ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ανάπτυξη μοντέλου εκτίμησης του κινδύνου εμφάνισης καρδιαγγειακής νόσου σε άτομα με Σακχαρώδη Διαβήτη, βασισμένου σε ταξινομητή Τυχαίων Δασών

Διπλωματική Εργασία

του

ΕΥΑΓΓΕΛΟΥ Γ. ΤΣΙΡΚΑ

Επιβλέπουσα:

Κωνσταντίνα Σ. Νικήτα

Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Ανάπτυξη μοντέλου εκτίμησης του κινδύνου εμφάνισης καρδιαγγειακής νόσου σε άτομα με Σακχαρώδη Διαβήτη, βασισμένου σε ταξινομητή Τυχαίων Δασών

Διπλωματική Εργασία

ΕΥΑΓΓΕΛΟΣ Γ. ΤΣΙΡΚΑΣ

Επιβλέπουσα:

Κωνσταντίνα Σ. Νικήτα

Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την…….

…….	…….	…….
Κωνσταντίνα Νικήτα	Διονύσιος-Δημήτριος Κουτσούρης	Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγήτρια Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2015

……

Ευάγγελος, Γ. Τσίρκας
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η παρούσα διπλωματική εργασία αποσκοπεί στη σχεδίαση, ανάπτυξη και αξιολόγηση ενός εξατομικευμένου μοντέλου εκτίμησης του κινδύνου ανάπτυξης καρδιαγγειακής νόσου ως μακροπρόθεσμη επιπλοκή του Σακχαρώδους Διαβήτη Τύπου 2 (ΣΔΤ2). Ο ΣΔΤ2 χαρακτηρίζεται από τη μειωμένη ευαισθησία των κυττάρων του οργανισμού στη δράση της ινσουλίνης, της κύριας ορμόνης που ρυθμίζει τα επίπεδα σακχάρου στο αίμα ή από την τάση του παγκρέατος για περιορισμένη παραγωγή και έκκρισή της. Τα αίτια εμφάνισης της νόσου παραμένουν αδιευκρίνιστα, ενώ οι βραχυπρόθεσμες και οι μακροπρόθεσμες επιπτώσεις της στην υγεία των ατόμων με ΣΔΤ2καθιστούνεπιτακτική την ανάγκη για μελέτη και αντιμετώπισή της.

Το μοντέλο λαμβάνει είσοδο δημογραφικά, ανθρωπομετρικά, κλινικά, και περιβαλλοντικά δεδομένα καθώς και πληροφορίες σχετικά με τις θεραπευτικές αγωγές που ακολουθεί ο ασθενής για την αντιμετώπιση της νόσου και των συνοδών νοσημάτων, ενώ εξάγει την πιθανότητα εμφάνισης καρδιαγγειακού νοσήματος σε βάθος χρόνου ίσο με 5 έτη. Η ανάπτυξη του μοντέλου βασίστηκε στη μεθοδολογία των Τυχαίων Δασών, μιας τεχνικής εξόρυξης γνώσης από δεδομένα με μηχανική μάθηση. Τα Τυχαία Δάση βασίζονται στην συνδυασμένη χρήση πολλαπλών Δέντρων Αποφάσεων, τα οποία μπορούν να επιλύσουν προβλήματα Ταξινόμησης και Παλινδρόμησης. Στο σύστημα που αναπτύξαμε τα Δέντρα Απόφασης λειτουργούν ανεξάρτητα, χειριζόμενα ένα πρόβλημα Παλινδρόμησης, και το καθένα από αυτά εξάγει την πιθανότητα εμφάνισης καρδιαγγειακής νόσου. Τα ανεξάρτητα αποτελέσματα κάθε Δέντρου Απόφασης συλλέγονται για τον μετέπειτα υπολογισμό της τελικής πιθανότητας. Για την ανάπτυξη και την αξιολόγηση του μοντέλου χρησιμοποιήθηκαν δεδομένα 560 ατόμων με ΣΔΤ2, τα οποία παραχωρήθηκαν από το Διαβητολογικό Κέντρο του Ιπποκράτειου Νοσοκομείου Αττικής. Το μοντέλο αξιολογήθηκε ως προς τη διακριτική του ικανότητα καθώς και ως προς την ικανότητά του να παράγει ακριβείς εκτιμήσεις του κινδύνου. Το τελικό σύστημα συγκρίθηκε με την απόκριση ενός έτοιμου μοντέλου (από τις βιβλιοθήκες του λογισμικού πακέτου Matlab), προκειμένου να διαπιστωθεί η ορθότητα της σχεδίασής του.

Λέξεις Κλειδιά:

Σακχαρώδης Διαβήτης Τύπου 2, Τυχαία Δάση, Δέντρα Απόφασης, Ταξινόμηση, Παλινδρόμηση, Μοντέλα Πρόβλεψης Κινδύνου, Συστήματα Κλινικών Αποφάσεων.

Abstract

The present thesis aims at the design, development and evaluation of a model for the assessment of Cardiovascular Disease risk as long term complication of Type 2 Diabetes Mellitus (T2DM). T2DM is characterized by elevated blood glucose levels due to insulin resistance and limited insulin secretion. The causes of the occurrence of the disease remain yet unclear, while its evolution and long-term complications pose the need for further research toward identifying an optimal treatment plan.

The model receives demographic, anthropometric, clinical and environmental data as input, along with information related to the therapeutic scheme of a patient, in order for the disease and its complications to be treated, while it produces the probability of the cardiovascular disease occurrence in a 5 year time span. The development of the proposed model has been based on the Random Forests methodology, which is a Machine Learning technique strongly connected to the concept of Data Mining. Random Forests are constructed with Decision Trees that are capable of solving Classification and Regression problems. The system presented in this dissertation involves the parallel and independent function of Decision Trees that deal with a Regression problem, and the response of every tree is a probability of cardiovascular disease occurrence in T2DM individuals. The independent results of every Decision Tree are then collected for the calculation of the final probability. The information on the aforementioned 560 individuals with T2DM was granted by the Diabetes Center, Hippokration Hospital of Athens, Greece. The evaluation of the model includes the use of efficiency measures for the notions of Discrimination and Calibration. The final system is compared to the response of a given model (from the libraries of Matlab software package), in order for us to validate the correctness of our design.

Keywords: Type 2 Diabetes Mellitus, Random Forests, Decision Trees, Classification, Regression, Risk Prediction Models, Clinical Decision Systems.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα την επιβλέπουσα καθηγήτριά της διπλωματικής μου εργασίας, κυρία Κωνσταντίνα Νικήτα, για την εμπιστοσύνη που μου έδειξε και τη δυνατότητα που μου έδωσε να εκπονήσω ένα θέμα με πρακτικό και λειτουργηματικό χαρακτήρα, καθώς και με μεγάλο ερευνητικό ενδιαφέρον. Επίσης, ευχαριστώ τη διδάκτορα κυρία Κωνσταντία Ζαρκογιάννη για τη συμμετοχή της στην επίβλεψη της εργασίας, καθώς και τους υποψήφιους διδάκτορες Ελένη Λίτσα και Κωνσταντίνο Μήτση για τις εύστοχες παρατηρήσεις τους.

Ακόμη, απευθύνω ιδιαίτερες ευχαριστίες προς το Διαβητολογικό Κέντρο του Ιπποκράτειου Νοσοκομείου Αθηνών, για την παραχώρηση ιατρικών δεδομένων ασθενών με Σακχαρώδη Διαβήτη Τύπου 2.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και την αδελφή μου για την αμέριστη υποστήριξη, την ανιδιοτελή υπομονή και την ηθική ενίσχυση που μου προσέφεραν σε όλη τη διάρκεια των σπουδών μου.

Ευάγγελος Τσίρκας

Περιεχόμενα

Σχήματα

Σχήμα 5.2 AUCs του συστήματός μας για την περίπτωση μη ισορροπημένων δεδομένων

Σχήμα 5.3 AUCs του συστήματός μας για ισορροπημένα δεδομένα

Πίνακες

Εισαγωγή

Ο Σακχαρώδης Διαβήτης είναι μία νόσος που οφείλεται στη διαταραχή του μεταβολισμού της γλυκόζης, η οποία έχει σοβαρές συνέπειες για ορισμένα ζωτικά όργανα εφόσον δεν αντιμετωπιστεί. Η παρούσα διπλωματική εργασία εστιάζει στο Σακχαρώδη Διαβήτη Τύπου 2 (ΣΔΤ2), και πραγματεύεται το ζήτημα της συσχέτισής του με την πιθανότητα εμφάνισης καρδιαγγειακής νόσου, προκειμένου να παραχθεί ένα αξιόπιστο υπολογιστικό εργαλείο υποστήριξης ιατρικών αποφάσεων για την διαχείριση του ΣΔΤ2. Η δομή της παρούσας εργασίας είναι η ακόλουθη:

Κεφάλαιο 1. Συνοπτική περιγραφή του Σακχαρώδους Διαβήτη, με έμφαση στο δεύτερο και πιο συνήθη τύπο του. Παρατίθενται στατιστικά στοιχεία που αφορούν την εξάπλωσή του στο σύγχρονο κόσμο, καθώς και στοιχεία νοσολογίας, φαρμακολογίας, πρόληψης και αντιμετώπισης της ασθένειας.

Κεφάλαιο 2. Στο δεύτερο κεφάλαιο παρουσιάζονται τα μοντέλα που προβλέπουν τον κίνδυνο εμφάνισης καρδιαγγειακής νόσου σε άτομα που πάσχουν από ΣΔΤ2. Συγκεκριμένα αναλύονται τα λειτουργικά χαρακτηριστικά των μοντέλων πρόβλεψης καρδιαγγειακής νόσου, που χρήζουν αποδοχή στην κλινική πρακτική.

Κεφάλαιο 3. Στο τρίτο κεφάλαιο γίνεται αναφορά στις μεθόδους μηχανικής μάθησης, του τρόπου με τον οποίο λειτουργούν, της απόδοσής τους και των εφαρμογών τους σε πρακτικές εφαρμογές. Ιδιαίτερη έμφαση δίνεται στην περίπτωση της μεθόδου συλλογικής μάθησης, στην οποία ανήκει και η μέθοδος των Τυχαίων Δασών που εφαρμόζουμε.

Κεφάλαιο 4. Στο τέταρτο κεφάλαιο περιγράφεται αναλυτικά η μεθοδολογία Τυχαίων Δασών και αναφέρονται τα χαρακτηριστικά απόδοσης της. Αναλυτική επισκόπηση της μεθοδολογίας που χρησιμοποιήθηκε στην εργασία, με παρουσίαση αλγορίθμων και περιγραφή σχετικών εννοιών και δομών.

Κεφάλαιο 5. Παρουσίαση των ποσοτικών αποτελεσμάτων του συστήματος που αναπτύχθηκε και σύγκριση με αποτελέσματα παρόμοιων μεθόδων.

Κεφάλαιο 6. Παράθεση των συμπερασμάτων της διπλωματικής εργασίας και αιτιολόγησή τους. Προτάσεις για μελλοντικές επεκτάσεις.

Κεφάλαιο 1. Διαβήτης

1.1 Ορισμός και Τύποι

Ο Σακχαρώδης Διαβήτης [1] (ΣΔ) είναι ένα σύνολο μεταβολικών διαταραχών πολλαπλής αιτιολογίας που χαρακτηρίζεται από χρόνια υπεργλυκαιμία. Η υπεργλυκαιμία συνοδεύεται από διαταραχές στο μεταβολισμό των υδατανθράκων, λιπών και πρωτεϊνών, που οφείλονται σε προβλήματα στην έκκριση της ή στη δράση της ινσουλίνης ή ακόμη και στη συνδυαστική εμφάνιση των δύο καταστάσεων. Ο ΣΔ συνδέεται με χρόνιες επιπλοκές που μακροπρόθεσμα οδηγούν σε βλάβη, δυσλειτουργία και ανεπάρκεια διάφορων οργάνων, όπως τα μάτια, οι νεφροί, τα νεύρα, η καρδιά και τα αιμοφόρα αγγεία. (Παγκόσμιος Οργανισμός Υγείας 1999).

Ο ΣΔ ταξινομείται σε τρεις κύριες κατηγορίες [3]:

- Ο **Σακχαρώδης Διαβήτης τύπου 1** (γνωστός και ως ΣΔΤ1; μέχρι πρόσφατα επονομαζόμενος και ινσουλινοεξαρτώμενος ή νεανικός διαβήτης) είναι ένα αυτοάνοσο νόσημα που προκαλείται από την καταστροφή των β-κυττάρων του παγκρέατος. Πρόκειται για τα κύτταρα που παράγουν την ινσουλίνη, η οποία ρυθμίζει τα επίπεδα του σακχάρου του αίματος στους υγιείς οργανισμούς. Το αποτέλεσμα είναι αυξημένα επίπεδα σακχάρου στο αίμα και τα ούρα. Ο όρος νεανικός διαβήτης προέρχεται από το γεγονός ότι η ασθένεια εμφανίζεται κυρίως σε παιδιά, προδιαγράφοντας έτσι κάποια κληρονομικά αίτια. Η έρευνα για τη διαχείριση και την αντιμετώπιση του ΣΔΤ1 είναι εκτεταμένη [4] [5] [6].
- Ο **Σακχαρώδης Διαβήτης τύπου 2** (γνωστός και ως ΣΔΤ2; μέχρι πρόσφατα επονομαζόμενος και ινσουλινοανεξάρτητος διαβήτης) προέρχεται από το συνδυασμό της μειωμένης παραγωγής ινσουλίνης και της απώλειας ευαισθησίας των κυττάρων στην ινσουλίνη (ινσουλινοαντίσταση). Είναι ο πιο συνηθισμένος τύπος διαβήτη και αναφέρεται στο 90% των ασθενών με ΣΔΤ2, ενώ τις περισσότερες φορές εμφανίζεται σε ενήλικες, παρόλο που υπάρχουν κάποιες εκδηλώσεις του στις παιδικές ηλικίες.
- Ο **Διαβήτης κύησης**, που εμφανίζεται κατά τη διάρκεια της κύησης(συνηθέστερα κατά τη διάρκεια του τρίτου τριμήνου) και υποχωρεί μετά τον τοκετό. Ο διαβήτης κύησης εμφανίζεται

στο 3-10% των εγκύων και έχει κοινή αιτιολογία και συμπτωματολογία με το ΣΔΤ2. Η παχυσαρκία αυξάνει την πιθανότητα εμφάνισης του διαβήτη κύησης, ενώ το 30-40% των γυναικών που έχουν εμφανίσει την ασθένεια θα εμφανίσουν ΣΔΤ2 αργότερα στη ζωή τους.

1.2 Επίγνωση της ασθένειας και πρόοδος στην αντιμετώπισή της

- **2002-2015**

Η αντιμετώπιση με το αντι-CD3 μονοκλωνικό αντίσωμα, hOKT3gamma1 (Ala-Ala), επιβραδύνει την επιδείνωση στην παραγωγή ινσουλίνης και βελτιώνει το μεταβολικό έλεγχο κατά τη διάρκεια του πρώτου έτους εμφάνισης του ΣΔΤ1 στην πλειονότητα των ασθενών.

Η Αμερικανική Ένωση για το Διαβήτη (AmericanDiabetesAssociation) ορίζει τον προδιαβήτη ως διαταραγμένη γλυκόζη νηστείας (impairedfastingglucose - IFG) και/ή ως μειωμένη ανοχή στη γλυκόζη (impairedglucosetolerance - IGT). Η IFG ορίζεται στα 100-125 mg/dl και η IGT στο εύρος 140 mg/dl – 199 mg/dl, δύο ώρες μετά από την κατανάλωση υγρού υψηλής περιεκτικότητας σε γλυκόζη

Αργότερα, τιμές της A1C από 5.7% ως 6.4% χρησιμοποιήθηκαν για να αναγνωριστούν τα άτομα με προδιαβήτη.

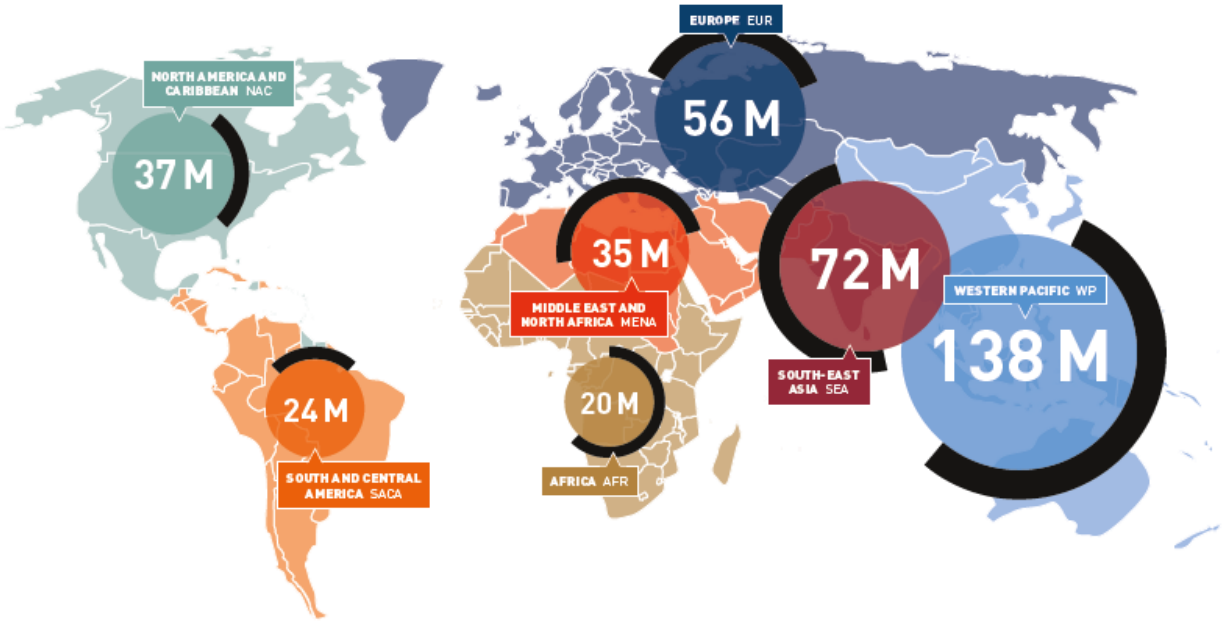
Το FDA εγκρίνει τη χρήση του JANUVIA (φωσφατιδική σιταγλιπτίνη), που είναι το πρώτο από μια νέα κατηγορία φαρμάκων, γνωστή ως αναστολείς του ενζύμου διπεπτυδιλική πεπτιδάση-4 (DDP-4), που ενισχύει την ικανότητα του σώματος να ελαττώνει τα αυξημένα επίπεδα σακχάρου στο αίμα.

Τα αποτελέσματα των μελετών ACCORD, ADVANCE και VADT δημοσιεύτηκαν και παρουσιάστηκαν στα επιστημονικά συνέδρια της Αμερικάνικης Εταιρίας Διαβήτη. Και οι τρεις μελέτες απέτυχαν να αποδείξουν ότι υπάρχει όφελος από την αυστηρό γλυκαιμικό έλεγχο στα καρδιαγγειακά συμβάντα σε ασθενείς με ΣΔΤ2 και υψηλό καρδιαγγειακό κίνδυνο. Από τα αποτελέσματα αυτών των μελετών προκύπτουν κλινικές συστάσεις για μια πιο εξατομικευμένη προσέγγιση των θεραπευτικών στόχων και των κατώτερων τιμών σακχάρου που επιθυμούμε να επιτύχουμε.

Επίσης, το FDA εγκρίνει τη χρήση του INVOKANA (καναγλιφλοζίνη), που είναι το πρώτο φάρμακο μιας νέας κατηγορίας, γνωστής ως αναστολέας του συμμεταφορέα 2 νατρίου και γλυκόζης (SGLT-2), για μείωση του σακχάρου στο αίμα των ασθενών με ΣΔΤ2. Οι αναστολείς SGLT-2 μπλοκάρουν τη δράση

των πρωτεϊνών-μεταφορέων νατρίου και γλυκόζης στο νεφρό, μειώνοντας την επαναπρόσληψη της γλυκόζης και αυξάνοντας την απέκκριση της στα ούρα [2].

Στις μέρες μας, ο ΣΔ είναι η τέταρτη από τις πέντε κύριες αιτίες θανάτου στις οικονομικά ανεπτυγμένες χώρες. Αυτό το δυσοίωνο θέμα που στοχεύει τα πολιτισμένα κράτη του σύγχρονου κόσμου, πλέον φαίνεται να απασχολεί και πολλά αναπτυσσόμενα και πρόσφατα βιομηχανοποιημένα έθνη. Υπάρχουν σοβαρές ενδείξεις ότι ο ΣΔ θα μπορούσε δικαίως να θεωρηθεί επιδημία γι' αυτές τις πιο ευάλωτες χώρες. Οι επιπλοκές του ΣΔ είναι πολυάριθμες και απειλητικές για τη φυσιολογική λειτουργία οργάνων ζωτικής σημασίας: στεφανιαία νόσος, περιφερική αγγειακή νόσος, εγκεφαλικό, διαβητική νευροπάθεια, ακρωτηριασμοί, νεφρική ανεπάρκεια και τύφλωση είναι ασθένειες που βρίσκονται στην κορυφή της λίστας. Οι προαναφερόμενες νόσοι είναι υπεύθυνες για αναπηρία, μειωμένο προσδόκιμο επιβίωσης και τεράστιο υγειονομικό κόστος της εκάστοτε κοινωνίας. Αυτά τα δεδομένα τοποθετούν το διαβήτη μεταξύ των πιο απαιτητικών προβλημάτων που απασχολούν το χώρο της υγείας τον 21^{ου} αιώνα. Υπάρχει ένας αξιοσημείωτος αριθμός μελετών με σκοπό να περιγράψουν την επιδημιολογία του διαβήτη τις τελευταίες δύο δεκαετίες, ωστόσο, την ίδια στιγμή παρατηρείται μια έντονη έλλειψη ευαισθητοποίησης και ενημέρωσης σχετικά με αυτό το σοβαρό θέμα. Η εξάπλωση όμως της νόσου είναι πολύ μεγάλη, όπως φαίνεται ουσιαστικά στην παρακάτω εικόνα:

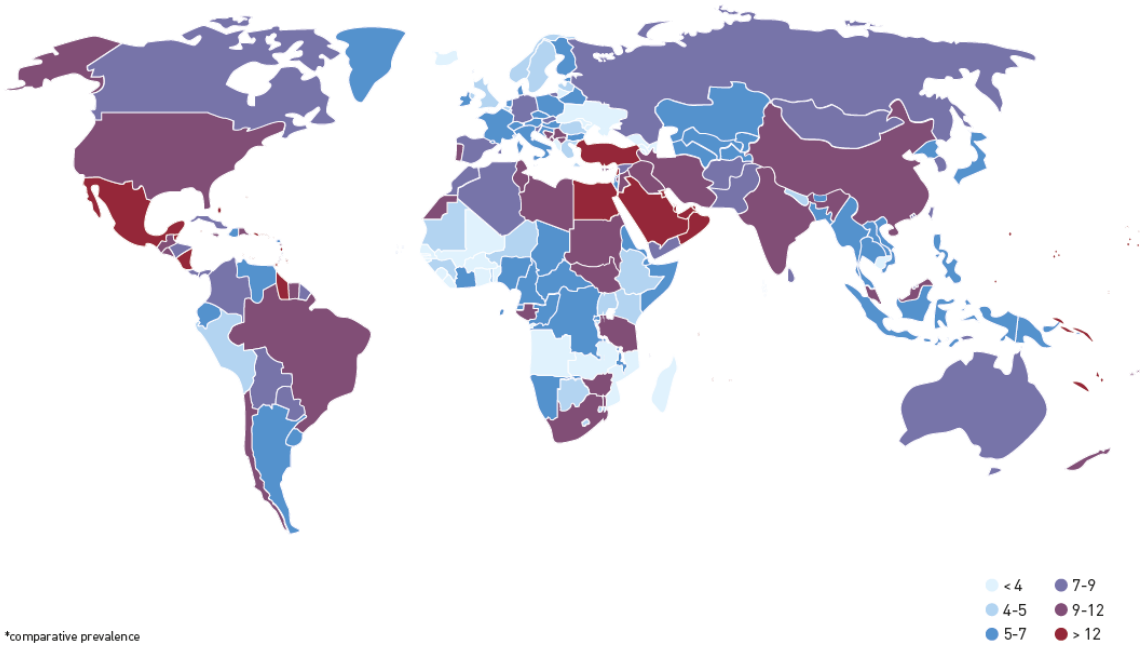


Σχήμα 1.1 Αριθμός ατόμων με διαβήτη ανά περιοχή, 2013

1.3 Στατιστικά στοιχεία

Ο διαβήτης είναι ένα διεθνές θέμα, που εξελίσσεται ραγδαίως, καθώς η πρόληψη και η διαχείρισή του δεν είναι ακόμα επαρκώς αποτελεσματικές. Ενδεικτικά, ο ΣΔΤ2 αντιστοιχεί στο 85%-95% όλων των ατόμων με ΣΔΤ2 στις ανεπτυγμένες χώρες, ενώ το ποσοστό είναι ακόμα υψηλότερο στις υπανάπτυκτες και αναπτυσσόμενες χώρες. Οι κύριοι παράγοντες αύξησης της επίπτωσης του ΣΔΤ2 είναι πολιτισμικοί και κοινωνικοί, όπως η πλειονότητα των πιο ηλικιωμένων ατόμων στις αστικές περιοχές, οι σύγχρονες διαιτητικές συνήθειες και η καθιστική ζωή. Ο διαβήτης τύπου 1 επίσης αυξάνεται στις χώρες τόσο με υψηλό όσο και με χαμηλό εισόδημα. Στις περισσότερες χώρες με υψηλό εισόδημα, ο διαβήτης στα παιδιά και τους εφήβους είναι ο διαβήτης τύπου 1. Επιπρόσθετα, ο διαβήτης της κύησης είναι εξίσου συχνός και εξαπλώνεται όπως η παχυσαρκία και ο ΣΔΤ2. Είναι ευρέως γνωστό από στατιστικές αναλύσεις πως γυναίκες με διαβήτη κύησης είναι πιθανότερο να εμφανίσουν ΣΔΤ2 στη μετέπειτα ζωή τους, με διακυμάνσεις στη συχνότητα εμφάνισης μεταξύ των διάφορων πληθυσμών. Αυτό εξηγείται εν μέρει από τη συσχέτιση των διαφορετικών προσεγγίσεων διάγνωσης με την πολυμορφία των πληθυσμών που μελετώνται.

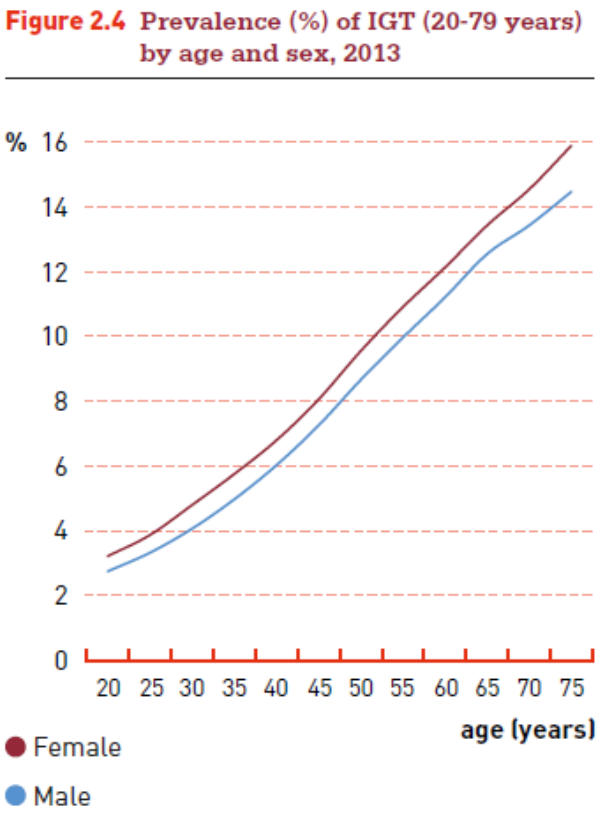
Ο επιπολασμός του διαβήτη μπορεί να γίνει κατανοητός από αριθμητικά δεδομένα. Περίπου 382 εκατομμύρια ανθρώπων παγκοσμίως, ή 8.3% των ενηλίκων, υπολογίζεται να έχουν διαβήτη. Ένα ποσοστό 80% από αυτούς ζουν σε οικονομικά υπανάπτυκτες ή αναπτυσσόμενες χώρες. Αν αυτό το πρότυπο συνεχίσει να υπάρχει, αναμένεται πως ως το 2035, περίπου 592 εκατομμύρια άνθρωποι ή ένας στους δέκα ενήλικες θα έχουν διαβήτη. Αυτό ισοδυναμεί με την ανακάλυψη τριών νέων περιπτώσεων κάθε 10 δευτερόλεπτα ή περίπου δύο εκατομμυρίων το χρόνο. Αυτό το κύμα νέων περιστατικών αναμένεται να συμβεί κυρίως σε περιοχές με αναπτυσσόμενες οικονομίες [7].



Σχήμα 1.2 Εξάπλωση του διαβήτη σε ενήλικους, 2013

Το 50% περίπου των ενηλίκων με διαβήτη έχουν ηλικία από 40 έως 59 χρονών. Σε αυτή την ηλικιακή ομάδα, υπάρχει ένα ποσοστό 80% (184 εκατομμύρια άνθρωποι) με διαβήτη που ζουν σε χώρες χαμηλού ή μεσαίου εισοδήματος. Αυτά τα άτομα φαίνεται να αντιπροσωπεύουν την πλειοψηφία των ατόμων με ΣΔΤ2 στα επόμενα χρόνια. Αν θέλαμε να μιλήσουμε με αριθμούς, αυτό θα σήμαινε ότι ως το 2035 το ποσό των 184 εκατομμυρίων θα αυξανόταν περίπου κατά 50% (264 εκατομμύρια). Ομοίως, περισσότεροι από το 86% θα ζούσαν σε χαμηλού ή μεσαίου εισοδήματος χώρες.

Η εξάπλωση της νόσου δε φαίνεται να δείχνει ιδιαίτερη προτίμηση σε κάποιο από τα δύο φύλα σε παγκόσμιο επίπεδο για τα άτομα με διαβήτη το 2013 ή το 2035. Η αριθμητική διαφορά μεταξύ ανδρών και γυναικών είναι περίπου 14 εκατομμύρια σε χάρη των γυναικών (198 εκατομμύρια άνδρες έναντι 184 εκατομμύριων γυναίκες). Παρ' όλα αυτά, αυτή η διαφορά αναμένεται να αυξηθεί στα 15 εκατομμύρια (303 εκατομμύρια άνδρες έναντι 288 εκατομμυρίων γυναίκες) ως το 2035 [7].



Σχήμα 1.3 Εξάπλωση του διαβήτη ανάλογα με την ηλικία και το φύλο, 2013

Η πλειοψηφία των ανθρώπων με διαβήτη κατοικούν σε αστικές περιοχές (246 εκατομμύρια). Τα άτομα με ΣΔΤ2 σε αγροτικές περιοχές (136 εκατομμύρια) αποτελούν περίπου τους μισούς από την προαναφερθείσα ομάδα με ένα ωστόσο συνεχώς αυξανόμενο ρυθμό. Αν λάβουμε υπ' όψιν μας τη διάκριση των χωρών σε χαμηλού και μεσαίου εισοδήματος, ο αριθμός των ατόμων με διαβήτη στις αστικές περιοχές είναι 181 εκατομμύρια, ενώ στην ύπαιθρο 122. Η διαφορά φαίνεται να διευρύνεται ως το 2035, με 347 εκατομμύρια ανθρώπων να ζουν στις αστικές και 145 εκατομμύρια στις αγροτικές περιοχές.

Τα δεδομένα που προέρχονται από τη στατιστική έρευνα πιθανώς επηρεάζονται από το γεγονός ότι καμιά χώρα δεν έχει διαγνώσει όλα τα άτομα που πάσχουν από διαβήτη. Για παράδειγμα στην Αφρική νότια της Σαχάρας, η διαλογή των ασθενών για διάγνωση διαβήτη δεν είναι πρωταρχικής σημασίας και οι πόροι είναι συχνά περιορισμένοι, οπότε ένα μεγάλο ποσοστό ατόμων με ΣΔΤ2 παραμένουν αδιάγνωστα, φθάνοντας ακόμα και το 90% σε ορισμένες χώρες. Ωστόσο, περίπου το ένα τρίτο των ατόμων με διαβήτη δεν έχουν επίσης διαγνωστεί στις αναπτυγμένες χώρες. Για να κατανοήσουμε τη σημασία του προβλήματος αρκεί να επισημάνουμε ότι η περιοχή της νοτιοανατολικής Ασίας και του Δυτικού

Ειρηνικού, με 35.1 και 74.7 εκατομμύρια μη διαγνωσμένων ατόμων με ΣΔΤ2 αντίστοιχα, υπολογίζεται ότι μαζί αντιστοιχούν σε ένα σύνολο 60% όλων των αδιάγνωστων ατόμων. Παγκοσμίως το 84% όλων των ανθρώπων που δεν έχουν διαγνωστεί ζουν σε χώρες χαμηλού και μεσαίου εισοδήματος [7].

Πίνακας 1.1 10 πρώτες χώρες εξάπλωσης του διαβήτη, 2013 και 2035

Table 2.4 Top 10 countries/territories for prevalence* (%) of IGT (20-79 years), 2013 and 2035

COUNTRY/ TERRITORY	2013 (%)	COUNTRY/ TERRITORY	2035 (%)
Kuwait	17.9	Poland	19.3
Qatar	17.1	Kuwait	18.1
United Arab Emirates	16.6	Qatar	17.4
Poland	16.5	United Arab Emirates	17.0
Bahrain	16.3	Bahrain	16.7
Malaysia	15.2	Malaysia	15.3
Hong Kong SAR	13.3	Hong Kong SAR	13.2
Nicaragua	12.9	Anguilla	13.0
Japan	12.6	Guadeloupe	13.0
Singapore	12.4	Macau SAR	12.9

*comparative prevalence

Ο διαβήτης είναι μια νόσος με τεράστιες δαπάνες για τις κυβερνήσεις και τα άτομα. Η ανάγκη για εξειδικευμένες υπηρεσίες υγείας, η έλλειψη και η ανικανότητα παραγωγικότητας μπορούν να θεωρηθούν σημαντική επιβάρυνση για το ίδιο το άτομο, την οικογένειά του και την κοινωνία. Τα πιθανά οφέλη της πρόληψης, πρώιμης διάγνωσης και θεραπείας δεν έχουν σημασία όταν οι ασθενείς με διαβήτη παραμένουν αδιάγνωστοι για μεγάλο χρονικό διάστημα. Αυτό αποτελεί ένα επιπλέον κόστος, αφού υπάρχουν μελέτες από τις ΗΠΑ που δείχνουν πως ο μη διαγνωσμένος διαβήτης είναι υπεύθυνος 18 εκατομμύρια δολάρια το χρόνο στο κόστος για την υγεία, που θα μπορούσε να αποφευχθεί αν η διάγνωση είχε γίνει εγκαίρως. Υπολογίζεται ότι 175 εκατομμύρια άνθρωποι παγκοσμίως (οι μισοί περίπου από αυτούς με ΣΔΤ2) δεν έχουν διαγνωσθεί. Όταν οριστικοποιηθεί η διάγνωση και αρχίσει η θεραπεία είναι συχνά εφικτό να προληφθούν οι επιπλοκές που εκτός από δαπανηρές είναι και επιβλαβείς και απειλητικές για τη ζωή.

1.4 Σακχαρώδης Διαβήτης Τύπου 2

Συνοψίζοντας τα χαρακτηριστικά του ΣΔΤ2 που αναφέρθηκαν προηγουμένως, η αιτία του είναι η αντίσταση του οργανισμού στη δράση της ινσουλίνης ή η μειωμένη έκκριση ινσουλίνης από το πάγκρεας, που οδηγούν σε παθολογικά επίπεδα γλυκόζης. Η ακριβής αιτιολογία εμφάνισης της νόσου είναι ακόμη άγνωστη, αν και τόσο γενετικοί όσο και περιβαλλοντικοί παράγοντες φαίνεται να συνεισφέρουν. Εκτός από τον όρο ΣΔΤ2, η ασθένεια είναι επίσης γνωστή ως διαβήτης των ενηλίκων αφού είναι συχνότερος σε πιο μεγάλης ηλικίας άτομα [8].

Αν γινόταν προσπάθεια σύγκρισης των δύο τύπων διαβήτη, το γενετικό τους προφίλ θα μπορούσε να περιγραφεί συνοπτικά από μελέτες σε δίδυμα: τα μονοζυγωτικά δίδυμα έχουν υψηλότερο ποσοστό εμφάνισης ΣΔΤ2 (60%-100% έναντι 36% για τον τύπο 1) ενώ το ίδιο ποσοστό για τα διζυγωτικά είναι πιο χαμηλό, αποδεικνύοντας ισχυρότερη γενετική συσχέτιση από τον τύπο 1. Ένας από τους αναγνωρισμένους γενετικούς υπότυπους του ΣΔΤ2 είναι ο λεγόμενος μιτοχονδριακός διαβήτης, που αποτελεί το 1-3% των ασθενών με ΣΔΤ2 και κληρονομείται από τη μητέρα [9].

