ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διάγνωση καρδιακών ασθενειών με μηχανική μάθηση

Διπλωματική Εργασία

Μητσέας Νικόλαος

Επιβλέπων: Σταμούλης Γεώργιος

Ιούλιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

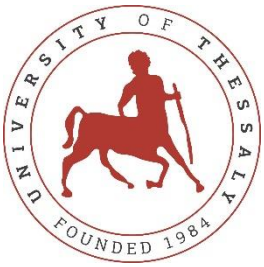
Διάγνωση καρδιακών ασθενειών με μηχανική μάθηση

Διπλωματική Εργασία

Μητσέας Νικόλαος

Επιβλέπων: Σταμούλης Γεώργιος

Ιούλιος 2022



UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Diagnosis of heart disease by machine learning

Diploma Thesis

Mitseas Nikolaos

Supervisor: Stamoulis Georgios

July 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων	**Σταμούλης Γεώργιος** Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Κολομβάτσος Κωσταντίνος** Επίκουρος Καθηγητής, Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας
Μέλος	

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελούν αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλουν οποιασδήποτε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχουν έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Μέσα στην εργασία με την χρήση κατάλληλων παραπομπών δηλώνονται ευδιάκριτα όποιες ιδέες που έχω αξιοποιήσει καθώς και οποιαδήποτε άλλα αρχεία, πήγες, κείμενα ή και έργα άλλων συγγραφέων. Στο τέλος της εργασίας στο τμήμα των βιβλιογραφικών αναφορών περιλαμβάνονται και οι σχετικές αναφορές των σημείων αυτών και αποδίδονται με πλήρη περιγραφή. Συγχρόνως δηλώνω με υπευθυνότητα ότι τα παρόντα αποτελέσματα δεν έχουν ή δεν προορίζονται να χρησιμοποιηθούν για την απόκτηση κάποιου άλλου προπτυχιακού ή μεταπτυχιακού τίτλου. Στην περίπτωση που στο μέλλον αποδειχθεί πως το παρόν κείμενο είναι προϊόν λογοκλοπής και όχι δικής μου αποκλειστικής προσπάθειας αναλαμβάνω υπεύθυνα και απολύτως όλες τις συνέπειες που μπορεί να προκύψουν σε διοικητικό και νομικό επίπεδο.

Ο Δηλών

Μητσέας Νικόλαος

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

The Declarant

Mitseas Nikolaos

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογένειά μου , η οποία με στήριξε πάρα πολύ καθ’ όλη τη διάρκεια των σπουδών μου .

Διπλωματική Εργασία

Διάγνωση καρδιακών ασθενειών με μηχανική μάθηση

Μητσέας Νικόλαος

Περίληψη

Η εργασία αυτή έχει ως θέμα τη μελέτη της διάγνωσης των καρδιακών ασθενειών με τη μέθοδο της μηχανικής μάθησης. Μελετώνται βασικά ζητήματα της καρδιολογίας και της μηχανικής μάθησης όπως η επιβλεπόμενη μάθηση, η κατηγοριοποίηση, η παλινδρόμηση, ο μετασχηματισμός και η συσταδοποίηση. Επίσης μελετώνται οι αλγόριθμοι κατηγοριοποίησης και οι αλγόριθμοι συσταδοποίησης. Στόχος της διπλωματικής εργασίας είναι η εφαρμογή αλγορίθμων classification πάνω σε ιατρικά δεδομένα, και πιο συγκεκριμένα πάνω σε δεδομένα τα οποία σχετίζονται με την ύπαρξη εγκεφαλικού επεισοδίου. Οι αλγόριθμοι οι οποίοι χρησιμοποιήθηκαν είναι οι Random Forests, Extra Trees, kNN και SVM. Η υλοποίηση έγινε σε γλώσσα Python. Το περιβάλλον το οποίο χρησιμοποιήθηκε για την συγγραφή του κώδικα είναι το Anaconda Spyder, ενώ οι βιβλιοθήκες οι οποίες βοήθησαν στην ολοκλήρωση της πειραματικής διαδικασίας είναι οι pandas, numpy, sklearn, seaborn και matplotlib. Τα αποτελέσματα οδήγησαν στο συμπέρασμα ότι οι αλγόριθμοι Random Forests και Extra Trees παρουσιάζουν καλύτερα αποτελέσματα σε σχέση τους αλγόριθμους kNN και SVM, ενώ ο kNN παρουσιάζει υψηλότερο precision σε σχέση με το μοντέλο SVM.

Λέξεις-κλειδιά:

Διάγνωση, καρδιακές ασθένειες, μηχανική μάθηση, αλγόριθμοι, κατηγοριοποίηση, συσταδοποίηση.

Diploma Thesis

Diagnosis of heart disease by machine learning

Mitseas Nikolaos

Abstract

The aim of this study is to study the diagnosis of heart disease by the method of machine learning. Basic issues of cardiology and machine learning such as supervised learning, categorization, regression, transformation and clustering are studied. Categorization algorithms and clustering algorithms are also studied. The aim of the dissertation is the application of classification algorithms on medical data, and more specifically on data related to the existence of a stroke. The algorithms used are Random Forests, Extra Trees, kNN and SVM. Implemented in Python. The environment used to write the code is Anaconda Spyder, while the libraries that helped complete the experimental process are pandas, numpy, sklearn, seaborn and matplotlib. The results led to the conclusion that the Random Forests and Extra Trees algorithms show better results than the kNN and SVM algorithms, while the kNN has a higher precision than the SVM model.

Keywords:

Diagnosis, heart disease, machine learning, algorithms, categorization, clustering.

Πίνακας περιεχομένων

Κατάλογος εικόνων

Κατάλογος Πινάκων

1. Καρδιολογία

1.1. Βασικές Έννοιες Καρδιολογίας

Με τον όρο καρδιολογία αναφερόμαστε στον επιστημονικό εκείνο κλάδο όποιος ασχολείται με τη μελέτη του καρδιαγγειακού συστήματος της καρδιάς και των αιμοφόρων αγγείων καθώς και των παθήσεων που αφορούν τα συστήματα αυτά. Η καρδιολογία αποτελεί κλάδο της εσωτερικής ιατρικής. Ο αρμόδιος γιατρός ο οποίος ασχολείται με τέτοια ζητήματα ονομάζεται καρδιολόγος και επομένως ένα άτομο με καρδιοπάθεια ή όποια άλλη σχετική ασθένεια απευθύνεται σε τέτοιο γιατρό. Έτσι οποιοσδήποτε γιατρός μπορεί να παραπέμψει τον ασθενή του σε καρδιολόγο εφόσον εμφανίσει κάποιο σύμπτωμα καρδιοπάθειας. Υπάρχει διάκριση μεταξύ καρδιολόγου και καρδιοχειρουργού καθώς ο δεύτερος ασχολείται με εγχειρήσεις καρδιάς.

Αντιθέτως ένας καρδιολόγος δεν ειδικεύεται σε θέματα εγχειρήσεων αλλά αντίθετα σε θέματα διάγνωσης και θεραπείας παθήσεων του συστήματος της καρδιάς και των αγγείων. Η διάγνωση πραγματοποιείται μέσα από εξετάσεις με χαρακτηριστικά παραδείγματα να αποτελούν αυτά της τοποθέτησης βηματοδότη ή της αγγειοπλαστικής.

Οι καρδιακές παθήσεις σχετίζονται ειδικά με την καρδιά, ενώ οι καρδιαγγειακές παθήσεις επηρεάζουν την καρδιά, τα αιμοφόρα αγγεία ή και τα δύο Συμπτώματα όπως οι πόνοι στο στήθος, υψηλή αρτηριακή πίεση, δύσπνοια ή ζαλάδα είναι μερικά από τα πιο χαρακτηριστικά συμπτώματα τα οποία μπορεί να αποτελούν ένδειξη κάποιας πάθησης του καρδιαγγειακού συστήματος.

Επιπρόσθετα στις αρμοδιότητες του καρδιολόγου περιλαμβάνονται εξετάσεις που αφορούν μη φυσιολογικό καρδιακό ρυθμό ή κάποιο καρδιακό φύσημα. Ασθένειες τις οποίες καλούνται να αντιμετωπίσουν μπορεί να είναι η καρδιακή ανεπάρκεια ή η καρδιακή προσβολή. Συνεργάζεται με τον καρδιοχειρουργό για την απόφαση εγχειρήσεις καρδιάς αλλά και άλλες παρεμβατικές ενέργειες όπως το stening ή η αγγειοπλαστική.

Πιο συγκεκριμένα οι παθήσεις στις οποίες αναλαμβάνει να θεραπεύσει ένας καρδιολόγος είναι η υπέρταση, η κοιλιακή ταχυκαρδία, η υψηλή χοληστερόλη και τα τριγλυκερίδια, η υψηλή αρτηριακή πίεση, η περικαρδίτιδα, οι αρρυθμίες, η κολπική μαρμαρυγή, η αθηροσκλήρωση, η στεφανιαία νόσος και η συμφορητική καρδιακή νόσος.

Συγχρόνως είναι σημαντικό να σημειωθεί ότι ένας άνθρωπος δεν χρειάζεται να φέρει συμπτώματα καρδιακής πάθησης για να επισκεφτεί έναν καρδιολόγο. Παράγοντες όπως η κληρονομικότητα, τα υψηλά επίπεδα χοληστερόλης, ο διαβήτης, η καθιστική ζωή, το κάπνισμα ή η επιθυμία για την έναρξη κάποιας αθλητικής δραστηριότητας αποτελούν παράγοντες παραγωγής σε κάποιον καρδιολόγο.

Προκειμένου ένας καρδιολόγος να ειδικευτεί σε θέματα μελέτης των καρδιακών παθήσεων ενηλίκων και παιδιών χρειάζεται να ακολουθήσει εξειδικευμένες πανεπιστημιακές σπουδές διαφορετικού χαρακτήρα. Ως εκ τούτου, ένας καρδιολόγος ενηλίκων (συχνά αποκαλούμενος απλώς «καρδιολόγος») είναι ανεπαρκώς εκπαιδευμένος για τη φροντίδα των παιδιών και οι παιδοκαρδιολόγοι δεν είναι εκπαιδευμένοι για τη φροντίδα της καρδιακής νόσου των ενηλίκων (Constantine, Shan, Flamm & Sivananthan, 2004).

Όπως αναφέρθηκε και παραπάνω ο καρδιολόγος επιτελεί διαφορετικά καθήκοντα από τον καρδιοχειρουργό. Ο δεύτερος αναλαμβάνει όλες τις χειρουργικές επεμβατικές διαδικασίες όπως για παράδειγμα η καρδιοπνευμονική παράκαμψη ή η αντικατάσταση μιτροειδούς βαλβίδας. Ωστόσο, η τοποθέτηση στεντ και βηματοδοτών γίνεται από καρδιολόγους.

Ως καρδιακή ηλεκτροφυσιολογία ορίζεται εκείνος ο κλάδος της επιστήμης ο οποίος ασχολείται με την διάγνωση και την θεραπεία προβλημάτων που σχετίζονται με την ηλεκτρική δραστηριότητα του καρδιαγγειακού συστήματος. Ο όρος χρησιμοποιείται συνήθως για να περιγράψει μελέτες τέτοιων φαινομένων με καταγραφή της αυθόρμητης δραστηριότητας με επεμβατικό (ενδοκαρδιακό) καθετήρα, καθώς και καρδιακές αποκρίσεις σε προγραμματισμένη ηλεκτρική διέγερση (PES). Οι παραπάνω μελέτες σχετίζονται με αρκετές διαδικασίες όπως τον εντοπισμό συμπτωμάτων των παθήσεων της καρδιάς, την αξιολόγηση και μελέτη ηλεκτροκαρδιογραφημάτων με ύποπτα ευρήματα, την αξιολόγηση του κινδύνου εμφανίσεις καρδιαγγειακών προβλημάτων όπως αρρυθμίες ενώ ταυτόχρονα αξιοποιούνται και στον τομέα της θεραπείας. Συγχρόνως περιλαμβάνουν και μία σειρά από διάφορες θεραπευτικές μεθόδους όπως είναι για παράδειγμα η κρυοκατάλυση, η χρήση βηματοδότη η εμφύτευση αυτόματων απινιδωτών καρδιά αναγωγής καθώς και η θεραπεία ασθενών με φάρμακα αντιαρρυθμικά. (Cheng, 2004).

Μέσω της μελέτης της καρδιακής ηλεκτροφυσιολογίας μετράται η απόκριση του μυοκαρδίου το οποίο φέρει βλάβες σε συγκεκριμένα φάρμακα. Σκοπός της αξιολόγησης αυτής είναι να φανεί εάν το συγκεκριμένο φαρμακευτικό σχήμα είναι αποτελεσματικό στη μείωση της η πιθανότητας εμφάνισης προβλημάτων όπως η κοιλιακή μαρμαρυγή η παρατεταμένη κοιλιακή ταχυκαρδία. Για τον σχεδιασμό ενός αποτελεσματικού θεραπευτικού πλάνου πολλές φορές χρειάζεται να γίνουνε διάφορες δοκιμές φαρμάκων ώστε καρδιολόγος να καταλήξει το φαρμακευτικό εκείνο σχήμα που ταιριάζει καλύτερα στις ανάγκες του εκάστοτε ασθενή και το

οποίο θα αποτρέπει την ανάπτυξη VF η VT. Οι αξιολογήσεις αυτές χρησιμοποιούνται επίσης και στην περίπτωση τοποθέτησης Ή αντικατάστασης του καρδιακού βηματοδότη. (Cheng, 2004).

