμίζουν τις στεφανιαίες αρτηρίες της καρδιάς με αποτέλεσμα να εμποδίζεται η ροή του αίματος σε αυτήν (αθηροσκλήρωση[2]). Η στένωση του αυλού της αρτηρίας έχει ως συνέπεια τη μειωμένη παροχή οξυγόνου και θρεπτικών ουσιών στους ιστούς της καρδιάς. Ο συνεχιζόμενος σχηματισμός του αποφρακτικού θρόμβου, με την πάροδο του χρόνου, έχει ως αποτέλεσμα την παντελή και παρατεταμένη έλλειψη οξυγόνου στο μυοκάρδιο, η οποία με τη σειρά της προκαλεί νέκρωση του μυοκαρδίου (έμφραγμα).



Εικόνα 2: Η εξέλιξη της στεφανιαίας νόσου.
Πηγή : http://www.montcopa.org/2319/Heart-Disease-Coronary-Artery-Disease

[1] Ταξινόμηση ICD-9 414 και ICD-10 I25.8

[2] *Αθήρωμα*, λατ. *atheroma*, ονομάζεται η απόφραξη της αρτηρίας —πιο συγκεκριμένα του αυλού της αρτηρίας— από υπολείμματα κυττάρων και περιλαμβάνει λιπίδια, ασβέστιο και ινώδη συνδετικό ιστό. Έχει χρώμα «αθηρό», δηλαδή υπόξανθο ή υποκίτρινο, από το οποίο παίρνει το όνομά του

2.3 Αιτίες καρδιακών παθήσεων

Οι σημαντικότεροι παράγοντες κινδύνου που ευθύνονται για τις καρδιακές παθήσεις και τα εγκεφαλικά επεισόδια είναι η ανθυγιεινή διατροφή, η σωματική αδράνεια και η χρήση καπνού και αλκοόλ. Οι επιδράσεις αυτών των παραγόντων κινδύνου συνήθως εμφανίζονται στα άτομα ως αυξημένη αρτηριακή πίεση, αυξημένη γλυκόζη αίματος, αυξημένα λιπίδια αίματος και ως παχυσαρκία. Αυτοί οι «ενδιάμεσοι παράγοντες κινδύνου» εντοπίζονται στις εγκαταστάσεις πρωτοβάθμιας περίθαλψης και υποδηλώνουν υψηλότερο κίνδυνο καρδιακής προσβολής, εγκεφαλικού επεισοδίου, καρδιακής ανεπάρκειας και άλλων επιπλοκών.

Η διακοπή της χρήσης του τσιγάρου, η μείωση του αλατιού στη διατροφή, η κατανάλωση περισσότερων φρούτων και λαχανικών, η τακτική σωματική άσκηση και η αποφυγή της κατάχρησης αλκοόλ έχουν αποδειχθεί ότι μειώνουν τον κίνδυνο των καρδιαγγειακών παθήσεων. Συνεπώς, η χάραξη πολιτικής προσανατολισμένης στην υγεία, που έχει ως στόχο να δημιουργήσει ευνοϊκά περιβάλλοντα ώστε να παρακινηθούν οι άνθρωποι να υιοθετήσουν και να διατηρήσουν υγιείς συμπεριφορές, αποτελεί επιτακτική ανάγκη.

Υπάρχει επίσης ένας αριθμός υποκείμενων καθοριστικών παραγόντων των καρδιακών παθήσεων. Οι παράγοντες αυτοί αντικατοπτρίζονται σε συνθήκες που οφείλονται στην κοινωνική, οικονομική και πολιτιστική αλλαγή, στην παγκοσμιοποίηση, στην αστικοποίηση και στη γήρανση του πληθυσμού. Άλλοι καθοριστικοί παράγοντες των καρδιαγγειακών νοσημάτων έχουν ως αιτία τη φτώχεια, το άγχος και την κληρονομικότητα.

Επιπλέον, η λήψη φαρμακευτικής θεραπείας της υπέρτασης, του διαβήτη και των υψηλών λιπιδίων στο αίμα είναι απαραίτητη για τη μείωση του καρδιαγγειακού κινδύνου και την πρόληψη καρδιακών προσβολών και εγκεφαλικών επεισοδίων σε άτομα με αυτές τις παθήσεις.

2.4 Το πρότζεκτ Statlog

Το πρότζεκτ statlog [47] συστήθηκε στις αρχές της δεκαετίας του 1990 και χρηματοδοτήθηκε από την Ευρωπαϊκή Ένωση. Περιείχε 20 αλγορίθμους ταξινόμησης και 23 σετ δεδομένων. Οι σκοποί του ήταν τρεις:

• Αξιολόγηση των διαφορετικών προσεγγίσεων των μεθόδων ταξινόμησης
• Σύγκριση των αποδόσεων αυτών των μεθόδων σε σχέση με ένα ευρύ φάσμα σετ δεδομένων
• Εξαγωγή συμπερασμάτων σχετικά με την δυνατότητα εφαρμογής των υπό εξέταση σετ δεδομένων σε ρεαλιστικές εφαρμογές της βιομηχανίας.

Το σετ δεδομένων statlog που αφορά τις καρδιακές παθήσεις περιέχει 270 παρατηρήσεις, 13 προσδιοριστές και 2 κλάσεις και προέρχεται από τη βάση δεδομένων του Κλίβελαντ (Cleveland database).Οι 13 προσδιοριστές εξήχθησαν από ένα σύνολο 75 αρχικών προσδιοριστών. Αρχικά αυτή η βάση δεδομένων περιείχε 303 περιπτώσεις, αλλά επειδή σε 6 από αυτές λείπανε τιμές, κατέληξε με 297. Επιπρόσθετα, προκειμένου να ελαχιστοποιηθεί ο χρόνος επεξεργασίας και εμφάνισης των αποτελεσμάτων, το statlog περιέχει 2 κλάσεις, αντί 4 που είχε αρχικά. Με άλλα λόγια, ενώ στο σετ δεδομένων του Κλίβελαντ υπήρχαν 4 προγνωστικές περιπτώσεις (0=απουσία καρδιακής πάθησης, 1=ήπια ή μέτρια πιθανότητα, 2=μέτρια ή σοβαρή πιθανότητα, 3=εμφάνιση νόσου), στο statlog σετ δεδομένων οι περιπτώσεις περιορίστηκαν σε 2 (1=εμφάνιση, 0=απουσία νόσου) .

Το πιο σημαντικό στοιχείο είναι ότι η αξιολόγηση του statlog έγινε με βάση το κόστος του σφάλματος ταξινόμησης. Αποδόθηκε σημασία, δηλαδή, στο κόστος της λανθασμένης ταξινόμησης (misclassification), αναφορικά με τη βαρύτητα ενός σφάλματος. Ο Elkan [48] όρισε μαθηματικά τη διαδικασία λήψης αποφάσεων βασισμένες σε έναν πίνακα κόστους. Αν θεωρήσουμε (i, j) τις προσθήκες ενός πίνακα κόστους C, τότε ορίζουμε ως i το κόστος της κλάσης πρόβλεψης όταν η πραγματική κλάση είναι j . Αν $i= j$ **.** τότε η πρόβλεψη είναι σωστή. Αν $i\neq j$ τότε η πρόβλεψη είναι λανθασμένη. Η βέλτιστη πρόβλεψη για ένα παράδειγμα x είναι όταν η κλάση i ελαχιστοποιεί την εξίσωση:

$$\sum_j L(x, i) P(j \mid x)\, C(i, j)$$

Για κάθε i , το $L(x, i)$ είναι ένα άθροισμα των εναλλακτικών πιθανοτήτων για την πραγματική τιμή του x Μέσα σε αυτό το πλαίσιο, ο ρόλος του αλγόριθμου μάθησης είναι να παράγει έναν ταξινομητή, που για κάθε παράδειγμα x να υπολογίζει την πιθανότητα $P(j \mid x)$ κάθε κλάσης j δεδομένου ότι το x αποτελεί πραγματική κλάση. Στα προηγούμενα σετ δεδομένων καρδιακών παθήσεων του Ιρβάιν, δεν είχε ληφθεί υπόψη ή ανάλυση κόστους, παράμετρος που είναι ιδιαίτερα σημαντική για τις ιατρικές εφαρμογές.

Για να υπολογιστεί το μέσο σφάλμα ταξινόμησης χρησιμοποιήθηκε διασταυρούμενη επικύρωση 9 τμημάτων (nine fold cross-validation). Για όλες τις κατηγορίες αλγορίθμων που χρησιμοποιήθηκαν, το μέσο κόστος υπολογίστηκε πολλαπλασιάζοντας τις τιμές του πίνακα σύγχυσης (confusion matrix) με τα βάρη των τιμών του πίνακα κόστους. Στην συνέχεια, αθροίστηκαν τα γινόμενα και διαιρέθηκαν με τον αριθμό των παρατηρήσεων.

Πίνακας κόστους της βάσης δεδομένων του statlog

	πρόβλεψη απουσίας	πρόβλεψη παρουσίας
απουσία εκδήλωσης	0	1
παρουσία εκδήλωσης	5	0

Οι περιπτώσεις μοναδιαίου κόστους, δεν λαμβάνονται υπόψη και γενικά θεωρείται ότι απλά δημιουργούν μια δυσαρέσκεια στον εξεταζόμενο. Για σφάλματα ταξινόμησης όμως που αρχικά η πρόβλεψη παρουσίας της νόσου ήταν αρνητική και στη συνέχεια ο εξεταζόμενος παρουσίασε την πάθηση, το κόστος είναι πενταπλάσιο. Σύμφωνα με τα δεδομένα της ανάλυσης, [47], οι πέντε πρώτοι αλγόριθμοι με βάση το μικρότερο κόστος, ήταν ο απλός Μπευζ Ταξινομητής (1), η Γραμμική Διαχωριστική Ανάλυση (2), η Λογιστική Διαχωριστική Ανάλυση (3), ο στατιστικός αλγόριθμος ALLOC80 (4) και η Τετραγωνική Διαχωριστική Ανάλυση (5).

2.5 Το πλαίσιο δεδομένων Statlog (Heart)

Τα σετ δεδομένων που αφορούν την διάγνωση των καρδιακών παθήσεων είναι τέσσερα και βρίσκονται αναρτημένα στο διαδικτυακό αποθετήριο βάσεων δεδομένων για χρήση σε μεθόδους μηχανικής μάθησης «UCI» (https://archive.ics.uci.edu/ml/index.html). Αυτές οι βάσεις δεδομένων συλλέχθηκαν από τον καθηγητή David Aha [3] του Πανεπιστημίου της Καλιφόρνια στο Ιρβάιν το 1988. Τα δεδομένα αντλήθηκαν από τις ακόλουθες τοποθεσίες :

1. Ιατρικό κέντρο του Κλίβελαντ στο Οχαϊο των ΗΠΑ (cleveland.data)
2. Ουγγρικό ινστιτούτο καρδιολογίας στη Βουδαπέστη της Ουγγαρίας (hungarian.data)
3. Ιατρικό κέντρο στο Λονγκ Μπιτς της Καλιφόρνια των ΗΠΑ (long-beach-va.data)
4. Πανεπιστημιακό νοσοκομείο της Ζυρίχης στην Ελβετία (switzerland.data)

Από αυτές τις τέσσερις βάσεις δεδομένων έχει επικρατήσει και χρησιμοποιείται ευρύτατα στη βιβλιογραφία το σετ δεδομένων του Κλίβελαντ. Οι μεταβλητές εισόδου αφορούν τόσο αποτελέσματα εργαστηριακών εξετάσεων, όσο και δημογραφικά στοιχεία.

Ωστόσο, στις δημοσιεύσεις χρησιμοποιείται ένα μικρότερο σετ 14 μεταβλητών. Καταργήθηκαν δηλαδή μεταβλητές που θεωρήθηκαν πλεονάζουσες και διατηρήθηκαν οι περισσότερο σημαντικές.

Στον παρακάτω πίνακα 5, απεικονίζονται οι 14 μεταβλητές, η σημασία των οποίων εξηγείται αναλυτικά σε επόμενο κεφάλαιο.

Πίνακας 4 Σετ δεδομένων καρδιακών παθήσεων

Σετ Δεδομένων:	0	1	2	3	4	Σύνολο Περιπτώσεων
Κλίβελαντ:	164	55	36	35	13	303
Βουδαπέστης:	188	37	26	28	15	294
Ζυρίχης:	8	48	32	30	5	123
Λονγκ Μπιτς:	51	56	41	42	10	200
Σύνολο	411	196	135	135	43	920

Αρκετές μέθοδοι ταξινόμησης έχουν δημιουργηθεί με βάση το σετ δεδομένων του Κλίβελαντ. Το 1989 ο Detrano δημιούργησε το Cleveland heart disease dataset και εφαρμόζοντας παράλληλα μέθοδο λογιστικής παλινδρόμησης κατάφερε ακρίβεια της τάξης του 77% [4]. Οι Chen και άλλοι [5] χρησιμοποίησαν εμπροσθοτροφοδοτούμενο πολλών επιπέδων νευρωνικό δίκτυο με δυαδικό ταξινομητή Perceptron (MLP) με ακρίβεια 80% . Ο Cheung [6] συνδύασε τον αλγόριθμο C4.5, Bayesian ταξινομητή, BNND και BNNF αλγορίθμους και επέτυχε ακρίβεια ταξινόμησης 81.11%, 81.48%, 81.11% και 80.96%, αντίστοιχα. Ο Senthil [7] με συνδυασμό των τεχνητών νευρωνικών δικτύων και ασαφούς λογικής σύστημα συμπερασμού επέτυχε ακρίβεια της τάξης 91.83%. Ο Bhuvaneswari [8] δημιούργησε ένα μηχανικό μοντέλο συνδυασμού γενετικών αλγορίθμων και τεχνητών νευρωνικών δικτύων. Χρησιμοποιήθηκαν γενετικοί αλγόριθμοι για τον υπολογισμό των βαρών ενός εμπροσθοτροφοδοτούμενου νευρωνικού δικτύου για την ταξινόμηση των καρδιακών παθήσεων. Η ακρίβεια ταξινόμησης γι' αυτό το μοντέλο ήταν 94.17%. Οι Das, Trukoglu και Sengur [9], χρησιμοποιώντας μια συγκεντρωτική μέθοδο νευρωνικών δικτύων, έφτασαν τα επίπεδα ακρίβειας στο 89.01%. Ο Can [10] με τη χρήση ανάλυσης πρωτευουσών συνιστωσών (PCA) και ενός συστήματος παράλληλων MLP νευρωνικών δικτύων πέτυχε ακρίβεια της τάξεως του 88.5%. Οι Karaduzovic και Köker [11], συνδύασαν δίκτυα ακτινικών συναρτήσεων βάσης (RBF) και εμπροσθοτροφοδοτούμενο πολλών επιπέδων νευρωνικό δίκτυο με δυαδικό ταξινομητή Perceptron (MLP) με ακρίβεια 95,5 %. Οι Dangare και Apte [12] προσθέσανε δυο μεταβλητές (obesity και smoking) στο σετ πρόβλεψης του Κλίβελαντ και κατάφεραν ακρίβεια 99.62% με πολυεπίπεδο νευρωνικό δίκτυο Perceptron. Βασιζόμενοι σε αυτή τη βάση δεδομένων και χρησιμοποιώντας απλό ταξινομητή Μπέϋζ, ο Subbalakshmi [13] και οι συνεργάτες του ανέπτυξαν ένα σύστημα υποβοήθησης απόφασης με την ονομασία DSHDPS, εξάγοντας με αυτό το μοντέλο κρυμμένη και χρήσιμη πληροφορία από τη βάση δεδομένων των καρδιακών παθήσεων. Ο E.P. Ephzibah [14] και άλλοι ερευνητές, προκειμένου να μειώσουν τον αριθμό των διαγνωστικών εξετάσεων, χρησιμοποίησαν 6 μεταβλητές από το αρχικό σετ δεδομένων του Κλίβελαντ και συνδύασαν γενετικούς αλγορίθμους για επιλογή χαρακτηριστικών και κανόνες ασαφούς λογικής για την ταξινόμηση.

2.6 Μηχανική μάθηση

Η μηχανική μάθηση είναι ένα υποπεδίο της τεχνητής νοημοσύνης (AI). Ο στόχος της μηχανικής μάθησης είναι γενικά η κατανόηση της δομής των δεδομένων και η προσαρμογή τους σε μοντέλα που μπορούν να γίνουν κατανοητά και να χρησιμοποιηθούν από τους ανθρώπους. [29]:

Αν και η μηχανική μάθηση ως πεδίο ανήκει στην επιστήμη των υπολογιστών, διαφέρει από τις παραδοσιακές υπολογιστικές προσεγγίσεις. Στην παραδοσιακή υπολογιστική, οι αλγόριθμοι είναι σύνολα ρητά προγραμματισμένων οδηγιών που χρησιμοποιούνται από τους υπολογιστές για τον υπολογισμό ή την επίλυση προβλημάτων. Οι αλγόριθμοι μηχανικής μάθησης επιτρέπουν στους υπολογιστές να εκπαιδεύονται σε δεδομένα εισόδου και να χρησιμοποιούν στατιστική ανάλυση για να εξάγουν τιμές που εμπίπτουν σε ένα συγκεκριμένο εύρος. Εξαιτίας αυτού, η μηχανική μάθηση διευκολύνει τους υπολογιστές στη δημιουργία μοντέλων από δείγματα δεδομένων, προκειμένου να αυτοματοποιηθούν οι διαδικασίες λήψης αποφάσεων με βάση τις εισαγωγές δεδομένων.

Στη μηχανική μάθηση, οι εργασίες ταξινομούνται γενικά σε ευρείες κατηγορίες. Αυτές οι κατηγορίες βασίζονται στο πώς προσλαμβάνεται η μάθηση ή πώς δίδεται ανατροφοδότηση για τη μάθηση στο εκπαιδευόμενο σύστημα.

Δύο από τις πιο διαδεδομένες μεθόδους μηχανικής μάθησης είναι η **εποπτευόμενη μάθηση** (Supervised Learning), η οποία εκπαιδεύει αλγορίθμους που βασίζονται σε παραδείγματα δεδομένων εισόδου και εξόδου που επισημαίνονται από τον άνθρωπο, και η **μη εποπτευόμενη μάθηση** (Unsupervised Learning), η οποία παρέχει στον αλγόριθμο δεδομένα χωρίς την επιθυμητή έξοδο. Ας εξερευνήσουμε αυτές τις μεθόδους με περισσότερες λεπτομέρειες.

Στην εποπτευόμενη μάθηση, ο υπολογιστής διαθέτει παραδείγματα εισόδων που επισημαίνονται με τις επιθυμητές εξόδους τους. Ο σκοπός αυτής της μεθόδου είναι ο αλγόριθμος να μπορεί να «μάθει», συγκρίνοντας την πραγματική του απόδοση με τις «διδαγμένες» εξόδους, για να βρει σφάλματα και να τροποποιήσει ανάλογα το μοντέλο. Συνεπώς, η εποπτευόμενη μάθηση χρησιμοποιεί μοτίβα για την πρόβλεψη τιμών.

Για παράδειγμα, με την εποπτευόμενη μάθηση ένας αλγόριθμος μπορεί να τροφοδοτεί δεδομένα με εικόνες καρχαριών που φέρουν την ετικέτα «ψάρια» και εικόνες των ωκεανών που φέρουν την ένδειξη «νερό». Εκπαιδευόμενος σε αυτά τα δεδομένα, ο εποπτευόμενος αλγόριθμος μάθησης θα πρέπει να είναι σε θέση να αναγνωρίσει αργότερα τις εικόνες των καρχαριών χωρίς ετικέτα ως «ψάρια» και τις εικόνες των ωκεανών χωρίς νερό ως «νερό».

Μία κοινή περίπτωση εφαρμογής της εποπτευόμενης μάθησης είναι η χρήση ιστορικών δεδομένων για την πρόβλεψη στατιστικά πιθανών μελλοντικών γεγονότων. Για παράδειγμα, να μπορεί να χρησιμοποιήσει ιστορικές πληροφορίες για το χρηματιστήριο για να προβλέψει επερχόμενες διακυμάνσεις ή να χρησιμοποιηθεί για να φιλτράρει μηνύματα spam. Στην εποπτευόμενη μάθηση, οι φωτογραφίες των σκύλων με ετικέτα μπορούν να χρησιμοποιηθούν ως δεδομένα εισόδου για την ταξινόμηση των μη επισημασμένων φωτογραφιών σκύλων.

Στη μη εποπτευόμενη μάθηση, τα δεδομένα δεν έχουν επιθυμητή έξοδο, οπότε ο αλγόριθμος μάθησης αφήνεται να βρει κοινά στοιχεία μεταξύ των δεδομένων εισόδου του. Ο στόχος της μη εποπτευόμενης μάθησης μπορεί να είναι τόσο απλός όσο η ανακάλυψη κρυφών μοτίβων σε ένα σύνολο δεδομένων. Όμως μπορεί, επίσης, να έχει ως στόχο τη μάθηση χαρακτηριστικών, η οποία επιτρέπει στην υπολογιστική μηχανή να ανακαλύψει αυτόματα τις αναπαραστάσεις που απαιτούνται για την ταξινόμηση των ακατέργαστων δεδομένων.

Χωρίς να αναφέρουν μια «σωστή» απάντηση, οι μέθοδοι της μη εποπτευόμενης μάθησης μπορούν να εξετάσουν πολύπλοκα δεδομένα που είναι πιο μεγάλα και φαινομενικά άσχετα, προκειμένου να τα οργανώσουν με δυνητικά ουσιαστικούς τρόπους. Η μη εποπτευόμενη μάθηση χρησιμοποιείται συχνά για τον εντοπισμό ανωμαλιών, συμπεριλαμβανομένων δόλιων αγορών πιστωτικών καρτών και συστημάτων που προτείνουν ποια προϊόντα θα αγοραστούν στη συνέχεια, με βάση τα αναδυόμενα μοτίβα καταναλωτικής συμπεριφοράς.

2.7 Επικύρωση μοντέλου

Σε κάθε περίπτωση, είναι αναγκαία η επικύρωση του μοντέλου έτσι ώστε να διασφαλίσουμε την αποτελεσματικότερη λήψη αποφάσεων βάσει του μοντέλου. Ουσιαστικά, για να χαρακτηρισθεί το μοντέλο ακριβές θα πρέπει να μπορεί να ταξινομεί ορθά τις παρατηρήσεις τις οποίες λαμβάνει. Υπάρχει ωστόσο το ενδεχόμενο να προκύψουν άγνωστες παρατηρήσεις στις οποίες το μοντέλο δεν έχει προηγουμένως εκπαιδευθεί. Για να επιτευχθεί αυτό θα πρέπει να έχουν αναπτυχθεί γενικευμένοι κανόνες.

Εξάλλου, εκεί έγκειται και η πραγματική αξία του μοντέλου, δηλαδή στη δυνατότητα του να προβλέπει άγνωστες παρατηρήσεις. Ένα πρόβλημα το οποίο ανακύπτει σε αυτό το σημείο είναι αυτό της ***υπερπροσαρμογής (overfitting)***. Όταν το μοντέλο είναι πολύπλοκο, τότε εστιάζει στην κατανόηση των επιμέρους σχέσεων και όχι στην ανάπτυξη γενικευμένων. Ως εκ τούτου, αν και ανταποκρίνεται με επιτυχία στις παρατηρήσεις που αποτελούν το σύνολο εκπαίδευσης, εντούτοις αδυνατεί όταν πρόκειται για άγνωστες παρατηρήσεις.

Εν γένει, μόνο εάν το μοντέλο είναι ακριβές μπορεί να υιοθετηθεί για την λήψη αποφάσεων σε πραγματικά προβλήματα. Άλλωστε, η ακρίβεια αποτελεί το βασικότερο κριτήριο για τη σύγκριση μεταξύ διαφορετικών μοντέλων.

Για την αξιολόγηση της ακρίβειας ενός μοντέλου όσον αφορά τη διαχείριση άγνωστων παρατηρήσεων, έχουν αναπτυχθεί κατά καιρούς διάφορες μέθοδοι, οι σημαντικότερες εκ των οποίων παρουσιάζονται στη συνέχεια [58].

Μέθοδος holdout

Στα πλαίσια της εν λόγω μεθόδου, τα δεδομένα τα οποία έχουμε στη διάθεσή μας ταξινομούνται σε δύο επιμέρους σύνολα. Το ένα εξ αυτών χρησιμοποιείται για την εκπαίδευση του μοντέλου, ενώ το δεύτερο για την επικύρωσή του. Συνήθως, τηρείται μια αναλογία 2 προς 1 όσον αφορά την κατανομή των παρατηρήσεων μεταξύ των συνόλων. Η ακρίβεια του μοντέλου προκύπτει από το ποσοστό των ορθών προβλέψεων. Μάλιστα, για τα καλύτερα δυνατά αποτελέσματα υπάρχει η δυνατότητα επανάληψης της μεθόδου holdout μεταβάλλοντας κάθε φορά την κατανομή των παρατηρήσεων μεταξύ των συνόλων.

Διασταυρούμενη επικύρωση δέκα τμημάτων

Για τις ανάγκες της εν λόγω μεθόδου, οι παρατηρήσεις κατανέμονται σε δέκα υποσύνολα με τυχαίο τρόπο. Τα εννέα εξ αυτών χρησιμοποιούνται για την εκπαίδευση του μοντέλου και το τελευταίο για την επικύρωσή του. Εναλλάσσοντας κάθε φορά το σύνολο επικύρωσης, η διαδικασία επαναλαμβάνεται δέκα φορές και εν τέλει εκτιμάται η επίδοση του μοντέλου. Εναλλακτικά είναι δυνατό να χρησιμοποιηθεί διαφορετικός αριθμός τμημάτων n και ανάλογα μπορούμε να αναπροσαρμόσουμε τον αριθμό των επαναλήψεων. Επίσης, υπάρχει η στρωματοποιημένη επικύρωση, κατά την οποία τα επιμέρους υποσύνολα συντίθεται από ίσο αριθμό παρατηρήσεων για κάθε κλάση.

Μέθοδος «άφησε ένα έξω»

Παραλλαγή της διασταυρούμενης επικύρωσης n τμημάτων αποτελεί η εν λόγω μέθοδος, στα πλαίσια της οποίας επιλέγεται μια παρατήρηση για την επικύρωση και όλες οι υπόλοιπες χρησιμοποιούνται για την εκπαίδευση. Η διαδικασία επαναλαμβάνεται για όλες τις παρατηρήσεις, μία προς μία.

Μέθοδος bootstrap

Κατ' αντιστοιχία με τα παραπάνω, στα πλαίσια της εν προκειμένω μεθόδου δημιουργούνται επίσης πολλαπλά σύνολα επικύρωσης. Ωστόσο, η διαφοροποίηση εδράζεται στο γεγονός πως οι παρατηρήσεις επικύρωσης επανατοποθετούνται στο δείγμα.

Όσον αφορά την αξιοπιστία των μεθόδων, η καταλληλότερη όλων φαίνεται να είναι η μέθοδος της διασταυρούμενης επικύρωσης δέκα τμημάτων σύμφωνα με σχετική έρευνα του Kohavi [30].

2.8 Επιδόσεις ανά κλάση

Στο σημείο αυτό κρίνεται σκόπιμη η αποσαφήνιση ορισμένων όρων στην περίπτωση δυαδικής κλάσης. Οι σχετικοί ορισμοί δίνονται παρακάτω.[57],[58]

Θετικές Παρατηρήσεις	Οι παρατηρήσεις, οι οποίες ανήκουν σε μια τιμή της κλάσης (απουσία καρδιακής νόσου).
Αρνητικές Παρατηρήσεις	Ονομάζονται οι παρατηρήσεις, οι οποίες ανήκουν στην άλλη τιμή της κλάσης (εμφάνιση καρδιακής νόσου)
Αληθινές Θετικές Προβλέψεις	Το πλήθος των επιτυχών προβλέψεων για θετικές παρατηρήσεις (π.χ. ο εξεταζόμενος πάσχει από την αθηρωματική νόσο και ο κατηγοριοποιητής προβλέπει σωστά την κλάση).
Αληθινές Αρνητικές Προβλέψεις	Το πλήθος των επιτυχημένων προβλέψεων για αρνητικές παρατηρήσεις (π.χ. ο εξεταζομένος δεν έχει αθηρωματική νόσο και ο κατηγοριοιποιητής προβλέπει σωστά την κλάση).
Ψευδείς Θετικές Προβλέψεις	Είναι το πλήθος των αποτυχημένων προβλέψεων για αρνητικές παρατηρήσεις (ο ασθενής δεν παρουσιάζει αθηρωματική νόσο, ο κατηγοριοποιητής όμως προβλέπει ότι την έχει).
Ψευδείς Αρνητικές Προβλέψεις	Είναι το πλήθος των αποτυχημένων προβλέψεων για θετικές παρατηρήσεις (ο ασθενής δεν έχει καρδιακή νόσο, ο κατηγοριοποιητής όμως προβλέπει ότι δεν έχει).

Όσον αφορά την αποτύπωση των επιδόσεων, ως πλέον ενδεδειγμένη λύση προβάλλει ο πίνακας σύγχυσης. Ουσιαστικά πρόκειται για έναν πίνακα δύο διαστάσεων στον οποίο στήλες αντιπροσωπεύουν τις προγνώσεις και οι γραμμές τις πραγματικές τιμές. Ο Πίνακας που ακολουθεί αποτελεί ένα παράδειγμα πίνακα σύγχυσης, όπου αποτυπώνονται οι ακόλουθες τιμές:

	Πρόβλεψη Αρνητικής Κλάσης	**Πρόβλεψη Θετικής Κλάσης**
Πραγματική Αρνητική Κλάση	**tn**	**fp**
Πραγματική Θετική Κλάση	**fn**	**tp**

Αληθινές αρνητικές – tn

Αληθινές θετικές – tp

Ψευδείς αρνητικές – fn

Ψευδείς θετικές - fp

Ορίζονται ως εξής :

$$sensitivity = \frac{tp}{pos}$$

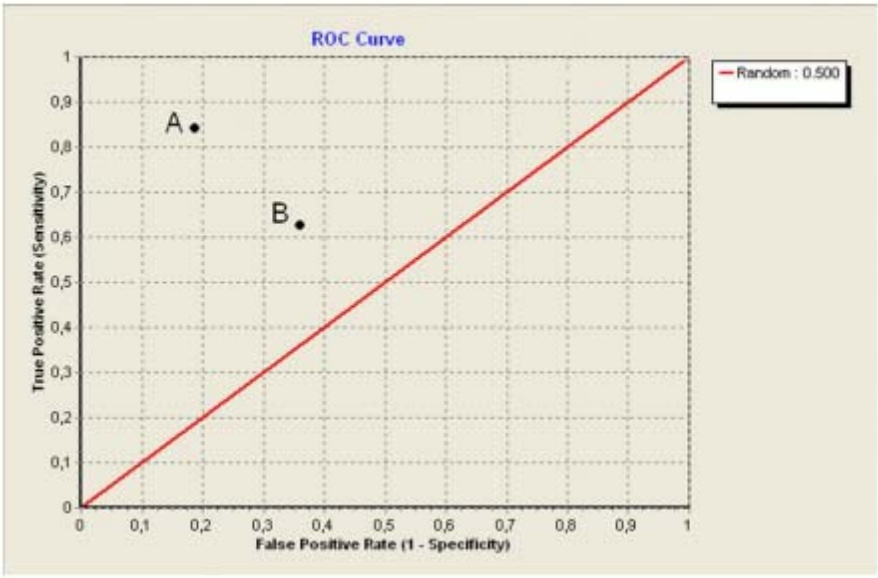
$$specificity = \frac{tn}{negat}$$

$$precision = \frac{tp}{tp + fp}$$

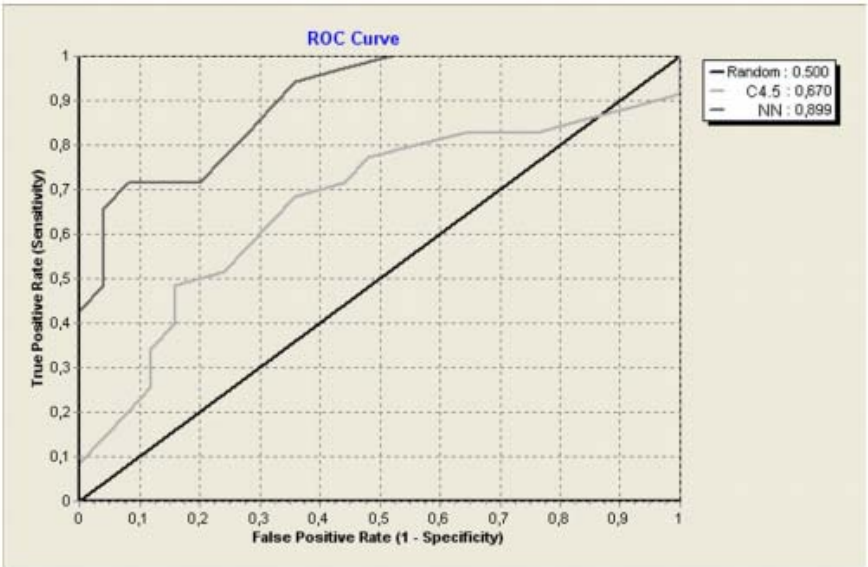
$$accuracy = \frac{tp + tn}{pos + negat}$$

Για τη μέτρηση της ακρίβειας είναι επίσης η δυνατή η χρήση των καμπυλών ROC (Receiver Operating Characteristics). Στο σύστημα συντεταγμένων, ο οριζόντιος άξονας αντιπροσωπεύει το στοιχείο specificity και ο κάθετος άξονας το στοιχείο sensitivity. Άρα λοιπόν, ο οριζόντιος άξονας δείχνει τις αρνητικές παρατηρήσεις οι οποίες ταξινομήθηκαν με λανθασμένο τρόπο και ο κάθετος τις θετικές οι οποίες ταξινομήθηκαν με σωστό τρόπο.

Στα παρακάτω Σχήματα δίνονται δύο παραδείγματα καμπυλών ROC. Στο σημείο τομής των αξόνων βρίσκεται ένας κατηγοριοποιητής ο οποίος δεν έχει τη δυνατότητα να προγνώσει ποτέ θετική παρατήρηση. Αντίθετα, στο σημείο (1,1) βρίσκεται ένας κατηγοριοποιητής ο οποίος πάντοτε προβλέπει θετική παρατήρηση. Εν γένει οι κατηγοριοποιητές οι οποίοι βρίσκονται στο άνω αριστερό τμήμα του διαγράμματος παρουσιάζουν τις καλύτερες επιδόσεις.



Εικόνα 2 Σημεία στο χώρο με καμπύλες ROC



Εικόνα 3 Σύγκριση κατηγοριοποιητών

2.9 Ανάλυση του πλαισίου δεδομένων

Η λογική της επιλογής των προσδιοριστών του πλαισίου δεδομένων statlog βασίζεται σε μεγάλο βαθμό στις μελέτες των Diamond και Forrester το 1979. Σύμφωνα με αυτές τις μελέτες [22], μετά από έλεγχο δεδομένων 23.996 ασθενών, ορίστηκε η προ του τεστ πιθανότητα (pretest probability) για στεφανιαία νόσο με βάση το φύλο και την ηλικία. Στη συνεχεία, οι Diamond και Forrester, λαμβάνοντας υπόψη αγγειογραφικά δεδομένα 4.952 ασθενών και διαχωρίζοντας τη θωρακαλγία σε τυπική στηθάγχη, άτυπη στηθάγχη και μη-στηθαγχικό άλγος, υπολόγισαν την πιθανότητα για στεφανιαία νόσο με βάση την ηλικία, το φύλο και τα χαρακτηριστικά του πόνου, ενώ σχετικά πρόσφατη αναθεώρηση από τους Βρετανούς πρόσθεσε και κλασικούς παράγοντες καρδιαγγειακού κινδύνου. Ο διαχωρισμός της στηθάγχης σε τυπική, άτυπη και μη σχετιζόμενη με στηθάγχη επικυρώθηκε και από μελέτη 20.391 ασθενών με βάση τα αγγειογραφικά τους ευρήματα [23].

Έτσι, λοιπόν, για τη μεταβλητή **CHESTPAIN**, έχουμε τέσσερις περιπτώσεις που αντιστοιχούν: 1 = τυπική στηθάγχη, 2 = άτυπη στηθάγχη, 3 = μη-στηθαγχικό άλγος και 4 = το ασυμπτωματικό θωρακικό άλγος. Για τη μεταβλητή **SEX**, οι δυνατές τιμές είναι 1= αρσενικό και 0 = θηλυκό. Για τη μεταβλητή **AGE** που αφορά την ηλικία, οι τιμές κυμαίνονται από 35 έως 70 χρονών, ακριβώς έτσι όπως περιγράφονται στην έρευνα των Diamond και Forrester. Άλλη μια μεταβλητή είναι η συστολική πίεση του αίματος **RESTBP**. Η ιδανική αρτηριακή πίεση σε υγιείς ενήλικες είναι κάτω από 120 για την συστολική και κάτω από 80 για τη διαστολική. Οποιαδήποτε τιμή αρτηριακής πίεσης άνω του 140 για τη συστολική και άνω του 90 για τη διαστολική θεωρείται υπέρταση. Οι μονάδες μέτρησης της πίεσης είναι τα χιλιοστά της στήλης υδραργύρου (mmHg). Σε αυτή την παράμετρο, οι τιμές κυμαίνονται από 94 μέχρι 200 mmHg.

Η ολική χοληστερόλη αναφέρεται στο πλαίσιο δεδομένων ως **CHOL**. Η χοληστερόλη είναι μία λιπαρή ουσία, η οποία αποτελεί βασικό συστατικό των λιποπρωτεϊνικών συμπλόκων που σχηματίζουν τις κυτταρικές μεμβράνες. Είναι αναγκαία για τον οργανισμό, αφού αποτελεί πρόδρομη ουσία μιας σειράς ορμονών. Εάν τα επίπεδα χοληστερόλης είναι υψηλά, διατρέχουμε μεγαλύτερο κίνδυνο εμφάνισης στεφανιαίας νόσου. Τα επίπεδα της χοληστερόλης είναι <200 mg/dL επιθυμητή, 200 – 239 mg/dL οριακά υψηλή και >240 mg/dL αρκετά υψηλή [24]. Το πεδίο τιμών αυτής της παραμέτρου παίρνει τιμές 126 μέχρι 564 mg/dL.

Το σάκχαρο νηστείας συμβολίζεται με **SUGAR**. Οι φυσιολογικές τιμές στο σάκχαρο νηστείας κυμαίνονται από 60 mg/dl μέχρι 120 mg/dl. Αυξημένη τιμή θεωρείται πάνω από 125 mg/dl, σύμφωνα με την Αμερικανική Διαβητολογική Εταιρεία, ή πάνω από 140 mg/dl, σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας. Η εμφάνιση των αθηροσκληρωτικών εκδηλώσεων που χαρακτηρίζουν τα καρδιαγγειακά νοσήματα τείνει να είναι συχνότερη και πρωιμότερη και η εξέλιξή της ταχύτερη στα διαβητικά σε σύγκριση με τα μη διαβητικά άτομα [25]. Επομένως για τιμές >120 mg/dl έχουμε λογικό 1, ενώ για τιμές <120 mg/dl έχουμε λογικό 0.

Στη μεταβλητή που περιγράφει το ηλεκτροκαρδιογράφημα ηρεμίας **ECG**, έχουμε τρεις τιμές. Το ηλεκτροκαρδιογράφημα είναι η απεικόνιση της ηλεκτρικής δραστηριότητας της καρδιάς. Η συστολή του μυοκαρδίου είναι αποτέλεσμα των ηλεκτρικών δυναμικών που δημιουργούνται στις καρδιακές κοιλότητες. Στο stalog πλαίσιο δεδομένων, η τιμή 0 αντιστοιχεί στο φυσιολογικό ηλεκτροκαρδιογράφημα, ενώ η τιμή 1 αντιστοιχεί σε κάποια ανωμαλία στο συμπλέγματος ST. Η τιμή 2 αντιπροσωπεύει την ανάσπαση του ST από την ισοηλεκτρική γραμμή μεγαλύτερης από 1 σε δύο τουλάχιστον απαγωγές αντιστοιχεί με οξύ έμφραγμα του μυοκαρδίου.

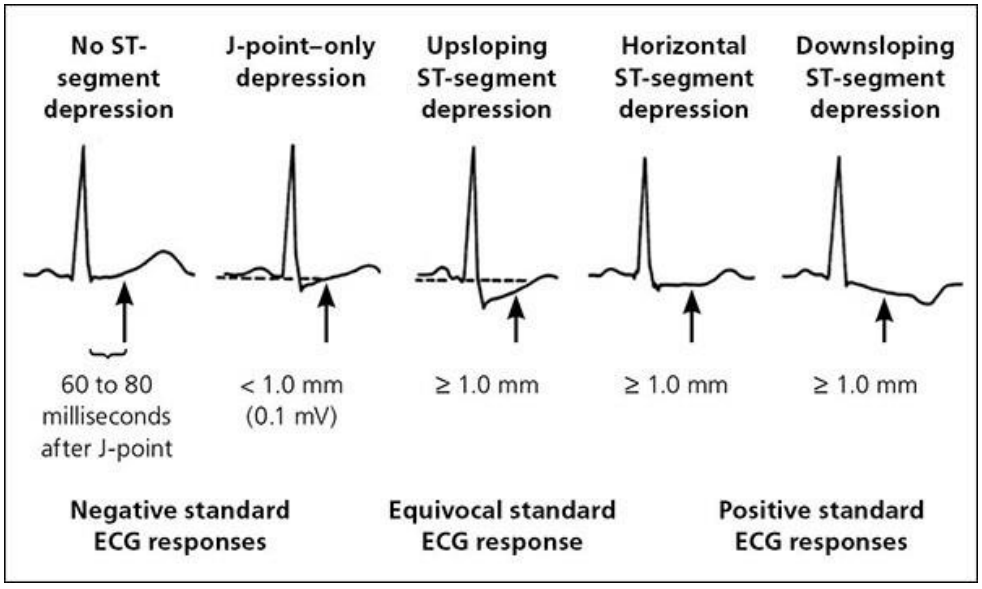
Η παράμετρος **MAXHR** εκφράζει την μέγιστη καρδιακή συχνότητα. Η αυξημένη καρδιακή συχνότητα επηρεάζει δραστικά την στεφανιαία ροή, καθώς βραχύνει την διαστολική περίοδο. Η επίδραση της αυξημένης καρδιακής συχνότητας στην στεφανιαία ροή αποκτά ιδιαίτερη βαρύτητα σε περίπτωση στενώσεων των επικαρδίων αγγείων [26]. Το πεδίο τιμών της MAXHR είναι από 71 μέχρι 202 BPM.

Η σταθερή στηθάγχη **ANGINA** είναι συνήθως προβλέψιμη, και γενικά αναπτύσσεται όταν η καρδιά πρέπει να εργαστεί σκληρότερα, όπως κατά τη διάρκεια της άσκησης ή το ανέβασμα σκάλας. Η σταθερή στηθάγχη εκλύεται με την άσκηση και αποτελεί τη συνήθη εκδήλωση της ανεπαρκούς αιμάτωσης του μυοκαρδίου και εκδηλώνεται με δυσφορία στο κέντρο του θώρακα. Είτε εκδηλώνεται (λογικό 1), είτε δεν εκδηλώνεται (λογικό 0) στο υπο-μελέτη πλαίσιο δεδομένων.

Η **DEP** συμβολίζει την κατάσπαση ST ενός ηλεκτροκαρδιογραφικού σήματος στη δοκιμασία κόπωσης. Μια περίπτωση, η δοκιμασία κόπωσης να θεωρείται παθολογική ή θετική είναι η εμφάνιση κατασπάσεων ST ισχαιμικού τύπου σε μικρό σωματικό έργο ή σε σχετικά χαμηλή καρδιακή συχνότητα. Η αδυναμία επίτευξης του 85% της προβλεπομένης για την ηλικία καρδιακής συχνότητας και γενικά η αδυναμία επαρκούς ανόδου της καρδιακής συχνότητας με την άσκηση (ανεπαρκής χρονότροπη απάντηση) θεωρείται παθολογική ένδειξη.