Τα συμπτώματα του ΣΔΤ2 εξελίσσονται αργά και μπορεί να περάσουν απαρατήρητα από τον ασθενή για μεγάλο χρονικό διάστημα:

- **Πολυδιψία και πολυουρία:** αυξημένη ποσότητα υγρών μετακινείται από τους ιστούς στην κυκλοφορία του αίματος εξαιτίας της περίσσειας γλυκόζης, προκαλώντας δίψα, αυξημένη πρόσληψη ύδατος και συχνότητα ούρησης.
- **Συνεχής και έντονη πείνα:** η μεταφορά γλυκόζης στα κύτταρα αποτελεί πρόβλημα, αφού η ινσουλίνη που είναι υπεύθυνη γι’ αυτό είναι ανεπαρκής. Το αποτέλεσμα είναι η έλλειψη ενέργειας και η πολυφαγία.
- **Απώλεια βάρους:** ακόμα και με το προηγούμενο σύμπτωμα της αυξημένης όρεξης, η απώλεια βάρους μπορεί να παρατηρηθεί ως συνέπεια της κακής λειτουργίας του μεταβολισμού της γλυκόζης. Ο οργανισμός ψάχνει για άλλες πηγές ενέργειας όπως αυτές που είναι αποθηκευμένες στους μύες και το λιπώδη ιστό. Η θερμιδική ισορροπία διαταράσσεται αφού η γλυκόζη απελευθερώνεται στα ούρα.
- **Εξάντληση και κόπωση:** λόγω στέρησης σακχάρου από τα κύτταρα.

- **Θόλωση όρασης:** όταν συμβαίνει η προαναφερθείσα μεταφορά υγρών από τους ιστούς στο αίμα, οι φακοί των οφθαλμών χάνουν ένα μέρος των υγρών τους και παρατηρείται θόλωση της όρασης που επηρεάζει την ικανότητα συγκέντρωσης.
- **Καθυστερημένη επούλωση τραυμάτων και συχνές λοιμώξεις**
- **Περιοχές μελάγχρωσης δέρματος:** σκούρα σημάδια στις πτυχές του δέρματος παρατηρούνται σε ορισμένους ανθρώπους με ΣΔΤ2 (acanthosisnigricans)

Υπάρχουν πολλοί παράγοντες που επηρεάζουν τα επίπεδα του σακχάρου στο αίμα και μερικές φορές τα προβλήματα που προκύπτουν απαιτούν άμεση θεραπεία όπως:

- **Υπεργλυκαιμία (υψηλό σάκχαρο στο αίμα):** η συνήθεια της πολυφαγίας και της μειωμένης πρόσληψης γλυκόζης σε συνδυασμό με την ελάττωση της αγωγής μπορεί να αυξήσει σημαντικά τα επίπεδα σακχάρου στο αίμα. Αυτό μπορεί επίσης να πυροδοτηθεί από μια ασθένεια. Η περιοδική μέτρηση του σακχάρου και η έγκαιρη συνειδητοποίηση των σημείων και συμπτωμάτων της υπεργλυκαιμίας- πολυουρία, πολυδιψία, ξηροστομία, θολή όραση, κόπωση και ναυτία- μπορούν να αποτρέψουν τα υψηλά επίπεδα γλυκόζης στο αίμα. Αν παρατηρηθούν υψηλές τιμές ή εμφανιστούν κάποια από αυτά τα συμπτώματα, το θεραπευτικό πλάνο και η αγωγή του ασθενή πρέπει να τροποποιηθούν.

- **Υπεργλυκαιμικό υπερωσμωτικό μη κετοτικό σύνδρομο (HHNS):** πρόκειται για μια απειλητική κατάσταση για τη ζωή που αναγνωρίζεται από πολύ υψηλές μετρήσεις σακχάρου (υψηλότερες από 600mg/dl (33.3mmol/l) , ξηροστομία, εξαιρετική δίψα, υψηλό πυρετό (πάνω από 38^{o}C), υπνηλία, σύγχυση, απώλεια όρασης, παραισθήσεις και σκούρα ούρα. Οι ατομικές συσκευές μέτρησης σακχάρου ίσως να μην καταφέρουν να υπολογίσουν την τιμή σε αυτές τις περιπτώσεις εφ' όσον υπολογίζεται πως είναι εκτός του εύρους τους. Η αιτία αυτής της επισφαλούς για τη ζωή επιπλοκής είναι η αυξημένη τιμή σακχάρου που αυξάνει τη γλοιότητα του αίματος και το κάνει να μοιάζει με σιρόπι. Παρατηρείται συχνότερα σε ηλικιωμένους ασθενείς με ΣΔΤ2 και συχνά είναι επακόλουθο λοίμωξης ή νόσου.

- **Αυξημένες κετόνες στα ούρα (διαβητική κετοξέωση):** αυτή η κατάσταση συμβαίνει όταν τα κύτταρα στερούνται ενέργειας και το σώμα αρχίζει να διασπά το λιπώδη ιστό με σκοπό να

καλύψει τις ανάγκες του. Αυτό όμως έχει σαν αποτέλεσμα την παραγωγή τοξικών οξέων γνωστών ως κετόνες. Οι ασθενείς θα πρέπει να είναι ενήμεροι ώστε να αναγνωρίσουν την έλλειψη όρεξης, την αδυναμία, τον εμετό, τον πόνο στο στομάχι και τον πυρετό. Υπάρχουν εξετάσεις που πραγματοποιούνται εύκολα και γρήγορα με σκοπό να ελέγξουν τα ούρα για την περίσσεια κετονών. Αν τα αποτελέσματα δείξουν παθολογικά υψηλές τιμές απαιτείται επείγουσα κι εντατική φροντίδα. Η κετοξέωση συναντάται συνηθέστερα σε ασθενείς με ινσουλινοεξαρτώμενο διαβήτη χωρίς να αποκλείεται η εμφάνισή της και σε ασθενείς με ΣΔΤ2.

- **Χαμηλάεπίπεδα γλυκόζης (υπογλυκαιμία):** είναι μια κατάσταση που χαρακτηρίζεται από μειωμένα επίπεδα σακχάρου στο αίμα. Η παράλειψη ενός γεύματος ή η έντονη σωματική άσκηση μπορούν να ρίξουν απότομα το σάκχαρο σε τιμές κάτω του φυσιολογικού. Μια άλλη αιτία υπογλυκαιμίας είναι η χρήση αντιδιαβητικών δισκίων (με μηχανισμό δράσης που διεγείρει την έκκριση ινσουλίνης) ή η ίδια η ινσουλίνη που μπορεί να χρησιμοποιεί ο ασθενής.

Τα άτομα με ΣΔΤ2 πρέπει να βρίσκονται πάντα σε επαγρύπνηση όσον αφορά τα σημεία και συμπτώματα της υπογλυκαιμίας (ιδρώτας, τρόμος, αδυναμία, πείνα, ζάλη, πονοκέφαλος, θολή όραση, καρδιακές αρρυθμίες, κολλώδης ομιλία, υπνηλία, σύγχυση και σπασμοί) όπως επίσης και να ελέγχουν τακτικά τα επίπεδα σακχάρου στο αίμα τους.

Η πρώτη αντίδραση σε ένα υπογλυκαιμικό επεισόδιο είναι η κατανάλωση κάποιας ουσίας που θα αυξήσει το σάκχαρο των ασθενών- χυμός, ταμπλέτες σακχάρου, σόδα ή κάποια άλλη πηγή ζάχαρης. Ένας επανέλεγχος απαιτείται μετά από 15 λεπτά με σκοπό να επιβεβαιώσει την επιστροφή των επιπέδων σακχάρου σε φυσιολογικές τιμές. Αν η υπογλυκαιμία δεν υποχωρεί, η διαδικασία πρέπει να επαναληφθεί. Αν ο ασθενής χάσει τις αισθήσεις του οι συνοδοί του ίσως χρειαστεί να παρέμβουν κάνοντας μια ένεση γλυκαγόνης (ορμόνη που διεγείρει την απελευθέρωση σακχάρου στην κυκλοφορία του αίματος).

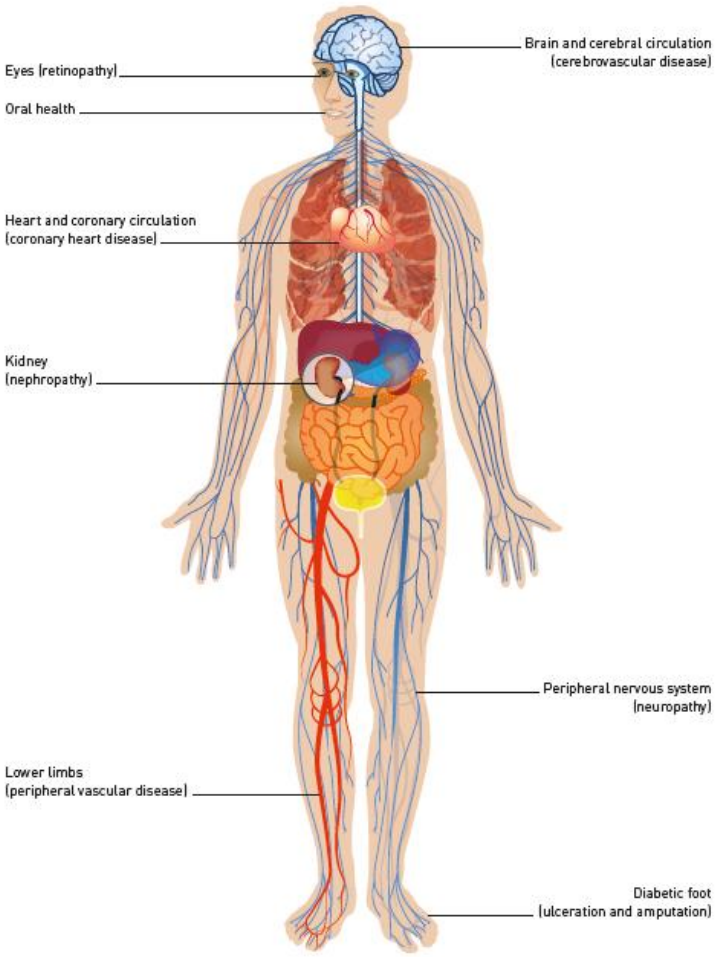
Είναι ακόμα ασαφές στους ερευνητές γιατί ορισμένοι άνθρωποι εμφανίζουν ΣΔΤ2 και άλλοι όχι. Ο μόνος ικανοποιητικός λόγος ωστόσο μπορεί να είναι το γεγονός ότι συγκεκριμένοι παράγοντες αυξάνουν τον κίνδυνο. Τέτοιοι παράγοντες είναι:

- Το βάρος: οι υπέρβαροι άνθρωποι είναι πιο επιρρεπείς στον ΣΔΤ2. Η αντίσταση των κυττάρων στην ινσουλίνη αυξάνεται όταν υπάρχει περίσσεια λιπώδους ιστού. Όμως, αν και είναι επαρκής δεν είναι αναγκαίος παράγοντας για την εμφάνιση του διαβήτη.

- Η κατανομή του λίπους: όχι μόνο η ποσότητα του σωματικού λίπους αλλά και η κατανομή του είναι παράγοντας κινδύνου για την εμφάνιση της νόσου. Για παράδειγμα το κοιλιακό λίπος αυξάνει τις πιθανότητες σε αντίθεση με τη συγκέντρωσή του σε άλλα μέρη του σώματος.
- Η έλλειψη άσκησης: η καθιστική ζωή είναι μία από τις αιτίες του διαβήτη τύπου 2 κυρίως επειδή ακόμα και η ήπια άσκηση επιτρέπει τον έλεγχο του βάρους ενώ παράλληλα ο οργανισμός χρησιμοποιεί τη γλυκόζη σα μορφή ενέργειας κάνοντας τα κύτταρα πιο ευαίσθητα στην ινσουλίνη.
- Το οικογενειακό ιστορικό: η κληρονομικότητα συνδέεται με τον κίνδυνο εμφάνισης διαβήτη ιδίως αν ο ένας γονέας ή αδερφός έχει ΣΔΤ2.
- Η φυλή: χωρίς ακόμα να είναι γνωστή η αιτία, άτομα από συγκεκριμένες φυλές όπως οι μαύροι, οι Ισπανοί, οι Ινδιάνοι της Αμερικής και οι ασιάτες Αμερικάνοι, είναι πιο επιρρεπείς στη νόσο από τους λευκούς.
- Η ηλικία: όσο πιο ηλικιωμένος είναι κάποιος τόσο μεγαλύτερη πιθανότητα έχει να εμφανίσει διαβήτη. Ειδικά μετά την ηλικία των 45, που ο άνθρωπος έχει μειωμένη δραστηριότητα, απώλεια μυϊκής μάζας και αύξηση του βάρους όχι απαραίτητα λόγω διατροφικών συνηθειών αλλά λόγω γήρατος. Παρ' όλα αυτά παρατηρείται δυστυχώς μια δραματική αύξηση μεταξύ των παιδιών και γενικότερα των νέων ανθρώπων.
- Ο προδιαβήτης: είναι μια κατάσταση υψηλών επιπέδων σακχάρου στο αίμα, όχι όμως αρκετά για να χαρακτηριστεί διαβήτης. Αν τον αγνοήσουμε και μείνει αθεράπευτος, μπορεί να προκαλέσει ΣΔΤ2.
- Ο διαβήτης της κύησης: αποτελεί σημαντική αιτία ΣΔΤ2. Επιπλέον, αν το βάρος του μωρού υπερβαίνει τα 4 kg, ο κίνδυνος εμφάνισης διαβήτη είναι μεγάλος.
- Σύνδρομο πολυκυστικών ωοθηκών: μια κατάσταση που προκαλεί ακανόνιστους κύκλους έμμηνης ρύσης, αυξημένη τριχοφυΐα και παχυσαρκία.

Ο ΣΔΤ2 μπορεί εύκολα να ξεφύγει της προσοχής, κυρίως στα πρώιμα στάδια που ο ασθενής αισθάνεται καλά. Ο διαβήτης όμως επηρεάζει τη λειτουργία πολλών βασικών οργάνων, όπως της καρδιάς, των αγγείων, των νεύρων, των ματιών και των νεφρών. Ο έλεγχος συνεπώς των επιπέδων σακχάρου, μπορεί να προλάβει την εμφάνιση αυτών των επιπλοκών. Αν και οι μακροπρόθεσμες επιπλοκές του διαβήτη αναπτύσσονται σταδιακά, μπορεί τελικά να οδηγήσουν σε σοβαρού βαθμού αναπηρία ακόμα και να γίνουν απειλητικές για τη ζωή. Μερικές από τις επιπλοκές του διαβήτη περιλαμβάνουν:

- Καρδιαγγειακή νόσος: ο διαβήτης θεωρείται σημαντικός παράγοντας κινδύνου για τις νόσους της καρδιάς και των αγγείων, όπως η στεφανιαία νόσος με προκάρδιο άλγος (στηθάγχη), το έμφραγμα του μυοκαρδίου, το εγκεφαλικό, η αρτηριοσκλήρυνση και η αρτηριακή υπέρταση. Η ανακάλυψη αυτής της σύνδεσης του ΣΔΤ2 και της καρδιοπάθειας είναι ο στόχος αυτής της εργασίας.
- Νευροπάθεια: μια από τις ανεπιθύμητες επιδράσεις του υψηλού σακχάρου είναι ο τραυματισμός των πολύ μικρών αγγείων (τριχοειδή) και κυρίως η καταστροφή των τοιχωμάτων τους. Τα τριχοειδή είναι τα αγγεία που τροφοδοτούν τα νεύρα με ουσίες απαραίτητες για τη φυσιολογική λειτουργία τους. Η καταστροφή τους οδηγεί σε αιμωδίες των άκρων, καύσο και πόνο που ξεκινούν από τις άκρες των δακτύλων των ποδιών και των χεριών και σταδιακά επεκτείνονται στα εγγύς τμήματα των άκρων. Αν ο έλεγχος της νόσου είναι ανύπαρκτος ή ελλιπής, μπορεί τελικά να οδηγήσει σε πλήρη απώλεια της αισθητικότητας του προσβεβλημένου άκρου. Αυτή η κατάσταση μπορεί να επεκταθεί και σε άλλες περιοχές, όπως τα νεύρα της κοιλιακής χώρας που ελέγχουν την πέψη προκαλώντας ναυτία, εμετό, διάρροια ή δυσκοιλιότητα. Για τους άνδρες ένα εξίσου σημαντικό θέμα είναι η στυτική δυσλειτουργία.
- Νεφροπάθεια: οι νεφροί είναι υπεύθυνοι για το φιλτράρισμα των άχρηστων προϊόντων του αίματος. Αποτελούνται από εκατομμύρια συμπλέγματα μικροσκοπικών αγγείων, που καταστρέφονται από το διαβήτη οδηγώντας μερικές φορές σε νεφρική ανεπάρκεια ή σε μη αναστρέψιμη τελικού σταδίου χρόνια νεφρική νόσο, που απαιτεί αιμοκάθαρση ή μεταμόσχευση.



Σχήμα 1.4 Επιπλοκές του διαβήτη

- Βλάβη των οφθαλμών: η διαβητική αμφιβληστροειδοπάθεια είναι μια από τις κοινές επιπλοκές του διαβήτη. Καταστρέφονται τα αγγεία του αμφιβληστροειδούς με αποτέλεσμα να επέλθει η τύφλωση. Άλλες πιθανές επιπλοκές είναι ο καταρράκτης και το γλαύκωμα.
- Βλάβη των ποδιών: η μειωμένη ροή αίματος στα πόδια ή η τοπική βλάβη των νεύρων δημιουργούν πολλά προβλήματα όπως καθυστέρηση της επούλωσης τραυμάτων και φλυκταινών, προκαλώντας σοβαρές λοιμώξεις. Αν δεν αντιμετωπιστούν, εξελίσσονται σε σοβαρές βλάβες που μπορεί να οδηγήσουν μέχρι και στον ακρωτηριασμό του δακτύλου ή ακόμα και του ποδιού.
- Ακουστική βλάβη: αυξημένες είναι οι πιθανότητες βλάβης της ακοής σε άτομα με ΣΔΤ2.
- Δερματοπάθεια: ο διαβήτης κάνει το δέρμα πιο ευάλωτο σε βακτηριακές και μυκητιασικές λοιμώξεις.

- Νόσος Alzheimer: φαίνεται να συνδέεται με τον κακό γλυκαιμικό έλεγχο και κατ' επέκταση με το διαβήτη αν και παραμένει ακόμα ασαφές.

Η διάγνωση του διαβήτη περιλαμβάνει έναν αριθμό εξετάσεων που μπορούν να πραγματοποιηθούν:

- Γλυκοζυλιωμένη αιμοσφαιρίνη (A1C): το ποσοστό των επιπέδων του σακχάρου στο αίμα για τους τελευταίους δύο ή τρεις μήνες υπολογίζεται μετρώντας το ποσοστό του σακχάρου που βρίσκεται συνδεδεμένο με την αιμοσφαιρίνη, την πρωτεΐνη μεταφοράς οξυγόνου στα ερυθρά αιμοσφαίρια. Όσο αυξημένα είναι τα επίπεδα της γλυκόζης του αίματος τόσο περισσότερη αιμοσφαιρίνη θα κυκλοφορεί με συνδεδεμένο σάκχαρο. Οι ενδεικτικές τιμές για το διαβήτη είναι 6.5% η παραπάνω σε δύο ξεχωριστές μετρήσεις. Η δεύτερη ζώνη χαρακτηρίζεται προδιαβήτης και περιλαμβάνει τιμές μεταξύ 5.7 και 6.4%. Οι φυσιολογικές τιμές είναι κάτω από 5.7%. Αν η εξέταση για την A1C δεν είναι διαθέσιμη ή επικρατούν συνθήκες όπως η εγκυμοσύνη ή υπάρχουν άλλες ασυνήθιστες μορφές αιμοσφαιρίνης (γνωστές ως παραλλαγές), η εξέταση δεν θα είναι ακριβής και τότε θα πρέπει να διενεργηθεί μια σειρά άλλων εξετάσεων προκειμένου να τεθεί η διάγνωση.
- Τυχαία μέτρηση γλυκόζης στο αίμα: ένα δείγμα αίματος λαμβάνεται τυχαία και υπολογίζεται το επίπεδο του σακχάρου. Οι μονάδες που χρησιμοποιούνται για να εκφράσουν την τιμή του σακχάρου είναι τα mg ανά dl ή τα mmol ανά lt. Μια τυχαία τιμή πάνω από 200mg/dl (11.1 mmol/lt) ή περισσότερο είναι ενδεικτική διαβήτη και η διάγνωση επιβεβαιώνεται περισσότερο αν συνυπάρχουν σημεία και συμπτώματα της νόσου, όπως πολυδιψία και πολυουρία.
- Γλυκόζη νηστείας: δείγμα αίματος λαμβάνεται μετά το πέρας μιας νύχτας που ο ασθενής είναι νηστικός. Αν το αποτέλεσμα της εξέτασης είναι μια τιμή κάτω από 100mg/dl (5.6 mmol/l) το άτομο δε διαγιγνώσκεται με υη νόσο. Ωστόσο αν η τιμή της γλυκόζης νηστείας είναι μεταξύ 100 και 125 mg/dl (5.6 με 6.9 mmol/l) ο ασθενής θεωρείται πως πάσχει από προδιαβήτη. Υψηλότερες τιμές σε δύο διαφορετικά δείγματα υποδηλώνουν διαβήτη.
- Τεστ ανοχής γλυκόζης: κι αυτή η εξέταση περιλαμβάνει επίσης ολονύκτια νηστεία. Μετά μετράται το σάκχαρο νηστείας στο αίμα. Έπειτα το άτομο καταναλώνει ένα υγρό με ζάχαρη και μετρώνται περιοδικά για τις επόμενες δύο ώρες τα επίπεδα του σακχάρου στο αίμα. Ομοίως, υπάρχουν ενδεικτικές τιμές, που προσδιορίζουν το φυσιολογικό μεταβολισμό του σακχάρου, τον προδιαβήτη και τον διαβήτη. Αν τα επίπεδα του σακχάρου στο αίμα είναι κάτω από 140mg/dl (7.8mmol/l) δεν τίθεται η διάγνωση του ΣΔΤ2. Μία μέτρηση πάνω από

200mg/dl (11.1mmol/l) δύο ώρες μετά αποδεικνύει διαβήτη, ενώ μία μέτρηση μεταξύ 140 και 199mg/dl (7.8 και 11mmol/l) χαρακτηρίζεται προδιαβήτης.

Μετά τη διάγνωση, μια ομάδα από γιατρούς και ειδικούς στο διαβήτη που προτείνονται από το θεράποντα ιατρό, έρχονται σε επαφή με τον ασθενή προκειμένου να δημιουργήσουν την ομάδα παρακολούθησής του. Αυτή περιλαμβάνει:

- Πιστοποιημένο εκπαιδευτή για διαβήτη
- Ποδίατρο
- Ενδοκρινολόγο
- Οφθαλμίατρο
- Διαιτολόγο

Η σωστή προσέγγιση για την αντιμετώπιση του διαβήτη ίσως προλάβει μερικές από τις ανεπιθύμητες συνέπειές του και περιλαμβάνει τα εξής:

- **Υγιεινή διατροφή**: λανθασμένα επικρατεί η άποψη από την πλειοψηφία των ανθρώπων ότι η δίαιτα για τον ΣΔΤ2 είναι ιδιαίτερα αυστηρή. Είναι όμως καίριας σημασίας η δίαιτα να είναι πλούσια σε τροφές με ίνες και μειωμένη ποσότητα λίπους όπως φρούτα, λαχανικά, δημητριακά ολικής άλεσης ενώ θα πρέπει παράλληλα να συμπληρώνεται τηρώντας το μέτρο από ζωικά προϊόντα, επεξεργασμένους υδατάνθρακες και γλυκά. Τροφές χαμηλού γλυκαιμικού δείκτη μπορούν επίσης να βοηθήσουν τα άτομα που πάσχουν από διαβήτη τύπου 2. Ο γλυκαιμικός δείκτης είναι ένας αριθμός που κατηγοριοποιεί τις τροφές ανάλογα με το πόσο γρήγορα αυξάνουν τα επίπεδα της γλυκόζης στο αίμα. Τροφές με υψηλό γλυκαιμικό δείκτη αυξάνουν το σάκχαρο πολύ γρήγορα και θα πρέπει να καταναλώνονται με προσοχή. Οι χαμηλού γλυκαιμικού δείκτη τροφές επιτρέπουν πιο σταθερά επίπεδα σακχάρου. Τέτοιες είναι οι τροφές πλούσιες σε φυτικές ίνες. Έτσι, ένα διαιτητικό πλάνο μπορεί να δημιουργηθεί από ένα διαιτολόγο κρατώντας την ισορροπία ανάμεσα στο τι πρέπει και τι θέλει να φάει ο κάθε ασθενής. Αυτή η διαδικασία θα κάνει πιο εύκολο τον υπολογισμό και την παρακολούθηση της πρόσληψης υδατανθράκων, αφού ο ασθενής θα ενημερώνεται για την ποσότητα των υδατανθράκων που χρειάζεται να καταναλώσει προκειμένου να διατηρεί σταθερά επίπεδα σακχάρου στο αίμα.

- **Ήπια άσκηση**: οι άνθρωποι με διαβήτη συμβουλεύονται να ασκούνται και να επικεντρώνονται κυρίως στην αερόβια άσκηση. Με την έγκριση του γιατρού τους καλό θα ήταν να καθιερώσουν ένα πρόγραμμα με ασκήσεις ρουτίνας όπως περπάτημα, κολύμβηση, ποδηλασία και παρόμοιες δραστηριότητες. Οι εκτάσεις και οι ασκήσεις ενδυνάμωσης μπορούν επίσης να αποδειχθούν χρήσιμες αν αρχίζουν ήπια και εντείνονται σταδιακά. Ωστόσο, είναι ζωτικής σημασίας για τα άτομα αυτά να έχουν κατά νου πως με τη σωματική άσκηση επέρχεται κατανάλωση θερμίδων άρα και πτώση του σακχάρου. Επομένως τα επίπεδα γλυκόζης θα πρέπει να ελέγχονται πριν από κάθε άσκηση και μια πιθανή πτώση αυτών μπορεί να αποφευχθεί με την κατανάλωση ενός μικρού γεύματος, ιδίως αν ο ασθενής λαμβάνει αγωγή που μειώνει τα επίπεδα του σακχάρου στο αίμα.
- **Παρακολούθηση του σακχάρου**: οι ασθενείς θα πρέπει να ελέγχουν συχνά τα επίπεδα και αν είναι σε αγωγή με ινσουλίνη, πολλές φορές τη μέρα σύμφωνα με τις οδηγίες του θεράποντα ιατρού. Επίσης είναι σημαντική η καταγραφή των μετρήσεων γιατί είναι ο μόνος τρόπος να διαπιστώσουν αν οι τιμές κυμαίνονται στο επιθυμητό εύρος. Επιπλέον η αντίδραση του ανθρώπινου σώματος είναι απρόβλεπτη που σημαίνει πως ο ασθενής θα πρέπει να μάθει πώς τα επίπεδα γλυκόζης επηρεάζονται από την διατροφή, την άσκηση, το αλκοόλ και τη φαρμακευτική αγωγή.
- **Φάρμακα για το διαβήτη και θεραπεία με ινσουλίνη**: από όλους τους ασθενείς με διαβήτη τύπου 2, υπάρχουν μερικοί που είναι ικανοί να τηρήσουν τα όρια των επιπέδων γλυκόζης μόνο με διατροφή και άσκηση. Ωστόσο ο μεγαλύτερος αριθμός ασθενών χρειάζεται αντιδιαβητικά φάρμακα ή ινσουλίνη. Η επιλογή εξαρτάται τόσο από την ανταπόκριση του οργανισμού στον καθημερινό τρόπο ζωής, όσο και στο ιατρικό προφίλ του κάθε ασθενή. Το θεραπευτικό σχήμα ίσως είναι πιο πολύπλοκο συνδυάζοντας φάρμακα διαφορετικών κατηγοριών με σκοπό την πιο αποτελεσματική θεραπεία των συμπτωμάτων. Πιθανά σχήματα αγωγής για τον ΣΔΤ2είναι:
- **Μετφορμίνη** (Glucophage, Glumetza και άλλα). Η μετφορμίνη είναι η πρώτη θεραπεία που περιγράφτηκε για τον ΣΔΤ2. Αλλάζει την ευαισθησία των σωματικών κυττάρων στην ινσουλίνη για να χρησιμοποιείται πιο αποτελεσματικά και συγχρόνως να εμποδίζει την παραγωγή γλυκόζης από το ήπαρ. Είναι μια ουσία που δε θα μείωνε επαρκώς τα επίπεδα σακχάρου από μόνη της. Η αλλαγή του τρόπου ζωής είναι επίσης απαραίτητη, όπως η απώλεια βάρους και η άσκηση. Οι πιο συχνές παρενέργειές της είναι η ναυτία και η διάρροια. Ωστόσο, εξαφανίζονται στις περισσότερες περιπτώσεις όσο η ουσία εισέρχεται στον οργανισμό τακτικά. Αν η συνδυασμένη δράση της μετφορμίνης με τον υγιεινό τρόπο ζωής δεν επαρκούν για τον έλεγχο των τιμών σακχάρου, ο γιατρός μπορεί να προτείνει άλλη θεραπεία από του στόματος ή ενέσιμη αγωγή.

- **Σουλφονυλουρίες**: αυτά τα φάρμακα διεγείρουν την έκκριση ινσουλίνης. Οι ουσίες που χρησιμοποιούνται πιο συχνά είναι η γλιβενκλαμίδη (DiaBeta, Glynase), γλιπιζίδη (Glucotrol) και γλιμεπιρίδη (Amaryl). Οι παρενέργειές τους είναι η υπογλυκαιμία και η αύξηση του σωματικού βάρους.
- **Μεγλιτινίδες**: δρουν σαν τις σουλφονυλουρίες, διεγείροντας το μηχανισμό έκκρισης ινσουλίνης, αλλά με γρηγορότερη δράση και μικρότερη διάρκεια. Ο κίνδυνος με αυτές είναι η μεγάλη μείωση του σακχάρου, αλλά όχι τόσο δραστικά όσο οι σουλφονυλουρίες. Ένα άλλο κοινό χαρακτηριστικό με τις σουλφονυλουρίες είναι η πιθανή αύξηση του σωματικού βάρους. Τέτοιες είναι η ρεπαγλινίδη (Pradin) και η νατεγλινίδη (Starlix).
- **Θειαζολιδινεδιόνες**: αυτές έχουν παρόμοιο μηχανισμό δράσης με τη μετφορμίνη αυξάνοντας την ευαισθησία των ιστών στην ινσουλίνη. Ένας αριθμός παρενεργειών, όπως η αύξηση του βάρους και ο κίνδυνος καρδιακής ανεπάρκειας και καταγμάτων έχουν αναφερθεί με αποτέλεσμα οι γιατροί να αποφεύγουν τη χορήγηση αυτών των φαρμάκων για τη θεραπεία του ΣΔΤ2. Η ροσιγλιταζόνη (Avandia) και η πιογλιταζόνη (Actos) ανήκουν σε αυτή την κατηγορία.
- **Αναστολείς τουDDP-4**: ελαττώνουν τα επίπεδα σακχάρου στο αίμα αλλά δεν έχουν πολύ ισχυρή δράση. Η σιταγλιπτίνη (Januvia), η σαξαγλιπτίνη (Onglyza) και η λιναγλιπτίνη (Tradjenta) είναι ορισμένα παραδείγματα αυτής της κατηγορίας.
- **Αγωνιστές του υποδοχέα GLP-1**: κάνουν πιο αργή την πέψη και μειώνουν τα επίπεδα σακχάρου, όχι όμως με τον τρόπο που δρουν οι σουλφονυλουρίες κι επομένως είναι πιο ήπια η δράση τους. Έτσι δε συστήνονται ως αποκλειστική θεραπεία. Αγωνιστές του GLP-1 είναι η εξανατίδη (Byetta) και η λιραγλουτίδη (Victoza). Η ναυτία και η πιθανή εμφάνιση παγκρεατίτιδας είναι οι πιο συχνές παρενέργειες αυτής της κατηγορίας φαρμάκων.
- **SGLT-2 αναστολείς**: πρόκειται για σχετικά καινούρια φάρμακα στη θεραπεία του διαβήτη κι επηρεάζουν τον τρόπο που λειτουργούν οι νεφροί, εμποδίζοντας την επαναρρόφηση της γλυκόζης στο αίμα και ενισχύοντας την απέκκρισή της στα ούρα. Η καναγλιφλοζίνη (Invokana) και η δαπαγλιφλοζίνη (Farxiga) είναι παραδείγματα αυτής της κατηγορίας. Αναφερόμενες παρενέργειες είναι οι μυκητιασικές λοιμώξεις και οι λοιμώξεις του ουροποιητικού συστήματος.
- **Ινσουλίνη**: όταν η φαρμακευτική αγωγή δεν είναι επαρκής για τον έλεγχο των επιπέδων σακχάρου συστήνεται η θεραπεία με ινσουλίνη. Ο τρόπος με τον οποίο οι επαγγελματίες υγείας αντιμετώπιζαν τους διαβητικούς ασθενείς όσον αφορά την ινσουλίνη έχει αλλάξει δραματικά, αφού πλέον δε θεωρείται η έσχατη λύση. Η πέψη εμποδίζει τη δράση της ινσουλίνης οδηγώντας στην ανάγκη της ένεσής της. Συνταγογραφείται ένα μείγμα ινσουλίνης, ανάλογα με τις ανάγκες του κάθε ασθενή και ρυθμίζονται οι δόσεις τόσο για τη μέρα όσο και για τη νύχτα. Η έναρξη της χρήσης της ινσουλίνης σε ΣΔΤ2 συνήθως γίνεται με ένεση ινσουλίνης μακράς δράσης τη νύχτα

και περιλαμβάνει τη χρήση μιας λεπτής βελόνας ή ένα στυλό ινσουλίνης (παρόμοιο με το κανονικό στυλό, αλλά με φυσίγγιο που αντί για μελάνι έχει ινσουλίνη).Υπάρχει ποικιλία στις μορφές ινσουλίνης που χρησιμοποιούνται με καθεμιά από αυτές να έχει διαφορετική δράση. Τέτοιες είναι η ινσουλίνη glulisine (Apidra), η lispro (Humalog), η aspart (Novolog), η glargine(Lantus), η detemir (Levemir), και η isophane (HumulinN, NovolinN).

Βαριατρική Χειρουργική: η παχυσαρκία κάνει τη θεραπεία της νόσου πολύ πιο δύσκολη, με αποτέλεσμα να αναγκάζονται οι ασθενείς με ΣΔΤ2 και δείκτη μάζας σώματος (BMI) μεγαλύτερο από 35 να προβούν σε χειρουργεία απώλειας βάρους (βαριατρική χειρουργική). Η επέμβαση επηρεάζει σημαντικά τα επίπεδα σακχάρου στο αίμα, ώστε να επιστρέφουν στις φυσιολογικές τιμές σε ποσοστό 55 με 95% των ατόμων με ΣΔΤ2. Ο τύπος της επέμβασης καθορίζει επίσης την επίδραση στα επίπεδα σακχάρου αφού τα χειρουργεία με παράκαμψη τμήματος του λεπτού εντέρου έχουν αποδειχθεί πιο επαρκή σχετικά με την επίδραση στο μεταβολισμό της γλυκόζης, σε σύγκριση με άλλες επεμβάσεις απώλειας βάρους. Παρ’ όλα αυτά τα χειρουργεία έχουν αρκετό κόστος και κινδύνους, ένας από τους οποίους είναι και ο θάνατος. Επιπλέον απαιτούν δραστικές αλλαγές του τρόπου ζωής ενώ μπορεί να προκαλέσουν μακροπρόθεσμα ελλείψεις θρεπτικών συστατικών και οστεοπόρωση.

Κεφάλαιο 2. Μοντέλα Εκτίμησης Κινδύνου Εμφάνισης Καρδιαγγειακής νόσου σε άτομα με Σακχαρώδη Διαβήτη Τύπου 2

Σκοπός αυτού του κεφαλαίου είναι η παρουσίαση των μοντέλων (state of the art) εκτίμησης του κινδύνου εμφάνισης καρδιαγγειακής νόσου ως επιπλοκής του ΣΔΤ2, καθώς επίσης και να διερευνηθεί εάν υπάρχει ανάγκη εφαρμογής προηγμένων μεθόδων για την ανάπτυξη πρωτότυπων μοντέλων με μεγαλύτερη απόδοση και ακρίβεια. Στις επόμενες παραγράφους παρουσιάζονται τα πιο κλινικώς αποδεκτά μοντέλα, καθώς επίσης και ορισμένα διεθνή guidelines που τα προτείνουν.

Τα στατιστικά στοιχεία υποδεικνύουν ότι περίπου 360 εκατομμύρια άτομα υποφέρουν από ΣΔ, αριθμός που αναμένεται να αυξηθεί απότομα, ενώ οι δυσμενείς αυτές συνθήκες συνυπάρχουν με την απουσία διάγνωσης για μεγάλο αριθμό ατόμων που ήδη πάσχουν. Ο όρος μεταβολικό σύνδρομο χρησιμοποιείται για να περιγράψει τη σύνδεση μεταξύ αγγειακού κινδύνου και ινσουλινοαντίστασης, ενώ το ενδιαφέρον γύρω από τα αίτιά του προέρχεται από το γεγονός ότι ο προαναφερθείς κίνδυνος προηγείται της ανάπτυξης ΣΔΤ2. Ωστόσο, η σύνδεση αυτή δε φαίνεται να προηγείται της εμφάνισης υπεργλυκαιμίας. Το κλειδί για την αποτελεσματική πρόληψη είναι η αντιμετώπιση του κινδύνου εμφάνισης καρδιαγγειακής νόσου με τρόπο που λαμβάνει υπόψη τις ιδιαιτερότητες κάθε ατόμου και τη δική του στάση απέναντι στη νόσο. Για το λόγο αυτό, διαφορές που εξαρτώνται από παράγοντες όπως το φύλο, η φυλή και ο τρόπος ζωής είναι καίριας σημασίας, όμως, η πιο σημαντική σχέση είναι αυτή μεταξύ της θνητότητας σε ασθενείς με ΣΔΤ2 και της ανάπτυξης καρδιαγγειακής νόσου . Αυτό οδήγησε την παγκόσμια κοινότητα στη δημιουργία κλινικών οδηγιών για τη διαχείριση του ΣΔΤ2, οι οποίες συνηγορούν στη χρήση υπολογιστικών μοντέλων εκτίμησης του κινδύνου εμφάνισης καρδιαγγειακής νόσου για τον έγκαιρο καθορισμό κατάλληλου θεραπευτικού σχήματος [10].