Στον κλάδο της καρδιολογίας περιλαμβάνεται και η κλινική καρδιακή ηλεκτροφυσιολογία. Αντικείμενο του κλάδου αυτού είναι η διάγνωση και η θεραπεία διαταραχών που σχετίζονται με το ρυθμό της καρδιάς. Οι καρδιολόγοι με εξειδίκευση σε αυτόν τον τομέα αναφέρονται συνήθως ως ηλεκτροφυσιολόγοι. Οι ηλεκτροφυσιολόγοι εκπαιδεύονται στον μηχανισμό, τη λειτουργία και την απόδοση των ηλεκτρικών δραστηριοτήτων της καρδιάςΑντικείμενο μελέτης και ενασχόλησης των ηλεκτροφυσιολόγων είναι η λειτουργία του μηχανισμού της ηλεκτρικής δραστηριότητας του καρδιαγγειακού συστήματος. Οι επαγγελματίες αυτοί συνεργάζονται με τους καρδιολόγους αλλά και τους καρδιοχειρουργούς σχετικά με ζητήματα που αφορούν διαταραχές του ρυθμού της καρδιάς όπως οι αρρυθμίες. Είναι εκπαιδευμένοι να εκτελούν επεμβατικές και χειρουργικές επεμβάσεις για τη θεραπεία της καρδιακής αρρυθμίας (Garcia, Faber, Cooke, Folks, Chen & Santana (2007).

Ο επιστημονικός κλάδος ο οποίος καταπιάνεται με τις διαταραχές του καρδιαγγειακού συστήματος σε άτομα προχωρημένης ηλικίας ονομάζεται καρδιογηριατρική η γηριατρική καρδιολογία.

Ένας σημαντικός αριθμός ηλικιωμένων συχνά χάνει τη ζωή του από παθήσεις του καρδιαγγειακού συστήματος. Μερικές από τις πιο συνήθεις διαταραχές του καρδιαγγειακού συστήματος που σχετίζεται με τη θνησιμότητα της Τρίτης ηλικίας είναι τα εμφράγματα του μυοκαρδίου, οι αρρυθμίες, η καρδιακή ανεπάρκεια και η κολπική μαρμαρυγή. Αντίθετα οι διαταραχές όπως η περιφερειακή αρτηριακή νόσος ή η άρθρο σκλήρυνση εμφανίζουν συννοσηρότητα με τις παραπάνω διαταραχές και συμβάλλουν επίσης στα υψηλά επίπεδα θνησιμότητας (Marr & Bowen, 2011).

Η καρδιακή απεικόνιση περιλαμβάνει ηχοκαρδιογραφία (ηχώ), μαγνητική τομογραφία καρδιάς (CMR) και αξονική τομογραφία καρδιάς (CCT). Οι επαγγελματίες εκείνοι που έχουν εκπαιδευτεί στην μέθοδο της καρδιακής απεικόνισης έχουν την επιλογή να εξειδικευτούν σε μία συγκεκριμένη απεικονιστική μέθοδο ή εκπαιδευτούν σε όλους τους διαφορετικούς τρόπους καρδιακής απεικόνισης (Niederer, Lumens & Trayanova, 2019).

Η ηχοκαρδιογραφία χρησιμοποιεί τυπικό δισδιάστατο, τρισδιάστατο και υπερηχογράφημα Doppler για τη δημιουργία εικόνων της καρδιάς. Η ηχοκαρδιογραφία συνήθως απασχολεί ένα αρκετά μεγάλο μέρος της κλινικής αξιολόγησης. Η διαδικασία αξιολόγησης και ανάγνωσης της ηχοκαρδιογραφίας καθώς και της μεθόδου της διοισοφαγικης ήχους αξιοποιούνται σε διαδικασίες όπως για παράδειγμα η εισαγωγή συσκευής απόφραξης του αριστερού προσαρτήματος.

Η καρδιακή μαγνητική τομογραφία χρησιμοποιεί ειδικά πρωτόκολλα για την απεικόνιση της δομής και της λειτουργίας της καρδιάς με συγκεκριμένες αλληλουχίες για ορισμένες ασθένειες όπως η αιμοχρωμάτωση και η αμυλοείδωση.

Για την απεικόνιση της λειτουργίας του οργάνου της καρδιάς χρησιμοποιείται η μέθοδος της αξονικής τομογραφίας. Η μέθοδος αυτή αξιοποιεί συγκεκριμένα ειδικά πρωτόκολλα και χρησιμοποιείται κυρίως για την διάγνωση παθήσεων των στεφανιαίων αρτηριών (Niederer, Lumens & Trayanova, 2019).

Η επεμβατική καρδιολογία είναι ένας κλάδος της καρδιολογίας που ασχολείται ειδικά με τη θεραπεία δομικών καρδιακών παθήσεων με βάση τον καθετήρα (Patton, Slomka, Germano & Berman, 2007 Ο καθετηριασμός αξιοποιείται για την πραγματοποίηση σημαντικών επεμβάσεων του καρδιαγγειακού συστήματος. Συνήθως η διαδικασία αυτή περιλαμβάνει την χρήση μίας θήκης σε κάποια μεγάλη Περιφερειακή φλέβα η αρτηρία όπως η μηριαία αρτηρία. Παράλληλα χρησιμοποιούνται ακτίνες Χ για την διασωλήνωση της καρδιάς υπό οπτικοποίηση.

Η μέθοδος της επεμβατικής καρδιολογίας είναι εκείνη η μέθοδος που προτείνεται για την μετέπειτα ταχύτερη ανάρρωση και φροντίδα ασθενών που πάσχουν από οξύ έμφραγμα του μυοκαρδίου. Συγχρόνως η επεμβατική καρδιολογία ή η ακτινολογική μέθοδος προτιμάται καθώς ο ασθενής εξασφαλίζει έτσι γρηγορότερη ανάρρωση αλλά και λιγότερο πόνο ή σημάδια όπως ουλές. Αυτή η διαδικασία μπορεί επίσης να γίνει προληπτικά, όταν περιοχές του αγγειακού συστήματος αποφραχθούν από την αθηροσκλήρωση.

Σκοπός της διαδικασίας αυτής είναι ο καρδιολόγος να μπορέσει να αποκτήσει πρόσβαση στην καρδιά. Για το λόγο αυτόν μέσα από το αγγειακό σύστημα εισάγει ένα περίβλημα το οποίο έχει έναν πολύ μικρό συρμάτινο σωλήνα γύρω από αυτόν και ένα μπαλόνι. Κατά τη διάρκεια της διαδικασίας εάν ο γιατρός εντοπίσει σημάδια απόφραξης χρησιμοποιεί το μπαλόνι φουσκώνοντας του προκειμένου να συμπιέσει το συγκεκριμένο σημείο εντός των αγγείων. Μετά από την διαδικασία αυτή ο καρδιολόγος προκειμένου να κρατήσει το αγγείο ανοιχτό τοποθετείται ένα stent.

Η εξειδίκευση της γενικής καρδιολογίας σε αυτήν ακριβώς των καρδιομυοπαθειών οδηγεί επίσης στην εξειδίκευση, στη μεταμόσχευση καρδιάς και την πνευμονική υπέρταση. Ένας άλλος σχετικά πρόσφατα αναδυόμενος κλάδος της επιστήμης της καρδιολογίας είναι η καρδιοογκολογία. Στους επιστήμονες του κλάδου αυτού απευθύνονται οι ασθενείς που πάσχουν από καρκίνο και οι οποίοι πρόκειται να εισέλθουν σε σχήματα χημειοθεραπειών ή σε ασθενείς που εμφανίζουν καρδιακές παθήσεις εξαιτίας των θεραπειών αυτών.

1.2. Βασικές Καρδιακές Ασθένειες

Η καρδιά και τα αιμοφόρα αγγεία είναι το δύο όργανα που συναπαρτίζουν το καρδιαγγειακό σύστημα. Μερικές από τις παθήσεις που απασχολούν το καρδιαγγειακό μας σύστημα είναι η ρευματική καρδιοπάθεια ανωμαλίες στο σύστημα αγωγιμότητας ενώ η πιο συχνή κατηγορία παθήσεων είναι αυτή της καρδιαγγειακής νόσου. (Antman& Loscalzo, 2016):

1. Η καρδιαγγειακή νόσος εμφανίζεται μέσα από τέσσερις διαφορετικές διαγνωστικές κατηγορίες.
2. Στεφανιαία νόσος η στεφανιαία καρδιοπάθεια. Η νόσος αυτή προκαλείται κατά κανόνα λόγω μειωμένης αιμάτωσης του μυοκαρδίου. Λόγω της μειωμένης αιμάτωσης μπορεί να προκληθεί έμφραγμα, καρδιακή ανεπάρκεια η στηθάγχη. Η στεφανιαία νόσος είναι η πιο συχνή μορφή καρδιαγγειακής νόσου καθώς το ένα τρίτο ή οι μισές περιπτώσεις καρδιαγγειακών νόσων έγκειται σε αυτή την κατηγορία.
3. Εγκέφαλο αγγειακή νόσος. Στην κατηγορία αυτή συναντάμε παθήσεις όπως η παροδική ισχαιμική προσβολή η το γνωστό εγκεφαλικό επεισόδιο.Περιφερική αρτηριακή νόσος (PAD): Ιδιαίτερα αρτηριακή νόσος που αφορά τα άκρα που μπορεί να οδηγήσει σε χωλότητα.
4. Αθηροσκλήρωση αορτής: Συμπεριλαμβανομένων ανευρυσμάτων θωρακικού και κοιλιακού

Αν και η καρδιαγγειακή νόσος μπορεί να προκύψει άμεσα από διαφορετικές αιτιολογίες, όπως έμβολα σε ασθενή με κολπική μαρμαρυγή που έχει ως αποτέλεσμα ισχαιμικό εγκεφαλικό επεισόδιο, ρευματικό πυρετό που προκαλεί βαλβιδοπάθεια, μεταξύ άλλων, επειδή η αντιμετώπιση των περισσότερων παραγόντων κινδύνου σχετίζεται με την ανάπτυξη της αθηρίτιδας είναι ένας κοινός παρονομαστής στην παθοφυσιολογία της καρδιαγγειακής νόσου.

Υπάρχουν πολλοί παράγοντες που έχουν οδηγήσει σε σταθερά υψηλά ποσοστά καρδιαγγειακών παθήσεων τα τελευταία χρόνια ειδικότερα στον δυτικό κόσμο. Πολλές έρευνες καταδεικνύουν ότι η σύγχρονη κοινωνία με την ανάπτυξη της τεχνολογίας την αύξηση του καταναλωτισμού και την εντατικοποίηση της εργασίας έχουν αλλάξει τις καθημερινές συνήθειες των ανθρώπων. Ο περιορισμένος ελεύθερος χρόνος και μετακίνησης και η ανάδυση νέων δραστηριοτήτων που σχετίζονται με την καθιστική ζωή είναι μερικά από τα χαρακτηριστικά της σύγχρονης της κοινωνίας που συμβάλλουν στην παραπάνω αύξηση τέτοιων παθήσεων. Ο σύγχρονος άνθρωπος τείνει να εργάζεται λιγότερο ενώ ολοένα και περισσότερες θέσεις εργασίας υιοθετούν έναν πιο καθιστικό και όχι ενεργητικό και σωματικά απαιτητικό χαρακτήρα. Συγχρόνως οι διατροφικές συνήθειες διαδραματίζουν πολύ σημαντικό ρόλο στην αύξηση εμφάνισης καρδιαγγειακών νοσημάτων. Η πρόσληψη τροφών με πολλές θερμίδες σάκχαρα και κορεσμένα λίπη οδηγούνε σε μία πληθώρα διαταραχών όπως υπέρταση και ο διαβήτης η αθηροσκλήρωση και το μεταβολικό σύνδρομο. (Ashley & Niebauer, 2004).

Μία πρόσφατη μελέτη που αξιοποίησε το παραπάνω θέμα είναι η μελέτη INTERHEART. Η μελέτη αυτή έλαβε χώρα σε 52 διαφορετικές χώρες Και φρόντισε να περιλαμβάνει πληθυσμούς τόσο χαμηλού αλλά όσο και μεσαίου και υψηλού εισοδήματος. Τα αποτελέσματα της έρευνας κατέληξαν σε εννέα συγκεκριμένους παράγοντες οι οποίοι σχετίζονται με το 90% της εμφάνισης καρδιαγγειακών παθήσεων. Οι παράγοντες αυτοί είναι η σωματική αδράνεια και η καθιστική ζωή, η κατανάλωση αλκοόλ και το κάπνισμα, οι διατροφικές συνήθειες και η μειωμένη κατανάλωση φρούτων και λαχανικών η υπέρταση, ο διαβήτης, η κοιλιακή παχυσαρκία η δυσλιπιδαιμία και τέλος κάποιοι ψυχοκοινωνικοί παράγοντες. (Antman & Loscalzo, 2016).