Σε μία μελέτη διαπιστώθηκε, ότι μεταξύ των ασθενών που είχαν κατάσπαση ST ισχαιμικού τύπου βάθους 2 mm ή περισσότερο, εκείνοι που πέτυχαν διάρκεια σωματικής άσκησης < από 6 λεπτά, είχαν σαφώς αυξημένο κίνδυνο εμφράγματος ή θανάτου [27]. Αντίθετα, οι ασθενείς με την ίδια κατάσπαση ST που πέτυχαν διάρκεια άσκησης > από 9 λεπτά είχαν σαφώς μικρότερα ποσοστά κινδύνου. Το πεδίο τιμών της DEP είναι από 0.0 έως 6.2 mm.

Η **EXERCISE** συμβολίζει την κλίση του τμήματος ST του ηλεκτροκαρδιογραφικού σήματος. Σύμφωνα με τη βιβλιογραφία [28], τα ηλεκτροκαρδιογραφικά κριτήρια θετικής για ισχαιμία (δηλ παθολογικής) δοκιμασίας κοπώσεως είναι η εμφάνιση, κατά την κόπωση ή και μετά την κόπωση κατάσπασης, ST με οριζόντια ή κατιούσα φορά τουλάχιστον 1mm υπό την ισοηλεκτρική γραμμή, 60-80 msec μετά το σημείο J, κατάσπασης ST με βραδέως ανιούσα φορά (ελαφρώς ανιούσα) τουλάχιστον 1,5 mm υπό την ισοηλεκτρική γραμμή 60-80 msec μετά το σημείο J, ή ανάσπασης ST τουλάχιστον 1mm πάνω από το ύψος της ισοηλεκτρικής γραμμής σε απαγωγή που δεν έχει κύμα Q προϋπάρχοντος εμφράγματος. Ως ισοηλεκτρική γραμμή θεωρείται η οριζόντια γραμμή που διέρχεται από το σημείο ένωσης του διαστήματος PR (ή PQ) με το QRS και σημείο J είναι το σημείο σύνδεσης του συμπλέγματος QRS με το διάστημα ST. Συνεπώς, για τη EXERCISE μεταβλητή έχουμε τρεις τιμές : (1= κατάσπαση ST με ανιούσα φορά, 2= κατάσπαση ST με οριζόντια φορά, 3= κατάσπαση ST με κατιούσα φορά).



Εικόνα 4 Ευρήματα ηλεκτροκαρδιογραφήματος κατα την κόπωση με διαγνωστικό ενδιαφέρον

Η **FLUOR** εκφράζει τον αριθμό της επεμβατικής εκλεκτικής σκιαγράφησης των στεφανιαίων αρτηριών κατά τις εξετάσεις της Στεφανιαιογραφίας. Πρόκειται για μία ελάχιστα επεμβατική τεχνική ακτινοσκιερής απεικόνισης των στεφανιαίων αρτηριών μετά από έγχυση ακτινοσκιερού μέσου, μέσω ειδικών καθετήρων. Η διαγνωστική αξία της αξονικής στεφανιογραφίας έγκειται κυρίως στον αποκλεισμό της ύπαρξης στεφανιαίας νόσου σε ασθενείς με χαμηλή ή ενδιάμεση πιθανότητα. Το πεδίο τιμών της FLUOR είναι από 0 έως 3 .

Η **THAL** πιθανότατα συμβολίζει την εξέταση Σπινθηρογραφήματος καρδιάς με ραδιενεργό θάλλιο. Για αυτή την παράμετρο, η περιγραφή που βρίσκεται στο αποθετήριο του UCI δεν είναι ακριβής. Σε αυτή την εξέταση, ο στόχος είναι να μάθουμε εάν κάποιες περιοχές του μυοκαρδίου δεν αιματώνονται αρκετά στην άσκηση. Αν η αρτηρία έχει μερική στένωση, το θάλλιο προσλαμβάνεται πλήρως ή σχεδόν πλήρως από το μειωμένης πρόσληψης τοίχωμα (αναστρέψιμη ισχαιμία). Αυτή η κατάσταση αντιστοιχεί στο νούμερο 7 του πλαισίου δεδομένων (7 = reversible defect). Αντίθετα, αν έπειτα από τρεις με τέσσερις ώρες το τοίχωμα δεν προσλάβει καθόλου φάρμακο, σημαίνει ότι η περιοχή έχει νεκρωθεί και έχει δημιουργηθεί ουλή (σταθερό έλλειμμα).

Η περίπτωση αυτή αντιπροσωπεύεται με το νούμερο 6 στο πλαίσιο δεδομένων (6 = fixed defect). Τέλος, οι πλήρως αποφραγμένες αρτηρίες εμφανίζονται σαν «ψυχρά σημεία» και αντιστοιχούν στο σημείο 3 (3= normal). Η μεταβλητή **OUTPUT** είναι η έξοδος. Αν υπάρχει αθηρωματική νόσος, η έξοδος παίρνει την τιμή 1 (presence), ενώ αν δεν υπάρχει παίρνει την τιμή 0 (absence).

Πίνακας 5 Συγκεντρωτική περιγραφή μεταβλητών πλαισίου δεδομένων statlog

Μεταβλητές					Παρουσία (Θετικό)	Απουσία (Αρνητικό)
No.	**Σύμβολο**	**Περιγραφή**	**Τύπος**	**Πεδίο τιμών**	**Μέσος ± STD**	**Διάμεσος ±**
1	AGE	Ηλικία σε χρόνια	Αριθμητική	[29,77]	56.76 ± 7.9	52.64 ±9.55
2	SEX	Φύλο	Δυαδική	0 = Θηλυκό, 1 = Αρσενικό	-	-
3	CHESTPAIN	Θωρακικό άλγος	Ονομαστική	1 = τυπική στηθάγχη, 2 = άτυπη στηθάγχη, 3=μη-στηθαγχικός πόνος, 4 =ασυμπτωματικός	-	-
4	RESTBP	Αρτηριακή πίεση σε mmHg	Αριθμητική	[94,200]	134.64 ±18.9	129.18
5	CHOL	Χοληστερόλη στον ορό σε mg/dl	Αριθμητική	[126,564]	251.85 ±49.68	243.49
6	SUGAR	Σάκχαρο νηστείας >120 mg/dl	Δυαδική	0 = Ψευδές, 1 = Αληθές	-	-
7	ECG	Ηλεκτροκαρδιογράφημα	Ονομαστική	Η = ανωμαλία στο συμπλέγματος ST, Η = ανάσπαση του ST από την ισοηλεκτρική γραμμή > 1 σε δύο τουλάχιστον απαγωγές	-	-
8	MAXHR	Μέγιστος καρδιακός παλμός	Αριθμητική	[71,202]	139.11 ±22.71	158.58 ±
9	ANGINA	Σταθερή στηθάγχη	Δυαδική	0 = δεν εκδηλώνεται, 1 = εκδηλώνεται	-	-
10	DEP	Εύρημα ΗΚΓ σε δοκιμασία κόπωσης	Αριθμητική	[0, 6.2]	1.59 ±1.31	0.6 ±0.79
11	EXERCISE	Κλίση του τμήματος ST του ΗΚΓ σήματος	Ονομαστική	1 = ST με ανιούσα φορά, 2 = ST με οριζόντια φορά, 3 = ST με κατιούσα φορά	-	-
12	FLUOR	Σκιαγράφηση των στεφανιαίων αρτηριών	Ονομαστική	0-3	-	-
13	THAL	Σπινθηρογράφημα	Ονομαστική	3 = ψυχρά σημεία, 6 = σταθερό έλλειμμα, 7 = αναστρέψιμη ισχαιμία	-	-

2.10 Λογιστική Παλινδρόμηση

2.10.1 Εισαγωγικά στοιχεία

Η παλινδρόμηση δεν αποτελεί μια καινούρια έννοια στη διεθνή βιβλιογραφία, καθώς η εμφάνισή της τοποθετείται στα τέλη του 19ου αιώνα όταν ο Galton [31] επιχείρησε να καταλήξει σε μια σχέση μεταξύ του ύψους μιας ομάδας παιδιών και των γονέων τους. Τα συμπεράσματα της εν λόγω έρευνας ήταν τα ακόλουθα:

- ✓ Οι γονείς με υψηλό ανάστημα γεννούν παιδιά με υψηλό ανάστημα, ωστόσο ο μέσος όρος του αναστήματος των παιδιών αυτών είναι χαμηλότερος από το μέσο όρο του αναστήματος των γονέων τους.
- ✓ Οι γονείς με χαμηλό ανάστημα γεννούν παιδιά με χαμηλό ανάστημα, ωστόσο ο μέσος όρος του αναστήματος των παιδιών αυτών είναι μεγαλύτερος από το μέσο όρο του αναστήματος των γονέων τους.

Η εν προκειμένω «αναντιστοιχία» προσδιορίστηκε ως «*παλινδρόμηση προς τη μετριότητα*». Στις μέρες μας, ο εν λόγω όρος χρησιμοποιείται για να περιγράψει τη σχέση μιας εξαρτημένης μεταβλητής με ένα σύνολο ανεξάρτητων μεταβλητών.

2.10.2 Γραμμική παλινδρόμηση

Η σχέση μεταξύ των μεταβλητών περιγράφεται από μια ευθεία γραμμή, η οποία καλείται γραμμή παλινδρόμησης και αντιπροσωπεύει μια εκτίμηση για τις μέσες τιμές του άξονα των τεταγμένων σε σχέση με τις τιμές του άξονα των τετμημένων. Στην περίπτωση αυτή κάνουμε λόγο για απλή γραμμική παλινδρόμηση, σύμφωνα με τον ακόλουθο τύπο:

$$Y = A_0 + A_1 X_1 + \varepsilon$$

Όπου,

Y: η εξαρτημένη μεταβλητή

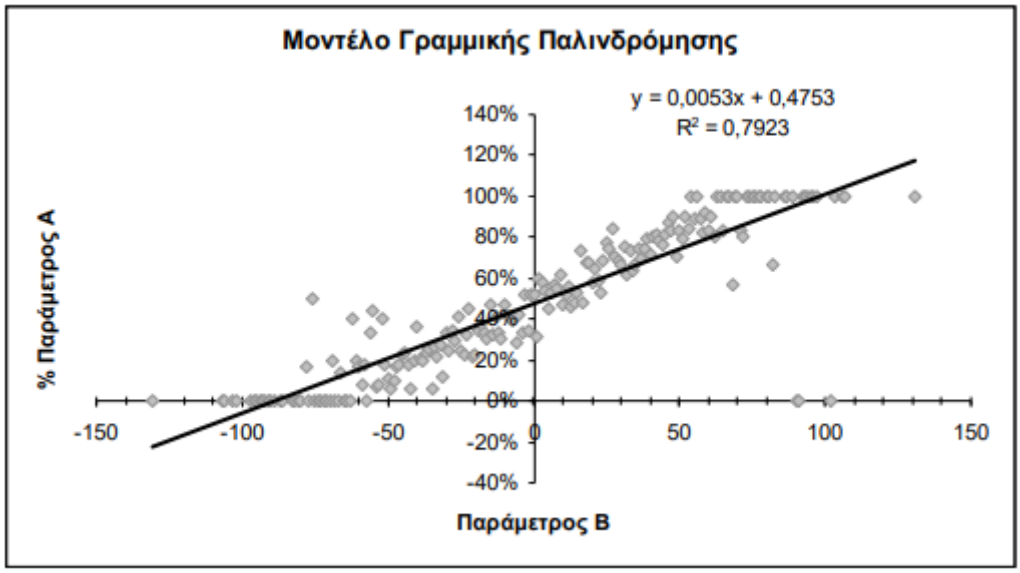
X_1: η ανεξάρτητη μεταβλητή

A_o: η σταθερά της απλή γραμμικής παλινδρόμησης

A_1: η κλίση της ευθείας

ε: το τυχαίο σφάλμα

Ένα παράδειγμα απλής γραμμικής παλινδρόμησης δίνεται στο Σχήμα που ακολουθεί:



Εικόνα 5 Γραμμική παλινδρόμηση

2.10.3 Λογιστική παλινδρόμηση

Σε ένα μοντέλο απλής γραμμικής παλινδρόμησης, η εξαρτημένη μεταβλητή Υ λαμβάνει τιμές από το –οο έως το +οο. Όταν πρόκειται για τη χρήση μιας συχνότητας ως εξαρτημένη μεταβλητή και όχι ενός ποσοστού, τότε μπορούμε να πάμε σε ένα εκθετικό μοντέλο, στο οποίο οι τιμές της εξαρτημένης μεταβλητής κυμαίνονται μεταξύ 0 και +οο. Ένας συνήθης μετασχηματισμός είναι ο ακόλουθος:

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

Ως πεδίο ορισμού της παραπάνω συνάρτησης δίνεται το [0,1]. Στην περίπτωση που το x ισούται με το μηδέν, η λογιστική συνάρτηση λαμβάνει την τιμή 0.5. Αντίθετα, όταν το x τείνει στο +οο η συνάρτηση λαμβάνει τιμή ίση με τη μονάδα. Τέλος, όταν το x τείνει στο –οο η συνάρτηση τείνει στο μηδέν. Σε κάθε περίπτωση, οι τιμές της συνάρτησης βρίσκονται αυστηρά μεταξύ του [0,1], γεγονός το οποίο επιτρέπει το χαρακτηρισμό ως πιθανότητα.

2.10.4 Γενικευμένα γραμμικά μοντέλα

Η λογιστική παλινδρόμηση συμπεριλαμβάνεται στην ευρύτερη ομάδα των γενικευμένων γραμμικών μοντέλων, τα οποία συνδυάζουν τα εξής τρία χαρακτηριστικά:

- ✓ Οι μεταβλητές εισόδου συνδυάζονται γραμμικά
- ✓ Οι έξοδοι ακολουθούν εκθετική κατανομή πιθανότητας (πχ διωνυμική κατανομή ή κατανομή Poisson)
- ✓ Η μέση τιμή της εξόδου προκύπτει ως γραμμικός συνδυασμός των μεταβλητών εισόδου και συνδέεται με αυτές μέσω μιας συνάρτησης που ονομάζεται link function

Οι είσοδοι εκφράζονται από την ακόλουθη σχέση:

$$x = \beta_0 + \beta_1 \cdot \mathrm{X}_1$$

Ακολούθως, η λογιστική παλινδρόμηση εκφράζεται ως εξής:

$$P(Y = 1 \mid X) = \frac{e^{\beta_0 + \beta_1 X_1}}{e^{\beta_0 + \beta_1 X_1} + 1}$$

Με την παραπάνω σχέση, ουσιαστικά υπολογίζουμε την πιθανότητα η έξοδος του μοντέλου να αποδίδει ως τιμή τη μονάδα, σε σχέση με την αντίστοιχη μεταβλητή εισόδου. Το επόμενο βήμα είναι η εκτίμηση της συνάρτησης link η οποία προαναφέρθηκε. Κάνοντας τους κατάλληλους μετασχηματισμούς καταλήγουμε στην παρακάτω σχέση:

$$\ln\left(\frac{P(Y=1\mid X)}{1-P(Y=1\mid X)}\right)=\beta_0+\beta_1 X_1$$

Το αριστερό μέρος της παραπάνω εξίσωσης ονομάζεται **logit**, ενώ το δεξί παρατηρούμε πως προσομοιάζει στη γραμμική παλινδρόμηση. Εν γένει, η λογιστική παλινδρόμηση βρίσκει εφαρμογή σε περιπτώσεις όπου αναζητούμε τη σχέση η οποία ισχύει μεταξύ ενός συνόλου ανεξάρτητων μεταβλητών και μιας δυαδικής μεταβλητής η οποία μπορεί να λάβει μια εκ δύο τιμών.

Ο υπολογισμός των συντελεστών της λογιστικής παλινδρόμησης προκύπτει μέσω του υπολογισμού της **μέγιστης πιθανοφάνειας**

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

Αντίστοιχα, η τιμή κάθε παρατήρησης εκτιμάται ως εξής:

$$l=\frac{1}{n}\log_e L$$

Η συνάρτηση της **πιθανοφάνειας** όσον αφορά την έκβαση ενός συγκεκριμένου γεγονότος εκτιμά το κατά πόσο το υπό μελέτη δείγμα περιγράφεται από συγκεκριμένα μέτρα όπως ο μέσος όρος ή η τυπική απόκλιση. Η εν λόγω μέθοδος ενδείκνυται για μεγάλα δείγματα. Εξάλλου, το μέγεθος του δείγματος επιδρά σημαντικά στην αποτελεσματικότητα της λογιστικής παλινδρόμησης

2.10.5 Μετρήσεις καλής προσαρμογής

Σε κάθε περίπτωση, πριν την υιοθέτησή του, το εκάστοτε μοντέλο με *n* ανεξάρτητες μεταβλητές θα πρέπει να αξιολογηθεί όσον αφορά την αξιοπιστία των αποτελεσμάτων του. Τα κριτήρια τα οποία χρησιμοποιούνται προς αυτήν την κατεύθυνση είναι τα εξής:

1. Pearson x^2

Ο έλεγχος x^2 στηρίζεται στον υπολογισμό των υπολειμμάτων, σύμφωνα με τον ακόλουθο τύπο:

$$x^2 = \sum_j r_j^2$$

$$r_j = \frac{y_i - m_j p_j}{\sqrt{m_j} p(1-p_j)}$$

Όπου,

y_j: ο αριθμός των επιτυχημένων εκβάσεων για τη μεταβλητή j

m_j: αριθμός των προσπαθειών ή επαναληπτικών μετρήσεων για την j μεταβλητή εκτιμώμενη (προσαρμοσμένη) πιθανότητα για την j μεταβλητή

r_j: τυποποιημένο υπόλειμμα του Pearson

2. Κριτήριο απόκλισης *D*

Το συγκεκριμένο κριτήριο διερευνά την προσαρμοστικότητα του μοντέλου στα χαρακτηριστικά της κατά περίπτωσης μελέτης. Ο υπολογισμός γίνεται ως εξής:

$$D(y, \hat{\mu}) = 2 \cdot \left(\log(p(y \mid \hat{\vartheta}_s)) - \log(p(y \mid \hat{\vartheta}_0)) \right)$$

Όπου το y είναι η έξοδος του μοντέλου

$\hat{\mu}$ είναι ο προσδιοριστής του μοντέλου

ϑs και ϑ_0 είναι το κορεσμένο μοντέλο και το εμφωλευμένο μοντέλο αντίστοιχα. Κορεσμένο μοντέλο καλείται το μοντέλο που έχει τόσες παραμέτρους όσες είναι και τα δεδομένα.

Η απόκλιση δείχνει τον βαθμό τον οποίο η πιθανοφάνεια του κορεσμένου μοντέλου $\log(p(y \mid \hat{\vartheta}_s)$ υπερβαίνει την πιθανοφάνεια του προτεινόμενου μοντέλου $\log(p(y \mid \hat{\vartheta}_0)$. Εάν το μοντέλο έχει καλή προσαρμογή, η απόκλιση είναι μικρή. Αν το προτεινόμενο μοντέλο έχει κακή προσαρμογή, η απόκλιση παίρνει μεγάλες τιμές.

3. Κριτήριο *Akaike*

Στα πλαίσια του εν λόγω κριτηρίου διενεργείται ο εξής υπολογισμός:

$$AIC = \frac{-2\log_e L(M_k) + 2p}{N}$$

Όπου, AIC είναι η εκτίμηση της μέγιστης πιθανοφάνειας του μοντέλου προσαρμογής, M_k και p ο αριθμός των παραμέτρων στο μοντέλο. Το μοντέλο με τη μικρότερη τιμή AIC θεωρείται ότι παρέχει την καλύτερη προσαρμογή και χρησιμοποιείται κυρίως στις περιπτώσεις σύγκρισης διαφορετικών μοντέλων.

4. Δείκτης λόγου Πιθανοφανειών

Για τις ανάγκες του εν λόγω μοντέλου διενεργείται η σύγκριση μεταξύ ενός μοντέλου με *k* ανεξάρτητες μεταβλητές και ενός αντίστοιχου μοντέλου στο οποίο οι εν λόγω μεταβλητές δεν υπάρχουν.

$$R^2{}_{MF} = 1 - \frac{\log_e L_M - k}{L_0}$$

5. Έλεγχος *Hosmer-Lemeshow*

Ο συγκεκριμένος έλεγχος εκτιμά την κατανομή των επιμέρους παρατηρήσεων, βάσει των αντίστοιχων πιθανοτήτων.

$$G^2{}_{HL} = \sum_{k-1}^{g} \frac{(O_k - n'_k \hat{p}_k)^2}{n'_k \hat{p}_k \cdot (1 - \hat{p}_k)}$$

n_k= ο αριθμός των ανεξάρτητων μεταβλητών στην k ομάδα
O_k= ο αριθμός των αποκρίσεων μεταξύ των n'k μεταβλητών
p_k=μέση τιμή πιθανότητας σε κάθε ομάδα.

6. Πολυσυγγραμικότητα

Η πολυγγραμικότητα αποτελεί ένα από τα βασικά ζητήματα όσον αφορά την εκτίμηση του μοντέλου. Ουσιαστικά, το φαινόμενο αυτό έγκειται στο γεγονός πως στα πλαίσια ενός γενικού γραμμικού μοντέλου, υπάρχει το ενδεχόμενο μία ή και περισσότερες εκ των ανεξάρτητων μεταβλητών να είναι εξαρτημένες. Στην περίπτωση αυτή οδηγούμαστε σε τυπικά σφάλματα. Βέβαια η συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών δεν είναι απαραίτητα γραμμική.

Ως εκ τούτου, θα πρέπει οι μεταβλητές οι οποίες χαρακτηρίζονται ως στατιστικά μη σημαντικές να παραλειφθούν. Προς αυτήν την κατεύθυνση είναι δυνατή η εφαρμογή διαφόρων τεχνικών όπως η παλινδρόμηση κορυφογραμμής και η παλινδρόμηση LASSO.

Η *παλινδρόμηση κορυφογραμμής* (ridge) είναι μια μέθοδος κατά την οποία εισάγεται μεροληψία (bias) στους συντελεστές αλλά με τέτοιο τρόπο έτσι ώστε να ελαχιστοποιείται η διακύμανση του πλαισίου δεδομένων. Για ένα μοντέλο με *k*-επεξηγηματικές μεταβλητές, χωρίς να λαμβάνουμε υπόψη το σταθερό όρο β_0, και ένα πλαίσιο δεδομένων με *n* παρατηρήσεις, η παλινδρόμηση Ridge ελαχιστοποιεί την παρακάτω ποσότητα:

$$RSS + \lambda \sum_{j=1}^{k} \beta_j{}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} \beta_j{}^2$$

Πολύ μεγάλη τιμή του λ θα ελαχιστοποιήσει τον συντελεστή RSS[3] και θα θέσει τους συντελεστές στο μηδέν. Μια μικρή τιμή του λ δεν θα εξαλείψει το πρόβλημα της μεροληψίας, ενώ μια τιμή του λ ίση με το μηδέν, απλά θα οδηγήσει σε εκπαίδευση του συστήματος σε γραμμική παλινδρόμηση.

Άλλη μια σημαντική τεχνική που περιορίζει το πρόβλημα της πολυσυγγραμμικότητας για να έχουμε καλύτερα αποτελέσματα, είναι η *LASSO* (Least Absolute Shrinkage and Selection Operator), η οποία σχεδιάστηκε το 1996 από τον Tibshirani [32] και είναι από τις κορυφαίες μεθόδους σήμερα. Ξεκίνησε με εφαρμογές στα γενικά γραμμικά μοντέλα, αλλά και στα γενικευμένα γραμμικά μοντέλα, όμως πλέον εφαρμόζεται και στα μοντέλα επιβίωσης, όπως το μοντέλο του *Cox*, της *Poisson* κλπ.

Θεωρείται ελκυστική ως μέθοδος, διότι έχει την ιδιότητα να εκτελεί ταυτόχρονα επιλογή μεταβλητών και συρρίκνωση του μοντέλου, εξ ου και η ονομασία της, γεγονός που την καθιστά πολύ χρήσιμη για την εξεύρεση ερμηνεύσιμων κανόνων πρόβλεψης για μεγάλων διαστάσεων δεδομένα.

Η διαφορά εμφανίζεται μόνο στον όρο ποινής, που περιλαμβάνει την ελαχιστοποίηση του αθροίσματος των απόλυτων τιμών των συντελεστών:

$$RSS + \lambda \sum_{j=1}^{k} \left|\beta_j\right| = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} \left|\beta_j\right|$$

2.11 Μηχανές Διανυσμάτων Υποστήριξης SVM

Στη συνέχεια αναφέρεται ένας άλλος αλγόριθμος ο οποίος επιλύει μη γραμμικά προβλήματα και ταυτόχρονα έχει μικρούς χρόνους εκπαίδευσης. Οι Μηχανές Διανυσμάτων Υποστήριξης, όπως ονομάζονται ,έχουν αποδειχθεί αρκετά αποδοτικές σε επίλυση προβλημάτων πρακτικών εφαρμογών.

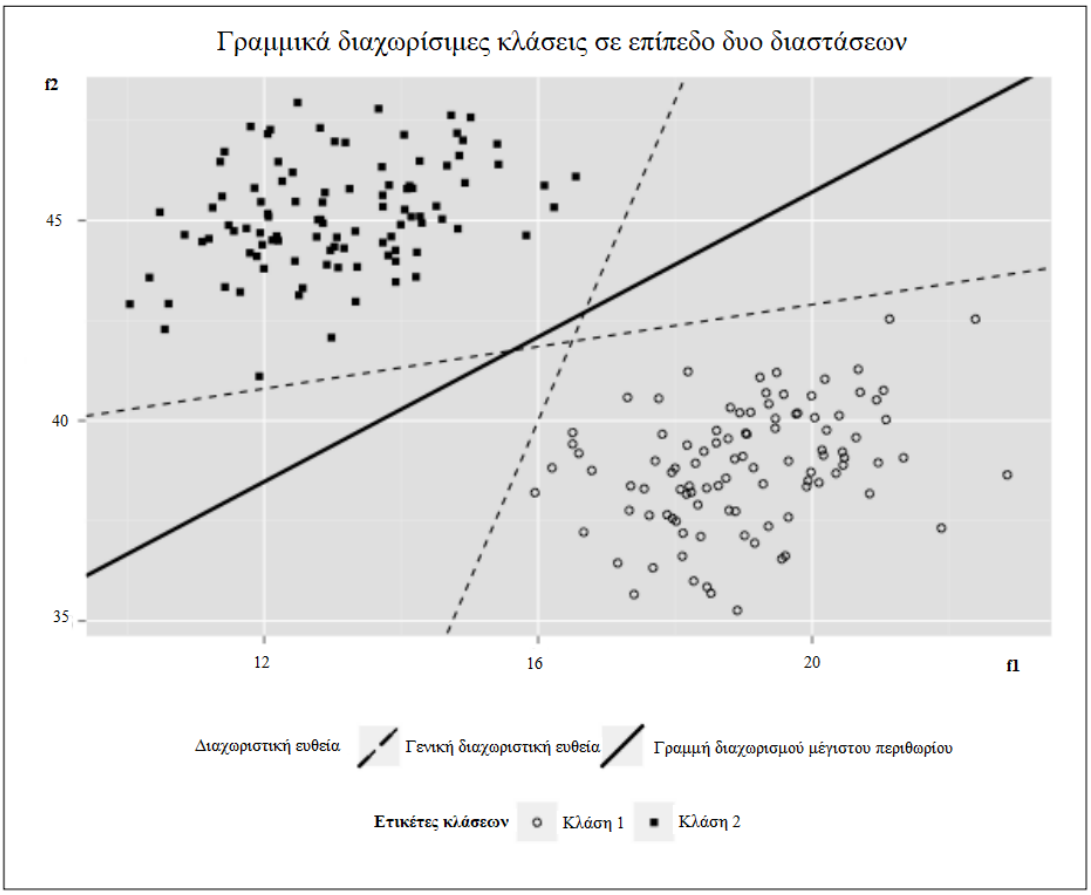
2.11.1 Γραμμικά διαχωρίσιμες κλάσεις σε επίπεδο δυο διαστάσεων

Για άλλη μία φορά, έχουμε να κάνουμε με ένα πρόβλημα δυαδικής ταξινόμησης, δηλαδή, το πρόβλημα του τρόπου σχεδιασμού ενός μοντέλου που θα προβλέπει σωστά εάν μια παρατήρηση ανήκει σε μία από δύο πιθανές κατηγορίες.

[3] Το άθροισμα των τετραγώνων των τιμών της παλινδρόμησης (RSS), εκφράζει τη μεταβλητότητα μεταξύ των προσαρμοσμένων τιμών

Έχουμε ήδη δει ότι αυτή η εργασία είναι απλούστερη όταν δύο κλάσεις είναι γραμμικά διαχωρίσιμες, δηλαδή, όταν μπορούμε να βρούμε ένα διαχωριστικό υπερεπίπεδο (επίπεδο σε πολυδιάστατο χώρο) στον χώρο των χαρακτηριστικών μας, έτσι ώστε όλες οι παρατηρήσεις στη μία πλευρά του υπερεπιπέδου να ανήκουν σε μία τάξη, και όλες οι παρατηρήσεις που βρίσκονται στην άλλη πλευρά να ανήκουν στη δεύτερη πλευρά του υπερεπιπέδου.

Αν οπτικοποιήσουμε αυτό το σενάριο, χρησιμοποιώντας ορισμένα δεδομένα σε ένα δυσδιάστατο χώρο χαρακτηριστικών, όπου το διαχωριστικό υπερεπίπεδο είναι απλώς μία διαχωριστική γραμμή (διαχωριστική ευθεία), θα λάβουμε την παρακάτω εικόνα:



Εικόνα 6 Γραμμικά διαχωρίσιμες κλάσεις σε επίπεδο δυο διαστάσεων.

Στην προηγούμενη εικόνα, μπορούμε να δούμε δύο ομάδες παρατηρήσεων, καθεμία από τις οποίες ανήκει σε διαφορετική κατηγορία. Χρησιμοποιούμε διαφορετικά σύμβολα για τις δύο κλάσεις, προκειμένου να είναι ξεκάθαρη η διαφοροποίηση. Στη συνέχεια, εμφανίζουμε τρεις διαφορετικές γραμμές που θα μπορούσαν να χρησιμεύσουν ως το *όριο απόφασης (decision boundary)* ενός ταξινομητή, με θεωρητική ακρίβεια ταξινόμησης 100% σε ολόκληρο το σύνολο δεδομένων. Επισημαίνεται ότι η εξίσωση ενός υπερεπιπέδου *(hyperplane)* μπορεί να εκφραστεί από το λεγόμενο *περιθώριο ταξινόμησης (margin)* [60] που συμβολίζεται με γ:

$$\gamma = \beta_0 + \sum_{k=1}^{p} \beta_k x_k$$

Όπου το β_0 ονομάζεται *βάρος κατωφλίου* και το β_k αντιπροσωπεύει το *διάνυσμα των βαρών*, ισχύουν: $x_k \in \Re^d$, $\beta_0 \in \Re^d$ και $\beta_k \in \Re$.

Επιπρόσθετα, για ένα διαχωρίσιμο υπερεπίπεδο, με σημεία δεδομένων (διανυσμάτων) x_{ik} και αντίστοιχων ζευγών τους y_i, παρουσιάζεται η παρακάτω ιδιότητα:

$$\beta_0 + \sum_{\kappa=1}^{p} \beta_k x_{ik} > 0 \text{ , αν } y_i = 1$$

$$\beta_0 + \sum_{\kappa=1}^{p} \beta_k x_{ik} < 0 \text{ , αν } y_i = -1$$

Στην πρώτη εξίσωση διαφαίνεται ότι τα σημεία δεδομένων που ανήκουν στην κλάση 1 βρίσκονται πάνω από το υπερεπίπεδο, ενώ τα σημεία δεδομένων που ανήκουν στην κλάση -1 βρίσκονται κάτω από το υπερεπίπεδο. Ο δείκτης *i* αντιπροσωπεύει τον αριθμό των παρατηρήσεων και ο δείκτης *k* χρησιμοποιείται για το πλήθος των χαρακτηριστικών, έτσι ώστε η μεταβλητή x_{ik} να σημαίνει την προσπέλαση του στοιχείου ενός δυσδιάστατου πίνακα. Μπορούμε να συνδυάσουμε αυτές τις δύο εξισώσεις σε μία εξίσωση για απλότητα, ως εξής:

$$y_i \cdot \left(\beta_0 + \sum_{\kappa=1}^{p} \beta_k x_{ik} \right) > 0, \forall i$$

2.11.2 Γραμμική Κατηγοριοποίηση (Hard margin)

Για να καταλάβουμε πώς βρήκαμε το υπερεπίπεδο μέγιστου περιθωρίου (maximal margin hyperplane), πρέπει να δούμε το πρόβλημα ως πρόβλημα βελτιστοποίησης με *p* χαρακτηριστικές (ή αλλιώς πρόβλημα τετραγωνικού προγραμματισμού) [59],[60], χρησιμοποιώντας τον ακόλουθο αλγόριθμο:

$$\textit{Επέλεξε } \beta_0, \beta_1 \ldots \beta_p \textit{ που μεγιστοποιεί το } M$$

$$\textit{κατά τέτοιο τρόπο ώστε } \sum_{k=1}^{p} \beta_k^2 = 1$$

$$\textit{και } \forall \iota : y_i \cdot \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) \geq M$$

Αυτές οι δύο σταθερές στο πρόβλημα βελτιστοποίησης εκφράζουν την ιδέα ότι οι παρατηρήσεις μας, σε ένα πλαίσιο δεδομένων, χρειάζεται όχι απλώς να είναι ορθά ταξινομημένες, αλλά να απέχουν *M* μονάδες απόσταση από το υπερεπίπεδο διαχωρισμού. Ο στόχος αυτός επιτυγχάνεται με κατάλληλη επιλογή του συντελεστή β_i. Κατά συνέπεια, χρειαζόμαστε μια διαδικασία βελτιστοποίησης που θα μπορεί να επιλύσει αυτό το πρόβλημα με τον καλύτερο τρόπο.

Σε αντίθεση με την ταξινόμηση στη λογιστική παλινδρόμηση όπου μεγιστοποιούμε την παράμετρο της Πιθανοφάνειας σε όλα τα δεδομένα, στον ταξινομητή μέγιστου περιθωρίου υπάρχει το λεγόμενο «*όριο απόφασης*» που υποστηρίζεται μόνο από τα σημεία εντός του περιθωρίου. Με άλλα λόγια, στο παράδειγμά μας, μπορούμε ελεύθερα να προσαρμόσουμε τη θέση οποιασδήποτε παρατήρησης, εκτός από τις τρεις στο περιθώριο και με την προϋπόθεση ότι η προσαρμογή δεν έχει ως αποτέλεσμα μια παρατήρηση να «πέσει» μέσα στο περιθώριο. Με αυτή τη συνθήκη, η διαχωριστική γραμμή θα παραμείνει ακριβώς στην ίδια θέση. Για τον λόγο αυτό, ορίζουμε τα κάθετα διανύσματα από τα σημεία που βρίσκονται στο περιθώριο έως το διαχωριστικό υπερεπίπεδο ως *διανύσματα υποστήριξης (support vectors)*.

2.11.3 Μη Γραμμική Κατηγοριοποίηση (Soft margin)

Είναι απαραίτητο τα δεδομένα μας να είναι γραμμικά διαχωρίσιμα, για να ταξινομηθούν με ταξινομητή μέγιστου περιθωρίου. Σε πραγματικές συνθήκες, τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, αλλά μπορούμε να χρησιμοποιήσουμε την έννοια του γραμμικού περιθωρίου με πρόσθετη εισαγωγή μεταβλητών.
Έτσι, ουσιαστικά, ορίζουμε ένα *μαλακό περιθώριο (soft margin)*, γεγονός που σημαίνει ότι ορισμένες από τις παρατηρήσεις στο σύνολο δεδομένων μπορούν να παραβιάσουν τον περιορισμό. Για να το επιτύχουμε αυτό, εισάγουμε τις λεγόμενες *μεταβλητές χαλαρότητας (slack variables)* στην αρχική εξίσωση. Όσο μεγαλύτερο είναι το περιθώριο τόσο πιο σίγουροι είμαστε για την ικανότητά μας να κάνουμε σωστά ταξινομήσεις σε νέες παρατηρήσεις, επειδή οι κλάσεις θα απέχουν περισσότερο μεταξύ τους, στο πλαίσιο δεδομένων εκπαίδευσης. Σε αυτή την περίπτωση ο αλγόριθμος περιγράφεται ως εξής:

$$\textit{Επέλεξε } \beta_0, \beta_1 \ldots \beta_p \textit{ που μεγιστοποιεί το } M$$

$$\textit{κατά τέτοιο τρόπο, ώστε } \sum_{k=1}^{p} \beta_k^2 = 1$$

$$\textit{και } \forall \iota : y_i \cdot \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) \geq M \cdot (1 - \xi_i)$$

$$\textit{και } \forall \iota : \xi_i \geq 0, \cdot \sum_{i=1}^{n} \xi_i \leq C$$

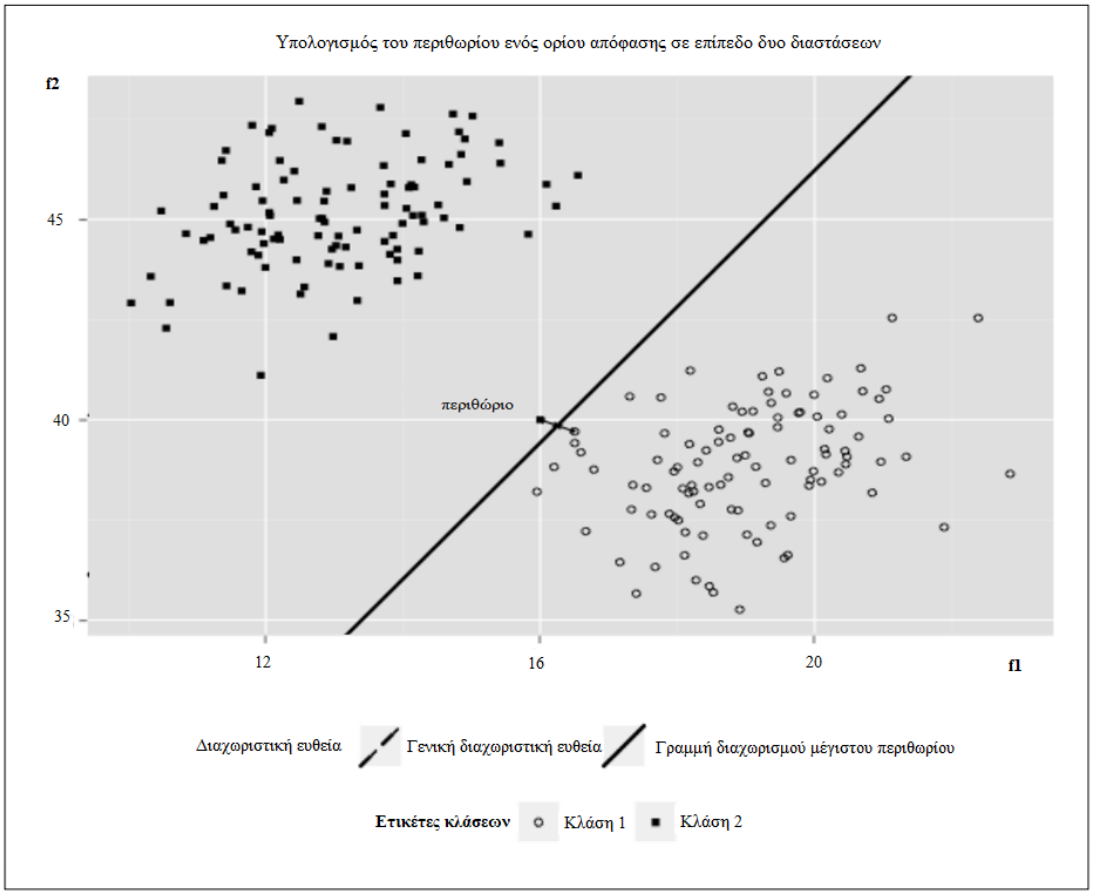
Όταν μια παρατήρηση βρίσκεται στη σωστή πλευρά του υπερεπιπέδου διαχωρισμού και έξω από το περιθώριο, η *μεταβλητή χαλαρότητας* για αυτή την παρατήρηση λαμβάνει την τιμή 0. Όταν μια παρατήρηση ταξινομείται σωστά, αλλά πέφτει σε απόσταση εντός του περιθωρίου, η αντίστοιχη μεταβλητή χαλαρότητας λαμβάνει μια μικρή θετική τιμή μικρότερη από 1. Όταν μια παρατήρηση δεν είναι ορθά ταξινομημένη, πέφτει, έτσι, στη «λάθος πλευρά» του υπερεπιπέδου και η μεταβλητή χαλαρότητας παίρνει τιμή μεγαλύτερη από 1.

Συνοψίζοντας:

$\xi_i = 0$, x_i είναι ταξινομημένο σωστά, και εκτός περιθωρίου

$0 \prec \xi_i \leq 1$, x_i είναι ταξινομημένο σωστά, αλλά πέφτει μέσα στο περιθώριο

$\xi_i \succ 1$, x_i δεν είναι ταξινομημένο σωστά



Εικόνα 7 Γραφική απεικόνιση περιθωρίου

Όταν μια παρατήρηση έχει ταξινομηθεί λανθασμένα, το μέγεθος των μεταβλητών χαλάρωσης είναι ανάλογο με την απόσταση μεταξύ αυτής της παρατήρησης και του ορίου του υπερεπιπέδου διαχωρισμού. Το γεγονός ότι το άθροισμα των μεταβλητών χαλάρωσης πρέπει να είναι μικρότερο από μια σταθερά C σημαίνει ότι μπορούμε να θεωρήσουμε αυτή την παράμετρο ως «*σταθερά σφάλματος*» που θα πρέπει να τη διαχειριστούμε ως προς την επιλογή της τιμής. Μεγάλη τιμή του C έχει ως αποτέλεσμα πολλές παρατηρήσεις να ταξινομούνται εσφαλμένα. Αυτό έχει ως συνέπεια το μοντέλο μας να παρουσιάζει μικρή διακύμανση και υψηλότερη μεροληψία. Από την άλλη, με μικρές τιμές του C έχουμε μικρότερη μεροληψία, αλλά υψηλά ποσοστά διακύμανσης σε διάφορα σετ δεδομένων εκπαίδευσης. Επομένως, η επιλογή του C θα πρέπει να είναι βέλτιστη, ώστε να μην έχουμε ούτε μεροληψία ούτε διακύμανση στο πλαίσιο εκπαίδευσης.