Ο Ευρωπαϊκός Οργανισμός έρευνας του Διαβήτη προτείνει τη χρήση των μοντέλων Framingham και DECODE ως των πιο κατάλληλων για πρόβλεψη της καρδιαγγειακής νόσου . Το πρώτο ορίστηκε το 1967 και περιελάμβανε σημαντικούς παράγοντες κινδύνου, όπως φύλο, ηλικία, συστολική πίεση, συνολική χοληστερόλη, κάπνισμα και διαβήτη. Η εφαρμογή του σε πολλούς πληθυσμούς και η σύγκριση με άλλα μοντέλα οδήγησε στο συμπέρασμα ότι παρατηρείται μια συνέπεια (consistency) στα αποτελέσματα μεταξύ πληθυσμών όταν χρησιμοποιείται το εν λόγω μοντέλο. Στο πλαίσιο μιας ακόμη πολύ σημαντικής μελέτης γνωστής ως DECODE, αποδείχθηκε ότι η γλυκόζη νηστείας αλλά και η μεταγευματική σε διάστημα 2 ωρών αποτελούν παράγοντα κινδύνου. Επομένως, το μοντέλο DECODE αποτέλεσε καινοτομία στον υπολογισμό του κινδύνου εμφάνισης καρδιαγγειακής νόσου [11]. Η εκτίμηση

του κινδύνου εμφάνισης καρδιαγγειακής νόσου θα πρέπει να γίνεται ετησίως, λαμβάνοντας υπόψη κάποια προηγούμενη καρδιαγγειακή νόσο, ηλικία, Δείκτη Μάζας Σώματος, καθώς και καθιερωμένους παράγοντες κινδύνου ένας το κάπνισμα, τα τριγλυκερίδια, το οικογενειακό ιστορικό, η χαμηλή HDL χοληστερόλη και η παρελθούσα κολπική μαρμαρυγή. Η χρήση εξισώσεων πρόβλεψης κινδύνου δεν προτείνεται για μη διαβητικούς πληθυσμούς, ενώ η εκτίμηση και εξαγωγή του θα μπορούσε να γίνει με το μοντέλο UKPDS [12]

Υπάρχουν πολλά μοντέλα πρόβλεψης κινδύνου της καρδιαγγειακής νόσου , ωστόσο θα παρουσιάσουμε μόνο τα πιο γνωστά και τα οποία είναι ταυτόχρονα συμβατά με τα διεθνή guidelines:

- Το Framingham Score [13] από τη μελέτη Framingham Heart Study. Παρά την διαθεσιμότητα ορισμένων δοκιμασμένων αλγορίθμων πρόβλεψης κινδύνου, η υιοθέτησή τους στην πρωτοβάθμια φροντίδα είναι περιορισμένη. Μια πιθανή αιτία πίσω από την έλλειψη δράσης από το χώρο ένας υγείας στη χρήση αυτού του είδους μοντέλων ίσως είναι η τάση των τελευταίων να εξειδικεύονται στην πρόβλεψη συγκεκριμένων χαρακτηριστικών καρδιαγγειακής νόσου . Πολύ συχνά, οι γιατροί χρειάζεται να καθορίσουν ένα σχήμα στρατηγικής που να αναφέρεται σε μια συγκεκριμένη ασθένεια, αλλά συνήθως οι γιατροί πρωτοβάθμιας φροντίδας στοχεύουν στην πρόληψη κάθε καρδιαγγειακής επιπλοκής. Το τεστ Framingham ακολουθεί αυτή την κατεύθυνση και υλοποιεί ένα μοντέλο πρόβλεψης πολυπαραγοντικού κινδύνου που επιτρέπει στους γιατρούς να αποφασίζουν ποια άτομα χαρακτηρίζονται από υψηλό κίνδυνο εμφάνισης όλων των καρδιαγγειακών περιστατικών. Προκειμένου να θεωρηθεί το Framingham έγκυρο για εκτίμηση, συμμετείχαν άτομα χωρίς σημαντικά περιστατικά καρδιαγγειακής νόσου στο ιατρικό ένας ιστορικό, με ηλικίες από 30 έως 74 χρόνων. Η μελέτη δηλώνει ότι η καρδιαγγειακή νόσος αναφέρεται σε ένα αριθμό προβλημάτων υγείας, ένας CHD (στεφανιαίος θάνατος, έμφραγμα του μυοκαρδίου, στεφανιαία ανεπάρκεια και στηθάγχη), αγγειοεγκεφαλικά περιστατικά (που περιλαμβάνουν το ισχαιμικό, το αιμορραγικό και το παροδικό ισχαιμικό εγκεφαλικό επεισόδιο), περιφερική αρτηριοπάθεια (διαλείπουσα χωλότητα), και καρδιακή ανεπάρκεια. Ακόμη, διεξήχθησαν έρευνα ιατρικού ιστορικού, σωματική εξέταση και επαφή με προσωπικούς γιατρούς, ώστε να συγκεντρωθούν πληροφορίες για καρδιαγγειακά περιστατικά κατά την παρακολούθηση, ενώ ένα πάνελ έμπειρων επιστημόνων αξιολόγησε κάθε ιατρικό φάκελο. Επιπρόσθετα, μια ανεξάρτητη επιτροπή που περιελάμβανε και ένα νευρολόγο μελέτησε τα αγγειοεγκεφαλικά περιστατικά, ενώ εξετάστηκαν και οι περισσότεροι συμμετέχοντες με πιθανό εγκεφαλικό.

Παλινδρομήσεις τύπου αναλογικών κινδύνων Cox [14] που λαμβάνουν υπόψη το φύλο χρησιμοποιήθηκαν για να συνδέσουν παράγοντες κινδύνου με τον κίνδυνο εμφάνισης πρώτου συμβάντος καρδιαγγειακής νόσου για μια περίοδο παρακολούθησης 12 ετών, αφού επιβεβαιώθηκε η υπόθεση αναλογικού κινδύνου που απαιτείται από την προαναφερόμενη μέθοδο. Το επόμενο βήμα ήταν η κατασκευή μαθηματικών συναρτήσεων καρδιαγγειακής νόσου ρίσκου από αυτά τα μοντέλα. Οι συναρτήσεις υπολόγισαν τον εκτιμώμενο κίνδυνο διάρκειας 10 ετών. Οι μεταβλητές που περιλαμβάνονται στα μοντέλα Cox ήταν η ηλικία, η συνολική χοληστερόλη, η HDL χοληστερόλη, η συστολική πίεση του αίματος, η χρήση φαρμακευτικής αγωγής κατά ένας υπέρτασης, το κάπνισμα την τρέχουσα περίοδο και η σοβαρότητα του διαβήτη. Παράγοντες όπως η διαστολική πίεση του αίματος, ο Δείκτης Μάζας Σώματος (ΔΜΣ) και τα τριγλυκερίδια λήφθηκαν υπόψη, αν και η στατιστική σημαντικότητα ήταν πολύ μικρή. Ένας λογαριθμικός μετασχηματισμός επιλέχθηκε για να αυξηθεί η απόδοση και η ακρίβεια του μοντέλου. Ακόμη, αξιολογήθηκε η ικανότητα του μοντέλου να διακρίνει άτομα που έχουν περάσει ένα καρδιαγγειακό περιστατικό από αυτούς που δεν πέρασαν κάτι αντίστοιχο, χρησιμοποιώντας μια γενική c στατιστική. Η στατιστική αυτή αντιστοιχεί στην καμπύλη ROC. Η λειτουργία της συνοψίζεται στην εξής πρόταση: δύο άτομα θεωρούνται συγκρίσιμα αν κάποιος μπόρεσε να προβλέψει ποιο από τα δύο επιβίωσε περισσότερο και αρμονικά αν έχουν παρόμοιες πιθανότητες επιβίωσης. Συνεπώς, η γενική c στατιστική είναι η πιθανότητα της αρμονίας δεδομένης της συγκρισιμότητας.

Το αρχικό σύνολο υπέστη δειγματοληψία με τη χρήση της bootstrap λογικής: τη βαθμονόμηση του μοντέλου πρόβλεψης κινδύνου αξιολογήθηκε σε ένα διάστημα 10 ετών, χρησιμοποιώντας το Hosmer-Lemeshow τεστ. Το μέτρο Kaplan-Meier χρησιμοποιήθηκε για να υπολογιστεί η συχνότητα εμφάνισης των καρδιαγγειακών περιστατικών, και στη συνέχεια συγκρίθηκε με το καρδιαγγειακό κίνδυνο.

Αφού δημιουργήθηκαν οι γενικές συναρτήσεις καρδιαγγειακού κινδύνου, διεξήχθη η πρόβλεψη του κινδύνου ατομικών χαρακτηριστικών καρδιαγγειακής νόσου. Αυτές οι συναρτήσεις μετασχηματίστηκαν σε καταγραφές των τιμών κινδύνου, χρησιμοποιώντας τη μέθοδο που περιγράψαμε. Στη συνέχεια δημιουργήθηκαν καταγραφές «καρδιακής ηλικίας» με σκοπό να κατανοηθεί η έννοια του κινδύνου. Η καρδιακή ηλικία ενός ατόμου θεωρείται η ηλικία του με τον ίδιο υπολογισμένο κίνδυνο αλλά με όλους τους άλλους παράγοντες κινδύνου να παίρνουν φυσιολογικές τιμές. Η καρδιακή ηλικία μπορεί να θεωρηθεί αντίστοιχη της αγγειακής ηλικίας, και αντιμετωπίζονται ως η ίδια έννοια σε όλη τη μελέτη.

Εκτός από τα γενικά μοντέλα πρόβλεψης κινδύνου καρδιαγγειακής νόσου που αναφέραμε ως τώρα, αναπτύχθηκε και ένας αριθμός απλοποιημένων μοντέλων. Αυτά χρησιμοποίησαν

χαρακτηριστικά που λαμβάνονται στην πρωτοβάθμια περίθαλψη σε καθημερινή βάση και δεν εμπλέκονται σε εργαστηριακές διαδικασίες. Το καινοτόμο αποτέλεσμα της μελέτης Framingham ήταν ότι πέτυχε την εκτίμηση τόσο του γενικού καρδιαγγειακού ρίσκου όσο και του κινδύνου ατομικών καρδιαγγειακών περιστατικών [13].

- Η μελέτη DECODE [15]. Έχουν διεξαχθεί μελέτες της θνητότητας που περιελάμβαναν συγκεντρώσεις γλυκόζης σε νηστεία και 2 ώρες μετά από ένα Τεστ Ανοχής στη Γλυκόζη που προσλαμβάνεται από το στόμα (75-g Oral Glucose Tolerance Test - OGTT). Οι μελέτες αυτές πραγματοποιήθηκαν στη Μονάδα Διαβήτη και Γενετικής Επιδημιολογίας του Δημόσιου Εθνικού Ινστιτούτου Υγείας στη Φινλανδία. Ένα σύνολο 14 ομάδων παρείχε δεδομένα καρδιαγγειακής θνητότητας με τους εξής παράγοντες καρδιαγγειακού κινδύνου: ηλικία, κάπνισμα, αρτηριακή πίεση, συγκέντρωση συνολικής χοληστερόλης και ΔΜΣ όπως επίσης και συγκεντρώσεις γλυκόζης νηστείας και μεταγευματική δύο ωρών. Άλλοι παράγοντες κινδύνου, όπως η HDL χοληστερόλη, τα τριγλυκερίδια, και η περίμετρος του κορμού δεν ήταν διαθέσιμοι σε όλες τις ομάδες. Το follow-up των υποψηφίων διήρκεσε το ελάχιστο 4.8 έτη για θνητότητα, ενώ σε κάποιες μελέτες κράτησε πάνω από 10 έτη. Τα άτομα ταξινομήθηκαν ως προς την ηλικία, τη γλυκόζη νηστείας και τη μεταγευματική γλυκόζη 2 ωρών (FPG και 2hPG), τη γλυκόζη νηστείας στο πλάσμα (FPG), το κάπνισμα, τις συστολικές πιέσεις, τη συνολική χοληστερόλη και το δείκτη μάζας σώματος. Οι υπολογισμοί των αναλογιών κινδύνου για καρδιαγγειακή θνητότητα πραγματοποιήθηκαν ξεχωριστά για τα δύο φύλα και προέκυψαν από Cox μοντέλα αναλογικού κινδύνου. Όλα τα μοντέλα προσαρμόστηκαν για την ηλικία (σε ηλικιακές ομάδες). Πιθανές μη γραμμικές επιδράσεις των παραγόντων κινδύνου λήφθηκαν υπόψη, καθώς κανένας από αυτούς δεν αγνοήθηκε, μια επιλογή που επέτρεψε στην επιστημονική ομάδα να αποτιμήσει το ρίσκο για τα άτομα συσσωρεύοντας απλά τα σκορ όλων των κατηγοριών ανά παράγοντα. Υπολογίστηκαν οι αναλογίες κινδύνου που συνδέονται με κάθε παράγοντα και μετά ορίστηκαν δύο πολυμεταβλητά μοντέλα που περιελάμβαναν όλους τους παράγοντες, το πρώτο με κατηγορίες για FPG και 2hPG και το δεύτερο μόνο με FPG κατηγορίες, για να αντιστοιχεί στις δύο περιπτώσεις όπου το OGTT είναι ή δεν είναι διαθέσιμο. Το τεστ αναλογίας λογαριθμικής πιθανοφάνειας (log-likelihoodratio) χρησιμοποιήθηκε για να ελέγξει αν οι παράγοντες κινδύνου είχαν διαφορετική επίδραση σε άντρες και γυναίκες σε πολυμεταβλητά μοντέλα, με όλους τους παράγοντες να συμπεριλαμβάνονται εκτός από το ΔΜΣ, που δε θεωρήθηκε σημαντικό χαρακτηριστικό. Ένας δείκτης κινδύνου αποδόθηκε σε κάθε κατηγορία κάθε παράγοντα κινδύνου. Αυτοί οι δείκτες είναι οι συντελεστές βήτα, όπως προκύπτουν από το μοντέλο Cox αναλογικών κινδύνων, πολλαπλασιασμένοι με το 10 και στρογγυλοποιημένοι στον κοντινότερο ακέραιο.

Το μέτρο κινδύνου για ένα άτομο μπορεί να αποκτηθεί αν αθροιστούν οι κίνδυνοι για το κατάλληλο επίπεδο καθενός από τους παράγοντες κινδύνου. Αυτό το απλό και εύκολα υπολογιζόμενο μέτρο είναι ένας σχετικός δείκτης κινδύνου για δεδομένο πληθυσμό. Για να υπολογίσουμε έναν απόλυτο κίνδυνο καρδιαγγειακής θνητότητας, βασισμένο σε στατιστικές θνητότητας του 1995, χρησιμοποιήθηκαν ως δεδομένα μία ομάδα ατόμων από το Μιλάνο. Η συγκεκριμένη ομάδα επιλέχθηκε επειδή περιελάμβανε άνδρες και γυναίκες και όλων των ηλικιακών ομάδων. Σε αυτή την ομάδα η πιθανότητα καρδιαγγειακού θανάτου δίνεται από το μοντέλο Cox από τη σχέση

$$1 - Se^{\frac{R}{10}}$$

Όπου S είναι η πιθανότητα επιβίωσης, υπολογισμένη για παρακολούθηση 5 ή 10 ετών, και της οποίας η μέση τιμή υπολογίζεται για τα δύο μοντέλα, το πρώτο με FPG και 2hPG και το δεύτερο μόνο με FPG.

Καθώς η θνητότητα λόγω καρδιαγγειακής νόσου διαφέρει από χώρα σε χώρα, χρησιμοποιούνται οι στατιστικές αιτίων θανάτου από το WHO για να βαθμονομηθεί ο απόλυτος κίνδυνο για άλλες χώρες, με την υπόθεση ότι η καρδιαγγειακή θνητότητα στην ιταλική ομάδα της μελέτης DECODE είναι αντιπροσωπευτική ολόκληρης της χώρας. Οι στατιστικές θνητότητας του 1995 που είναι διαθέσιμα από τον ιστόχωρο του WHO [1] μας επέτρεψε να αποφασίσουμε έναν πολλαπλασιαστικό παράγοντα για τη θνητότητα κάθε χώρας σε σχέση με την καρδιαγγειακή θνητότητα της Ιταλίας. Έτσι δημιουργήθηκαν πολλαπλασιαστικοί παράγοντες (M) για τα δύο φύλα χρησιμοποιώντας της μέση τιμή των λόγων της καρδιαγγειακής θνητότητας του 1995, σε ηλικιακές κατηγορίες βήματος 5 ετών, σε συγκεκριμένες χώρες έχοντας ως βάση την Ιταλία, στο ηλικιακό εύρος 35 έως 79 ετών. Η πλήρης εξίσωση για την πρόβλεψη του θανάτου είναι τότε ίδια με αυτή που ορίσαμε προηγουμένως.

Η μελέτη που έγινε αφορούσε 16.506 άνδρες και 8907 γυναίκες ηλικίας από 30 έως 74 ετών. Οι αριθμοί των ατόμων μέσα στις ομάδες ποικίλουν, και το πλήρες ηλικιακό εύρος καλύφθηκε στις 7 από τις 14 μελέτες. Για 5 έτη, ο κίνδυνος καρδιαγγειακής θνητότητας διέφερε μεταξύ ερευνητικών κέντρων για τους άνδρες, ανάμεσα στους οποίους οι συγκεντρώσεις γλυκόζης, το κάπνισμα, η συστολική πίεση και η συγκέντρωση χοληστερόλης υπήρξαν σημαντικά χαρακτηριστικά καρδιαγγειακής θνητότητας 5 ετών, με το ΔΜΣ να μην είναι χαρακτηριστικό. Όλοι αυτοί οι παράγοντες διατήρησαν την ικανότητα πρόβλεψής τους όταν μπήκαν σε ένα πολυμεταβλητό μοντέλο. Στις γυναίκες, μόνο οι συγκεντρώσεις γλυκόζης, η συστολική πίεση και το ΔΜΣ ήταν σημαντικά χαρακτηριστικά καρδιαγγειακής θνητότητας. Ωστόσο, όταν όλοι οι

παράγοντες εισήχθησαν στο μοντέλο, μόνο οι συγκεντρώσεις γλυκόζης και η ηλικία διατήρησαν την προβλεπτική τους ικανότητα. Σε ένα διάστημα 10 ετών, το ΔΜΣ δεν ήταν παράγοντας πρόβλεψης της καρδιαγγειακής θνητότητας σε άντρες και γυναίκες. Και στα δύο φύλα, οι συγκεντρώσεις γλυκόζης, το κάπνισμα και η συστολική πίεση ήταν προβλεπτικοί παράγοντες, ενώ η συγκέντρωση χοληστερόλης συνεισφέρει επίσης στην πρόβλεψης θανατηφόρου καρδιαγγειακού κινδύνου στους άνδρες αλλά όχι στις γυναίκες. Αυτά τα αποτελέσματα δεν άλλαξαν όταν όλοι οι παράγοντες εισήχθησαν σε ένα πολυμεταβλητό μοντέλο. Το ΔΜΣ δεν συμπεριλαμβάνεται στον τελικό υπολογισμό καθώς δεν ήταν σημαντικός παράγοντας καρδιαγγειακής θνητότητας σε πολυμεταβλητά μοντέλα. Από τεστ λόγων λογαριθμικής πιθανοφάνειας, οι παράγοντες γλυκόζης βελτίωσαν την πρόβλεψη για την καρδιαγγειακή νόσο είτε η γλυκόζη ήταν FPG είτε ήταν 2hPG. Οι πολλαπλασιαστικοί παράγοντας (M) για τα δύο φύλα και για κάθε χώρα και οι συσσωρευμένες πιθανότητες επιβίωσης (S) επιτρέπουν τον υπολογισμό του απόλυτου κινδύνου καρδιαγγειακού θανάτου, χρησιμοποιώντας το δείκτη κινδύνου R. Ως παράδειγμα χρήσης αυτού του δείκτη κινδύνου, υπολογίστηκε το πενταετές και δεκαετές ρίσκο καρδιαγγειακής θνητότητας για έναν άνδρα και μια γυναίκα 55 ετών από την Ιταλία, οι οποίοι κάπνιζαν, είχαν συστολική πίεση 145 mmHg και μια συγκέντρωση χοληστερόλης 6.5 mmol/l. Το ρίσκο αυτό ήταν μικρότερο από 1%, πλην της περίπτωσης του διαβητικού άνδρα. Σε follow-up 10 ετών υπήρξε μια βαθμιαία αύξηση καθώς χειροτέρευε αυξανόταν και η ινσουλινοαντίσταση (ειδικά στις γυναίκες).

Το συμπέρασμα είναι ότι για ασθενείς με ΣΔΤ2 αυτά τα μέτρα καρδιαγγειακού κινδύνου παίζουν ένα σημαντικό ρόλο στην κλινική πρακτική. Σε μια κλινική, ο υπολογισμός του κινδύνου οδήγησε σε μια αύξηση της χορήγησης φαρμάκων που συνεισφέρουν στη μείωσή του, και μάλιστα μόνο σε ασθενείς με υψηλό κίνδυνο, όπως είναι το επιθυμητό. Οι συγκεντρώσεις γλυκόζης, είτε συνδυάζουμε συγκεντρώσεις γλυκόζης νηστείας και μεταγευματική 2 ωρών είτε χρησιμοποιούμε μόνο γλυκόζη νηστείας, παίζουν ένα σημαντικό ρόλο στην πρόβλεψη της καρδιαγγειακής θνητότητας σε άτομα με υπογλυκαιμία.

- Η Μηχανή πρόβλεψης UKPDS. Η μηχανή αυτή αναφέρεται ειδικά στο ΣΔΤ2 και μπορεί να χρησιμοποιηθεί για την πρόβλεψη του κινδύνου εμφάνισης εμφράγματος του μυοκαρδίου ή εγκεφαλικού που συμβαίνει μέσα σε μια συγκεκριμένη χρονική περίοδο. Για αυτή τη μηχανή πρόβλεψης έχει αναπτυχθεί μια μελέτη [16] που επιτυγχάνει τον υπολογισμό πρόβλεψης θανάτου από CHD και εγκεφαλικό επεισόδιο.

Η μελέτη αναφέρεται σε 5.102 ασθενείς ηλικίας από 25 έως 65 ετών που παρουσίασαν νέα διάγνωση διαβήτη μεταξύ 1977 και 1992 και τους ακολούθησε από 6 έως και 20 χρόνια. Μία επιτροπή έκρινε τις νόσους κατά άτομο και όρισε το έμφραγμα του μυοκαρδίου (myocardial infarction–MI) ως μη θανατηφόρο MI, το θανατηφόρο MI ή τον ξαφνικό καρδιακό θάνατο. Το εγκεφαλικό ορίστηκε ως θανατηφόρο ή μη με συμπτώματα που διαρκούν πάνω από ένα μήνα. Με τον τρόπο αυτό, το MI ορίστηκε ως θανατηφόρο αν ο θάνατος επήλθε 6 μήνες μετά το γεγονός και το εγκεφαλικό επεισόδιο χαρακτηρίστηκε με το ίδιο σκεπτικό. Η μελέτη διεξήγαγε fitting μοντέλων στατιστικής ανάλυσης για θνητότητα από MI σε 674 περιστατικά MI που συνέβησαν σε 597 ασθενείς. Ακόμη 206 περιστατικά συνέβησαν σε 148 ασθενείς για τους οποίους όλες οι τιμές μεταβλητών είναι διαθέσιμες ή για τους οποίους το MO συνέβη πριν από τη μέτρηση των τιμών αυτών. Το fitting για την περίπτωση του εγκεφαλικού έγινε για 234 περιστατικά εγκεφαλικού που συνέβησαν σε 199 ασθενείς. Ακόμη 69 εγκεφαλικά συνέβησαν σε 60 ασθενείς για τους οποίους δεν ήταν διαθέσιμες όλες οι μεταβλητές ή για τους οποίους το εγκεφαλικό έλαβε χώρα πριν από τη μέτρηση των τιμών τους. Χρησιμοποιήθηκε πολυμεταβλητή λογιστική παλινδρόμηση για να συγκρίνει τα επίπεδα των πιθανών παραγόντων κινδύνου μεταξύ αυτών με θανατηφόρο MI και αυτών με μη θανατηφόρο MI και, ομοίως, μεταξύ αυτών με θανατηφόρο εγκεφαλικό και μη θανατηφόρο εγκεφαλικό. Οι πιθανοί παράγοντες κινδύνου μετρήθηκαν κατά τη διάγνωση του διαβήτη με τις ακόλουθες εξαιρέσεις: για κάθε άτομο, η γλυκοζυλιωμένη αιμοσφαιρίνη, HbA1c, η συστολική πίεση sBP, ο λιπιδικός λόγος (συνολική προς HDL χοληστερόλη), το ΔΜΣ, η αλβουμίνη των ούρων και τα τριγλυκερίδια ορίστηκαν ως οι μέσες τιμές των τιμών που προσελήφθησαν 1 και 2 έτη μετά από τη διάγνωση του διαβήτη. Αυτές οι μέσες τιμές μπορούν να έχουν μεγάλη προβλεπτική ικανότητα. Ο χρόνος μεταξύ διάγνωσης και εμφάνισης διαβήτη θεωρήθηκε συνεχής μεταβλητή για αποφυγή σύγχυσης. Επίσης, ελέγχθηκε αν ένα δεύτερο MI ή εγκεφαλικό ήταν πιο πιθανό να είναι θανατηφόρο από ένα πρώτο με την εισαγωγή στο σύστημα μιας μεταβλητής με τιμή 0 για ένα πρώτο γεγονός και 1 για ένα δεύτερο. Για ασθενείς με τρία ή περισσότερα MI, μόνο τα δύο χρησιμοποιήθηκαν στην ανάλυση. Κανένας από τους ασθενείς δεν είχε περισσότερα από δύο εγκεφαλικά. Η μοντελοποίηση έλαβε χώρα στο Ινστιτούτο SAS (Cary, NC) χρησιμοποιώντας μία βηματική συνάρτηση σύνδεσης αλγόριθμο επιλογής και ένα επίπεδο σημαντικότητας του 5%. Η καταλληλότητα της λογιστικής συνάρτησης σύνδεσης επιβεβαιώθηκε χρησιμοποιώντας το Hosmer-Lemeshow τεστ. Από τα στοιχεία του συνόλου, δεν ήταν όλα ανεξάρτητα, επειδή κάποιοι ασθενείς εμφανίζονται πάνω από μια φορά λόγω περισσότερων του ενός MI ή εγκεφαλικών. Αυτή η εξάρτηση θα μπορούσε να επηρεάσει τις p-values της λογιστικής παλινδρόμησης. Για να περιοριστεί αυτό το γεγονός για ασθενείς με τρία ή περισσότερα

περιστατικά, μόνο τα δύο πρώτα συμπεριλήφθηκαν. Επίσης χρησιμοποιήθηκε έλεγχος αντιμετάθεσης, για να επιβεβαιώσει ότι τα p-values από τη λογιστική παλινδρόμηση δεν επηρεάστηκαν από αυτή την εξάρτηση.

Με δεδομένα αυτά, αναπτύχθηκαν οι UKPDSεξισώσεις για τη θνητότητα από στεφανιαία νόσο και εγκεφαλικό. Αυτές χρησιμοποιούν μοντέλα λογιστικής παλινδρόμησης όπως προηγουμένως, αλλά διαφέρουν σε δύο σημεία. Πρώτον, για λόγους συμβατότητας με τις υπάρχουσες εξισώσεις πρόβλεψης κινδύνου, οι μεταβλητές περιορίστηκαν στις εξής: ηλικία διάγνωσης ΣΔΤ2, διάρκεια από διάγνωση μέχρι συμβάν, φύλο, εθνικότητα, κάπνισμα, HbA1c, sBP και λιπιδικός λόγος. Δεύτερον, για να ελαχιστοποιήσει την πιθανότητα σφάλματος, οι μεταβλητές ελέγχθηκαν στο επίπεδο του 0.5% αντί του 5%, πλην της περίπτωσης που η βιβλιογραφία υποστήριζε μια μεταβλητή ως κύριο παράγοντα κινδύνου. Διεξήχθησαν έλεγχοι των αλληλεπιδράσεων όλων των μεταβλητών που βρέθηκαν σημαντικές. Για να επιβεβαιωθεί ότι τα δεδομένα που απορρίφθηκαν λόγω τιμών που έλειπαν ήταν συμβατά με τα δεδομένα που χρησιμοποιήθηκαν.

Από τα 674 περιστατικά MI που χρησιμοποιούνται στο fittingτου μοντέλου, 52% ήταν θανάσιμα, ενώ από τα 234 εγκεφαλικά το 21% ήταν θανάσιμα. Το πολυμεταβλητό μοντέλο MI αναγνώρισε την ηλικία στη διάγνωση του διαβήτη, το χρονικό διάστημα μεταξύ διάγνωσης και συμβάντος, την HbA1c, την sBP, και την αλβουμίνη ούρων ως καίριους παράγοντες κινδύνου για τη θνητότητα από MI. Αυτό σημαίνει ότι καθένας από αυτούς τους παράγοντες ήταν ανεβασμένος σε αυτούς με θανάσιμο MI σε σύγκριση με αυτούς που έχουν πέρασαν μη θανάσιμο MI. Δε βρέθηκε σημαντική διαφορά μεταξύ του ρυθμού θνητότητας στα πρώτα MIs (51%, 304 θανάσιμα από 597 MIs) και τα δεύτερα MIs (61%, 47 από 77). Ο χρόνος από τη διάγνωση μέχρι το περιστατικό ήταν σημαντικός με μεγαλύτερη θνητότητα σε περιστατικά που συμβαίνουν πολύ αργότερα από τη διάγνωση του διαβήτη. Το Hosmer and Lemeshow τεστ έδειξε ότι η logistic link συνάρτηση είναι μια κατάλληλη επιλογή. Το ημερολογιακό έτος στο οποίο έλαβε χώρα το MI δε συνδέεται με τη θνητότητα του ατόμου. Όταν ο λόγος των δειγμάτων εκπαίδευσης προς το λόγο των δειγμάτων αξιολόγησης είναι μεγάλος, οι πιθανότητες που εξάγονται μπορούν να θεωρηθούν αντιπροσωπευτικές για τον κίνδυνο: σε αυτά τα δεδομένα ο λόγος των θανάσιμων προς τα μη θανάσιμα MI είναι κοντά στο 1.

Το πολυμεταβλητό μοντέλο strokeαναγνώρισε το φύλο, την αυξημένη HbA1c, την sBP, τον αριθμό των λευκών αιμοσφαιρίων, και το προηγούμενο εγκεφαλικό ως σημαντικούς παράγοντες κινδύνου για το εγκεφαλικό επεισόδιο. Τα πρώτα εγκεφαλικά ήταν λιγότερο πιθανό να είναι

θανατηφόρα από τα επόμενα εγκεφαλικά: 18% σε σύγκριση με 60%, αντίστοιχα. Η διάρκεια μεταξύ διάγνωσης και εμφάνισης του προβλήματος δεν ήταν σημαντική όπως και το έτος στο οποίο έλαβε χώρα το επεισόδιο. Και σε αυτή την περίπτωση το Hosmer and Lemeshowτεστ υπέδειξε πως η λογιστική συνάρτηση σύνδεσης είναι κατάλληλη για το μοντέλο.

Όσον αφορά τις εξισώσεις κινδύνου UKPDS, η εξίσωση για τη θνητότητα από MI είναι:

$$\textit{Πιθανότητα να οδηγήσει σε θάνατο ένα MI} = \frac{1}{1+e^{(0.713-0.048)(\textit{ηλικία}-55)-0.178(HbA1c-6.86)-\frac{0.141(sBP-141)}{10}-0.104*\textit{διάρκεια}}}$$

206 περιπτώσεις MI σε 148 ασθενείς δε λήφθηκαν υπόψη στο fitting λόγω ελλιπών δεδομένων. Στις 82 περιπτώσεις (74 ασθενείς) που είχα τα απαιτούμενα δεδομένα για την παραπάνω εξίσωση, ο προβλεπόμενος ρυθμός θνητότητας από MI ήταν 50.1%. Από την άλλη πλευρά, το μοντέλο πρόβλεψης κινδύνου θανάτου από εγκεφαλικό είναι:

$$\textit{Πιθανότητα να οδηγήσει σε θάνατο ένα εγκεφαλικό επεισόδιο}: = \frac{1}{1+e^{(1.684-0.249)-\frac{(sBP-144)}{10}-2.210*\textit{προηγούμενο\_επεισόδιο}}}$$

Οι εξισώσεις θνητότητας μπορούν να συνδυαστούν με τις εξισώσεις κινδύνου για MI ή εγκεφαλικό προκειμένου να προβλεφθούν οι κίνδυνοι για θανατηφόρα MI και εγκεφαλικά. Οι εξισώσεις κινδύνου από τη UKPDS μελέτη μπορούν να χρησιμοποιηθούν για αυτό το σκοπό.

$$\textit{Πιθανότητα θανατηφόρου MI σε } t \textit{ έτη} = (\textit{κίνδυνος MI σε } t \textit{ έτη}) \times (\textit{θνητότητα σε } \frac{t}{2} \textit{έτη})$$

Στα συμπεράσματα αυτής της ανάλυσης συγκαταλέγεται το γεγονός ότι στο ΣΔΤ2, η HbA1c είναι σημαντικός παράγοντας κινδύνου για τη θνητότητα από MI: αυτό σημαίνει ότι αυτοί που

έχουν θανατηφόρο είχαν υψηλότερη HbA1 στα προηγούμενα χρόνια από αυτούς με μη θανατηφόρο ΜΙ. Από την άλλη πλευρά, η υπεργλυκαιμία μετά από ΜΙ φαίνεται μόνο να συνεισφέρει στη βαρύτητα του επεισοδίου. Ένα ακόμη συμπέρασμα είναι ότι η HbA1cεπηρεάζει και την πιθανότητα για εγκεφαλικό επεισόδιο. Το κάπνισμα δε φάνηκε να συνδέεται με θνητότητα από ΜΙ, σε αντίθεση με άλλες μελέτες στο διαβήτη και στο γενικό πληθυσμό. Το εύρημα της μελέτης ότι τα λευκά αιμοσφαίρια προστατεύουν από τον κίνδυνο εγκεφαλικού οφείλεται στο γεγονός ότι το πλήθος τους αποτελεί ένδειξη εμβολικού ισχαιμικού εγκεφαλικού και όχι αιμορραγικού, άρα παρέχει γνώση για τη φύση του προβλήματος. Κλείνοντας, αυτό που έχει να προσφέρει η μελέτη είναι οι συναρτήσεις εκτίμησης κινδύνου ΜΙ και εγκεφαλικού σε ασθενείς με ΣΔΤ2, που θα είναι χρήσιμες για την κατασκευή στρατηγικής από γιατρούς.

Από την άλλη πλευρά η Διεθνής Ομοσπονδία για το Διαβήτη (International Diabetes Federation - IDF) όρισε guidelines [12] που προτείνουν τη χρήση της μηχανής πρόβλεψης κινδύνου UKPDS. Τα χαρακτηριστικά των τριών δεικτών που προαναφέραμε συνοψίζονται στους Πίνακες 2.1,2.2, 2.3, αντίστοιχα. Αξιοσημείωτο είναι το γεγονός ότι τα Framingham και DECODE μπορούν να εφαρμοστούν στο γενικό πληθυσμό ενώ το UKPDS αναφέρεται μόνο στους ασθενείς με ΣΔΤ2.

Οι μεθοδολογίες που χρησιμοποιήθηκαν για την ανάπτυξη των μοντέλων βασίστηκαν σε ανάλυση επιβίωσης και λογιστική παλινδρόμηση. Όσον αφορά την απόδοση της αξιολόγησης, προηγούμενες μελέτες έδειξαν ότι τα μέτρα καρδιαγγειακού κινδύνου, όπως τα Framingham και Decode, που βασίζονται στο γενικό πληθυσμό, υποτιμούν τον κίνδυνο καρδιαγγειακής νόσου σε άτομα με διαβήτη [17], [18]. Η μηχανή UKPDS επιβεβαιώθηκε σε δύο μελέτες με πολύ διαφορετικά αποτελέσματα [19]. Η μία μελέτη παρατήρησε μια AUC ίση με 0.61 και κακό calibration, ενώ η άλλη παρατήρησε μια AUC ίση μια AUC ίση με 0.86 και καλό calibration. Η μηχανή UKPDS για CHD επιβεβαιώθηκε σε 8 μελέτες. Το discrimination είχε εύρος από 0.65 μέχρι 0.76, και οι περισσότερες από αυτές τις μελέτες παρατήρησαν κακό calibration και υπερεκτίμηση του κινδύνου.

Δεδομένου ότι η απόδοση των καρδιαγγειακών δεικτών κινδύνου είναι γενικά μέτρια, επιλέξαμε να ερευνήσουμε τη δυνατότητα εφαρμογής πιο εκλεπτυσμένων στρατηγικών για να αναπτύξουμε ένα καινούριο μοντέλο πρόβλεψης κινδύνου καρδιαγγειακής νόσου που να μην έχει τις αδυναμίες των μέτρων που μελετήσαμε σε αυτό το κεφάλαιο.