Ο κίνδυνος εμφράγματος του μυοκαρδίου καταλογίστηκε στο κάπνισμα. Δύο έρευνες μεγάλου βεληνεκούς και πιο συγκεκριμένα η τρίτη Εθνική έρευνα εξέτασης υγείας και διατροφής και η καρδιολογική Μελέτη framingham εντόπισαν στατιστικώς σημαντική συσχέτιση στο κάπνισμα στη δυσανεξία στη γλυκόζη στην αρτηριακή πίεση και στη δυσλιπιδαιμία όσον αφορά την πιθανότητα εμφάνισης παθήσεων του καρδιαγγειακού συστήματος. (Johnson, Torres, Glicksberg, Shameer, Miotto, Ali & Dudley, 2018).

Τα αποτελέσματα από τις παραπάνω μελέτες αναδεικνύουν ότι ένα συντριπτικό ποσοστό της τάξης του 60 έως 90% των περιστατικών πάθησης του καρδιαγγειακού συστήματος εμφανίστηκαν σε ασθενείς που είχαν υιοθετήσει τουλάχιστον έναν από τους παραπάνω παράγοντες κινδύνου. Το σημαντικότερο όμως από όλα είναι πως τα ευρήματα αυτά έχουνε αξιοποιηθεί από την Αμερικανική Καρδιολογική Εταιρεία για την ανάπτυξη προγραμμάτων πρόληψης και προαγωγής της υγείας. Τα προγράμματα αυτά περιλαμβάνουν συγκεκριμένες συστάσεις όπως την αύξηση της σωματικής δραστηριότητας την υιοθέτηση υγιεινής διατροφής τη διατήρηση του σωματικού βάρους της γλυκόζης της χοληστερόλης και της αρτηριακής πίεσης σε φυσιολογικά επίπεδα και την αποφυγή του καπνίσματος..

Αντίθετα είναι τα αποτελέσματα ερευνών για παράγοντες μη ελέγξιμους όπως το φύλο η ηλικία ή το οικογενειακό ιστορικό. (Cheng, 2004). Το οικογενειακό ιστορικό, ιδιαίτερα η πρόωρη αθηροσκληρωτική νόσος που ορίζεται ως καρδιαγγειακή νόσος ή ο θάνατος από καρδιαγγειακή νόσο σε συγγενή πρώτου βαθμού πριν από 55 έτη (σε άνδρες) ή 65 έτη (στις γυναίκες) θεωρείται ανεξάρτητος παράγοντας κινδύνου.

Ερευνητικά δεδομένα επίσης καταδεικνύουν ότι το φύλλο επηρεάζεται διαφορετικά από διαφορετικούς παράγοντες κινδύνου. Για παράδειγμα το κάπνισμα 20 τσιγάρων καθημερινά και η πάθηση του διαβήτη επηρεάζουν πιο σημαντικά τις γυναίκες αναφορικά με τους άνδρες και αυξάνουν τις πιθανότητες εμφάνισης καρδιαγγειακής νόσου (Patton,Slomka, Germano, & Berman, 2007). Ο επιπολασμός της καρδιαγγειακής νόσου αυξάνεται σημαντικά με κάθε δεκαετία της ζωής. Η παρουσία HIV (ιός ανθρώπινης ανοσοανεπάρκειας), (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020) ιστορικό ακτινοβολίας μεσοθωρακίου ή θωρακικού τοιχώματος, μικρολευκωματινουρία, αυξημένοι δείκτες φλεγμονής

έχουν επίσης συσχετιστεί με αυξημένο ποσοστό και επίπτωση καρδιαγγειακής νόσου.

Αξίζει να αναφερθεί ότι υπάρχει μία σημαντική σύγχυση αναφορικά με την σύνδεση συγκεκριμένων ειδών διατροφής και τις συσχετίσεις τους με τις παθήσεις του καρδιαγγειακού συστήματος. Τροφές όπως φυτικές ίνες το κρέας ή ο καφές εμφανίζουν αμφιλεγόμενα αποτελέσματα σε διαφορετικές μελέτες. Συγχρόνως έχει καταγραφεί ότι στις Ηνωμένες πολιτείες της Αμερικής σε μία περίοδο πενήντα ετών από το 1975 έως το 2015 οι καρδιακές νόσοι ήταν η πρώτη αιτία θανάτου με τον καρκίνο να ακολουθεί δεύτερος. (Antman & Loscalzo, 2016).

Τα ερευνητικά αυτά δεδομένα μπορούν να γενικευτούν και στον παγκόσμιο πληθυσμό αφού το 2015 σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας οι παθήσεις του καρδιαγγειακού συστήματος αποτέλεσαν την Πρώτη αιτία θανάτου. Παράλληλα το κόστος αντιμετώπισης και περίθαλψης των ασθενών με καρδιαγγειακά νοσήματα αποτελεί την πιο δαπανηρή νοσολογική κατηγορία με το παγκόσμιο κόστος να ανέρχεται στα 237 δισεκατομμύρια δολάρια εννοώ μέχρι το 2035 το κόστος αυτό προβλέπεται να εκτιναχθεί στα 368 δισεκατομμύρια δολάρια.

Η θνησιμότητα από έμφραγμα του μυοκαρδίου μειώνεται με την πάροδο του χρόνου, αντανακλώντας την πρόοδο της διάγνωσης και της θεραπείας τις τελευταίες δύο δεκαετίες, ο κίνδυνος καρδιακής νόσου παραμένει υψηλός με υπολογισμένο κίνδυνο 50% έως την ηλικία των 45 ετών στο γενικό πληθυσμό (Johnson,Torres Soto,Glicksberg, Shameer,Miotto, Ali, & Dudley, 2018 Είναι επιστημονικά αποδεδειγμένο πως παράλληλα με την αύξηση της ηλικίας αυξάνεται και η συχνότητα εμφάνισης καρδιαγγειακών νοσημάτων. Συγχρόνως παρατηρούνται διακυμάνσεις ανάμεσα στα δύο φύλα. Οι άνδρες μικρότερης ηλικίας είναι πιο επιρρεπείς στην εμφάνιση τέτοιων οχημάτων από ότι οι γυναίκες. (Antman & Loscalzo, 2016). Η διαφορά στη συχνότητα περιορίζεται προοδευτικά στην μετεμμηνοπαυσιακή κατάσταση.

Ως αθηροσκλήρωση ονομάζεται εκείνη η διαγνωστική κατηγορία που αφορά την μειωμένη ή την παντελή απουσία ροής του αίματος στις αρτηρίες και στην αορτή εξαιτίας της στένωσης των αιμοφόρων αγγείων. Η αθηροσκλήρωση προκαλείται από πολλούς και διαφορετικούς παράγοντες όπως ανοσολογικά φαινόμενα ενδοθηλιακή δυσλειτουργία η φλεγμονή. Αυτοί οι παράγοντες πιστεύεται ότι προκαλούν το σχηματισμό λιπώδους ράβδου, η οποία είναι το χαρακτηριστικό στην ανάπτυξη της αθηρωματικής πλάκας (Goodfellow, Bengio & Courville, 2016) μια προοδευτική διαδικασία που μπορεί να συμβεί ήδη από την παιδική ηλικία.

Η διαδικασία ανάπτυξης αθηρωματικής πλάκας συμβαίνει εξαιτίας της πάχυνσης του εσωτερικού χιτώνα των αγγείων. Συγχρόνως μικροφάγα αφρώδη κύτταρα φορτισμένα με λιπίδια συσσωρεύονται στην κυτταρική μήτρα παράλληλα με την συσσώρευση λείων μυϊκών κυττάρων. Καθώς αυτές οι βλάβες συνεχίζουν να επεκτείνονται, μπορεί να προκύψει απόπτωση των εν τω βάθει στιβάδων,

επιταχύνοντας περαιτέρω επιστράτευση μακροφάγων που μπορεί να ασβεστοποιηθεί και μετάβαση σε αθηροσκληρωτικές πλάκες (Harrington, 2012).

1.3. Χαρακτηριστικά τους

Ως καρδιαγγειακές παθήσεις ορίζεται εκείνη η ομάδα των παθήσεων που αφορούν το καρδιαγγειακό σύστημα δηλαδή την καρδιά και τα αιμοφόρα αγγεία δηλαδή τις φλέβες αρτηρίες και το τριχοειδή αγγεία. Μερικά παραδείγματα καρδιαγγειακών παθήσεων είναι η εγκέφαλο αγγειακές παθήσεις η ρευματοειδής καρδιακές παθήσεις οι περιφερειακές αρτηριακές και οι στεφανιαίες παθήσεις. (Antman & Loscalzo, 2016).

Επιπλέον, οι καρδιακές προσβολές και τα εγκεφαλικά είναι συνήθως οξέα συμβάντα, που προκαλούνται από απόφραξη που εμποδίζει τη ροή του αίματος προς την καρδιά και τον εγκέφαλο, αντίστοιχα (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020)

Η κύρια αιτία για αυτό είναι η συσσώρευση λιπών στα εσωτερικά τοιχώματα των BVs που παρέχουν αίμα στην καρδιά και τον εγκέφαλο.

Η Αιμορραγία των αιμοφόρων αγγείων του εγκεφάλου και οι θρομβώσεις στο αίμα μπορεί να οδηγήσουν στην εμφάνιση εγκεφαλικών επεισοδίων. Παράγοντες όπως η υπέρταση και η αρτηριοσκλήρυνση θεωρούνται ως οι πιο σημαντικοί σχετικά με την εμφάνιση καρδιαγγειακών νόσων. Η ηλικία επίσης είναι ένας πολύ σημαντικός παράγοντας καθώς με την πάροδο του χρόνου προκαλούνται αλλαγές στη λειτουργία του καρδιαγγειακού συστήματος μορφολογικού και φυσιολογικού χαρακτήρα. Οι αλλαγές αυτές αυξάνουν την πιθανότητα εμφάνισης τέτοιων νοσημάτων ακόμα και αν το άτομο της Τρίτης ηλικίας είναι υγιές. (Garcia, Faber, Cooke, Folks Chen & Santana, 2007).

Παρόλο που οι παθήσεις του καρδιαγγειακού συστήματος αποτελούν την πρώτη αιτία θανάτου σε παγκόσμιο επίπεδο ένα θετικό στοιχείο είναι πως τα επίπεδα θνησιμότητας έχουν μειωθεί στο δυτικό κόσμο από το 1970 και μετά και κυρίως σε χώρες ο πληθυσμός λαμβάνει υψηλό εισόδημα. (Limbacher, Douglas & Germano, 1998). Εν τω μεταξύ, οι καρδιαγγειακοί θάνατοι και οι ασθένειες έχουν αυξηθεί με ταχύτερο ρυθμό σε χώρες χαμηλού και μεσαίου εισοδήματος (Marr & Bowen, 2011) και έχει υπολογιστεί ότι πάνω από το 80% των θανάτων παγκοσμίως από καρδιαγγειακά νοσήματα συμβαίνουν σε χώρες χαμηλού και μεσαίου εισοδήματος (Niederer, Lumens & Trayanova, 2019). Στις χώρες αυτές χαμηλού και μεσαίου εισοδήματος ο μέσος όρος ηλικίας όπου οι άνθρωποι πεθαίνουν από καρδιαγγειακά νοσήματα είναι αρκετά χαμηλότερος. Συγχρόνως τις χώρες αυτές

χαμηλότερος είναι και ο μέσος όρος ηλικίας θανάτου από άλλες ασθένειες μη μεταδοτικές.

Ο Παγκόσμιος Οργανισμός Υγείας παρουσιάζει επίσης συναφή ερευνητικά δεδομένα που αποδεικνύουν ότι τα καρδιαγγειακά νοσήματα αποτελούν την πρώτη αιτία θανάτου σε όλο τον κόσμο. Ενδεικτικά το 2008 17,3 εκατομμύρια άνθρωποι έχασαν τη ζωή τους με το ποσοστό αυτών αποτελεί το 30% όλων των θανάτων παγκοσμίως (Topol & Teirstein, 2015). Από αυτούς τους θανάτους, τον Παγκόσμιο άτλαντα για την πρόληψη και τον έλεγχο της καρδιαγγειακής νόσου, ο ΠΟΥ έχει υπολογίσει ότι 7,3 εκατομμύρια οφείλονταν σε στεφανιαία νόσο και 6,2 εκατομμύρια οφείλονταν σε εγκεφαλικό (Niederer, Lumens, Trayanova, 2019).