2.11.4 Εσωτερικά γινόμενα

Το υπό εξέταση μοντέλο μπορεί να απλοποιηθεί σε μια πιο «βολική» έκδοση, με τη χρήση των εσωτερικών γινομένων των παρατηρήσεων. Ένα εσωτερικό γινόμενο δύο διανυσμάτων, v_1 και v_2, πανομοιότυπου μήκους, υπολογίζεται πολλαπλασιάζοντας στοιχείο στοιχείο των δύο διανυσμάτων και λαμβάνοντας το άθροισμα των αποτελεσμάτων. Ένα παράδειγμα, γραμμένο στη γλώσσα R, φαίνεται παρακάτω:

```
v1 <- c(1.2, 3.3, -5.6, 4.5, 0, 9.0)
v2 <- c(-3.5, 0.1, -0.2, 1.0, -8.7, 0)
v1 * v2
[1] -4.20    0.33   1.12   4.50   0.00   0.00
inner_product <- sum(v1 * v2)
inner_product
[1] 1.75
```

Σύμφωνα με τη μέθοδο αυτή, μετασχηματίζουμε κατάλληλα τα διανύσματα εισόδου, ώστε να επιτύχουμε πιο εύκολη/γενικεύσιμη λύση του προβλήματος ταξινόμησης. Για να γίνει αυτό, μεταβαίνουμε, αρχικά, από τον χώρο εισόδου (input space) σε έναν μετασχηματισμένο χώρο χαρακτηριστικών (feature space) με πιθανόν υψηλότερη διάσταση, σε μία διάσταση με μία (μη γραμμική) αναπαράσταση. Σε μαθηματική μορφή, το εσωτερικό γινόμενο δύο διανυσμάτων μπορεί να αποτυπωθεί ακολούθως:

$$\langle \nu_1, \nu_2 \rangle = \sum_{i=1}^{p} \nu_{1i} \cdot \nu_{2i}$$

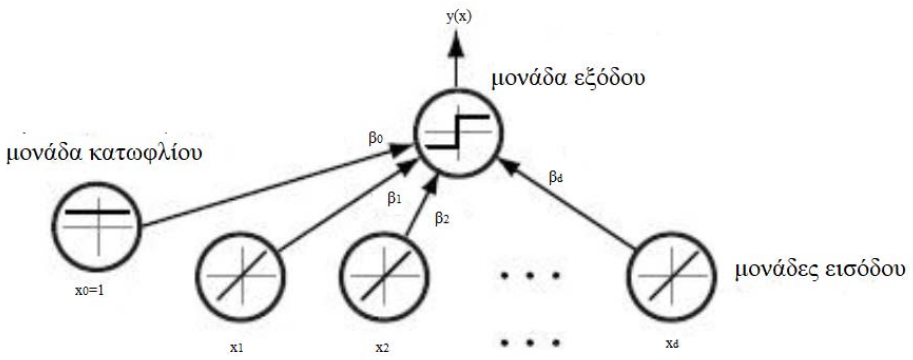
Για την παραπάνω εξίσωση, για τα δύο διανύσματα v_1 και v_2, ισχύει το άθροισμα των γινομένων v_1 και v_2, με τον δείκτη *i* να τρέχει από 1 έως *p*. Το *p* αντιπροσωπεύει είτε χαρακτηριστικές είτε διαστάσεις.

Είδαμε παραπάνω ότι η σχέση που ισχύει για τον ταξινομητή διανυσματικής υποστήριξης είναι:

$$y = \beta_0 + \sum_{\iota=1}^{p} \beta_k x_k$$

Αν θέλουμε να εκφράσουμε την παραπάνω εξίσωση με όρους εσωτερικού γινομένου μεταξύ μιας *x* παρατήρησης, την οποία προσπαθούμε να ταξινομήσουμε, και όλων των άλλων x_i παρατηρήσεων στο σετ δεδομένων εκπαίδευσης, θα πάρουμε την παρακάτω σχέση:

$$y(x) = \beta_0 + \sum_{i=1}^{n} a_i \langle x, x_i \rangle$$



Εικόνα 8 Γραμμικός ταξινομητής που έχει x μονάδες εισόδου, καθεμία από τις οποίες αντιστοιχεί στην τιμή ενός στοιχείου του διανύσματος εισόδου. Κάθε είσοδος (τιμή χαρακτηριστικού x_i πολλαπλασιάζεται με το αντίστοιχο βάρος της β_i

Δηλαδή, το μοντέλο προβλέπει την τιμή του *y* ως συνάρτηση μιας μεταβλητής εισόδου *x* που αντιστοιχεί σε μια παρατήρηση. Ο τελεστής του αθροίσματος υπολογίζει το σταθμισμένο άθροισμα των εσωτερικών γινομένων μεταξύ της υπό μελέτη παρατήρησης με όλες τις άλλες στο σετ δεδομένων, μέχρι έναν *n* αριθμό παρατηρήσεων. Η παράμετρος *α* ονομάζεται *πολλαπλασιαστής Lagrange* και προκύπτει από την ελαχιστοποίηση της τετραγωνικής συνάρτησης.

Σε ένα πραγματικό σενάριο, ο αριθμός των παρατηρήσεων στο σύνολο δεδομένων, το *n*, είναι, συνήθως, πολύ μεγαλύτερος από τον αριθμό των παραμέτρων *p*, με επακόλουθο ο αριθμός των συντελεστών *α* να είναι φαινομενικά μεγαλύτερος από τον αριθμό των συντελεστών *β*. Ο αριθμός των διανυσμάτων υποστήριξης στο πλαίσιο δεδομένων είναι, τυπικά, πολύ μικρότερος από τον συνολικό αριθμό των παρατηρήσεων και η εξίσωση απλοποιείται ως εξής:

$$y(x) = \beta_0 + \sum_{s \in S} a_s \langle x, x_s \rangle$$

Όπου το *S* αντιπροσωπεύει το πλαίσιο δεδομένων εκπαίδευσης.

2.11.5 Συναρτήσεις Πυρήνα (Kernel Functions)

Τα προβλήματα που μέχρι τώρα έχουμε αντιμετωπίσει με τις μηχανές διανυσματικής υποστήριξης (SVMs) είναι αυτά με διαχωριστικές γραμμές (ή υπερεπίπεδα) διανυσμάτων υποστήριξης και σκοπό τον βέλτιστο διαχωρισμό δύο κλάσεων, όταν αυτές μπορούν να διαχωριστούν γραμμικά (hard margin) ή ακόμη και όταν αυτό δεν είναι εφικτό (soft margin). Είδαμε ότι στην περίπτωση που τα δεδομένα μας δεν είναι γραμμικώς διαχωρίσιμα, τότε, ανάλογα με τις τιμές που μπορεί να δώσει ο χρήστης στο C, ελέγχεται και η σωστή ταξινόμηση των δεδομένων. Όμως, ο γραμμικός διαχωρισμός χωρίς λάθη, σε τέτοιες περιπτώσεις, δεν είναι πραγματοποιήσιμος και θα υπήρχε αποτυχία διαχωρισμού αρκετών στιγμιοτύπων.

Ένας τρόπος να ξεπεράσουμε αυτό το πρόβλημα είναι να εισάγουμε στο μοντέλο μας έναν μη γραμμικό μετασχηματισμό. Ορίζουμε μια γενική συνάρτηση K, που την ονομάζουμε *συνάρτηση πυρήνα (kernel function)*, η οποία επενεργεί σε δύο διανύσματα, με συνέπεια να παράγεται το αποτέλεσμα:

$$y(x) = \beta_0 + \sum_{s \in S} \alpha_s K\langle x, x_s \rangle$$

Ακολουθώντας αυτή την τεχνική, μετασχηματίζονται τα διανύσματα εισόδου, ώστε να έχουμε πιο εύκολη λύση στο πρόβλημα ταξινόμησης. Πηγαίνοντας σε έναν χώρο υψηλότερης διάστασης, τα δεδομένα γίνονται περισσότερο «διαχωρίσιμα». Για να επιτύχουμε κάτι τέτοιο, μεταφερόμαστε, σε πρώτη φάση, από τον *χώρο εισόδου* (input space) σε έναν άλλο *χώρο χαρακτηριστικών (feature space),* σε υψηλότερη διάσταση.

Μπορούμε να χρησιμοποιήσουμε τις συναρτήσεις πυρήνα σε αλγορίθμους ταξινόμησης με το λεγόμενο «τέχνασμα πυρήνα» (kernel trick): Ο αλγόριθμος ταξινόμησης, που εκφράζεται σε σχέση με κάποιο εσωτερικό γινόμενο, αντικαθίσταται με μια οποιαδήποτε συνάρτηση πυρήνα, όπως αυτές που αναφέρονται παρακάτω:

Γραμμικός πυρήνας (linear kernel)

$$\mathrm{K}_{linear}(x_i, x_j) = \sum_{k=1}^{p} x_{ik} x_{jk}$$

Πολυωνυμικός πυρήνας (polynomial kernel)

$$K_{polynomial}(x_i, x_j) = \left(1 + \sum_{k=1}^{p} x_{ik} x_{jk}\right)^d$$

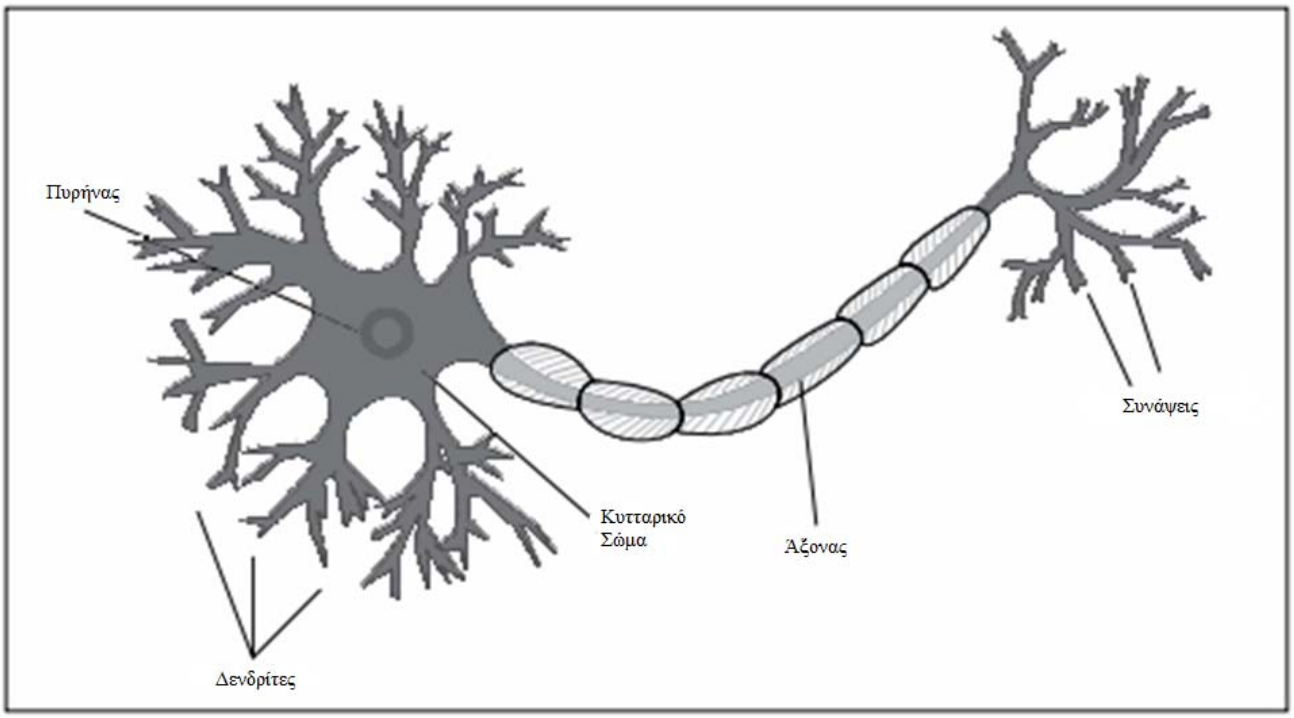
Πυρήνας ακτινικής βάσης ή Γκαουσιανός (radial basis kernel)

$$K_{radial}(x_i, x_j) = e^{-\frac{1}{2\sigma^2} \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$$

2.12 Νευρωνικό Δίκτυο

2.12.1 Ο βιολογικός νευρώνας

Τα μοντέλα νευρωνικών δικτύων αντλούν την αναλογία τους από την οργάνωση των νευρώνων στον ανθρώπινο εγκέφαλο. Γι' αυτό τον λόγο συχνά αναφέρονται και ως *τεχνητά νευρωνικά δίκτυα* (ANN). Θα μπορούσε να θεωρηθεί ότι ένας μόνο βιολογικός νευρώνας λειτουργεί ως απλή υπολογιστική μονάδα και ότι, όταν ένας μεγάλος αριθμός νευρώνων συνδυάζονται μαζί, το αποτέλεσμα είναι μια εξαιρετικά ισχυρή και μαζικά κατανεμημένη μηχανή επεξεργασίας, που είναι ικανή να παράγει μάθηση. Η μηχανή αυτή είναι γνωστή ως ανθρώπινος εγκέφαλος.



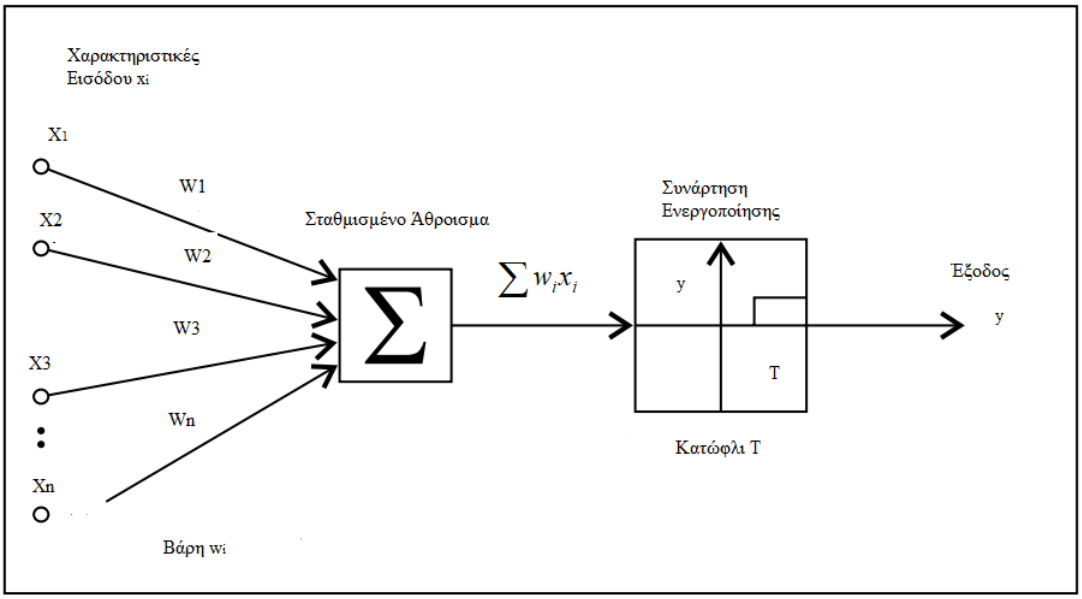
Εικόνα 9 Το ανθρώπινο νευρικό κύτταρο

Με λίγα λόγια, μπορούμε να σκεφτούμε έναν ανθρώπινο νευρώνα ως μια υπολογιστική μονάδα που αποτελείται από σειρά παράλληλων εισόδων ηλεκτρικών σημάτων –γνωστών ως *συναπτικών νευροδιαβιβαστών (synaptic neurotransmitters)*– που προέρχονται από τους *δενδρίτες*. Οι δενδρίτες μεταδίδουν χημικά σήματα στο σώμα του νευρώνα ως απόκριση στα σήματα των συναπτικών νευροδιαβιβαστών. Αυτή η μετατροπή ενός εξωτερικού σήματος εισόδου σε «εσωτερικό» μπορεί να θεωρηθεί ως διαδικασία στην οποία οι δενδρίτες εφαρμόζουν ένα *βάρος* (το οποίο μπορεί να είναι αρνητικό ή θετικό, ανάλογα αν οι χημικές ουσίες που παράγονται είναι *αναστολείς* ή *ενεργοποιητές*, αντίστοιχα) σε σχέση με τις εισόδους τους.

Το σώμα του νευρώνα, που φιλοξενεί τον *πυρήνα* ή *κεντρικό επεξεργαστή*, αναμειγνύει τα σήματα εισόδου σε μία διαδικασία που μπορεί να θεωρηθεί ως άθροιση όλων των σημάτων. Κατά συνέπεια, οι αρχικές είσοδοι από τους δενδρίτες μετατρέπονται βασικά σε ένα ενιαίο *γραμμικά σταθμισμένο άθροισμα*. Αυτό το άθροισμα αποστέλλεται στον άξονα του νευρώνα, ο οποίος λειτουργεί ως πομπός του νευρώνα. Το σταθμισμένο άθροισμα των ηλεκτρικών εισόδων δημιουργεί ένα αντίστοιχο ηλεκτρικό δυναμικό στον νευρώνα και το δυναμικό αυτό υφίσταται επεξεργασία στον άξονα (*λειτουργία ενεργοποίησης)*, που θα καθορίσει εάν ο νευρώνας θα πυροδοτηθεί.

2.12.2 Ο τεχνητός νευρώνας (μοντέλο ΜακΚάλοχ-Πιτς)

Με βάση την παραπάνω περιγραφή, χρησιμοποιώντας αυτή τη βιολογική αναλογία, μπορούμε να κατασκευάσουμε ένα μοντέλο υπολογιστικού νευρώνα, το οποίο είναι γνωστό ως *μοντέλο νευρώνα ΜακΚάλοχ-Πιτς* (*McCulloch-Pitts*) [63].



Εικόνα 10 Το μοντέλο McCulloch-Pitts για τον τεχνητό νευρώνα

Ο υπολογιστικός νευρώνας αποτελεί το απλούστερο παράδειγμα νευρωνικού δικτύου. Μπορούμε να «κατασκευάσουμε» τη συνάρτηση εξόδου, y, του νευρωνικού μας δικτύου απευθείας. Αν θεωρήσουμε ότι $x_1,x_2...x_v$ είναι οι *είσοδοι του νευρώνα* και w_i τα *συναπτικά βάρη*, τότε το άθροισμα του φορτίου που δέχεται ο νευρώνας είναι:

$$y = g(w_0 + \sum_{i=1}^{p} w_i x_i)$$

Η συνάρτηση *g ()* στο νευρωνικό μας δίκτυο είναι η **συνάρτηση ενεργοποίησης**. Εδώ, στο συγκεκριμένο παράδειγμα, η συνάρτηση ενεργοποίησης που επιλέγεται είναι η βηματική.

$$g(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

Όταν το γραμμικό σταθμισμένο άθροισμα των εισόδων υπερβεί το μηδέν, η συνάρτηση του βήματος εξάγει 1· όταν δεν το υπερβεί, η συνάρτηση εξάγει -1. Αρκετές φορές θεωρούμε ότι, εκτός από τα σήματα εισόδου και τα αντίστοιχα βάρη *w*, ο νευρώνας έχει και ένα *εσωτερικό βάρος* που τον χαρακτηρίζει, το οποίο πρέπει, επίσης, να ληφθεί υπόψη στην εξίσωση.

Αυτό το εσωτερικό βάρος ονομάζεται *παράγοντας προδιάθεσης του νευρώνα* (*dummy input)*, είναι ξεχωριστό από τα άλλα βάρη, παρόλα αυτά επενεργεί κατά τον ίδιο τρόπο με τα άλλα βάρη. Κατά συνέπεια, υπάρχει η θεώρηση ότι η τιμή του σήματος που περνάει σε όλα τα εσωτερικά βάρη είναι 1. Έτσι, λοιπόν, οι μονάδες του w_0 θα είναι οι ίδιες με τις μονάδες του γινομένου ($w_i x_i$), με αποτέλεσμα η εξίσωση, στην πιο γενική μορφή της, τώρα να γίνεται:

$$y = g \cdot \left(\sum_{i=0}^{p} w_i x_i \right)$$

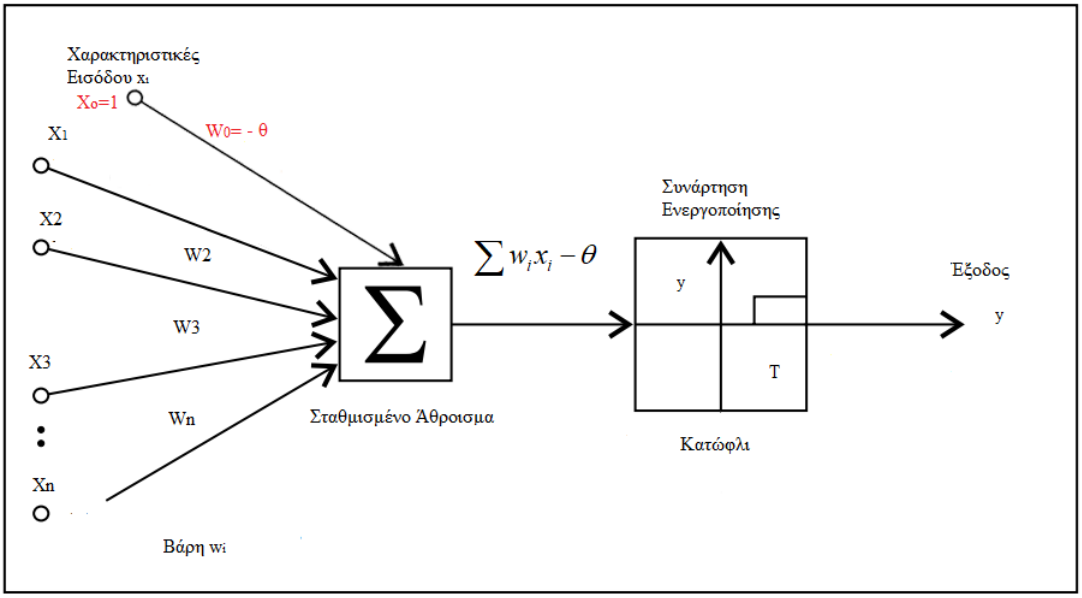
2.12.3 Ο αισθητήρας perceptron (μοντέλο Ρόζενμπλατ)

Ο Αμερικανός ψυχολόγος Frank Rosenblatt [62] πρότεινε το κλασικό μοντέλο του αισθητήρα perceptron το 1958. Ακολούθησε περαιτέρω επεξεργασία και προσεκτική ανάλυση από τους Minsky και Papert (1969), οι οποίοι τελειοποίησαν το μοντέλο που αναφέρεται ως μοντέλο perceptron.

Το μοντέλο perceptron είναι πιο γενικό υπολογιστικό μοντέλο από τον νευρώνα McCulloch-Pitts (M-P). Ξεπερνά μερικούς από τους περιορισμούς του νευρώνα M-P εισάγοντας την έννοια των συναπτικών βαρών (ένα μέτρο σπουδαιότητας) για τις εισόδους και έναν μηχανισμό για την εκμάθηση αυτών των βαρών. Οι είσοδοι δεν περιορίζονται πλέον σε δυαδικές (boolean) τιμές, όπως στην περίπτωση ενός νευρώνα M-P, γεγονός που καθιστά το μοντέλο αυτό πιο χρήσιμο και γενικευμένο.

$$y = \sum_{i=1}^{n} w_i x_i - \theta$$

Σε αυτή την περίπτωση παίρνουμε το σταθμισμένο άθροισμα των εισόδων και ορίζουμε έξοδο, όταν το άθροισμα είναι μεγαλύτερο από ένα αυθαίρετο όριο, που συμβολίζεται με θ (*theta*). Πρέπει να σημειωθεί ότι, αντί να ρυθμίσουμε χειροκίνητα την παράμετρο κατωφλίου, την προσθέτουμε σε μία από τις εισόδους, με το βάρος *θήτα* όπως φαίνεται παρακάτω, γεγονός που την καθιστά ικανή για εκμάθηση.



Εικόνα 11 Το μοντέλο Minsky-Papert για τον τεχνητό νευρώνα perceptron

Το μοντέλο perceptron μπορεί να χρησιμοποιηθεί μόνο για την εφαρμογή γραμμικά διαχωρίσιμων συναρτήσεων. Παίρνει και πραγματικές και boolean τιμές εισόδου και τις πολλαπλασιάζει με ένα σύνολο βαρών.

Το πλεονέκτημα του μοντέλου perceptron έναντι του νευρώνα M-P είναι ότι μπορεί να μάθει μέσω της διόρθωσης σφάλματος και να διαχωρίσει γραμμικά το πρόβλημα χρησιμοποιώντας ένα υπερεπίπεδο. Με αυτό τον τρόπο, οτιδήποτε πέφτει κάτω από το υπερεπίπεδο είναι 0 και οτιδήποτε πάνω από αυτό είναι 1. Αυτή η διόρθωση σφάλματος επιτρέπει στον perceptron να ρυθμίσει τα βάρη και να μετακινήσει τη θέση του υπερεπιπέδου ώστε να μπορεί να ταξινομήσει σωστά τα δεδομένα. Νωρίτερα, αναφέρθηκε ότι το perceptron μαθαίνει να ταξινομεί γραμμικά ένα πρόβλημα, όπως επίσης μαθαίνει και την επίδραση της εισόδου στην έξοδο. Έτσι, όσο μεγαλύτερο είναι το βάρος που σχετίζεται με μία συγκεκριμένη είσοδο τόσο μεγαλύτερη είναι η επίδραση στην πρόβλεψη (ταξινόμηση).

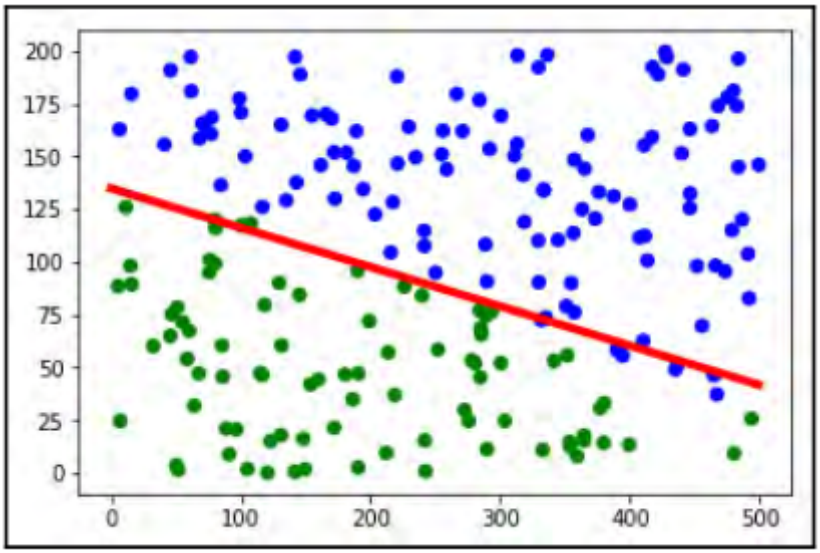
Η ενημέρωση για τα βάρη (μάθηση) γίνεται ως εξής:

$$w_{new} = w_{old} + \delta x$$

Όπου δ = αναμενόμενη τιμή - προβλεπόμενη τιμή.
Θα μπορούσαμε επίσης να προσθέσουμε ένα ποσοστό εκμάθησης ($0 < \eta \leq 1$) εάν θέλουμε να επιταχύνουμε τη μάθηση, ώστε η ενημέρωση για τα βάρη να είναι ως εξής:

$$w_{new} = w_{old} + \eta \cdot \delta x$$

Κατά τη διάρκεια αυτών των ενημερώσεων, το perceptron υπολογίζει την απόσταση του υπερεπιπέδου από τα σημεία που είναι να ταξινομηθούν και αυτοπροσαρμόζεται ώστε να βρει τη βέλτιστη θέση, δηλαδή την τέλεια γραμμική ταξινόμηση για τις δύο κλάσεις. Έτσι, χωρίζει στο μέγιστο και τα δύο σημεία εκατέρωθεν, τα οποία μπορούμε να δούμε στο παρακάτω διάγραμμα:



Εικόνα 12 Ο τρόπος διαχωρισμού δυο κλάσεων με βάση τον νευρώνα perceptron

2.12.4 Στοχαστική κατάβαση κλίσης

Σε αυτή την ενότητα θα εξετάσουμε μία σημαντική μέθοδο που χρησιμοποιείται στα τεχνητά νευρωνικά δίκτυα: την **κατάβαση κλίσης** (gradient descent). Με αυτή τη μαθηματική πράξη ουσιαστικά ελαχιστοποιείται το τετράγωνο της διαφοράς της εξόδου και της επιθυμητής τιμής για όλους τους νευρώνες εξόδου.[61]

Στα μοντέλα που είδαμε μέχρι τώρα, όπως η λογιστική παλινδρόμηση, αναφερθήκαμε σε ένα κριτήριο ή μία συνάρτηση που πρέπει να ελαχιστοποιήσει το μοντέλο ενώ εκπαιδεύεται. Αυτό το κριτήριο είναι γνωστό ως **συνάρτηση κόστους** (cost function). Για παράδειγμα, η συνάρτηση κόστους ελαχίστων τετραγώνων για ένα μοντέλο μπορεί να διατυπωθεί ως εξής:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(\bar{y}_i - y_i\right)^2$$

Αν υποθέσουμε ότι τα δεδομένα μας έχουν σταθερές τιμές και τα βάρη είναι μεταβλητά ως προς τις τιμές τους και ότι πρέπει να επιλεγούν ώστε να ελαχιστοποιηθεί το κριτήριο, μπορούμε να διατυπώσουμε τη συνάρτηση κόστους σε σχέση με τις τιμές των βαρών ως ακολούθως:

$$J(\vec{w}) = \frac{1}{2n}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{p} w_j x_j\right) - y_i\ \right)^2$$

Για τα βάρη του μοντέλου χρησιμοποιούμε το γράμμα *w*. Δεδομένου ότι οι μεταβλητές του μοντέλου μας είναι τα βάρη, μπορούμε να θεωρήσουμε τη συνάρτηση ως συνάρτηση ενός διανύσματος βάρους *w*. Προκειμένου να βρούμε το ελάχιστο αυτής της συνάρτησης πρέπει απλώς να λάβουμε το μερικό διαφορικό της συνάρτησης κόστους σε σχέση με το διάνυσμα βάρους. Για συγκεκριμένο βάρος w_k, αυτή η σχέση μερικής παραγώγου δίνεται από:

$$\frac{\partial J(\vec{w})}{\partial w_k} = \frac{1}{n}\sum_{i=1}^{n}((\sum_{j=1}^{p} w_j x_j) - y_i)\cdot x_{ik}$$

Αντικαθιστώντας στην παραπάνω εξίσωση το εσωτερικό άθροισμα με την απόδοση του μοντέλου $\bar{y}_i$, η σχέση απλοποιείται αρκετά:

$$\frac{\partial J(\vec{w})}{\partial w_k} = \frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)\cdot x_{ik}$$

Δηλαδή το μερικό διαφορικό της συνάρτησης κόστους που προσπαθούμε να ελαχιστοποιήσουμε για ένα συγκεκριμένο βάρος w_k , στο μοντέλο μας, είναι απλώς η διαφορά μεταξύ της τιμής πρόβλεψης και της επιθυμητής εξόδου, πολλαπλασιασμένο κατά έναν παράγοντα x_{ik}. (για την i^{th} παρατήρηση, η τιμή της ανεξάρτητης μεταβλητής εισόδου που αντιστοιχεί στο βάρος w_k). Όπου *n*, είναι το σύνολο των παρατηρήσεων στο πλαίσιο δεδομένων.

Για να βρούμε τις βέλτιστες τιμές βαρών, πρέπει να λύσουμε αυτή την εξίσωση για κάθε βάρος στον αντίστοιχο πίνακα. Με άλλα λόγια, παράγεται ένα πλήρες σύστημα γραμμικών εξισώσεων που είναι συχνά πολύ μεγάλο, συνεπώς η άμεση επίλυσή του είναι πολλές φορές απαγορευτική, καθώς απαιτείται μεγάλη υπολογιστική ισχύς. Αντ' αυτού, πολλές υλοποιήσεις μοντέλων χρησιμοποιούν προσεγγιστικές μεθόδους βελτιστοποίησης (*iterative optimization procedures*) που έχουν σχεδιαστεί για να προσεγγίζουν σταδιακά τη σωστή λύση. Μία τέτοια μέθοδος είναι η κατάβαση κλίσης. Για μία συγκεκριμένη τιμή του διανύσματος βάρους, η κατάβαση κλίσης βρίσκει την κατεύθυνση στην οποία η κλίση της συνάρτησης κόστους είναι περισσότερο απότομη και προσαρμόζει τα βάρη προς αυτή την κατεύθυνση κατά ένα μικρό ποσό, το οποίο καθορίζεται από μία παράμετρο γνωστή ως **ρυθμός εκμάθησης** (learning rate). Έτσι, η εξίσωση ενημέρωσης είναι:

$$w_k \longleftarrow w_k - \eta\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)\cdot x_{ik}$$

Δηλαδή, ανά βήμα, αλλάζονται οι τιμές των βαρών *w* με ρυθμό εκμάθησης «*η*» προς την κατεύθυνση που προκαλείται η μεγαλύτερη μείωση σφάλματος. Εάν επιλέξουμε μια τιμή για το «*η*» που είναι πολύ μικρή, ο αλγόριθμος θα ενημερώσει τα βάρη κατά ελάχιστο ποσό κάθε φορά και έτσι θα διαρκέσει πολύ να τελειώσει. Εάν χρησιμοποιήσουμε μια τιμή για το «*η*» που είναι πολύ μεγάλη, ενδέχεται να προκαλέσουμε αλλαγή στα βάρη πολύ δραστικά, ωστόσο ο ρυθμός εκμάθησης θα ταλαντεύεται μεταξύ των τιμών και έτσι ο αλγόριθμος μάθησης πάλι θα χρειάζεται πολύ χρόνο για να συγκλίνει.

Μία παραλλαγή της μεθόδου στοχαστικής κλίσης που κάνει παρόμοιο υπολογισμό αλλά παίρνει τις παρατηρήσεις μία-μία κάθε φορά, αντί για όλες μαζί, ονομάζεται **στοχαστική κατάβαση κλίσης** (stochastic gradient descent). Η βασική ιδέα είναι ότι κατά μέσο όρο υπολογίζεται η κλίση της συνάρτησης κόστους για μία συγκεκριμένη παρατήρηση και αυτή ισούται με την κλίση όλων των παρατηρήσεων. Αυτό, φυσικά, αποτελεί προσέγγιση, παρ' όλα αυτά σημαίνει ότι μπορούμε να επεξεργαζόμαστε μεμονωμένες παρατηρήσεις, μία τη φορά, με πλεονέκτημα το μικρότερο υπολογιστικό κόστος. (tricks-2012.pdf (microsoft.com))

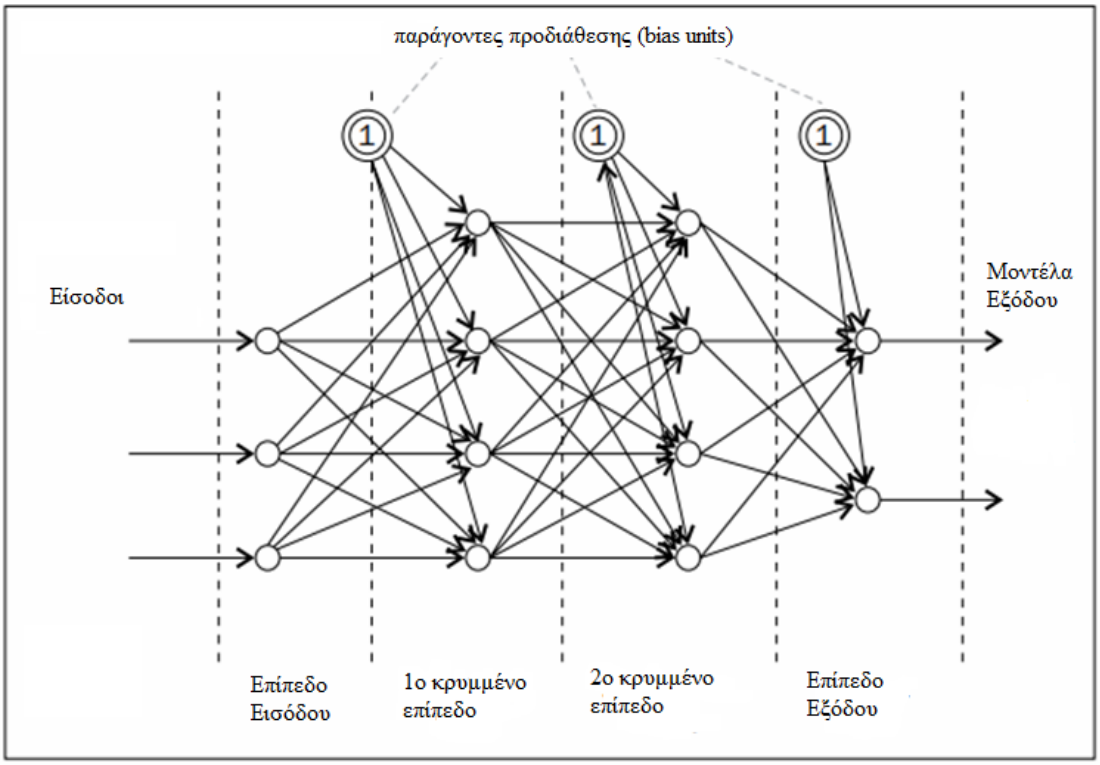
Η στοχαστική κατάβαση κλίσης ενημερώνει ένα συγκεκριμένο βάρος κατά την επεξεργασία με την ακόλουθη εξίσωση:

$$w_k \longleftarrow w_k - \eta\left(\bar{y}_i - y_i\right)\cdot x_{ik}$$

2.12.5 Πολυεπίπεδα δίκτυα perceptron

Τα πολυστρωματικά νευρωνικά δίκτυα είναι μοντέλα που συνδέουν πολλούς νευρώνες προκειμένου να δημιουργούν τη λεγόμενη **νευρωνική αρχιτεκτονική**. Οι νευρώνες είναι οι δομικές μονάδες του δικτύου, αλλά, όταν συνδεθούν μεταξύ τους, μπορούμε να δημιουργήσουμε ένα μοντέλο αρκετά ισχυρό σε σχέση με τις μεμονωμένες περιπτώσεις νευρώνων.

Τα νευρωνικά δίκτυα σχεδιάζονται σε διαφορετικά **επίπεδα** (layers) και γίνεται διάκριση μεταξύ διαφορετικών ειδών νευρωνικών δικτύων, κυρίως με βάση τις **συνδέσεις** (connections) που υπάρχουν μεταξύ αυτών των στρωμάτων και των τύπων νευρώνων που χρησιμοποιούνται. Σε πολλές περιπτώσεις, στην αρχιτεκτονική, υπάρχουν και ενδιάμεσα επίπεδα, τα οποία ονομάζονται **κρυμμένα** (hidden layers). Μεταξύ των επιπέδων διακρίνουμε **κόμβους** (nodes) και **μονάδες** (units). Οι μεταβλητές εισόδου βρίσκονται στο λεγόμενο **επίπεδο εισόδου** (input layer), ενώ η εξαρτημένη μεταβλητή, αντίστοιχα, στο **επίπεδο εξόδου** (output layer). Το παρακάτω διάγραμμα δείχνει τη γενική δομή ενός **πολυστρωματικού perceptron (MLP)** νευρωνικού δικτύου, για δύο κρυμμένα επίπεδα:



Εικόνα 13 Πολυστρωματικό Perceptron με δυο κρυμμένα επίπεδα

Το πρώτο χαρακτηριστικό του δικτύου MLP είναι ότι οι πληροφορίες ρέουν σε μία μόνο κατεύθυνση από το επίπεδο εισόδου στο επίπεδο εξόδου. Γι' αυτό είναι γνωστό ως **πρόσθιας τροφοδότησης νευρωνικό δίκτυο** (feedforward neural network). Σε άλλους τύπους νευρωνικών δικτύων, οι ρέουσες πληροφορίες επιστρέφουν σε προηγούμενους νευρώνες στο δίκτυο και θεωρούνται σήμα ανατροφοδότησης. Αυτά τα δίκτυα είναι γνωστά ως **οπίσθιας τροφοδότησης νευρωνικά δίκτυα** (feed backward) ή **αναδρομικά νευρωνικά δίκτυα** (recurrent). Τα αναδρομικά νευρωνικά δίκτυα είναι γενικά πολύ δύσκολο να εκπαιδευτούν. Παρόλα αυτά, βρίσκουν έναν αριθμό εφαρμογών, ιδίως με προβλήματα που αφορούν χρονικά στοιχεία, όπως, για παράδειγμα, πρόβλεψη και επεξεργασία σήματος.

Επιστρέφοντας στην αρχιτεκτονική MLP που φαίνεται στο διάγραμμα, αναφέρουμε ότι η πρώτη ομάδα των νευρώνων στα αριστερά είναι γνωστοί ως **νευρώνες εισόδου** και σχηματίζουν το **στρώμα εισόδου**. Πάντα έχουν τόσους νευρώνες εισόδου όσες είναι και οι είσοδοι. Στην άκρη δεξιά του διαγράμματος έχουμε το **στρώμα εξόδου** με τους **νευρώνες εξόδου**. Συνήθως έχουμε τόσους νευρώνες εξόδου για όσες εξόδους μοντελοποιούμε.

Μεταξύ των επιπέδων εισόδου και εξόδου έχουμε τα **κρυμμένα επίπεδα**. Οι νευρώνες είναι οργανωμένοι σε επίπεδα. Για παράδειγμα, οι νευρώνες στο πρώτο κρυφό στρώμα συνδέονται άμεσα με τουλάχιστον έναν νευρώνα στο στρώμα εισόδου, ενώ οι νευρώνες στο δεύτερο κρυφό στρώμα συνδέονται επίσης με έναν ή περισσότερους νευρώνες στο πρώτο κρυφό στρώμα. Το διάγραμμά μας είναι ένα παράδειγμα αρχιτεκτονικής 4-4, που σημαίνει ότι υπάρχουν δύο κρυμμένα επίπεδα με τέσσερις νευρώνες το καθένα. Παρόλο που δεν αποτελούν νευρώνες, το διάγραμμα δείχνει και τα εσωτερικά βάρη, που ονομάζονται **μονάδες προδιάθεσης** ή **παράγοντες προδιάθεσης** (bias units). Οι μονάδες προδιάθεσης έχουν πάντα την τιμή 1 για όλα τα εσωτερικά βάρη.

Δεν έχουν όλοι οι νευρώνες στην αρχιτεκτονική την ίδια συνάρτηση ενεργοποίησης. Επιλέγεται ξεχωριστά η συνάρτηση ενεργοποίησης για τους νευρώνες στα κρυφά στρώματα και ξεχωριστά σε αυτούς που βρίσκονται στο επίπεδο εξόδου. Η συνάρτηση ενεργοποίησης για το επίπεδο εξόδου επιλέγεται με βάση τον επιθυμητό τύπο εξόδου, δηλαδή αν επιτελείται παλινδρόμηση ή ταξινόμηση.

Η συνάρτηση ενεργοποίησης για τους νευρώνες των κρυμμένων στρωμάτων είναι γενικά μη γραμμική, επειδή η συνέλιξη των γραμμικών νευρώνων μπορεί να απλοποιηθεί αλγεβρικά και να καταλήξει σε έναν μόνο γραμμικό νευρώνα με διαφορετικά βάρη. Η πιο κοινή συνάρτηση ενεργοποίησης είναι η λογιστική συνάρτηση, αλλά, επίσης, χρησιμοποιείται και η υπερβολική εφαπτόμενη συνάρτηση.

Η έξοδος του νευρωνικού δικτύου λαμβάνεται διαδοχικά, υπολογίζοντας τις εξόδους των νευρώνων κάθε στρώματος. Ένα από τα πλεονεκτήματα των νευρωνικών δικτύων είναι αυτή η δύναμη να αποκτούν γνώση μέσω της εκμάθησης των βαρών στα κρυμμένα στρώματα. Αυτή η διαδικασία επαναλαμβάνεται για κάθε στρώμα στο νευρωνικό δίκτυο μέχρι το τελικό στρώμα, όπου λαμβάνουμε την έξοδο στο σύνολό της. Η διαδικασία διάδοσης των σημάτων από την είσοδο στο επίπεδο εξόδου είναι γνωστή ως **εμπρόσθια διάδοση** (forward propagation).

2.12.6 Εκπαίδευση πολυστρωματικού νευρωνικού δικτύου perceptron

Τα πολυστρωματικού perceptron (MLP) δίκτυα που περιλαμβάνουν κρυμμένα επίπεδα είναι πιο περίπλοκα στην εκπαίδευση σε σχέση με τον απλό νευρώνα perceptron. Ο αλγόριθμος που χρησιμοποιήθηκε για την εκπαίδευσή τους, ο οποίος υπήρχε από τη δεκαετία του 1980, είναι γνωστός ως **αλγόριθμος οπισθοδιάδοσης** (backpropagation) [64].