Πίνακας 2.1 Χαρακτηριστικά του μοντέλου Framingham

Framingham	
Μοντέλο Εκτίμησης Κινδύνου	Μοντέλο επιβίωσης βασισμένο στην κατανομή Weibull
Αποτέλεσμα	Επιπλοκές: - Καρδιαγγειακή Νόσος
Ορίζοντας Πρόβλεψης	10 έτη
Χαρακτηριστικά	- Ηλικία - Φύλο - Συστολική ή διαστολική πίεση - Κάπνισμα - Παρουσία ή όχι ΣΔ - ΔΜΣ

Πίνακας 2.2 Χαρακτηριστικά του μοντέλου DECODE

DECODE	
Μοντέλο Εκτίμησης Κινδύνου	Μοντέλο Cox παλινδρόμησης αναλογικών κινδύνων
Αποτέλεσμα	Επιπλοκές: - Θνητότητα λόγω καρδιαγγειακής νόσου
Ορίζοντας Πρόβλεψης	5 και 10 έτη
Χαρακτηριστικά	- Ηλικία - Γλυκόζη νηστείας και μεταγευματική 2 ωρών - Γλυκόζη νηστείας - Χοληστερόλη - Κάπνισμα - Συστολική πίεση - ΔΜΣ

Πίνακας 2.3 Χαρακτηριστικά του μοντέλου UKPDS

UKPDS Risk Engine	
Μοντέλο Εκτίμησης Κινδύνου	Πολυπαραγοντική λογιστική παλινδρόμηση
Ορίζοντας Πρόβλεψης	5 και 10 έτη
Αποτέλεσμα	Επιπλοκές: - Μη-θανάσιμη και θανάσιμη στεφανιαία νόσος - Θανάσιμη στεφανιαία νόσος - Μη-θανάσιμο και θανάσιμο εγκεφαλικό - Θανάσιμο εγκεφαλικό
Χαρακτηριστικά	- Ηλικία - Φύλο - Συστολική πίεση - Κάπνισμα - Κολπική Μαρμαρυγή - Εθνικότητα - Γλυκοζυλιωμένη Αιμοσφαιρίνη - Συνολική χοληστερόλη - HDL χοληστερόλη

Κεφάλαιο 3 Μηχανική Μάθηση

3.1 Εισαγωγή

Η Μηχανική Μάθηση είναι ένας όρος που αναφέρεται σε ένα πολύ σημαντικό πεδίο της επιστήμης των υπολογιστών, καθώς συνδυάζει δημοφιλείς τεχνικές που ανήκουν στις επιστημονικές περιοχές της αναγνώρισης προτύπων και της τεχνητής νοημοσύνης [20]. Πιο συγκεκριμένα, η μηχανική μάθηση περιέχει τη συσσωρευμένη γνώση αλγορίθμων που στοχεύουν στην εκμάθηση από δεδομένα καθώς και στην παραγωγή προβλέψεων για αυτά. Η Μηχανική Μάθηση είναι γνωστή και ως προβλεπτική ανάλυση και μοντελοποίηση σε βιομηχανικές εφαρμογές, με αλγορίθμους που δέχονται εισόδους και τις αξιοποιούν έτσι, ώστε να κατασκευάσουν μοντέλα που παράγουν αποκρίσεις οδηγούμενες από τα δεδομένα (data-driven).

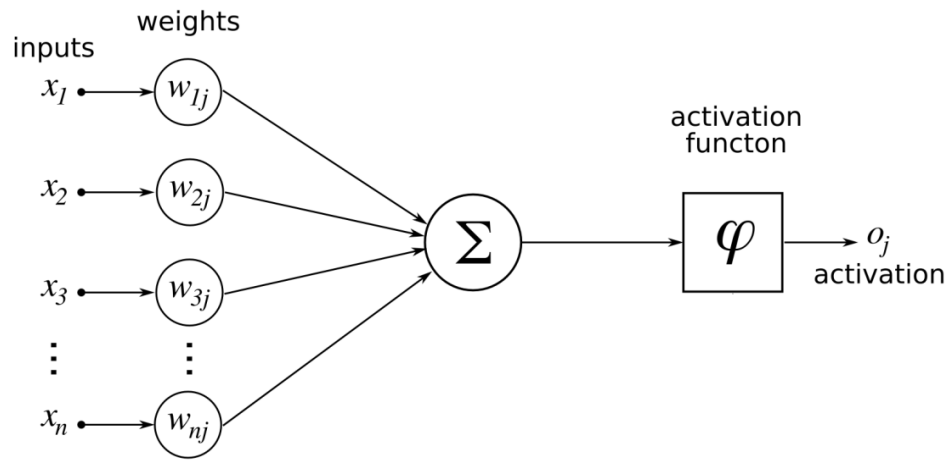
Η Ανάλυση Δεδομένων και η Μηχανική Μάθηση είναι πεδία που αλληλεπικαλύπτονται, καθώς η ταξινόμηση και η ανάλυση παλινδρόμησης χρησιμοποιούνται και από τα δύο προαναφερόμενα πεδία. Συνεπώς, η μηχανική μάθηση συνδέεται με αυστηρά μαθηματικές τεχνικές βελτιστοποίησης, χωρίς όμως να χάνει τα πλεονεκτήματα που παρέχονται από την επιστήμη των υπολογιστών. Η Μηχανική Μάθηση μπορεί να χειρίζεται προβλήματα με ασαφή δομή, γεγονός που την καθιστά πολύ χρήσιμη σε απαιτητικές εφαρμογές όπως το φιλτράρισμα spam, οι μηχανές αναζήτησης και η όραση υπολογιστών. Επιπλέον, αξιοσημείωτο είναι το γεγονός ότι παρόλο που η μηχανική μάθηση συχνά θεωρείται ταυτόσημη του όρου Εξόρυξη Δεδομένων, η τελευταία επιδεικνύει περισσότερο ενδιαφέρον σε μαθηματικές και όχι τόσο σχετικές με τους ηλεκτρονικούς υπολογιστές πτυχές των προβλημάτων. Στα επόμενα παρατίθενται τεχνικές μηχανικής μάθησης, πολλές από τις οποίες έχουν χρησιμοποιηθεί για την πρόβλεψη του κινδύνου, παρόλο που είναι πιο σύνηθες τα περισσότερα σχετικά μοντέλα να βασίζονται σε ανάλυση επιβίωσης (survival analysis) και πολυπαραγοντική παλινδρόμηση (multivariate regression).

3.2 Μέθοδοι

3.2.1 Τεχνητά Νευρωνικά Δίκτυα

Πρόκειται για μια πολύ δημοφιλή μέθοδο που έλαβε το όνομα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks - ANNs) από το γεγονός ότι ο ανθρώπινος εγκέφαλος εκτελεί υπολογισμούς με έναν τελείως διαφορετικό τρόπο από αυτόν που αφορά τους ψηφιακούς υπολογιστές [21]. Τα μοναδικά χαρακτηριστικά του εγκεφάλου είναι η αυξημένη πολυπλοκότητα, η μη γραμμικότητα και η προσαρμοστικότητα. Εκτός από την κληρονόμηση εγκεφαλικών νευροφυσιολογικών δυνατοτήτων, τα ANNs αναπτύσσονται στο εσωτερικό των υπολογιστικών συστημάτων, κερδίζοντας επίσης πολλά από τα πλεονεκτήματα των τελευταίων. Για το λόγο αυτό χαρακτηρίζονται από ανοχή στα σφάλματα, ομοιομορφία ανάλυσης και σχεδίασης, καθώς και πιθανή υλοποίηση σε VLSI.

Το γενικό μοντέλο ενός νευρώνα, που είναι η μονάδα επεξεργασίας πληροφοριών, αποτελείται από ένα σύνολο διασυνδέσεων, τις συνάψεις, οι οποίες έχουν τη δική τους ισχύ (συχνά αποκαλούμενη συναπτικό βάρος) και εισέρχονται σε ένα γραμμικό συνδυαστή (αθροιστή). Μιμούμενοι τη λειτουργία των δυναμικών δράσης στους φυσικούς νευρώνες, οι τεχνητοί νευρώνες έχουν έξοδο της οποία το μέγεθος καθορίζεται από μια συνάρτηση ενεργοποίησης, η οποία είναι κατά βάση μια συνάρτηση κατωφλίου ή μια σιγμοειδής συνάρτηση. Ωστόσο, το μοντέλο του νευρώνα ενδέχεται να μην είναι ντετερμινιστικό, όπως αυτό που μόλις περιγράφηκε. Η απόφαση για ενεργοποίηση ενός στοχαστικού νευρώνα υπόκειται σε πιθανοτικές διαδικασίες.



Σχήμα 3.1 Ένας μη γραμμικός νευρώνας

Οι νευρώνες που περιγράψαμε έως τώρα αποτελούσαν μέρη συστημάτων ανοιχτού βρόχου. Υπάρχει επίσης μια ειδική κατηγορία νευρωνικών δικτύων που αναφέρεται ως επαναλαμβανόμενα (recurrent) και περιλαμβάνουν σχήματα ανάδρασης ελέγχου. Η λειτουργική διαφορά αυτής της υποκατηγορίας τεχνητών νευρωνικών δικτύων συμπληρώνεται και από μια ξεχωριστή αρχιτεκτονική, καθώς η έξοδος ανατροφοδοτείται στην είσοδο κάθε νευρώνα. Οι βασικές αρχιτεκτονικές ANNs είναι οι προαναφερθείσες, με τα νευρωνικά δίκτυα ευθέος κλάδου να μπορούν να ταξινομηθούν σε δύο βασικές υποκατηγορίες:

- Νευρωνικά δίκτυα πρόσθιας τροφοδότησης ενός επιπέδου, όπου ένα επίπεδο εισόδου συνδέεται άμεσα σε ένα επίπεδο νευρώνων εξόδου.
- Δίκτυα ευθέος κλάδου πολλών επιπέδων, όπου υπάρχουν ένα ή περισσότερα κρυφά επίπεδα. Οι εσωτερικού νευρώνες ονομάζονται επίσης κρυφοί, αφού αυτό το τμήμα του δικτύου δεν είναι ορατό από την είσοδο ή την έξοδο. Ο λόγος για τον οποίο εισήχθησαν περισσότερα από ένα επίπεδα ήταν η στατιστική υπεροχή στην απόδοση των νευρωνικών δικτύων μετά από αυτή την προσθήκη.

Μια άλλη κατηγοριοποίηση των νευρωνικών δικτύων προέρχεται από την παρουσία ή την απουσία "δασκάλου" κατά την εκμάθηση:

- Επιβλεπόμενη μάθηση, που περιλαμβάνει την παρουσία δασκάλου, ο οποίος είναι μια οντότητα που διαθέτει μερική γνώση του περιβάλλοντος μεταξύ των εισόδων και εξόδων, την οποία αξιοποιεί για να καθοδηγήσει τη συμπεριφορά του δικτύου χρησιμοποιώντας διόρθωση σφάλματος.

- Μάθηση χωρίς επίβλεψη, όπου δύο γενικές μέθοδοι μπορούν να οριστούν. Η πρώτη είναι γνωστή ως ενισχυμένη μάθηση και περιλαμβάνει ένα μηχανισμό γνωστό ως κριτή, ο οποίος επιχειρεί να μειώσει ένα κόστος. Η δεύτερη ονομάζεται εκμάθηση χωρίς επίβλεψη, όπου δεν υπάρχουν οντότητες όπως ο δάσκαλος ή ο κριτής. Αυτή η μέθοδος συνοδεύεται από ένα ανεξάρτητο μέτρο ποιότητας, που παρατηρεί το δίκτυο και σύμφωνα με αυτές τις παρατηρήσεις διορθώνει το δίκτυο.

Υπάρχουν πολλοί τύποι νευρωνικών δικτύων, με πιο δημοφιλείς τους εξής:

- Το perceptron του Rosenblatt και το perceptron πολλών επιπέδων. Το πρώτο αναφέρεται σε ένα μη γραμμικό νευρώνα, όπως ήταν αυτός που είδαμε στο Σχήμα 3.1. Το perceptron πολλών επιπέδων είναι ένα δίκτυο με περισσότερες από μία εισόδους, κρυφά στρώματα και εξόδους, και, όπως προαναφέρθηκε, προσφέρει υψηλότερης ποιότητας υλοποιήσεις σε σύγκριση με το perceptron του Rosenblatt, που διαθέτει μόνο ένα επίπεδο.

- Οι Kohonen αυτο-οργανούμενοι χάρτες ή SOM, που αξιοποιούν την ιδέα της ανταγωνιστικής μάθησης. Οι νευρώνες εξόδου του χάρτη ανταγωνίζονται ο ένας τον άλλο για την απόκτηση άδειας ενεργοποίησης, με αποτέλεσμα μόνο ένας νευρώνας να είναι ενεργός κάθε στιγμή. Ο νικητής νευρώνας αποκτά τότε προνόμιο της λογικής winner-takes-all και το αποτέλεσμα της διαδικασίας είναι ένας χάρτης με τη μορφή δικτύου, ισοδύναμος με ένα σύστημα λογικών συντεταγμένων.

- Τα Δίκτυα Συναρτήσεων Ακτινικής Βάσης (Radial Basis Function - RBF), που αποτελούνται από μεθόδους παρεμβολής σε πολυδιάστατους χώρους και προέρχονται από την ιδέα ότι ένας προβλεπόμενος στόχος για την τιμή ενός δείγματος θα είναι πιθανότατα σχεδόν ο ίδιος με άλλων δειγμάτων που έχουν παρόμοιες τιμές χαρακτηριστικών.

Ένα παράδειγμα μοντέλων πρόβλεψης κινδύνου που χρησιμοποιείται για πρόβλεψη της εμφάνισης καρδιοπάθειας προτείνεται από τους Jain και Singh [22], αποδεικνύοντας πως τα ANNs μπορούν με αποδοτικό τρόπο να δημιουργήσουν συστήματα για να προβλέπουν τη στεφανιαία νόσο (Coronary Heart Disease - CHD). Είναι πολύ σημαντικό να σημειώσουμε ότι ο διαβήτης ήταν ανάμεσα στους παράγοντες κινδύνου που λήφθηκαν υπόψη για το μοντέλο (μαζί με την ηλικία, την αρτηριακή υπέρταση, το κάπνισμα, την παχυσαρκία, την κληρονομικότητα, τη γλυκόζη, τη χοληστερίνη, τα τριγλυκερίδια κλπ.). Για το λόγο αυτό κατασκευάστηκε ένα διαγνωστικό μοντέλο, ικανό να συνδέσει με ακρίβεια γενετικούς και μη παράγοντες με τη CHD, θέτοντας τις προϋποθέσεις για ανάπτυξη σχετικού λογισμικού.

3.2.2 Εκμάθηση κανόνων συσχέτισης

Η εκμάθηση κανόνων συσχέτισης (Association Rules Learning – ARL) είναι μια τεχνική εξόρυξης δεδομένων χωρίς επίβλεψη, η οποία εξειδικεύεται στην ανακάλυψη σημαντικών σχέσεων και εξαρτήσεων (γνωστών ως συσχετίσεων) σε μεγάλα σύνολα αντικειμένων, που αποτελούν τα δεδομένα της μεθόδου. Ο όρος συναλλαγές (transactions) είναι υψίστης σημασίας και αναφέρεται στην κατάσταση των αντικειμένων που εμπλέκονται στην εκμάθηση, γεγονός που σημαίνει ότι οι συναλλαγές, που παράγονται εξωτερικά ή λαμβάνονται από σχεσιακές βάσεις δεδομένων, είναι το κύριο ζήτημα που πραγματεύεται η μέθοδος.

Τα χαρακτηριστικά των κανόνων συσχέτισης επιτρέπουν τον ορθό χειρισμό ακόμη και μετά από κλιμάκωση στον όγκο των δεδομένων, κάτι που καθιστά τη μέθοδο ARLένα πολύ ουσιώδες εργαλείο για την εξόρυξη δεδομένων.

Τα πεδία εφαρμογών της μεθόδου ανήκουν σε ένα ευρύ φάσμα. Για το λόγο αυτό, η ARL αποδεικνύεται χρήσιμη σε πολλές περιοχές εξειδίκευσης, από την επιχειρηματική ανάληψη αποφάσεων μέχρι τη συσταδοποίηση (clustering) και την ανάλυση γονιδιωματικών δεδομένων [23].

Για να κατανοήσει κανείς τη σύνδεση μεταξύ των αντικειμένων και των συναλλαγών, το επόμενο παράδειγμα μπορεί να φανεί χρήσιμο: αν αντί για ένα αντικείμενο θεωρήσουμε την παρουσία ή την απουσία του σε ένα σύνολο δεδομένων, τότε μια Boolean μεταβλητή θα μπορούσε να οριστεί για κάθε αντικείμενο και η οποία θα ήταν η μεταβλητή ενδιαφέροντος. Συνεπώς, ο σκοπός της μεθόδου θα ήταν να αποφασίσει ποιος συναλλαγές συσχετίζονται στο τρέχον περιβάλλον. Για παράδειγμα ο σκοπός της

μεθόδου θα ήταν να αποφασίσει ποια αντικείμενα είναι συνήθως παρόντα ή απόντα την ίδια χρονικά στιγμή.

Η έννοια που περιγράφεται στα προηγούμενα μπορεί να αναπαρασταθεί από λογικές συνεπαγωγές με τη μορφή κανόνων συσχέτισης:

$$LHS \Rightarrow RHS \text{ [support, confidence]},$$

όπου η αριστερή πλευρά (left-hand side- LHS) της συνεπαγωγής καταλήγει στη δεξιά πλευρά της (right-hand side- RHS), με την παρουσία δύο μέτρων ποιότητας, που ονομάζονταισupportκαιconfidence.

Τα Supportκαιconfidenceμετρούν τη σημαντικότητα ενός κανόνα συσχέτισης, σε όρους χρηστικότητας ή ισχύος και βεβαιότητας. Το supportαναφέρεται στον αριθμό των συναλλαγών που χρησιμοποιήθηκαν για να παραχθεί ένας κανόνας συσχέτισης και περιλαμβάνει αντικείμενα τόσο από το LHSόσο και από τοRHS, ενώ το confidence αναφέρεται στον αριθμό των συναλλαγών με αντικείμενα από το LHS που επίσης περιλαμβάνουν αντικείμενα από το RHS. Συνήθως εκφράζονται ως ποσοστά και καθορίζουν το ενδιαφέρον ενός κανόνα συσχέτισης. Οι κανόνες που υπερβαίνουν το ελάχιστο κατώφλι του support και το ελάχιστο κατώφλι του confidence ταυτόχρονα θεωρούνται ισχυροί κανόνες συσχέτισης. Η επιλογή των κατωφλίων ανήκει στο χρήστη, που αναμένεται να έχει κάποια εμπειρία στο πεδίο.

Υπάρχουν τρία σχήματα κατηγοριοποίησης για τους κανόνες συσχέτισης [23]:

- Για να οριστούν οι κανόνες συσχέτισης, τα γεγονότα των συναλλαγών πρέπει να λάβουν χώρα την ίδια χρονική στιγμή ή σε μικρό χρονικό διάστημα.Όταν ένα μόνο γεγονός απαιτείται για να παραχθεί ένας κανόνας συσχέτισης, ο κανόνας ονομάζεται μονοδιάστατος. Αντίθετα, όταν περισσότερες συναλλαγές πρέπει να λαμβάνουν χώρα ταυτόχρονα για να επικυρωθεί ο κανόνας, τότε ο τελευταίος ονομάζεται πολυδιάστατος.
- Μια συναλλαγή μπορεί να είναι Boolean ή ποσοτική. Μια Boolean μεταβλητή δηλώνει τη παρουσία ή την απουσία ενός αντικειμένου, ενώ μια ποσοτική μεταβλητή αναφέρεται σε περιοχές τιμών των αντικειμένων.

- Ο τρίτος τύπος ταξινόμησης των κανόνων συσχέτισης αναφέρεται στους όρους ενός επιπέδου και πολυεπίπεδος. Ο πρώτος όρος αφορά ένα επίπεδο αφαίρεσης, ενώ ο δεύτερος αναφέρεται σε αντικείμενα που ανήκουν σε διάφορα επίπεδα μιας ιεραρχίας.

Όπως αναφέρθηκε και προηγουμένως, τα δεδομένα εισόδου σε αυτή τη μέθοδο έχουν τη μορφή συσχετίσεων. Κάθε συναλλαγή συνοδεύεται από ένα αναγνωριστικό, καθώς και από πληροφορίες για κάθε αντικείμενο που συμμετέχει στη συναλλαγή. Η πηγή άντλησης αυτών των αντικειμένων μπορεί να είναι μια σχεσιακή βάση δεδομένων ή αυτά μπορεί να προέρχονται από ένα μετασχηματισμό σχεσιακών δεδομένων. Μια βάση δεδομένων που συμπεριλαμβάνει συναλλαγές στη μορφή λίστας μπορεί να χρησιμοποιηθεί ως είσοδος σε ένα σύστημα που προσπαθεί να ορίσει κανόνες συσχέτισης που συνδέουν την παρουσία ενός συνόλου αντικειμένων με ένα άλλο σύνολο αντικειμένων.

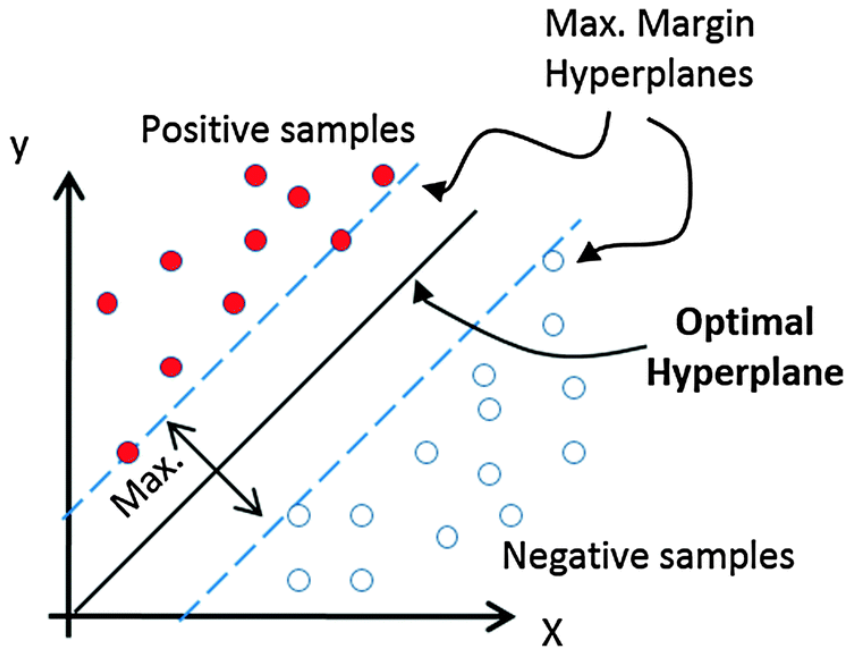
Στην εξόρυξη γνώσης από δεδομένα, ένα σύνολο αντικειμένων ονομάζεται itemset, και ένα itemset με πληθικότητα k είναι ένα k-itemset. Το support count (γνωστό και ως συχνότητα ή συχνότητα εμφάνισης) ενός itemset είναι ένα μέτρο όλων των συναλλαγών που περιέχουν το itemset. Ένα itemset θεωρείται συχνό (frequent) όταν ξεπερνά ένα ελάχιστο επίπεδο support. Αυτό το γεγονός ισοδυναμεί με το να είναι το support count μεγαλύτερο ή ίσο του γινομένου του αριθμού όλων των συναλλαγών και του κατωφλίου του support. Είναι εύκολο να συμπεράνει κανείς ότι η παραγωγή ισχυρών κανόνων συσχέτισης απαιτεί την εύρεση συχνών itemsets, που είναι μια δύσκολη διαδικασία όταν το πλήθος των αντικειμένων είναι μεγάλο. Η πιο συχνή μορφή κανόνων συσχέτισης είναι boolean, ενός επιπέδου και μονοδιάστατη. Ωστόσο, μπορεί να γίνουν πιο περίπλοκοι σε μεθόδους που απαιτούν να είναι πολυεπίπεδοι, πολυδιάστατοι και ποσοτικοί [23].

Η μέθοδος ARLαποδεικνύεται χρήσιμη σε μοντέλα πρόβλεψης του κινδύνου εμφάνισης καρδιαγγειακής νόσου, όπως μπορεί κάποιος να συμπεράνει από βιβλιογραφικές πηγές. Οι R. Thanigaivel και Δρ. K. Ramesh Kumar έχουν αναπτύξει ένα σύστημα πρόβλεψης της εμφάνισης καρδιαγγειακής νόσου, λαμβάνοντας υπόψη πλήθος παραγόντων κινδύνου, ανάμεσα στους οποίους βρίσκεται και ο διαβήτης. Χρησιμοποιώντας ένα μοντέλο εκμάθησης κανόνων συσχέτισης, δημιούργησαν ένα σύστημα επαρκώς υψηλής απόδοσης, και το οποίο είναι κατάλληλο για την πρόγνωση της νόσου σε πρώιμο στάδιο [24].

3.2.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support Vector Machines –SVMs) είναι μία δημοφιλής επιβλεπόμενη μέθοδος μηχανικής μάθησης, που αποτελείται από μοντέλα και αλγορίθμους ικανούς να αναλύουν πληροφορία και να ενσωματώνουν τεχνικές αναγνώρισης προτύπων, για προβλήματα ταξινόμησης και ανάλυσης παλινδρόμησης. Όταν ένα σύνολο δυαδικών δεδομένων εκπαίδευσης είναι διαθέσιμο, ένα ΜΔΥ ταξινόμησης κατασκευάζει ένα μοντέλο που αποφασίζει αν ένα νέο δείγμα πρέπει να αποδοθεί σε μια κατηγορία ή στις άλλες υπάρχουσες.

Για το λόγο αυτό, κάθε SVNθεωρείται ένας ντετερμινιστικός δυαδικός γραμμικός ταξινομητής. Ο σκοπός τέτοιων δικτύων είναι να καθοριστούν όλα τα δείγματα σαν σημεία στο χώρο με ένα συγκεκριμένο τρόπο. Η μέθοδος προσπαθεί να δημιουργήσει ομάδες δειγμάτων που να διαχωρίζονται από ένα κενό τόσο ευρύ όσο αυτό είναι εφικτό. Μετά από τη διαδικασία εκπαίδευσης, τα νέα δείγματα προβλέπεται πως ανήκουν σε μια κατηγορία και τοποθετούνται ως σημεία στον ίδιο χώρο. Η πρόβλεψη εξαρτάται από την πλευρά του κενού στην οποία καταλήγουν. Το κενό είναι γνωστό και σαν περιθώριο (margin)και είναι η περιοχή στο κέντρο της οποίας ορίζεται ένα υπερεπίπεδο, γνωστό σαν βέλτιστο υπερεπίπεδο, καθώς εφαρμόζει βέλτιστο διαχωρισμό των διαφορετικών ομάδων. Ένα δισδιάστατο παράδειγμα του διαχωρισμού απεικονίζεται στο Σχήμα 2.2.



Σχήμα3.2 Ένας δυαδικός γραμμικός ταξινομητής που πραγματοποιεί διαχωρισμό

Το παράδειγμα της δυαδικής ταξινόμησης θα αναλυθεί περαιτέρω για λόγους κατανόησης: Μια συνάρτηση απώλειας ορίζεται και αντιστοιχεί στο κόστος που θα επιβαρύνει το μοντέλο όταν μια λάθος επιλογή κατηγορίας λάβει χώρα. Όπως προαναφέρθηκε, οι κατηγορίες είναι δύο στη δυαδική ταξινόμηση. Ένα παράδειγμα συστήματος που απαιτεί αυτή τη μέθοδο είναι ένα διαγνωστικό μοντέλο για μια ασθένεια, και το οποίο έχει μόνο δύο καταστάσεις σαν αποτέλεσμα. Ο στόχος της εν λόγω μεθόδου μηχανικής μάθησης είναι να βρεθεί μια συνάρτηση που προσεγγίζει το ελάχιστο πιθανό ρίσκο. Ωστόσο, αυτή η συνάρτηση δεν μπορεί να κατασκευαστεί άμεσα, καθώς η κατανομή πιθανότητας των εισόδων και εξόδων είναι άγνωστη. Αν κάποιος προσπαθήσει να εξαγάγει μια συνάρτηση χρησιμοποιώντας μόνο τα διαθέσιμα δείγματα (επιλέγοντας, για παράδειγμα, τη συνάρτηση που ελαχιστοποιεί το μέσο κόστος), μπορεί να οδηγηθεί σε υπερεκπαίδευση, που οφείλεται στην πεποίθηση ότι αυτός ο πεπερασμένος αριθμός δειγμάτων μπορεί να αντιπροσωπεύσει επαρκώς την κατανομή, προκειμένου να γίνει χρήση σε στατιστικές διαδικασίες. Μια καλύτερη προσέγγιση θα ήταν να οριστεί ένα μικρό σύνολο συναρτήσεων που προσπαθούν να ελαχιστοποιήσουν την τιμή τους σε αυτό το σύνολο (εμπειρική ελαχιστοποίηση κόστους – empirical risk minimization-ERM).

Η ανάπτυξη του αλγορίθμου εκμάθησης για το ΜΔΥ βασίζεται σε μια θεμελιώδη ιδέα, που είναι γνωστή ως πυρήνας εσωτερικού γινομένου. Αναφέρεται στο εσωτερικό γινόμενο ενός διανύσματος υποστήριξης και σε ένα διάνυσμα που προέρχεται από το χώρο των εισόδων του συστήματος. Τα διανύσματα υποστήριξης αποτελούνται από ένα μικρό υποσύνολο σημείων που εξάγονται από το δείγμα εκπαίδευσης. Αυτή η ιδιότητα οδήγησε στην εισαγωγή του όρου «μέθοδος πυρήνα» (kernel method), που αναφέρεται στον αλγόριθμο εκμάθησης και επιτυγχάνει βέλτιστα αποτελέσματα.[21] [25] [26].

Μια εφαρμογή των ΜΔΥ στην προσπάθεια συσχέτισης του ΣΔΤ2 με τις καρδιοπάθειες αναπτύσσεται στο [27]. Πολυάριθμοι παράγοντες κινδύνου λαμβάνονται υπόψη, ενώ η διαδικασία περιλαμβάνει προεπεξεργασία δεδομένων και μείωση διάστασης μέσω Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis– PCA). Με το πέρας της προεπεξεργασίας, το μοντέλο περιλαμβάνει δύο μεγάλα τμήματα, ένα για εκμάθηση χωρίς επίβλεψη χρησιμοποιώντας αλγορίθμους συσταδοποίησης βασιζόμενους σε Ασαφή Λογική και ένα για πρόβλεψη του ΣΔΤ2 με τη χρήση ΜΔΥ.

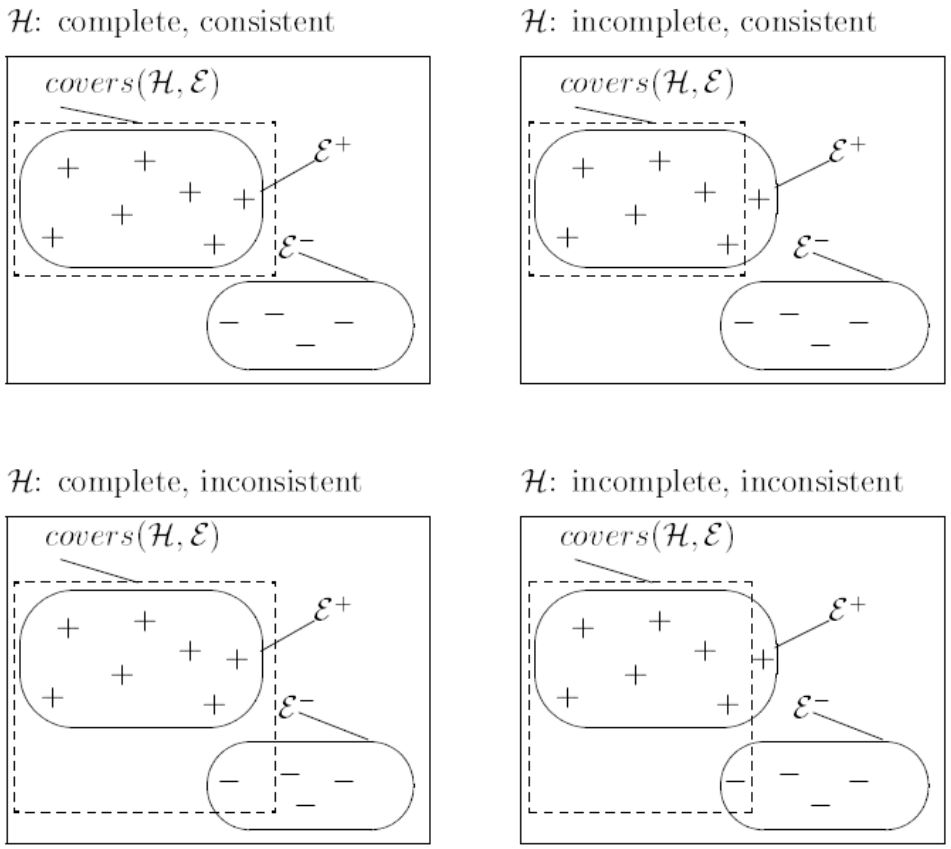
3.2.4 Προγραμματισμός Επαγωγικής Λογικής

Ο Προγραμματισμός Επαγωγικής Λογικής (Inductive Logic Programming - ILP), σε αντίθεση με άλλες μεθόδους μηχανικής μάθησης, πραγματεύεται την εκμάθηση κανόνων πρώτης τάξης (γνωστών και ως σχεσιακών κανόνων). Αυτή η μοναδική ιδιότητα της μεθόδου αποδίδει πιο εκφραστικές αναπαραστάσεις σε σχέση με άλλες που προκύπτουν από δέντρα απόφασης και νευρωνικά δίκτυα. Η εξέλιξη τουILP είναι αρκετά πρόσφατη, και η σημειολογία και ορολογία των στοιχείων που τον αποτελούν δεν έχουν πλήρως αποφασιστεί. Ωστόσο, πρόκειται για μια υποσχόμενη μέθοδο, όπως θα συμπέραινε κάποιος αν επιχειρούσε να κατανοήσει τη δομή και τους στόχους της.
Όταν πρόκειται για σχεσιακή εκμάθηση, η πρότερη γνώση θεωρείται καίριος παράγοντας. Η μέθοδος προσπαθεί να ανακαλύψει μια σχέση δεδομένης αυτής της γνώσης και ενός συνόλου παραδειγμάτων. Τα παραδείγματα μπορεί είτε να συνεισφέρουν στη διαδικασία εκμάθησης (θετικά παραδείγματα – positive examples ή E^+) είτε όχι (negative examples or E^-). Οι μηχανές εκμάθησης σχέσεων πρέπει να χρησιμοποιούν μια γλώσσα υπόθεσης *L*, προκειμένου να δηλώσουν τα χαρακτηριστικά τους. Όταν πρόκειται για τη γλώσσα των λογικών προγραμμάτων, η μέθοδος εκμάθησης ονομάζεται ILP. Τα παραδείγματα που δίνονται, η πρότερη γνώση και η υπόθεση δηλώνονται όλα σε μια μορφή λογικού προγράμματος.

Αν δεν υπάρχει πρότερη γνώση για να υποστηρίξει τη διαδικασία εκμάθησης, η μηχανή αξιοποιεί μόνο τα παραδείγματα που είναι διαθέσιμα. Ωστόσο, η πρότερη γνώση τυπικά απαιτείται, προκειμένου η μηχανή να επιτύχει μια πιο γενική και ακριβή έκφραση των παραδειγμάτων.

Η μεθοδολογία ILP προέρχεται από την Επαγωγική μέθοδο Εκμάθησης Εννοιών (Inductive Concept Learning– ICL), που συνοψίζεται από την εξής απλουστευμένη ιδέα: δεδομένου ενός συνόλου θετικών και αρνητικών παραδειγμάτων μιας έννοιας *C*, πρέπει να βρεθεί μια υπόθεση*H*, εκφρασμένη σε μια δεδομένη γλώσσα περιγραφής εννοιών *L*, ώστε κάθε θετικό παράδειγμα να καλύπτεται από την *H*και κανένα αρνητικό παράδειγμα να μην καλύπτεται από την*H* [28] [29].

Μια υπόθεση *H* είναι πλήρης (complete) αν καλύπτει όλα τα θετικά παραδείγματα, και συνεπής (consistent) αν δεν καλύπτει κανένα αρνητικό παράδειγμα. Δεδομένων αυτών των όρων, κάποιες πιθανές καταστάσεις της ICL, όσον αφορά τις έννοιες completeness και consistency για τη συγκεκριμένη υπόθεση, απεικονίζονται στο Σχήμα 2.3 [29].



Σχήμα 3.3 Πιθανές καταστάσεις της ICL, όσον αφορά τα completeness και consistency

Υπάρχουν δύο σημαντικοί τύποι του ILP [28]:

- Ο εμπειρικός ILP, στον οποίο ένα σύνολο παραδειγμάτων *E* δίνεται σε μια μηχανή εκμάθησης. Τα παραδείγματα αυτά θεωρούνται πειστικά γεγονότα ενός κατηγορήματος *p*, και συνυπάρχουν με μια πρότερη γνώση *B*. Η μηχανή εκμάθησης του εμπειρικού ILP θα μάθει μια υπόθεση *H* που είναι συνεπής όσον αφορά τα*E*και*B*.
- Ο διαδραστικός ILP, που είναι πολύ κοντινός με την έννοια του εμπειρικού ILP. Ωστόσο, σε αυτόν τον τύπο η μηχανή εκμάθησης δέχεται μόνο ένα παράδειγμα *e*, μια υπόθεση η οποία μπορεί να είναι και εσφαλμένη (*H*), καθώς και ένα «μάντη» που μπορεί να απαντήσει σε ερωτήσεις συμμετοχής για τα δεδομένα εισόδου. Μια νέα υπόθεση *H'* παράγεται από τη μηχανή μέσω τροποποιήσεων της αρχικής υπόθεσης .