Οι οικονομικές επιπτώσεις που επιφέρουν οι μεταδοτικές ασθένειες όπως η καρδιαγγειακή νόσος είναι πολύ σημαντικές ιδιαίτερα σε χώρες χαμηλού και μεσαίου εισοδήματος. Στις χώρες αυτές τα αποτελέσματα είναι πολύ πιο έντονα καθώς υπολογίζεται ότι το ΑΕΠ συρρικνώνεται μέχρι και 6,77%. Οι Mathers et al. ανέφεραν ότι τα καρδιαγγειακά νοσήματα (κυρίως από έμφραγμα και εγκεφαλικά) αναμένεται να παραμείνουν η μοναδική κύρια αιτία θανάτου και θα αυξηθούν και θα φτάσουν τα 23,3 εκατομμύρια έως το 2030 (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020)

Επιπλέον, οι Lim et al. ανέλυσε μια συστηματική και συγκριτική εκτίμηση κινδύνου για την επιβάρυνση των ασθενειών και των τραυματισμών που αποδίδονται σε παράγοντες κινδύνου και παρατήρησε ότι κάθε χρόνο 9,4 εκατομμύρια θάνατοι ή το 16,5% όλων των θανάτων προκαλούνται λόγω υψηλής αρτηριακής πίεσης, μεταξύ των οποίων το 51% των θανάτων περιλαμβάνει πνεύμονες και το 45% των θανάτων που αποδίδονται σε καρδιαγγειακά νοσήματα. Η υιοθέτηση ενός υγιεινού τρόπου ζωής από δεν θα περιλαμβάνει συνήθειες όπως το κάπνισμα η παχυσαρκία και η ανθυγιεινή διατροφή και ο καθιστικός τρόπος ζωής αποτελούν μερικούς από τους πιο αποτελεσματικούς τρόπους πρόληψης και αντιμετώπισης των καρδιαγγειακών νοσημάτων. (Ayodele, 2010; Bonaccorso, 2017).

Η πρόσφατη έλευση της νανοτεχνολογίας είχε τεράστιο αντίκτυπο σε πολλούς τομείς της επιστήμης και της μηχανικής, ειδικά στην πρόοδο της ιατρικής επιστήμη και φροντίδα υγείας (Butler, Davies, Cartwright, Isayev & Walsh, 2018).

Στη νανοτεχνολογία, χρησιμοποιούνται μηχανικά υλικά ή συσκευές για τη μικρότερη λειτουργική οργάνωση στην κλίμακα νανομέτρου (1-100 nm) σε τουλάχιστον μία διάσταση (Goldberg & Holland, 1988). Επίσης τόσο νανοσυσκευές όσο και τα νανοϋλικά συνδυάζονται και αλληλεπιδρούν σε μοριακό επίπεδο με διάφορες οντότητες βιολογικού χαρακτήρα. Η αλληλεπίδραση αυτή περιλαμβάνει έναν υψηλό βαθμό τόσο ειδικότητας όσο και αντιδραστικότητα σας. Μέσα από την αλληλεπίδραση αυτή οι νανοσυσκευές και τα νανοϋλικά μπορούν να προκαλέσουν κάποιες φυσιολογικές αποκρίσεις με παράλληλη μείωση ανεπιθύμητων ενεργειών καθώς διεγείρουν συγκεκριμένα κύτταρα στόχους. Έτσι γίνεται εύκολα αντιληπτό

πως η νανοτεχνολογία μπορεί να έχει σημαντικές και ουσιώδεις επαναστατικές εφαρμογές στον τομέα των καρδιαγγειακών παθήσεων. (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020).

1.4. Τρόποι Διάγνωσής τους

Η διαχείριση της καρδιαγγειακής νόσου είναι πολύ εκτεταμένη ανάλογα με την κλινική κατάσταση (θρομβόλυση κατευθυνόμενη από καθετήρα για οξύ ισχαιμικό εγκεφαλικό επεισόδιο, αγγειοπλαστική για περιφερική αγγειακή νόσο, στεφανιαία ενδοπρόθεση για ΣΝ Παρόλα αυτά γίνεται εύκολα αντιληπτό πως οι ασθενείς με νοσήματα του καρδιαγγειακού συστήματος είναι επιτακτικό να εκπαιδεύονται αναφορικά με την υιοθέτηση ενός πιο υγιεινού τρόπου ζωής καθώς και την υιοθέτηση δραστηριοτήτων δευτερογενούς πρόληψης.

Ο θάνατος αποτελεί την σημαντικότερη επίπτωση που μπορεί να προκληθεί από τις παθήσεις του καρδιαγγειακού συστήματος. Παρόλο που τις τελευταίες δεκαετίες η εξέλιξη της ιατρικής έχει συμβάλει σημαντικά στην αντιμετώπιση της νόσου αυτής η καρδιαγγειακή νόσος παραμένει Η Πρώτη αιτία θανάτου παγκοσμίως. (Antman & Loscalzo, 2016).

Άλλες επιπλοκές όπως η ανάγκη για μεγαλύτερη νοσηλεία, η σωματική αναπηρία και το αυξημένο κόστος περίθαλψης είναι σημαντικές και αποτελούν το επίκεντρο για τους υπεύθυνους χάραξης πολιτικής στον τομέα της υγείας καθώς πιστεύεται ότι θα συνεχίσουν να αυξάνονται τις επόμενες δεκαετίες (El Naqa & Murphy, 2015).

Η λίστα με τις πιθανές επιπτώσεις που μπορεί να προκαλέσουν τα εγκεφαλικά επεισόδια είναι μακρά. Μερικές από τις πιο χαρακτηριστικές επιπτώσεις είναι η αφασία η δυσφαγία και η δυσαρθρία. Άλλες μορφές αναπηρίας είναι γενικευμένη η εστιακή μυϊκή αδυναμία είτε προσωρινή είτε μόνιμη. Η μορφή αυτή αναπηρίας μπορεί να προκαλέσει κάποιες επιπρόσθετες επιπλοκές αναφορικά με την κινητικότητα του ασθενούς όπως για παράδειγμα τα θρομβοεμβολικά επεισόδια ή η ανάπτυξη ουροποιητικού ct. Κάποιες επιπρόσθετες επιπλοκές που δεν συνιστούν αναπηρία είναι ο σημαντικός σωματικός περιορισμός πιθανές πληγές στο σώμα η ισχαιμία στα άκρα..

Τα ερευνητικά δεδομένα επίσης καταδεικνύουν ότι για την αποτελεσματικότερη αντιμετώπιση παθήσεων όπως η στεφανιαία νόσος η καρδιακή ανεπάρκεια είναι ιδιαίτερα αποτελεσματική η συνεργασία μεταξύ γιατρών πρωτοβάθμιας περίθαλψης διαφορετικών ειδικοτήτων αλλά κι άλλων επαγγελματιών. Μία ολοκληρωμένη και πολύπλευρη διεπιστημονική ομάδα που περιλαμβάνει γιατρούς διαιτολόγους νοσηλευτές καρδιολόγους και άλλες ειδικότητες έχει αποδειχθεί ότι μπορεί να επιφέρει πολύ καλύτερα και θετικά αποτελέσματα όχι μόνο στις παραπάνω διαγνωστικές κατηγορίες αλλά και σε άλλες παθήσεις. (Davies, Cartwright, Isayev & Walsh, 2018).

Βέβαια είναι δεδομένο πως γιατί είναι καλύτερη και αποτελεσματικότερη αντιμετώπιση των παθήσεων καρδιαγγειακής φύσεως το σημαντικότερο ρόλο παίζει η πρωτογενής πρόληψη και διαρκής ενημέρωσης των πολιτών. Στόχος είναι η διαρκής παραίνεση για την υιοθέτηση ενός υγιεινού τρόπου ζωής προκειμένου να μειωθούν σημαντικά οι πιθανότητες για την δημιουργία αθηροσκλήρωσης η οποία αποτελεί σημαντικό παράγοντα επικινδυνότητας για την ανάπτυξη καρδιαγγειακής νόσου στο μέλλον.

Μερικές από τις σημαντικότερες προτροπές για την υιοθέτηση ενός υγιεινού τρόπου ζωής είναι οι εξής:

- Αποφυγή καπνίσματος
- Τακτική σωματική δραστηριότητα
- Διατήρηση το δείκτη μάζας σώματος σε φυσιολογικά επίπεδα
- Χαμηλά επίπεδα χοληστερόλης
- Διατήρηση της αρτηριακής πίεσης σε χαμηλά επίπεδα χωρίς θεραπεία
- Διατήρηση της γλυκόζης στο αίμα σε επίπεδα μικρότερα από τα 100 mg/dl

Φυσικά ιδιαίτερη βάση πρέπει να δοθεί στις ομάδες υψηλού κινδύνου δηλαδή σε άτομα μεγάλης ηλικίας άτομα που καπνίζουν είναι παχύσαρκα ή πάσχουν από άλλες νόσους που ενισχύουν την πιθανότητα εμφάνισης καρδιοπάθειας όπως υπέρταση δυσλιπιδαιμία ή διαβήτη. Ένας τρόπος για την προστασία των ευπαθών αυτών ομάδων είναι ο διαρκής έλεγχος της κατάστασης τους μέσω ιατρικών εξετάσεων καθώς και η προτροπή για την υιοθέτηση υγιεινών συνηθειών όπως η διατήρηση φυσιολογικού δείκτη μάζας σώματος και η απώλεια βάρους, η αποφυγή του καπνίσματος και η άσκηση. Συγχρόνως στις ομάδες αυτές έχει αποδειχθεί αποτελεσματική και η προτροπή για χρήση φαρμάκων όπως η ασπιρίνη σε χαμηλή δόση.σημασίας (Niederer, Lumens &Trayanova, 2019).

Όπως αναφέρθηκε και παραπάνω οι καρδιαγγειακές παθήσεις αποτελούν την Πρώτη αιτία θανάτων σε παγκόσμιο επίπεδο και χωρίζονται σε τέσσερις κατηγορίες: CAD CVD PVD η αθηροσκλήρωση της αορτής. Ο παράγοντας κινδύνου και η τροποποίηση του τρόπου ζωής είναι πρωταρχικής σημασίας για την πρόληψη της καρδιαγγειακής νόσου. Όταν μιλάμε για μέτρα πρωτογενούς πρόληψης των καρδιαγγειακών παθήσεων μιλάμε ως επί το πλείστον για μέτρα που στοχεύουν στην πρόληψη δημιουργίας αθηροσκλήρωσης της αορτής. (Davies, Cartwright, Isayev & Walsh, 2018).

Πληθώρα επιστημονικών δεδομένων καταδεικνύουν πως για την αποτελεσματικότερη αντιμετώπιση καρδιαγγειακών παθήσεων όπως η καρδιακή ανεπάρκεια ή η Στεφανία νόσος είναι απαραίτητη η συμμετοχή μιας πολύπλευρης και διευρυμένης επιστημονικής ομάδας επαγγελματιών οι οποίοι παράλληλα να συνεργάζεται πάντα με επίκεντρο τον ασθενή. Μία ομάδα αποτελούμενη από γιατρούς διαφορετικών ειδικοτήτων όπως οι καρδιολόγοι αλλά και άλλοι επαγγελματίες όπως νοσηλευτές φαρμακοποιοί διαιτολόγοι συνθέτουν με τις

γνώσεις τους μία άριστα θεωρητικά καταρτισμένη ομάδα και με τον τρόπο αυτόν αυξάνονται οι πιθανότητες επιτυχούς αντιμετώπισης τέτοιων παθήσεων. Παρόμοια είναι και τα ερευνητικά αποτελέσματα που καταδεικνύουν πως μία ομάδα παρέμβασης ακολουθούμενη στη συνέχεια από μία διευρυμένη διεπιστημονική ομάδα διαφόρων ειδικοτήτων είναι περισσότερο αποτελεσματική. (Niederer, Lumens, Trayanova, 2019). Αυτή η ομάδα είχε μια μείωση στη θνησιμότητα όλων των αιτιών που σχετίζεται με ΣΝ κατά 76% σε σύγκριση με την ομάδα ελέγχου. Επιπρόσθετα οι υποχρεώσεις των επαγγελματιών στο χώρο της υγείας είναι να ενημερώνει και να προτρέπει τους ασθενείς να υιοθετούν ένα πιο υγιεινό τρόπο ζωής αλλά και να προσπαθούν να τροποποιούν διαρκώς τους παράγοντες κινδύνου που οδηγούν στην εμφάνιση καρδιαγγειακών παθήσεων. (Marr & Bowen, 2011).

Τα τελευταία χρόνια έχει καταδειχθεί επανειλημμένα ο ρόλος της φλεγμονής στην ανάπτυξη και εξέλιξη παθήσεων του καρδιαγγειακού συστήματος. Η φλεγμονή αποδεδειγμένα αυξάνει τις πιθανότητες για την ανάπτυξη αθηροσκλήρωσης ενώ παράλληλα έχει αποδειχθεί ότι μειώνει την βιολογική υπόσταση των αρτηριών του καρδιαγγειακού συστήματος. Η τροποποίηση των παραπάνω παραγόντων κινδύνου μπορεί να συμβάλλει σημαντικά στην αντιμετώπιση του φαινομένου της φλεγμονής και τις συνακόλουθες επιπτώσεις που αυτή προκαλεί στην υγεία μας. (Marsland, 2011.