Υπάρχουν δύο πολύ σημαντικά στοιχεία προκειμένου να γίνει κατανοητός αυτός ο αλγόριθμος. Κάθε παρατήρηση εξελίσσεται σε δύο βήματα. Με το βήμα εμπρόσθιας διάδοσης, που ξεκινά από το επίπεδο εισόδου και τελειώνει στο επίπεδο εξόδου, υπολογίζεται η τιμή της εξόδου του δικτύου για μία συγκεκριμένη παρατήρηση. Είναι μια σχετικά ευθεία διαδικασία, καθώς μπορεί να γίνει με χρήση της εξίσωσης για την έξοδο κάθε νευρώνα, η οποία είναι απλώς η εφαρμογή της συνάρτησης ενεργοποίησης στο γραμμικά σταθμισμένο άθροισμα των εισόδων του.

Το βήμα οπίσθιας διάδοσης έχει σχεδιαστεί για να τροποποιεί τα βάρη του δικτύου όταν η πραγματική έξοδος δεν ταιριάζει με την επιθυμητή. Αυτό το βήμα ξεκινά από το επίπεδο εξόδου, υπολογίζοντας το σφάλμα στους κόμβους εξόδου και επιτελώντας τις απαραίτητες ενημερώσεις στα βάρη των νευρώνων εξόδου. Στη συνέχεια, μετακινείται προς τα πίσω μέσω του δικτύου, ενημερώνοντας αντίστροφα τα βάρη κάθε κρυμμένου στρώματος έως ότου φτάσει στο πρώτο κρυφό επίπεδο.

Το δεύτερο στοιχείο είναι ότι η ενημέρωση των βαρών των νευρώνων στο κρυφό στρώμα είναι ουσιαστικά πιο δύσκολη από την ενημέρωση των βαρών στο επίπεδο εξόδου. Δηλαδή, όταν θέλουμε να ενημερώσουμε τα βάρη των νευρώνων στο επίπεδο εξόδου, γνωρίζουμε ακριβώς ποια θα είναι η επιθυμητή έξοδος για τον συγκεκριμένο νευρώνα για μία δεδομένη είσοδο. Αυτό συμβαίνει επειδή οι επιθυμητές έξοδοι των νευρώνων εξόδου είναι οι έξοδοι του ίδιου του δικτύου, οι οποίες είναι διαθέσιμες σε εμάς στο σετ δεδομένων εκπαίδευσης. Αντιθέτως, δεν γνωρίζουμε πραγματικά ποια πρέπει να είναι η σωστή έξοδος ενός νευρώνα σε ένα κρυφό στρώμα για μία συγκεκριμένη είσοδο. Επιπλέον, αυτή η έξοδος κατανέμεται σε όλους τους νευρώνες του επόμενου στρώματος στο δίκτυο, συνεπώς επηρεάζει και όλες τις εξόδους τους.

Η βασική ιδέα εδώ είναι ότι το σφάλμα διαδίδεται από τους νευρώνες εξόδου πίσω στους νευρώνες στα κρυμμένα στρώματα. Αυτό το κάνουμε βρίσκοντας την κατάβαση κλίσης της συνάρτησης κόστους προκειμένου να ρυθμιστούν τα βάρη των νευρώνων προς την κατεύθυνση του μεγαλύτερου σφάλματος. Στη συνέχεια, με εφαρμογή του κανόνα διαφόρισης αλυσίδας, εκφράζεται η κλίση κατάβασης με τους όρους εξόδου του μεμονωμένου νευρώνα που μας ενδιαφέρει. Αυτή η διαδικασία εκφράζεται με μια γενική εξίσωση που είναι γνωστή ως **κανόνας δέλτα** (delta update rule):

$$w_{ji}^{(n)} \leftarrow w_{ji}^{(n)} + \eta \cdot \delta_j^{(n)} \cdot y_i^{(n)}$$

Αυτή η εξίσωση μας πληροφορεί με ποιο τρόπο να ενημερώσουμε το βάρος μεταξύ του νευρώνα *j* που βρίσκεται στο στρώμα l και του νευρώνα *i* που βρίσκεται στο στρώμα πριν από αυτόν (στρώμα 1-1). Το *n* υποδηλώνει τη νιοστή παρατήρηση στο σύνολο δεδομένων μας. Η παράμετρος «*η*» είναι ο ρυθμός εκμάθησης (learning rate), ενώ το δ_j αντιπροσωπεύει την τοπική κλίση (local gradient). Το σφάλμα του νευρώνα *j* συμβολίζεται με e_j και η κλίση της συνάρτησης ενεργοποίησης αντιπροσωπεύεται με το σύμβολο *g()*:

$$\delta_j = e_j \cdot g'(z_j)$$

Εδώ, υποδηλώνεται η έξοδος του νευρώνα *j* πριν εφαρμόσουμε τη συνάρτηση ενεργοποίησης του z_j ώστε να ισχύει η ακόλουθη σχέση:

$$y_i = g\left(z_j\right)$$

Τέλος, ο τρίτος όρος στον κανόνα ενημέρωσης δέλτα είναι η είσοδος στον νευρώνα *j* από τον νευρώνα *i*, που είναι απλώς η έξοδος του νευρώνα *i*, y_i. Όταν ο νευρώνας *j* είναι έξοδος νευρώνα, η τοπική κλίση δίνεται από την παρακάτω εξίσωση:

$$\delta_j = (t_j - y_j) \cdot y_j \cdot (1 - y_j)$$

Ο πρώτος όρος σε παρενθέσεις εκφράζει το σφάλμα του νευρώνα εξόδου, δηλαδή τη διαφορά μεταξύ της επιθυμητής εξόδου, t_j, και της πραγματικής εξόδου, y_j. Οι άλλοι δύο όροι προκύπτουν από τη διαφόριση της λογιστικής συνάρτησης ενεργοποίησης (logistic activation function). Όταν ο νευρώνας *j* είναι ένας νευρώνας κρυμμένου στρώματος, η κλίση της λογιστικής συνάρτησης ενεργοποίησης είναι η ίδια, αλλά το σφάλμα υπολογίζεται ως το σταθμισμένο άθροισμα των τοπικών κλίσεων *k* νευρώνων στο επόμενο στρώμα που λαμβάνουν είσοδο από τον νευρώνα *j*:

$$\delta_j = \left(\sum_k \delta_k \cdot w_{kj} \right) \cdot y_j \cdot \left(1 - y_j\right)$$

2.13 Ταξινομητής Μπέϋζ

Ο ταξινομητής Μπέϋζ είναι ένας δημοφιλής αλγόριθμος των προβλημάτων ταξινόμησης. Σε αυτή την περίπτωση η «απόφαση» ερμηνεύεται με πιθανολογικούς όρους και αναφερόμαστε σε στοχαστικά μοντέλα.

2.13.1 Το θεώρημα Μπέυζ

Ας υποθέσουμε ότι μας ενδιαφέρουν δύο γεγονότα, το Α και το Β. Για παράδειγμα, το γεγονός Α μπορεί αντιπροσωπεύει το γεγονός ότι ένας ασθενής έχει σκωληκοειδίτιδα και το γεγονός Β μπορεί να αντιπροσωπεύει το γεγονός ότι ο συγκεκριμένος ασθενής παρουσιάζει στη διαγνωστική εξέταση υψηλό αριθμό λευκών αιμοσφαιρίων.

Η **υπό συνθήκη πιθανότητα** (conditional probability) του ενδεχόμενου Α και δοθέντος του Β είναι, ουσιαστικά, η πιθανότητα να συμβεί το γεγονός Α, όταν γνωρίζουμε ότι το γεγονός Β έχει ήδη συμβεί. Τυπικά, ορίζουμε την υπό συνθήκη πιθανότητα του συμβάντος Α δεδομένου του γεγονότος Β ως την τομή της πιθανότητας να συμβούν και τα δύο συμβάντα (Α τομή Β) ή (Α και Β) διαιρούμενα με την πιθανότητα, να συμβεί το συμβάν Β:

$$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P}(\mathrm{A} \cap \mathrm{B})}{\mathrm{P}(\mathrm{B})}$$

Να σημειωθεί ότι αυτή η θεώρηση είναι σύμφωνη με τον τρόπο που ορίζεται και η στατιστική ανεξαρτησία. Η στατιστική ανεξαρτησία εμφανίζεται, όταν η υπό συνθήκη πιθανότητα δυο ενδεχομένων είναι απλά το γινόμενο των επιμέρους πιθανοτήτων αυτών των δυο ενδεχομένων. Με αυτό το σκεπτικό, η προηγούμενη εξίσωση μετασχηματίζεται:

$$P_{ανεξάρτητωνενδεχόμενων}(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{P}(\mathrm{A}) \cdot \mathrm{P}(\mathrm{B})}{\mathrm{P}(\mathrm{B})} = \mathrm{P}(\mathrm{A})$$

Δηλαδή, γνωρίζουμε ότι δύο γεγονότα είναι ανεξάρτητα το ένα από το άλλο ξέροντας ότι, αν έχει ήδη συμβεί το ενδεχόμενο Β, δεν αλλάζει η πιθανότητα το ενδεχόμενο Α να εμφανιστεί. Στη βιβλιογραφία αυτή η ιδιότητα είναι γνωστή ως «αντιστροφή της θέσης των ενδεχομένων». Η παραπάνω εξίσωση μπορεί να μετασχηματιστεί και να πάρουμε την παρακάτω μορφή:

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

Έτσι, τελικά με αντικατάσταση στην πρώτη εξίσωση καταλήγουμε στην **τελική μορφή του θεωρήματος Μπέυζ**:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Στην προηγούμενη εξίσωση, το P (A) αναφέρεται ως η **εκ των προτέρων πιθανότητα** (prior probability) του συμβάντος Α, καθώς αντιπροσωπεύει την πιθανότητα να συμβεί το ενδεχόμενο Α, πριν από οποιαδήποτε προσθήκη νέας πληροφορίας P (A | B) στην εξίσωση, πριν δηλαδή την εκτέλεση του πειράματος.

Ανατρέχοντας στο προηγούμενο παράδειγμα, αν υποθέσουμε ότι με επιστημονικά ερευνητικά δεδομένα ότί 1 στις 100 γυναίκες παρουσιάζουν καρκίνο του μαστού (0.01), και ένα τεστ ελέγχου παρουσιάζει 11% (0.11) ψευδός θετικά αποτελέσματα (11% από τις γυναίκες που ελέγχονται έχουν διάγνωση καρκίνου χωρίς να υπάρχει) και 10% (0.1) ψευδός αρνητικά (10% από τις γυναίκες που έχουν καρκίνο το τεστ δεν τον ανιχνεύει). Τότε οι πιθανότητες η γυναίκα να έχει θετικά αποτελέσματα (διάγνωση καρκίνου του μαστού) με μια πρώτη εκτίμηση θα ήταν 90% (με περιθώριο λάθους 10%). Αυτή όμως είναι μια λάθος εκτίμηση της πραγματικότητας. Σύμφωνα με τη Μπεϋζιανή προσέγγιση η πιθανότητα να έχει καρκίνο δεδομένου του θετικού τεστ θα είναι

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} = \frac{0.9 \cdot 0.01}{0.11} = 0.82 \ (8.2\%)$$

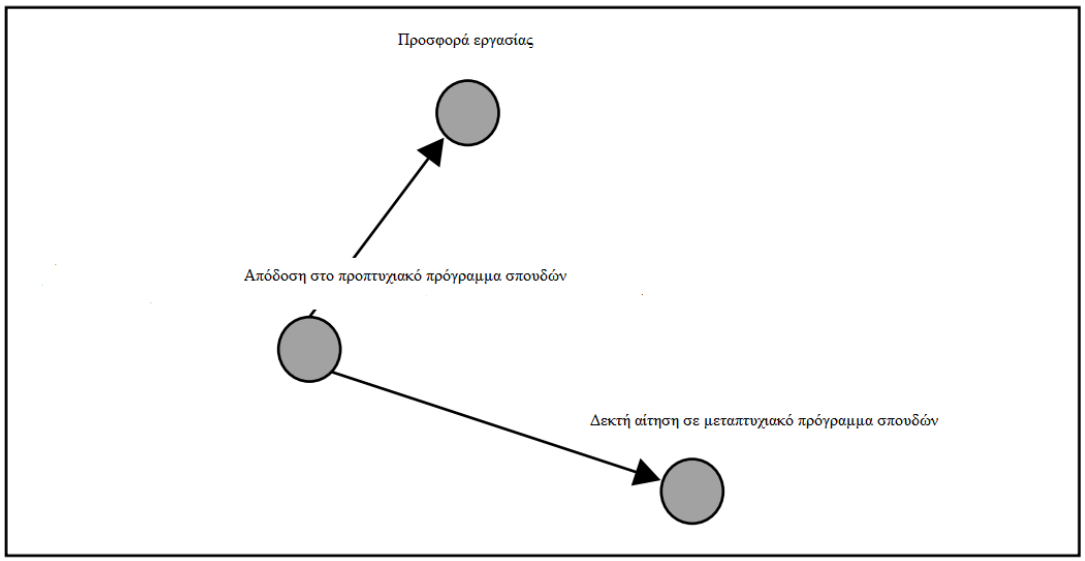
Δηλαδή η πιθανότητα η γυναίκα να έχει καρκίνο είναι μόλις 8.2% και όχι 90% που αρχικά είχε υπολογιστεί.
Θα πρέπει δηλαδή να ανατρέξουμε σε αρχεία περιπτώσεων παρελθόντων ασθενών, που παρουσίασαν τη νόσο και να προσθέσουμε πληροφορία στην εξίσωση. Το θεώρημα του Μπέυζ είναι πολύ σημαντικό για προβλέψεις, επειδή μας επιτρέπει να εκτιμήσουμε την αιτία παρατηρώντας το αποτέλεσμα.

2.13.2 Υπό όρους ανεξαρτησία

Γνωρίζουμε από τη Στατιστική ότι, όταν αναφερόμαστε στην έννοια της *στατιστικής ανεξαρτησίας*, εννοούμε την πιθανότητα δύο τυχαίων μεταβλητών, A και B να εκφράζονται απλά από το γινόμενο των ενδεχομένων τους. Μερικές φορές δύο μεταβλητές που ενδέχεται να μην είναι στατιστικά ανεξάρτητες μεταξύ τους, αλλά συγκρίσιμες με μια τρίτη μεταβλητή, τη C, μπορεί να έχουν ως αποτέλεσμα να γίνουν εν τέλει ανεξάρτητες η μια από την άλλη. Εν ολίγοις λέμε ότι τα γεγονότα A και B είναι υπό όρους ανεξάρτητα δοθέντος του C και μπορούμε να το εκφράσουμε ως εξής:

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B} \mid C) = \mathrm{P}(\mathrm{A} \mid C) \cdot \mathrm{P}(\mathrm{B} \mid C)$$

Για παράδειγμα, ας υποθέσουμε ότι το J αντιπροσωπεύει την πιθανότητα κάποιος να δεχθεί πρόταση εργασίας και το G αντιπροσωπεύει την πιθανότητα να γίνει δεκτός σε κάποιο μεταπτυχιακό πρόγραμμα σπουδών σε ένα συγκεκριμένο Πανεπιστήμιο. Και τα δύο μπορεί να εξαρτώνται από μια μεταβλητή U, που εκφράζει την απόδοση του ατόμου στο προπτυχιακό πρόγραμμα. Αυτό μπορεί να συνοψιστεί σε γράφημα ως εξής:



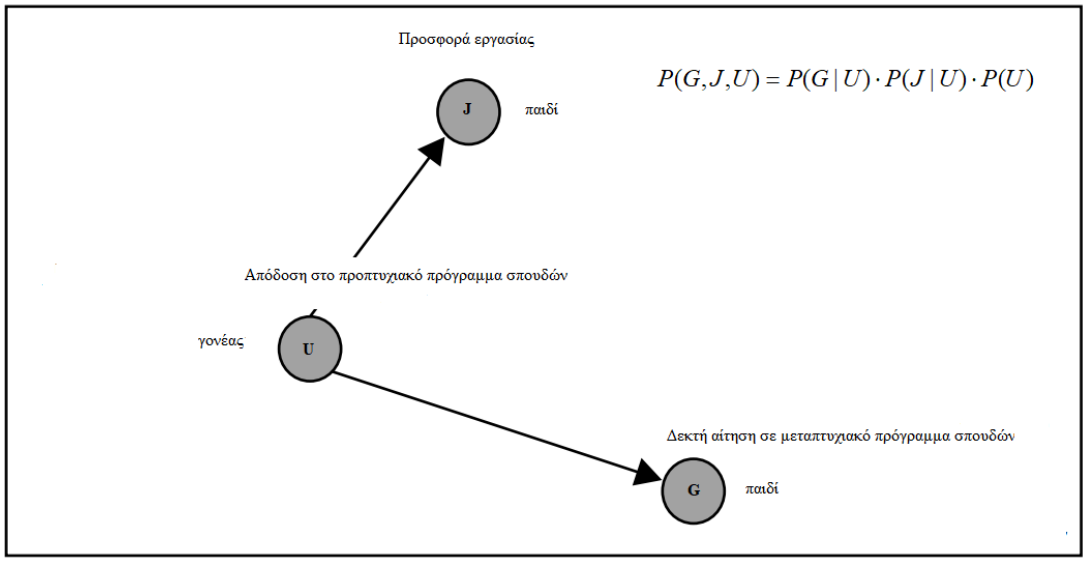
Εικόνα 14 Γραφική απεικόνιση παραδείγματος με τρεις ανεξάρτητες μεταβλητές

Όταν δεν γνωρίζουμε το U, την απόδοση ενός ατόμου στο προπτυχιακό του, γνωρίζοντας, όμως, ότι έγινε δεκτός στο μεταπτυχιακό, μπορεί να αυξήσει την πεποίθησή μας ότι το συγκεκριμένο πρόσωπο έχει αυξημένη πιθανότητα να βρει εργασία. Αυτό συμβαίνει, επειδή έχουμε την τάση να πιστεύουμε πως, επειδή παρουσίασε υψηλή απόδοση στη βαθμολογία του προπτυχιακού του προγράμματος, αυξάνονται και οι πιθανότητες ενός ατόμου να απορροφηθεί από την αγορά εργασίας. Έτσι, τα δύο γεγονότα J και G δεν είναι ανεξάρτητα του U.

Εάν πληροφορηθούμε για την απόδοση ενός ατόμου στο προπτυχιακό του, μπορεί να υποθέσουμε ότι η πιθανότητα του ατόμου αυτού να δεχθεί προσφορά για μια θέση εργασίας, ενδέχεται να είναι *ανεξάρτητη* από την πιθανότητά του να εισαχθεί στο πρόγραμμα μεταπτυχιακών σπουδών. Αυτό οφείλεται σε άλλους παράγοντες που μπορεί να επηρεάσουν την κατάσταση, όπως η συνέντευξη εργασίας του ατόμου σε μια συγκεκριμένη ημέρα ή η ποιότητα των άλλων δυνητικών υποψηφίων για τη δουλειά, δηλαδή, παράμετροι οι οποίες δεν επηρεάζονται από το αίτημα για μεταπτυχιακές σπουδές.

2.13.3 Μπεϋζιανά δίκτυα

Τα δίκτυα Μπέυζ είναι ένας τύπος γράφου που περιλαμβάνει ένα κατευθυνόμενο ακυκλικό γράφημα. Αναφερόμαστε συχνά στον κόμβο της ουράς μιας κατευθυνόμενης ακμής σε ένα γραφικό μοντέλο ως «**γονέας**» (parent) και στον κόμβο της κεφαλής ως «**παιδί**» (child) ή «**απόγονος**» (descendant). Στην πραγματικότητα, γενικεύουμε αυτήν την τελευταία έννοια, έτσι, ώστε, εάν υπάρχει μια διαδρομή από τον κόμβο Α στον κόμβο Β στο μοντέλο, ο κόμβος Β είναι «απόγονος» του κόμβου Α. Μπορούμε να διακρίνουμε την ειδική περίπτωση αυτή, αναφέροντας ότι ο κόμβος Α συνδέεται με τον κόμβο Β λέγοντας ότι ο τελευταίος είναι «**άμεσος απόγονος**» (direct descendant).



Εικόνα 15 Απλό Μπευζιανό δίκτυο με βάση το προηγούμενο παράδειγμα

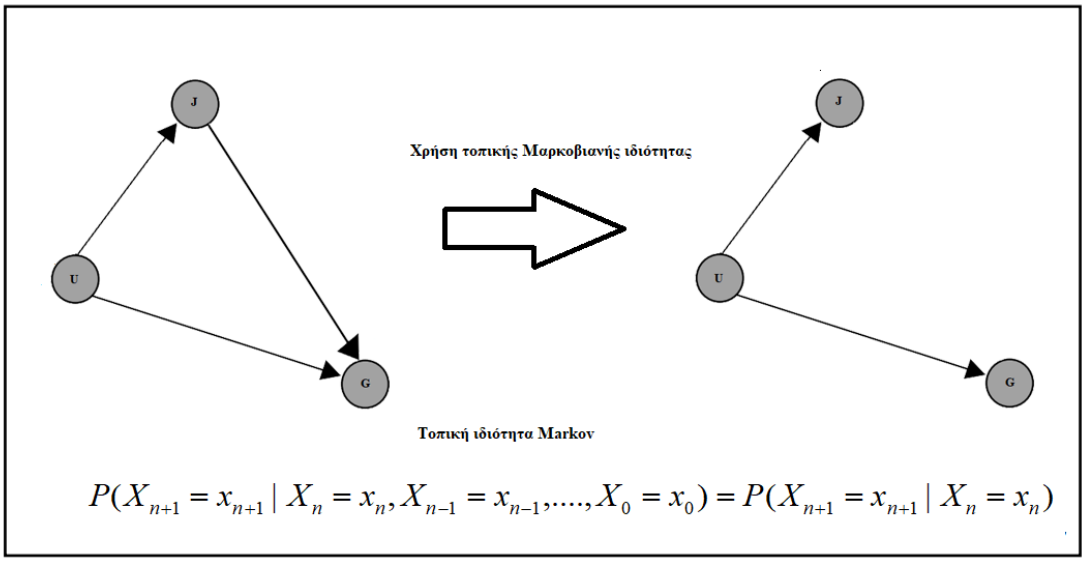
Η σχέση «γονέα» και «απογόνων» είναι αλληλοαποκλειόμενη σε ένα Μπεϋζιανό δίκτυο, επειδή αναφερόμαστε σε ακυκλικό γράφο. Τα δίκτυα Μπέυζ έχουν την ιδιότητα, ότι δοθέντων των «γονέων» κάθε κόμβος στο δίκτυο είναι υπό όρους ανεξάρτητος όλων των άλλων κόμβων του δικτύου αυτού που δεν είναι «απόγονοί» του. Αυτή η ιδιότητα αναφέρεται ως **τοπική Μαρκοβιανή ιδιότητα** (local Markov property). Είναι μια σημαντική ιδιότητα, γιατί σημαίνει ότι μπορούμε εύκολα να παραγοντοποιήσουμε τη συνάρτηση πιθανότητας από κοινού όλων των τυχαίων μεταβλητών στο μοντέλο λαμβάνοντας απλά υπόψιν τις αιχμές σε ένα γράφημα.

Για να καταλάβουμε πώς λειτουργεί αυτή η αρχή θα λάβουμε τις μεταβλητές από το προηγούμενο παράδειγμα G, J και U:

$$P(G,J,U) = P(G \mid J,U) \cdot P(J,U) = P(G \mid J,U) \cdot P(J \mid U) \cdot P(U)$$

Αυτό είναι στην πραγματικότητα ένα απλό Μπευζιανό δίκτυο, όπου οι G και J έχουν U ως «γονέα». Χρησιμοποιώντας την τοπική Μαρκοβιανή ιδιότητα μπορούμε να απλοποιήσουμε την εξίσωση για την κοινή κατανομή πιθανότητας ως εξής:

$$P(G,J,U) = P(G \mid U) \cdot P(J \mid U) \cdot P(U)$$

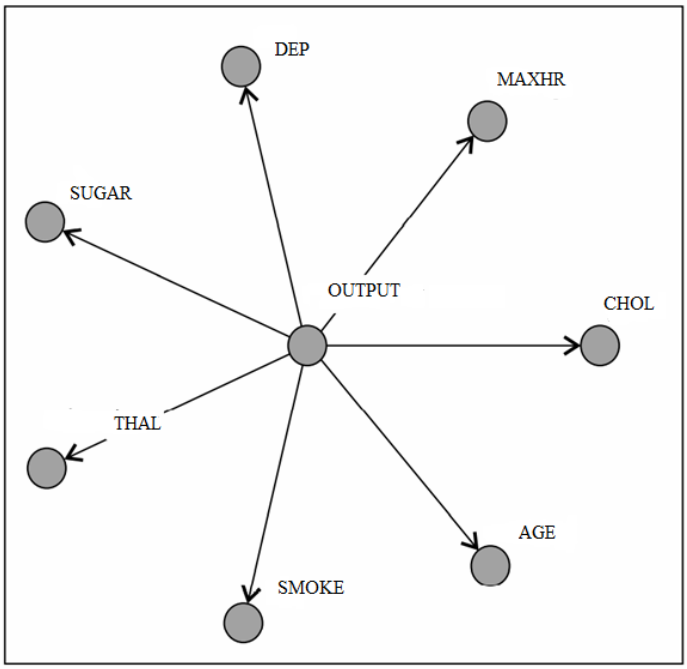


Εικόνα 16 Αριστερά το αρχικό Μπευζιανό δίκτυο και δεξιά το απλοποιημένο μετά την εφαρμογή της τοπικής ιδιότητας Μαρκόβ

Η ιδιότητα παραγοντοποίησης μιας κατανομής πιθανότητας με αυτόν τον τρόπο είναι χρήσιμη, καθώς απλοποιεί τους υπολογισμούς που πρέπει να κάνουμε. Δηλαδή, καταλήγουμε σε μια απλούστερη και ευέλικτη μορφή.

2.13.4 Ο απλός ταξινομητής Μπέυζ

Ο απλός ταξινομητής Μπέυζ (Naïve Bayes classifier) είναι ένα κατευθυνόμενο γραφικό μοντέλο που περιέχει έναν κόμβο «γονέα» και μια σειρά θυγατρικών κόμβων «παιδιών» που αντιπροσωπεύονται από τυχαίες μεταβλητές και εξαρτώνται μόνο από αυτόν τον κόμβο. Αναφέρουμε παράδειγμα:



Εικόνα 17 Παράδειγμα ενός δικτύου Μπέυζ

Συνήθως ερμηνεύουμε το κόμβο «γονέα» ως αιτιώδη κόμβο (causal node), έτσι και στο δικό μας παράδειγμα η τιμή του κόμβου OUTPUT θα επηρεάσει την τιμή του κόμβου SMOKE, του κόμβου THAL και ούτω καθεξής. Δεδομένου ότι αυτό είναι ένα Μπεϋζιανό δίκτυο, χρησιμοποιείται η τοπική ιδιότητα Μαρκόβ προκειμένου να ερμηνευτεί η βασική παραδοχή του μοντέλου.

Στην πράξη, χρησιμοποιούμε τον απλό ταξινομητή Μπέυζ σε ένα πλαίσιο, όπου μπορούμε να παρατηρήσουμε και να μετρήσουμε τις τιμές των θυγατρικών κόμβων, υπολογίζοντας τον κόμβο «γονέα» ως έξοδο. Κατά συνέπεια, οι θυγατρικοί κόμβοι θα είναι οι χαρακτηριστικές μεταβλητές εισόδου του μοντέλου και ο κόμβος «γονέα» θα αποτελεί τη μεταβλητή εξόδου. Για παράδειγμα, οι κόμβοι «παιδί» μπορεί να αντιπροσωπεύουν διάφορα ιατρικά συμπτώματα και ο κόμβος «γονέα» μπορεί να φανερώνει αν υπάρχει μια συγκεκριμένη ασθένεια.

Για να καταλάβουμε πώς λειτουργεί το μοντέλο στη πράξη, κάνουμε χρήση του θεωρήματος Μπέυκαι θεωρούμε, όπου C τον γονικό κόμβο και όπου F_i το σύνολο των χαρακτηριστικών μεταβλητών εισόδου:

$$P(C \mid F_1,....,F_n) = \frac{P(C) \cdot P(F_1,....,F_n \mid C)}{P(F_1,.....,F_n)}$$

Η οποία μπορεί να απλοποιηθεί περαιτέρω:

$$P(C \mid F_1,....,F_n) = \frac{P(C) \cdot P(F_1 \mid C) \cdot ....\cdot P(F_n \mid C)}{P(F_1,.....,F_n)}$$

Προκειμένου να πραγματοποιηθεί ταξινόμηση για αυτό το μοντέλο πιθανοτήτων, ο στόχος είναι να επιλέξουμε μια τιμή για την κλάση C_i που μεγιστοποιεί την **εκ των υστέρων πιθανότητα** (posterior probability) P ($C_i|F_{1...}F_n$) με βάση τα παρατηρούμενα χαρακτηριστικά. Ο παρονομαστής αντιπροσωπεύει την **από κοινού κατανομή πιθανότητας** (joint probability) των παρατηρούμενων χαρακτηριστικών.

$$|\ \mathrm{T}\alpha\xi\iota\nu\acute{o}\mu\eta\sigma\varepsilon\ C_i = \arg\max_c P(C)\cdot\prod_{i=1}^{n} P(F_i \mid C)$$

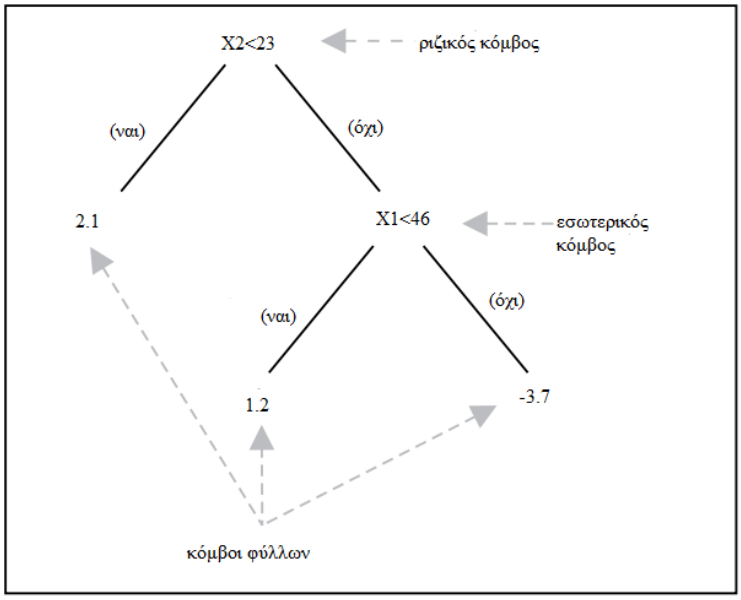
Αν δοθούν δεδομένα, μπορούμε να εκτιμήσουμε τις πιθανότητες P ($F_i|\ C_j$), για όλες τις διαφορετικές τιμές της μεταβλητής F_i ως τη σχετική αναλογία (relative proportion) των παρατηρήσεων της κλάσης C_j. Μπορούμε, επίσης, να θεωρήσουμε το P (C_j) ως τη σχετική αναλογία των παρατηρήσεων που ανατίθενται στην κλάση C_j. Αυτές είναι οι εκτιμήσεις μέγιστης Πιθανοφάνειας (maximum likelihood estimates).

2.14 Δέντρα απόφασης

2.14.1 Περιγραφή των Δέντρων Απόφασης

Το δέντρο αποφάσεων πρόκειται για ένα μοντέλο με πολύ απλή δομή που μας επιτρέπει να προβλέψουμε μια μεταβλητή εξόδου, βασισμένη σε σειρά κανόνων με δομή που μοιάζει με δέντρο. Η μεταβλητή εξόδου που μπορούμε να μοντελοποιήσουμε ενδέχεται να είναι κατηγορηματική, επιτρέποντας να χρησιμοποιήσουμε ένα δέντρο αποφάσεων για να διαχειριστούμε προβλήματα ταξινόμησης. Ομοίως, μπορούμε να χρησιμοποιήσουμε δέντρα αποφάσεων για να προβλέψουμε μια αριθμητική έξοδο. Με αυτό τον τρόπο θα αντιμετωπίσουμε επίσης προβλήματα που η προγνωστική εργασία τους βασίζεται στην παλινδρόμηση.

Τα δέντρα αποφάσεων αποτελούνται από μία σειρά διαχωρισμένων σημείων, τα οποία συχνά αναφέρονται ως **κόμβοι** (nodes). Για να κάνουμε πρόβλεψη χρησιμοποιώντας ένα δέντρο αποφάσεων ξεκινάμε στην κορυφή του δέντρου σ' έναν μόνο κόμβο, που είναι γνωστός ως **ριζικός κόμβος** (root node). Ο ριζικός κόμβος αντιπροσωπεύει μία απόφαση ή ένα **σημείο διάσπασης** (split point), γιατί αυτό θέτει μία συνθήκη ως προς την τιμή ενός εκ των χαρακτηριστικών εισόδου. Με βάση αυτή την απόφαση γνωρίζουμε αν θα συνεχίσουμε με το αριστερό ή με το δεξιό μέρος του δέντρου. Επαναλαμβάνουμε τη διαδικασία επιλογής, πηγαίνοντας αριστερά ή δεξιά σε κάθε εσωτερικό κόμβο που συναντάμε, μέχρι να φτάσουμε σε έναν από τους **κόμβους φύλλων** (leaf nodes). Πρόκειται για τους κόμβους στη βάση του δέντρου, που μας δίνουν μία συγκεκριμένη τιμή εξόδου η οποία αποτελεί την πρόβλεψή μας.

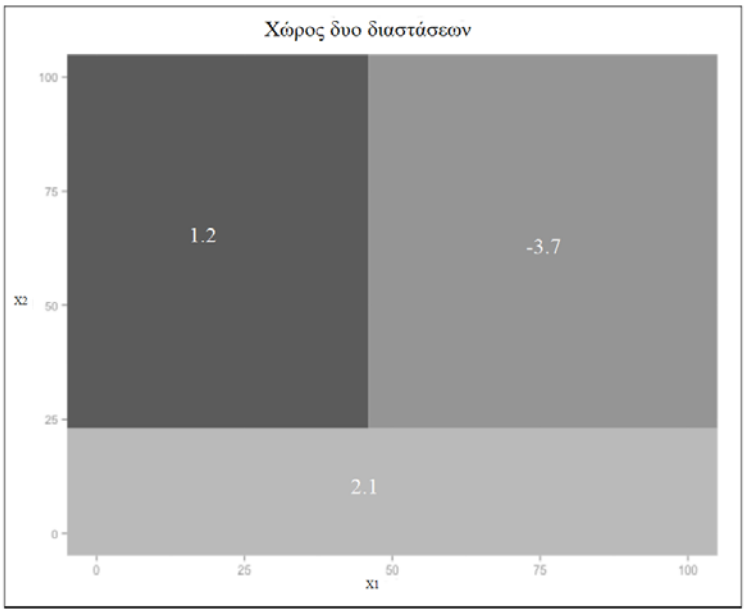


Εικόνα 18 Ένα απλό παράδειγμα δέντρου απόφασης με δυο μεταβλητές X1 και X2

Σημειώνεται ότι το δέντρο είναι μία αναδρομική δομή, δηλαδή το αριστερό και το δεξιό μέρος του δέντρου που βρίσκονται κάτω από έναν συγκεκριμένο κόμβο αποτελούν και αυτά δέντρα. Αναφέρονται ως **αριστερό υποδέντρο** (left subtree) και **δεξιό υποδέντρο** (right subtree), αντίστοιχα, και οι κόμβοι τους οποίους οδηγούν ονομάζονται **αριστερό παιδί** (left child) και **δεξιό παιδί** (right child). Για να καταλάβουμε πρακτικά πώς χρησιμοποιούμε ένα δέντρο αποφάσεων, μπορούμε να δοκιμάσουμε ένα απλό παράδειγμα. Ας υποθέσουμε ότι θέλουμε να χρησιμοποιήσουμε το δέντρο μας για να προβλέψουμε την έξοδο για μία παρατήρηση όπου η τιμή του *x1* είναι *96,0* και η τιμή του *x2* είναι *79,9*. Ξεκινάμε από τη ρίζα και αποφασίζουμε ποιο υποδέντρο θα ακολουθήσουμε. Η τιμή του *x2* είναι μεγαλύτερη από *23*, οπότε ακολουθούμε τον σωστό κλάδο και ερχόμαστε σε έναν νέο κόμβο με μια νέα προϋπόθεση για έλεγχο. Η τιμή του *x1* είναι μεγαλύτερη από *46*, οπότε παίρνουμε για άλλη μια φορά τον σωστό κλάδο και φτάνουμε σε έναν κόμβο φύλλων. Έτσι, η έξοδος που υποδεικνύεται από το φύλλο κόμβο είναι *-3,7*. Η τιμή αυτή προβλέπει το μοντέλο μας με βάση το ζεύγος εισόδων που καθορίσαμε. Ένας τρόπος για να αντιληφθούμε τα δέντρα αποφάσεων είναι ότι στην πραγματικότητα κωδικοποιούν μία σειρά κανόνων *if-then* που οδηγούν σε ξεχωριστές εξόδους. Μπορούμε να εξαγάγουμε όλους αυτούς τους κανόνες *if-then* ξεκινώντας από τον ριζικό κόμβο και ακολουθώντας κάθε διαδρομή κάτω από το δέντρο που οδηγεί σε έναν κόμβο φύλλων. Για παράδειγμα, τα μικρά δέντρα παλινδρόμησης μας οδηγούν στους ακόλουθους τρεις κανόνες, έναν για κάθε κόμβο φύλλου:

```
If (x2 < 23) Then Output 2.1
If (x2 > 23) AND (x1 < 46) Then Output 1.2
If (x2 > 23) AND (x1 > 46) Then Output -3.7
```

Ένας άλλος τρόπος για να σκεφτούμε τα δέντρα αποφάσεων είναι ότι «διαμερισματοποιούν» έναν **χώρο χαρακτηριστικών** (feature space), σε ορθογώνιες περιοχές στο επίπεδο των δύο διαστάσεων, σε κύβους στο επίπεδο των τριών διαστάσεων, και σε υπερκύβους σε υψηλότερες διαστάσεις. Ο χώρος χαρακτηριστικών για το παράδειγμά μας περιλαμβάνει δέντρο παλινδρόμησης με δύο διαστάσεις, συνεπώς παρατηρούμε ότι ο χώρος χωρίζεται σε ορθογώνιες περιοχές:



Εικόνα 19 Απεικόνιση προηγούμενου παραδείγματος σε χώρο δυο διαστάσεων

2.14.2 Ο αλγόριθμος CART

Ο αλγόριθμος CART είναι ένας αλγόριθμος ταξινόμησης για τη δημιουργία δέντρων απόφασης, του οποίου το κριτήριο διάσπασης βασίζεται σ' έναν **συντελεστή καθαρότητας** (purity index) που ονομάζεται **δείκτης Gini**. Τα δέντρα απόφασης CART είναι δυαδικά δέντρα που περιλαμβάνουν σημείο διάσπασης με δύο «κόμβους παιδί» επαναλαμβανόμενα. Για μία μεταβλητή εξόδου *K* διαφορετικών κλάσεων, ο συντελεστής Gini ορίζεται ως:

$$\Delta\text{είκτης } Gini = \sum_{k=1}^{K} \hat{p}_k \cdot \left(1 - \hat{p}_k\right)$$

Για τη μέτρηση του δείκτη Gini υπολογίζουμε την πιθανότητα κάθε κλάσης και την πολλαπλασιάζουμε με την πιθανότητα να μην ανήκει σ' αυτή την κλάση. Στη συνέχεια αθροίζουμε όλα τα γινόμενα. Για ένα πρόβλημα δυαδικής ταξινόμησης είναι εύκολο να αποδειχθεί ότι ο δείκτης Gini ισούται με *2 pˆ (1- pˆ)*, όπου *pˆ* είναι η εκτιμώμενη πιθανότητα μίας εκ των κλάσεων.

Για να υπολογίσουμε τον δείκτη Gini σε έναν συγκεκριμένο κόμβο σε ένα δέντρο, μπορούμε απλώς να χρησιμοποιήσουμε την αναλογία του αριθμού των σημείων δεδομένων –που χαρακτηρίζονται ως κλάση *K* επί του συνολικού αριθμού των σημείων δεδομένων– ως εκτίμηση για την πιθανότητα ενός σημείου δεδομένων να ανήκει στην κλάση *K* στον κόμβο.

Ακολουθεί μία απλή συνάρτηση *R* για τον υπολογισμό του ευρετηρίου Gini:

```
gini_index <- function(v) {
 t <- table(v)
 probs <- t / sum(t)
 terms <- sapply(probs, function(p) p * (1 - p) )
 return(sum(terms))}
```

Για οποιονδήποτε δεδομένο κόμβο στο δέντρο, συμπεριλαμβανομένου του **ριζικού κόμβου** (root node), ξεκινάμε με την εκχώρηση ορισμένων σημείων δεδομένων σε αυτό τον κόμβο. Στον ριζικό κόμβο εκχωρούνται όλα τα σημεία δεδομένων, αλλά μόλις κάνουμε έναν διαχωρισμό, ορισμένα από τα σημεία δεδομένων ανατίθενται σε κόμβο «αριστερού παιδιού» και τα υπόλοιπα σημεία σε κόμβο «δεξιού παιδιού». Η αρχική τιμή του **μέσου τετράγωνου σφάλματος** SSE είναι το άθροισμα τετραγώνου του σφάλματος που υπολογίζεται χρησιμοποιώντας τη μέση τιμή $\bar{y}$ της μεταβλητής εξόδου y_i για n σημεία δεδομένων που έχουν εκχωρηθεί στον τρέχοντα κόμβο:

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Εάν χωρίσουμε αυτά τα σημεία δεδομένων σε δύο ομάδες μεγέθους n_1 και n_2 ώστε $n_1 + n_2 = n$ και υπολογίσουμε το νέο SSE για όλα τα σημεία δεδομένων ως το άθροισμα των τιμών SSE για καθένα από τις δύο νέες ομάδες, έχουμε:

$$SSE = \sum_{j=1}^{n_1} (y_j - \bar{y}_1)^2 + \sum_{k=1}^{n_2} (y_k - \bar{y}_2)^2$$

Εδώ, το πρώτο άθροισμα από τα αριστερά αντιστοιχεί στο αριστερό παιδί και το δεύτερο άθροισμα στο δεξιό παιδί. Η ιδέα πίσω από το CART είναι ότι βρίσκουμε έναν τρόπο να σχηματίσουμε αυτές τις δύο ομάδες δεδομένων εξετάζοντας κάθε πιθανό χαρακτηριστικό και κάθε πιθανό σημείο διάσπασης, ώστε να ελαχιστοποιείται αυτή η νέα ποσότητα. Έτσι, μπορούμε να σκεφτούμε τη συνάρτηση σφάλματος στο CART ως το SSE.

Η μεθοδολογία CART εφαρμόζει πάντα την ίδια λογική για να καθορίσει εάν θα κάνουμε νέο διαχωρισμό σε κάθε κόμβο καθώς και πώς θα επιλέξουμε ποια μεταβλητή και ποια τιμή θα χωριστεί. Αυτή η αναδρομική προσέγγιση του διαχωρισμού των σημείων δεδομένων σε κάθε κόμβο για την κατασκευή του δέντρου απόφασης είναι ο λόγος για τον οποίο αυτή η διαδικασία είναι επίσης γνωστή ως **αναδρομική κατάτμηση** (recursive partitioning).

Αν επρόκειτο να επιτρέψουμε να συντελείται η επανάληψη της αναδρομικής κατάτμησης αόριστα, η διαδικασία θα τερματιζόταν τελικά έχοντας κόμβους φύλλων σε ένα μόνο σημείο δεδομένων, γιατί τότε δεν θα μπορούσαμε να χωρίσουμε τα δεδομένα περαιτέρω. Αυτό το μοντέλο θα ταίριαζε τα δεδομένα κατάρτισης σε μέγιστο βαθμό, αλλά είναι πολύ απίθανο να είχε υψηλή απόδοση σε νέα δεδομένα. Συνεπώς, τα μοντέλα που βασίζονται σε δέντρα είναι ευαίσθητα στην **υπερπροσαρμογή**. Για να το καταπολεμήσουμε αυτό πρέπει να ελέγξουμε το βάθος του τελικού δέντρου αποφάσεων.