Τα μοντέλα πρόβλεψης κινδύνου για το διαβήτη δεν έχουν ακόμη αξιοποιήσει ευρέως τον ILP, καθώς δεν υφίσταται εκτεταμένη ερευνητική δραστηριότητα που να συνδέει αυτά τα δύο πεδία.

3.2.5 Clustering

Όταν δεν υπάρχει πρότερη γνώση της φύσης των δεδομένων, η μη επιβλεπόμενη συσταδοποίηση καλύπτει της απαιτήσεις και παρέχει αποδοτικότητα. Οι εφαρμογές της είναι ευρείες και βοηθητικές στο χειρισμό πολλών προβλημάτων όπως η διάγνωση ασθενειών και η κλινική αντιμετώπιση. Για παράδειγμα, αποδεικνύεται χρήσιμη όταν πρόκειται για ταξινόμηση όμοιων υποτύπων μιας συγκεκριμένης ασθένειας, από μορφολογική σκοπιά ή ακόμη από λειτουργικά χαρακτηριστικά που είναι συνδεδεμένα με πρότυπα έκφρασης γονιδίων.

Η γενική μεθοδολογία ενός αλγορίθμου συσταδοποίησης περιλαμβάνει το διαμερισμό ορισμένων αντικειμένων δεδομένων (data objects – πρότυπα, οντότητες, παρατηρήσεις και μονάδες) σε κατηγορίες, συχνά γνωστών ως συστάδες (clusters). Ωστόσο, ο όρος συστάδα είναι αντιφατικός όσον αφορά τον ορισμό του. Ο Everitt (1980) αποπειράθηκε να συνοψίσει αυτό τον ορισμό στην εξής δήλωση: «μια συστάδα είναι ένα σύνολο οντοτήτων που είναι όμοιες, και οντότητες από διαφορετικές συστάδες είναι ανόμοιες». Η κύρια ιδέα πίσω από την τεχνική της συσταδοποίησης είναι μάλλον απλή: η περιοχή της συστάδας καθορίζεται από τη μέτρηση των αποστάσεων των σημείων δεδομένων. Η απόσταση μεταξύ οποιωνδήποτε δύο σημείων του ίδιου cluster πρέπει να είναι μικρότερη από την απόσταση μεταξύ οποιουδήποτε σημείου στο clusterκαι οποιουδήποτε σημείου που δεν ανήκει σε αυτό.

Όλοι οι προαναφερθέντες αντιφατικοί ορισμοί συμφωνούν μεταξύ τους στο γεγονός ότι σημεία δεδομένων στην ίδια συστάδα θα πρέπει να είναι όμοια μεταξύ τους, ενώ σημεία δεδομένων σε διαφορετικές συστάδες θα πρέπει να είναι ανόμοια μεταξύ τους. Γενικά, υπάρχει ένα σύνολο προτύπων εισόδου $\mathbf{X}= \{ \mathbf{x}_1 , \dots , \mathbf{x}_j, \dots , \mathbf{x}_N\}$, where $\mathbf{x}_j= (x_{j1} , x_{j2} , \dots , x_{jd}) \in R^d$, με κάθε μέτρο$x_{ji}$ να λέγεται χαρακτηριστικό [20].

Η ανάλυση συστάδων ως διαδικασία περιλαμβάνει τα εξής τέσσερα βασικά βήματα [30]:

- *Επιλογή χαρακτηριστικών ή εξαγωγή (extraction).* Η πρώτη μέθοδος επιλέγει διακεκριμένα χαρακτηριστικά μεταξύ υποψηφίων. Από την άλλη πλευρά υπάρχει η εξαγωγή χαρακτηριστικών, όπου ένας αριθμός μετασχηματισμών λαμβάνει χώρα, με σκοπό να παραχθούν νέα χαρακτηριστικά που μπορούν να αξιοποιηθούν στο εσωτερικό της μεθόδου. Η φάση της

εξαγωγής αποκαλύπτει τις ιδιαιτερότητες της δομής δεδομένων, δημιουργώντας τα προαναφερθέντα νέα χαρακτηριστικά. Ένας παράγοντας που πρέπει να ληφθεί υπόψη σε αυτό το σημείο είναι η πιθανότητα παραγωγής μη ερμηνεύσιμων χαρακτηριστικών , τα οποία δεν είναι παρόντα στη φάση της επιλογής, καθώς η τελευταία εξασφαλίζει ότι οι φυσικές πτυχές των αρχικών χαρακτηριστικών θα παραμείνουν στο σύστημα, Είναι αρκετά συνηθισμένο για αυτούς τους δύο όρους να χρησιμοποιούνται ταυτόσημα, αν και η λειτουργία τους είναι αρκετά διαφορετική, Η συνεισφορά τους είναι καίρια, καθώς βελτιώνουν την αποδοτικότητα των μεθόδων συσταδοποίησης. Αυτό σημαίνει ότι μπορούν να ελαχιστοποιήσουν το κόστος μέτρησης, και τις ανάγκες για αποθήκευση, να μειώσουν την πολυπλοκότητα της διαδικασίας σχεδίασης και να κάνουν τη σύλληψη της φύσης των δεδομένων πιο κατανοητή. Η ύπαρξη ιδανικών χαρακτηριστικών είναι υψίστης σημασία, καθώς αυτά θα χρησιμοποιηθούν σε διακεκριμένα πρότυπα των διαφορετικών συστάδων. Αυτά τα πρότυπα είναι σθεναρά σε ότι αφορά το θόρυβο, καθώς επίσης και εύκολα στην απόκτηση και την ερμηνεία.

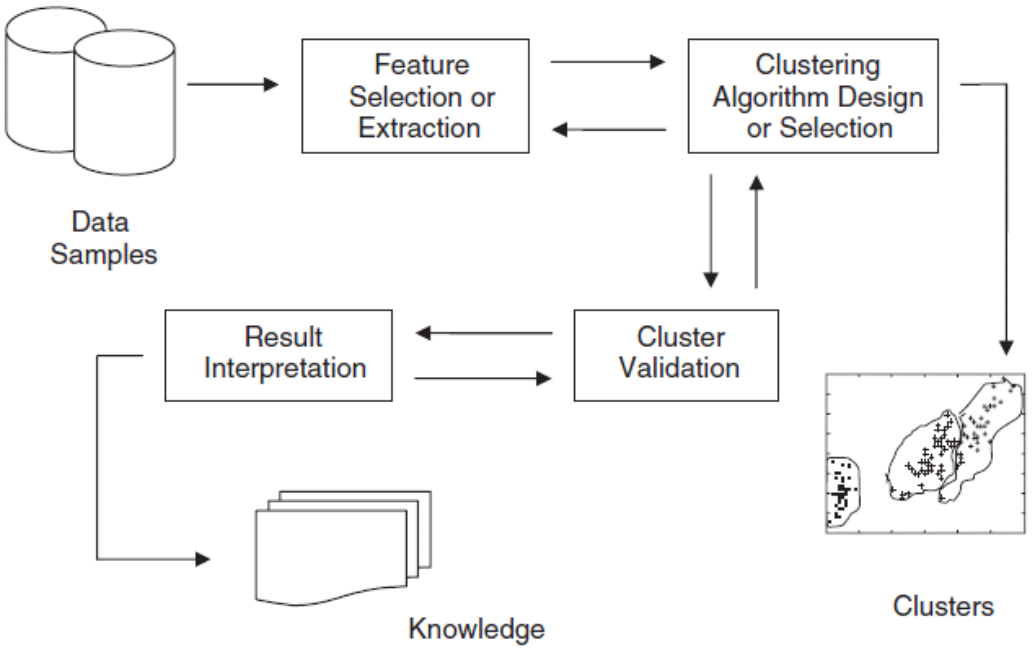
- Σχεδίαση ή επιλογή αλγορίθμου συσταδοποίησης. Αυτό το μέρος της μεθόδου περιλαμβάνει την επιλογή ενός αποδοτικού μέτρου προσέγγισης (proximity measure) και τον ορισμό μιας συνάρτησης κριτηρίου. Όπως αναφέρθηκε πριν, η ομαδοποίηση των δεδομένων σε συστάδες εξαρτάται από την ομοιότητα συγκεκριμένων σημείων δεδομένων. Για το λόγο αυτό, ο ορισμός ενός μέτρου προσέγγισης (όπως ο πίνακας προσέγγισης) απαιτείται στις περισσότερες αλγοριθμικές διαδικασίες συσταδοποίησης. Μετά από αυτό τον ορισμό, μια συνάρτηση κριτηρίου κατασκευάζεται και επιτρέπει τη θεώρηση της συσταδοποίησης ως ενός προβλήματος βελτιστοποίησης. Η συνάρτηση κριτηρίου αποφασίζει υποκειμενικά τη διαδικασία κατασκευής των συστάδων.

Υπάρχει ένας μεγάλος αριθμός αλγορίθμων συσταδοποίησης διαθέσιμος στη βιβλιογραφία, και αντιστοιχεί στην ποικιλία των απαιτήσεων μεταξύ των διάφορων επιστημονικών πεδίων. Μια μεγάλη προσπάθεια έχει καταβληθεί για την απόπειρα κατασκευής ενός ενοποιημένου πλαισίου που θα μπορούσε να χειριστεί όλα τα προβλήματα. Ως αποτέλεσμα, είναι κρίσιμο ζήτημα η μελέτη και η κατανόηση των πτυχών του προβλήματος ώστε να επιλεγεί ο πιο αποδοτικός αλγόριθμος συσταδοποίησης. Είναι σύνηθες αυτοί οι αλγόριθμοι να προβαίνουν σε υποθέσεις από τις οποίες θα επωφελούνταν προβλήματα που προέρχονται από ένα συγκεκριμένο πεδίο. Για παράδειγμα, ο αλγόριθμος K-means έχει την τάση να δημιουργεί υπερσφαιρικές συστάδες, αφού χρησιμοποιεί την ευκλείδεια νόρμα. Ωστόσο, αυτές οι υποθέσεις δεν είναι πάντα ευνοϊκές, αφού ορισμένες πραγματικές συστάδες ενδέχεται να ακολουθούν άλλες γεωμετρικές μορφές, γεγονός

που θα καθιστούσε άμεσα τον K-meansμη αποδοτικό. Το ίδιο πρότυπο σκέψης θα πρέπει να χρησιμοποιείται και στη mixture-model συσταδοποίηση, όπου τα δεδομένα θεωρείται πως προέρχονται από ήδη γνωστά μοντέλα.

- *Αξιολόγηση συστάδων*. Είτε υπάρχει μια συγκεκριμένη δομή δεδομένων είτε όχι, μια διαδικασία συσταδοποίησης είναι πάντα ικανή να δημιουργήσει ένα διαμερισμό του συνόλου δεδομένων. Είναι επίσης σύνηθες για διαφορετικές προσεγγίσεις συσταδοποίησης να αποδίδουν διαφορετικές συστάδες, καθώς επίσης και, ακόμη και όταν χρησιμοποιείται ο ίδιος αλγόριθμος, τα τελικά αποτελέσματα να είναι πολύ διαφορετικά, αν μια κρίσιμη παράμετρος επιλέγεται με διαφορετική τιμή ή αν η σειρά παρουσίασης των εισόδων στο σύστημα αλλάζει σε διαφορετικές εκτελέσεις του αλγορίθμου.

Οι προαναφερθείσες σκέψεις αποδεικνύουν τη σημαντικότητα των κριτηρίων αξιολόγησης που ενισχύουν τη μέθοδο με βεβαιότητα για την ορθότητα των αποτελεσμάτων της. Η αξιολόγηση θα πρέπει να χαρακτηρίζεται από αντικειμενικότητα, δε θα πρέπει δηλαδή να κάνει διακρίσεις προς όφελος ή σε βάρος ορισμένων αλγορίθμων. Ακόμη, θα πρέπει να διευκολύνει την κατανόηση των πτυχών της μεθόδου, παρέχοντας πληροφορίες σχετικά με τον αριθμό των κρυμμένων συστάδων στα δεδομένα ή ακόμη σχετικά με το αν οι συστάδες που παρήχθησαν έχουν κάποιο νόημα για τη μέθοδο. Επιπροσθέτως, θα πρέπει να απαντά σε ερωτήσεις όπως γιατί διαλέγουμε έναν αλγόριθμο αντί για κάποιον άλλο. Τα κριτήρια αξιολόγησης μπορούν να ανατεθούν σε τρεις διαφορετικές κατηγορίες: εξωτερικοί δείκτες, εσωτερικοί δείκτες και σχετικοί δείκτες, και ορίζονται στη βάση τριών δομικών τύπων συσταδοποίησης: διαιρετική συσταδοποίηση, ιεραρχική συσταδοποίηση και ατομικές συστάδες. Κάποιες φορές διενεργούνται τεστ που ελέγχουν την ύπαρξη μια δομής συσταδοποίησης στα δεδομένα αλλά η χρήση τους είναι σπάνια καθώς οι επιστήμονες είναι συνήθως βέβαιοι για την ύπαρξη των συστάδων. Πιο συγκεκριμένα, οι εξωτερικοί δείκτες λειτουργούν βασισμένοι σε μια προκαθορισμένη δομή, που προέρχεται από πρότερη γνώση των δεδομένων και αποδεικνύεται χρήσιμη για την αξιολόγηση των λύσεων που παρέχει η συσταδοποίηση. Από την άλλη πλευρά, οι εσωτερικοί δείκτες δεν εξαρτώνται από την πρότερη γνώση, καθώς εξερευνούν τη δομή της συσταδοποίησης από το αρχικό σύνολο δεδομένων. Τέλος, οι σχετικοί δείκτες συγκρίνουν διαφορετικές δομές συσταδοποίησης, προκειμένου να αποφασίσουν ποια δομή θα αποκάλυπτε τις ιδιαιτερότητες των αντικειμένων με τον πιο αποδοτικό τρόπο.



Σχήμα 3.4 Ένα γενικό σχηματικό διάγραμμα Συσταδοποίησης

- *Ερμηνεία Αποτελεσμάτων*. Το τελικό βήμα μιας μεθόδου συσταδοποίησης εξυπηρετεί τον τελικό σκοπό της συσταδοποίησης, που είναι να επιτρέψει στους χρήστες να αποκτήσουν ολοκληρωμένη άποψη για τα δεδομένα, ώστε να επιτύχουν ένα επαρκές επίπεδο κατανόησής τους, γεγονός που επιτρέπει την επίλυση σχετικών προβλημάτων, Σύμφωνα με τον Anderberg (1973), η ανάλυση συστάδων είναι μία «συσκευή παραγωγής προτάσεων για υποθέσεις». Το Σχήμα 2.4 περιλαμβάνει επίσης έναν κλάδο ανάδρασης, που υποδεικνύει ότι η ανάλυση συστάδων δεν είναι διαδικασία μίας προσπάθειας. Ανάλογα με τη φύση των δεδομένων, αρκετές προσπάθειες μπορεί να είναι απαραίτητες, και ύστερα από καθεμιά από αυτές τα αποτελέσματα ανατροφοδοτούνται στην είσοδο της διαδικασίας για περαιτέρω επεξεργασία.

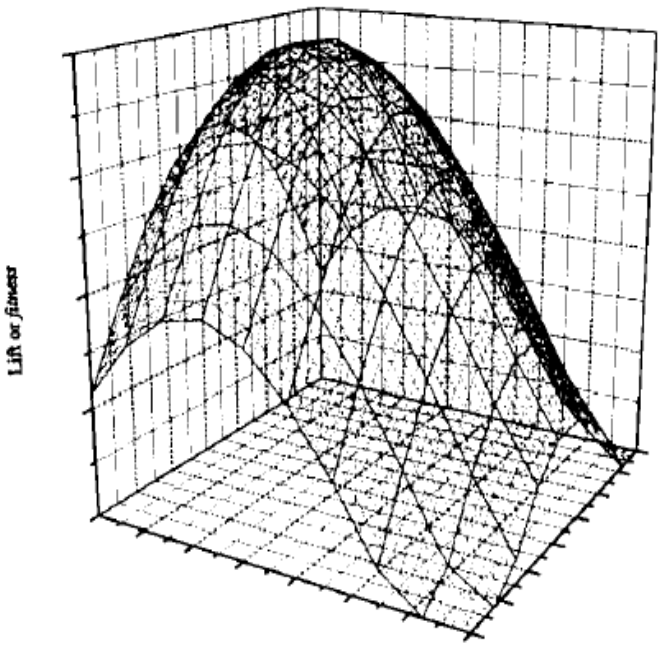
Στο [31] μπορούμε να δούμε ότι η υπερινσουλιναιμία και η συσταδοποίηση των καρδιαγγειακών παραγόντων κινδύνου με ινσουλινοαντίσταση συνδέονται με το θάνατο από στεφανιαία νόσο σε ασθενείς με ΣΔΤ2. Η εφαρμογή ανάλυσης παραγόντων και Ανάλυσης Κύριων Συνιστωσών είχε ως καινοτόμο αποτέλεσμα την απόδειξη ότι η συστάδα της υπερινσουλιναιμίας είναι ένας παράγοντας ικανός να προβλέψει το θάνατο από στεφανιαία νόσο σε ΣΔΤ2 ασθενείς, ενισχύοντας της πεποίθηση ότι η

συσταδοποίηση μπορεί να αποδειχθεί χρήσιμη στην κατεύθυνση της ανάλυσης πρόβλεψης κινδύνου για το διαβήτη.

3.2.6 Γενετικοί Αλγόριθμοι

Οι Γενετικοί Αλγόριθμοι (Genetic algorithms - GAS) είναι μια οικογένεια μεθόδων μηχανικής μάθησης που αποτελείται από αλγορίθμους αριθμητικής βελτιστοποίησης, οι οποίοι προέρχονται από φυσική επιλογή και από τη γενετική. Ένα μεγάλο πλεονέκτημα των ΓΑ είναι η ευελιξία, που τους επιτρέπει να αντιμετωπίζουν μεγάλο εύρος προβλημάτων, πολλά από τα οποία αντιστοιχούν σε πρακτικές εφαρμογές στον πραγματικό κόσμο. Ένα ακόμη θετικό χαρακτηριστικό τους είναι το γεγονός ότι είναι πολύ κατανοητοί και ότι μετασχηματίζονται εύκολα σε κώδικες σχετικά μικρής πολυπλοκότητας. Ωστόσο, δεν είναι η πρώτη επιλογή μεταξύ των μεθόδων μηχανικής μάθησης, καθώς επισκιάζονται συχνά από τα νευρωνικά δίκτυα που είναι πιο δημοφιλή, αν και ο ουσιώδης λόγος για τον οποίο συμβαίνει αυτό δεν είναι ακόμη σαφής.

Η προσέγγιση αυτή περιλαμβάνει την αναζήτηση ενός συνόλου πιθανών λύσεων, που στοχεύουν να ανακαλύψουν της λύση που περιγράφει το πρόβλημα με βέλτιστο τρόπο. Για παράδειγμα, όπως βλέπουμε στο Σχήμα 2.5, κάποιος μπορεί να επιθυμεί να βρει τις βέλτιστες τιμές παραμέτρων για να μεγιστοποιήσει το μέγεθος της ανύψωσης του φτερού ενός αεροπλάνου. Αν υποθέσουμε πως μόνο δύο παράμετροι, a και b, απαιτούν ρύθμιση, τότε πολλοί συνδυασμοί μπορούν να διερευνηθούν και να παραχθεί η απεικόνιση μιας επιφάνειας, στην οποία τα a,b και η ανύψωση βρίσκονται στους άξονες x, yκαι z, αντίστοιχα [32].



Σχήμα3.5 Η ανύψωση του φτερού ενός αεροπλάνου ως επιφάνεια εξαρτώμενη από δύο παραμέτρους

Οι ΓΑ συνήθως ξεκινούν με έναν αριθμό τυχαίων εικασιών, που είναι διασκορπισμένες στο χώρο αναζήτησης. Τρεις σημαντικοί τελεστές, η επιλογή (selection), η διασταύρωση (crossover) και η μετάλλαξη (mutation) εφαρμόζονται για να οδηγήσουν αυτές τις εικασίες μέσα από μια σειρά γενεών στο γενικό βέλτιστο. Η πιο συχνή περίπτωση περιλαμβάνει μόνο δυαδικές αναπαραστάσεις των αρχικών εικασιών, αν και γίνεται όλο και πιο συχνή η τάση των ΓΑ να τις κωδικοποιούν σε πραγματικές τιμές. Μετά το πέρας της κωδικοποίησης, εφαρμόζονται οι προαναφερθέντες τελεστές [32]:

- Η επιλογή είναι ένας τελεστής που διαχειρίζεται τις εικασίες με έναν τρόπο όμοιο με αυτόν που εφαρμόζεται από τη φυσική επιλογή στα φυσιολογικά συστήματα. Λειτουργεί σαν κριτής που απορρίπτει άτομα που αποδίδουν λίγο και επιτρέπει την είσοδο σε άτομο που αποδίδουν περισσότερο, παρέχοντάς του επίσης μια μέση πιθανότητα διέλευσης της πληροφορίας τους στην επόμενη γενιά.

- Η διασταύρωση επιτρέπει την ανταλλαγή πληροφορίας μεταξύ ατόμων με τρόπο όμοιο με αυτόν που χρησιμοποιούν οι βιολογικοί οργανισμοί κατά τη διάρκεια της αναπαραγωγής. Μια ευρέως χρησιμοποιούμενη μέθοδος διασταύρωσης, η διασταύρωση ενός σημείου (single-point crossover) περιλαμβάνει την επιλογή ζευγών ατόμων που πέρασαν από τον τελεστή της επιλογής, και την αυθαίρετη επιλογή ενός σημείου μέσα στις δυαδικές ακολουθίες. Μετά το πέρας της επιλογής,

όλη η κωδικοποιημένη πληροφορία αντιμετατίθεται στα δεξιά αυτού του σημείου μεταξύ των δύο ατόμων.

- Η μετάλλαξη διαφέρει ουσιωδώς από τη διασταύρωση, αφού αντί για αντιμετάθεση αδιάσπαστης πληροφορίας μεταξύ των ατόμων, περιλαμβάνει τυχαίες αλλαγές της τιμής συγκεκριμένων bitsμέσα στις ακολουθίες. Η μετάλλαξη γενικά δε χρησιμοποιείται συχνά.

Όταν και οι τρεις τελεστές έχουν εφαρμοστεί στις εικασίες, μια νέα γενιά του αρχικού πληθυσμού έχει προκύψει. Το ίδιο πρότυπο ενεργειών συνεχίζεται, μέχρι ένα ή περισσότερα κριτήρια τερματισμού να εκπληρωθεί, όπως η συμπλήρωση συγκεκριμένου αριθμού γενεών ή να έχει λάβει χώρα μια έννοια σύγκλισης [32].

Οι γενετικοί αλγόριθμοι έχουν κυρίως χρησιμοποιηθεί για ταξινόμηση καρδιοπαθειών, όπως στο [33], όπου ένας αλγόριθμος συνδυάζει τη μέθοδο εκμάθησης των K-κοντινότερων γειτόνων με ένα γενετικό αλγόριθμο για αποδοτική ταξινόμηση. Το μέρος που αναφέρεται σε γενετικούς αλγορίθμους πραγματοποιεί γενική αναζήτηση σε μεγάλους και σύνθετους χώρους και παρέχει μια βέλτιστη λύση. Επιπλέον, τα αποτελέσματα βελτιώνουν την ακρίβεια στη διάγνωση των καρδιοπαθειών, συνεισφέροντας έτσι και στην επιστημονική περιοχή της πρόβλεψης κινδύνου. Η δημοσίευση δίνει ακόμη έμφαση στη σύνδεση μεταξύ διαβήτη και καρδιοπαθειών.

3.2.7 Άλλες Μέθοδοι

Η οικογένεια των μεθόδων μηχανικής μάθησης περιλαμβάνει ακόμη [20]:

- τα Bayesian δίκτυα. Πρόκειται για πιθανοτικά μοντέλα γραφημάτων που απεικονίζουν τις εξαρτήσεις μεταξύ τυχαίων μεταβλητών μέσω ακυκλικών γράφων. Σε θέματα σχετικά με την υγεία, οι εξαρτήσεις αυτές ενδέχεται να αναφέρονται στη συμπτωματολογία και τη συχνότητα εμφάνισης των ασθενειών. Αυτό σημαίνει ότι η πιθανότητα της ανάπτυξης μιας ασθένειας σε ένα άτομο θα μπορούσε να προβλεφθεί με αυξημένη ακρίβεια με την αξιοποίηση τέτοιων δικτύων.

- Η Εκμάθηση Ενίσχυσης (Reinforcement learning) είναι βασισμένη σε εκμάθηση με τη χρήση πρακτόρων. Οι έννοιες του πράκτορα και του περιβάλλοντός του είναι καλά ορισμένες, και τον ορισμό τους ακολουθεί μια σειρά ενεργειών, που στοχεύουν στη μεγιστοποίηση ενός μέτρου

επιβράβευσης. Η μέθοδος αυτή περιλαμβάνει στρατηγικές μεθόδους που αντιστοιχίζουν καταστάσεις σε προτεινόμενες ενέργειες για τον πράκτορα.

- Η Εκμάθηση Εκπροσώπησης (Representation Learning) αναφέρεται σε μια οικογένεια αλγορίθμων εκμάθησης χωρίς επίβλεψη, και επιχειρεί να βελτιώσει την αντιστοίχιση των εισόδων κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Η ανάλυση κύριων συνιστωσών και η ανάλυση συστάδων είναι μέθοδοι που ανήκουν σε αυτή την οικογένεια μεθόδων. Είναι πολύ συνηθισμένο οι αλγόριθμοι να προστατεύουν την πληροφορία που προέρχεται από τις εισόδους του συστήματος, ενώ την ίδια στιγμή προσπαθούν να την κάνουν πιο χρηστική μετατρέποντας τη μορφή της. (προεπεξεργασία δεδομένων). Αυτή η τεχνική βελτιώνει πολύ συχνά τη διαδικασία ταξινόμησης και ανάλυσης παλινδρόμησης που ακολουθεί.

- Μετρική Εκμάθηση και Εκμάθηση Ομοιότητας: σε αυτή την περίπτωση, το σύστημα τροφοδοτείται με δεδομένα στη μορφή ζευγών. Κάποια από τα ζεύγη έχουν παρόμοιες ιδιότητες ενώ άλλα θεωρούνται ανόμοια. Μετά την είσοδό τους σε μια μηχανή, μια συνάρτηση ομοιότητας ορίζεται και είναι ικανή να αναγνωρίζει τις ομοιότητες και τις διαφορές σε νέα αντικείμενα. Η μέθοδος αυτή χρησιμοποιείται συχνά από την προηγούμενη μέθοδο, την Εκμάθηση Εκπροσώπησης.

- Sparse dictionary μάθηση: περιλαμβάνει την αναπαράσταση δεδομένων ως γραμμικών συνδυασμών συναρτήσεων βάσης. Οι συντελεστές του γραμμικού συνδυασμού θεωρούνται αραιοί (sparse). Αν το x είναι ένα d-διάστατο δεδομένο, ο D θα είναι ένας d επί n πίνακας, όπου οι στήλες του D είναι συναρτήσεις βάσης. Οι συντελεστές συμβολίζονται με το r. Όταν η μέθοδος χρησιμοποιείται για ταξινόμηση, το κύριο πρόβλημα είναι να αποφασιστεί σε ποιες κλάσεις ανήκουν τα νέα δεδομένα. Ένας αναλυτής δεδομένων πρέπει πάντα να έχει υπόψη ότι τα προβλήματα που επιχειρεί να λύσει η μέθοδος είναι ισχυρά NP-hard και οι λύσεις τους προσεγγίζονται εύκολα.

3.3 Συλλογική Μάθηση

Η Συλλογική Μάθηση (ensemble learning) προέρχεται από μια σχετικά πρόσφατη τάση στο πεδίο της μηχανικής μάθησης, που χαρακτηρίζεται από τη λήψη αποφάσεων βασιζόμενη σε πολλές οντότητες. Η ιδέα της συλλογικής μάθησης απέκτησε απήχηση που προήλθε από την τάση της να περιορίζει τη μεταβλητότητα των ταξινομητών, βελτιώνοντας παράλληλα τη σθεναρότητα και της ακρίβεια των συστημάτων την ίδια στιγμή. Οι πιο συγκροτημένες πτυχές αυτής της έννοιας αναπτύχθηκαν πρόσφατα μέσω τεχνικών όπως το boostingκαι τα Τυχαία Δάση (Random Forests). Οι εφαρμογές της είναι ποικίλες, και περιλαμβάνουν την κατάτμηση εικόνων, την ανίχνευση αντικειμένων, την αναγνώριση προτύπων, την εξόρυξη δεδομένων, τη βιοπληροφορική και πολλές ακόμη.

Το ενδιαφέρον που έχουν τραβήξει τα συλλογικά συστήματα οφείλεται στη βελτιωμένη τους αποδοτικότατα και στην αξιοσημείωτη ευελιξία που επιδεικνύουν (οι εφαρμογές τους σχετίζονται με προβλήματα ευρέος φάσματος, και τα προβλήματα αυτά αντιστοιχούν σε απαιτήσεις του πραγματικού κόσμου). Για το λόγο αυτό ο αρχικός τους σκοπός της βελτίωσης ακρίβειας έχει επεκταθεί στο σημείο να θεωρούνται μια από τις σημαντικότερες προσεγγίσεις για προβλήματα όπως η επιλογή χαρακτηριστικών, η διόρθωση σφαλμάτων και ο χειρισμός προβλημάτων με ανισόρροπα δυαδικά δεδομένα. Ένα ακόμη καίριο πλεονέκτημά τους είναι η ικανότητα που έχουν να ενισχύουν την πεποίθηση ότι οι σωστές αποφάσεις έχουν ληφθεί, κάνοντας τα συστήματα και τους σχεδιαστές πιο ισχυρούς. Το σκοπό αυτό πετυχαίνουν θεωρώντας πολλαπλές επιλογές και αξιοποιώντας τες για να αποφασίσουν πιο αποδοτικές λειτουργικές διαδικασίες [34].

Η Συλλογική Μάθηση βασίζεται σε τρεις πυλώνες[34]:

- *Δειγματοληψία/Επιλογή Δεδομένων*. Η παρουσία σφαλμάτων είναι ένας σημαντικός παράγοντας που επηρεάζει την απόδοση τον συλλογικών συστημάτων. Μια συνηθισμένη μορφή σφαλμάτων στα συστήματα αυτά είναι η τάση για παραγωγή ίδιων εξόδων, δηλαδή η έλλειψη ποικιλομορφίας. Η ιδανική περίπτωση θα περιελάμβανε ανεξάρτητα ή ακόμη καλύτερα, αρνητικά συσχετισμένες εξόδους. Ένας αριθμός τεχνικών έχει αναπτυχθεί ώστε να επιτευχθεί ποικιλομορφία, ωστόσο η πιο συνηθισμένη μέθοδος είναι η χρήση διαφορετικών υποσυνόλων των δεδομένων εκπαίδευσης για τους ταξινομητές. Η διαδικασία δειγματοληψίας και τα χαρακτηριστικά της επηρεάζουν την αλγοριθμική διαδικασία και τα αποτελέσματα του συστήματος. Πολλές ευρέως γνωστές συλλογικές μέθοδοι διαφέρουν ουσιωδώς στη φάση της δειγματοληψίας. Για παράδειγμα, το bootstrapping (η δειγματοληψία δεδομένων με παραγωγή

διαφορετικών συνόλων χρησιμοποιώντας τυχαία επιλογή με αντικατάσταση)των δεδομένων εκπαίδευσης είναι μια πολύ χαρακτηριστική επεξεργασία που ανήκεις στη μέθοδο bagging (bootstrap aggregating – συσσώρευσης των bootstraps), ενώ η δειγματοληψία από μια κατανομή που δίνει προβάδισμα σε δείγματα που ταξινομήθηκαν εσφαλμένα προηγουμένως είναι χαρακτηριστική συμπεριφορά του boosting. Μια άλλη οικογένεια προσεγγίσεων της δειγματοληψίας, γνωστή και ως μέθοδοι τυχαίου υποχώρου, αναθέτουν διαφορετικά υποσύνολα χαρακτηριστικών σε κάθε ταξινομητή. Οι τόσο πολλές μέθοδοι δειγματοληψίας συνοδεύονται από έναν αριθμό μέτρων ποικιλομορφίας, για τα οποία θα μπορούσε να βρει κανίς μεγάλο όγκο επιστημονικής έρευνας, παρόλο που μια σύνδεση μεταξύ ποικιλομορφίας και ακρίβειας δεν έχει οριστεί επισήμως.

- *Εκπαίδευση ατομικών ταξινομητών*. Η διαδικασία που βρίσκεται στον πυρήνα κάθε συλλογικής μεθόδου είναι η εκπαίδευση των ατομικών ταξινομητών Ο τρόπος που οι ταξινομητές αυτοί λειτουργούν συνεργατικά υποδεικνύεται από έναν αλγόριθμο ανταγωνισμού που παραλαμβάνει τα ατομικά αποτελέσματα και παράγει μια γενική έξοδο, που είναι και η έξοδος του συστήματος. Κάποιοι από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους είναι το bagging(και άλλοι αλγόριθμοι που προέρχονται από αυτό, όπως ο arc-x4 και τα τυχαία δάση), το boosting και η γενίκευση στοίβας.

- *Συνδυασμός ταξινομητών*. Η προαναφερθείσα παραγωγή της εξόδου ακολουθεί μια στρατηγική που επιλέγεται θεωρώντας τη φύση των μελών του συλλογικού μηχανισμού, δηλαδή των ατομικών ταξινομητών του συστήματος. Για παράδειγμα, υπάρχουν οι μηχανές διανυσμάτων υποστήριξης που παράγουν μόνο διακριτές τιμές εξόδου. Συνεπώς, η έξοδος του συλλογικού μηχανισμού υπολογίζεται από μια απλή ή σταθμική πλειοψηφία στο πρώτο βήμα και από το μέτρο Borda στο δεύτερο βήμα. Άλλα παραδείγματα, όπως το πολυεπίπεδο perceptron παράγουν συνεχείς τιμές εξόδου που είναι συγκεκριμένες για κάθε κλάση, και οι οποίες αναφέρονται στο βαθμό που η μέθοδος ευνοεί την κάθε κλάση. Πέρα από όσα αναφέρθηκαν ως τώρα, υπάρχουν πολύ περισσότερες επιλογές που μπορούν να αξιοποιηθούν στο συνδυασμό των ταξινομητών, όπως οι γραμμικοί συνδυαστές (π.χ. μέση τιμή), και πιο σύνθετες διαδικασίες απόφασης που συμπληρώνουν την καθιερωμένη μέθοδο πλειοψηφίας. Κάποιος θα μπορούσε επίσης να διακρίνει τους συνδυαστές όσον αφορά τις απαιτήσεις τους σε πόρους: κάποιοι από αυτούς μπορούν να υπολογιστούν άμεσα (όπως η πλειοψηφία και η μέση τιμή), ενώ άλλοι ενδέχεται να χρειάζονται μια πιο σύνθετη διαδικασία, που περιλαμβάνει ένα περαιτέρω βήμα εκπαίδευσης (όπως η μέθοδος της γενίκευσης στοίβας).

3.3.1 Βέλτιστος ταξινομητής Bayes

Ένας απλός τρόπος ταξινόμησης ενός δείγματος **x** είναι ο υπολογισμός της ύστερης πιθανότητας P(y|**x**) διαφορετικών εξόδων που σημειώνονται ως y (και εκπροσωπούν κατηγορίες), χρησιμοποιώντας ένα πιθανοτικό μοντέλο, και υπολογίζουν τη μία με τη μεγαλύτερη ύστερη πιθανότητα. Το θεώρημα του Bayes αποδεικνύεται χρήσιμο εδώ:

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$$

Όπου η P(y) μπορεί να υπολογιστεί ως η αναλογία της ύπαρξης της κλάσης y στο σύνολο εκπαίδευσης, ενώ η P(**x**) μπορεί να αγνοηθεί, αφού η σύγκριση των y λαμβάνει χώρα δεδομένου του ίδιου (συνεπώς, είναι κοινός παρονομαστής για κάθε δείγμα). Άρα, η μόνη ποσότητα που πρέπει να υπολογιστεί είναι η P(**x**|y). Αν η εκτίμηση είναι ακριβής, ο πιο αποδοτικός ταξινομητής είναι το αποτέλεσμα της διαδικασίας δεδομένων των δεδομένων εκπαίδευσης. Ο ταξινομητής αυτός ονομάζεται ταξινομητής του Bayes και έχει το μικρότερο ρυθμό σφάλματος, γνωστό και ως ρυθμό Bayes.

Κάποιος μπορεί να παρατηρήσει ότι ο Βέλτιστος Ταξινομητής Bayes είναι στην πραγματικότητα ένας συλλογικός μηχανισμός όλων των υποθέσεων στο χώρο υποθέσεων. Σε καθεμιά από αυτές τις υποθέσεις ανατίθεται μια ψήφος που είναι ανάλογη της πιθανότητας το δεδομένο σύνολο εκπαίδευσης να υφίσταται δειγματοληψία από ένα σύστημα αν η υπόθεση αυτή ήταν ορθή. Προκειμένου ένας αναλυτής δεδομένων να είναι ικανός να χειρίζεται σύνολα δεδομένων εκπαίδευσης πεπερασμένης πληθικότητας, η ψήφος κάθε υπόθεσης πολλαπλασιάζεται επίσης με την πρότερη πιθανότητα αυτής της υπόθεσης. Συνεπώς, ο Βέλτιστος Ταξινομητής Bayes δίνεται από την επόμενη εξίσωση:

$$y = argmax_{c_j \epsilon C} \sum_{h_i \epsilon H} P(c_j|h_i) P(T|h_i) P(h_i),$$

Όπου y είναι η προβλεπόμενη κλάση που αναφέραμε προηγουμένως, C είναι το σύνολο όλων των πιθανών κλάσεων, Heίναι ο χώρος της υπόθεσης και T είναι το σύνολο εκπαίδευσης. Η υπόθεση του Βέλτιστου Ταξινομητή Bayes δεν ανήκει πάντα στον H, αν και αποδεικνύεται ότι είναι η βέλτιστη

υπόθεση που γίνεται στο εσωτερικού του χώρου συλλογικής μάθησης(του χώρου όλων των πιθανών συλλογικών μηχανισμώνσπου αποτελούνται μόνο από υποθέσεις στον H) [35].