Μέσα από την κλινική πράξη και τα πειράματα έχουν αναλυθεί επιστημονικά μετά που καταδεικνύουν την αξιοποίηση της φλεγμονώδους κατάστασης ως μέτρο πρόληψης και αντιμετώπισης των καρδιαγγειακών παθήσεων. Εντελώς, η έννοια της φλεγμονής έχει υλοποιηθεί από τον τομέα της θεωρίας και των εργαστηριακών ερευνών για να αναλάβει έναν πολλά υποσχόμενο ρόλο ως χρήσιμο εργαλείο στην κλινική για να βοηθήσει την πρόληψη και τη διαχείριση των καρδιαγγειακών παθήσεων (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020),

Η καρδιαγγειακή νόσος και οι σχετικές διαταραχές χαρακτηρίζονται από ένα φλεγμονώδες συστατικό σε κάποιο στάδιο της παθολογίας. Οι παράγοντες κινδύνου για καρδιαγγειακά νοσήματα προκαλούν μια ποικιλία ερεθισμάτων που εξάγουν την έκκριση των διαλυτών λευκοκυττάρων μορίων προσκόλλησης όπως η Ε-σελεκτίνη, τα οποία προάγουν την προσκόλληση μονοκυττάρων σε ενδοθηλιακά κύτταρα (τα ενδοθηλιακά κύτταρα επίσης ενεργοποιούν και εκφράζουν μια ποικιλία μορίων προσκόλλησης) και διευκολύνουν τη μετανάστευση των μονοκυττάρων στον υποενδοθηλιακό χώρο.

Επιπλέον, στον υποενδοθηλιακό χώρο, ο μετασχηματισμός των μονοκυττάρων σε μακροφάγους και η πρόσληψη λιποπρωτεϊνών χοληστερόλης πιστεύεται ότι αναδύουν λιπαρές ραβδώσεις. Επιπλέον, τα βλαβερά ερεθίσματα μπορεί να συνεχίσουν τη συσσώρευση και την έλξη μακροφάγων, μαστοκυττάρων και ενεργοποιημένων Τ κυττάρων εντός της αναπτυσσόμενης αθηροσκληρωτικής βλάβης.

Ισχαιμικός τραυματισμούς με συνακόλουθη επαναιμάτωση της περιοχής του εμφράγματος μπορεί να προκληθεί από εγκεφαλικά επεισόδια η εμφράγματα του μυοκαρδίου τα οποία όμως δεν είναι θανατηφόρα. Τα περιστατικά αυτά φράζουν τα στεφάνια και τα εγκεφαλικά αγγεία και συχνά οδηγούν στα παραπάνω αποτελέσματα. (Niederer, Lumens, Trayanova, 2019).

Καθώς συμβαίνει η επαναιμάτωση της περιοχής του φράγματος ενεργοποιούνται λευκοκύτταρα. Τα λευκοκύτταρα αυτά στη συνέχεια ελευθερώνουν στο αίμα διάφορα μόρια απόκρισης στο οξειδωτικό στρες καθώς και πεπτικά και προφλεγμονώδει λιπιδικά μέσα. Οι οξειδωμένες λιποπρωτεΐνες χαμηλής πυκνότητας (LDL-οξειδωμένες) μπορεί να είναι ένας από τους πολλούς παράγοντες που προκαλούν απώλεια λείων μυϊκών κυττάρων μέσω απόπτωσης στο καπάκι της αθηρωματικής πλάκας.

Το κολλαγόνο μπορεί να διασπαστεί μέσα από την απελευθέρωση μεγαλοπρωτεινάσης αλλά και ενζύμων του συνδετικού ιστού καθώς και ενεργοποιούνται τα μακροφάγα. Η διάσπαση αυτή του κολλαγόνου οδηγεί σε αποδυνάμωση αλλά και πιθανής ρήξης του καπακιού της αθηρωματικής Πλάκας. Οι θρομβώσεις προκαλούνται ακριβώς από αυτή την ρήξη της Πλάκας. Για τους παραπάνω λόγους θεωρείται ότι η αθηροσκλήρωση έχει κάποια χαρακτηριστικά κύτταρα της φλεγμονής όπως βιοενεργά μόρια και κυτοκίνες..

Ο σημαντικός και καίριος ρόλος της φλεγμονής στην ανάπτυξη καρδιαγγειακών παθήσεων καθιστά σημαντικό και το ρόλο της στην πρόβλεψη και έγκαιρη διάγνωση των νόσων αυτών. Οι χαρακτηριστικοί δείκτες ανάπτυξης φλεγμονής όπως τα μακροφάγα και οι κυτοκίνες η πεπτιδικοι μεσολαβητές και τα λιπίδια ενεργοποιημένα Τ κύτταρα και τα μαστοκύτταρα μπορούν να αξιοποιηθούν και την έγκαιρη πρόβλεψη παθήσεων του καρδιαγγειακού συστήματος. Οι ενημερωμένες πληροφορίες σχετικά με τις φλεγμονές της καρδιαγγειακής νόσου, όπως οι αιτιολογικοί παράγοντες, τα διαδοχικά συμβάντα και οι πιθανοί μηχανισμοί, θα ωφεληθούν στη διάγνωση και στις κλινικές εφαρμογές των καρδιαγγειακών νοσημάτων (Davies, Cartwright, Isayev & Walsh, 2018).

1.5. Τρόποι αντιμετώπισής τους

Τα καρδιαγγειακά νοσήματα συνεχίζουν να αποτελούν ακόμα και σήμερα την πρώτη αιτία νοσηρότητας αλλά και θανάτων παγκοσμίως παρόλες τις επιστημονικές εξελίξεις τόσο στην πρόσληψη όσο και στη θεραπεία τα τελευταία χρόνια. Παρόλα αυτά η επιστημονική έρευνα έχει οδηγήσει στην αναζήτηση καινούργιων πρακτικών όπως οι μοριακές θεραπευτικές πρακτικές αντιμετώπισης των καρδιαγγειακών νοσημάτων. Τέτοια παραδείγματα αποτελούν οι θεραπείες γονιδίων οι θεραπείες

με βλαστοκύτταρα η μεταμόσχευση κυττάρων οι αναστολές αισθητικού παράγοντα καθώς και τα micro RNSs.

Από τις παραπάνω θεραπείες η πρακτική της θεραπείας με βλαστοκύτταρα θεωρείται ως η επικρατέστερη αναδυόμενη θεραπεία των καρδιαγγειακών νοσημάτων. Πιο συγκεκριμένα μέσω της μεθόδους αυτής μπορεί να αποκατασταθεί η ροή του αίματος καθώς προγονικά κύτταρα όπως εκείνα του μυελού των οστών διαφοροποιούνται σε άλλους τύπους αγγειακών κυττάρων. Τα βλαστοκύτταρα σε αυτά είναι ιδιαίτερα αποτελεσματικά στα εμφράγματα του μυοκαρδίου καθώς ενισχύουν την δημιουργία νέων αγγείων μετά από ισχαιμικό επεισόδιο του μυοκαρδίου. Η αποτελεσματικότητα αυτή υποστηρίζεται ότι συμβαίνει καθώς τα κύτταρα του μυελού των οστών διαφοροποιούνται σε κύτταρα του μυοκαρδίου καθώς μεταμοσχεύονται σε κοιλιακό ουλώδη ιστό. Η διαφοροποίηση αυτή μπορεί να αποκαταστήσει την φυσιολογική λειτουργία του μυοκαρδίου. (Davies, Cartwright, Isayev & Walsh, 2018).

Αντίθετα η μέθοδος και η τεχνική των εμφυτευμάτων κυττάρων φαίνεται να έχει περιορισμένη αποτελεσματικότητα και επίδραση όσον αφορά τη λειτουργία της καρδιάς. Η περιορισμένη αυτή επίδραση συμβαίνει εξαιτίας ελλιπούς βεβαιότητας σχετικά με τη διαφοροποίηση και τον πολλαπλασιασμό των καρδιακών κυττάρων. Παράλληλα ενώ έχει καταδειχθεί επιστημονικά ότι τα προγονικά κύτταρα του μυελού των οστών μπορεί να διαφοροποιηθούν σε ενδοθηλιακά κύτταρα δεν ισχύει το ίδιο σχετικά με την ιδέα διαφοροποίηση των προγονικών κυττάρων σε καρδιομυοκύτταρα.

Τις τελευταίες δεκαετίες έχει αναπτυχθεί ιδιαίτερα η επιστήμη των μοριακών τεχνικών και θεραπειών αντιμετώπισης των καρδιαγγειακών νοσημάτων. Η ανάπτυξη των τεχνικών αυτών στοχεύει ουσιαστικά να τροποποιήσει τις μοριακές διεργασίες και τη μοριακή βιολογία του καρδιαγγειακού συστήματος και να στοχεύσει τα αποτυχημένα και μη λειτουργικά κύτταρα του μυοκαρδίου.

Οι χρόνιες και παραδοσιακές φαρμακολογικές θεραπείες καρδιαγγειακών νοσημάτων τα τελευταία χρόνια να αντικαθιστώνται από γονιδιακές θεραπείες. Η γονιδιακή θεραπεία θεωρείται ως μία καινούργια και ελκυστική λύση στην αντιμετώπιση των νοσημάτων αυτών. Επιπλέον, οι εξελίξεις στην τεχνολογία του ανασυνδυασμένου DNA, συμπεριλαμβανομένης της μεταφοράς γονιδίων, έχουν υποκινήσει την ελπίδα ότι αυτή η τεχνολογία μπορεί να χρησιμοποιηθεί για τη βελτίωση της πρακτικής της καρδιαγγειακής ιατρικής (Marr & Bowen, 2011).

Η ανάπτυξη εμπλοκών της μοριακής γενετικής για τη θεραπεία των καρδιαγγειακών νοσημάτων εξαρτάται από την τεχνική πρόοδο στην ανάπτυξη μεθόδων μεταφοράς γονιδίων. Οι διάφορες τεχνικές μεταφορά γονιδίων έχουν ως στόχο να τροποποιήσουν και να παρέμβουν σε συγκεκριμένα κύτταρα του καρδιαγγειακού συστήματος με τρόπο αποτελεσματικό μακροπρόθεσμο και στοχευμενο..

Τρεις είναι οι τρόποι μεταφοράς γονιδίων για την αντιμετώπιση των παθήσεων αυτών του καρδιαγγειακού συστήματος. Πρώτα μέσα από ειδικούς φορείς όπως ρετροιοι και οι αδενοιοι δεύτερον μέσα από μυϊκούς φορείς όπως τα κανονικά λιποσωματα και τρίτον μέσα από χορήγηση η οποία γίνεται άμεσα διαγγειακά και ex vivo.

Επιπλέον, γονιδιακή θεραπεία με ισομορφές αυξητικών παραγόντων όπως οι αγγειακοί ενδοθηλιακοί αυξητικοί παράγοντες (VEGFs), οι αυξητικοί παράγοντες ινοβλαστών (FGFs) και οι αυξητικοί παράγοντες ηπατοκυττάρων (HGFs) επάγουν αγγειογένεση, μειώνουν την απόπτωση και οδηγούν σε προστασία στην ισχαιμική καρδιά. Η γονιδιακή θεραπεία που κωδικοποιεί αντιοξειδωτικά, πρωτεΐνες θερμικού σοκ (HSPs), ενεργοποιημένη με μιτογόνο πρωτεϊνική κινάση (MAPK) και πολλές άλλες αντι-αποπτωτικές πρωτεΐνες έχει επιδείξει σημαντική καρδιοπροστασία σε ζωικά μοντέλα (Davies, Cartwright, Isayev & Walsh, 2018).

Ωστόσο, έχει αναφερθεί ότι έχει ξεκινήσει μια καρδιαγγειακή γονιδιακή θεραπεία για τη διέγερση της αγγειογένεσης σε ασθενείς, την περιφερική αγγειακή νόσο. Αντιθέτως πιο αργά θα λέγαμε ότι προχωράει η μέθοδος της μεταφοράς γονιδίων αντιμετώπισης νοσημάτων του καρδιαγγειακού συστήματος. Συγχρόνως κυτταρολυτικές αποκρίσεις που στοχεύουν μολυσμένα η κατεστραμμένα κύτταρα οδηγούν σε μία έλλειψη εμμονής αναφορικά με την γονιδιακή έκφραση.

Πολλές παθήσεις του καρδιαγγειακού συστήματος έχουν αποδοθεί στον παράγοντα θρομβοπλαστίνη ή αλλιώς ιστικό παράγοντα. Η θρομβοπλαστίνη θεωρείται ως ο κύριος παράγοντας έναρξης της πήξης του αίματος. Πολλοί ασθενείς με νοσήματα που σχετίζονται με την καρδιά και τα αγγεία εμφανίζουν αυξημένα επίπεδα που παράγοντα αυτού.

Η θρομβοπλαστίνη επίσης έχει καταδειχθεί ότι οδηγεί στην ανάπτυξη αθηροσκλήρωσης μέσα από την δημιουργία θρόμβων. Η θρομβοπλαστίνη θεωρείται ότι ενεργοποιείται μέσα από διάφορους μηχανισμούς όπως η κινάση MAP η πρωτεϊνική κινάση C ή η Pi3 κινάση.

Συγχρόνως η γονιδιακή θεραπεία έχει αποδειχθεί ιδιαίτερα αποτελεσματική στην προστασία ισχαιμικού επεισοδίου της καρδιάς και στη μείωση της απόπτωσης. Πιο συγκεκριμένα η γονιδιακή θεραπεία που αξιοποιεί ισόμορφες αυξητικών παραγόντων όπως οι αυξητικοί παράγοντες ινοβλαστών και η αγγειακή ενδοθηλιακοί αυξητικοί παράγοντες οδηγούν σε αγγειογένεση. Συγχρόνως σημαντική καρδιακή προστασία μπορεί να προσφέρει η γονιδιακή θεραπεία που κωδικοποιεί διάφορες αντί αποπτωτικες πρωτεΐνες όπως αντιοξειδωτικά πρωτεΐνες θερμικού σοκ και πρωτεϊνικές κινάσης ενεργοποιημένες με μιτογονο. (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020).