Η διαδικασία αφαίρεσης κόμβων από το δέντρο για τον περιορισμό του μεγέθους και της πολυπλοκότητάς του είναι γνωστή ως **κλάδεμα** (pruning). Μια πιθανή μέθοδος κλαδέματος είναι να επιβληθεί ένα όριο για τον μικρότερο αριθμό σημείων δεδομένων που μπορούν να χρησιμοποιηθούν, ώστε να δημιουργηθεί μια νέα διάσπαση στο δέντρο αντί για έναν κόμβο φύλλων. Αυτό θα δημιουργήσει κόμβους φύλλων νωρίτερα στη διαδικασία και τα σημεία δεδομένων που τους έχουν εκχωρηθεί ενδέχεται να μην έχουν όλα την ίδια έξοδο. Σε μια τέτοια περίπτωση μπορούμε απλώς να προβλέψουμε τη μέση τιμή για παλινδρόμηση (και την πιο δημοφιλή κατηγορία για ταξινόμηση). Η μέθοδος αυτή ονομάζεται **προ-κλάδεμα** (pre-pruning).

Μπορούμε να δούμε ότι όσο μεγαλύτερο είναι το βάθος του δέντρου και όσο μικρότερος ο μέσος αριθμός σημείων δεδομένων που εκχωρούνται στους κόμβους των φύλλων, τόσο μεγαλύτερος είναι ο βαθμός υπερπροσαρμογής. Φυσικά, εάν έχουμε λιγότερους κόμβους στο δέντρο, πιθανότατα δεν είμαστε αρκετά ακριβείς στη μοντελοποίηση των δεδομένων. Το ερώτημα *πόσο μεγάλο δέντρο πρέπει να επιτραπεί να αναπτυχθεί* έχει ουσιαστικά να κάνει με τον τρόπο που μοντελοποιούμε τα δεδομένα μας, ελέγχοντας τον βαθμό υπερπροσαρμογής. Στην πράξη, η χρήση του προ-κλαδέματος δεν εφαρμόζεται εύκολα καθώς είναι δύσκολο να βρεθεί κατάλληλο κατώφλι.

Προκειμένου να ξεπεραστεί το πρόβλημα, μια άλλη διαδικασία κανονικοποίησης που χρησιμοποιείται από τη μεθοδολογία CART είναι γνωστή ως **ρύθμιση κόστους πολυπλοκότητας** (cost-complexity tuning). Στην πραγματικότητα, τα δέντρα συχνά αφήνονται να αναπτυχθούν χρησιμοποιώντας πλήρως την αναδρομική προσέγγιση καταμερισμού που περιγράφεται στην προηγούμενη ενότητα.

Μόλις αυτό ολοκληρωθεί, το δέντρο που προκύπτει κλαδεύεται, δηλαδή αρχίζουμε να αφαιρούμε διαχωρισμένα σημεία και να συγχωνεύουμε κόμβους φύλλων. Έτσι το δέντρο συρρικνώνεται σύμφωνα με ένα συγκεκριμένο κριτήριο. Αυτό είναι γνωστό ως **μετα-κλάδεμα** (post-pruning), καθώς κλαδεύουμε το δέντρο μετά την κατασκευή του. Όταν κατασκευάζουμε το αρχικό δέντρο, η συνάρτηση σφάλματος που χρησιμοποιούμε είναι η SSE.

Για τη διαδικασία κλαδέματος του δέντρου χρησιμοποιείται η έκδοση με ποινή (penalized version) της SSE, με στόχο την ελαχιστοποίηση του μέσου τετράγωνου σφάλματος:

$$SSE_{\,\mu\varepsilon\ \pi\omega\iota\nu\acute{\eta}} = SSE + c_p \cdot T_p$$

Εδώ, το c_p είναι η **παράμετρος πολυπλοκότητας** (complexity parameter), που ρυθμίζει τον βαθμό κανονικοποίησης, και το T_p αντιπροσωπεύει τον αριθμό των κόμβων στο δέντρο, που είναι ένας τρόπος για να μοντελοποιήσουμε το μέγεθος του δέντρου. Αυτή η μέθοδος είναι παρόμοια με τη LASSO παλινδρόμηση που εφαρμόσαμε στη Λογιστική Παλινδρόμηση, κατά την οποία περιορίζεται το μέγεθος των συντελεστών παλινδρόμησης. Σε αυτή τη διαδικασία κανονικοποίησης, αναλόγως, περιορίζεται το μέγεθος του δέντρου που προκύπτει. Μια πολύ μικρή τιμή του c_p έχει ως αποτέλεσμα μικρό βαθμό κλαδέματος. Από την άλλη πλευρά, αν χρησιμοποιηθεί υψηλή τιμή αυτής της παραμέτρου θα οδηγηθούμε σε ένα δέντρο μηδενικού μεγέθους, χωρίς καθόλου διασπάσεις, το οποίο προβλέπει μόνο τη μέση τιμή της εξόδου για όλες τις πιθανές εισόδους.

2.14.3 Αλγόριθμος C5.0

Ο Ross Quinlan [69] ανέπτυξε τον αλγόριθμο C5.0 για τη δόμηση ενός δέντρου απόφασης με σκοπό την ταξινόμηση. Αυτός ο αλγόριθμος είναι ο τελευταίος σε μια αλυσίδα διαδοχικά βελτιωμένων εκδόσεων που ξεκινούν από έναν αλγόριθμο γνωστό ως ID3, ο οποίος εξελίχθηκε σε C4.5 (και υλοποιήθηκε σε ανοιχτό κώδικα στη γλώσσα προγραμματισμού Java, γνωστό ως J48) πριν κορυφωθεί στο C5.0. Υπάρχουν πολλά ακρωνύμια που χρησιμοποιούνται για δέντρα αποφάσεων, αλλά ευτυχώς πολλά από αυτά σχετίζονται μεταξύ τους. Η αλυσίδα αλγορίθμων C5.0 έχει αρκετές διαφορές από τη μεθοδολογία CART, κυρίως στην επιλογή του **κριτηρίου διάσπασης** (splitting criterion) αλλά και στη **διαδικασία κλαδέματος** (pruning procedure). [68]

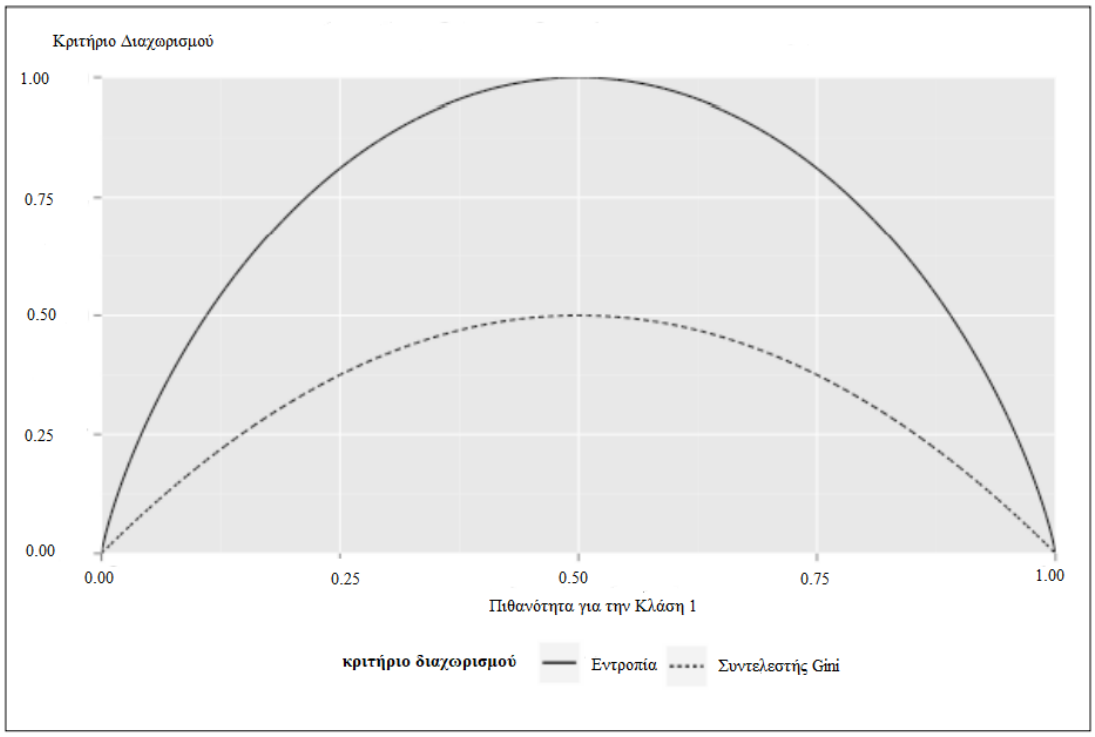
Το κριτήριο διάσπασης που χρησιμοποιείται στον αλγόριθμο C5.0 είναι γνωστό ως **εντροπία** (entropy) και έχει τις ρίζες της στη θεωρία επικοινωνιών. *Η εντροπία ορίζεται ως ο μέσος όρος δυαδικών ψηφίων (bits) που απαιτούνται για την επικοινωνία πληροφοριών μέσω μηνυμάτων η οποία εκφράζεται ως συνάρτηση των πιθανοτήτων των διαφορετικών κλάσεων που χρησιμοποιούνται.* Η εντροπία έχει επίσης ρίζες στη στατιστική φυσική, όπου χρησιμοποιείται για να αναπαραστήσει τον βαθμό του χάους και της αβεβαιότητας σε ένα σύστημα. Ο επίσημος ορισμός της εντροπίας σε bits για το σενάριο K πολλαπλών κλάσεων είναι:

$$\text{Εντροπία} = -\sum_{k=1}^{K} p_k \cdot \log_2 p_k$$

Όπου $p_1, p_2, \ldots p_i$ είναι οι πιθανότητες του κάθε ενδεχομένου που περιλαμβάνεται στο σύνολο.

Για δυαδικές περιπτώσεις, η εξίσωση απλοποιείται αρκετά (όπου το *p* αναφέρεται αυθαίρετα στην πιθανότητα μίας εκ των δύο κατηγοριών):

$$\text{Εντροπία Δυαδικής Κλάσης} = -[p \cdot \log_2 p + (1-p) \cdot \log_2 (1-p)]$$



Εικόνα 20 Κριτήριο διαχωρισμού για δυαδική ταξινόμηση.
Για x=50, δηλαδή για διαχωρισμό 50/50, έχουμε μέγιστη εντροπία

Από το γράφημα διαπιστώνουμε ότι και οι δύο συναρτήσεις έχουν το ίδιο γενικό σχήμα για ένα δυαδικό πρόβλημα κλάσης. Όσο χαμηλότερη είναι η εντροπία τόσο χαμηλότερη είναι η αβεβαιότητα που έχουμε για την κατανομή των κλάσεων. Ως εκ τούτου έχουμε υψηλότερη **καθαρότητα κόμβων** (node purity). Κατά συνέπεια, απαιτείται να ελαχιστοποιηθεί το κριτήριο της εντροπίας καθώς «χτίζουμε» το δέντρο μας. Στο ID3, το κριτήριο διάσπασης που χρησιμοποιείται είναι η σταθμισμένη μείωση εντροπίας, γνωστή και ως **κέρδος πληροφορίας** (information gain): [68]

$$\text{Κέρδος Πληροφορίας} = \text{Εντροπία} - \sum_{i=1}^{p} \frac{n_i}{n} \cdot \text{Εντροπία}_i$$

Όπου η *εντροπία* είναι η συνάρτηση εντροπίας, ενώ το *p* εκφράζει το πλήθος των τιμών n_i που παίρνει το *n* σε ένα σύνολο δειγμάτων. Ο λόγος n_i/n αντιπροσωπεύει το ποσοστό των δειγμάτων, ενώ η εντροπία_i αποτελεί το υποσύνολο του συνόλου δειγμάτων, όπου η τιμή του *n* είναι n_i.

Όπως διαφαίνεται, το κριτήριο λειτουργεί με μεροληψία (bias), καθώς τείνει να ευνοεί τις κατηγορικές μεταβλητές λόγω του μεγάλου αριθμού πιθανών ομαδοποιήσεων σε σύγκριση με το γραμμικό **εύρος διασπάσεων** (range of splits) που βρίσκουμε σε συνεχή χαρακτηριστικά. Προκειμένου να καταπολεμηθεί αυτό το φαινόμενο, από τον αλγόριθμο C4.5 και ύστερα υπήρξε μια βελτίωση του κριτηρίου *κέρδους πληροφορίας*, που ονομάστηκε **Λόγος Κέρδους Πληροφορίας** (Information Gain Ratio).

Πρόκειται για μία διαφορετική εκδοχή του *κέρδους πληροφορίας* σε σχέση με μια ποσότητα, γνωστή ως **Τιμή Διαχωρισμένης Πληροφορίας** (Split Information Value).[70] Αυτό με τη σειρά του αντιπροσωπεύει την πιθανή αύξηση των πληροφοριών που μπορούμε να πάρουμε μόνο από το μέγεθος των ίδιων των διαμερισμάτων. Η υψηλή τιμή διαχωρισμού πληροφορίας εμφανίζεται όταν έχουμε διαμερίσματα ομοιόμορφου μεγέθους και η χαμηλή τιμή εμφανίζεται όταν τα περισσότερα σημεία δεδομένων συγκεντρώνονται σε έναν μικρό αριθμό διαμερισμάτων. Συνοψίζοντας, έχουμε τις ακόλουθες σχέσεις:

$$\text{Λόγος Κέρδους Πληροφορίας} = \frac{\text{Κέρδος Πληροφορίας}}{\text{Τιμή Διαχωρισμένης Πληροφορίας}}$$

$$\text{Τιμή Διαχωρισμένης Πληροφορίας} = -\sum_{i-1}^{p} \frac{n_i}{n} \cdot \log_2\left(\frac{n_i}{n}\right)$$

Το C5.0, ειδικότερα, είναι ένας πολύ ισχυρός αλγόριθμος που περιέχει επίσης βελτιώσεις στην ταχύτητα, τη χρήση της μνήμης, καθώς και τη δυνατότητα καθορισμού ενός πίνακα κόστους, ώστε ο αλγόριθμος να αποφύγει ορισμένους τύπους λανθασμένων ταξινομήσεων έναντι άλλων, όπως ακριβώς είδαμε με τις μηχανές διανυσμάτων υποστήριξης.

3. Μεθοδολογία

Σε αυτό το σημείο θα αναλυθούν λεπτομερώς τα βήματα που ακολουθήθηκαν για την συγγραφή της μεταπτυχιακής εργασίας. Ο στόχος είναι η βελτιστοποίηση του υπο-μελέτη πλαισίου δεδομένων και των διαφόρων παραμέτρων των αλγορίθμων μηχανικής μάθησης, έτσι ώστε να βρεθεί ποιος από αυτούς έχει την υψηλότερη ακρίβεια, με σκοπό να υλοποιηθεί κατάλληλη εφαρμογή (λογισμικό) με προγνωστικό χαρακτήρα. Το εργαλείο της ερευνητικής μεθόδου είναι η γλώσσα προγραμματισμού R και συγκεκριμένα το προγραμματιστικό πακέτο caret που περιέχει σύνολο συναρτήσεων για τη δημιουργία προγνωστικών μοντέλων, όπως για παράδειγμα διάφοροι ταξινομητές, διαχωρισμός δεδομένων, επιλογή χαρακτηριστικών, προεπεξεργασία μοντέλου κ.α. Μέσα από αυτή την μεθοδολογία, θα αξιολογηθούν οι μέθοδοι και τα εργαλεία μηχανικής μάθησης που χρησιμοποιήθηκαν στην έρευνα.

3.1 Προετοιμασία Μοντέλου

Κατέβασμα πλαισίου δεδομένων από το αποθετήριο του UCI

```
library(data.table)
heart <- fread('http://archive.ics.uci.edu/ml/databases/statlog/heart/heart.dat')
```

Προβολή των 6 πρώτων στοιχείων του πλαισίου δεδομένων

```
head(heart)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
1	70	1	4	130	322	0	2	109	0	2.4	2	3	3	2
2	67	0	3	115	564	0	2	160	0	1.6	2	0	7	1
3	57	1	2	124	261	0	0	141	0	0.3	1	0	7	2
4	64	1	4	128	263	0	0	105	1	0.2	2	1	7	1
5	74	0	2	120	269	0	2	121	1	0.2	1	1	3	1
6	65	1	4	120	177	0	0	140	0	0.4	1	0	7	1

Εισαγωγή ετικετών σαν επικεφαλίδες του πλαισίου δεδομένων

```
names(heart) <- c("AGE", "SEX", "CHESTPAIN", "RESTBP", "CHOL",
 "SUGAR", "ECG", "MAXHR", "ANGINA", "DEP", "EXERCISE", "FLUOR",
 "THAL", "OUTPUT")
```

Μετατρέπουμε τις τιμές της μεταβλητής OUTPUT σε 0 (απουσία πάθησης) και 1 (παρουσία πάθησης), από 1 και 2 που ήταν αντίστοιχα, για να είναι πιο ευανάγνωστες .

```
heart$OUTPUT = heart$OUTPUT - 1
```

Τώρα μπορούμε να δούμε τις έξι πρώτες γραμμές του πλαισίου δεδομένων μας:

```
head(heart)
```

	AGE	SEX	CHESTPAIN	RESTBP	CHOL	SUGAR	ECG	MAXHR	ANGINA	DEP	EXERCISE	FLUOR	THAL	OUTPUT
1	70	1	4	130	322	0	2	109	0	2.4	2	3	3	1
2	67	0	3	115	564	0	2	160	0	1.6	2	0	7	0
3	57	1	2	124	261	0	0	141	0	0.3	1	0	7	1
4	64	1	4	128	263	0	0	105	1	0.2	2	1	7	0
5	74	0	2	120	269	0	2	121	1	0.2	1	1	3	0
6	65	1	4	120	177	0	0	140	0	0.4	1	0	7	0

Αποθήκευση πλαισίου δεδομένων

```
saveRDS(heart, file = "heart.rds")
```

Ανάκτηση πλαισίου δεδομένων

```
readRDS(file = "heart.rds")
```

Πριν εκπαιδεύσουμε το μοντέλο, χωρίζουμε τα δεδομένα σε δυο μέρη. Το ένα αφορά το πλαίσιο δεδομένων εκπαίδευσης και το άλλο, ελέγχου. Χρησιμοποιούμε τον κανόνα 85-15 (230 παρατηρήσεις στο σετ εκπαίδευσης και 40 παρατηρήσεις στο τεστ σετ. Για την τμηματοποίηση, θα χρειαστεί να

«φορτώσουμε» τη βιβλιοθήκη της R που ονομάζεται caret (http://topepo.github.io/caret/index.html).

Δημιουργία πλαισίων εκπαίδευσης και δοκιμής με κανόνα 85-15

```
install.packages("caret")
install.packages("ggplot2")
library(caret)
set.seed(987954)
heart_sampling_vector <- createDataPartition(heart$OUTPUT, p = 0.85, list = FALSE)
heart_train <- heart[heart_sampling_vector,]
heart_train_labels <- heart$OUTPUT[heart_sampling_vector]
heart_test <- heart[-heart_sampling_vector,]
heart_test_labels <- heart$OUTPUT[-heart_sampling_vector]
summary(heart)
```

```
+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
|      AGE      |      SEX      |   CHESTPAIN   |    RESTBP     |      CHOL     |     SUGAR     |      ECG      |
+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
| Min.   :29.00 | Min.   :0.0000 | Min.   :1.000 | Min.   : 94.0 | Min.   :126.0 | Min.   :0.0000 | Min.   :0.000 |
| 1st Qu.:48.00 | 1st Qu.:0.0000 | 1st Qu.:3.000 | 1st Qu.:120.0 | 1st Qu.:213.0 | 1st Qu.:0.0000 | 1st Qu.:0.000 |
| Median :55.00 | Median :1.0000 | Median :3.000 | Median :130.0 | Median :245.0 | Median :0.0000 | Median :2.000 |
| Mean   :54.43 | Mean   :0.6778 | Mean   :3.174 | Mean   :131.3 | Mean   :249.7 | Mean   :0.1481 | Mean   :1.022 |
| 3rd Qu.:61.00 | 3rd Qu.:1.0000 | 3rd Qu.:4.000 | 3rd Qu.:140.0 | 3rd Qu.:280.0 | 3rd Qu.:0.0000 | 3rd Qu.:2.000 |
| Max.   :77.00 | Max.   :1.0000 | Max.   :4.000 | Max.   :200.0 | Max.   :564.0 | Max.   :1.0000 | Max.   :2.000 |
+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
|  MAXHR        | ANGINA        | DEP           | EXERCISE      |  FLUOR        | THAL          | OUTPUT        |
| Min.   : 71.0 | Min.   :0.0000 | Min.   :0.00 | Min.   :1.000 | Min.   :0.0000 | Min.   :3.000 | Min.   :0.0000 |
| 1st Qu.:133.0 | 1st Qu.:0.0000 | 1st Qu.:0.00 | 1st Qu.:1.000 | 1st Qu.:0.0000 | 1st Qu.:3.000 | 1st Qu.:0.0000 |
| Median :153.5 | Median :0.0000 | Median :0.80 | Median :2.000 | Median :0.0000 | Median :3.000 | Median :0.0000 |
| Mean   :149.7 | Mean   :0.3296 | Mean   :1.05 | Mean   :1.585 | Mean   :0.6704 | Mean   :4.696 | Mean   :0.4444 |
| 3rd Qu.:166.0 | 3rd Qu.:1.0000 | 3rd Qu.:1.60 | 3rd Qu.:2.000 | 3rd Qu.:1.0000 | 3rd Qu.:7.000 | 3rd Qu.:1.0000 |
| Max.   :202.0 | Max.   :1.0000 | Max.   :6.20 | Max.   :3.000 | Max.   :3.0000 | Max.   :7.000 | Max.   :1.0000 |
|               |               |               |               |               |               |               |
+---------------+---------------+---------------+---------------+---------------+---------------+---------------+
```

Με χρήση της εντολής supply από την εργαλειοθήκη εντολών της R παίρνουμε αντίστοιχα τις ελάχιστες και μέγιστες τιμές των προσδιοριστών (πεδίο τιμών) του πλαισίου δεδομένων :

```
> sapply(heart, max)

AGE       SEX CHESTPAIN    RESTBP      CHOL     SUGAR       ECG     MAXHR    ANGINA       DEP  EXERCISE     FLUOR
77.0      1.0       4.0     200.0     564.0       1.0       2.0     202.0       1.0       6.2       3.0       3.0
THAL    OUTPUT
7.0       1.0

-------------------------------------------------------------------------------------------------------------------

> sapply(heart, min)

 AGE       SEX CHESTPAIN    RESTBP      CHOL     SUGAR       ECG     MAXHR    ANGINA       DEP  EXERCISE     FLUOR
  29         0         1        94       126         0         0        71         0         0         1         0
THAL    OUTPUT
   3         0
```

Από τα παραπάνω στατιστικά μπορούμε να συμπεράνουμε ότι:

- Η μέση ηλικία των ασθενών του πλαισίου δεδομένων είναι 54.3 χρόνια

- Οι περισσότεροι νοσούντες είναι άντρες 68%

- Η μέση τιμή της αρτηριακής πίεσης είναι 132 mm/Hg, με ελάχιστη τιμή 94 mm/Hg και μέγιστη στα 200mm/Hg

- Τα μέσα επίπεδα χοληστερόλης είναι 246 mg/dl με ελάχιστες τιμές 126 mg/dl και μέγιστες στα 564 mg/dl

Με χρήση του πακέτου summarytools προχωρήσαμε σε κάποια χρήσιμα στατιστικά συμπεράσματα, όπως ενδεικτικά φαίνονται παρακάτω :

```
library(summarytools)
freq(OUTPUT$SEX)
```

	OUTPUT		
SEX	**Απουσία**	**Παρουσία**	**Σύνολα**
Γυναίκες	67 (77.0%)	20 (23.0%)	87 (100.0%)
Άντρες	83 (45.4%)	100 (54.6%)	183 (100.0%)
Σύνολα	150 (55.6%)	120 (44.4%)	270 (100.0%)

Δηλαδή, περισσότεροι άντρες (54.6%) σε σχέση με γυναίκες (23.0%) παρουσιάζουν το κίνδυνο εμφάνισης της καρδιακής ισχαιμικής νόσου

```
freq(OUTPUT$THAL)
```

	OUTPUT		
THAL	**Απουσία**	**Παρουσία**	**Σύνολα**
Αποφραγμένες αρτηρίες	119 (78.3%)	33 (21.7%)	152 (100.0%)
Περιοχή που έχει νεκρωθεί	6 (42.9%)	8 (57.1%)	14 (100.0%)
Αναστρέψιμη ισχαιμία	25 (24.0%)	79 (76.0%)	104 (100.0%)
Σύνολα	150 (55.6%)	120 (44.4%)	270 (100.0%)

Το 76% των περιπτώσεων που παρουσίασαν τη νόσο είχαν αρτηρίες με μερική στένωση (αναστρέψιμη ισχαιμία)

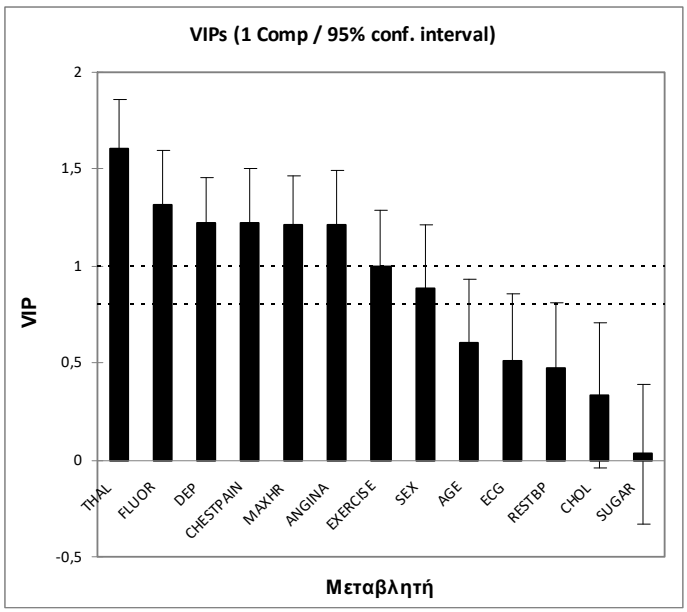
Παθογόνος παράγοντας είναι και η αρτηριακή υπέρταση. Τιμές πάνω από 14.4 mmHg οδηγούν σε πιθανότητα 58.2% εμφάνισης της καρδιακής αθηρωματικής νόσου..

```
freq(OUTPUT$RESTBP)
```

	OUTPUT		
RESTBP (mmHg)	**Απουσία**	**Παρουσία**	**Σύνολα**
94-142	127 (59.7.%)	88 (40.9%)	215 (100.0%)
144-200	23 (41.8%)	32 (58.2%)	55 (100.0%)
Σύνολα	150 (55.6%)	120 (44.4%)	270 (100.0%)

Τα συγκεντρωτικά αποτελέσματα του πίνακα συχνοτήτων αναφέρονται στο κεφάλαιο 3.

Στη συνέχεια, με εφαρμογή «*Παλινδρόμησης Μερικών Ελάχιστων Τετραγώνων*» (Partial Least Squares Regression) από το πακέτο pls4 της R, μπορούμε να δούμε ποιες μεταβλητές συνεισφέρουν την περισσότερη πληροφορία στο μοντέλο μας (VIP) :

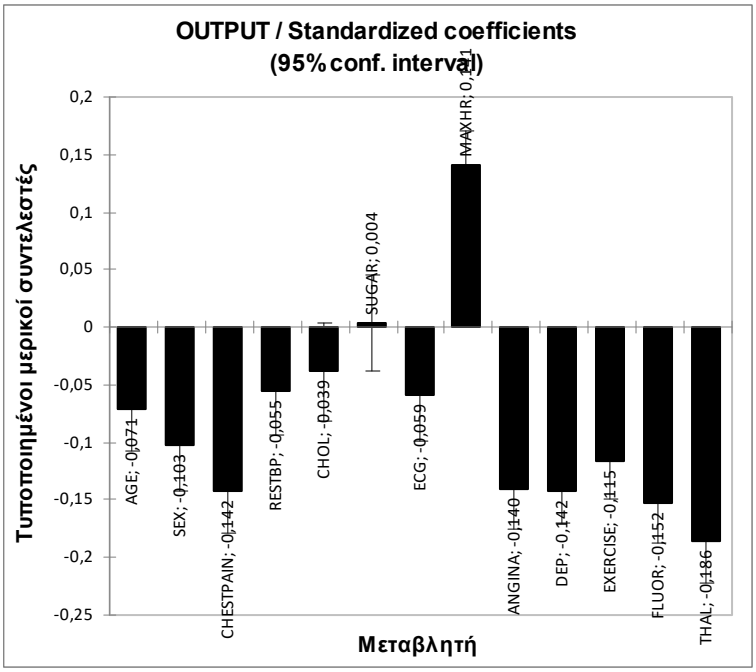


Εικόνα 21 Συνεισφορά μεταβλητών του πλαισίου δεδομένων Statlog

Variable	VIP	Standard deviation	Lower bound(95%)	Upper bound(95%)
THAL	1,607	0,130	1,351	1,863
FLUOR	1,315	0,145	1,030	1,600
DEP	1,224	0,120	0,988	1,459
CHESTPAIN	1,223	0,145	0,937	1,508
MAXHR	1,216	0,125	0,970	1,462
ANGINA	1,210	0,144	0,926	1,493
EXERCISE	0,995	0,151	0,698	1,293
SEX	0,889	0,164	0,566	1,212
AGE	0,608	0,166	0,282	0,934
ECG	0,513	0,176	0,166	0,860
RESTBP	0,476	0,171	0,139	0,813
CHOL	0,335	0,191	-0,041	0,710
SUGAR	0,032	0,183	-0,328	0,393

Ενώ οι μεταβλητές που αλληλεπιδρούν περισσότερο με τη μεταβλητή εξόδου OUTPUT φαίνονται στο παρακάτω σχήμα. Να σημειωθεί ότι οι αρνητικές τιμές απλά δηλώνουν το βαθμό αρνητικής συσχέτισης της εκάστοτε μεταβλητής με την OUTPUT.

[4] https://cran.r-project.org/web/packages/pls/pls.pdf



Εικόνα 22 Μεταβλητές που αλληλεπιδρούν περισσότερο με την OUTPUT

3.2 Εκπαίδευση Πλαισίου Δεδομένων με Λογιστική Παλινδρόμηση

Στη συνέχεια θα εκπαιδεύσουμε το μοντέλο μας με λογιστική παλινδρόμηση. Η συνάρτηση glm βρίσκεται στο πακέτο caret της R και χρησιμοποιείται για την εφαρμογή γενικευμένων γραμμικών μοντέλων, όπως για παράδειγμα η γραμμική παλινδρόμηση. Η πρώτη παράμετρος αφορά την εξαρτημένη μεταβλητή. Η παράμετρος data είναι το πλαίσιο δεδομένων και η τελευταία παράμετρος (family) χρησιμοποιείται για να διευκρινιστεί ότι επιθυμούμε να εφαρμόσουμε λογιστική παλινδρόμηση. Στη συνέχεια, το αποτέλεσμα δίνει τις εκτιμήσεις των συντελεστών των παραμέτρων, την απόκλιση (deviance) του μηδενικού μοντέλου και των υπολοίπων μαζί με τους βαθμούς ελευθερίας τους και την τιμή του κριτηρίου AIC :

```
glm(formula = OUTPUT ~ ., family = binomial("logit"), data = heart_train,
    start = NULL, model = TRUE, method = "glm.fit", x = FALSE,
    y = TRUE, singular.ok = TRUE, contrasts = NULL)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6033  -0.5531  -0.1588   0.4345   2.3479

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.311729   3.368364  -1.874  0.06095 .
AGE         -0.021761   0.027754  -0.784  0.43300
SEX          1.873713   0.586490   3.195  0.00140 **
CHESTPAIN    0.570693   0.222371   2.566  0.01028 *
RESTBP       0.022860   0.011945   1.914  0.05565 .
CHOL         0.007795   0.004319   1.805  0.07107 .
SUGAR       -0.832775   0.636006  -1.309  0.19040
ECG          0.276340   0.210234   1.314  0.18870
MAXHR       -0.027802   0.012058  -2.306  0.02113 *
ANGINA       1.078435   0.477019   2.261  0.02377 *
DEP          0.193149   0.235218   0.821  0.41156
EXERCISE     0.301623   0.421520   0.716  0.47426
FLUOR        1.170562   0.308754   3.791  0.00015 ***
THAL         0.321890   0.110447   2.914  0.00356 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 317.11  on 229  degrees of freedom
Residual deviance: 158.82  on 216  degrees of freedom
AIC: 186.82

Number of Fisher Scoring iterations: 6
```

Αρχικά παρατηρούμε ότι ο μέσος (median) έχει τιμή κοντά στο μηδέν (-0.1588), κάτι που είναι ενθαρρυντικό καθώς αυτό υποδεικνύει ότι το μοντέλο μας δεν παρουσιάζει μεροληψία. Στη συνέχεια οι συντελεστές παλινδρόμησης που έχουν δημιουργηθεί για το πλαίσιο δεδομένων, παρουσιάζονται με τις αντίστοιχες τυπικές τιμές (z-values). Όσο μεγαλύτερη είναι η απόλυτη τιμή των τυπικών τιμών, τόσο πιθανότερο είναι να υπάρχει συσχέτιση σε σχέση με την εξαρτημένη μεταβλητή OUTPUT. Επιπρόσθετα, οι παράμετροι FLUOR, SEX, CHESTPAIN και THAL είναι οι πιο ισχυροί προσδιοριστές για τις καρδιακές παθήσεις. Για παράδειγμα, η CHESTPAIN που αντιστοιχεί στο ασυμπτωματικό θωρακικό άλγος έχει ισχυρή συσχέτιση με την εμφάνιση της καρδιακής νόσου. Επίσης, εμφανίζεται ένας μεγάλος αριθμός μεταβλητών εισόδου που βρίσκονται πέρα από στατιστικό επίπεδο σημαντικότητας (p-value). Αυτό σημαίνει ότι δεν είναι και τόσο καλοί προσδιοριστές για τις καρδιακές παθήσεις σε σχέση με το συγκεκριμένο πλαίσιο δεδομένων. Δηλαδή για παράδειγμα, η ανεξάρτητη μεταβλητή AGE που δεν παρουσιάζει στατιστική σημαντικότητα (p=0.43300), δεν σημαίνει απαραίτητα ότι δεν είναι καλός προσδιοριστής των καρδιακών παθήσεων, αλλά σε σχέση με το υπάρχον πλαίσιο δεδομένων, αυτή η μεταβλητή δεν προσθέτει ικανοποιητική πληροφορία στο σετ δεδομένων. Επιπρόσθετα, τα τυπικά σφάλματα είναι μικρά (<2) και δεν εμφανίζεται κανένα που να υποδηλώνει αριθμητικό πρόβλημα.

Μπορούμε να επαληθεύσουμε τους συλλογισμούς για την συνεισφορά των ανεξάρτητων μεταβλητών, εφαρμόζοντας t έλεγχο, όπου και παίρνουμε αντίστοιχα αποτελέσματα :

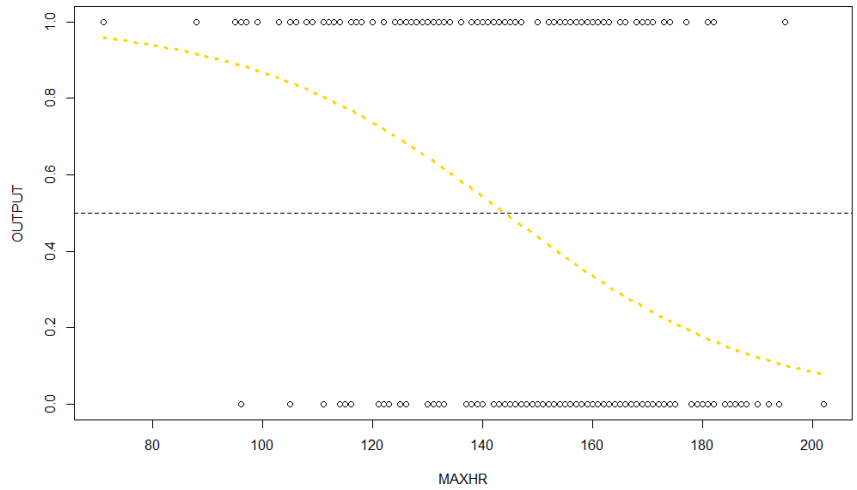
```
anova(heart_model)
```

	Df	Deviance	Resid. Df	Resid. Dev
NULL			229	317.11
AGE	1	7.546	228	309.56
SEX	1	32.061	227	277.50
CHESTPAIN	1	40.904	226	236.60
RESTBP	1	4.254	225	232.34
CHOL	1	2.941	224	229.40
SUGAR	1	0.863	223	228.54
ECG	1	4.492	222	224.05
MAXHR	1	19.513	221	204.53
ANGINA	1	5.230	220	199.30
DEP	1	8.847	219	190.46
EXERCISE	1	0.093	218	190.36
FLUOR	1	22.903	217	167.46
THAL	1	8.643	216	158.82

Δηλαδή, ξανά, παρατηρούμε σημαντική συνεισφορά των μεταβλητών CHESTPAIN, SEX, MAXHR, DEP, FLUOR και THAL στην απόκλιση όταν αυτές προστίθενται στο μοντέλο

Με τη βοήθεια της R μπορούμε να απεικονίσουμε τη χαρακτηριστική παράσταση της λογιστικής παλινδρόμησης ενός εκ τον προσδιοριστών του πλαισίου δεδομένων και της εξαρτημένης μεταβλητής εξόδου.

```
fit = glm(OUTPUT ~ MAXHR, data=heart, family=binomial)
newdat <- data.frame(MAXHR=seq(min(heart$MAXHR), max(heart$MAXHR),len=100))
newdat$OUTPUT = predict(fit, newdata=newdat, type="response")
plot(OUTPUT~MAXHR, data=heart, col="black")
lines(OUTPUT ~ MAXHR, newdat, col="gold", lwd=3,type="l",lty=3)
abline(h=.5, lty=2)
```



Εικόνα 23 Γραφική παράσταση της μεταβλητής OUTPUT σε σχέση με την MAXHR

Στη συνέχεια θα ελέγξουμε αν το πλαίσιο δεδομένων παρουσιάζει πολυσυγγραμμικότητα (multicollinearity). Θα χρησιμοποιήσουμε την εντολή vif που βρίσκεται στο πακέτο car. Η vif έχει ένα εύρος τιμών που προσδιορίζεται εμπειρικά στη βιβλιογραφία. Αν είναι από το 1 μέχρι 4 τότε δεν υπάρχει βαθμός πολυσυγγραμμικότητας μεταξύ των μεταβλητών. Από τιμές 5 μέχρι 10 υπάρχει πολυσυγραμμικότητα μεταξύ συγκεκριμένων προσδιοριστών του πλαισίου δεδομένων.