Ωστόσο, ο Βέλτιστος Ταξινομητής Bayes έχει ένα μεγάλο μειονέκτημα: εν γένει μπορεί μόνο να εφαρμοστεί στα πιο απλά προβλήματα, αφού [35]:

- Οι περισσότεροι χώροι υπόθεσης είναι τόσο ευρείς που δεν είναι πρακτικό να παραχθούν επαναλήψεις.
- Η φύση κάποιων υποθέσεων οδηγεί στην παραγωγή μόνο μιας προβλεπόμενης κλάσης, ενώ απαιτείται πρακτικά η πιθανότητα κάθε κλάσης.
- Είναι πολύ δύσκολο να υπολογίσει κανείς την πρότερη πιθανότητα του συνόλου δεδομένων χωρίς bias.
- Η σωστή εκτίμηση της πρότερης πιθανότητας κάθε υπόθεσης του χώρου υποθέσεων είναι πολύ σπάνια.

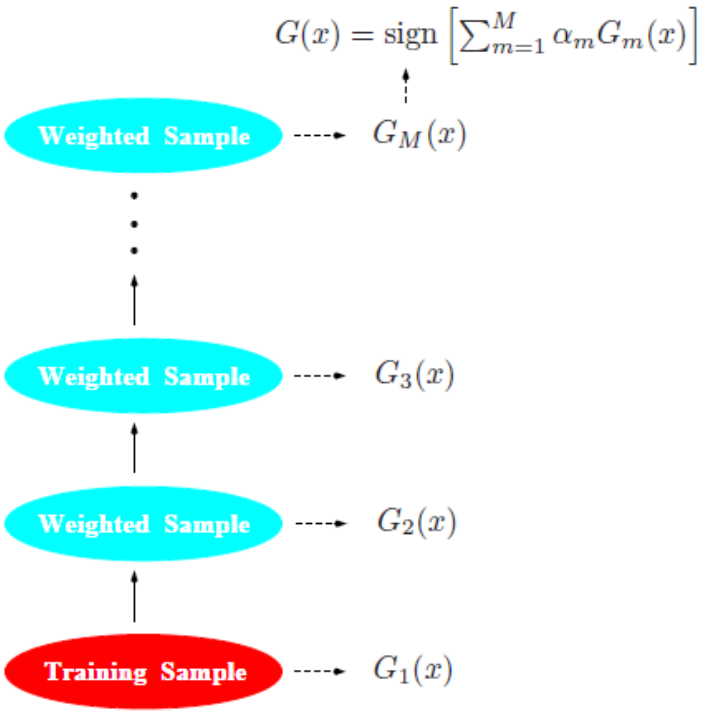
3.3.2 Boosting

Το Boosting ήταν μια μέθοδος που άλλαξε τη θεώρηση της ensemble μάθησης από τη στιγμή που παρουσιάστηκε για πρώτη φορά, όντας ένα εργαλείο υψηλής απόδοσης, γεγονός που εξηγείται από την ιδιαιτερότητα της λειτουργίας του: συνδυάζει τα αποτελέσματα των ασθενών ταξινομητών με σκοπό να δημιουργήσει μια αποδοτική «επιτροπή». Παρόλο που αρχικά σκόπευε να χειρίζεται μόνο προβλήματα ταξινόμησης, επεκτείνεται και στην ανάλυση παλινδρόμησης διατηρώντας τα χαρακτηριστικά της απόδοσής του Προκειμένου να διευκολύνουμε την κατανόηση της συνεισφοράς του boostingστην ensembleμάθηση –και γενικότερα στη μηχανική μάθηση- θα περιγράψουμε τον πιο δημοφιλή αλγόριθμο boosting, τον AdaBoost.M1 [36].

Δεδομένου ενός προβλήματος δύο κλάσεων, με τη μεταβλητή εξόδου να παίρνει τιμές στο$Y \in \{-1,1\}$, και ένα διάνυσμα χαρακτηριστικών X, ένας ταξινομητής G(X) παράγει μια πρόβλεψη που παίρνει μία από τις δύο τιμές τουΥκαι έχει ρυθμό σφάλματος ίσο με:

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq G(x_i))$$

Ενώ το αναμενόμενο σφάλμα για τις μελλοντικές προβλέψεις είναι $E_{XY}I(Y \neq G(X))$. Αν ένας ταξινομητής έχει ρυθμό σφάλματος που δεν είναι ουσιωδώς καλύτερος από την εικασία, τότε θεωρείται αδύναμος. Το Boosting είναι μια μέθοδος που εκτελεί έναν ιδιαίτερο αλγόριθμο ταξινόμησης σε επαναλαμβανόμενα τροποποιημένες εκδοχές των δεδομένων, δημιουργώντας έτσι έναν αριθμό αδύναμων ταξινομητών$G_m(x), m = 1,2 \dots, M$.



Σχήμα 3.6Η μέθοδος AdaBoost ως διάγραμμα ροής

Μετά από την κατασκευή των Μ αδύναμων ταξινομητών, οι προβλέψεις συγκεντρώνονται και χρησιμοποιούνται σαν είσοδοι σε έναν αθροιστή πλειοψηφίας, παράγοντας την τελική πρόβλεψη:

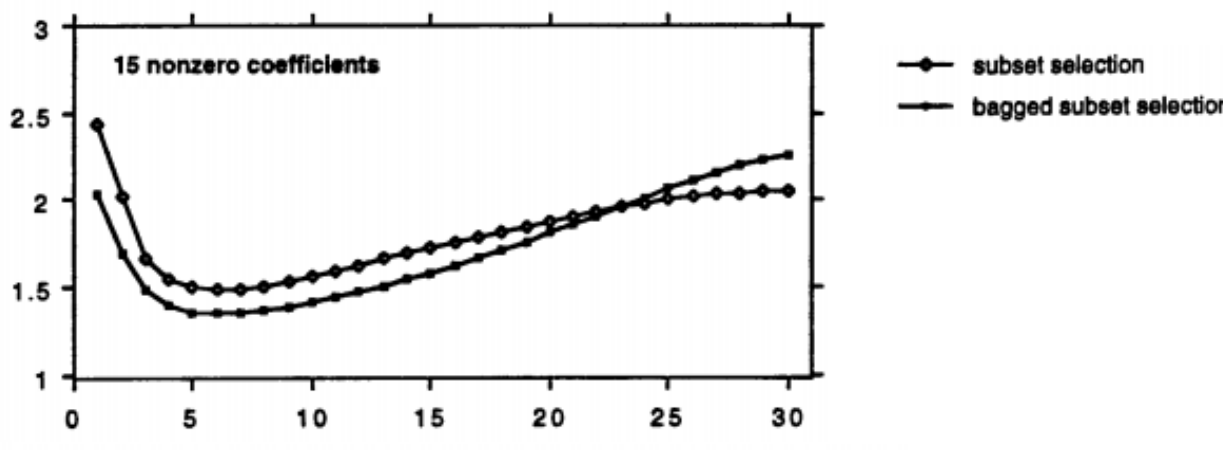
$$G(x) = sign(\sum_{m=1}^{M} a_m G_m(x))$$

Σε αυτό το σημείο, τα $a_1, a_2, \dots, a_m$ είναι βάρη που υπολογίζονται από τον αλγόριθμο. Συγκεκριμένα, υπολογίζουν σταθμικά τη συνεισφορά κάθε $G_m(x)$. Ο σκοπός είναι να ενισχύσουν τους πιο ακριβείς ταξινομητές της ακολουθίας.

Το Σχήμα 3.1 απεικονίζει τη διαδικασία AdaBoost. Οι μετατροπές των δεδομένων σε κάθε βήμα του boostingπεριλαμβάνουν την εφαρμογή των βαρών $w_1, w_2, \dots, w_N$, στα ζεύγη εκπαίδευσης $(x_i, y_i), i = 1,2, \dots, N$. Κάθε βάρος δίνεται αρχικά από την τιμή $w_i = \frac{1}{N}$. Για τις εναπομένουσες M-1 επαναλήψεις τα βάρη θα αλλάζουν και ο αλγόριθμος θα επαναλαμβάνεται λαμβάνοντας υπόψη τις νέες τιμές βαρών. Ένα βήμα της διαδικασίας, m, περιλαμβάνει μια αύξηση των βαρών των παρατηρήσεων που ταξινομήθηκαν εσφαλμένα. Ισοδύναμα, τα βάρη των ορθά ταξινομημένων παρατηρήσεων θα μειωθούν. Αυτή η λειτουργία οδηγεί σε μια αυξημένη επιρροή των παρατηρήσεων των οποίων η ταξινόμηση δεν ήταν απλή, κάτι που αναγκάζει τον επόμενο ταξινομητή να δαπανήσει περισσότερη προσπάθεια στο χειρισμό τέτοιων παρατηρήσεων [36].

3.3.3 Συσσώρευση των Bootstraps (Bagging)

Η συσσώρευση των bootstraps, συχνά αναφερόμενη και ως bagging για συντομία είναι μια πολύ δημοφιλής μέθοδος ensemble, που εκπαιδεύει πολλές μηχανές είτε για ταξινόμηση είτε για ανάλυση παλινδρόμησης ως πρώτο βήμα, και τις συσσωρεύει με σκοπό να σχηματίσει τη γενική μηχανή που παράγει τα αποτελέσματα του μοντέλου. Για την πρώτη περίπτωση, την ταξινόμηση, το baggingπεριλαμβάνει πλειοψηφία όταν οι κλάσεις προβλέπονται από ατομικές μηχανές εκμάθησης, ενώ για τη δεύτερη περίπτωση, την παλινδρόμηση, όπου αριθμητικά αποτελέσματα είναι οι έξοδοι των μηχανών μάθησης, χρησιμοποιείται εξαγωγή μέσης τιμής των αποτελεσμάτων. Προκειμένου να παραχθούν πολλαπλές μηχανές μάθησης, το αρχικό σύνολο δεδομένων εκπαίδευσης κλωνοποιείται με bootstrapping, με αποτέλεσμα να δημιουργούνται τόσα δεδομένα εκπαίδευσης όσος και ο αριθμός των ατομικών μηχανών. Έχει αποδειχθεί ότι αυτή η διαδικασία βελτιώνει ουσιωδώς την ακρίβεια, ειδικά όταν το αρχικό σύνολο εκπαίδευσης έχει υποστεί διαταραχές. Κάποιος μπορεί να διαπιστώσει τη βελτίωση στην ακρίβεια του baggingαν κοιτάξει τις τιμές του σφάλματος πρόβλεψης στο Σχήμα 3.2, όπου επιχειρείται μια σύγκριση μεταξύ κανονικών και baggedυποσυνόλων.



Σχήμα 3.7Επιλογή Bagged υποσυνόλου έναντι κανονικής επιλογής

Περιγράφουμε σύντομα τη διαδικασία συσσώρευσης των bootstraps:

Ας είναι το αρχικό σύνολο$L = \{(y_n, \boldsymbol{x}_n), n-1, \dots, N\}$, όπου (ανάλογα με τον τύπο μάθησης) οι έξοδοι Y είναι κλάσεις ή αριθμητικές τιμές. Η διαδικασία τότε συμβολίζεται ως$\varphi(\boldsymbol{x}, \boldsymbol{L})$. Ας υποθέσουμε ότι υπάρχει ένας αριθμός συνόλων μάθησης $\{L_k\}$, καθένα από τα οποία έχει N παρατηρήσεις και διατηρεί την κατανομή του L. Ο στόχος του bagging είναι να παραγάγει μια καλύτερη μηχανή μάθησης από αυτή που θα δημιουργούνταν αν το L είχε χρησιμοποιηθεί ως σύνολο εκμάθησης.

Όταν το y είναι αριθμητικό (δηλαδή η μηχανή μάθησης θα χρησιμοποιηθεί για ανάλυση παλινδρόμησης), κάποιος θα μπορούσε απλά να αξιοποιήσει την προαναφερθείσα προσέγγιση μέσης τιμής για την απόκτηση του $\varphi(\boldsymbol{x}, L_k)$: $\varphi_A(\boldsymbol{x}) = E_{\boldsymbol{L}}\varphi(\boldsymbol{x}, L)$**,**όπου το Εείναι το σύμβολο της μέσης τιμής στο L, ενώ ο υποδείκτης A συμβολίζει τη συσσώρευση. Αν υπάρχουν J κλάσεις και το $\varphi(\boldsymbol{x}, L)$ προβλέπει μία από αυτές, την$j \in \{1, \dots, J\}$, τότε η ψήφιση θεωρείται μια αποδοτική μέθοδος πρόβλεψης της εξόδου του συνολικού συστήματος. Ο αριθμός των συνόλων μάθησης για τα οποία η γενική μηχανή απέδωσε την κλάση jείναι$N_j = |\{k; \varphi(\boldsymbol{x}, L_k) = j\}|$, και $\varphi_A(\boldsymbol{x}) = argmax_j N_j$.

Ωστόσο, είναι πιο σύνηθες να έχουμε ένα μόνο σύνολο L, χωρίς να υπάρχουν αρχικά τα bootstrap σύνολα. Συνεπώς, κάποιος θα πρέπει να δημιουργήσει αυτά τα bootstrap δείγματα, $\{L^{(B)}\}$, και επίσης να σχηματίσει το$\{\varphi(\boldsymbol{x}, L^{(B)})\}$, ώστε να προκύψει το αποτέλεσμα της διαδικασίας του bagging:

- Ταξινόμηση: το$\{\boldsymbol{\varphi}(\boldsymbol{x}, L^{(B)})\}$ψηφίζει για να σχηματιστεί το $\boldsymbol{\varphi_B(x)}$.
- Παλινδρόμηση, το$\boldsymbol{\varphi_B(x)}$ υπολογίζεται ως μέσος όρος, $\boldsymbol{\varphi_B(x) = av_B \varphi(x, L^{(B)})}$.

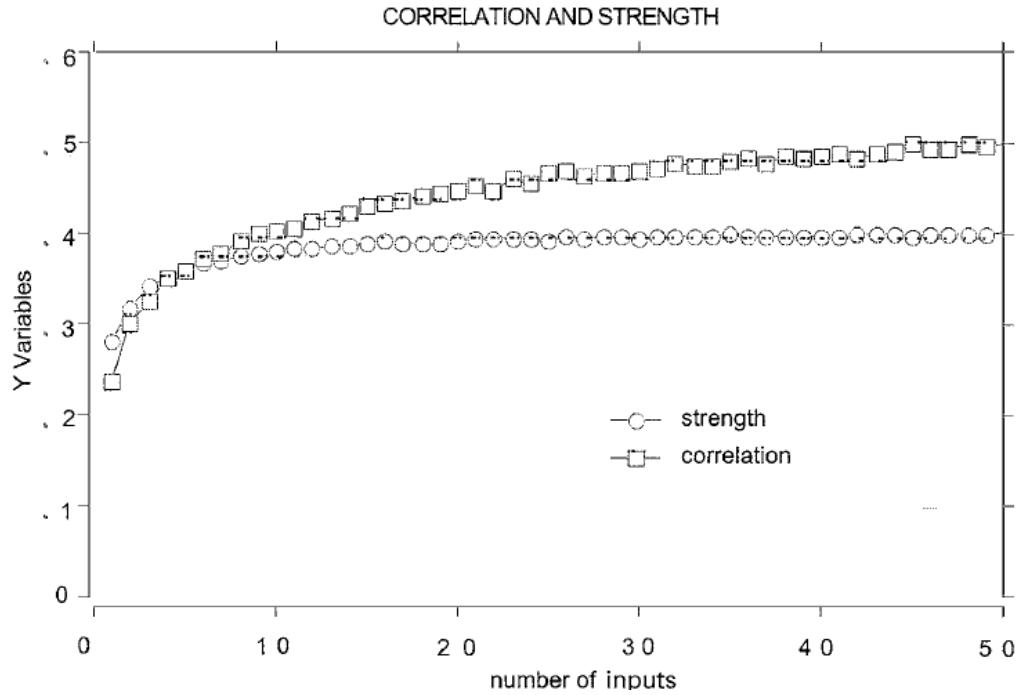
Αυτό είναι το σημείο όπου η διαδικασία που είναι γνωστή ως συσσώρευση των bootstraps ή ως bagging τελειώνει, παράγοντας τα αποτελέσματα της συσσώρευσης ενός συστήματος από ατομικές bootstrapped μηχανές μάθησης [37].

Κεφάλαιο 4. Μεθοδολογία

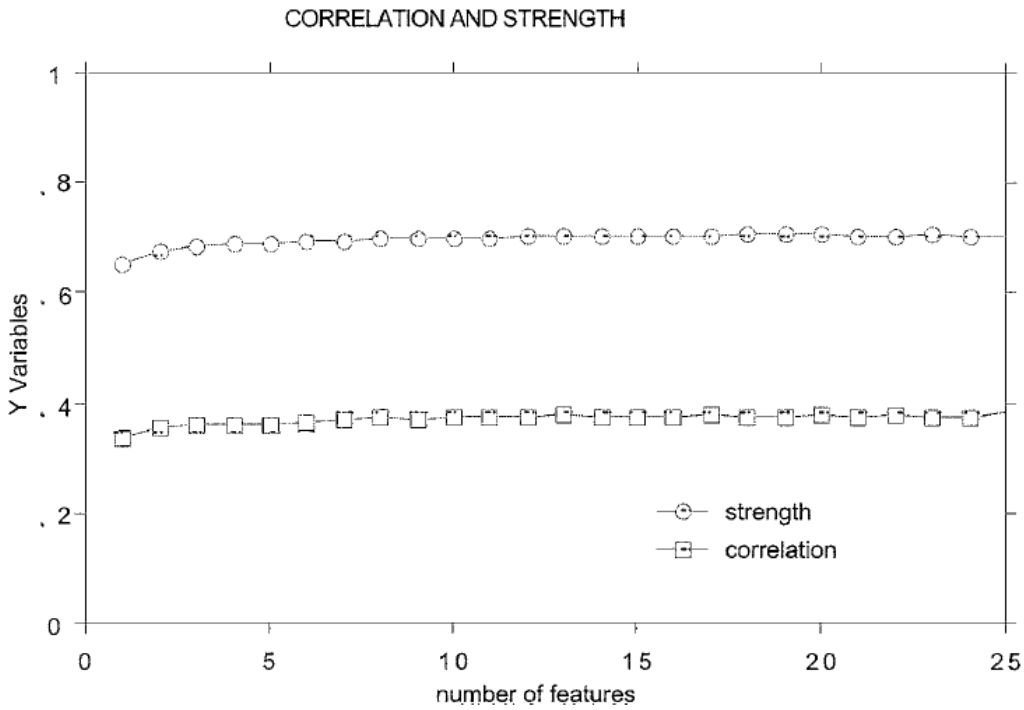
4.1 Κίνητρο

Η μέθοδος των Τυχαίων Δέντρων (Random Forests – RF) που ορίστηκε από το Leo Breiman θα μπορούσε να θεωρηθεί ένας συνδυασμός του baggingμε μεθόδους τυχαίων υποχώρων. Οι ατομικοί ταξινομητές του είναι δέντρα απόφασης. Κατά τη διαδικασία της εκπαίδευσης, επιλέγεται τυχαία ένα υποσύνολο χαρακτηριστικών για κάθε κόμβο, ώστε να λάβει χώρα ο πιο αποδοτικός διαχωρισμός του κόμβου. Η μέθοδος περιλαμβάνει μόνο δύο παραμέτρους, τον αριθμό των στοιχείων στα προαναφερθέντα υποσύνολα, που είναι ο ίδιος για όλους τους κόμβους όλων των δέντρων, ενώ η δεύτερη παράμετρος είναι ο αριθμός των δέντρων που θα παραχθούν. [38]. Στα επόμενα θα παρουσιάσουμε τη δομή των δεδομένων μας καθώς επίσης και τη γενική μεθοδολογία του συστήματός μας.

Ο κύριος λόγος για τον οποίο επιλέξαμε τη συγκεκριμένη μεθοδολογία ήταν η αποδεδειγμένη ικανότητά της να αποδίδει καλύτερα από άλλες ευρέως χρησιμοποιούμενες μεθόδους, όπως οι SVNκαι τα νευρωνικά δίκτυα. Ένα άλλο πλεονέκτημα της μεθόδου είναι η σθεναρότητά της όσον αφορά την υπερκπαίδευση (που σημαίνει ότι δεν προσκολλάται στα τρέχοντα δεδομένα ώστε να παράγει αποδοτικά αποτελέσματα μόνο για αυτά, αλλά μπορεί να γενικευτεί ώστε να μπορεί να ανακαλύπτει και να περιγράφει σχέσεις μεταξύ πληροφοριών που δεν έχουν ακόμη εισαχθεί στο σύστημα). Επιπρόσθετα, είναι πολύ κατανοητή σαν τεχνική, καθώς ακολουθεί απλά πρότυπα σκέψης, και αποτρέπει την επίπονη διαδικασία της ρύθμισης παραμέτρων, αφού περιλαμβάνει μόνο δύο, ενώ ταυτόχρονα έχει μικρή ευαισθησία ως προς τις τιμές τους. Κάποιες ενδεικτικές τιμές της ισχύος και της συσχέτισης που επιδεικνύει η μέθοδος απεικονίζονται στα Σχήματα 4.4 και 4.5, που αναφέρονται σε εμπειρικά αποτελέσματα από δορυφορικά δεδομένα και δεδομένα από σόναρ. [44]:



Σχήμα 4.1Εμπειρικά αποτελέσματα εφαρμογής των ΤΔ σε δεδομένα σόναρ. Απεικόνιση ισχύος και συσχέτισης.



Σχήμα 4.2 Εμπειρικά αποτελέσματα εφαρμογής των ΤΔ σε δορυφορικά δεδομένα. Απεικόνιση ισχύος και συσχέτισης.

Ενδιαφέρουσες είναι και οι παρατηρήσεις που προκύπτουν όταν τα ΤΔ συγκρίνονται με άλλες μεθόδους. Στον Πίνακα 4.1, μπορούμε να δούμε την υπεροχή των ΤΔ σε σχέση με τα νευρωνικά δίκτυα, το baggingτη λογιστική παλινδρόμηση και τη μέθοδο των πλησιέστερων γειτόνων, σε μελέτες που διεξήχθησαν για τη σκωληκοειδίτιδα και το διαβήτη σε δεδομένα από Ινδιάνους Pima [43].

Πίνακας 4.1 Σύγκριση των ΤΔ με άλλες μεθόδους

Table 2
Area under the ROC curves (AUC), Brier score, and nonparametric 95% bootstrap confidence intervals (in parenthesis) for the appendicitis and the Pima Indian diabetes data sets.

Machine	Appendicitis data		Pima Indian diabetes data	
	AUC	Brier score	AUC	Brier score
b-NN	0.847 (0.672 – 1.000)	0.102 (0.066 – 0.145)	0.819 (0.779 – 0.858)	0.180 (0.167 – 0.197)
classRF	0.931 (0.846 – 0.900)	0.075 (0.038 – 0.121)	0.952 (0.853 – 0.913)	0.163 (0.147 – 0.184)
lboost	0.976 (0.928 – 0.900)	0.043 (0.023 – 0.073)	0.863 (0.825 – 0.897)	0.173 (0.155 – 0.198)
logreg	0.853 (0.672 – 0.900)	0.088 (0.050 – 0.136)	0.839 (0.802 – 0.875)	0.160 (0.145 – 0.181)
k-NN	0.844 (0.694 – 0.969)	0.106 (0.066 – 0.149)	0.843 (0.777 – 0.855)	0.182 (0.168 – 0.199)
regRF	0.976 (0.934 – 0.982)	0.061 (0.037 – 0.088)	0.971 (0.862 – 0.919)	0.163 (0.151 – 0.179)

Η μεθοδολογία Δέντρων Ταξινόμησης και Παλινδρόμησης (Classification and Regression Trees - CART), που χρησιμοποιήθηκε για την κατασκευή ατομικών ταξινομητών, δηλαδή των δέντρων, συνδέεται στενά με την ΤΔ μεθοδολογία. Μια άλλη τεχνική για κατασκευή δέντρων απόφασης είναι ο αλγόριθμος ID3, που έχει χρησιμοποιηθεί σε πολλές εφαρμογές, παρόλο που τα μειονεκτήματά του τον κατέστησαν λιγότερο δημοφιλή τα τελευταία χρόνια. Το κύριο πρόβλημα με τον ID3 ήταν ότι δεν μπορούσε να λάβει υπόψη την ισχυρή παρουσία της εξάρτησης μεταξύ των διάφορων μεταβλητών, ένα γεγονός που δυσχεραίνει τη συνολική απόφαση του αλγόριθμου. Ένας άλλος περιορισμός του ID3 είναι ότι οι μεταβλητές μπορούσαν να είναι μόνο διακριτές. Ωστόσο, η συνεισφορά του στην κατασκευή απλοποιημένων μοντέλων είναι πολύ σημαντική [42].

Μεγάλες προσπάθειες έχουν καταβληθεί στην κατεύθυνση της εύρεσης τρόπων βελτίωσης της απόδοσης του προαναφερθέντος αλγορίθμου. Το αποτέλεσμα αυτών των προσπαθειών ήταν μια μέθοδος που ονομάζεται Έξυπνος Αλγόριθμος Δέντρων Απόφασης (Intelligent Decision Tree Algorithm-IDA). Η μέθοδος καθιέρωσε μια ιδέα γενικής εξάρτησης μεταξύ των διαθέσιμων μεταβλητών, καθώς επίσης και μια αλγοριθμική διαδικασία πρόβλεψης με σκοπό να επιλεγούν μόνο τα πιο χρήσιμα χαρακτηριστικά. Πέρα από αυτά τα πλεονεκτήματα έναντι του ID3, ο IDAέχει επίσης τη μισή υπολογιστική πολυπλοκότητα του πρώτου. Ωστόσο δεν ήταν ακόμη επαρκώς ικανοποιητική μέθοδος, λόγω των δικών

της μειονεκτημάτων: μόνο διακριτά σύνολο μπορούν να χρησιμοποιηθούν ως είσοδοι του IDA, και καμία τιμή δεν μπορεί να απουσιάζει από αυτά.

Η απόκριση της επιστημονικής κοινότητας ήταν η εισαγωγή ενός αλγορίθμου που δε θα επηρεαζόταν από αυτούς του περιορισμούς. Αυτός ο αλγόριθμος ονομάζεται C4.5 και παρουσιάζει αυξημένη απόδοση, ενώ μπορεί επίσης να χειριστεί τόσο διακριτές όσο και συνεχείς τιμές. Ακόμη, οι απούσες τιμές χαρακτηριστικών δε θεωρούνται πλέον πρόβλημα. Ένα ακόμη πολύ θετικό χαρακτηριστικό του C4.5 είναι η σθεναρότητά του όταν υπάρχει θόρυβος, όπως επίσης και η τάση του να αποφεύγει την υπερεκπαίδευση. Όλα αυτά τα χαρακτηριστικά τον κάνουν να εφαρμόζεται εύκολα σε πολλά προβλήματα του πραγματικού κόσμου και αυξάνουν τη δημοτικότητά του [42].

Τα ΤΔ είναι μια τεχνική συλλογικής μάθησης με πολυάριθμες εφαρμογές στη βιοπληροφορική. Η φύση των προβλημάτων που θα αναφερθούν περιλαμβάνει υψηλή πολυπλοκότητα της σχέσης μεταξύ χαρακτηριστικών και εξόδων του συστήματος, καθώς επίσης και ισχυρή συσχέτιση μεταξύ των μεταβλητών. Αυτά τα δύο χαρακτηριστικά των προβλημάτων δικαιολογούν απόλυτα τη δημοτικότητα των ΤΔ σε αυτές τις εφαρμογές. Ωστόσο, αξιοσημείωτο είναι ότι οι περισσότερες μελέτες επιχειρούν να αξιοποιήσουν περισσότερες από μία μεθόδους, τόσο για σύγκριση όσο και για λόγους πλουραλισμού της λύσης που αναζητείται. Αυτό σημαίνει ότι οι έρευνες λαμβάνουν υπόψη τα πλεονεκτήματα και τις αδυναμίες διαφορετικών μεθόδων, και για το λόγο αυτό χρησιμοποιούν περισσότερες από μια για να αποκτήσουν σφαιρική άποψη για ασθένειες με σύνθετους μηχανισμούς. Κάποιες από τις εφαρμογές τους είναι [39]:

- Γενετική επιδημιολογία. Η δομή των ΤΔ διευκολύνει το χειρισμό των προβλημάτων γενετικής διασύνδεσης μεγάλης κλίμακας. Πολύ συχνά η έξοδος του συστήματος είναι ένας φαινότυπος, κατηγορικός (για παράδειγμα, υγιής ή όχι) ή αριθμητικός. Τα χαρακτηριστικά του συνόλου είναι γενετικοί δείκτες, όπως κατηγορικοί σηματοθορυβικοί λόγοι. Το αποτέλεσμα της διαδικασίας είναι ένα αξιόπιστο μοντέλο πρόβλεψης όπως επίσης και μια αναφορά αξιολόγησης των σηματοθορυβικών λόγων όσον αφορά την ικανότητα αξιολόγησής τους.

- Όταν γονιδιωματικά δεδομένα εισέρχονται στο σύστημα, κάποιοι ΤΔ αλγόριθμοι προσπαθούν να ανιχνεύσουν περιοχές όμοιες με αυτές που προέρχονται από τυποποιημένες αναλύσεις, ενώ άλλοι ενδιαφέρονται για σχέσεις μεταξύ μεμονωμένων γονιδίων.

- Άλλες τεχνικές δημιουργούν γενετικές περιοχές, όπου εκατοντάδες σηματοθορυβικοί λόγοι χρησιμοποιούνται κάθε φορά για να κατασκευαστεί το μοντέλο. Ωστόσο, αυτές οι τεχνικές

απαιτούν περαιτέρω έρευνα, καθώς αν οι γενετικές περιοχές δεν ανιχνεύονται από τυποποιημένες μεθόδους, η λειτουργία τους θα απέδιδε ανεπιθύμητα αποτελέσματα.

- Μια άλλη οικογένεια εφαρμογών των ΤΔ στη βιοπληροφορική προσπαθεί να προβλέψει τις ιδιότητες μορίων έχοντας ως βάση πληροφορίες αλληλουχίας, προσπαθεί για παράδειγμα να εντοπίσει την ικανότητα αντιγραφής του HIV.

- Τέλος, τα ΤΔ έχουν εφαρμοστεί στην οικολογία, ώστε η παρουσία ενός είδους να προβλεφθεί από κλιματικές και τοπογραφικές μεταβλητές, και φαίνεται να αποδίδουν στην πρόβλεψη από περιβαλλοντικές μεταβλητές.

4.2 Δεδομένα

Το αρχικό σύνολο δεδομένων που χρησιμοποιήθηκε ως είσοδος του μοντέλου παραχωρήθηκε από το Ιπποκράτειο Νοσοκομείο Αθηνών. Το σύνολο περιλαμβάνει 560 ασθενείς με διαβήτη και 17 χαρακτηριστικά πρόβλεψης. Το πρώτο χαρακτηριστικό ήταν ένας μοναδικός αναγνωριστικός αριθμός που χρησιμοποιείται μόνο για λόγους διάκρισης, και για το λόγο αυτό δεν εισέρχεται στη διαδικασία. Οι υπόλοιπες 16 μεταβλητές εισόδου παρουσιάζονται στον πίνακα 4.2:

Πίνακας 4.2 Οι 16 μεταβλητές εισόδου των δεδομένων

Συνεχείς Μεταβλητές	
Παράγοντας κινδύνου	**Μέση Τιμή ± Τυπική Απόκλιση**
Ηλικία	58.56 ± 10.70 (έτη)
Διάρκεια ΣΔ	7.67 ± 7.37 (*έτη*)
ΔΜΣ	29.49 ± 5.54
Γλυκοζυλιωμένη Αιμοσφαιρίνη	7.43 ± 1.81 (%)

Παλμική Πίεση	56.75 ± 15.80 (*mmHg*)
Γλυκόζη Νηστείας	165.15 ± 56.15 (*mg/dL*)
Συνολική Χοληστερόλη	226.64 ± 50.04 (*mg/dL*)
Τριγλυκερίδια	167.39 ± 110.81 (*mg/dL*)
HDL Χοληστερόλη	48.35 ± 16.46 (*mg/dL*)

Κατηγορικές Μεταβλητές

Παράγοντας κινδύνου	**Αριθμός ασθενών (Ποσοστό)**
Κάπνισμα	
Μη καπνιστές	289 (51.61%)
Καπνιστές	146 (26.07%)
Πρώην καπνιστές	125 (22.32%)
Φύλο	
Άνδρες	263 (46.96%)
Γυναίκες	297 (53.04%)
Υπέρταση	260 (46.42%)
Θεραπεία μείωσης λιπιδίων	
Όχι	469 (83.75%)
Στατίνες	74 (13.21%)
Φιβράτες	17 (3.04%)
Ασπιρίνη	
Όχι	509 (90.89%),
100 *mg*	44 (7.85%),
325 *mg*	7 (3.03%)

Θεραπεία με ινσουλίνη	
Όχι	494 (88.21%),
Ναι	66 (11.79%)
Οικογενειακό Ιστορικό ΣΔ	
Όχι	304 (54.28%)
Ναι	256 (45.72%)

Ο πίνακας επίσης σημειώνει ποια χαρακτηριστικά είναι ποσοτικά και ποια κατηγορικά. Στην δική μας υλοποίηση, χρησιμοποιήσαμε ένα δυαδικό σύστημα ετικετών (labels), αφού η διαδικασία εκπαίδευσης διαφοροποιείται ως προς τη φύση των μεταβλητών. Ακόμη, μια δυαδική τιμή εξόδου είναι διαθέσιμη για κάθε ασθενή και δηλώνει την εμφάνιση ή όχι της καρδιαγγειακής νόσου.

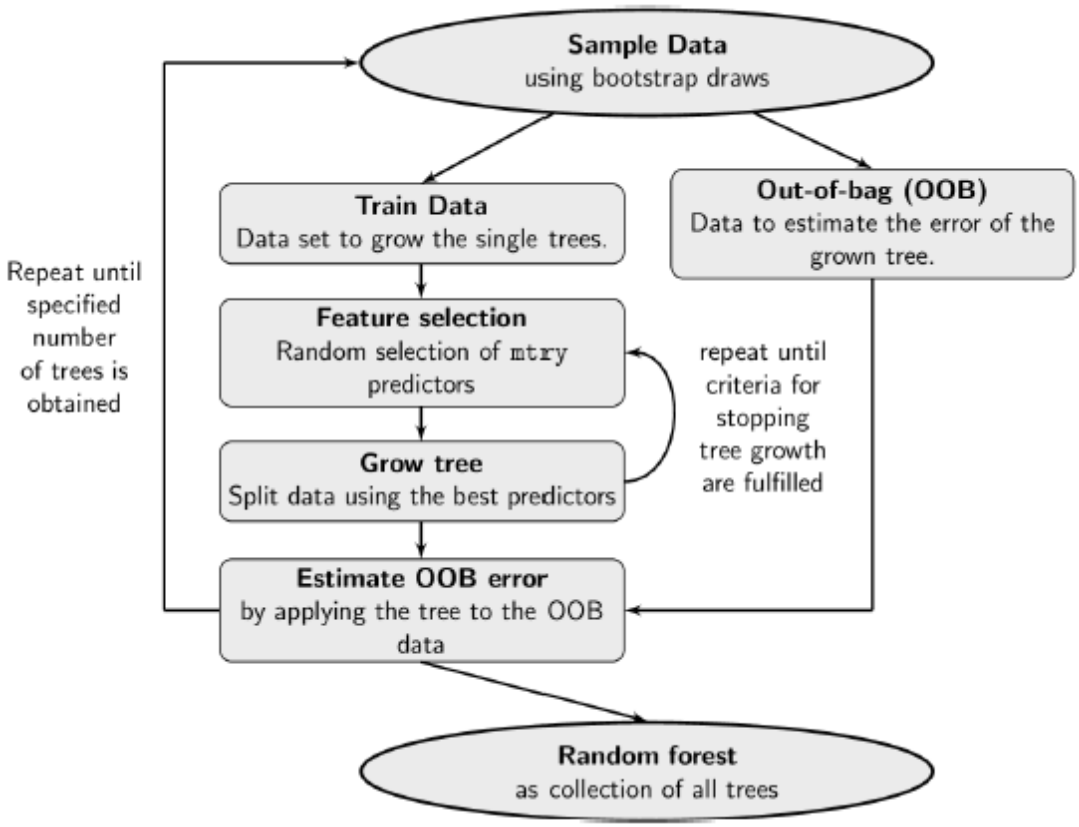
Προκειμένου να εκτιμήσουμε τον τρόπο με τον οποίο τα αποτελέσματα της διαδικασίας θα γενικεύονταν σε ένα ανεξάρτητο σύνολο δεδομένων, εφαρμόστηκε το κριτήριο διασταυρωμένης επικύρωσης σε 10 ίσα μέρη (10-fold cross-validation). Ο σκοπός αυτής της προεπεξεργασίας ήταν να μειώσουμε τη μεταβλητότητα, και αφού τελειώσει για κάθε εκδοχή η προτεινόμενη διαδικασία, τα κριτήρια αξιολόγησης που επιλέγονται συνυπολογίζονται σε μια μέση τιμή προκειμένου να παράγουν ένα ανεξάρτητο και γενικευμένο μέτρο.