Η παραγωγή ιστικού παράγοντα μπορεί επίσης να μειωθεί και μέσα από μόρια όπως τα ριβοένζυμα η αντί πληροφοριακά ολιγονουκλεοτιδια. Τα μόρια αυτά

καθώς και τα πολυκλωνικά και μονοκλωνικα αντισωματα οδηγούν στην αδρανοποίηση της θρομβοπλαστίνης. Έτσι ο σχηματισμός θρομβοπλαστίνης αναστέλλεται μέσα από τον σχηματισμό ενός τέταρτοταγούς ανασταλτικού συμπλέγματος καθώς στο σύμπλεγμα του ηλεκτρικού παράγοντα παρεμβαίνει η ανασυνδυασμένη αντιπηκτική πρωτεΐνη.

Τα τελευταία χρόνια οι έρευνες έχουν οδηγήσει στην ανάδειξη πολλών και διαφορετικών θεραπευτικών προτάσεων με σκοπό την αντιμετώπιση της δράσης της θρομβοπλαστίνης. Παρόλα αυτά η επίδραση και η συμβολή του αισθητικού παράγοντα στην γέννηση θρόμβων αποτελεί αντικείμενο επιστημονικής συζήτησης.

Μία άλλη σύγχρονη μοριακή μέθοδος αντιμετώπισης των καρδιαγγειακών παθήσεων είναι τα micro RNA. Τα μόρια αυτά είναι μη πρωτεϊνικά ρυθμιστικά μόρια που περιέχουν από 20 έως 23 νουκλεοτίδια. Τα μόρια αυτά τα οποία παρεμβαίνουν και ρυθμίζουν την αδρανοποίηση του RNA υπάρχουν σε όλους τους οργανισμούς. Συγχρόνως τα micro RNA μόρια έχουνε παραμείνει αναλλοίωτα στη διαδικασία της εξέλιξης και διαδραματίζουν πολύ σημαντικό ρόλο στις βασικές βιολογικές λειτουργίες. Περιληπτικά η διαδικασία θεραπεία μέσω των μορίων miRNA είναι η εξής. Τα αρχικά πρωτογενή mi RNA μόρια μετά από επεξεργασία σχηματίζονται σε πρόδρομα mRNA μόρια. Τα μόρια αυτά στη συνέχεια υποβάλλονται σε περαιτέρω επεξεργασία και στη συνέχεια σχηματίζονται μικρά δίπλα mRNA μόρια που αποτελούνται από 20 έως 23 νουκλεοτίδια.

Τέλος, τα miRNA εμπλέκονται στη σίγαση της γονιδιακής έκφρασης στο μετα-μεταφραστικό επίπεδο με αποτέλεσμα την αποικοδόμηση του mRNA ή με τη μεταφραστική αναστολή που τελικά οδηγεί σε καταστολή της πρωτεΐνης στόχου. Ανάλογα με το είδος της καρδιαγγειακής πάθησης τα mRNA μόρια εμφανίζουν και διαφορετικά πρότυπα έκφρασης. Για παράδειγμα διαφορετικά πρότυπα εμφανίζονται στην περίπτωση της καρδιακής υπερτροφίας ή ανεπάρκειας και διαφορετικά στο έμφραγμα του μυοκαρδίου. Το γεγονός αυτό δημιουργεί σημαντικές προσδοκίες αναφορικά με την χρησιμοποίηση των μορίων αυτών ως διαγνωστικούς δείκτες καρδιαγγειακών παθήσεων. (Davies, Cartwright, Isayev & Walsh, 2018).

Επίσης έχει καταδειχθεί επιστημονικά ότι για μεγάλα χρονικά διαστήματα τα mRNA μόρια μπορούν με αποτελεσματικότητα να ανασταλούν μέσα από αξιοποίηση τεχνολογιών αντιπληροφορίας. Αυτές οι τεχνολογίες έχουν τροφοδοτήσει ένα αυξανόμενο ενδιαφέρον για την αναστολή συγκεκριμένων miRNAs (που αναφέρονται ως antimiRs) και έχουν προκαλέσει ενθουσιασμό για τα miRNA ως νέους θεραπευτικούς στόχους.

Τα microRNA μόρια της καρδιάς που είναι υπεύθυνα για την αλυσίδα ασθενειών αδρανοποιούνται μέσα από την παροχή antimiRs. Η χορήγηση των ουσιών αυτών μπορούν να οδηγήσουν στην άμεση τροποποίηση των βιολογικών μηχανισμών που οδηγούν στην ανάδυση παθήσεων της καρδιάς. Ωστόσο, η φύση της σταθερότητας

των miRNAs που κυκλοφορούν στην κυκλοφορία του αίματος έθεσε πιθανές προκλήσεις των προσεγγίσεων που βασίζονται σε microRNA θεραπείες που βασίζονται σε φάρμακα (Timmis, Townsend, Gale, Torbica, Lettino, Petersen & Vardas, 2020).

2. Μηχανική Μάθηση

2.1. Ορισμός

Στο πιο βασικό της επίπεδο, η μηχανική μάθηση αναφέρεται σε κάθε τύπο προγράμματος υπολογιστή που μπορεί να «μάθει» από μόνο του χωρίς να χρειάζεται να προγραμματιστεί ρητά από άνθρωπο. Η ιδέα της ηλεκτρονικής μάθησης έχει αναπτυχθεί εδώ και δεκαετίες με σημαντικότερο σημείο εκκίνησης την εργασία Computing Machinery Intelligence του Alan Turing το 1950. Στην εργασία του αυτή αναφέρεται σε μία ενότητα πως η μηχανή μάθησης μπορεί να κάνει τον άνθρωπο και τον κάνει να πιστεύει ότι είναι πραγματική. (Ayodele, 2010).

Σήμερα, η μηχανική μάθηση είναι ένας ευρέως χρησιμοποιούμενος όρος που περιλαμβάνει πολλούς τύπους προγραμμάτων που θα συναντήσετε στην ανάλυση μεγάλων δεδομένων και την εξόρυξη δεδομένων. Στο τέλος της ημέρας, οι «εγκέφαλοι» που τροφοδοτούν στην πραγματικότητα τα περισσότερα προγνωστικά προγράμματα – συμπεριλαμβανομένων των φίλτρων ανεπιθύμητης αλληλογραφίας, των συστάσεων προϊόντων και των ανιχνευτών απάτης – είναι αλγόριθμοι μηχανικής μάθησης.

Ανάμεσα στην εποπτευόμενη μηχανική και την μη εποπτευόμενοι μηχανική μάθηση υπάρχουν συγκεκριμένες διαφορές με τις οποίες οι επιστήμονες της μηχανικής μάθησης πρέπει να γνωρίζουν και να είναι εξοικειωμένοι. Παράλληλα αναμένεται να γνωρίζουν και τη μοντελοποίηση του συνόλου καθώς διάφορες τεχνικές προσέγγισης τις ημι εποπτευόμενης μάθησης. Η τελευταία συνδυάζει μεθόδους και τεχνικές από την εποπτευόμενη και μη εποπτευόμενη μηχανική μάθηση. (Zhang, 2020).

Οι διαφορές μεταξύ εποπτευόμενης και μη εποπτευόμενης μάθησης θα μπορούσαμε να πούμε ότι είναι οι εξής. Στην πρώτη η χρήσης του προγράμματος του διαμορφώνεται με τέτοιο τρόπο προκειμένου να λαμβάνει μία απάντηση μέσα από μία συγκεκριμένη σειρά δεδομένων. Στην πρώτη κατηγορία αυτή συνήθως χρησιμοποιούνται αλγόριθμοι παλινδρόμησης και ταξινόμησης όπως είναι τα δέντρα αποφάσεων οι μηχανές διανυσμάτων και τα τυχαία δάση. Στην δεύτερη περίπτωση δηλαδή στην μη εποπτευόμενη μηχανική μάθηση οι απαντήσεις δημιουργούνται μέσα από δεδομένα τα οποία όμως είναι άγνωστα.

Οι επιστήμονες δεδομένων μπορούν να χρησιμοποιήσουν διάφορες μεθόδους προκειμένου να δημιουργήσουν αλγόριθμους μηχανικής μάθησης. Μεταξύ άλλων που μπορούν να χρησιμοποιήσουν οι γλώσσες και τεχνολογίες προγραμματισμού όπως για παράδειγμα της Java Scala και Python. Πέρα από τις γλώσσες μπορούν να χρησιμοποιηθούν και κάποια πλαίσια μηχανικής μάθησης τα οποία είναι προκατασκευασμένα όπως για παράδειγμα το Πλαίσιο Mahout ή η βιβλιοθήκη του apache spark η MLilb. Επίσης μπορεί να χρησιμοποιήσουμε και αλγόριθμος

ομαδοποίησης όπως τα k-means τα οποία συνδυάζονται σε μηχανική μάθηση μη εποπτευόμενη. (Sra, Nowozin & Wright, 2012).

Μία ευρέως διαδεδομένη μορφή μηχανικής μάθησης είναι αυτή της βαθιάς μάθησης. Η βαθιά μάθηση κυρίως τα τελευταία χρόνια αξιοποιείται για ζητήματα επίλυσης τύπων προβλημάτων του υπολογιστή τα οποία είναι σχετικά δύσκολα ενώ επίσης χρησιμοποιείται και στο πεδίο της όρασης υπολογιστή ή την επεξεργασία φυσικής γλώσσας. Η βαθιά μάθηση είναι ενδεικτικό ότι αξιοποιεί αλγόριθμους που προέρχονται τόσο από την εποπτευόμενη όσο και από την μη εποπτευόμενη μάθηση. Τα νευρωνικά δίκτυα μπορούν να έχουν πολλά κρυφά επίπεδα (all_is_magic/Shutterstock)

Πιο συγκεκριμένα η βαθιά μάθηση προέρχεται μέσα από τη θεωρία της μηχανικής μάθησης και βασίζεται στην εκμάθηση αναπαράστασης ή αλλιώς εκμάθηση χαρακτηριστικών. Ταυτόχρονα τα μοντέλα βαθιάς μάθησης είναι πιο αποδοτικά και πιο γρήγορα σε σχέση με τα παραδοσιακά μοντέλα μηχανικής μάθησης. Λειτουργούν μέσα από την εξαγωγή σύνθετων αφαιρέσεων υψηλού επιπέδου οι οποίες λειτουργούν ως αναπαραστάσεις δεδομένων. (Sra, Nowozin & Wright, 2012).

Η μεγαλύτερη αποδοτικότητα των συστημάτων βαθιάς μάθησης έγκειται στο γεγονός πως τα μοντέλα αυτά μπορούν να μάθουν τα σημαντικά χαρακτηριστικά από μόνα τους. Έτσι ο εκάστοτε επιστήμονας δεν χρειάζεται να επιλέξει τα χαρακτηριστικά αυτά με τρόπο μη αυτόματο προκειμένου να μπορέσεις να το μάθω το μοντέλο μηχανικής μάθησης. Το «βαθύ» στη βαθιά μάθηση προέρχεται από τα πολλά επίπεδα που είναι ενσωματωμένα στα μοντέλα βαθιάς μάθησης, τα οποία είναι συνήθως νευρωνικά δίκτυα.

Συνελικτικό νευρωνικό δίκτυο ή αλλιώς CNN ονομάζεται ένα δίκτυο το οποίο αποτελείται από πολλά μοντέλα. Στα δίκτυα αυτά το κάθε στρώμα λαμβάνει είσοδο από το προηγούμενο στρώμα. Στη συνέχεια μετά το στάδιο της επεξεργασίας το στρώμα αυτό εξάγεται στο επόμενο μέσω της μεθόδου της μαργαρίτας. Ένα από τα πιο διαδεδομένα συνελικτικά νευρωνικά δίκτυα ήταν αυτό της ομάδας Deepmind της Google. Το δίκτυο αυτό κατάφερε να κερδίσει τον παγκόσμιο πρωταθλητή του παιχνιδιού Go ενός αρχαίου κινεζικού παιχνιδιού. (Bonaccorso, 2017).

Ένας ακόμη λόγος για τον οποίον η βαθιά μάθηση έχει αποκτήσει τόσο μεγάλη δημοτικότητα είναι πως θα συνελικτικά νευρωνικά δίκτυα είναι πολύ πιο γρήγορα σε σχέση με τα παραδοσιακά δίκτυα GPU. Ένα τέτοιο χαρακτηριστικό παράδειγμα επεξεργαστή είναι ο επεξεργαστής Tesla k80 της Nvidia.. Δεύτερον, οι επιστήμονες δεδομένων συνειδητοποίησαν ότι τα τεράστια αποθέματα δεδομένων που συλλέγουμε μπορούν να χρησιμεύσουν ως ένα τεράστιο σώμα εκπαίδευσης και έτσι να επιβαρύνουν τα CNN ώστε να αποφέρουν ουσιαστική βελτίωση στην ακρίβεια της όρασης υπολογιστών και των αλγορίθμων NLP (Sra, Nowozin & Wright, 2012).