```
install.packages("car")
library(car)
vif(heart_model)
```

```
AGE       SEX       CHESTPAIN RESTBP    CHOL      SUGAR     ECG       MAXHR     ANGINA    DEP       EXERCISE  FLUOR
1.589769  1.501276  1.183750  1.213120  1.354031  1.194737  1.084041  1.531423  1.130819  1.475681  1.629894  1.220482
THAL
1.143869
```

Οπότε δεν υπάρχει υψηλή συσχέτιση μεταξύ των προσδιοριστών του πλαισίου δεδομένων και ως εκ τούτου πολυσυγραμμικότητα.

Παραπάνω μπορούμε να δούμε και τις τιμές μηδενικής απόκλισης (Null deviance) και υπολειμματικής απόκλισης (Residual deviance). Η μηδενική απόκλιση είναι 317,11 στους 229 βαθμούς ελευθερίας και η υπολειμματική απόκλιση είναι 158,82 στους 216 βαθμούς ελευθερίας. Η απόκλιση είναι ένα μέτρο της προσαρμογής του μοντέλου (goodness of fit). Η μηδενική απόκλιση δείχνει πόσο καλά η εξαρτημένη μεταβλητή προβλέπεται από ένα μοντέλο που περιλαμβάνει μόνο τον μεγάλο μέσο (grand mean), ενώ η υπολειμματική απόκλιση περιλαμβάνει το σύνολο των εξαρτημένων μεταβλητών. Σε έναν εμπειρικό κανόνα[5] αναφέρεται ότι η τιμή της υπολειμματικής απόκλισης (158.82) πρέπει να είναι όσο το δυνατόν κοντά στην τιμή των βαθμών ελευθερίας (229) προκειμένου να έχουμε καλή προσαρμογή.

Για να υπολογίσουμε την πιθανοφάνεια χρησιμποποιούμε το πακέτο "BaylorEdPsych" που δημιουργήθηκε από το Πανεπιστήμιο του Μπέιλορ.
(https://cran.rproject.org/web/packages/BaylorEdPsych/BaylorEdPsych.pdf)

```
install.packages("BaylorEdPsych")
```

Με τη βοήθεια της συνάρτησης "PseudoR2" του εν λόγω πακέτου υπολογίζουμε την πιθανοφάνεια :

```
library(BaylorEdPsych)
```

```
psR.glm<-glm(formula = OUTPUT ~ ., family = binomial("logit"), data =heart_train)
```

```
PseudoR2(psR.glm, c("McFadden"))
```

```
McFadden 0.5556977
```

Το δικό μας μοντέλο λογιστικής παλινδρόμησης εξηγεί μόλις το 55 % με χρήση του συντελεστή τύπου R2 του McFadden (δείκτης του λόγου πιθανοφανειών) της μηδενικής απόκλισης. Δεν είναι ιδιαίτερα υψηλή η τιμή, προφανώς της έλλειψης τιμών στο πλαίσιο δεδομένων, για να κάνουμε περισσότερο ακριβείς προβλέψεις. Να σημειωθεί ότι στη περίπτωση της λογιστικής παλινδρόμησης, είναι δυνατόν ο ψευδο-συντελεστής προσδιορισμού (pseudo-R2) να ξεπεράσει την μονάδα, αλλά αυτό συμβαίνει μόνο σε προβληματικές καταστάσεις και σε τέτοιες περιπτώσεις δεν θα πρέπει να εμπιστευόμαστε το μοντέλο.

3.2.1 Εφαρμογή παλινδρόμησης LASSO στο μοντέλο

Με τη συνάρτηση predict() μπορούμε να υπολογίσουμε την έξοδο του μοντέλου μας. Αυτή η έξοδος είναι η πιθανότητα της εισόδου να παίρνει την τιμή 1. Μπορούμε να εφαρμόσουμε δυαδική ταξινόμηση με την εφαρμογή ενός κατωφλίου (T= 0,5). Αυτό το πραγματοποιούμε τόσο για το πλαίσιο δεδομένου ελέγχου, όσο και όσο και το πλαίσιο εκπαίδευσης. Έπειτα, θα ελέγξουμε την ακρίβεια ταξινόμησης και για τα δυο πλαίσια ελέγχου.

[5] https://stats.stackexchange.com/questions/37732/when-someone-says-residual-deviance-df-should-1-for-a-poisson-model-how-appro

Πρώτα ξεκινάμε με το σετ δεδομένων της εκπαίδευσης :

```
predictions_heart_model <- predict(heart_model,
newdata = heart_train, type = "response")
predictions_T <- as.numeric(predictions_heart_model > 0.5)
mean(predictions_T == heart_train$OUTPUT)
```

```
[1] 0.8869565
```

Στην συνέχεια, με το σετ δεδομένων ελέγχου :

```
test_predictions_heart = predict(heart_model, newdata = heart_test, type = "response")
predictions_T_test = as.numeric(test_predictions_heart > 0.5)
mean(predictions_T_test == heart_test$OUTPUT)
```

```
[1] 0.775
```

Η ακρίβεια ταξινόμησης και για τα δυο σετ δεδομένων είναι κοντά στο 80%. Αυτό είναι ένα πολύ καλό σημείο εκκίνησης για να δουλέψουμε το μοντέλο μας. Ωστόσο, όπως προκύπτει από την παραπάνω ανάλυση, υπάρχουν αρκετοί συντελεστές που δεν προσφέρουν πληροφορία στο πλαίσιο δεδομένων. Το επόμενο βήμα είναι να επιλέξουμε τις κατάλληλες μεταβλητές, έτσι ώστε να βελτιστοποιήσουμε το μοντέλο μας.

Ένα μεγάλο πλεονέκτημα της ανάλυσης κορυφογραμμής είναι ότι ένα γράφημα, το λεγόμενο γράφημα ίχνους κορυφογραμμής, μπορεί να βοηθήσει τον αναλυτή να διαπιστώσει ποιοι συντελεστές είναι ευαίσθητοι στα δεδομένα. Το ίχνος κορυφογραμμής είναι το γράφημα της τιμής κάθε συντελεστή έναντι του λ. Το γράφημα θα έχει μια καμπύλη (ίχνος) για κάθε συντελεστή. Για να είναι πιο ξεκάθαρο, προτείνεται να μη σχεδιάζονται περισσότεροι από 10 συντελεστές στο ίδιο γράφημα,

Ο στόχος είναι να βρούμε μια τιμή του λ η οποία δίνει ένα σύνολο συντελεστών με το μικρότερο μέσο τετραγωνικό σφάλμα. Φυσικά, καθώς το λ αυξάνει, το άθροισμα των τετραγώνων των υπολοίπων (SSE) επίσης αυξάνεται. Αυτό δεν είναι ιδιαίτερα ανησυχητικό διότι ο στόχος δεν είναι να αποκτήσουμε ένα μοντέλο που να προσαρμόζεται όσο το δυνατόν καλύτερα στα δεδομένα, αλλά να αναπτύξουμε ένα "ευσταθές" σύνολο συντελεστών, οι οποίοι θα εκτιμούν αποτελεσματικά μελλοντικές παρατηρήσεις,

Με τη λέξη "ευσταθείς" εννοούμε συντελεστές οι οποίοι δεν είναι ευαίσθητοι σε μικρές αλλαγές στα δεδομένα. Αν οι μεταβλητές πρόβλεψης εμφανίζουν μεγάλες συσχετίσεις, οι συντελεστές θα αλλάζουν γρήγορα για μικρές τιμές του λ και σταδιακά θα σταθεροποιούνται (θα αλλάζουν λίγο) για μεγαλύτερες τιμές.

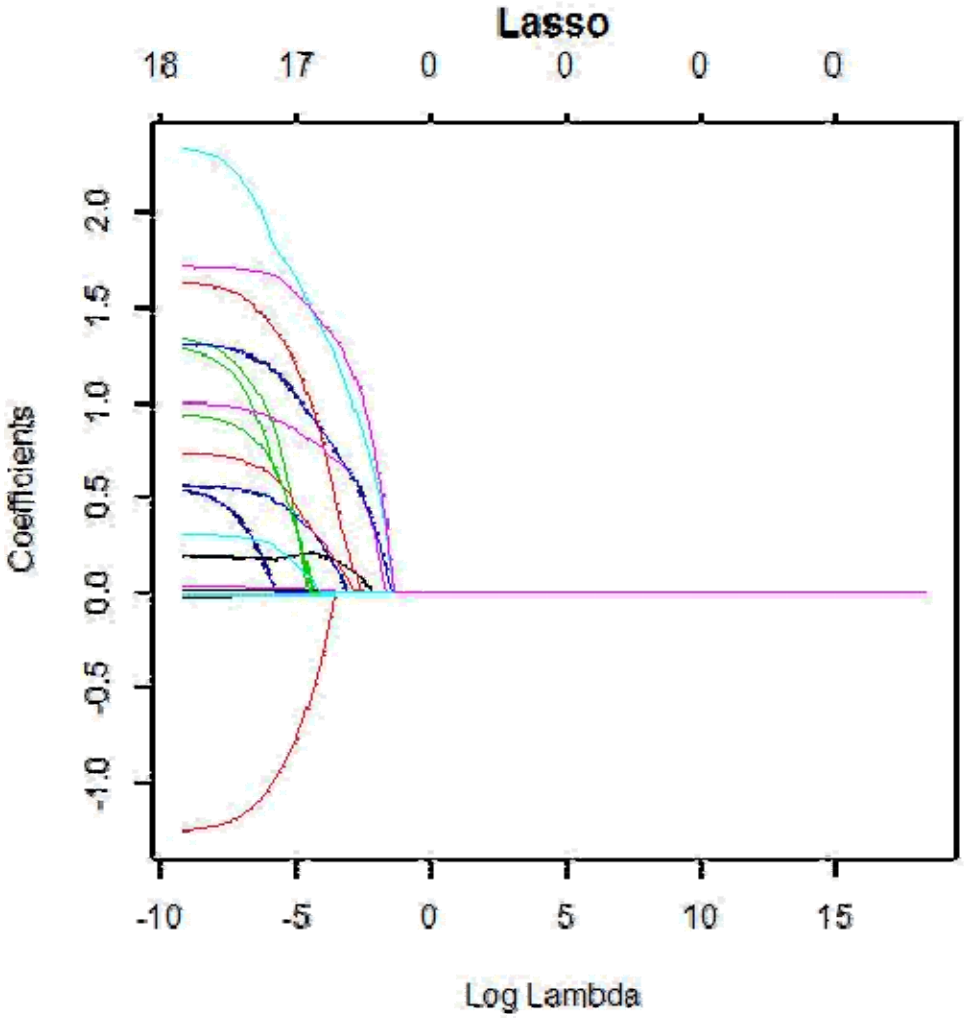
Η τιμή του λ, για την οποία οι συντελεστές έχουν σταθεροποιηθεί, δίνει τους συντελεστές. Αν οι μεταβλητές πρόβλεψης είναι ορθογώνιες, τότε οι συντελεστές θα αλλάζουν ελάχιστα (δηλαδή θα είναι ήδη σταθεροί). Πρέπει βέβαια εδώ να τονίσουμε ότι η βέλτιστη τιμή του λ είναι άγνωστη και μπορεί μόνο να εκτιμηθεί. Ένας εναλλακτικός τρόπος υπολογισμού του βέλτιστου λ είναι με τη μέθοδο της διασταυρούμενης επικύρωσης (cross-validation).

Με χρήση της συνάρτησης glmnet() χρησιμοποιείται ένα πλέγμα διαφορετικών τιμών του λ, προκειμένου να εκπαιδευτεί ένα μοντέλο παλινδρόμησης για κάθε τιμή του. Η πρώτη παράμετρος της glmnet() είναι ένας πίνακας δεδομένων, ο οποίος μπορεί να υλοποιηθεί με τη συνάρτηση model.matrix(). Η δεύτερη παράμετρος της glmnet() είναι ένα διάνυσμα που περιλαμβάνει μαζί και την τιμή της παραμέτρου εξόδου (σε αυτή την περίπτωση είναι η OUTPUT). Η τιμή alpha είναι ένας "διακόπτης" που για την τιμή 0 επιτελεί παλινδρόμηση Ridge, ενώ για την τιμή 1 επιτελεί παλινδρόμηση Lasso.

```
install.packages("glmnet", dependencies=TRUE)
library(glmnet)
heart_matrix <- model.matrix(OUTPUT ~ ., heart_train)[,-1]
lambdas <- 10 ^ seq(8, -4, length = 250)
lasso <- glmnet(heart_matrix,heart_train$OUTPUT, alpha = 1, lambda = lambdas, family =
"binomial")
plot(lasso, xvar = "lambda", main = "Lasso\n")
```

Δηλαδή δημιουργήσαμε μια αλληλουχία 250 τιμών του λ, και επί της ουσίας, εκπαιδεύσαμε 250 φορές το μοντέλο μας με παλινδρόμηση Lasso.

Για να βρούμε τη βέλτιστη τιμή του λ, το πακέτο της R glmnet(), προσφέρει τη συνάρτηση cv.glmnet(). Με αυτή τη συνάρτηση, χρησιμοποιείται διασταυρούμενη επικύρωση δέκα τμημάτων (cross-validation) στο σετ εκπαίδευσης έτσι ώστε να βρεθεί ένα βέλτιστο λ που ελαχιστοποιεί το μέσο τετράγωνο σφάλμα (MSE).



Εικόνα 24 Εφαρμογή παλινδρόμησης LASSO

Καθώς η τιμή της ρυθμιστικής παραμέτρου αυξάνεται, οι συντελεστές των μεταβλητών μεταβάλλονται. Περίπου από την τιμή 0.02 και μετά, οι συντελεστές σχεδόν σταθεροποιούνται. Συνεπώς .μπορούμε να συμπεράνουμε ότι η βέλτιστη τιμή της παραμέτρου είναι λ = 0.02.

Αφού βρήκαμε την βέλτιστη τιμή του λ (0.01474602), τώρα μπορούμε να δούμε το βελτιστοποιημένο μοντέλο :

```
predict(lasso, type = "coefficients", s = lasso_best)
```

SEX	0.883684017
CHESTPAIN2	.
CHESTPAIN3	.
CHESTPAIN4	1.419.778.212
RESTBP	0.010720684
CHOL	0.003302495
SUGAR	-0.427915086
ECG1	.
ECG2	0.281391031
MAXHR	-0.010101541
ANGINA	0.773432623
DEP	0.203377072
EXERCISE2	0.288792869
EXERCISE3	.
FLUOR	0.884411752
THAL6	0.014204741
THAL7	1.444.485.624

Παρατηρούμε ότι ένας μεγάλος αριθμός από προσδιοριστές έχει αφαιρεθεί από το μοντέλο, αφού οι συντελεστές είναι μηδέν. Από το νέο σετ δεδομένων έχουν αφαιρεθεί πέντε μεταβλητές AGE, CHESTPAIN2, CHESTPAIN3, ECG1,EXERCISE3 και THAL7. Σε αυτή την περίπτωση, η ακρίβεια ταξινόμησης είναι για το πλαίσιο εκπαίδευσης 0.8913043, ενώ για το πλαίσιο ελέγχου 0.925.

3.2.2 Απόδοση Συστήματος στη Λογιστική Παλινδρόμηση

Σε αυτό το σημείο, θα μελετήσουμε την ακρίβεια του νέου μοντέλου, όπως αυτό προέκυψε από την παλινδρόμηση Lasso

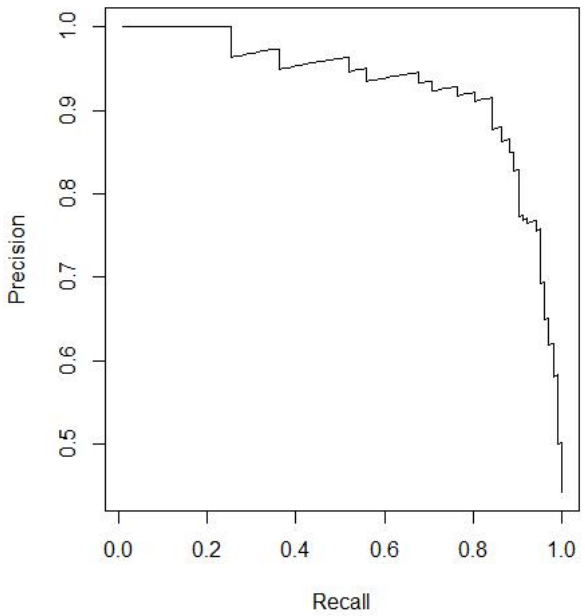
```
lasso_train_predictions <- predict(heart_models_lasso,
s = lambda_lasso, newx = heart_train_mat, type = "response")
lasso_train_class_predictions <-
as.numeric(lasso_train_predictions >
0.5)
mean(lasso_train_class_predictions ==
heart_train$OUTPUT)
[1] 0.8913043

heart_test_mat <- model.matrix(OUTPUT ~ ., heart_test)[,-1]
lasso_test_predictions <- predict(heart_models_lasso,
s = lambda_lasso, newx = heart_test_mat, type = "response")
lasso_test_class_predictions <-
as.numeric(lasso_test_predictions >
0.5)
```

```
mean(lasso_test_class_predictions ==
heart_test$OUTPUT) [1] 0.925
```

Δηλαδή πετύχαμε για το σετ εκπαίδευσης ακρίβεια της τάξης του 89%, ενώ για το σετ ελέγχου **92%**.

Χρησιμοποιούμε τον "στάνταρ" κώδικα του πακέτου ROCR για να απεικονίσουμε την καμπύλη ROC

```
pr <- predict(heart_model, newdata = heart_train, type = "response")
pd <- prediction(pr, heart_train$OUTPUT)
pf <- performance(pd, measure = "prec", x.measure = "rec")
plot(pf)
```



Εικόνα 25 Καμπύλη ROC για το μοντέλο Λογιστικής Παλινδρόμησης

Είναι χρήσιμο να χρησιμοποιήσουμε κάποιες τιμές κατωφλίου για να δούμε πως μεταβάλλεται η ακρίβεια του συστήματος . Θα θεωρήσουμε ότι χρειαζόμαστε τουλάχιστο 80 ακρίβεια και κλίση το κατ ελάχιστο 90%

```
thresholds <- data.frame(cutoffs = perf@alpha.values[[1]], recall
= perf@x.values[[1]], precision = perf@y.values[[1]])
subset(thresholds,(recall > 0.9) & (precision > 0.8))
```

Βλέπουμε δηλαδή ότι ένα κατώφλι με 0.35 ικανοποιεί τις απαιτήσεις μας

```
+-----------------------------------+
|     cutoffs   recall    precision |
+-----------------------------------+
| 112 0.3491857 0.9019608 0.8288288 |
| 113 0.3472740 0.9019608 0.8214286 |
| 114 0.3428354 0.9019608 0.8141593 |
| 115 0.3421438 0.9019608 0.8070175 |
+-----------------------------------+
```

3.3 Εκπαίδευση συστήματος με Μηχανές Διανυσμάτων Υποστήριξης

Η R έχει πολλά πακέτα που μπορούν να υποστηρίξουν τις Μηχανές Διανυσμάτων Υποστήριξης. Εμείς θα συνεχίσουμε με τη χρήση του trctl και του kernlab για την εκπαίδευση και αξιολόγηση του μοντέλου. Θα χρησιμοποιήσομε διασταυρούμενη επικύρωση δέκα τμημάτων και γραμμικό πυρήνα.

```
#Εκπαίδευση μοντέλου με μηχανές διανυσμάτων υποστήριξης
readRDS(file = "heart_train.rds")
```

Η εξαρτημένη μεταβλητή του πλαισίου δεδομένων (μεταβλητή εξόδου) παίρνει δυο τιμές, 0 ή 1. Για να "δουλέψει" σωστά όμως ο αλγόριθμος των μηχανών υποστήριξης, θα πρέπει να μετατραπεί η OUTPUT σε κατηγορική μεταβλητή. Αυτό μπορεί να γίνει με χρήση της συνάρτησης factor:

```
heart_train[["OUTPUT"]] = factor(heart_train[["OUTPUT"]])
```

3.3.1 Μελέτη περίπτωσης με Γραμμικό Πυρήνα

Σε αυτήν τη περίπτωση πρέπει να δώσουμε τιμές μόνο για την παράμετρο C.Το *C* όπως έχει αναφερθεί είναι ή σταθερά σφάλματος που θέτει ένα όριο μεταξύ του μαλακού περιθωρίου και της σωστής ταξινόμησης των σημείων εκπαίδευσης.

Στη συνέχεια προχωράμε στη εκπαίδευση του μοντέλου με γραμμικό πυρήνα :

```
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
set.seed(3233)
svm_Linear <- train(OUTPUT ~., data = heart_train, method =
"svmLinear", trControl=trctrl,
preProcess = c("center", "scale"),
                 tuneLength = 10)
```

```
Support Vector Machines with Linear Kernel

230 samples
 13 predictor
  2 classes: '0', '1'

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 208, 206, 207, 206, 207, 207, ...
Resampling results:

  Accuracy   Kappa
  0.8248408  0.6419096

Tuning parameter 'C' was held constant at a value of 1
```

Σε αυτό τον αλγόριθμο, δηλαδή, με επιλογή συντελεστή Κόεν *k=0.65* (που προέκυψε από τη διασταυρούμενη επικύρωση 10 τμημάτων) και *C=1*, η ακρίβεια για το σετ δεδομένων εκπαίδευσης που επιτυγχάνουμε είναι της τάξης του 83%

Για τον πίνακα σύγχυσης έχουμε:

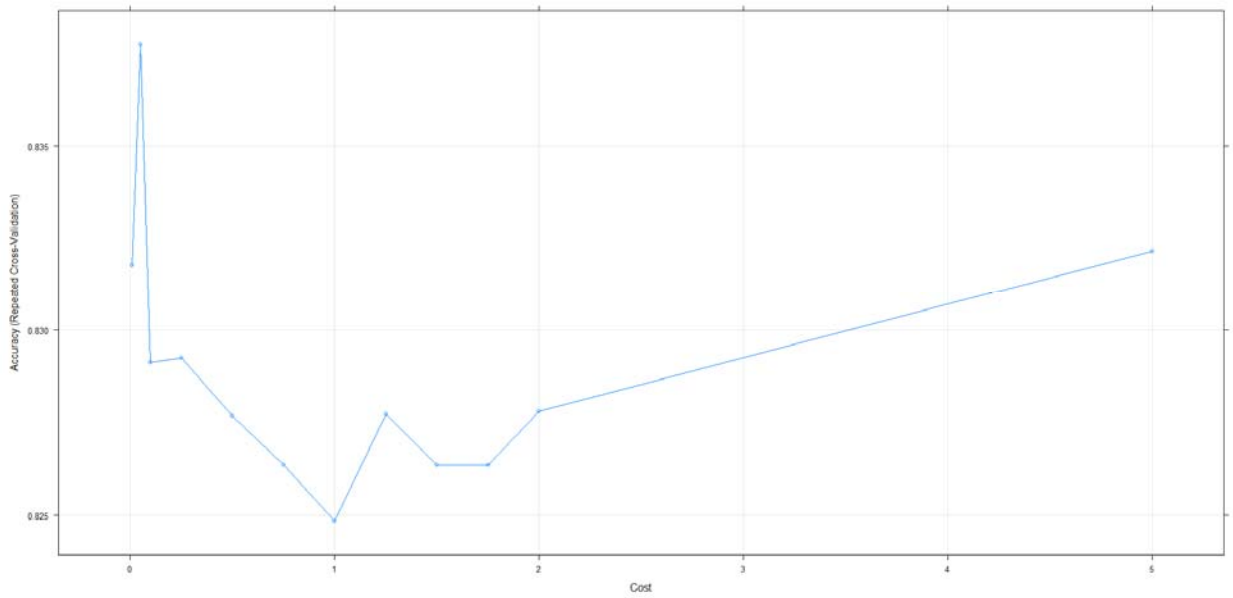
```
heart_test[["OUTPUT"]] = factor(heart_test[["OUTPUT"]])
pred <- predict(svm_Linear, newdata = heart_test)
confusionMatrix(pred, heart_test$OUTPUT )
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 19  5
         1  3 13

               Accuracy : 0.8
                 95% CI : (0.6435, 0.9095)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 0.0008833

                  Kappa : 0.5918

 Mcnemar's Test P-Value : 0.7236736

            Sensitivity : 0.8636
            Specificity : 0.7222
         Pos Pred Value : 0.7917
         Neg Pred Value : 0.8125
             Prevalence : 0.5500
         Detection Rate : 0.4750
   Detection Prevalence : 0.6000
      Balanced Accuracy : 0.7929

       'Positive' Class : 0
```

Επιτύχαμε ακρίβεια για το σετ εκπαίδευσης δηλαδή της τάξης των **80,0%**. Πρέπει να αναφερθεί ότι το πακέτο caret έχει προεπιλεγμένη τιμή C=1 για τη σταθερά σφάλματος. Θα επιχειρήσουμε να βελτιώσουμε ακόμα περισσότερο τα χαρακτηριστικά του μοντέλου. Έχουμε τη δυνατότητα να δώσουμε τιμές στην παράμετρο κόστους C (Cost) με χρήση της συνάρτησης expand.grid()

```
grid <- expand.grid(C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75,
2,5))
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
set.seed(3233)
svm_LinearGrid <- train(OUTPUT ~., data = heart_train, method = "svmLinear",
trControl=trctrl, preProcess = c("center", "scale"), tuneGrid = grid, tuneLength = 10)
```

```
Support Vector Machines with Linear Kernel

230 samples
 13 predictor
  2 classes: '0', '1'

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 208, 206, 207, 206, 207, 207, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.00       NaN       NaN
  0.01  0.8317688  0.6562179
  0.05  0.8377635  0.6684909
  0.10  0.8291337  0.6512064
  0.25  0.8292545  0.6514944
  0.50  0.8276790  0.6481015
  0.75  0.8263560  0.6445497
  1.00  0.8248408  0.6419096
  1.25  0.8277448  0.6477811
  1.50  0.8263614  0.6443166
  1.75  0.8263614  0.6443166
  2.00  0.8278107  0.6471644
  5.00  0.8321640  0.6563201

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 0.05.
```



Εικόνα 26 Γραφική παράσταση της ακρίβειας σε σχέση με το βάρος του κόστους των λανθασμένων ταξινομήσεων για SVM με γραμμικό πυρήνα

Βλέπουμε δηλαδή ότι ο ταξινομητής δίνει βέλτιστη ακρίβεια 83% για *C=0,05*. Θα δοκιμάσουμε να εκπαιδεύσουμε το πλαίσιο δεδομένων ελέγχου με τιμή *C =0,05* :

```
heart_test[["OUTPUT"]] = factor(heart_test[["OUTPUT"]])
pred_grid <- predict(svm_LinearGrid, newdata = heart_test)
confusionMatrix(pred_grid, heart_test$OUTPUT )
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 21  5
         1  1 13

               Accuracy : 0.85
                 95% CI : (0.7016, 0.9429)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 5.926e-05

                  Kappa : 0.6907

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.9545
            Specificity : 0.7222
         Pos Pred Value : 0.8077
         Neg Pred Value : 0.9286
             Prevalence : 0.5500
         Detection Rate : 0.5250
   Detection Prevalence : 0.6500
      Balanced Accuracy : 0.8384

       'Positive' Class : 0
```

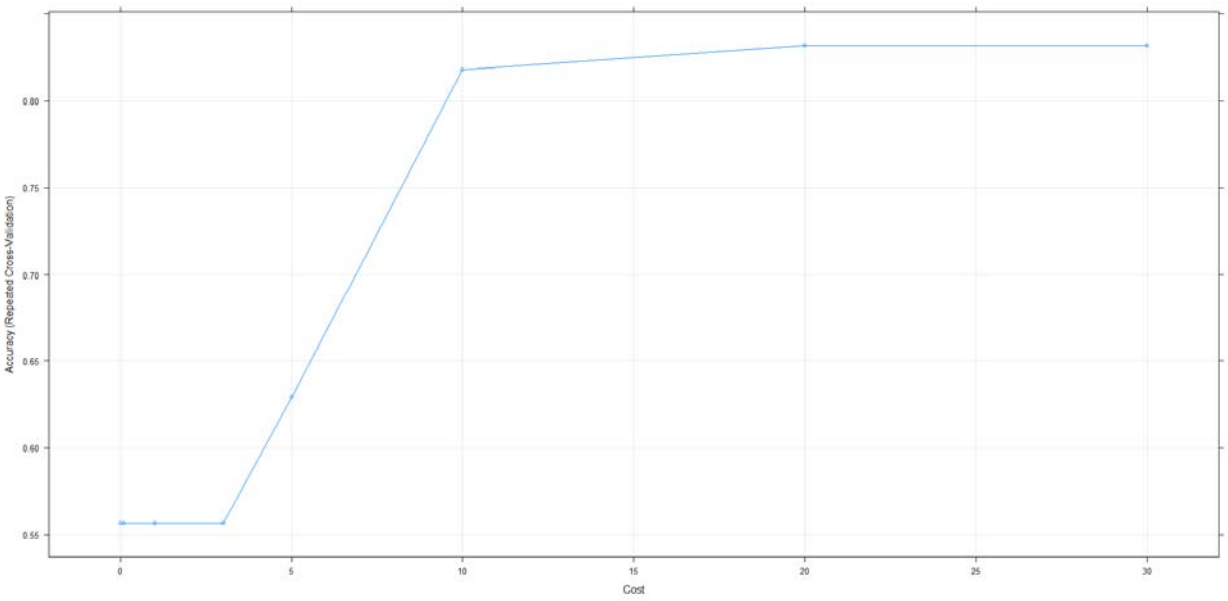
Δηλαδή με γραμμικό πυρήνα και παράμετρο C=0,01 έχουμε ακρίβεια πρόβλεψης **85%.**

3.3.2 Μελέτη περίπτωσης με Πυρήνα ακτινικής βάσης

Σε αυτό το σημείο θα δούμε αν μπορούμε να έχουμε ακόμα καλύτερα αποτελέσματα χρησιμοποιώντας μη γραμμικούς πυρήνες. Στην περίπτωση των πυρήνων RBF, εκτός από την παράμετρο c, υπάρχει ακόμη και η παράμετρος *sigma* (*σ*) (εύρος ζώνης της συνάρτησης πυρήνα). Αυτή η παράμετρος ελέγχει το επίπεδο μη γραμμικότητας που εισάγεται στο μοντέλο. Εάν η τιμή *sigma* είναι πολύ μικρή, τότε το όριο απόφασης είναι μη γραμμικό. Αρχικά πρέπει να βρούμε τις βέλτιστες τιμές για τις παραμέτρους *C* και sigma:

```
grid <-expand.grid(.sigma=.1,.C=c(0.01,0.1,1,3,5,10,20))
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
set.seed(3233)
svm_radialGrid <- train(OUTPUT ~., data = heart_train, method = "svmRadial",
trControl=trctrl, preProcess = c("center", "scale"), tuneGrid = grid, tuneLength = 10)
```

```
Support Vector Machines with Radial Basis Function Kernel

230 samples
 13 predictor
  2 classes: '0', '1'

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 208, 206, 207, 206, 207, 207, ...
Resampling results across tuning parameters:

  C      Accuracy   Kappa
   0.01  0.5565492  0.0000000
   0.10  0.5565492  0.0000000
   1.00  0.5565492  0.0000000
   3.00  0.5565492  0.0000000
   5.00  0.6293094  0.1782866
  10.00  0.8180281  0.6214182
  20.00  0.8318950  0.6562645
  30.00  0.8318347  0.6559609

Tuning parameter 'sigma' was held constant at a value of 1e-04
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 1e-04 and C = 20.
```



Εικόνα 27 Γραφική παράσταση της ακρίβειας σε σχέση με το βάρος του κόστους των λανθασμένων ταξινομήσεων για SVM με Πυρήνα ακτινικής βάσης

Ενώ ο πίνακας σύγχυσης για την περίπτωση της μηχανής διανυσμάτων υποστήριξης με χρήση με πυρήνα ακτινικής βάσης είναι :

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 22  5
         1  0 13

               Accuracy : 0.875
                 95% CI : (0.732, 0.9581)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 1.186e-05

                  Kappa : 0.7409

 Mcnemar's Test P-Value : 0.07364

            Sensitivity : 1.0000
            Specificity : 0.7222
         Pos Pred Value : 0.8148
         Neg Pred Value : 1.0000
             Prevalence : 0.5500
         Detection Rate : 0.5500
   Detection Prevalence : 0.6750
      Balanced Accuracy : 0.8611

       'Positive' Class : 0
```

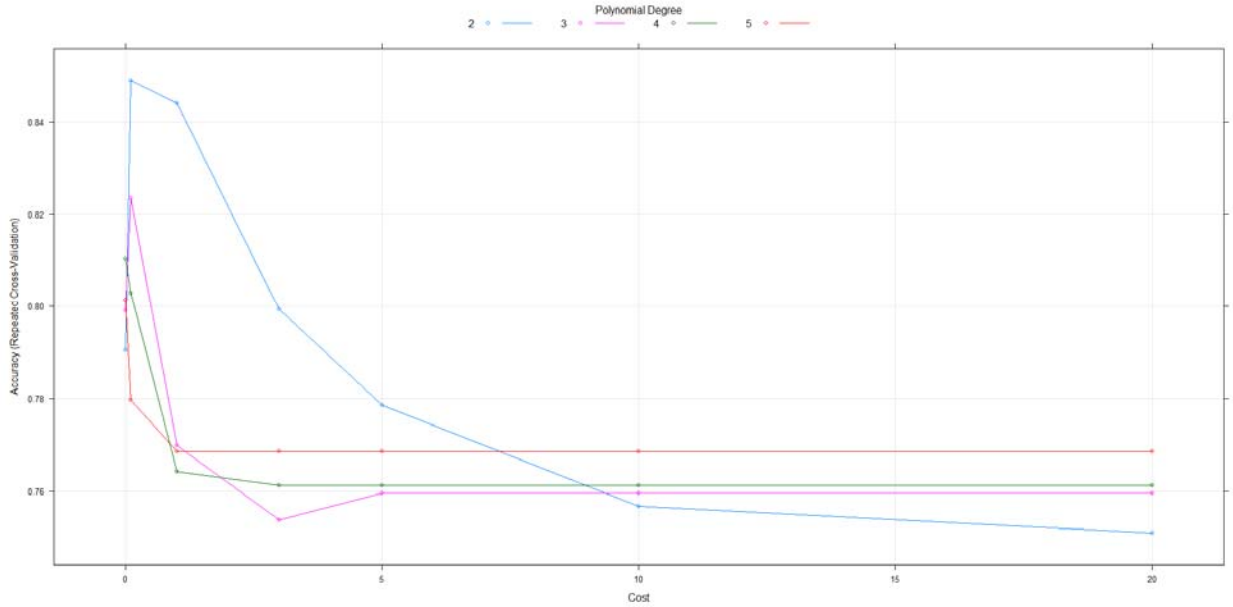
Σε αυτή την περίπτωση επιτύχαμε ακρίβεια **87.5 %** με μεγάλες τιμές του C=20. Δηλαδή μεγάλη τιμή του C σημαίνει ότι ο αλγόριθμος βελτιστοποίησης θα επιλέξει ένα υπερεπίπεδο μικρότερου περιθωρίου, επειδή κάνει καλύτερη δουλειά για να ταξινομήσει σωστά όλα τα σημεία εκπαίδευσης. Αυτό όμως κάνει το μοντέλο μας μεροληπτικό (biased) και ως εκ τούτου αυτός ο πυρήνας δεν αποτελεί καλή επιλογή.

3.3.3 Μελέτη περίπτωσης με Πολυωνυμικό πυρήνα

Το *scale* είναι μια παράμετρος κλιμάκωσης για τα δεδομένα εισόδου. Τα δεδομένα εισόδου συνιστάται να κλιμακώνονται σε σχέση με ένα χαρακτηριστικό πριν εφαρμοστούν στη συνάρτηση πυρήνα. Όταν οι απόλυτες τιμές ορισμένων χαρακτηριστικών κυμαίνονται ευρέως, το εσωτερικό τους γινόμενο μπορεί να είναι κυρίαρχο στον υπολογισμό του πυρήνα. Ο βαθμός πολυωνύμου *degree* ελέγχει την ευελιξία του ορίου απόφασης. Οι πυρήνες υψηλότερου βαθμού αποδίδουν ένα πιο ευέλικτο όριο απόφασης.
Εκπαιδεύοντας το σύνολο δεδομένων με πολυωνυμικό πυρήνα παίρνουμε :

```
grid <-expand.grid(.degree=(2:5), .scale=.1, .C=c(0.01,0.1,1,3,5,10,20))
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
set.seed(3233)
svm_polyGrid <- train(OUTPUT ~., data = heart_train, method = "svmPoly",
trControl=trctrl, preProcess = c("center", "scale"), tuneGrid = grid, tuneLength = 10)
```

```
Support Vector Machines with Polynomial Kernel

230 samples
 13 predictor
  2 classes: '0', '1'

Pre-processing: centered (18), scaled (18)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 208, 206, 207, 206, 207, 207, ...
Resampling results across tuning parameters:

  degree  C      Accuracy   Kappa
  2        0.01  0.7903491  0.5611225
  2        0.10  0.8490448  0.6918161
  2        1.00  0.8440108  0.6826962
  2        3.00  0.7992369  0.5922800
  2        5.00  0.7786232  0.5505321
  2       10.00  0.7567468  0.5061060
  2       20.00  0.7509497  0.4959013
  3        0.01  0.7990558  0.5813545
  3        0.10  0.8234464  0.6409829
  3        1.00  0.7698068  0.5318058
  3        3.00  0.7537220  0.5002301
  3        5.00  0.7595246  0.5117282
  3       10.00  0.7595246  0.5117282
  3       20.00  0.7595246  0.5117282
  4        0.01  0.8104084  0.6063329
  4        0.10  0.8028986  0.5983230
  4        1.00  0.7641359  0.5214488
  4        3.00  0.7612374  0.5155290
  4        5.00  0.7612374  0.5155290
  4       10.00  0.7612374  0.5155290
  4       20.00  0.7612374  0.5155290
  5        0.01  0.8014493  0.5906927
  5        0.10  0.7797047  0.5514811
  5        1.00  0.7685496  0.5289520
  5        3.00  0.7685496  0.5289520
  5        5.00  0.7685496  0.5289520
  5       10.00  0.7685496  0.5289520
  5       20.00  0.7685496  0.5289520

Tuning parameter 'scale' was held constant at a value of 0.1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were degree = 2, scale = 0.1 and C = 0.1.
```



Εικόνα 28 ακρίβεια σε σχέση την παράμετρο πολυπλοκότητας για διάφορες τιμές πολυωνύμων

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 19  5
         1  3 13

               Accuracy : 0.8
                 95% CI : (0.6435, 0.9095)
    No Information Rate : 0.55
    P-Value [Acc > NIR] : 0.0008833

                  Kappa : 0.5918

 Mcnemar's Test P-Value : 0.7236736

            Sensitivity : 0.8636
            Specificity : 0.7222
         Pos Pred Value : 0.7917
         Neg Pred Value : 0.8125
             Prevalence : 0.5500
         Detection Rate : 0.4750
   Detection Prevalence : 0.6000
      Balanced Accuracy : 0.7929

       'Positive' Class : 0
```

Δηλαδή πετύχαμε **ακρίβεια = 80%** με C=0.1 , degree = 2 και scale=0.1

3.4 Εκπαίδευση Πλαισίου Δεδομένων με Νευρωνικό Δίκτυο

Ο νευρώνας Perceptron είναι ένα είδος τεχνητού νευρωνικού δικτύου και μπορεί να χαρακτηριστεί ώς ένα απλό είδος ενός εμπροσθοτροφοδοτούμενου νευρωνικού δικτύου, δηλαδή ένας γραμμικός ταξινομητής. Ένα πολυεπίπεδο ή πολυστρωματικό εμπροσθοτροφοδοτούμενο νευρωνικό δίκτυο έχει πολύ περισσότερη επεξεργαστική ισχύ από ένα Perceptron ενός επιπέδου, για το λόγο αυτό και προτιμάται από έναν απλό Perceptron

3.4.2 Μελέτη Περίπτωσης με Πολυεπίπεδο Νευρωνικό Δίκτυο Perceptron

Κατά τη προεργασία του σετ, το πρώτο βήμα είναι να εξετάσουμε αν στο πλαίσιο δεδομένων υπάρχουν κατηγορικές μεταβλητές, έτσι ώστε να κατηγοριοποιηθούν σε αντίστοιχη αριθμητική τιμή:

```
> str(heart)
Classes ‘data.table’ and 'data.frame':  270 obs. of  14 variables:
 $ AGE      : num  70 67 57 64 74 65 56 59 60 63 ...
 $ SEX      : num  1 0 1 1 0 1 1 1 1 0 ...
 $ CHESTPAIN: num  4 3 2 4 2 4 3 4 4 4 ...
 $ RESTBP   : num  130 115 124 128 120 120 130 110 140 150 ...
 $ CHOL     : num  322 564 261 263 269 177 256 239 293 407 ...
 $ SUGAR    : num  0 0 0 0 0 0 1 0 0 0 ...
 $ ECG      : num  2 2 0 0 2 0 2 2 2 2 ...
 $ MAXHR    : num  109 160 141 105 121 140 142 142 170 154 ...
 $ ANGINA   : num  0 0 0 1 1 0 1 1 0 0 ...
 $ DEP      : num  2.4 1.6 0.3 0.2 0.2 0.4 0.6 1.2 1.2 4 ...
 $ EXERCISE : num  2 2 1 2 1 1 2 2 2 2 ...
 $ FLUOR    : num  3 0 0 1 1 0 1 1 2 3 ...
 $ THAL     : num  3 7 7 7 3 7 6 7 7 7 ...     `
 $ OUTPUT   : num  1 0 1 0 0 0 1 1 1 1 ...
```

Επειδή όλες οι μεταβλητές έχουν κατηγοριοποιηθεί σε αριθμητικές τιμές, δεν απαιτείται κάτι παραπάνω. Για την ορθή λειτουργία του νευρωνικού δικτύου, θα πρέπει να μετατρέψουμε όλες τις μεταβλητές σε διχότομες (dummy, δηλ. του τύπου 0-1). Ακολουθήσαμε μέθοδο κανονικοποίησης μεγίστου-ελαχίστου (min-max normalization) :

```
#Scaling [-1,1]#

max_heart <- apply(heart_data, 2, max)
min_heart <- apply(heart_data, 2, min)
heart_scaled <- scale(heart_data, center = min_heart, scale = max_heart - min_heart)

head(heart_scaled)
```

```
        AGE SEX CHESTPAIN    RESTBP      CHOL SUGAR ECG     MAXHR ANGINA        DEP
1 0.8541667   1 1.0000000 0.3396226 0.4474886     0   1 0.2900763      0 0.38709677
2 0.7916667   0 0.6666667 0.1981132 1.0000000     0   1 0.6793893      0 0.25806452
3 0.5833333   1 0.3333333 0.2830189 0.3082192     0   0 0.5343511      0 0.04838710
4 0.7291667   1 1.0000000 0.3207547 0.3127854     0   0 0.2595420      1 0.03225806
5 0.9375000   0 0.3333333 0.2452830 0.3264840     0   1 0.3816794      1 0.03225806
6 0.7500000   1 1.0000000 0.2452830 0.1164384     0   0 0.5267176      0 0.06451613
  EXERCISE     FLUOR THAL OUTPUT
1      0.5 1.0000000    0      1
2      0.5 0.0000000    1      0
3      0.0 0.0000000    1      1
4      0.5 0.3333333    1      0
5      0.0 0.3333333    0      0
6      0.0 0.0000000    1      0
```

Στη συνέχεια χωρίζουμε το πλαίσιο δεδομένων σε test και train με κανόνα 85/15:

```
index = sample(1:nrow(heart_data),round(0.85*nrow(heart_data)))
heart_train <- as.data.frame(heart_scaled[index,])
heart_test <- as.data.frame(heart_scaled[-index,])
```

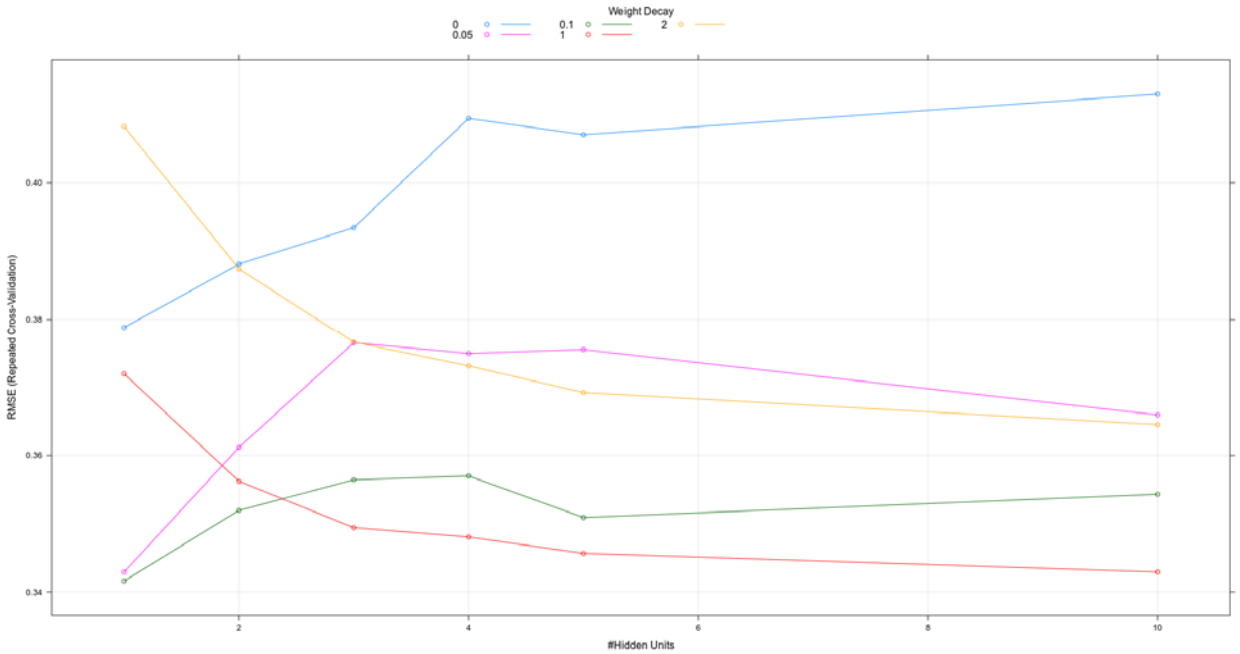
Για την εκπαίδευση του νευρωνικού δικτύου χρησιμοποιήθηκε η συνάρτηση nnet[6], καθώς με αυτή, δίνεται η δυνατότητα για διασταυρούμενη επικύρωση. Πρέπει να σημειωθεί επίσης, ότι ένας τρόπος για να περιορίσουμε το μέσο τετράγωνο σφάλμα ενός νευρωνικού δικτύου είναι μικραίνοντας την αρχιτεκτονική του, δηλαδή ουσιαστικά των αριθμό των βαρών του δικτύου. Η προσθήκη του όρου κανονικοποίησης (regularization) στην ουσία παρεμποδίζει τα βάρη να λάβουν υψηλές (κατ' απόλυτη τιμή) τιμές κατά την εκπαίδευση. Ο πιο απλός τρόπος για να επιτύχουμε κανονικοποίηση βασίζεται στη προσθήκη ενός όρου τιμωρίας (penalty term) στη συνάρτηση τετραγωνικού σφάλματος που ελαχιστοποιούμε κατά την εκπαίδευση του δικτύου. Κατα συνέπεια ακολουθήθηκε και σε αυτό το μοντέλο η μέθοδος πλέγματος με σκοπό να βρεθεί μια βέλτιστη τιμή αριθμού βαρών του δικτύου. Μπορούμε δηλαδή να θεωρήσουμε ότι οι τιμές των βαρών 'φθείρονται' κατά τη διάρκεια της εκπαίδευσης, για το λόγο αυτό η μέθοδος ονομάζεται εκπαίδευση "Αποσύνθεσης Βαρών" (weight decay).