4.3 Εκπαίδευση

Όπως αναφέραμε και στο προηγούμενο κεφάλαιο, οι μέθοδοι συλλογικής μάθησης έχουν κερδίσει μεγάλη δημοτικότητα. Τα boostingκαι baggingθεωρούνται αρκετά αντιπροσωπευτικά αυτής της οικογένειας τεχνικών. Τα δέντρα του boostingκατασκευάζονται διαδοχικά δίνοντας προβάδισμα σε εσφαλμένες προβλέψεις όσο λαμβάνει χώρα η διαδικασία, και η μέθοδος τελειώνει με ένα σταθμικό μέσο όρο στο συνολικό συλλογικό μηχανισμό. Από την άλλη πλευρά, το baggingπεριλαμβάνει την κατασκευή ανεξάρτητων δέντρων (η ιδέα της διαδοχής απουσιάζει στο bagging) και κάθε δέντρο απόφασης δημιουργείται ώστε να διαφέρει από τα άλλα όσον αφορά το σύνολο εισόδων του, που είναι ένα bootstrapδείγμα του αρχικού συνόλου εκπαίδευσης. Η έξοδος του συστήματος αποκτάται μετά από

εξαγωγή μέσης τιμής των αποτελεσμάτων όλων των δέντρων για ανάλυση παλινδρόμησης ή από πλειοψηφία για ταξινόμηση.

Η καινοτομία των ΤΔ αναφέρεται σε μια επιπλέον διάσταση τυχαιότητας: η τυχαιότητά τους μπορεί να εντοπιστεί κατά την κατασκευή των δέντρων και είναι η κύρια διαφορά τους με το bagging. Στην πραγματικότητα, όταν αυτή η επιπρόσθετη τυχαιότητα δε εισέρχεται στο σύστημα, τα bagging και ΤΔ είναι πανομοιότυπα. Τα ΤΔ περιλαμβάνουν bootstrapδειγματοληψία και τυχαίο διαχωρισμό κόμβων: προκειμένου να διασπαστεί ένας κόμβος, δε λαμβάνονται υπόψη όλα τα χαρακτηριστικά για την απόφαση (όπως στο bagging); Από την άλλη πλευρά, κάθε κόμβος διασπάται χρησιμοποιώντας ένα τυχαίο υποσύνολο των χαρακτηριστικών του διαθέσιμου συνόλου στο τρέχον σημείο της διαδικασίας. Είναι σημαντικό να αναφερθεί ότι ο αριθμός των στοιχείων του υποσυνόλου είναι σταθερός για κάθε κόμβο όλων των δέντρων (και συνήθως συμβολίζεται με mtry, που, μαζί με τον αριθμό των δέντρων, είναι οι παράμετροι της διαδικασίας). Το διάγραμμα ροής της ΤΔ τεχνικής απεικονίζεται στο Σχήμα 4.1 [39].

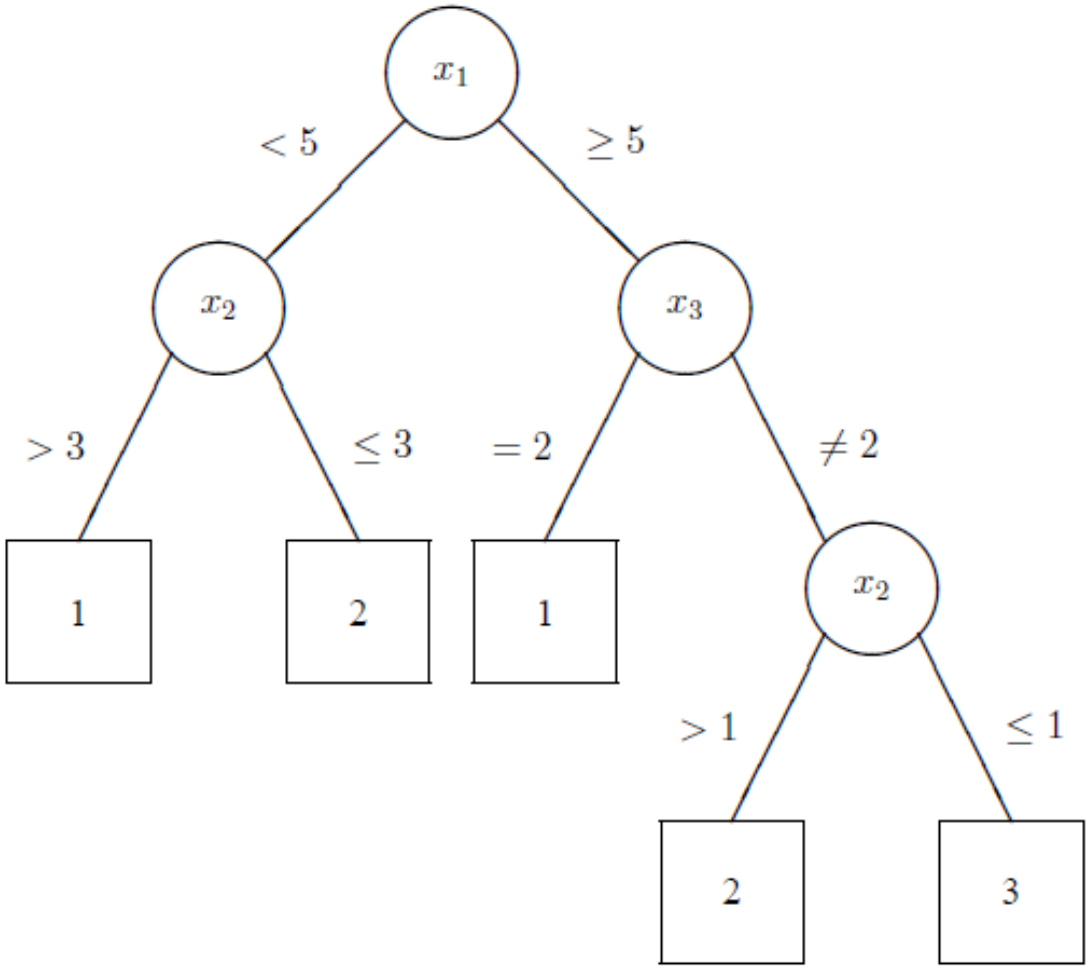


Σχήμα 4.3 Διάγραμμα ροής της ΤΔ προσέγγισης

Η εκμάθηση με δέντρα απόφασης είναι μια υποκατηγορία των μεθόδων μηχανικής μάθησης, που αξιοποιεί τη συμπαγή δομή των δέντρων και πραγματοποιεί μια σειρά απλών τεστ: κάθε ποσοτικό χαρακτηριστικό συγκρίνεται με μια τιμή κατωφλίου (ή ελέγχεται η παρουσία ενός κατηγορικού χαρακτηριστικού σε ένα σύνολο). Ένα καίριο πλεονέκτημα της μεθόδου είναι η απλότητά της: ένας μη εξοικειωμένος χρήστης μπορεί εύκολα να κατανοήσει την ιδέα της δομής του δέντρου, καθώς κάθε μονοπάτι του δέντρου μπορεί να μετασχηματιστεί σε μια σειρά ερωτήσεων με θετική απόκριση. Συνεπώς, πολλοί επιστήμονες προτιμούν την εκμάθηση με δέντρα απόφασης από άλλες μεθόδους (όπως τα νευρωνικά δίκτυα), λόγω της υπεροχής των πρώτων σε όρους κατανόησης.

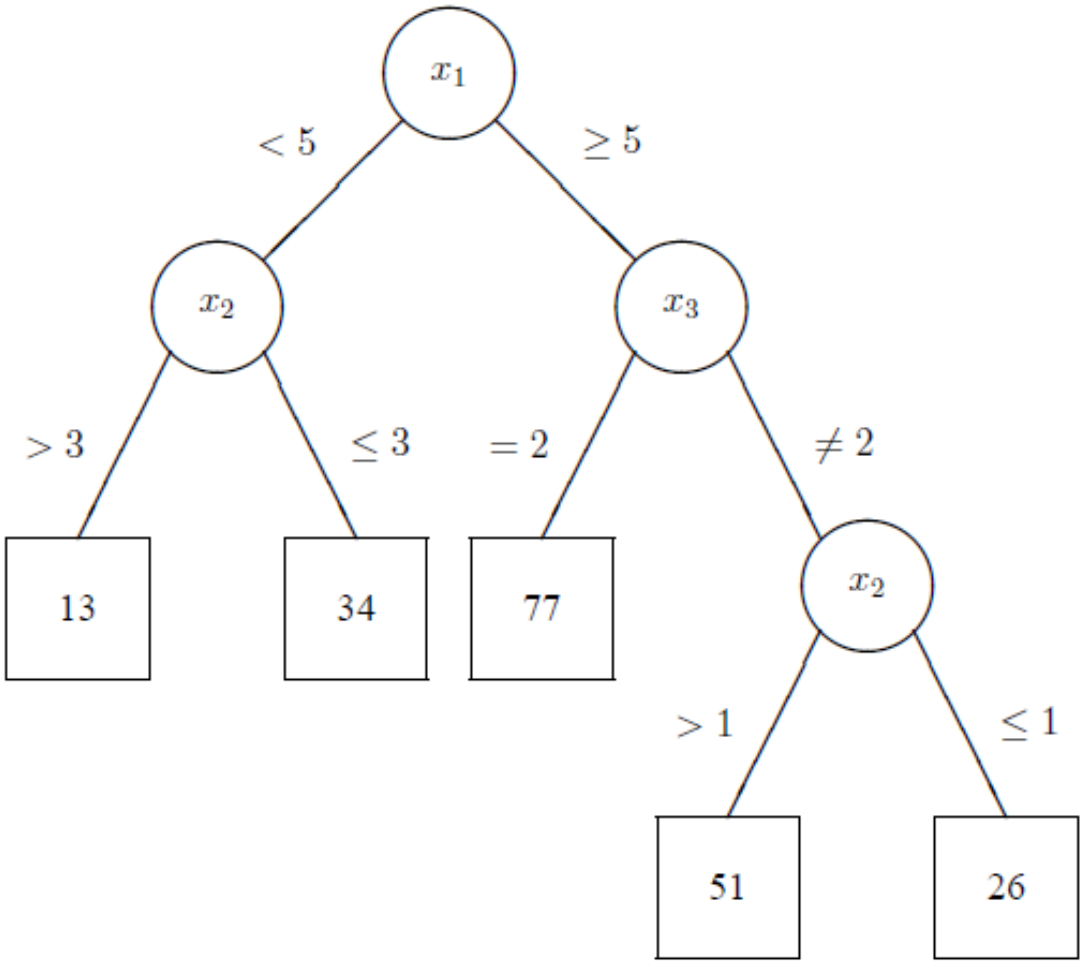
Στην ταξινόμηση, όταν ένας ασθενής βρεθεί σε έναν τελικό κόμβο, το δέντρο τον ταξινομεί στην πιο συνηθισμένη κλάση του κόμβου. Τότε ορίζεται το μέτρου του ρυθμού σφάλματος ως ο αριθμός όλων των ασθενών που ταξινομήθηκαν εσφαλμένα από το δέντρο προς το συνολικό αριθμό ασθενών. Αντίθετα, ο ρυθμός ευστοχίας είναι η μονάδα μείον το ρυθμό σφάλματος. Η διαδικασία ταξινόμησης (ή παλινδρόμησης) αποτελείται από μια greedyαναζήτηση, καθώς περιλαμβάνει αναζητήσεις μέσα σε ολόκληρο το χώρο πιθανών δέντρων απόφασης δεδομένης της αρχικής πληροφορίας. Η μέθοδος θεωρείται greedy για το λόγο ότι γίνονται προσπάθειες εύρεσης του σημείου διαχωρισμού του κόμβου που παρέχει το υψηλότερο κέρδος. Το επόμενο βήμα είναι η απόφαση για διαχωρισμό του δείγματος, και κατόπιν η συνολική διαδικασία επαναλαμβάνεται μέχρι ένας κόμβος να θεωρηθεί τερματικός [30].

Αν υποθέσουμε ότι έχουμε ένα διάνυσμα εισόδων **x** και την αντίστοιχη τιμή απόκρισης y, τότε ένα δέντρο ταξινόμησης θα μπορούσε να είναι το εξής:



Σχήμα 4.4 Παράδειγμα δέντρου ταξινόμησης

Η τιμή κάθε τερματικού κόμβου είναι η κατηγορία που θα ανατεθεί σε έναν ασθενή που θα βρεθεί στον εν λόγω τερματικό κόμβο. Οι ανισότητες πάνω από τα κλαδιά δηλώνουν το κατώφλι που αποφασίζει τον επόμενο κόμβο/προορισμό ενός ασθενή.



Σχήμα 4.5 Παράδειγμα ενός δέντρου παλινδρόμησης

Ένα δέντρο παλινδρόμησης ακολουθεί την ίδια διαδικασία, διαφέροντας μόνο στην κατασκευή των τερματικών κόμβων. Η παλινδρόμηση αποδίδει ένα αριθμητικό αποτέλεσμα, όπως μπορούμε να δούμε σε κάθε τερματικό κόμβο. Αυτό το αποτέλεσμα θα συνδεθεί με τον ασθενή που θα βρεθεί στο συγκεκριμένο κόμβο [40].

Ο Αλγόριθμος CART (Classification and Regression Trees) είναι η μέθοδος που χρησιμοποιήσαμε για την κατασκευή των δέντρων παλινδρόμησης. Ορίστηκε και περιγράφηκε εκτενώς το 1984 [41]. Σε αυτό σημείο θα αναλύσουμε τον αλγόριθμο για σκοπούς παλινδρόμησης μόνο. Ακολουθώντας τη γενική φιλοσοφία κατασκευής δέντρων απόφασης, ο CARTπεριλαμβάνει την αναζήτηση του βέλτιστου διαχωρισμού κόμβων μεταξύ όλων των διαθέσιμων. Η μέθοδος στοχεύει στη δημιουργία των «καθαρότερων» κόμβων, και ένας κανόνας διάσπασης είναι ότι οι πολυμεταβλητοί διαχωρισμοί δεν επιτρέπονται (δηλαδή στο τέλος μόνο ένα χαρακτηριστικό θα επιλεγεί για τον κάθε κόμβο). Συνεπώς, κάθε μοναδική τιμή του κάθε χαρακτηριστικού θεωρείται υποψήφιο σημείο διαχωρισμού, Αν ένα κατηγορικό χαρακτηριστικό αποτελείται από Ικατηγορίες, τότε υπάρχουν $2^{I-1}-1$πιθανοί διαχωρισμοί. Αν ένα χαρακτηριστικό είναι αριθμητικό με K μοναδικές τιμές στο τρέχον δείγμα, τότε υπάρχουν K-1 διαθέσιμοι διαχωρισμοί για το συγκεκριμένο χαρακτηριστικό. Η διαδικασία κατασκευής αρχίζει από τη «ρίζα» του δέντρου, που είναι ένας κόμβος που περιέχει όλο το δείγμα. Σε κάθε κόμβο, ακολουθείται μια συγκεκριμένη σειρά βημάτων:

Αλγόριθμος 1. CART

1. Βρες το βέλτιστο σημείο διαχωρισμού για κάθε χαρακτηριστικό. Αν το χαρακτηριστικό έχει ποσοτικές τιμές, αυτές ταξινομούνται σε αύξουσα σειρά και κάθε μοναδική τιμή (δηλαδή ίδιες τιμές δεν απαιτείται να επανελεγχθούν) θεωρείται υποψήφιο κατώφλι για το διαχωρισμό. Αν υποθέσουμε ότι αυτή η τιμή είναι v, τότε αν x $\leq$ v, όπου xη τιμή αυτού του χαρακτηριστικού για ένα δείγμα, τότε το δείγμα ανατίθεται στον αριστερό κόμβο-παιδί, αλλιώς ανατίθεται στο δεξιό κόμβο-παιδί.

   Το βέλτιστο σημείο διαχωρισμού είναι αυτό που μεγιστοποιεί το κριτήριο διαχωρισμού. Το σύνηθες κριτήριο για την παλινδρόμηση είναι το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error–MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

   Όπου N είναι ο αριθμός των δειγμάτων στον τρέχοντα κόμβο, y_iείναι η τιμή απόκρισης ενός δείγματος στον τρέχοντα κόμβο και $\bar{y}$είναι η μέση τιμή αποκρίσεων για όλα τα δείγματα στον

κόμβο. Συνεπώς, η τιμή ντου συγκεκριμένου χαρακτηριστικού επιλέγεται, αν οι δύο κόμβοι (δεξιός και αριστερός) που δημιουργεί έχουν το ελάχιστο συνδυασμένο MSEμεταξύ όλων των συνδυασμών (δηλαδή το άθροισμα των MSE τους είναι το ελάχιστο μεταξύ των αθροισμάτων MSE για όλους τους υποψήφιους διαχωρισμούς του κόμβου).

Μια παρόμοια τεχνική εφαρμόζεται για κατηγορικά χαρακτηριστικά: αντί για τη δοκιμή κάθε μοναδικής τιμής, θεωρείται ένας αριθμός υποσυνόλων του δυναμοσυνόλου των τιμών αυτών, όπως θα περιγράψουμε στην επόμενη παράγραφο. Αυτά τα υποσύνολα ανταγωνίζονται μεταξύ τους, χρησιμοποιώντας το ίδιο κριτήριο διαχωρισμού με πριν (MSE). Κατά τον ίδιο τρόπο σκέψης, αν το «βέλτιστο» υποσύνολο είναι το A και αν $x \in A$, το δείγμα ανατίθεται στον αριστερό κόμβο-παιδί, αλλιώς ανατίθεται στο δεξιό κόμβο-παιδί.

2. Βρες το βέλτιστο σημείο διαχωρισμού για κάθε κόμβο. Η διαδικασία επεκτείνεται σε σύγκριση όλων των χαρακτηριστικών Μεταξύ των βέλτιστων σημείων διαχωρισμού που βρέθηκαν για κάθε χαρακτηριστικό στο Βήμα 1, επίλεξε εκείνο που μεγιστοποιεί το κριτήριο διαχωρισμού για τον κόμβο.

3. Διαχώρισε τον κόμβο χρησιμοποιώντας το βέλτιστο σημείο διαχωρισμού που βρέθηκε στο βήμα 2, και μόνο αν δεν επιβεβαιώνονται τα κριτήρια τερματισμού.

Για κριτήρια τερματισμού υπάρχουν πολλές επιλογές. Έχουμε επιλέξει να λάβουμε υπόψη το αναπόφευκτο κριτήριο που προκαλείται από την απουσία διαθέσιμων σημείων διαχωρισμού (αυτό συμβαίνει όταν όλα τα δείγματα που έχουν μείνει έχουν τις ίδιες τιμές για κάθε χαρακτηριστικό), καθώς επίσης και το κριτήριο που υποδεικνύει τον τερματισμό της διαδικασίας όταν 5 η λιγότερα δείγματα πρόκειται να βρεθούν σε οποιονδήποτε από τους κόμβους-παιδιά.

Η εκδοχή του CART για ΤΔ δεν περιλαμβάνει τη συμμετοχή όλων των χαρακτηριστικών στη διαδικασία διαχωρισμού ενός κόμβου. Αντίθετα, όπως έχουμε ήδη αναφέρει, περιλαμβάνει την τυχαία επιλογή mtry χαρακτηριστικών, που θα ανταγωνιστούν μεταξύ τους για το διαχωρισμό του κόμβου. Με άλλα λόγια, ο CART υφίσταται μια μικρή αλλά ουσιώδη τροποποίηση όταν χρησιμοποιείται από την τεχνική ΤΔ, ενώ παραμένει αμετάβλητος σε τεχνικές bagging [41].

Όπως αναφέραμε, ο αλγόριθμος CART αναφέρεται στην κατασκευή των ατομικών μηχανών μάθησης. Ο αλγόριθμος ΤΔ περιγράφει τη συνολική διαδικασία που απαιτείται για την ανάπτυξη του γενικού συλλογικού συστήματος [43] [44]:

Αλγόριθμος 2. Random Forests

1. Δημιούργησε ntree bootstrap δείγματα από τα αρχικά δεδομένα.

2. Για καθένα από τα bootstrap δείγματα, κατασκεύασε ένα δέντρο ταξινόμησης ή παλινδρόμησης (ανάλογα με τη φύση του προβλήματος), με την εξής τροποποίηση: σε κάθε κόμβο, αντί της επιλογής του βέλτιστου σημείου διαχωρισμού μεταξύ όλων των χαρακτηριστικών, πάρε mtry τυχαία δείγματα από τα χαρακτηριστικά και επίλεξε το βέλτιστο σημείο διαχωρισμού μεταξύ των χαρακτηριστικών αυτού του τυχαίου υποσυνόλου (το bagging μπορεί να θεωρηθεί ως η ειδικά περίπτωση των ΤΔ όταν επιλέγεται mtry = p, όπου p ο αριθμός των χαρακτηριστικών του τρέχοντος δείγματος.

3. Πρόβλεψε τις αποκρίσεις νέων δεδομένων συσσωρεύοντας τις προβλέψεις των δέντρων (δηλαδή τα αποτελέσματα της πλειοψηφίας για δέντρα ταξινόμησης και της μέσης τιμής για δέντρα παλινδρόμησης).

Κάποια από τα χαρακτηριστικά όμως μπορεί να είναι κατηγορικά, που σημαίνει ότι έχουν διακριτές (συνήθως ακέραιες) τιμές που αντιστοιχούν σε κατηγορίες για παράδειγμα «καπνιστής» και «μη-καπνιστής». Επομένως, ο τρόπος που θα χειριστούμε αυτές τις μεταβλητές θα πρέπει να προσδιοριστεί, ώστε να συνεισφέρουν στη έξοδο του συστήματος μαζί με τα ποσοτικά χαρακτηριστικά. Η προσέγγιση που ακολουθείται από τον Breiman είναι η εξής: κάθε φορά που η επιλεγμένη μεταβλητή για το υποψήφιο σημείο διαχωρισμού είναι κατηγορική, κατασκευάζεται το δυναμοσύνολο των κατηγορικών μεταβλητών. Όταν παραχθεί το δυναμοσύνολο, κάποια από τα στοιχεία του (δηλαδή τα υποσύνολα που

περιέχουν συνδυασμούς των κατηγορικών μεταβλητών) επιλέγονται τυχαία. Κατά συνέπεια, ορίζεται μια μεταβλητή-υποκατάστατο για κάθε υποσύνολο, και αναπαριστά την παρουσία ή όχι της κατηγορικής τιμής του συγκεκριμένου χαρακτηριστικού για τον ασθενή στο συγκεκριμένο υποσύνολο. Για παράδειγμα, αν το χαρακτηριστικό παίρνει τιμές από το σύνολο {1,2,3}, και τα {1} και {1,2} είναι τα στοιχεία που επιλέγονται από το δυναμοσύνολο, τότε ένας ασθενής με τιμή 3 για το χαρακτηριστικό αυτό θα έχει μηδενική τιμή στη δυαδική μεταβλητή-υποκατάστατο και για τα δύο υποσύνολα, ενώ ένας ασθενής με τιμή 2 θα έχει μηδενική τιμή για τη δυαδική μεταβλητή του πρώτου υποσυνόλου και τιμή 1 για τη δυαδική μεταβλητή του δεύτερου υποσυνόλου. Αφού κατασκευαστεί η δυαδική μεταβλητή, ο περαιτέρω χειρισμός της είναι πανομοιότυπος με αυτόν που εφαρμόζεται στις ποσοτικές μεταβλητές [44].

Έχουμε πει ότι ένα από τα μεγάλα πλεονεκτήματα των ΤΔ είναι η απουσία πολλών και σύνθετων παραμέτρων. Οι παράμετροι της μεθόδου είναι το ntrees, ο αριθμός των δέντρων που θα κατασκευαστούν στο συλλογικό μηχανισμό, και mtry, ο αριθμός των χαρακτηριστικών που θα επιλεχθούν τυχαία από το χώρο των χαρακτηριστικών για κάθε υποψήφιο σημείο διαχωρισμού του κάθε κόμβου (και είναι σταθερός για όλα τα δέντρα στο συλλογικό μηχανισμό).

Η πρώτη παράμετρος, ntrees, ήταν σχετικά εύκολο να ρυθμιστεί, καθώς τα ΤΔ είναι μια μέθοδος που δεν υπερεκπαιδεύεται όσο αυξάνει ο αριθμός των δέντρων [44]. Αυτό εξηγείται από την παράλληλη και ανεξάρτητη κατασκευή των δέντρων στο ίδιο συλλογικό μηχανισμό. Η εκπαίδευση ενός δέντρου δεν επηρεάζει την εκπαίδευση των άλλων, καθώς η ιδέα της διαδοχής απουσιάζει στα ΤΔ (σε αντίθεση με τον αλγόριθμο επαναλαμβανόμενων βαρών του boostingγια τα διαδοχικά δέντρα ή με τις εποχές στα νευρωνικά δίκτυα). Συνεπώς, αρχίσαμε τη διαδικασία με ένα σχετικά μεγάλο αριθμός δέντρων και στη συνέχεια μειώσαμε σταδιακά μέχρι που παρατηρήσαμε μείωση στην απόδοση, προκειμένου να της εξισορροπήσουμε με την υπολογιστική πολυπλοκότητα. Η τελικά τιμή ήταν τα 500 δέντρα.

Η δεύτερη παράμετρος, mtry, δε θεωρείται τόσο εύκολη στη ρύθμιση όσο η πρώτη, καθώς μπορεί να προκαλέσει δυσχερέστερη λειτουργία του συστήματος και προς τις δύο κατευθύνσεις. Ωστόσο, προκειμένου να αποφευχθούν υπερβολικά υψηλές ή χαμηλές τιμές που θα μείωναν την αποδοτικότητα του συστήματος, υπάρχει ένας εμπειρικός βιβλιογραφικός κανόνας που είναι (στην περίπτωση της παλινδρόμησης) [46]:

$$mtry = \lceil \frac{1}{3} p \rceil,$$

όπου p είναι ο αριθμός των διαθέσιμων χαρακτηριστικών και τα σύμβολα που περιβάλλουν την έκφραση στο δεξί μέλος της ισότητας υποδεικνύουν την επιλογή του αμέσως επόμενου ακέραιου. Στη περίπτωση του δικού μας συστήματος, η προτεινόμενη mtry παράμετρος που παράχθηκε από τον κανόνα ήταν το 6, και η τελική τιμή που χρησιμοποιήσαμε ήταν 5 τυχαία επιλεγμένα χαρακτηριστικά (ανταγωνιζόμενα για το σημείο διαχωρισμού κάθε κόμβου).

Κάποιος θα μπορούσε να αποκτήσει περισσότερες πληροφορίες από ένα σύστημα αν σε αυτό έχει εφαρμοστεί η ΤΔ μέθοδος. Αυτή η πληροφορία απαντάται στη μορφή δύο αριθμητικών μέτρων που παρέχουν καλύτερη κατανόηση των πτυχών του συστήματος. [45]:

- **Importance και OOB σφάλμα.** Το μέτρο του importance των μεταβλητών είναι δύσκολο στην περιγραφή, ωστόσο παρέχει πολύτιμες πληροφορίες για το σύνολο εκπαίδευσης. Για το λόγο αυτό είναι καλύτερο να αναφερθεί πρώτα το Out-Of-Bag σφάλμα: ένα δέντρο κατασκευάζεται αρχικά και το σφάλμα πρόβλεψης υπολογίζεται για τα OOB δεδομένα (αυτά που δε χρησιμοποιήθηκαν στην εκπαίδευση και αφέθηκαν εκτός λόγω των διπλότυπων που εμφανίστηκαν στα bootstrap δείγματα). Τότε ένα χαρακτηριστικό x_j μετατίθεται, δηλαδή το κανάλι που μεταφέρει την πληροφορία του μεταξύ του χαρακτηριστικού και της εξόδου καταστρέφεται και υπολογίζεται η διαφορά μεταξύ του αρχικού και του OOB σφάλματος, παρέχοντας πληροφορίες για μεταβολή στην ακρίβεια. Η διαδικασία επαναλαμβάνεται για κάθε δέντρο στο συλλογικό μηχανισμό, και οι διαφορές που παράγονται συνεισφέρουν στον υπολογισμό μιας μέσης τιμής που αποτελεί το μέτρο του importance της μεταβλητής.

- **Proximity** Κατασκευάζεται ένας πίνακας εγγύτητας, και περιέχει (i, j) στοιχεία που αντιπροσωπεύουν το κλάσμα των δέντρων στα οποία τα δείγματα i και j καταλήγουν στον ίδιο τερματικό κόμβο. Δεν είναι δύσκολο να υποθέσουμε ότι παρόμοια δείγματα θα βρεθούν στους ίδιους τερματικούς κόμβους (τουλάχιστον πιο συχνά από τα ανόμοια). Ο πίνακας εγγύτητας αποκαλύπτει πτυχές της δομής δεδομένων στο χρήστη, ενώ μπορεί και να χρησιμοποιηθεί σε ΤΔ μάθηση χωρίς επίβλεψη.

Κεφάλαιο 5. Αποτελέσματα και Συζήτηση

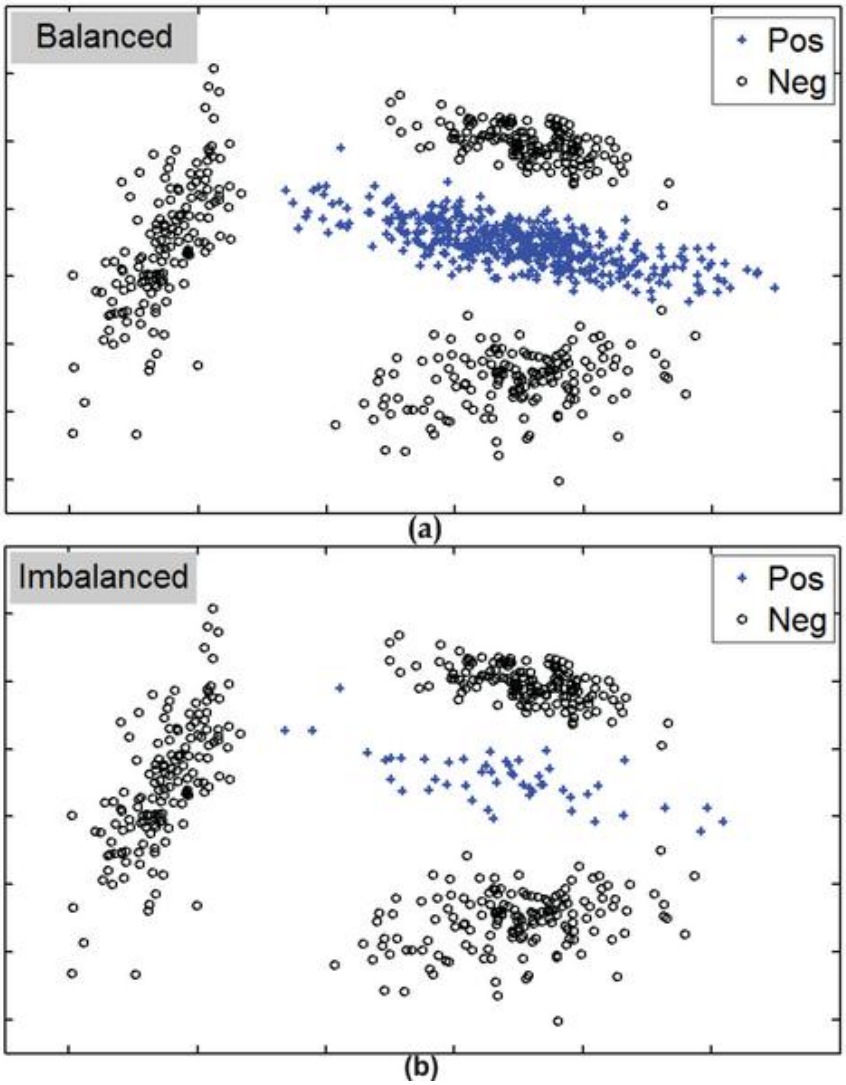
5.1 Γενική αξιολόγηση

Παρόλο που ο έλεγχος τους μοντέλου χρησιμοποιώντας ένα σύνολο αξιολόγησης θα μπορούσε να θεωρηθεί μέρος της μεθοδολογίας, έχουμε αποφασίσει να τον συμπεριλάβουμε στο τρέχον κεφάλαιο, καθώς ο κύριος στόχος του είναι να παράγει τα αποτελέσματα του συστήματος προκειμένου να τα αξιολογήσει ο χρήστης.

Η διαδικασία διασταυρωμένης επικύρωσης που αναφέραμε στο προηγούμενο κεφάλαιο έχει διαιρέσει τα δεδομένα σε δύο σύνολα για κάθε εκδοχή (ως εκδοχή αναφέρουμε τα folds): το πρώτο σύνολο είναι το σύνολο εκπαίδευσης, ενώ το δεύτερο σύνολο της εκδοχής ονομάζεται σύνολο αξιολόγησης. Συνεπώς, όταν τελειώσει η διαδικασία, το σύστημα που δημιουργήθηκε μπορεί να αξιολογηθεί. Το πρώτο βήμα προς αυτή την κατεύθυνση είναι να τροφοδοτηθεί το σύστημα με έναν αριθμό δειγμάτων που δεν έχουν εισαχθεί ξανά στο μοντέλο. Αυτό σημαίνει ότι δεν έχουν επηρεάσει την κατασκευή στου συστήματος σε κανένα σημείο. Σε αυτή την περίπτωση δεν παρέχονται στο σύστημα οι αποκρίσεις των δειγμάτων. Οι αποκρίσεις αυτές ονομάζονται αναμενόμενες έξοδοι και θα χρησιμοποιηθούν όταν ανατεθούν στο σύνολο αξιολόγησης οι πραγματικές έξοδοι του συστήματος, οι προβλέψεις. Αυτές μπορούν να ληφθούν ως εξής: κάθε δείγμα εισέρχεται σε κάθε δέντρο του συλλογικού μηχανισμού, και ακολουθεί τους κανόνες που είχαν παραχθεί σε κάθε κόμβο. Για παράδειγμα, αν συναντήσει έναν κόμβο όπου ο κανόνας υποδεικνύει ότι τα δείγματα με τιμή μικρότερη από ν για το (ποσοτικό) χαρακτηριστικόx_iθα πρέπει να ακολουθήσουν την αριστερή κατεύθυνση, ενώ αυτά που έχουν μεγαλύτερη τιμή από νθα πρέπει να ακολουθήσουν τη δεξιά κατεύθυνση στο δέντρο. Αυτό σημαίνει ότι το τρέχον δείγμα πρέπει να υπακούει στους κανόνες και να ακολουθεί την ανάλογη κατεύθυνση. Αυτή η σειρά βημάτων οδηγεί στον τερματικό κόμβο, όπου του ανατίθεται η μέση απόκριση των δειγμάτων του τερματικού κόμβου. Αυτές οι τιμές συλλέγονται από όλα τα δέντρα για το συγκεκριμένο δείγμα και στη συνέχεια υπολογίζεται η μέση τιμή τους για όλα τα δέντρα (ntrees). Αυτή είναι η εκτιμούμενη πιθανότητα για το συγκεκριμένο δείγμα. Η διαδικασία στη συνέχεια επαναλαμβάνεται για κάθε δείγμα και έτσι κατασκευάζεται ένα διάνυσμα πιθανοτήτων, σε αντίθεση με το αναμενόμενο διάνυσμα, που περιέχει την αρχική δυαδική απόκριση που δείχνει αν οι τιμές ήταν θετικές ή αρνητικές (στην περίπτωσή μας αν οι ασθενείς ανέπτυξαν καρδιαγγειακή νόσο ή όχι) [36].

5.2 Ισορροπία δεδομένων

Έχουμε αποφασίσει να δούμε τα δεδομένα από δύο σκοπιές: η πρώτη σκοπιά περιλαμβάνει τη θεώρηση των αρχικών δεδομένων ως ενός συνόλου. Ωστόσο, αυτό το σύνολο δεν είναι ισορροπημένο (όρος που αναφέρεται στην άνιση συχνότητα ύπαρξης θετικών και αρνητικών δειγμάτων στο σύνολο). Είναι σχετικά συνηθισμένο για δεδομένα που χρησιμοποιούνται στα μοντέλα πρόβλεψης κινδύνου να έχουν περισσότερες αρνητικές περιπτώσεις, δηλαδή τις περιπτώσεις ασθενών που δεν ανέπτυξαν την ασθένεια [47] (Σχήμα 5.1 [48]).



Σχήμα 5.1 Ισορροπημένα και μη ισορροπημένα δεδομένα

Αυτό θα μπορούσε να βλάψει το μοντέλο, καθώς θα γνωρίζει ότι ένα μεγάλο ποσοστό δειγμάτων είχαν αρνητική απόκριση, και θα οδηγούσε σε bias εναντίον της μικρής κλάσης (αυτής δηλαδή που είναι λιγότερο συχνή, η θετική κλάση στην περίπτωσή μας), και το σύστημα θα έτεινε να ευνοήσει τη μεγάλη κλάση (την πιο συχνή, την αρνητική κλάση στην περίπτωσή μας). Συνεπώς, το σύστημα θα είχε ως έξοδο πολύ χαμηλές πιθανότητες ανάπτυξης της ασθένειας και δε θα αντιστοιχούσαν στην πραγματική περίπτωση. Για το λόγο αυτό αποφασίσαμε να πραγματοποιήσουμε ελέγχους και σε ισορροπημένα δεδομένα, γεγονός που αποτελεί τη δεύτερη σκοπιά από την οποία βλέπουμε τα δεδομένα. Συγκεκριμένα, έχουμε αρχικά 560 διαφορετικούς ασθενείς με πολύ μικρή ισορροπία. Μόνο 41 από αυτούς ήταν θετικοί, ενώ οι υπόλοιποι δεν εμφάνισαν την ασθένεια μέχρι τη στιγμή που έγινε η συλλογή των στατιστικών δεδομένων. Προκειμένου να εξετάσουμε τη λειτουργία του συστήματός μας μετά από την τροφοδότησή του με ισορροπημένα δεδομένα, επιλέγουμε άλλες 41 τυχαίες αρνητικές περιπτώσεις και αφήνουμε τη διαδικασία αμετάβλητη.