Μερικά από τα πιο χαρακτηριστικά παραδείγματα πλαισίων ανάπτυξης λογισμικού είναι τα Caffe, Torch, Theano καθώς και toTensorflow που δημιουργήθηκε από την Google. Τα λογισμικά αυτά θα λέγαμε Ε πως έχουν διαδραματίσει πολύ σημαντικό ρόλο στην ταχεία ανάπτυξη αυτοκινήτων αυτοοδηγούμενων. Η πρόοδος αυτή οφείλεται σε ένα βαθμό στην βαθιά μηχανική εκμάθηση μέσω της χρήσης συλλεκτικών νευρωνικών δικτύων σε GPU. (Shobha & Rangaswamy, 2018).

2.2. Επιβλεπόμενη Μάθηση

Η εποπτευόμενη μάθηση είναι ένα παράδειγμα μηχανικής μάθησης για την απόκτηση των πληροφοριών σχέσης εισόδου-εξόδου ενός συστήματος που βασίζεται σε ένα δεδομένο σύνολο ζευγών δειγμάτων εκπαίδευσης εισόδου-εξόδου. Ταυτόχρονα ένα δείγμα εισόδου-εξόδου μπορεί να πάρει την ονομασία ετικετοποιημένα δεδομένα εκπαίδευσης καθώς και εποπτευόμενα δεδομένα εννοώ Σε πιο σπάνιες περιπτώσεις μπορεί να πάρει την ονομασία Μάθηση από Επισημασμένα Δεδομένα η Επαγωγική Μηχανική Μάθηση καθώς και Μάθηση με Δάσκαλο. Αυτό γίνεται διότι η έξοδος είναι η ετικέτα επίβλεψης η ετικέτα των δεδομένων εισόδου. (Bonaccorso, 2017).

Η επιβλεπόμενη η εποπτευόμενη μάθηση σε αντίθεση με άλλα συστήματα μάθησης διενεργείται με σκοπό την δημιουργία ενός τεχνητού συστήματος. Το σύστημα αυτό αναμένεται να μπορεί να προβλέψει την έξοδο του συστήματος καθώς και να μπορεί να μάθει τη χαρτογράφηση που συμβαίνει ανάμεσα στην είσοδο και στην έξοδο. Εάν η έξοδος λάβει ένα πεπερασμένο σύνολο διακριτών τιμών που υποδεικνύουν τις ετικέτες κλάσεων της εισόδου, η εκμάθηση αντιστοίχισης οδηγεί στην ταξινόμηση των δεδομένων εισόδου. Εάν η έξοδος λάβει συνεχείς τιμές, οδηγεί σε παλινδρόμηση της εισόδου.

Σε πολλές περιπτώσεις χρησιμοποιούνται παράμετροι μοντέλου μάθησης ώστε να αντιπροσωπευθούν οι πληροφορίες που προκύπτουν μέσα από τη σχέση εισόδου εξόδου. Παρόλα αυτά υπάρχει η περίπτωση να πραγματοποιηθεί μία διαδικασία λήψης αναφορικά με αυτές τις παραμέτρους στην περίπτωση που οι παραμετρικές αυτοί δεν είναι διαθέσιμες μέσα από τα διάφορα δείγματα εκπαίδευσης. Τα δεδομένα κατάρτισης για την εποπτευόμενη μάθηση χρειάζονται εποπτευόμενες ή επισημασμένες πληροφορίες, ενώ τα δεδομένα εκπαίδευσης για μάθηση χωρίς επίβλεψη είναι χωρίς επίβλεψη, καθώς δεν φέρουν ετικέτα (δηλαδή, απλώς οι εισροές). Αλγόριθμος ημιεποπτευόμενης μάθησης ονομάζεται εκείνος ο αλγόριθμος ο οποίος χρησιμοποιεί δεδομένα τα οποία είναι εποπτευόμενα αλλά και μη εποπτευόμενα. (Butler, Davies, Cartwright, Isayev & Walsh, 2018). . Ενεργή μάθηση από την άλλη πλευρά ονομάζεται εκείνη η μορφή επαναληπτικής εποπτευόμενης. μάθησης κατά την οποία ο αλγόριθμος ρωτά τον χρήστη δάσκαλο για τις διαφορές ετικέτες στη διαδικασία της εκπαίδευσης.

Η εποπτευόμενη μάθηση παρουσιάζει αρκετά πλεονεκτήματα αλλά και μία σειρά από μειονεκτήματα. Όσον αφορά τα πλεονεκτήματα το κυριότερο από αυτά είναι ότι όλες οι αναλογικές έξοδοι του αλγόριθμου μπορούν να αποκτήσουν νόημα για τον άνθρωπο. Επίσης οι αναλογικές έξοδοι που χρησιμοποιεί ο αλγόριθμος μπορούν να χρησιμοποιηθούν τόσο στην παλινδρόμηση δεδομένων όσο και για διακριτική ταξινόμηση προτύπων.

Από την άλλη πλευρά σχετικά με τα μειονεκτήματα το πρώτο και κυριότερο από αυτά είναι η δυσκολία που υπάρχει στην συλλογή ετικετών αλλά και στην εποπτεία. Ειδικότερα στην περίπτωση που ο όγκος δεδομένων που χειριζόμαστε είναι πολύ μεγάλος συνήθως το κόστος είναι πολύ μεγάλο προκειμένου να επισημανθούν όλες αυτές οι πληροφορίες. Άλλο ένα σημαντικό μειονέκτημα είναι πως σε πολλές περιπτώσεις υπάρχουν ασάφειες και αβεβαιότητα αναφορικά με τις ετικέτες αφού στον πραγματικό κόσμο δεν έχουμε όλα τα πράγματα μία ετικέτα που να τα χαρακτηρίζει. Ένα χαρακτηριστικό παράδειγμα του μειονεκτήματος αυτού είναι το δίπολο εννοιών ζεστό και κρύο. Ο διαχωρισμός ανάμεσα σε αυτές τις έννοιες είναι δύσκολος να διακριθεί. Τα μειονεκτήματα αυτά είναι κάποια από τα χαρακτηριστικότερα παραδείγματα της εποπτευόμενης μάθησης.(Shalev-Shwartz & Ben-David, 2014). Για να ξεπεραστούν αυτοί οι περιορισμοί στην πράξη, μπορούν να ληφθούν υπόψη και άλλα παραδείγματα μάθησης, όπως η μάθηση χωρίς επίβλεψη, η ημι-εποπτευόμενη μάθηση, η ενισχυτική μάθηση, η ενεργητική μάθηση ή ορισμένες μικτές προσεγγίσεις μάθησης.

Η βασική λειτουργία της εποπτευόμενης μάθησης είναι πως μία μηχανή ή ένας υπολογιστής μπορεί να μάθει τη συμπεριφορά που επιδεικνύει ένα αντικείμενο ή ένας άνθρωπος σε κάποιες συγκεκριμένες δραστηριότητες. Μέσα από την εκμάθηση των συμπεριφορών αυτόν το μηχάνημα στη συνέχεια μπορεί να πραγματοποιήσει παρόμοιες συμπεριφορές επάνω σε αυτές τις δραστηριότητες. Ένα σημαντικό στοιχείο είναι πως σε πολλές περιπτώσεις οι μηχανές αυτές μπορούν να παράγουν έργο και να λειτουργήσουν πολύ πιο γρήγορα από τον άνθρωπο αλλά και με μεγαλύτερη ακρίβεια. Αυτό γίνεται διότι ένας υπολογιστής πραγματοποιεί τις αντιστοιχίσεις εισόδου-εξόδου κατά μέσο όρο πιο γρήγορα από ότι ο άνθρωπος. Μάλιστα η ταχύτητα αυτή πολλαπλασιάζεται όταν η μηχανή συνοδεύεται από έναν καλό επόπτη (Butler, Davies, Cartwright, Isayev & Walsh, 2018).

Όλα αυτά όσον αφορά τις πιο πολύ περίπλοκες δραστηριότητες και εργασίες οι περισσότεροι αλγόριθμοι εποπτευόμενης μάθησης αδυνατούν να ακολουθήσουν την ικανότητα μάθησης του ανθρώπου. Αυτό συμβαίνει διότι υπάρχει σημαντικός περιορισμός στο σχεδιασμό τόσο του λογισμικού και τον αλγόριθμο αλλά και στην έλλειψη υλικού (Shobha & Rangaswamy, 2018). Μερικές από τις πιο σημαντικές εφαρμογές της εποπτευόμενης μάθησης είναι η βιοπληροφορική και η χημείοπληροφορική, η ανίχνευση ανεπιθύμητης αλληλογραφίας και η ανάλυση της αγοράς, η ανακάλυψη η δεδομένων, η ανάκτηση πληροφοριών αλλά και η αναγνώριση ομιλίας.

2.3. Κατηγοριοποίηση

Όλοι οι εποπτευόμενοι αλγόριθμοι μηχανικής εκμάθησης βασίζονται σε προκαθορισμένο σύνολο ετικετών C και σε ένα εκπαιδευτικό σύνολο που αποτελείται από άρθρα στα οποία έχουν εκχωρηθεί μία ή περισσότερες ετικέτες από το C.

Ως Μηχανές Διανυσμάτων Υποστήριξης η αλλιώς SVM ονομάζονται τα μοντέλα μάθησης τα οποία είναι εποπτευόμενα και τα οποία αξιοποιούνται τόσο για την ανάλυση παλινδρόμησης όσο και ανάλυση ταξινόμησης (Limbacher, Douglas & Germano, 1998). Επίσης, τα SVM έχουν πολλά πλεονεκτήματα, χώρο χαρακτηριστικών υψηλών διαστάσεων. Για να είναι αποτελεσματικές οι Μηχανές Διανυσμάτων Υποστήριξης θα πρέπει ο αριθμός των δειγμάτων να είναι μικρότερος από τον αριθμό των διαστάσεων σε όλο τον αριθμό των διαθέσιμων δεδομένων. Επιπλέον, τα SVM μπορούν να είναι αποτελεσματικά στη χρήση μνήμης (El Naqa & Murphy, 2015).

Ένα σημαντικό πλεονέκτημα των SVM είναι πως είναι συμβατά με διαφορετικές συναρτήσεις πυρήνα. Έτσι το μοντέλο θα μπορέσει να μάθει τις συναρτήσεις σύνθετης απόφασης. Όμως οι μηχανές διανυσμάτων υποστήριξης έχουν και κάποια σημαντικά μειονεκτήματα. Ένα από αυτά είναι η υπερβολική προσαρμογή. Η οποία λαμβάνει χώρα όταν ο αριθμός των δειγμάτων είναι μικρότερος από τον αριθμό των χαρακτηριστικών. Τα SVM έχουν χρησιμοποιηθεί στη βιβλιογραφία για την ταξινόμηση των ADL στον τομέα της υγείας.

Ένα από τα πιο γνωστά μη εποπτευόμενα παραγωγικά μοντέλα είναι το Hidden Markov Model η πιο σύντομα HMM. Το μοντέλο αυτό πραγματεύεται καταστάσεις οι οποίες είναι κρυφές δηλαδή που δεν είναι άμεσα παρατηρήσιμες. Τα διαδοχικά σύνολα δεδομένων είναι αυτά τα οποία αξιοποιούνται από το συγκεκριμένο μοντέλο. Στις περιπτώσεις αυτές οι καταστάσεις έχουν κατανομή πιθανότητας. Το μοντέλο αυτό έχει δώσει την ονομασία του στην διαδικασία μετάβασης από μία κρυφή κατάσταση σε μία άλλη οποιαδήποτε διαδικασία. Η διαδικασία αυτή που μένος ονομάζεται διαδικασία Markov.. Η χρήση του HMM προτάθηκε για την ταξινόμηση των ADL πολυκατοικιών σε έξυπνα σπίτια (Shalev-Shwartz & Ben-David, 2014). Οι κρυφές καταστάσεις διαμορφώθηκαν ως ετικέτες των δραστηριοτήτων και οι παρατηρήσεις είναι οι αναγνώσεις των αισθητήρων Το ARAS dataset είναι το εργαλείο εκείνο το οποίο κατασκευάστηκε προκειμένου να αξιολογηθεί η ακρίβεια του μοντέλου. Το σύνολο δεδομένων αντιπροσωπεύει πραγματικές δραστηριότητες που καταγράφηκαν από πολλούς κατοίκους σε δύο πραγματικά σπίτια.

Ως Δέντρα Αποφάσεων ορίζονται εκείνα τα μοντέλα τα οποία είναι εποπτευόμενα και μη παραμετρικά. Τα μοντέλα αυτά μάθησης αξιοποιούνται συνήθως τόσο για τις διαδικασίες παλινδρομήσεις όσο και ταξινόμησης. Τα μοντέλα αυτά είναι εύκολα

τόσο στην κατανόηση όσο και την ερμηνεία τους από τους ειδικούς επιστήμονες. Αυτό συμβαίνει διότι ένα μοντέλο DT μπορεί να μάθει κανόνες απλούς συνθήκης. Ένα άλλο σημαντικό πλεονέκτημα του μοντέλου αυτού είναι ότι έχει χαμηλή χρονική πολυπλοκότητα και πολύ υψηλή απόδοση. (Shalev-Shwartz & Ben-David, 2014).