```
tune_grid_neural <- expand.grid(size = c(1:5, 10),decay = c(0, 0.05, 0.1, 1, 2))
max_size_neural <- max(tune_grid_neural$size)
max_weights_neural <- max_size_neural*(nrow(heart_train) + 1) + max_size_neural + 1
train_control_neural <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

fit_nn <- train(OUTPUT ~.,
data = heart_train,
method = "nnet",
        tuneGrid = tune_grid_neural,
        trControl = train_control_neural,
```

[6] https://cran.r-project.org/web/packages/nnet/nnet.pdf

```
        preProcess = c("center", "scale"),
        tuneLength = 10,
        trace = FALSE,
        maxit = 100,
        MaxNWts = max_weights_neural)
```

Δηλαδή ο αλγόριθμος προτείνει βέλτιστο αριθμό βαρών 10 με τιμή μέσου τετράγωνου σφάλματος 0.36 και βήμα αποσύνθεσης βαρών 1.



Εικόνα 29 Ο βέλτιστος αριθμός βαρών με το ελάχιστο μέσο τετράγωνο σφάλμα

Τα αποτελέσματα μπορούμε να τα πάρουμε με χειροκίνητο τρόπο, καθώς η συνάρτηση nnet δεν έχει ενσωματωμένες ρουτίνες για αυτόματο υπολογισμό των μετρικών απόδοσης του μοντέλου :

```
train_predictions <- predict(fit_nn, newdata = heart_train,type = "raw")

train_class_predictions <- as.numeric(train_predictions > 0.5)

mean(train_class_predictions == heart_train$OUTPUT)
[1] 0.8695652

> test_predictions = predict(fit_nn, newdata = heart_test,type = "raw")

test_class_predictions = as.numeric(test_predictions > 0.5)

mean(test_class_predictions == heart_test$OUTPUT)
[1] 0.9250

(confusion_matrix <- table(predicted = test_predictions, actual = heart_test$OUTPUT))
```

```
          actual
predicted   0   1
        0 21   2
        1  1  16
```

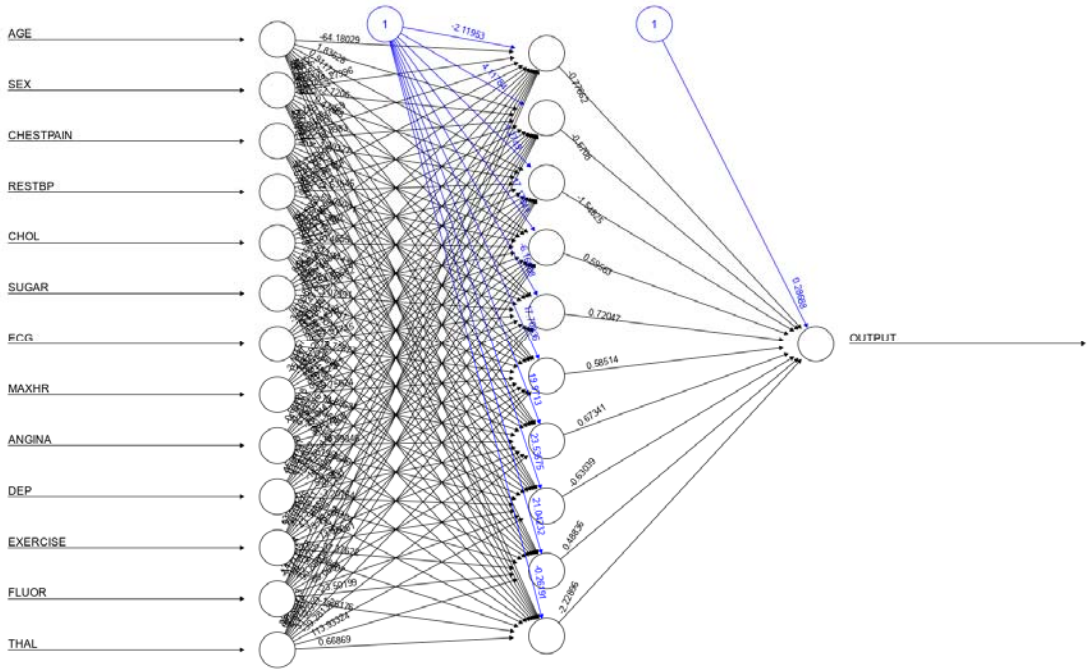
```
(precision <- confusion_matrix[2, 2] / sum(confusion_matrix[2,]))
[1] 0.9130
>
(recall <- confusion_matrix[2, 2] / sum(confusion_matrix[,2]))
[1] 0.8217822
```

```
(f = 2 * precision * recall / (precision + recall))
[1] 0.9333
```

```
(specificity <- confusion_matrix[1,1]/sum(confusion_matrix[1,]))
[1] 0.8889
```

Δηλαδή επιτύχαμε **ακρίβεια 91.3 %.** Για την απεικόνιση των αποτελεσμάτων επιλέχθηκε η συνάρτηση neuralnet[7] :

```
library(neuralnet)
n <- names(heart)
f <- as.formula(paste("OUTPUT ~", paste(n[!n %in% c("OUTPUT")], collapse = " + ")))
heart_model = neuralnet(f,data=heart_train ,hidden=10,linear.output=T)
plot(heart_model)
```



Εικόνα 30 Το νευρωνικό δίκτυο με βάση το πλαίσιο δεδομένων Statlog (heart)

[7] https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf

3.5 Εκπαίδευση Πλαισίου Δεδομένων με Ταξινομητή Μπέϋζ

Το πακέτο caret περιέχει τη βιβλιοθήκη εκπαίδευσης ενός ταξινομητή Μπέϋζ (method = 'naive_bayes'). Προκειμένου να τρέξει ο αλγόριθμος απαιτείται να δοθούν τιμές για τις τρεις παραμέτρους στοχαστικής βελτιστοποίησης: **usekernel**, **laplace** και **adjust**. Η μεταβλητή laplace αντιστοιχεί σε μια τεχνική που ονομάζεται **Λαπλασιανή λείανση** (Laplace smoothing). Μια μέθοδο, δηλαδή, λείανσης των κατηγορικών δεδομένων. Ενσωματώνεται μια μικρή διόρθωση-δείγματος σε κάθε εκτίμηση πιθανότητας και αυτό έχει ως αποτέλεσμα όλες οι πιθανότητες να έχουν τιμές διάφορες του μηδενός.[55] Με το **adjust** εννοούμε το τρόπο με τον οποίο προσαρμόζεται το μέγεθος εκτίμησης της πυκνότητας πυρήνα (Kernel density estimation-KDE). Χρησιμοποιείται μόνο όταν η επιλογή **usekernel** είναι TRUE, διαφορετικά, δεν χρειαζόμαστε αυτή την παράμετρο, αφού η μέθοδος εκτίμησης της πυκνότητας θα είναι κατανομής Γκάους (Gaussian). Το εύρος ζώνης στο KDE αντιπροσωπεύει το ποσό της διασποράς στη λειτουργία πυρήνα του KDE. Σύμφωνα με τα βιβλιογραφικά δεδομένα, είναι μια μέθοδος που μπορεί να αυξήσει σε μεγάλο βαθμό την ακρίβεια. [53],[54].

```
#Naive Bayes Machine learning algorithm
library(caret)
library(naivebayes)

heart <- fread('http://archive.ics.uci.edu/ml/machine-learning-
databases/statlog/heart/heart.dat')
names(heart) <- c("AGE", "SEX", "CHESTPAIN", "RESTBP", "CHOL",
 "SUGAR", "ECG", "MAXHR", "ANGINA", "DEP", "EXERCISE", "FLUOR",
 "THAL", "OUTPUT")

heart$OUTPUT = factor(heart$OUTPUT)
str(heart$OUTPUT)

head(heart)

library(caret)
set.seed(987954)
heart_sampling_vector <-
createDataPartition(heart$OUTPUT, p = 0.85, list = FALSE)
heart_train <- heart[heart_sampling_vector,]
heart_train_labels <- heart$OUTPUT[heart_sampling_vector]
heart_test <- heart[-heart_sampling_vector,]
heart_test_labels <- heart$OUTPUT[-heart_sampling_vector]

Grid<-   expand.grid(usekernel = c(TRUE, FALSE), laplace = c(0, 0.5, 1), adjust =
c(0.75, 1, 1.25, 1.5))
fitControl <- trainControl(method = "cv", number = 10)
```

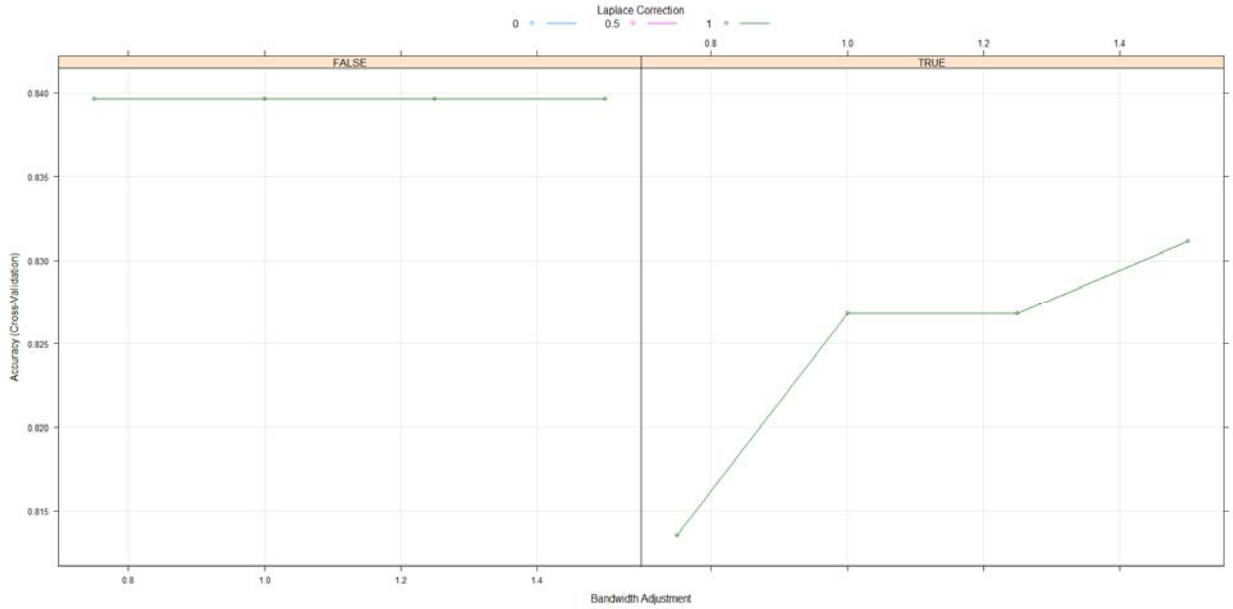
```
heartBayes <- train(OUTPUT ~.,
                    data=heart_train,
                    method = 'naive_bayes',
                    trControl = fitControl,
                    metric = "Accuracy",
                    tuneGrid = Grid)
```

και τα αποτελέσματα που παίρνουμε είναι τα εξής :

```
Naive Bayes

230 samples
 13 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 207, 207, 207, 207, 207, 207, ...
Resampling results across tuning parameters:
```

usekernel	laplace	adjust	Accuracy	Kappa
FALSE	0.0	0.75	0.8396410	0.6746947
FALSE	0.0	1.00	0.8396410	0.6746947
FALSE	0.0	1.25	0.8396410	0.6746947
FALSE	0.0	1.50	0.8396410	0.6746947
FALSE	0.5	0.75	0.8396410	0.6746947
FALSE	0.5	1.00	0.8396410	0.6746947
FALSE	0.5	1.25	0.8396410	0.6746947
FALSE	0.5	1.50	0.8396410	0.6746947
FALSE	1.0	0.75	0.8396410	0.6746947
FALSE	1.0	1.00	0.8396410	0.6746947
FALSE	1.0	1.25	0.8396410	0.6746947
FALSE	1.0	1.50	0.8396410	0.6746947
TRUE	0.0	0.75	0.8135540	0.6255153
TRUE	0.0	1.00	0.8267951	0.6522104
TRUE	0.0	1.25	0.8267951	0.6520249
TRUE	0.0	1.50	0.8311430	0.6613756
TRUE	0.5	0.75	0.8135540	0.6255153
TRUE	0.5	1.00	0.8267951	0.6522104
TRUE	0.5	1.25	0.8267951	0.6520249
TRUE	0.5	1.50	0.8311430	0.6613756
TRUE	1.0	0.75	0.8135540	0.6255153
TRUE	1.0	1.00	0.8267951	0.6522104
TRUE	1.0	1.25	0.8267951	0.6520249
TRUE	1.0	1.50	0.8311430	0.6613756

```
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = FALSE and adjust = 0.75.
```



Εικόνα 31 Γραφική παράσταση της ακρίβειας σε σχέση με τη ρύθμιση του εύρους. Από αριστερά για την περίπτωση που δεν εφαρμόζεται λείανση Laplace και δεξιά για την περίπτωση που εφαρμόζεται αυτή η τεχνική.

```
- Call: naive_bayes.default(x = x, y = y, laplace = param$laplace, usekernel = FALSE)
- Laplace: 0
- Classes: 2
- Samples: 230
- Features: 13
- Conditional distributions:
    - Gaussian: 13
- Prior probabilities:
    - 0: 0.5565
    - 1: 0.4435
```

Στη συνέχεια για το πλαίσιο δεδομένων ελέγχου έχουμε

```
heart_test_mat <- model.matrix(OUTPUT ~ ., heart_test)[,-1]
test_predictions = predict(heartBayes, newdata = heart_test_mat,type = "raw")
test_class_predictions = as.numeric(test_predictions > 0.5)
mean(test_class_predictions == heart_test$OUTPUT)
(confusion_matrix <- table(predicted = test_class_predictions, actual =
heart_test$OUTPUT))
```

```
         actual
predicted  0   1
        0 18   4
        1  3  15
```

```
(precision <- confusion_matrix[2, 2] / sum(confusion_matrix[2,]))
[1] 0.8182
```

```
(recall <- confusion_matrix[2, 2] / sum(confusion_matrix[,2]))
[1] 0.8571
```

```
(f = 2 * precision * recall / (precision + recall))
[1] 0.8372

(specificity <- confusion_matrix[1,1]/sum(confusion_matrix[1,]))
[1] 0.7895
```

Δηλαδή επιτύχαμε ακρίβεια 82,50 %

3.6 Εκπαίδευση Πλαισίου Δεδομένων με Δέντρα Απόφασης

3.6.1 C50

Σε αυτή τη περίπτωση απαιτούνται οι βιβλιοθήκες rJava, RWeka και η C50. Με το expand.grid πρέπει να καθορίσουμε τις μεταβλητές trials και winnow. Το trials αποτελεί ακέραιο αριθμό που καθορίζει μια μέθοδο μηχανικής μάθησης για κατηγοριοποίηση που ονομάζεται **επαναληπτική ώθηση** (boosting iteration). Ο αλγόριθμος C5 έχει επίσης μια μέθοδο που ονομάζεται «**winnow**» για να αναδείξει τις μεταβλητές που δίνουν την περισσότερη πληροφορία στο πλαίσιο δεδομένων. Η επιλογή κατά τη διάρκεια εκτέλεσης του προγραμματιστικού κώδικα, είναι είτε TRUE (εφαρμογή μεθόδου winnow), είτε FALSE (χωρίς την εφαρμογή της μεθόδου winnow) .

```
install.packages("rJava")
install.packages("RWeka")
install.packages("C50")
```

Στην συνέχεια και προκειμένου να τρέξει σωστά ο αλγόριθμος C50, είναι απαραίτητο να μετατρέψουμε τη μεταβλητή εξόδου OUTPUT σε κατηγορική μεταβλητή (factor):

```
heart$OUTPUT = factor(heart$OUTPUT)
str(heart$OUTPUT)
```

και ο προγραμματιστικός κώδικας για τον αλγόριθμο C50 είναι ο εξής :

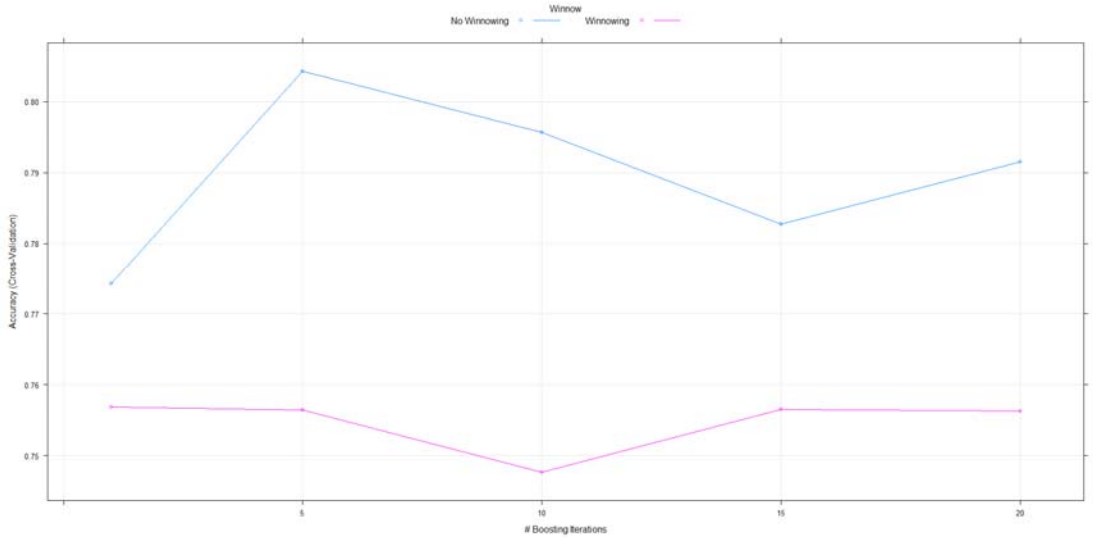
```
library(caret)
library(C50)
library(plyr)

heart <- fread('http://archive.ics.uci.edu/ml/machine-learning-
databases/statlog/heart/heart.dat')
names(heart) <- c("AGE", "SEX", "CHESTPAIN", "RESTBP", "CHOL",
 "SUGAR", "ECG", "MAXHR", "ANGINA", "DEP", "EXERCISE", "FLUOR",
 "THAL", "OUTPUT")

heart$OUTPUT = factor(heart$OUTPUT)
```

```
str(heart$OUTPUT)

head(heart)

library(caret)
set.seed(987954)
heart_sampling_vector <-
createDataPartition(heart$OUTPUT, p = 0.85, list = FALSE)
heart_train <- heart[heart_sampling_vector,]
heart_train_labels <- heart$OUTPUT[heart_sampling_vector]
heart_test <- heart[-heart_sampling_vector,]
heart_test_labels <- heart$OUTPUT[-heart_sampling_vector]

Grid <- expand.grid( .winnow = c(TRUE,FALSE), .trials=c(1,5,10,15,20), .model="tree" )
fitControl <- trainControl(method = "cv", number = 6)

heartC50 <- train(factor(OUTPUT)~.,
                  data=heart_train,
                  method = 'C5.0',
                  trControl = fitControl,
                  metric = "Accuracy",
                  tuneGrid = Grid)
```

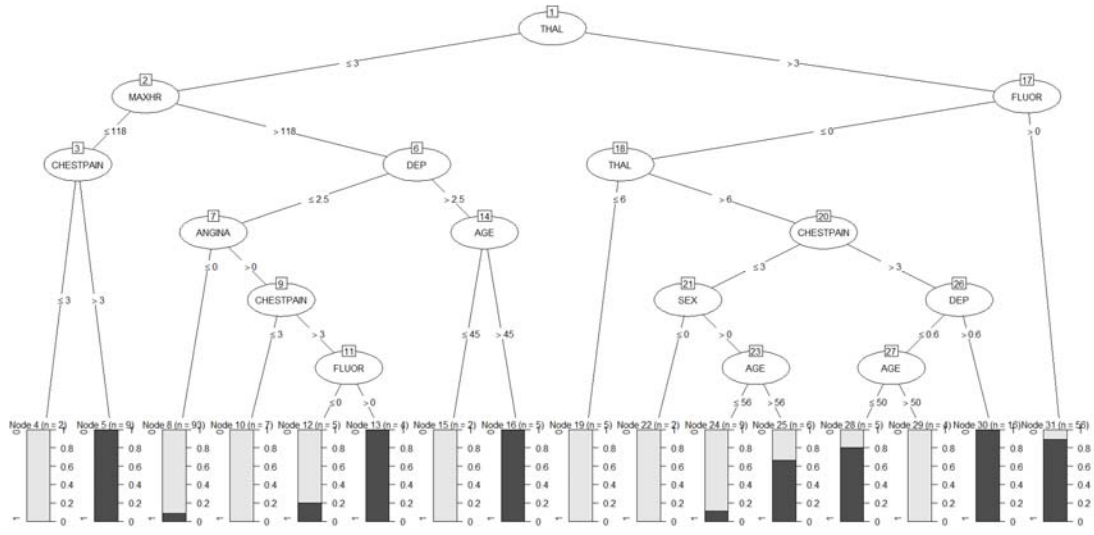
```
C5.0

230 samples
 13 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (6 fold)
Summary of sample sizes: 192, 192, 192, 191, 191, 192, ...
Resampling results across tuning parameters:

  winnow  trials  Accuracy   Kappa
  FALSE    1      0.7742915  0.5420027
  FALSE    5      0.8043185  0.6064609
  FALSE   10      0.7956590  0.5854453
  FALSE   15      0.7827260  0.5584615
  FALSE   20      0.7914980  0.5757064
   TRUE    1      0.7568601  0.5095831
   TRUE    5      0.7564103  0.5105318
   TRUE   10      0.7476383  0.4929054
   TRUE   15      0.7565227  0.5082672
   TRUE   20      0.7562978  0.5096859

Tuning parameter 'model' was held constant at a value of tree
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 5, model = tree and winnow = FALSE.
```

```
heart_predictions <- predict(heartC50, heart_test)
mean(heart_test$OUTPUT == heart_predictions)
(confusion_matrix <- table(predicted = heart_predictions, actual = heart_test$OUTPUT))
(precision <- confusion_matrix[2, 2] / sum(confusion_matrix[2,]))
0.8666667
(recall <- confusion_matrix[2, 2] / sum(confusion_matrix[,2]))
0.7222222
(f = 2 * precision * recall / (precision + recall))
0.7878788
(specificity <- confusion_matrix[1,1]/sum(confusion_matrix[1,]))
0.8
```

Επιλέγεται η τιμή trials =5 για **ακρίβεια μοντέλου 82,5%**



Εικόνα 32 Γραφική παράσταση ακρίβειας για 20 επαναληπτικές ωθήσεις



Εικόνα 33 Δέντρο απόφασης για αλγόριθμο C50

3.6.2 CART

Στο πακέτο caret, με τη προσθήκη στη συνάρτηση train του ορίσματος method = 'rpart' μπορούμε να υπολογίσουμε δέντρο απόφασης με τον αλγόριθμο CART σε σχέση με μια βέλτιστη τιμή του συντελεστή πολυπλοκότητας. Η παράμετρος πολυπλοκότητας (c_p) χρησιμοποιείται για τον έλεγχο και την επιλογή του καλύτερου μεγέθους του δέντρου απόφασης. Με άλλα λόγια η τιμή που υπολογίζεται για το c_p από τον αλγόριθμο προέρχεται από την εφαρμογή διασταυρούμενης επικύρωσης. Εάν το κόστος προσθήκης μιας άλλης μεταβλητής στο δέντρο αποφάσεων από τον τρέχοντα κόμβο, είναι πάνω από την τιμή του cp, τότε το δέντρο δεν συνεχίζεται. Τα αποτελέσματα επίσης, περιλαμβάνουν τιμές από τα πιο γνωστά μεγέθη αξιολόγησης. Στα μοντέλα παλινδρόμησης, οι πιο γνωστές μετρήσεις αξιολόγησης περιλαμβάνουν:

R-square (R2), το οποίο είναι το ποσοστό της διακύμανσης στο αποτέλεσμα που εξηγείται από τις μεταβλητές πρόβλεψης (προσδιοριστές). Σε μοντέλα πολλαπλής παλινδρόμησης, το R2 αντιστοιχεί στη τετραγωνική συσχέτιση μεταξύ των παρατηρούμενων τιμών αποτελεσμάτων και των προβλεπόμενων τιμών από το μοντέλο. Όσο υψηλότερο είναι το τετράγωνο R, τόσο καλύτερο το μοντέλο.

Root Mean Squared Error (RMSE), το οποίο μετρά το μέσο σφάλμα που εκτελεί το μοντέλο στη διαδικασία πρόβλεψης του αποτελέσματος για μια παρατήρηση. Όσο χαμηλότερο το RMSE, τόσο καλύτερο το μοντέλο.

Μέσο Απόλυτο Σφάλμα (MAE), όπως το RMSE, το MAE μετρά το σφάλμα πρόβλεψης. Μαθηματικά, είναι η μέση απόλυτη διαφορά μεταξύ των παρατηρούμενων και των προβλεπόμενων αποτελεσμάτων. Το MAE είναι λιγότερο ευαίσθητο σε ακραίες τιμές σε σύγκριση με το RMSE.

Ακολουθεί ο προγραμματιστικός κώδικας για το δέντρο απόφασης με CART

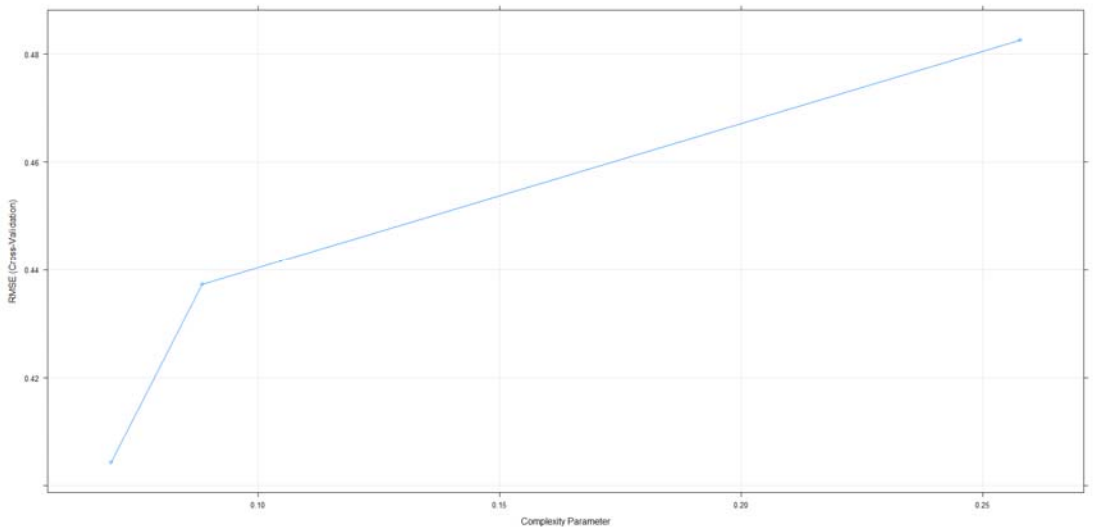
```
library(caret)

heart <- fread('http://archive.ics.uci.edu/ml/machine-learning-
databases/statlog/heart/heart.dat')
names(heart) <- c("AGE", "SEX", "CHESTPAIN", "RESTBP", "CHOL",
 "SUGAR", "ECG", "MAXHR", "ANGINA", "DEP", "EXERCISE", "FLUOR",
 "THAL", "OUTPUT")

head(heart)

library(caret)
set.seed(987954)
heart_sampling_vector <-
```

```
createDataPartition(heart$OUTPUT, p = 0.85, list = FALSE)
heart_train <- heart[heart_sampling_vector,]
heart_train_labels <- heart$OUTPUT[heart_sampling_vector]
heart_test <- heart[-heart_sampling_vector,]
heart_test_labels <- heart$OUTPUT[-heart_sampling_vector]

Grid <- expand.grid(cp=seq(0, 0.2, 0.001))
fitControl <- trainControl(method = "cv", number = 6)

heartCART <- train(factor(OUTPUT)~.,
                   data=heart_train,
                   method='rpart',
                   trControl = fitControl,
                   metric = "Accuracy",
                   tuneGrid = Grid,
                   na.action = na.omit,
                   parms=list(split='Gini'))
```

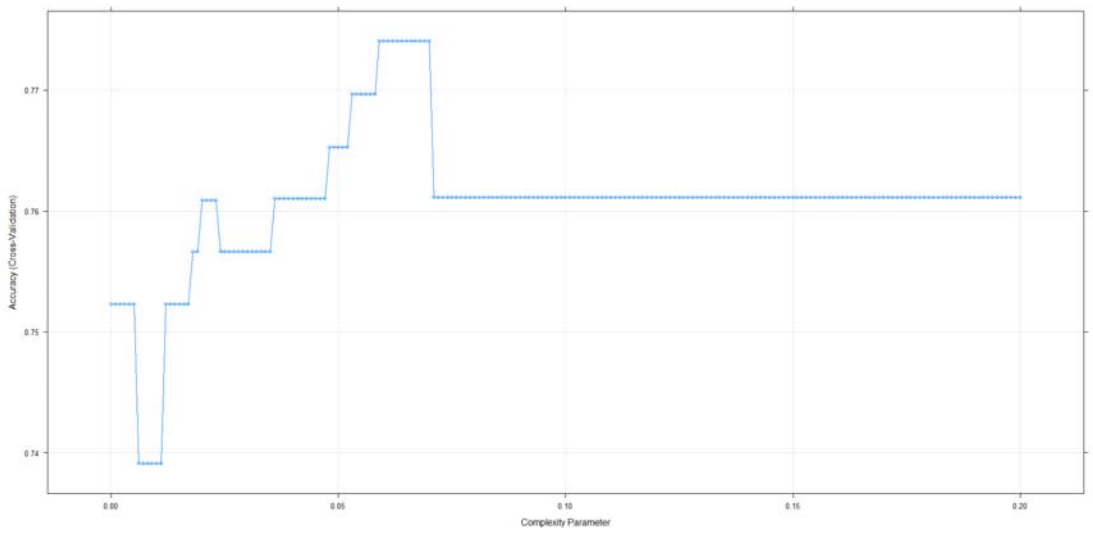
και τα αποτελέσματα είναι :

```
CART

230 samples
 13 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 207, 207, 207, 207, 207, 207, ...
Resampling results across tuning parameters:

  cp         RMSE       Rsquared   MAE
  0.0695685  0.4220265  0.3137964  0.3217973
  0.0884098  0.4472062  0.2292371  0.3672355
  0.2577310  0.4869450  0.1526044  0.4466210

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.0695685.
```

Με βάση τα προηγούμενα επιλέγουμε την τιμή $c_p = 0.0695685$ για το κλάδεμα του δέντρου.



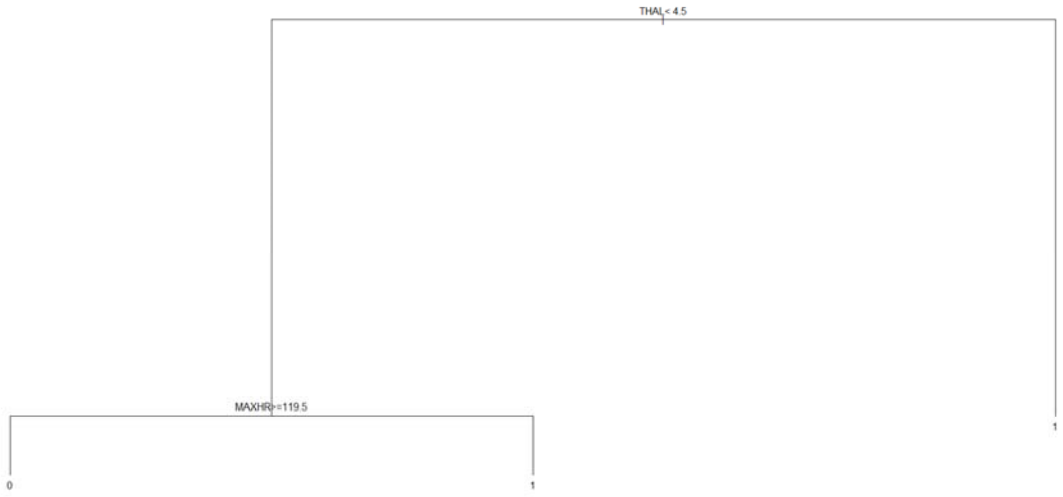
Εικόνα 34 Γραφική παράσταση του συντελεστή πολυπλοκότητας σε σχέση με το συντελεστή RMSE



Εικόνα 35 Γραφική παράσταση της ακρίβειας του μοντέλου σε σχέση με την παράμετρο πολυπλοκότητας

Για το σετ δεδομένων εκπαίδευσης , κατα συνέπεια έχουμε **72,5 % ακρίβεια**

```
heart_predictions <- predict(heartCART, heart_test)
mean(heart_test$OUTPUT == heart_predictions)
0.725
(precision <- confusion_matrix[2, 2] / sum(confusion_matrix[2,]))
[1] 0.7692308
(recall <- confusion_matrix[2, 2] / sum(confusion_matrix[,2]))
[1] 0.8571
(f = 2 * precision * recall / (precision + recall))
[1] 0.6451613
(specificity <- confusion_matrix[1,1]/sum(confusion_matrix[1,]))
0.7037037
```



Εικόνα 36 Δέντρο απόφασης για αλγόριθμο CART

3.7 Υλοποίηση της εφαρμογής στο προγραμματιστικό περιβάλλον της R

Το πακέτο Shiny χρησιμοποιείται για τη δημιουργία διαδραστικών διαδικτυακών εφαρμογών με χρήση της γλώσσας προγραμματισμού R. Η εφαρμογή Shiny εμπεριέχεται σε ένα script που καλείται app.R.Το script app.R βρίσκεται σε ένα directory (για παράδειγμα, newdir/) και η εφαρμογή μπορεί να τρέξει με την εντολή runApp("newdir").

Το app.R έχει περιλαμβάνει τρεις συναρτήσεις:

- Τη συνάρτηση διεπαφής του χρήστη (ui<) που μπορούμε να σχεδιάσουμε την εφαρμογή.
- Τη συνάρτηση του server (server<) που υπάρχουν οι οδηγίες για τη σύνθεση της εφαρμογής.
- Τη συνάρτηση shinyApp για να τρέξουν οι δυο παραπάνω, δηλαδή είναι για να καλούμε τις δυο παραπάνω συναρτήσεις.

Ο στόχος εδώ είναι να σχεδιαστεί μια web εφαρμογή που θα τρέχει έναν αλγόριθμο Μηχανικής Μάθησης στο background. Η συνάρτηση διεπαφής του χρήστη (ui<) θα επιτρέπει στο χρήστη να επιλέγει τιμές με βάση τους προσδιοριστές του πλαισίου δεδομένων και πατώντας του submit κουμπί θα γίνεται πρόβλεψη.

Δημιουργείται αρχικά ένα dataframe στο ui. Στη συνέχεια δημιουργούμε το αποθηκεύουμε, με τη βοήθεια της write.table, σαν input.csv φάκελο.

```
OUTPUT <- "OUTPUT"
df <- rbind(df, OUTPUT)
input <- transpose(df)
write.table(input,"input.csv", sep=",", quote = FALSE, row.names = FALSE, col.names = FALSE)
```

Αφού διαβαστούν τα δεδομένα και δομηθεί το μοντέλο, πατώντας το κουμπί submit στέλνονται οι τιμές εισόδου από το ui στη συνάρτηση του server. Ο server θα χρησιμοποιήσει αυτές τις τιμές για να τροφοδοτήσει το προγνωστικό μοντέλο με τις νέες τιμές από τους 13 προσδιοριστές :

```
mainPanel(
tags$label(h3('Status/Output')), # Status/Output Text Box
verbatimTextOutput('contents'),
tableOutput('tabledata'), # Prediction results table
```

Τα αποτελέσματα βρίσκονται στο tabledata και απεικονίζονται πάνω δεξιά της εφαρμογής.

Εικόνα 37 Αποτελέσματα από την εκτέλεση της εφαρμογής

Όλες οι τιμές από τις παραμέτρους εισόδου περνάνε από το ui στο server εδώ :

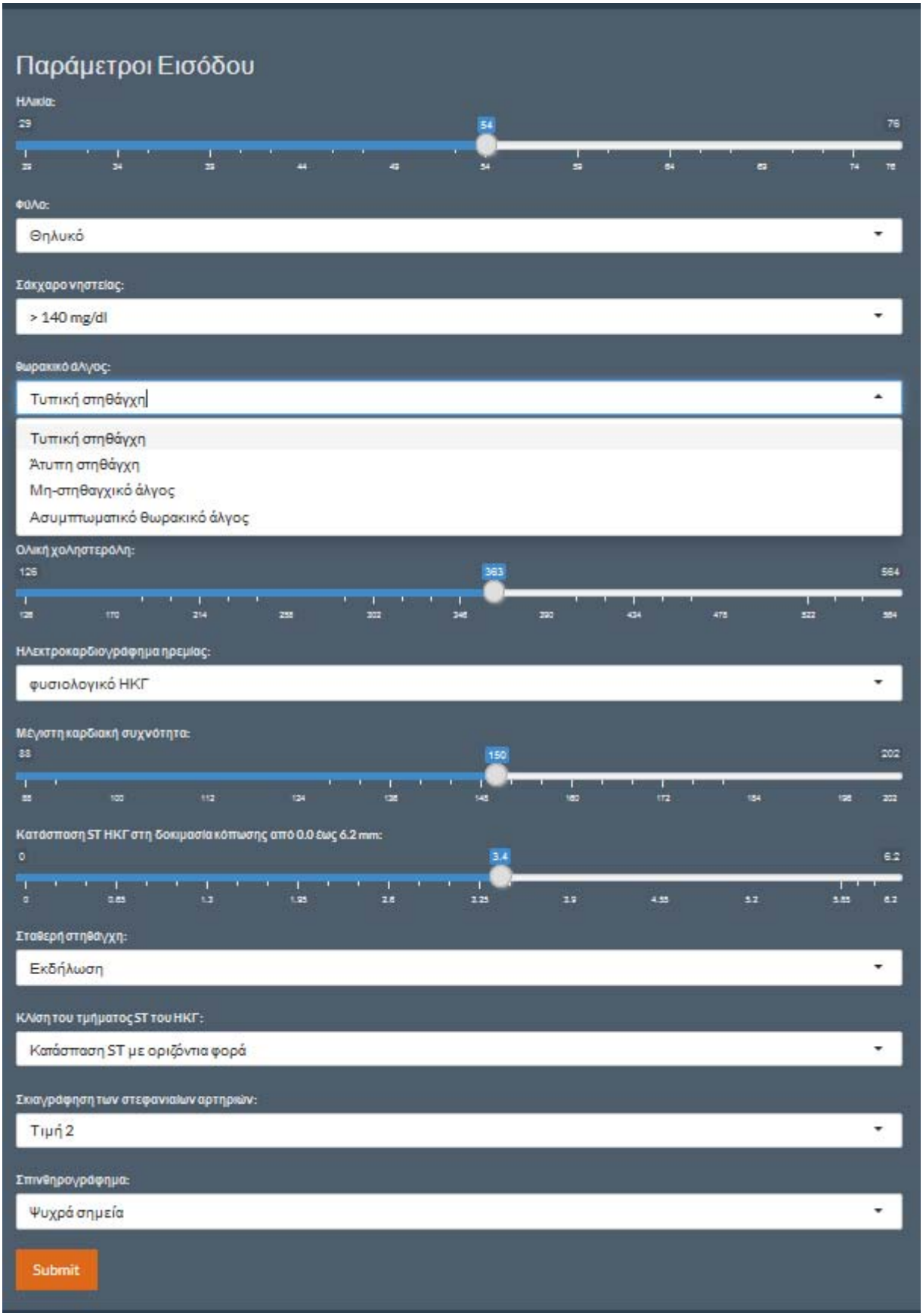
```
OUTPUT <- "OUTPUT"
df <- rbind(df, OUTPUT)
input <- transpose(df)
write.table(input,"input.csv", sep=",", quote = FALSE, row.names = FALSE, col.names =
FALSE)
test <- read.csv(paste("input", ".csv", sep=""), header = TRUE)
Output <- data.frame(Πρόβλεψη = predict(heartmodel, test),
round(predict(heartmodel,test,type="prob"), 2))
names(Output) <- c('Πρόβλεψη', 'ΑΠΟΥΣΙΑ', 'ΠΑΡΟΥΣΙΑ')
print(Output)
```

Όταν γίνει η πρόβλεψη από το μοντέλο μας, η τιμή βρίσκεται στο datasetInput