5.3 Διαδικασία Αξιολόγησης - Κριτήρια

Αφού ολοκληρωθεί η διαδικασία του testing και είναι πλέον διαθέσιμο το διάνυσμα των αναμενόμενων τιμών για κάθε εκδοχή, μπορεί να λάβει ψώρα μια αξιολόγηση του μοντέλου. Έχουμε αποφασίσει να χρησιμοποιήσουμε το Hosmer-Lemeshow τεστ ως ένδειξη του “goodness of fit” του μοντέλου μας, που είναι ένας όρος που περιγράφει όσο καλά ένα μοντέλο ταιριάζει με τις παρατηρήσεις μας.

Η διαδικασία αξιολόγησης θα μας δώσει δύο μέτρα, ένα για discrimination(που παράγεται από τον υπολογισμό από τη στατιστική c) και ένα για calibration (που παράγεται από το Hosmer-Lemeshow τεστ). Εξηγούμε σύντομα τη σημασία αυτών των όρων [50]:

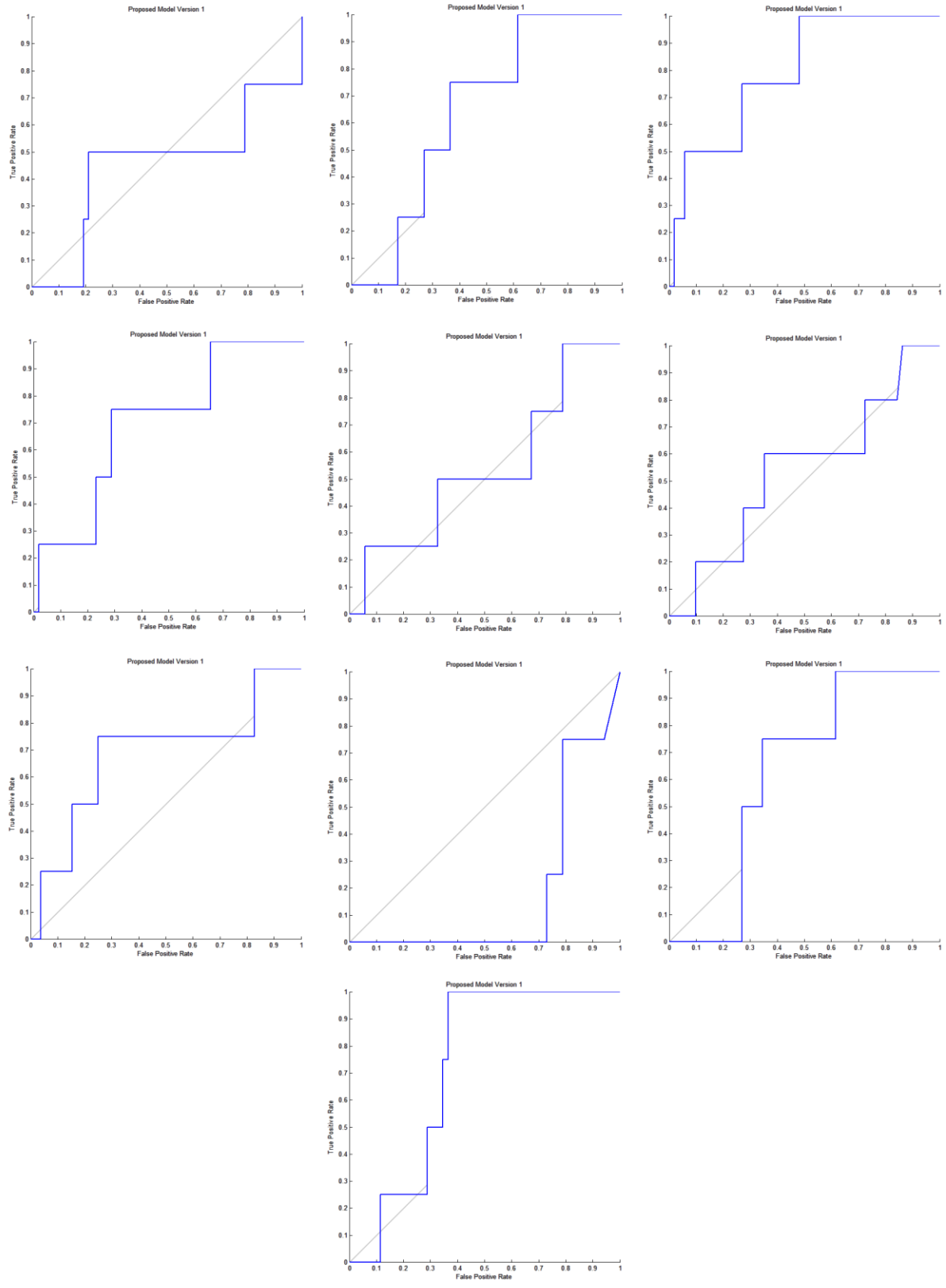
- Discrimination: Ο όρος αναφέρεται στην ικανότητα του μοντέλου να αποφασίσει σωστά το αποτέλεσμα. Το μέτρο που πρόκειται να χρησιμοποιήσουμε για αυτό μέρος της αξιολόγησης είναι η Επιφάνεια κάτω από τη Χαρακτηριστική Καμπύλη (Receiver Operating Characteristic - ROC). Η επιφάνεια είναι γνωστή ως Επιφάνεια Κάτω από την Καμπύλη (Area Under the Curve - AUC). Προέρχεται από τη στατιστική c, που είναι πανομοιότυπη με την AUC για προβλήματα με δυαδικές τιμές απόκρισης. Η Χαρακτηριστική Καμπύλη που κατασκευάζεται απεικονίζει το μέτρο της ευαισθησίας ή το ρυθμό αληθώς θετικών δειγμάτων (το λόγο των θετικών δειγμάτων που ταξινομούνται και ως θετικά) ως προς την έκφραση 1-(ρυθμός ψευδώς θετικών δειγμάτων).

Ο ρυθμός ψευδώς θετικών δειγμάτων είναι ο λόγος των θετικών δειγμάτων που δεν αναγνωρίζονται ως θετικά.

- Calibration: Αναφέρεται στη συμφωνία μεταξύ των αναμενόμενων και των προβλεπόμενων τιμών απόκρισης. Για παράδειγμα, αν προβλέψουμε ένα ρίσκο 20% για ανάπτυξη καρδιαγγειακής νόσου σε ασθενείς με ΣΔΤ2, η συχνότητα καρδιαγγειακής νόσου θα έπρεπε να είναι περίπου 20 θετικοί για 100 ασθενείς. Αποτελέσματα με παρόμοιες πιθανότητες μπορούν να κατηγοριοποιηθούν σε ομάδες, και η μέση εκτιμώμενη πιθανότητά τους μπορεί να συγκριθεί με τη μέση έξοδο που παρατηρήθηκε. Αυτός είναι και ο κύριος στόχος του Hosmer-Lemeshow τεστ. Επιπρόσθετα, το μέτρop-valueδείχνει αν το μοντέλο είναι συνεπές με την υπόθεση ότι το null hypothesis (δηλαδή η πεποίθηση ότι υπάρχει μια σύνδεση μεταξύ δύο φαινομένων, του ΣΔΤ2 και της καρδιαγγειακής νόσου στην περίπτωσή μας) είναι αληθής. Αν το p-value είναι μικρότερα από ένα επίπεδο (συνήθως τιμής 5%), τότε η μηδενική υπόθεση απορρίπτεται, αλλιώς γίνεται δεκτή.

5.4 Αποτελέσματα

- **Αποτελέσματα για μη ισορροπημένα δεδομένα:** Το επόμενο Σχήμα 5.2 απεικονίζει τις 10 AUCs που αντιστοιχούν στις 10 εκδοχές που δημιουργούνται από τη διαδικασία. Ακολουθείται από τον Πίνακα 5.1 που περιέχει τις AUC τιμές ως ποσοστά.



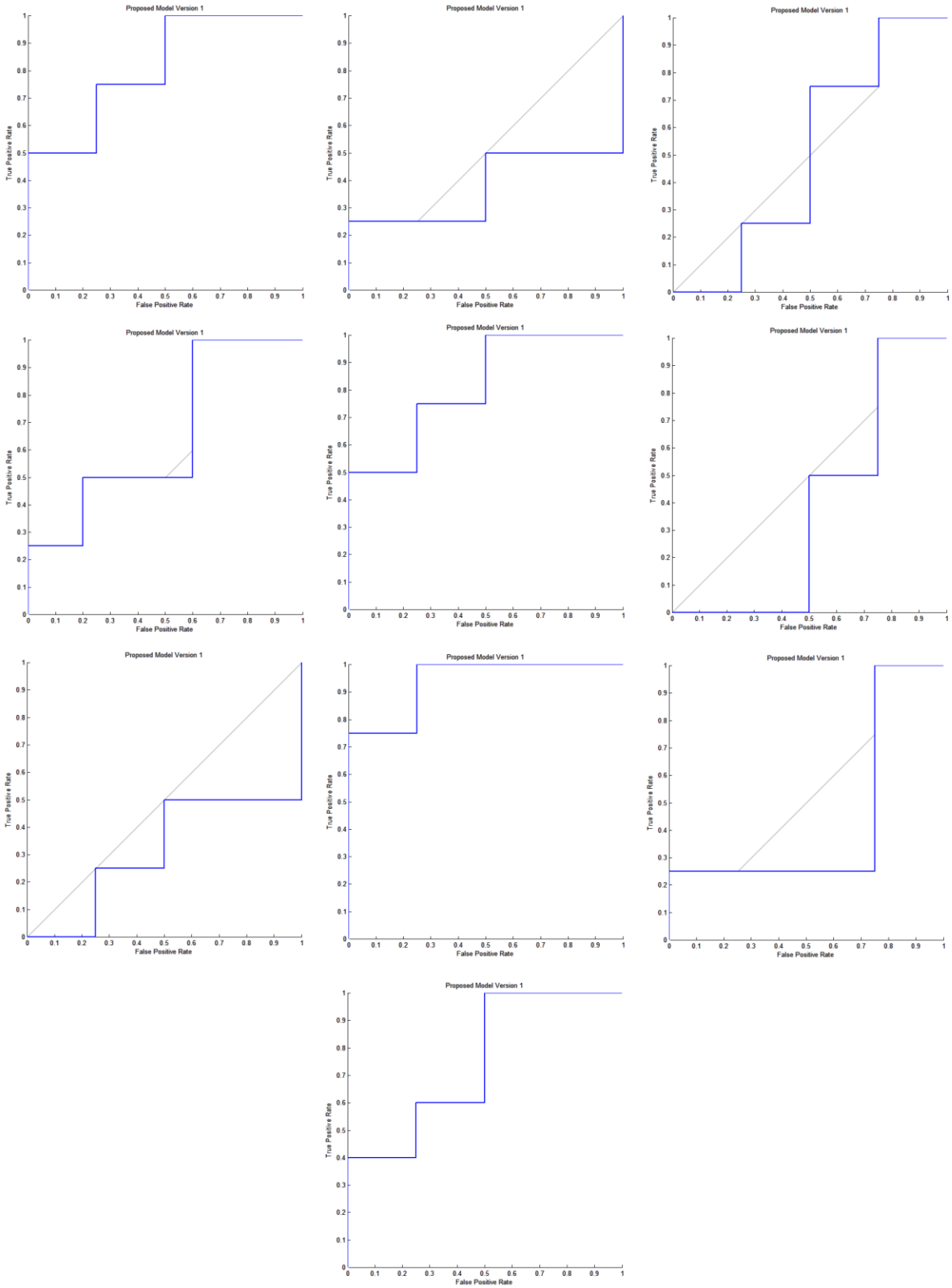
Σχήμα 5.2 AUCs του συστήματός μας για την περίπτωση μη ισορροπημένων δεδομένων

Πίνακας 5.1 AUC ποσοστά ως αποτέλεσμα του συστήματός μας για τα μη ισορροπημένα δεδομένα ανά εκδοχή

AUC
0.452
0.644
0.793
0.702
0.539
0.540
0.683
0.180
0.625
0.721

Το μέσο AUC ποσοστό είναι προσεγγιστικά ίσο με 0.59, ενώ η τυπική απόκλιση είναι περίπου 0.175. Το p-value είναι κοντά στο μηδέν για κάθε εκδοχή.

- **Αποτελέσματα για ισορροπημένα δεδομένα:**



Σχήμα 5.3 AUCs του συστήματός μας για ισορροπημένα δεδομένα

Πίνακας 5.2 AUC ποσοστά και p-values ως αποτέλεσμα του συστήματος για ισορροπημένα δεδομένα ανά εκδοχή

AUC	p - value
0.813	0.85
0.375	0.46
0.500	0.93
0.650	0.55
0.813	0.26
0.375	0.92
0.313	0.31
0.938	0.95
0.438	0.74
0.750	0.93

Το μέσο AUC ποσοστό είναι προσεγγιστικά ίσο με 0.60, αν η τυπική απόκλιση είναι 0.223 και το μέσο p-valueείναι 0.69, που σημαίνει ότι τα αποτελέσματα της μεθόδου είναι έγκυρα και μπορούν να ληφθούν υπόψη.

5.5Αποτελέσματα Matlab

Για λόγους πληρότητας και εγκυρότητας, επιλέξαμε να αναπτύξουμε το ίδιο σύστημα, αξιοποιώντας τα αντίστοιχα εργαλεία που παρέχονται από το Matlab (κλάση TreeBagger). Χρησιμοποιήσαμε τις ίδιες τιμές παραμέτρων με πριν, με σκοπό να μιμηθούμε τη διαδικασία όσο περισσότερο γίνεται. Αυτός ο τρόπος σκέψης θα μας επιτρέψει να συγκρίνουμε την ικανότητα των δύο προσεγγίσεων για discrimination. Παρουσιάζουμε τα αποτελέσματα με τον ίδιο τρόπο με προηγουμένως, τόσο για τα ισορροπημένα όσο και για τα μη ισορροπημένα δεδομένα:

- **Αποτελέσματα για μη ισορροπημένα δεδομένα:**

Πίνακας 5.3 Ποσοστά AUC (αποτελέσματα Matlab) για μη ισορροπημένα δεδομένα ανά εκδοχή

AUC
0.750
0.875
0.450
0.500
0.500
0.313
0.600
0.500
0.938
0.875

Το μέσο ποσοστό AUC είναι προσεγγιστικά ίσο με 0.63, ενώ η τυπική απόκλιση είναι 0.21.

- **Αποτελέσματα για ισορροπημένα δεδομένα:**

Πίνακας 5.4 Ποσοστά AUC (αποτελέσματα Matlab) για ισορροπημένα δεδομένα ανά εκδοχή

AUC
0.813
0.750
0.550
0.500
0.438
0.313
0.650
0.438
0.938
0.875

Το μέσο ποσοστό AUC είναι περίπου ίσο με 0.63, ενώ η τυπική απόκλιση είναι περίπου 0.21.

5.6 Αποτελέσματα Certainty

Ως απόκριση $y = f_i(x)$ ενός συστήματος μάθησης μπορεί να θεωρηθεί η πιθανότητα ενός δεδομένου xνα ανήκει σε μια κλάση, Όταν προσεγγίζονται οι ακραίες τιμές απόκρισης 0 και 1, ο σχεδιαστής του συστήματος θα είναι περισσότερο βέβαιος ότι το δεδομένο έχει κατηγοριοποιηθεί σωστά ως μέρος της κλάσης ή όχι.
Η πρακτική εκδοχή της ανωτέρω σκέψης είναι ο προσδιορισμός του certainty c(y) μιας απόκρισης y:

$$c(y) = \begin{cases} y, & if\ y \geq \frac{1}{2} \\ 1 - y, & otherwise \end{cases}$$

Αυτό μπορεί να ερμηνευθεί ως εξής: το certainty αυξάνεται όταν το y απομακρύνεται από τη λιγότερο certain τιμή (0.5). Μια τιμή απόκρισης y_1 είναι λιγότερο certain από μια άλλη y_2 , αν $c(y_1)<c(y_2)$.

Μια άλλη προσέγγιση θα ήταν να επιτρέψουμε στα βάρη να εξαρτώνται αναλογικά από τα certainties των τιμών απόκρισης. Αυτή η παρατήρηση οδηγεί στον ορισμό του dynamically averaged network (DAN):

$$f_{DAN} = \sum_{i=1}^{N} \omega_i f_i(x)$$

Όπου τα ω_i είναι:

$$\omega_i = \frac{c(f_i(x))}{\sum_{j=1}^{n} c(f_j(x))}$$

f_{DAN} είναι ένας σταθμικός μέσος των τιμών απόκρισης του δικτύου [50].

Πίνακας 5.5 Ποσοστά AUC του συστήματός μας (με certainty) για ισορροπημένα δεδομένα ανά εκδοχή

AUC
0.875
0.375
0.500
0.650
0.875
0.375
0.375
0.938
0.438
0.700

Το μέσο ποσοστό AUC είναι προσεγγιστικά ίσο με 0.61, ενώ η μέση τυπική απόκλιση περίπου 0.23.

Τα αποτελέσματά μας μπορούν επίσης να φανούν συγκεντρωτικά στους συγκριτικούς Πίνακες 5.6 και 5.7.

Πίνακας 5.6 Συγκριτικός Πίνακας για τα ισορροπημένα δεδομένα

	Προτεινόμενο Σύστημα (Ισορροπημένα)	TreeBagger(Ισορροπημένα)	Προτεινόμενο Σύστημα με Certainty (Ισορροπημένα)
Μέση AUC	0.60 ± 0.22	0.63 ± 0.21	0.61 ± 0.23
Μέσο Sensitivity	1	1	1
Μέσο Specificity	0	0.1	0
Μέσο Accuracy	0.5	0.5	0.5

Πίνακας 5.7 Συγκριτικός Πίνακας για τα μη ισορροπημένα δεδομένα

	Προτεινόμενο Σύστημα (Μη ισορροπημένα)	TreeBagger (Μη ισορροπημένα)
Μέση AUC	0.59 ± 0.18	0.63 ± 0.21
Μέσο Sensitivity	0.1	0.29
Μέσο Specificity	1	0.7
Μέσο Accuracy	0.93	0.51

Κεφάλαιο 6 Επίλογος

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία επιχειρήσαμε τη συσχέτιση του ΣΔΤ2 με τον κίνδυνο εμφάνισης καρδιαγγειακής νόσου, χρησιμοποιώντας ένα μοντέλο πρόβλεψης βασισμένο σε ΤΔ. Το κίνητρο για την υλοποίηση του μοντέλου ήταν η διαπίστωσή στην οποία καταλήξαμε μετά από εκτενή βιβλιογραφική επισκόπηση, ότι δηλαδή απαιτούνται προηγμένες τεχνικές ανάπτυξης συστημάτων μεγαλύτερης ακρίβειας για την αντιμετώπιση του προβλήματος που πραγματευόμαστε. Στη συνέχεια αναλύονται τα συμπεράσματά μας από τη μελέτη των αποτελεσμάτων του μοντέλου που δημιουργήσαμε.

Τα αποτελέσματα που πήραμε από την περίπτωση των μη ισορροπημένων δεδομένων δείχνουν ότι θα έπρεπε να απορρίψουμε τη μηδενική υπόθεση, καθώς το p-value προέκυψε πολύ χαμηλό. Ωστόσο, όταν τα δεδομένα είναι ισορροπημένα, αυτό το πρόβλημα παύει να ισχύει. Τα p-values αυτής της περίπτωσης δείχνουν ότι το calibration του συστήματος είναι ικανοποιητικό. Αυτό το γεγονός οδηγεί στο συμπέρασμα ότι το πρώτο σύστημα επιδέχεται περαιτέρω βελτίωση, η οποία μπορεί να επιτευχθεί με τη χρήση τεχνικών που να χειρίζονται τα μη ισορροπημένα δεδομένα.

Αν στρέψουμε την προσοχή μας στα ισορροπημένα δεδομένα ξανά, μπορούμε να επιβεβαιώσουμε ότι οι AUC τιμές έχουν μια πολύ υψηλή μεταβλητότητα για τις 10 εκδοχές. Το γεγονός ότι το σύστημα αποφασίζει με μεγάλη βεβαιότητα για κάποιες εκδοχές ενώ δεν το κάνει για άλλα μας οδηγεί στο συμπέρασμα ότι τα Random Forests ως μέθοδος είναι ασύμβατα με τη δομή του δοθέντος συνόλου δεδομένων. Αντίθετα, αν είχαμε παρατηρήσει ομοιομορφία στα ποσοστά AUC (ακόμα και για χαμηλές τιμές), θα μπορούσαμε να συμπεράνουμε ότι είναι θέμα απόδοσης και όχι ασυνέπειας του συστήματος.

Προκειμένου να αξιολογήσουμε τη μέθοδό μας, προσπαθήσαμε να τη μιμηθούμε, χρησιμοποιώντας τα κατάλληλα εργαλεία του Matlab. Αν είχαμε παρατηρήσει μια αξιοσημείωτη διαφορά στη συμπεριφορά του τελευταίου συστήματος, θα είχαμε καταλήξει στο συμπέρασμα ότι η μέθοδός μας έχει αναπτυχθεί εσφαλμένα, καθώς θα ήταν λάθος να αμφισβητήσουμε την ορθότητα του εργαλείου. Ωστόσο, τα αποτελέσματα των συστημάτων μας και αυτά που προέρχονται από την κλάση TreeBagger του Matlab ήταν σχεδόν πανομοιότυπα. Αυτός είναι ένας ακόμη λόγος για τον οποίο μπορούμε να πούμε με βεβαιότητα ότι υπάρχει ζήτημα ασυμβατότητας.

Εκτός από τα μέτρα AUC και p-value, επιλέξαμε να υπολογίσουμε τα μεγέθη Sensitivity, Specificity και Accuracy. Οι πρώτοι δύο όροι είναι οι λόγοι αληθώς θετικών και αληθώς αρνητικών δειγμάτων, ενώ το Accuracy είναι το άθροισμά τους στο γενικό πληθυσμό [51]. Το πολύ χαμηλό Sensitivity που παρατηρείται στα μη ισορροπημένα δεδομένα δηλώνει την τάση του μοντέλου να αναγνωρίζει ελάχιστα δεδομένα ως στοιχεία της θετικής κλάσης, το οποίο εν μέρει οφείλεται στην ύπαρξη πολύ λίγων αντίστοιχων δεδομένων κατά την εκπαίδευση (το σύστημα είναι biased). Για τον ίδιο λόγο το Specificity είναι υψηλό, αφού το μοντέλο έχει την τάση να αναγνωρίζει τα δεδομένα που εισέρχονται ως στοιχεία της αρνητικής κλάσης. Ένα αρνητικό στοιχείο επομένως πράγματι θα ταξινομηθεί ως αρνητικό λόγω της προαναφερθείσας τάσης, αυξάνοντας έτσι και το ρυθμό ψευδώς αρνητικών δειγμάτων. Από την άλλη πλευρά, στα ισορροπημένα δεδομένα παρατηρούμε την αντίθετη τάση, η οποία όμως δεν μπορεί να οφείλεται πλέον στην έλλειψη επαρκών θετικών δεδομένων. Πιθανά αίτια αυτής της συμπεριφοράς είναι η ιδιαιτερότητα του συγκεκριμένου πληθυσμού και η καταλληλότητα των χαρακτηριστικών. Το μέγεθος Accuracy αναφέρεται στη γενικότερη ακρίβεια του συστήματος, και σε συνδυασμό με την πληροφορία από τα άλλα δύο μεγέθη που το παράγουν δεν παίρνει ικανοποιητικές τιμές.

Κλείνοντας, αποφασίσαμε να ακολουθήσουμε μια διαφορετική στρατηγική βελτίωσης απόδοσης, προκειμένου να ενισχύσουμε την πεποίθησή μας ότι τα ΤΔ ως προσέγγιση και το διαθέσιμο σύνολο δεδομένων είναι ασύμβατα. Αν ίσχυε το αντίθετο, θα είχαμε παρατηρήσει μια διαφορά στα μέτρα ποιότητας, τα οποία θα είχαν ληφθεί υπόψη. Ωστόσο, καμία μεγάλη βελτίωση δεν παρατηρήθηκε, δικαιολογώντας τη βεβαιότητά μας για την αρχική μας υπόθεση.

6.2 Μελλοντική Έρευνα

Από τα ποσοστά θνητότητας και την υποβάθμιση της ποιότητας ζωής εξαιτίας των συνεπειών του ΣΔ, είναι αντιληπτό πως θα πρέπει να δοθεί η αρμόζουσα προσοχή στη διαχείριση της νόσου [52] [53].
Ειδικά για την επιστημονική περιοχή που μας απασχόλησε, συνοψίζονται ακολούθως οι μελλοντικές επεκτάσεις της παρούσας διπλωματικής εργασίας:

- Το μοντέλο που αναπτύξαμε θα μπορούσε να δοκιμαστεί σε άλλους πληθυσμούς και να καταγραφεί η απόδοσή του όταν αυτοί χρησιμοποιούνται ως είσοδος. Με τον τρόπο αυτό θα μπορούσε να διαφανεί πιο ορθά η σχέση μεταξύ της εισόδου του συστήματος και της εξόδου, αφού το σύστημα θα παρέμενε το ίδιο και θα διαπιστωνόταν η απόκρισή του για διάφορες εισόδους. Ακόμη, κάποιος θα μπορούσε να εξάγει συμπεράσματα για τα χαρακτηριστικά που έχουν επιλεγεί αν παρατηρούσε τις εξόδους του συστήματος για διαφορετικά δεδομένα εισόδου.

Η συλλογή δεδομένων θα μπορούσε να γίνει με τη χρήση πλαισίων συγκέντρωσης πληροφορίας [54].

- Μια πολύ ενδιαφέρουσα επέκταση στην αντιμετώπιση του προβλήματος θα ήταν η αναζήτηση πιο ισορροπημένων συνόλων εκπαίδευσης, προκειμένου οι πληροφορίες που παρέχονται και από την κλάση των ατόμων που δεν εμφάνισαν καρδιαγγειακή νόσο να είναι επαρκείς, όπως και αυτές που προέρχονται από τα άτομα που δε νόσησαν.

- Η έλλειψη της ισορροπίας δεδομένων θα μπορούσε να αντιμετωπιστεί και με την αναζήτηση τεχνικών που να περιορίζουν αυτή την αδυναμία, χωρίς να είναι απαραίτητο να βρεθούν πιο ισορροπημένοι πληθυσμοί. Πρόκειται για μεθόδους δίνουν στο αναπτυσσόμενο σύστημα την «εντύπωση» της επεξεργασίας ισορροπημένων δεδομένων.

- Οι μικροαγγειακές επιπλοκές των επικρατέστερων τύπων ΣΔ (ΣΔΤ1 και ΣΔΤ2) αποτελούν έναν παράγοντα που δε θα πρέπει να αγνοηθεί, δεδομένου ότι η μελέτη τους ενδέχεται να προσφέρει πληροφορίες για τη συσχέτιση των εν λόγω επιπλοκών με την καρδιαγγειακή νόσο. Ουσιαστικά αναφερόμαστε στη μελέτη του ενδιάμεσου βήματος μεταξύ ΣΔ και εμφάνισης μακροαγγειακής επιπλοκής, όπως η αμφιβληστροειδοπάθεια [55] (όπως υπαγορεύουν τα στάδια επιδείνωσης της νόσου από άποψη φυσιολογίας), προκειμένου να διερευνηθεί αν η πρόβλεψη θα μπορούσε να είναι πιο ακριβής εφόσον πραγματοποιείται σε βήματα.

- Τέλος, όπως η τεχνική που χρησιμοποιήσαμε δεν είχε αξιοποιηθεί έως τώρα για την πρόβλεψη καρδιαγγειακού κινδύνου εξαιτίας του ΣΔΤ2, θα μπορούσαν να χρησιμοποιηθούν και άλλες γνωστές μέθοδοι που δεν έχουν δοκιμαστεί ακόμη από τους ερευνητές σε διεθνές επίπεδο για το συγκεκριμένο πρόβλημα.

Κεφάλαιο 7. Βιβλιογραφία

[1] World Health Organization

[2] American Diabetes Association

[3] Intelligent Systems of Personalized Medical Decisions Support for the management of Diabetes Mellitus, Konstantia Zarkogianni

[4] Zarkogianni K, Mitsis K, Litsa E, Arredondo MT, Fico G, Fioravanti A, Nikita KS. Comparative assessment of glucose prediction models for Patients with Type 1 Diabetes Mellitus applying sensors for glucose and physical activity monitoring. Medical & Biological Engineering & Computing. In Press, 2015

[5] Mougiakakou, S.G.; Bartsocas, C.S.; Bozas, E.; Chaniotakis, N.; Iliopoulou, D.; Kouris, I.; Pavlopoulos, S.; Prountzou, A.; Skevofilakas, M.; Tsoukalis, A.; Varotsis, K.; Vazeou, A.; Zarkogianni, K.; Nikita, K.S., "SMARTDIAB: A Communication and Information Technology Approach for the Intelligent Monitoring, Management and Follow-up of Type 1 Diabetes Patients," in Information Technology in Biomedicine, IEEE Transactions on , vol.14, no.3, pp.622-633, May 2010

[6] Zarkogianni, K.; Vazeou, A.; Mougiakakou, S.G.; Prountzou, A.; Nikita, K.S., "An Insulin Infusion Advisory System Based on Autotuning Nonlinear Model-Predictive Control," in IEEE Transactions on Biomedical Engineering, vol.58, no.9, pp.2467-2477, Sept. 2011

[7] IDF Diabetes Atlas, International Diabetes Federation, 6th Edition, 2013

[8] http://www.mayoclinic.org/

[9] Oxford Handbook of Endocrinology and Diabetes, Helen E. Turner, John A. H. Wass, 2nd edition, 2009

[10] L. Rydén, P. Grant, S. Anker, C. Berne, F. Consentino, N. Danchin, et. al.,“Diabetes, Pre-Diabetes and Cardiovascular Diseases developed with the EASD,”*Eur Heart J.*, vol. 34, pp. 3035-87, 2013.

[11] L .Ryden, E. Standl, M. Bartnik, et al,"Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC); European Association for the Study of Diabetes (EASD). Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD),"*Eur Heart J*, vol. 28, pp. 88-136, 2007.

[12] International Diabetes Federation. IDF Clinical Guidelines Task Force. Global Guideline for Type 2 Diabetes. http://www.idf.org/webdata/docs/IDF%20GGT2D.pdf

[13] RB D'Agostino, RS Vasan, MJ Pencina, et al. "General cardiovascular risk profile for use in primary care: The Framingham heart study,"*Circulation*, vol. 117, pp. 745-753, 2008.

[14] Cox DR. Regression models and life tables. *J Royal Stat Soc*. 1972; 34(series B):187–220.

[15] B. Balkau, G. Hu, Q. Qiao, et al. "Prediction of the risk of cardiovascular mortality using a score that includes glucose as a risk factor. The DECODE Study,"*Diabetologia*, vol. 47, pp. 2118-2128, 2004.

[16] RJ Stevens, RL Coleman, AI Adler, IM Stratton, DR Matthews, RR Holman, "Risk factors for myocardial infarction case fatality and stroke case fatality in type 2 diabetes: UKPDS 66,"*Diabetes Care*, vol. 27, pp. 201-207, 2004.

[17] R. Coleman, R. Stevens, R.Retnakaran, R. Holman, "Framingham, SCORE, and DECODE Risk Equations Do Not Provide Reliable Cardiovascular Risk Estimates in Type 2 Diabetes,"*Diabetes Care*, vol. 30 no. 5, pp. 1292-1293, 2007.

[18] G. Collins, S. Mallett, O. Omar, and L. Yu, "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting," BMC Medicine, vol. 9, vol. 1-14, 2011.

[19] S. van Dieren, J. Beulens, A.Kengne, L.Peelen, G. Rutten, M. Woodward, Y. van der Schouw, K. Moons, "Prediction Models for the Risk of Cardiovascular Disease in Patients With Type 2 Diabetes,"*Heart*., vol. 98, no. 5, pp. 360-369, 2012.

[20] Machine Learning, Wikipedia, the free Encyclopedia.

[21] Neural Networks and Learning Machines, Simon Haykin, 3rd Edition, 2009

[22] A Neural Network based Approach for the Diabetes Risk Estimation, DeeptiJain, Divakarsingh, International Journal of Computer Applications, Vol. 73– No.10, July 2013

[23] Data Mining: A Knowledge Discovery Approach, Krzysztof J. Cios, WitoldPedrycz, Roman W. Swiniarski, Lukasz A. Kurgan, 2007

[24] A study on effective data mining association rules for heart disease prediction system, R.Thanigaivel, Dr. K.Ramesh Kumar, International Journal of Data Engineering (IJDE), Singaporean Journal of Scientific Research(SJSR), Vol.7.No.1 2015 Pp.371-379

[25] Support Vector Machines, Wikipedia, The free encyclopedia

[26] Support Vector Machines, Ingo Steinwart,Andreas Christmann, 2008

[27] Diagnosing Heart Diseases for Type 2 Diabetic Patients by Cascading the Data Mining Techniques, P. Radha, Tamil Nadu, Dr. B. Srinivasan, International Journal on Recent and Innovation Trends in Computing and Communication , Volume: 2 Issue: 8

[28] Review of "Inductive Logic Programming: Techniques and Applications"*, Nada Lavrač, SašoDžeroski, Machine Learning, 23, 103-108, (1996)*

[29] Inductive Logic Programming: Techniques and Applications, Nada Lavrač, SašoDžeroski, 1994

[30] Clustering, ruixu, donald c. Wunsch, II, 2009

[31] Cardiovascular risk factors clustering with endogenous hyperinsulinaemia predict death from coronary heart disease in patients with Type II diabetes, S.Lehto1, T.Rönnemaa2, K.Pyörälä1, M.Laakso1, Diabetologia (2000) 43: 148±155

[32] An introduction to genetic algorithms for scientists and engineers, David A. Coley, 1999

[33] Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, M.Akhiljabbar, B.L Deekshatulua, Priti Chandra B, International Conference on Computational Intelligence: Modeling Techniques and Applications, (CIMTA) 2013

[34] Ensemble Machine Learning, Methods and Applications, Cha Zhang, Yunqian Ma, 2012

[35] Ensemble Methods: Foundations and Algorithms, Ralf Herbrich, ThoreGraepel, 2012

[36] The elements of statistical learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd Edition, 2008

[37] Bagging Predictors, Leo Breiman, Technical Report No. 421, Department of Statistics, University of California, Berkeley, California, 1994.

[38] Learning on Complex Simulations, Robert E. Banfield, a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Department of Computer Science and Engineering, College of Engineering, University of South Florida, 2007

[39] Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics, Anne-Laure Boulesteix, SilkeJanitza, JochenKruppa, Inke R. König, Technical Report Number 129, Department of Statistics, University of Munich, 2012

[40] Decision trees: a recent overview, S. B. Kotsiantis, ArtifIntell Rev (2013) 39:261–283, 2011

[41] Classification and regression trees, Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, 1984

[42] Overview of Use of Decision Tree algorithms in Machine Learning, Arundhati Navada, AamirNizam Ansari, SiddharthPatil, BalwantA.Sonkamble, IEEE Control and System Graduate Research Colloquium, 2011

[43] Consistent Probability Estimation Using Nonparametric Learning Machines, J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, A. Ziegler, Schattauer, 2012

[44] Random Forests, LEO BREIMAN, Machine Learning, 45, 5–32, 2001

[45] Classification and Regression by RandomForest, Andy Liaw, Matthew Wiener, R News, 2002

[46] Mining data with random forests: Current options for real-world applications, Andreas Ziegler,Inke R. König, WIREs Data Mining KnowlDiscov2014, 4:55–63

[47] Data mining for imbalanced datasets: an overview, Nitesh V. Chawla, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Springer, 853-867

[48] Integrated Oversampling for Imbalanced Time Series Classification, Hong CAO, Xiao-Li LI, Yew-Kwong WOON and See-Kiong NG, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[49] Assessing the performance of prediction models: a framework for some traditional and novel measures, Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat, Gonen, Nancy Obuchowski, Michael J. Pencina, Michael W. Kattan, Epidemiology. 2010 January ; 21(1): 128–138.

[50] Dynamically Weighted Ensemble Neural Networks for Classification, Daniel Jiménez, The University of Texas Health Science Center at San Antonio, Department of Rehabilitation Medicine.

[51] Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations Wen Zhu , Nancy Zeng , Ning Wang, K&L consulting services, Inc, Fort Washington, PA Octagon Research Solutions, Wayne, PA, NESUG 2010.

[52] Konstantia Zarkogianni, Konstantina Nikita, “Personal Health Systems for Diabetes Management, Early Diagnosis and Prevention”, Handbook of Research on Trends in the Diagnosis and Treatment of Chronic Conditions, IGI Global book series Advances in Medical Diagnosis, 2015

[53] K. Zarkogianni, E. Litsa, K. Mitsis, P. Wu, C. D. Kaddi, C. Cheng, M.D. Wang, Senior Member, IEEE, and K.S. Nikita, “A Review of EmergingTechnologies for the Management of Diabetes Mellitus”, in IEEE Transactions on Biomedical Engineering, in Press, 2015

[54] Dagliati, A.; Sacchi, L.; Bucalo, M.; Segagni, D.; Zarkogianni, K.; Martinez Millana, A.; Cancela, J.; Sambo, F.; Fico, G.; Meneu Barreira, M.T.; Cerra, C.; Nikita, K.; Cobelli, C.; Chiovato, L.; Arredondo, M.T.; Bellazzi, R., "A data gathering framework to collect Type 2 diabetes patients data," in Biomedical

and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on , vol., no., pp.244-247, 1-4 June 2014

[55] Skevofilakas, M.; Zarkogianni, K.; Karamanos, B.G.; Nikita, K.S., "A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus," in Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE , vol., no., pp.6713-6716, Aug. 31 2010-Sept. 4 2010