```
# Prediction results table
output$tabledata <- renderTable({
if (input$submitbutton>0) {
isolate(datasetInput())
```

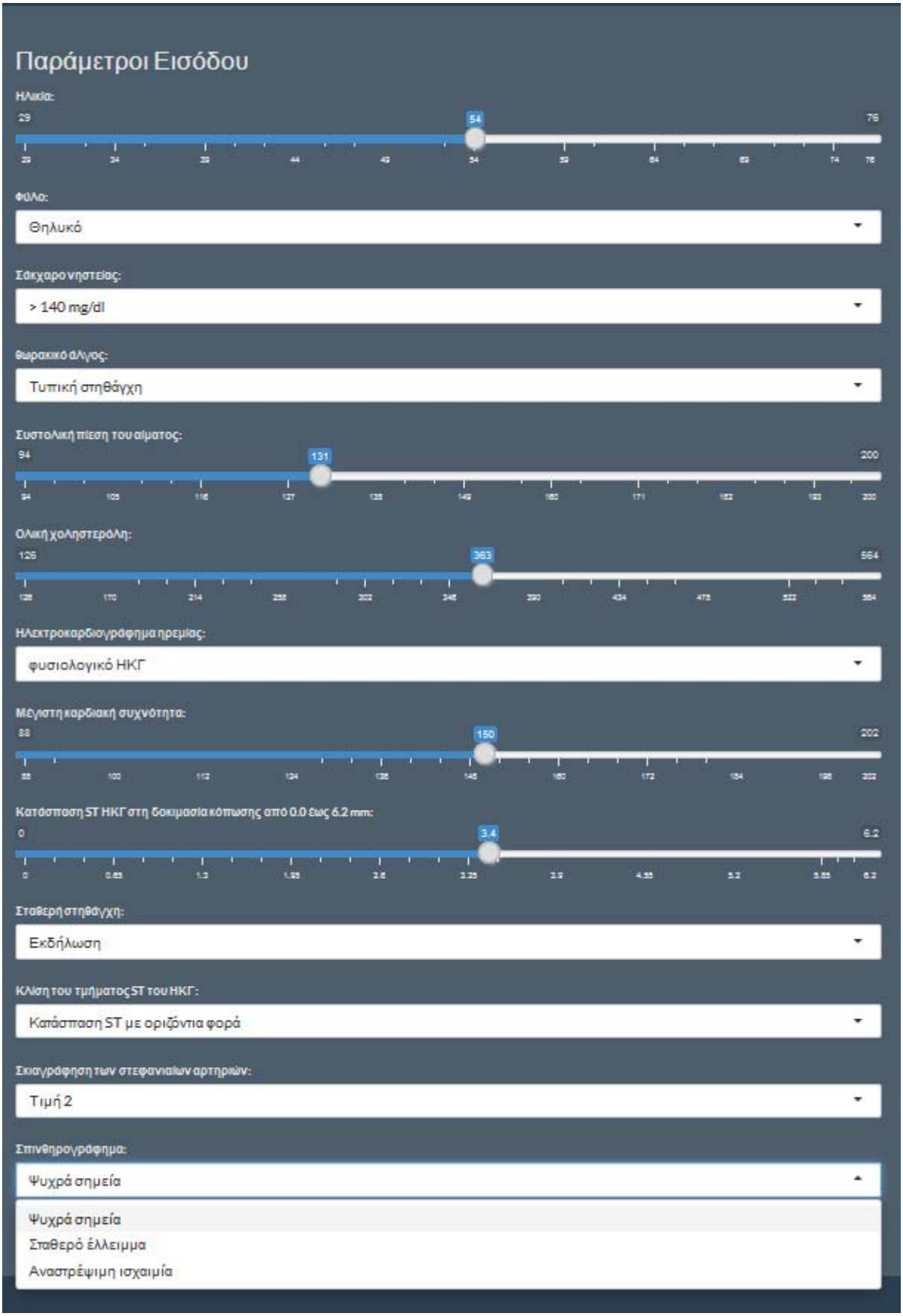
Στη συνέχεια το output$tabledata που περιέχει τη τιμή της πρόβλεψης, στέλνεται στο ui για να απεικονισθεί αυτή η τιμή. Παρακάτω ακολουθούν διάφορες εικόνες της εφαρμογής, για διάφορες τιμές παραμέτρων.

Εικόνα 38 Παράμετροι εισόδου πρώτη περίπτωση

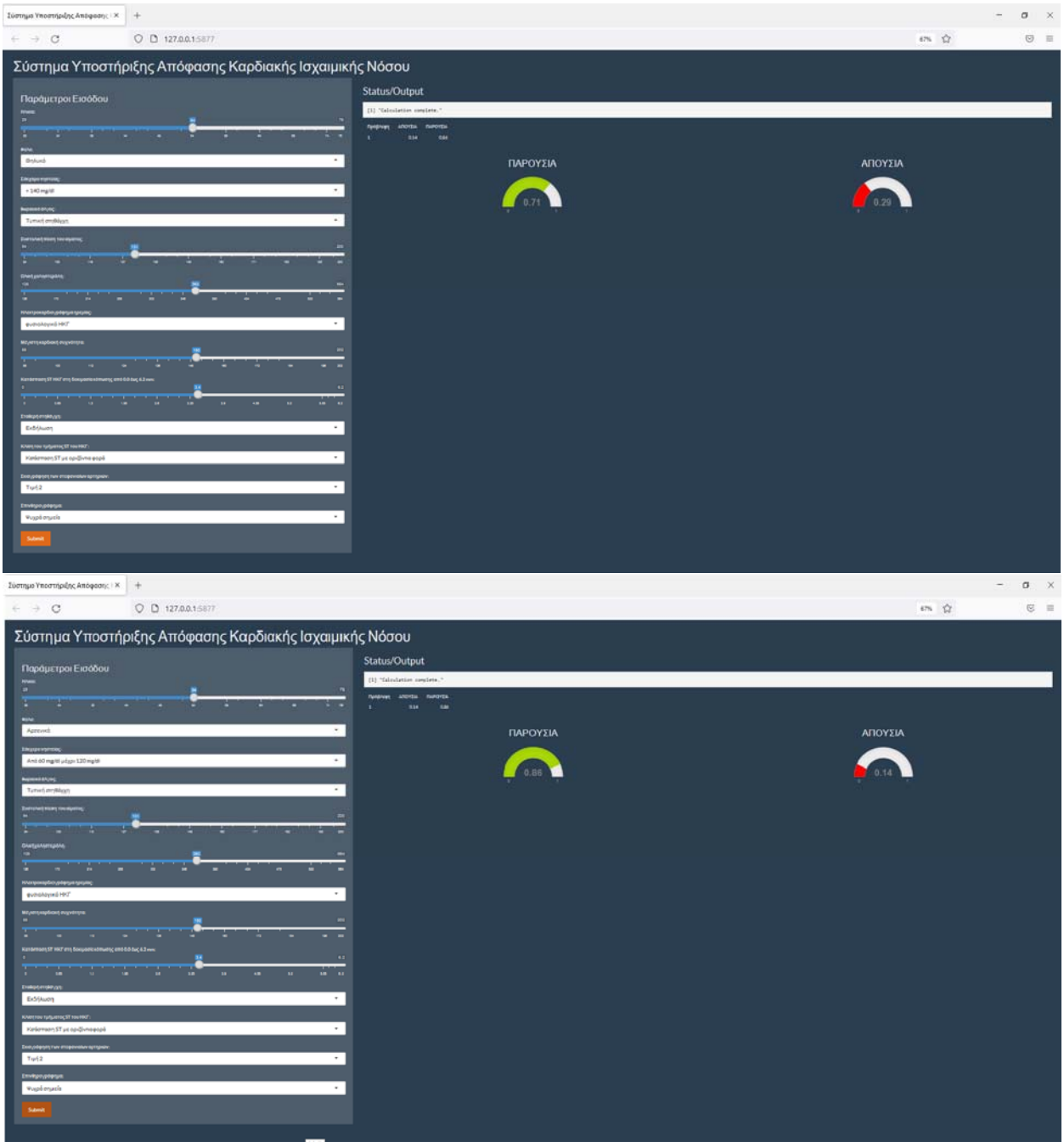


Εικόνα 39 Παράμετροι εισόδου δεύτερη περίπτωση

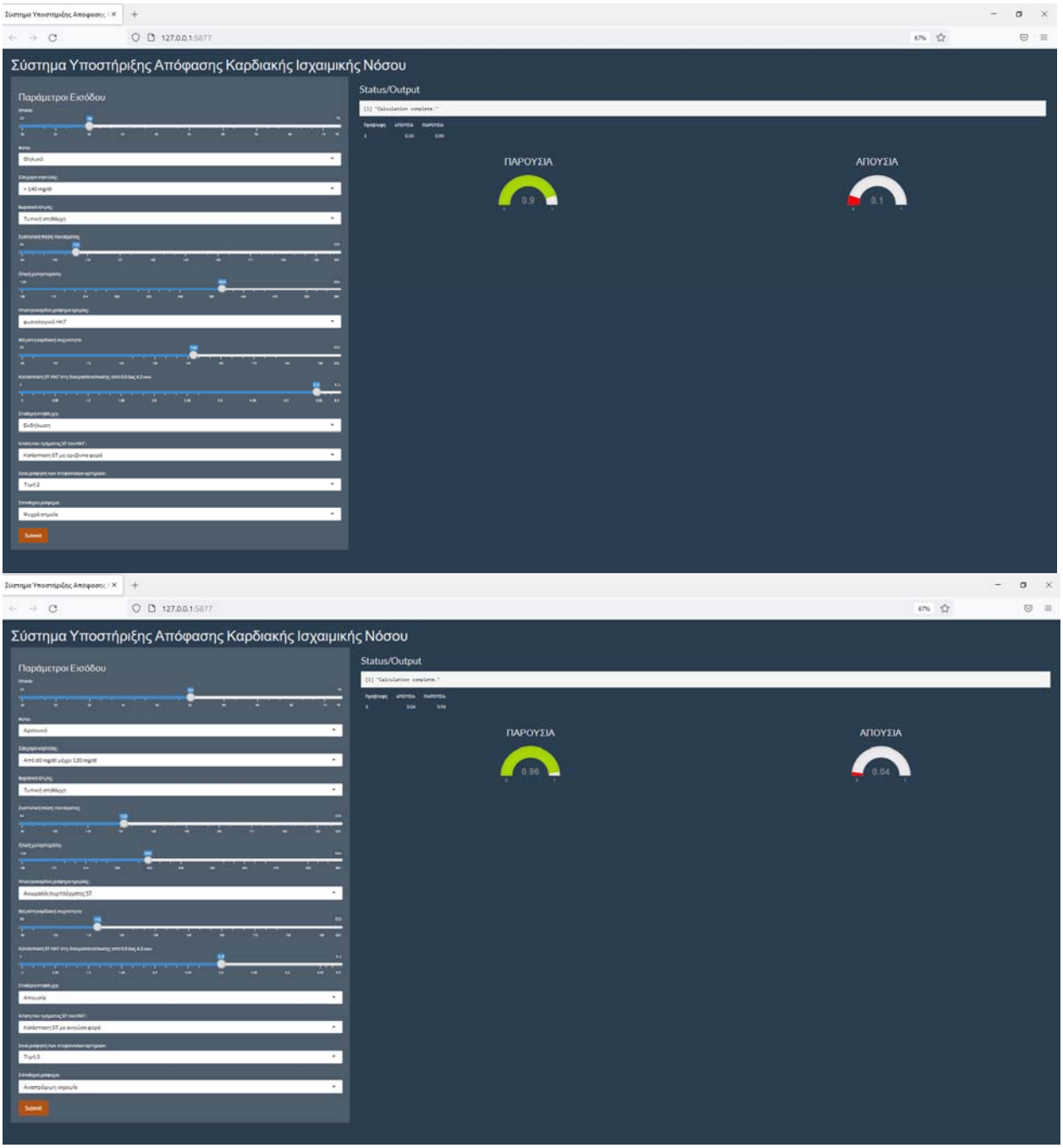
Εικόνα 40 Παράμετροι εισόδου τρίτη περίπτωση



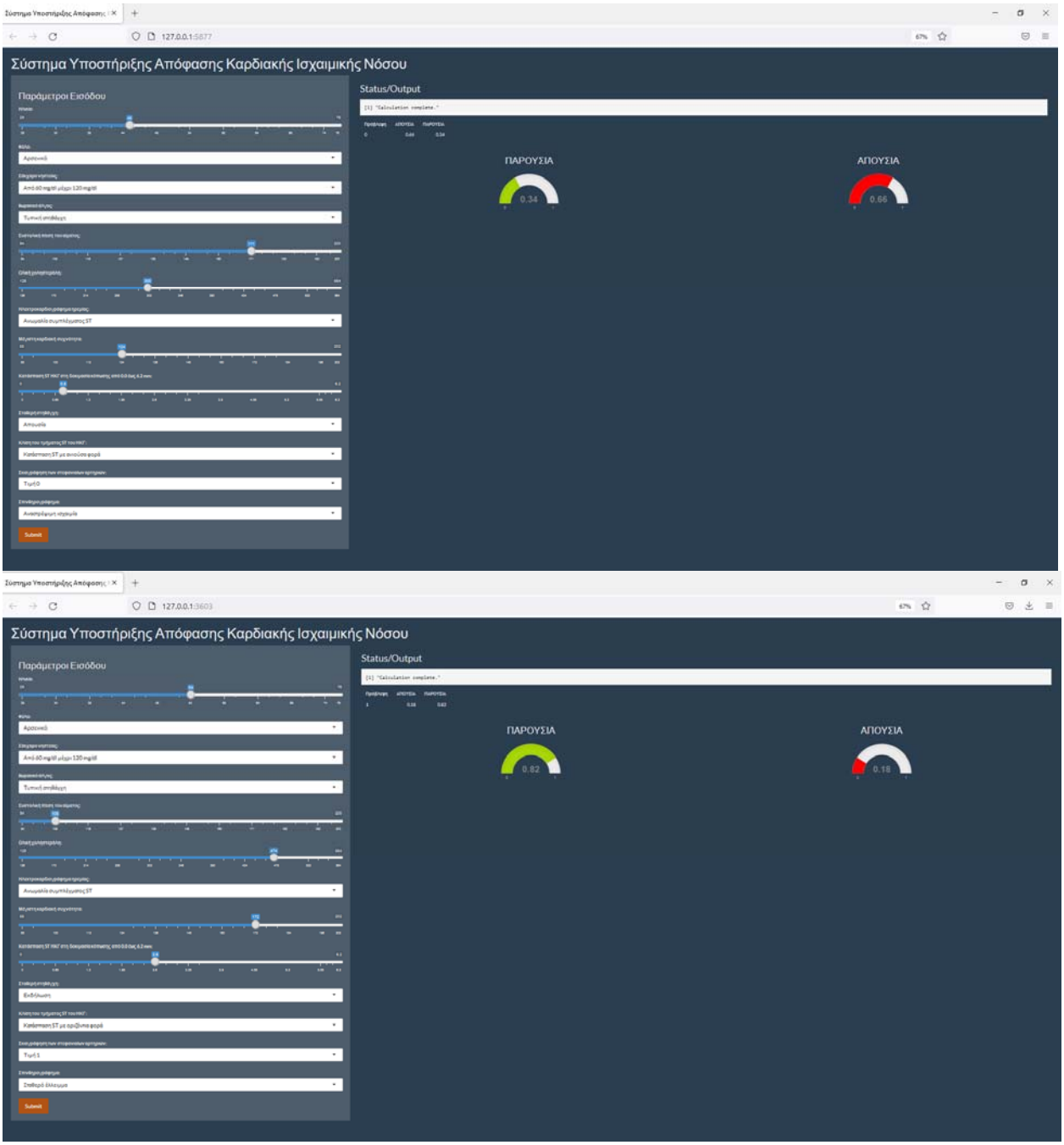
Εικόνα 41 Παράμετροι εισόδου τέταρτη περίπτωση



Εικόνα 42 Φωτογραφίες από την εφαρμογή 1



Εικόνα 43 Φωτογραφίες από την εφαρμογή 2



Εικόνα 44 Φωτογραφίες από την εφαρμογή

4. Αποτελέσματα

Πίνακας 6 Συγκεντρωτικός πίνακας συχνοτήτων

	OUTPUT			
SEX	**Απουσία**	**Παρουσία**	**Σύνολα**	**Ερμηνεία**
Γυναίκες	67 (77.0%)	20 (23.0%)	87 (100.0%)	Οι άντρες εμφανίζουν περισσότερο ποσοστό
Άντρες	83 (45.4%)	100 (54.6%)	183 (100.0%)	εμφάνισης της νόσου σε σχέση με τις γυναίκες
THAL				
Πλήρως αποφραγμένες αρτηρίες	119 (78.3%)	33 (21.7%)	152 (100.0%)	Σε ποσοστό περίπου 76% η ισχαιμία είναι αναστρέψιμη.
Περιοχή που έχει νεκρωθεί	6 (42.9%)	8 (57.1%)	14 (100.0%)	
Αναστρέψιμη ισχαιμία	25 (24.0%)	79 (76.0%)	104 (100.0%)	
EXERCISE (στη δοκιμασία κόπωσης)				
ST με ανιούσα φορά	98 (75.4%)	32 (24.6%)	130 (100.0%)	Το 75% με εύρημα ΗΚΓ ελαφρώς ανιούσα ST κλίση παρουσιάζει απουσία νόσου.
ST με οριζόντια φορά	44 (36.1%)	78 (63.9%)	122 (100.0%)	
ST με κατιούσα φορά	8 (44.4%)	10 (55.6%)	18 (100.0%)	
SUGAR (mg/dl)				
<120	127 (84.7%)	103 (44.8%)	230 (100.0%)	Φυσιολογικές τιμές στο σάκχαρο νηστείας αποτρέπουν την εκδήλωση της νόσου.
>120	23 (57.5%)	17 (42.5%)	40 (100.0%)	
RESTBP (mmHg)				
94-142	127 (59.7.%)	88 (40.9%)	215 (100.0%)	Φυσιολογικές τιμές αρτηριακής πίεσης αποτρέπουν την εκδήλωση της νόσου.
144-200	23 (41.8%)	32 (58.2%)	55 (100.0%)	
ANGINA				
Απουσία στηθάγχης	127 (70.2%)	54 (29.8%)	181 (100.0%)	Η εκδήλωση στηθάγχης αποτελεί ισχυρό προσδιοριστή ύπαρξης της νόσου.
Εκδήλωση στηθάγχης	23 (25.8%)	66 (74.2%)	89 (100.0%)	
FLUOR				
Δείκτης στεφανιαιογραφίας 0	120 (75.0%)	40 (25.0%)	160 (100.0%)	Εύρημα Στεφανιογραφίας με δείκτη 3 είναι ισχυρός παράγοντας εκδήλωσης καρδιακής ισχαιμικής νόσου.
Δείκτης στεφανιαιογραφίας 1	20 (34.5%)	38 (65.5%)	58 (100.0%)	
Δείκτης στεφανιαιογραφίας 2	7 (21.2%)	26 (78.8%)	33 (100.0%)	
Δείκτης στεφανιαιογραφίας 3	3 (15.8%)	16 (84.2%)	19 (100.0%)	
ECG (στην ηρεμία)				
Φυσιολογικό ΗΚΓ	85 (64.9%)	46 (35.1%)	131 (100.0%)	Εύρημα ΗΚΓ με ανάσπαση συμπλέγματος ST>1 αποτελεί ισχυρή ένδειξη εμφάνισης της νόσου.
Ανωμαλία συμπλέγματος ST	1 (50.0%)	1 (50.0%)	2 (100.0%)	
Ανάσπαση συμπλέγματος ST>1	64 (46.7%)	73 (53.3%)	137 (100.0%)	
CHESTPAIN				
Τυπική στηθάγχη	15 (75.0%)	5 (25.0%)	20 (100.0%)	Το ασυμπτωματικό θωρακικό άλγος συνδέεται με την αθηρωματική νόσο.
Άτυπη στηθάγχη	35 (83.3%)	7 (16.7%)	42 (100.0%)	
Μη-στηθαγχικό άλγος	62 (78.5%)	17 (21.5%)	79 (100.0%)	
Ασυμπτωματικό θωρακικό άλγος	38 (29.5%)	91 (70.5%)	129 (100.0%)	
DEP (mm/sec)				
0-0.9	106 (73.1%)	39 (26.9%)	145 (100.0%)	Εύρημα ΗΚΓ με κατάσπαση ST
1-1.9	35 (48.0%)	38 (52.0%)	73 (100.0%)	>3 mm/sec αποτελεί ισχυρή ένδειξη
2-2.9	6 (19.4%)	25 (80.6%)	31 (100.0%)	εμφάνισης της νόσου
3-3.8	2 (13.3%)	13 (86.7%)	15 (100.0%)	
4-6.2	1 (16.7%)	5 (83.3%)	6 (100.0%)	
AGE (years)				
29-45	43 (76.8%)	13 (23.2%)	56 (100%)	Άνδρες άνω των 55 ετών με ασυμπτωματικό
46-55	51 (62.2%)	31 (37.8%)	82 (100%)	θωρακικό άλγος έχουν ποσοστά άνω του 55%
56-77	56 (42.4%)	76 (57.6%)	132 (100%)	
MAXHR (bpm)				
71-120	6 (17.7%)	28 (82.3%)	34 (100%)	Σε υψηλά επίπεδα καρδιακών παλμών >121bpm

121-150	36 (40.9%)	52 (59.1%)	88 (100%)	παρουσιάζεται αυξημένος κίνδυνος εμφάνισης
151-202	108 (73.0%)	40 (27.0%)	148 (100%)	της νόσου (60%)
CHOL (mg/dL)				
126-200	25 (62.5%)	15 (37.5%)	40 (100%)	Αυξημένα επίπεδα χοληστερόλης >240 mg/dL
201-240	56 (65.1%)	30 (34.9%)	86 (100%)	αποτελούν σημαντικό προδιαθεσικό
242-564	69 (47.9%)	75 (52.1%)	144 (100%)	παράγοντα εκδήλωσης της νόσου

Πίνακας 7 Στατιστικά μεταβλητών πλαισίου δεδομένων

Μεγέθη	AGE	ANGINA	CHESTPAIN	CHOL	DEP	ECG	EXERCISE	FLUOR
Μέσος όρος	54.43	0.33	3.17	249.66	1.05	1.02	1.59	0.67
Τυπική απόκλιση	9.11	0.47	0.95	51.69	1.15	1.00	0.61	0.94
Ελάχιστο	29.00	0.00	1.00	126.00	0.00	0.00	1.00	0.00
Διάμεσος	55.00	0.00	3.00	245.00	0.80	2.00	2.00	0.00
Μέγιστο	77.00	1.00	4.00	564.00	6.20	2.00	3.00	3.00
Σύνολα	270.00	270.00	270.00	270.00	270.00	270.00	270.00	270.00

Μεγέθη	MAXHR	OUTPUT	RESTBP	SEX	SUGAR	THAL
Μέσος όρος	149.68	0.44	131.34	0.68	0.15	4.70
Τυπική απόκλιση	23.17	0.50	17.86	0.47	0.36	1.94
Ελάχιστο	71.00	0.00	94.00	0.00	0.00	3.00
Διάμεσος	153.50	0.00	130.00	1.00	0.00	3.00
Μέγιστο	202.00	1.00	200.00	1.00	1.00	7.00
Σύνολα	270.00	270.00	270.00	270.00	270.00	270.00

ΣΥΓΚΕΝΤΡΩΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Πίνακας 8 Συγκεντρωτικά αποτελέσματα για τη Λογιστική Παλινδρόμηση

Λογιστική Παλινδρόμηση

Ταξινομητής	Overall Accuracy %	Specificity (%)	Sensitivity (%)	Time (s)
Λογιστική Παλινδρόμηση	88.7	84.3	92.2	1.00
Λογιστική Παλινδρόμηση με LASSO κανονικοποίηση †	92.0	87.4	95.6	1.00

Πίνακας 9 Συγκεντρωτικά αποτελέσματα για τις Μηχανές Διανυσμάτων Υποστήριξης

Μηχανές Διανυσμάτων Υποστήριξης

Ταξινομητής	Overall Accuracy %	Specificity (%)	Sensitivity (%)	Time (s)
SVM με γραμμικό πυρήνα†	85.0	72.2	95.5	1.00
SVM με πυρήνα ακτινικής βάσης ††	87.5	72.2	98.0	>20
SVM με πολυωνυμικό πυρήνα†††	80.0	72.2	86.4	>20

Πίνακας 10 Συγκεντρωτικά αποτελέσματα για τα Νευρωνικά Δίκτυα

Νευρωνικά Δίκτυα

Ταξινομητής	Overall Accuracy %	Specificity (%)	Sensitivity (%)	Time (s)
Νευρωνικό Δίκτυο Perceptron	92.50	88.89	95.45	>5.00

Πίνακας 11 Συγκεντρωτικά αποτελέσματα για Μπεϋζιανό ταξινομητή

Μπεϋζιανός ταξινομητής

Ταξινομητής	Overall Accuracy %	Specificity (%)	Sensitivity (%)	Time (s)
Ταξινομητής Μπέϋζ†	82.5	79.0	85.7	1.00

Πίνακας 12 Συγκεντρωτικά αποτελέσματα για τα Δέντρα Απόφασης

Δέντρα Απόφασης

Ταξινομητής	Overall Accuracy %	Specificity (%)	Sensitivity (%)	Time (s)
Αλγόριθμος C50†	82.5	72.2	91.0	>5.00
Αλγόριθμος CART††	72.0	55.0	85.6	>5.00

Πίνακας 13 Συγκεντρωτικά αποτελέσματα για όλους τους αλγόριθμους Μηχανικής Μάθησης της εργασίας

Οι υψηλότερες τιμές ακρίβειας που επιτεύχθηκαν από πέντε διαφορετικές περιπτώσεις ταξινομητών

Ταξινομητής	Overall Accuracy %	Specificity (%)	Sensitivity (%)	Time (s)
Λογιστική Παλινδρόμηση με κανονικοποίηση LASSO	90.0	87.4	95.6	1.00
SVM	87.5	72.2	98.0	>20
Νευρωνικό Δίκτυο Perceptron	92.5	79.0	85.7	>5.00
Ταξινομητής Bayes	82.5	79.0	85.7	1.00
Δέντρα Απόφασης	82.5	72.2	91.0	>5.00

5. Συζήτηση & Συμπεράσματα

Η εκπαίδευση του πλαισίου δεδομένων των καρδιακών παθήσεων Statlog με εφαρμογή λογιστικής παλινδρόμησης επέδειξε χαμηλή τιμή για το συντελεστή μέγιστης πιθανοφάνειας R 2 (55%) και αρχική ακρίβεια μοντέλου 77.5%. Προς βελτίωση της ακρίβειας, εφαρμόστηκε η μέθοδος κανονικοποίησης LASSO, που επιβάλλει ένα είδος ποινής στη συνάρτηση πιθανοφάνειας με αποτέλεσμα οι συντελεστές των μεταβλητών στο μοντέλο να συρρικνώνονται. Με διασταυρούμενη επικύρωση δέκα τμημάτων, βρέθηκε η βέλτιστη τιμή για το χαρακτηριστικό μέγεθος "λ=0.02" της LASSO Παλινδρόμησης. Ο νέος έλεγχος απόδοσης ανέδειξε την τιμή της ακρίβειας στο 92.5%. Η λογιστική παλινδρόμηση όμως σαν ταξινομητής παρουσιάζει και μειονεκτήματα. Οι Pochet και Suykens [56], επισήμαναν ότι αυτή η τεχνική δεν είναι σε θέση να προσδιορίσει πιθανές μη γραμμικές δομές σε ένα σύνολο ασθενών. Επίσης, ανέφεραν ότι οι παραδοσιακές στατιστικές μέθοδοι, όπως η λογιστική παλινδρόμηση, σκοπεύουν να οικοδομήσουν ένα μοντέλο ταξινόμησης που να ταιριάζει βέλτιστα σε ένα σύνολο δεδομένων. Αυτό όμως οδηγεί συχνά σε υπερπροσαρμογή του σετ δεδομένων.

Στη συνέχεια, χρησιμοποιήσαμε έναν πολύ ισχυρό αλγόριθμο ταξινόμησης για τη μεγιστοποίηση του περιθωρίου μεταξύ των μεταβλητών κλάσης. Αυτό το περιθώριο (διάνυσμα υποστήριξης/SVM) αντιπροσωπεύει την απόσταση μεταξύ των διαχωριστικών υπερεπιπέδων (όριο απόφασης). Η SVM δεν είναι τόσο επιρρεπής σε ακραίες τιμές, όπως η λογιστική παλινδρόμηση αλλά δίνει βαρύτητα μόνο για τα σημεία που βρίσκονται πιο κοντά στο όριο απόφασης . Αλλάζει το όριο της απόφασης ανάλογα με την τοποθέτηση των νέων θετικών ή αρνητικών προσδιοριστών. Για τις Μηχανές Διανυσματικής Υποστήριξης, η βελτιστοποίηση επιτυγχάνεται ελαχιστοποιώντας το σφάλμα ταξινόμησης στο σετ εκπαίδευσης παράλληλα με μείωση της πολυπλοκότητας του μοντέλου. Η επιλογή των κατάλληλων παραμέτρων μάθησης είναι ένα κρίσιμο βήμα για τον υπολογισμό των βέλτιστων ρυθμίσεων των μηχανών διανυσματικής υποστήριξης. Συνήθως, οι ρυθμίσεις αυτών των παραμέτρων βασίζονται στη λεγόμενη αναζήτηση πλέγματος (grid-search). Δηλαδή, για κάθε παράμετρο καθορίζεται ένας πεπερασμένος αριθμός πιθανών τιμών και στη συνέχεια όλοι οι πιθανοί συνδυασμοί αυτών των τιμών, θεωρείται ότι βρίσκουν αυτόν που αποδίδει το καλύτερο αποτέλεσμα. Επομένως, η πολυπλοκότητα της αναζήτησης πλέγματος αυξάνεται εκθετικά με τον αριθμό των παραμέτρων. Αυτός είναι ο κύριος λόγος για το γεγονός ότι χρησιμοποιήθηκαν στην εργασία μόνο δυο χαρακτηριστικοί παράμετροι μάθησης : Η παράμετρος *C* για το συμβιβασμό μεταξύ της μεγιστοποίησης του περιθωρίου και της ανοχής σε σφάλματα και το «*σ*» ενός τυπικού πυρήνα Gaussian που εκφράζει το βαθμό καμπυλότητας στο όριο απόφασης . Με βέλτιστες τιμές C=0.50, sigma =0.25 καταφέραμε μέγιστη ακρίβεια 87,5 % (πυρήνας ακτινικής βάσης).

Ενώ με τα SVM χρησιμοποιώντας το κόλπο του πυρήνα (kernel trick) στέλνουμε τα δεδομένα σε μια υψηλότερη διάσταση, όπου εκεί γίνονται γραμμικά διαχωρίσιμα, σε μια άλλη μέθοδο Μηχανικής Μάθησης που ονομάζεται τεχνητά νευρωνικά δίκτυα, επιτελείται μια σειρά γραμμικών συνδυασμών που αναμιγνύονται με (συνήθως) μη γραμμικές συναρτήσεις ενεργοποίησης σε διάφορα κρυμμένα επίπεδα (hidden layers) και με παρόμοιο τρόπο πραγματοποιείται ο γραμμικός διαχωρισμός των μετασχηματισμένων δεδομένων Ένα νευρωνικό δίκτυο είναι μια σειρά αλγορίθμων που προσπαθούν να αναγνωρίσουν τις υποκείμενες σχέσεις σε ένα σύνολο δεδομένων μέσω μιας διαδικασίας που μιμείται τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Για τα νευρωνικά δίκτυα χρησιμοποιήθηκε η μέθοδος αποσύνθεσης βαρών (weight decay), ένας πρόσθετος δηλαδή όρος στη χαρακτηριστική εξίσωση του κανόνα ενημέρωσης βάρους που προκαλεί την εκθετική μείωση των βαρών έως το μηδέν. Πρέπει να σημειωθεί ότι μια ορθότερη προσέγγιση θα ήταν να χρησιμοποιηθεί ένας πιο προηγμένος αλγόριθμος βελτιστοποίησης (Levenberg-Marquardt ή των συζυγών κλίσεων), καθώς αυτοί είναι πολύ πιο γρήγοροι, και εκεί δεν χρειάζεται να οριστεί ο ρυθμός μάθησης. Ωστόσο η εργαλειοθήκη της R δεν έχει αναπτύξει ακόμα επαρκώς τέτοια προγραμματιστικά εργαλεία (εκτός από τη βιβλιοθήκη minpack.lm που υπάρχουν περιορισμένες αναφορές). Κατα την εκπαίδευση του πλαισίου δεδομένων, με βέλτιστο αριθμό βαρών 10, τιμή μέσου τετράγωνου σφάλματος 0.35 και βήμα αποσύνθεσης βαρών 0.1, επιτύχαμε ακρίβεια 92%

Στη παράγραφο 3.5 μελετήθηκε ο ταξινομητής Μπέυζ, που βασίζεται στην περιγραφή ενός προβλήματος ταξινόμησης με πιθανοτικούς όρους. Σύμφωνα με αυτό τον αλγόριθμο, ο κανόνας Μπέυζ επιτρέπει τον υπολογισμό της εκ των υστέρων πιθανότητας (που είναι δύσκολο να καθοριστεί) από την εκ των προτέρων πιθανότητα, την πιθανοφάνεια, και τις αποδείξεις (που υπολογίζονται ευκολότερα). Στη συνέχεια, σε προγραμματιστικό επίπεδο και με τη βοήθεια του ταξινομητή Μπεϋζ, εκπαιδεύτηκε το μοντέλο και με ειδικές τεχνικές εξομάλυνσης επετεύχθη ακρίβεια 84,0%

Ο τελευταίος ταξινομητής που χρησιμοποιήθηκε ήταν τα Δέντρα Απόφασης. Χρησιμοποιήθηκαν δυο αλγόριθμοι : CART και C5.0. Ο αλγόριθμος CART αποδίδει δυαδικά δέντρα (binary trees) ενώ ο C5.0 σχηματίζει δέντρα με κατάτμηση πολλάπλών δρόμων (multiway splitting). Η χαμηλή «ακρίβεια» (72%) του αλγόριθμου CART οφείλεται στην ίδια την τεχνική του, δηλαδή στη χρήση του «συντελεστή καθαρότητας» και της τεχνικής «κλαδέματος» που δεν ευθυγραμμίζεται με τη μείωση του ποσοστού εσφαλμένης ταξινόμησης, που είναι ο απώτερος στόχος της ανάπτυξης δέντρων απόφασης. (αυτή είναι και μια άποψη που έχουν υποστηρίξει και οι Bertsimas και Dunn [71]).

Όπως έχει λεχθεί στη παράγραφο 3.2, το μοντέλο περιγράφεται μόλις από το 55 % των συνολικών μεταβλητών του. Αυτό οφείλεται κυρίως στην έλλειψη ικανοποιητικού αριθμού μεταβλητών-προσδιοριστών στο πλαίσιο δεδομένων, που είναι αναγκαίοι προκειμένου κάνουμε περισσότερο ακριβείς προβλέψεις. Μια λύση θα ήταν η πρόσθεση περισσότερων ανεξάρτητων μεταβλητών εισόδου από το σετ δεδομένων του Cleveland στο πλαίσιο δεδομένων του Statlog. Για παράδειγμα οι τιμές των SMOKE

(καπνιστής ή όχι), FAMHIS (οικογενειακό ιστορικό), DIURETIC (χρήση διουρητικού αγκύλης κατά τη διάρκεια του τεστ κόπωσης) θα προσέθεταν πληροφορία που θα βελτίωνε την τιμή του συντελεστή R2 του McFadden. Ωστόσο επειδή σε αυτές τις μεταβλητές λείπουν τιμές, θα έπρεπε να υπάρχει διαχείριση όσων αφορά την αντικατάστασή τους. Δηλαδή μια μέθοδος συσταδοποιήσης, όπως αυτή του πλησιέστερου γείτονα ή ένα Μπεϋζιανό δίκτυο θα μπορούσαν να αναδείξουν τιμές για αυτά τα δείγματα χωρίς τιμές σε ορισμένα από τα χαρακτηριστικά τους. Αυτό όμως ίσως οδηγούσε το μοντέλο σε υπερπροσαρμογή ή μεροληψία. Για το λόγο αυτό θα έπρεπε να υπάρξει προσεκτική προετοιμασία προκειμένου να διαμορφωθεί ένα ομοιογενές πλαίσιο δεδομένων (balanced dataset).

Αναφορές - Πηγές

[1] National Center for Chronic Disease Prevention and Health Promotion, Know the Facts About Heart Disease. : National Center for Chronic Disease Prevention and Health Promotion, 2013.

[2] M. Pignone, A. Fowler-Brown, M. Pletcher, J. A. Tice, "Screening for asymptomatic coronary artery disease: A systematic review," Systematic Evidence Review, no. 22, 2003.

[3] David W. Aha & Dennis Kibler. Instance-based prediction of heart-disease presence with the Cleveland database

[4] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310

[5] A.H. Chen and S.Y. Huang and P.S. Hong and C.H. Cheng and E.J. Lin, HDPS: Heart Disease Prediction System, Computing in Cardiology 2011

[6] N. Cheung, Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering B. SC. Thesis University of Queenland, 2001

[7] A.V. Senthil Kumar, Diagnosis of Heart Disease using Fuzzy Resolution Mechanism, Journal of Artificial Intelligence 5 (1): pp. 47-55, 2012.

[8] A. N. G. Bhuvaneswari, Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network, Computing, Communication and Applications (ICCCA), 2012 International Conference on, vol., no., pp. 1-5, 22-24 Feb 2012.

[9] R. Das and I. Trukoglu and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications, Eslevier Lt,2008.

[10] Can Committee Machine Networks to Diagnose Cardivascular Diseases, Southeast Europe Journal of Soft Computing, pp. 76-93, 2013

[11] Kanita Karaduzovic-Hadziab, Raşit Köker b, Diagnosis of heart disease using a committee machine neural network, Proceedings of the 9th International Conference on Applied Informatics Eger, Hungary, January 29– February 1, 2014. Vol. 1. pp. 351–360.

[12] Chaitrali S Dangare and Sulabha S Apte. Article: Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications 47(10):44-48, June 2012.

[13] Mrs. G.Subbalakshmi, Mr.M.Chinna Rao "Decision Support in heart disease prediction system using naïve bays", IJCSE Indian journal of computer sciense and engineering, ISSN : 0976-5166 Vol. 2 No. 2. Apr-May 2011

[14] E.P.Ephzibah, Dr. V. Sundarapandian, "Framing Fuzzy Rules using Support Sets for Effective Heart Disease Diagnosis"; International Journal of Fuzzy Logic Systems (IJFLS) Vol.2, No.1, February 2012.

[15] R. Wu, W. Peters, M. W. Morgan, "The next generation clinical decision support: Linking evidence to best practice," Journal Healthcare Information Management, vol. 16, no. 4, pp. 5055, 2002.

[16] S. Palaniappan, R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in Computer Systems and Applications, Doha, Qatar, 2008.

[17] Dinevski, Dejan & Bele, Uros & Sarenac, Tomislav & Rajkovic, Uros & S?us?ters?ic, Olga. (2011). Clinical Decision Support Systems. 10.5772/25399.

[18] Rajendra Akerkar,Priti Sajja Knowledge-Based Systems 1st Edition 2010, Jones and Bartlett Publishers, ISBN-13: 978-0763776473

[19] Berlin A, Sorani M, Sim I. A taxonomic description of computer-based clinical decision support systems. J Biomed Inform. 2006;39(6):656-667. doi:10.1016/j.jbi.2005.12.003

[20] Dhandhania, Vinay & Choubey, Dilip & Paul, Sanchita. (2017). Rule based diagnosis system for diabetes. Biomedical Research (India). 28. 5196-5209.

[21] A. Tzavaras, P. R. Weller and B. Spyropoulos, "A Neuro-Fuzzy Controller for the estimation of Tidal Volume and Respiration Frequency ventilator settings for COPD patients ventilated in control mode," 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, 2007, pp. 3765-3768, doi: 10.1109/IEMBS.2007.4353151.

[22] Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. The New England journal of medicine. Jun 14 1979;300(24):1350-1358

[23] Pope JH, Aufderheide TP, Ruthazer R, et al. Missed diagnoses of acute cardiac ischemia in the emergency department. The New England journal of medicine. Apr 20 2000;342(16):1163-1170.

[24] Williamernard Kannel, William Peter Castelli, T. R. Gordon, Patricia Mannix Mcnamara, Serum cholesterol, lipoproteins, and the risk of coronary heart disease. The Framingham study, Published 1971 in Annals of internal medicine

[25] Sarwar N, Gao P, Kondapally Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies. The Lancet. 2010 Jan 1;375(9733):2215-2222. Available from, DOI: 10.1016/S0140-6736(10)60484-9

[26] Kim Fox, Jeffrey S. Borer, A. John Camm, Nicolas Danchin et al, Resting Heart Rate in Cardiovascular Disease, Journal of the American College of Cardiology, Volume 50, Issue 9, August 2007, DOI: 10.1016/j.jacc.2007.04.079.

[27] PhD Jacqueline M.Dekkera PhD Evert G.Schoutena MD Peter Klootwijk PhD Jan Pool PhD Daan Kromhout, ST segment and T wave characteristics as indicators of coronary heart disease risk: The Zutphen study, Journal of the American College of Cardiology, Volume 25, Issue 6, May 1995, Pages 1321-1326.

[28] Wagner, GS. Marriott's practical electrocardiography. 9th edn. Baltimore: Williams and Wilkens, 1994 Διαθέσιμο από το Google Scholar

[29] Mitchell, T.M. (1997). Machine Learning. Maidenhead, H.B.:McGraw-Hill International Editions

[30] Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2, 1137-1143. San Francisco, CA: Morgan Kaufmann

[31] Galton F. Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 1886;15: 246–63.

[32] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society Series B, 58: 267-288.

[33] LEDLEY RS, LUSTED LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959 Jul 3;130(3366):9-21. doi: 10.1126/science.130.3366.9. PMID: 13668531.

[34] Horrocks JC, McCann AP, Staniland JR, Leaper DJ, De Dombal FT. Computer-aided diagnosis: description of an adaptable system, and operational experience with 2,034 cases. Br Med J. 1972 Apr 1;2(5804):5-9. doi: 10.1136/bmj.2.5804.5. PMID: 4552593; PMCID: PMC1789007.

[35] Edward H. Shortliffe. 1974. A rule-based computer program for advising physicians regarding antimicrobial therapy selection. In Proceedings of the 1974 annual ACM conference - Volume 2 (ACM '74). Association for Computing Machinery, New York, NY, USA, 739. DOI:https://doi.org/10.1145/1408800.1408906

[36] Bates D, Leape L, Cullen DJ, Laird N, Petersen L, Teich JM, et al. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA 1998;280(15):1311–6.

[37] Baker AM, McCarthy B, Gurley VF, Yood MU. Influenza immunization in a managed care organization. J Gen Intern Med 1998;13(7):469–75

[38] Bogusevicius A, Maleckas A, Pundzius J, Skaudickas D. Prospective randomised trial of computer-aided diagnosis and contrast radiography in acute small bowel obstruction. Eur J Surg 2002;168(2):78–83

[39] Pryor, TA & Gardner, R & Clayton, P & Warner, H. (1983). The HELP system. Journal of medical systems. 7. 87-102. 10.1007/BF00995116.

[40] Manotti C, Moia M, Palareti G, Pengo V, Ria L, Dettori AG. Effect of computer-aided management on the quality of treatment in anticoagulated patients: a prospective, randomized, multicenter trial of APROAT (Automated PRogram for Oral Anticoagulant Treatment). Haematologica 2001;86(10):1060–70

[41] Boukhors Y, Rabasa-Lhoret R, Langelier H, Soultan M, Lacroix A, Chiasson JL. The use of information technology for the management of intensive insulin therapy in type 1 diabetes mellitus. Diabetes Metab 2003;29(6):619–27

[42] Branston LK, Greening S, Newcombe RG, Daoud R, Abraham JM, Wood F, et al. The implementation of guidelines and computerised forms improves the completeness of cancer pathology reporting. The CROPS project: a randomised controlled trial in pathology. Eur J Cancer 2002;38(6):764–72.

[43] Cannon DS, Allen SN. A comparison of the effects of computer and manual reminders on compliance with a mental health clinical practice guideline. J Am Med Inform Assoc 2000;7(2):196–203

[44] Lutz SF, Ammerman AS, Atwood JR, Campbell MK, DeVellis RF, Rosamond WD. Innovative newsletter interventions improve fruit and vegetable consumption in healthy adults. J Am Diet Assoc 1999;99(6):705–9

[45] Goodey RD, Brickley MR, Hill CM, Shepherd JP. A controlled trial of three referral methods for patients with third molars. Br Dent J 2000;189(10):556–60.

[46] Wan Abu Bakar, Wan Aezwani. (2020). HDP: Heart Disease Prediction Tool using Neural Network. International Journal of Emerging Trends in Engineering Research. 8. 1794-1797. 10.30534/ijeter/2020/50852020.

[47] Comparative Testing and Evaluation of Statistical and Logical Learning Algorithms for Large-Scale Applications in Classification, Prediction and Control. Project reference: 5170 Funded under FP2-ESPRIT 2

[48] Charles Elkan, The Foundations of Cost-Sensitive Learning, Department of Computer Science and Engineering 0114, University of California, San Diego, Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence

[49] G. C. Cawley and N. L. C. Talbot, Over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research, 2010. Research, vol. 11, pp. 2079-2107, July 2010.

[50] Tom Oluoch, Xenophon Santas, Daniel Kwaro, Martin Were, Paul Biondich, Christopher Bailey, Ameen Abu-Hanna, Nicolette de Keizer, The effect of electronic medical record-based clinical decision support on HIV care in resource-constrained settings: A systematic review, International Journal of Medical Informatics,Volume 81, Issue 10,2012,Pages e83-e92,ISSN 1386-5056,

[51] Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning, second edition. New York: Springer, 2008.

[52]] Christianini, N., and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press, 2000.

[53] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022. doi:10.1061/JPEODX.0000175.

[54] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2001). The Elements of Statistical Learning : Data Mining, Inference, and Prediction : with 200 full-color illustrations. New York: Springer. ISBN 0-387-95284-5. OCLC 46809224.

[55] M. Kikuchi, M. Yoshida, M. Okabe and K. Umemura, "Confidence interval of probability estimator of Laplace smoothing," *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2015, pp. 1-6, doi: 10.1109/ICAICTA.2015.7335387.

[56] Pochet NL, Suykens JA. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. Ultrasound Obstet Gynecol. 2006 Jun;27(6):607-8. doi: 10.1002/uog.2791. PMID: 16715467.

[57] Demertzis, Konstantinos. (2018). Ανάπτυξη Ευφυών Προτύπων και αντίστοιχων Πληροφοριακών Συστημάτων, εμπνευσμένων από Βιολογικά αντίστοιχα, με στόχο την αξιολόγηση Περιβαλλοντικών Προβλημάτων και Κινδύνων. 10.13140/RG.2.2.35268.27526.

[58] https://repository.kallipos.gr/bitstream/11419/1237/2/Kef._10.pdf

[59] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, second edition. New York: Springer, 2008.

[60] Christianini, N., and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[61] Neural Networks and Learning Machines 3rd Edition, Simon Haykin, Prentice Hall

[62] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408.

[63] McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943).

[64] Werbos, P. J. (1981). Applications of Advances in Nonlinear Sensitivity Analysis. Proceedings of the 10th IFIP Conference, 31.8 - 4.9, NYC (p./pp. 762--770), .

[65] Jierula, A.; Wang, S.; OH, T.-M.; Wang, P. Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. *Appl. Sci.* 2021, *11*, 2314. https://doi.org/10.3390/app11052314

[66] Bucchianico, A.D. (2014). Coefficient of Determination (R^2). In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat03980

[67] Kim S, Alizamir M, Zounemat-Kermani M, Kisi O, Singh VP. Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea. J Environ Manage. 2020 Sep 15;270:110834. doi: 10.1016/j.jenvman.2020.110834. Epub 2020 Jun 5. PMID: 32507742.

[68]] S.B. Kotsiantis “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31 (2007) 249-268, 2007.

[69] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[70] J.R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996

[71] Bertsimas, Dimitris, and Jack Dunn. “Optimal Classification Trees.” Machine Learning 106.7 (2017): 1039–1082.