Κάποιοι τύποι δεδομένων δεν είναι συμβατοί με κάποιους αλγόριθμους μηχανικής μάθησης. Ωστόσο, τα Decision Trees είναι σε θέση να λειτουργούν τόσο με κατηγορίες όσο και με αριθμητικά σύνολα δεδομένων. Το πρόβλημα αυτό καλείται να λύσει το σύστημα το οποίο αποδεδειγμένα έχει προσφέρει σημαντικές λύσεις στο ζήτημα ταξινόμησης όταν υπάρχουν πολλές ετικέτες. Ο λόγος για τον οποίον το μοντέλο αυτό είναι αποτελεσματικό είναι ότι ο αριθμός των υπό επεξεργασία δεδομένων που χρειάζεται είναι πολύ μικρή. Για την αντιμετώπιση ζητημάτων με πολλαπλές ταξινόμησης ετικετών έχει αναπτυχθεί μία καινούργια εκδοχή του μοντέλου των δέντρων αποφάσεων το οποίο ονομάζεται E-ID5 του οποίου η αποτελεσματικότητα έχει αποδεδειγμένα αποδείχθηκε αξιολογηθεί μέσα από το σύνολο δεδομένων ARAS. (Raschka, 2015).

Στην περίπτωση που επιθυμούμε να βρούμε την ελάχιστη αλλά και τη μέγιστη τιμή μιας συνάρτησης μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο Stochasting Gradient Descent η πιο σύντομα SGD. Ο αλγόριθμος αυτός ο οποίος είναι επαναληπτικός συνδυάζεται συνήθως με της συνάρτησης κυρτής απώλειας. Στόχος είναι ο ερευνητής να βρει το ελάχιστο σφάλμα. Παράλληλα μπορεί να συνδυαστεί και με γραμμικούς ταξινομητές. Στην περίπτωση αυτή σκοπός είναι η επίλυση προβλημάτων ταξινόμησης αλλά και ταξινόμησης πολλαπλών ετικετών. Παρόλα αυτά είναι σημαντικό πλεονέκτημα είναι πως πριν από την χρήση του κάποιες σημαντικές υπέρ παράμετροι όπως αριθμός των επαναλήψεων πρέπει πρώτα να διευθετηθούν. (Goldberg & Holland, 1988). Εφαρμόστηκε η λογιστική παλινδρόμηση με τον αλγόριθμο SGD προκειμένου να αναπτυχθεί μια επεκτάσιμη διάγνωση μοντέλο για εφαρμογές υγειονομικής περίθαλψης.

2.4. Παλινδρόμηση

Στη μηχανική μάθηση η διαδικασία της παλινδρόμησης αποτελείται από μία σειρά μαθηματικών μεθόδων. Οι μέθοδοι αυτοί δίνουν τη δυνατότητα στους εκάστοτε επιστήμονες να μπορέσουν με βάση την τιμή μιας μεταβλητής (x) να κάνουν προβλέψεις και να συνάγουν ένα συνεχές αποτέλεσμα (y). Η γραμμική παλινδρόμηση είναι ίσως η πιο δημοφιλής μορφή ανάλυσης παλινδρόμησης λόγω της ευκολίας χρήσης της στην πρόγνωση και την πρόβλεψη (Raschka, 2015).

Πιο συγκεκριμένα με τον όρο ανάλυση παλινδρόμησης εννοούμε την μέθοδο εκείνη για την απόδοση της σχέσης που υπάρχει ανάμεσα σε εξαρτημένες και ανεξάρτητες μεταβλητές. Οι ανεξάρτητες αυτές μεταβλητές που ονομάζονται και προβλεπτικές μπορούν να είναι μία ή περισσότερες. Μέσω της ανάλυσης παλινδρόμησης έχουμε

τη δυνατότητα να δούμε τον τρόπο με τον οποίον η τιμή της εξαρτημένης μεταβλητής η οποία σχετίζεται με μία ανεξάρτητη μεταβλητή μεταβάλλεται. Στην περίπτωση αυτή οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν αμετάβλητες. Η ανάλυση παλινδρόμησης μπορεί να χρησιμοποιεί μόνο συνεχής τιμές. Για παράδειγμα μία τέτοια τιμή είναι η θερμοκρασία. (El Naqa & Murphy, 2015).

Στον τομέα της μηχανικής μάθησης η παλινδρόμηση είναι ο στατιστικός εκείνος όρος με τον οποίον μπορούμε να ανακαλύψουμε τη συσχέτιση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών. Συγχρόνως μπορούμε να προβλέψουμε τη διακύμανση της τιμής της μεταβλητής συνεχούς εξόδου σε συνάρτηση με τις μεταβλητές προβλέψεις στις οποίες διαθέτουμε. Χρησιμοποιείται κυρίως για την πρόβλεψη, την πρόβλεψη, τη μοντελοποίηση χρονοσειρών και τον προσδιορισμό της σχέσης αιτιώδους αποτελέσματος μεταξύ των μεταβλητών.

Στα πλαίσια χρήσης της παλινδρόμησης κατασκευάζουμε πάντα ένα γράφημα που αποτυπώνει τις ανεξάρτητες και εξαρτημένες μεταβλητές. Με βάση το διάγραμμα αυτό το μοντέλο μηχανικής μάθησης που χρησιμοποιούμε πραγματοποιεί προβλέψεις. Στο διάγραμμα αυτό παρουσιάζεται συνήθως μία γραμμή ή μία καμπύλη. Η γραμμή αυτή για περνάει από όλα τα σημεία με τέτοιο τρόπο ώστε η απόσταση μεταξύ των σημείων δεδομένων και της γραμμής να είναι η μικρότερη δυνατή. Η απόσταση αυτή ανάλογα με το πόσο μεγάλη ή μικρή είναι καταδεικνύει αν υπάρχει ισχυρή σχέση μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών. (Goldberg and Holland, 1988)

Παρακάτω θα δούμε κάποια παραδείγματα παλινδρόμησης.

Απόδοση των τάσεων του πληθωρισμού.

Πρόβλεψη τροχαίων ατυχημάτων σε συνάρτηση με την κατανάλωση αλκοόλ.

Πρόβλεψη της θερμοκρασίας σε συνάρτηση με άλλους παράγοντες.

Ορολογίες που σχετίζονται με την ανάλυση παλινδρόμησης (Goodfellow, Bengio & Courville, 2016):

- Εξαρτημένη μεταβλητή: Ονομάζεται και μεταβλητή στόχος και είναι ο παράγοντας εκείνος τον οποίον θέλουμε να προβλέψουμε
- Ανεξάρτητη μεταβλητή: Ονομάζονται και προβλέψεις και είναι οι παράγοντες εκείνοι που μεταβάλλουν και επηρεάζουν τις τιμές των εξαρτημένων μεταβλητών.
- Outliers: Τα outliers είναι παράγοντες με πολύ ακραίες υψηλές και χαμηλές τιμές οι οποίες συνήθως αποφεύγονται και δεν λαμβάνονται υπόψη για να μην περιορίζουν το αποτέλεσμα μας.
- Πολυσυγγραματκότητα. Η συνθήκη κατά την οποία οι ανεξάρτητες μεταβλητές σχετίζονται μεταξύ τους σε πολύ μεγάλο βαθμό. Η κατάσταση

αυτή επίσης θα πρέπει να αποφεύγεται και να μην υπάρχει στο σύνολο των δεδομένων μας.

- Υποπροσαρμογή και υπερπροσαρμογή: Υποπροσαρμογή ονομάζεται η κατάσταση εκείνη κατά την οποία ο αλγόριθμος δεν λειτουργεί καλά με το σύνολο των δεδομένων δοκιμής. Παρόλα αυτά λειτουργεί καλά με το σύνολο δεδομένων εκπαίδευσης. Αντίθετα υπέρπροσαρμογή είναι η κατάσταση όπου αλγόριθμος λειτουργεί καλά Και με τις δύο ομάδες δεδομένων

Στην καθημερινότητά μας υπάρχουν αρκετές συνθήκες και καταστάσεις στις οποίες χρειάζεται να πραγματοποιήσουμε σταθερές και στοχευμένες προβλέψεις. Ένα τέτοιο παράδειγμα είναι ο καιρός οι τάσεις των πωλήσεων της αγοράς και του πληθωρισμού. Η ανάλυση παλινδρόμησης είναι η στατιστική εκείνη μέθοδος με βάση την οποία μπορούμε να κάνουμε τέτοιες προβλέψεις. Έτσι και στην μηχανική μάθηση η ανάλυση παλινδρόμησης χρησιμοποιείται προκειμένου να εξαχθούν ασφαλής και έγκυρες προβλέψεις όσον αφορά την ανάλυση των δεδομένων.(Murphy, 2012)

Παρακάτω παρουσιάζονται άλλες εφαρμογές της ανάλυσης παλινδρόμησης.

- Πρόβλεψη σχέσης μεταξύ οι εξαρτημένες και ανεξάρτητες μεταβλητές.
- Εξεύρεση τάσεων ανάμεσα στα παρεχόμενα δεδομένα.
- Δυνατότητα προσδιορισμού με ακρίβεια των σημαντικών και λιγότερο σημαντικών παραγόντων αλλά και των μεταξύ τους αλληλεπιδράσεων.
- Πρόβλεψη των συνεχών τιμών.

Όπως σε όλες τις επιστήμες έτσι και στην μηχανική μάθηση αξιοποιούνται διαφορετικοί τύποι παλινδρόμησης. Παρόλες τις διαφορές που έχουν μεταξύ τους βασικός στόχος όλων των τύπων παλινδρόμησης είναι η αποτύπωση της σχέσης εξάρτησης και αλληλεπίδρασης μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών. (Goodfellow, Bengio & Courville, 2016):

Μερικοί από τους πιο ευρέως χρησιμοποιούμενος τύπους παλινδρόμησης είναι η λογιστική παλινδρόμηση, η πολυωνυμική παλινδρόμηση, η παλινδρόμηση δέντρου απόφασης, η γραμμική παλινδρόμηση, η τυχαία παλινδρόμηση δασών, η παλινδρόμηση κορυφογραμμής και τέλος η υποστήριξη διανυσματικής παλινδρόμησης.

2.5 Μάθηση χωρίς επίβλεψη

Με τον όρο μάθηση χωρίς επίβλεψη ή μη επιβλεπόμενη μάθηση ονομάζεται η μέθοδος χρήσης αλγορίθμων τεχνητής νοημοσύνης (AI). Σκοπός της χρήσης αυτής είναι ο εντοπισμός μοτίβων μέσα από συγκεκριμένα σύνολα δεδομένων. Τα δεδομένα αυτά δεν είναι ούτε επισημασμένα ούτε ταξινομημένα.

Πιο συγκεκριμένα ή μη εποπτευόμενη μάθηση ανήκει επίσης στο χώρο της μηχανικής και εκμάθησης. Η διαφορά με τους υπόλοιπους τύπους είναι ότι στην περίπτωση αυτή ο αλγόριθμος δεν έχει ετικέτες προκαθορισμένες. Για τον λόγο αυτόν οι αλγόριθμοι στην περίπτωση αυτή βρίσκουν πρώτα από μόνοι τους μοτίβα μέσα στα παρεχόμενα σύνολα δεδομένων. (Goodfellow, Bengio & Courville, 2016).

Ένα πιο χαρακτηριστικά παραδείγματα αλγορίθμων μάθησης χωρίς επίβλεψης είναι η ανάλυση βασικών στοιχείων και το παράδειγμα της ομαδοποίησης. Στην πρώτη περίπτωση ο αλγόριθμος προσπαθεί να συμπιέσει τα παρεχόμενα δεδομένα. Παράλληλα προσπαθεί να αναδείξει τις λειτουργίες εκείνες που χρησιμοποιούνται προκειμένου να πραγματοποιηθεί Η διάκριση μεταξύ των διαφόρων παραδειγμάτων. Στο παράδειγμα της ομαδοποίησης ο αλγόριθμος καλείται να ομαδοποιήσει τα παραδείγματα αυτά. Η ομαδοποίηση γίνεται πάντα σε κατηγορίες με παρόμοια χαρακτηριστικά..

Αυτό έρχεται σε αντίθεση με την εποπτευόμενη μάθηση στην οποία τα δεδομένα εκπαίδευσης περιλαμβάνουν προκαθορισμένες ετικέτες κατηγοριών (συχνά από άνθρωπο ή από την έξοδο αλγορίθμου ταξινόμησης μη μάθησης). Η ενισχυτική μάθηση και η μη εποπτευόμενη μάθηση είναι τα υπόλοιπα ενδιάμεσα στάδια που περιλαμβάνονται στο φάσμα της εποπτείας. Στην ενισχυτική μάθηση έχουμε διαθέσιμες βαθμολογίες μόνο αριθμητικές για τα διαθέσιμα παραδείγματα ενός στην ημι εποπτευόμενη μάθηση έχουμε επισήμανση μόνο σε ένα συγκεκριμένο μέρος των διαθέσιμων δεδομένων (Mohri, Rostamizadeh & Talwalkar, 2018).

Τα πλεονεκτήματα της μάθησης χωρίς επίβλεψη περιλαμβάνουν τον ελάχιστο φόρτο εργασίας για την προετοιμασία και τον έλεγχο του συνόλου εκπαίδευσης, σε αντίθεση με τις τεχνικές εποπτευόμενης μάθησης όπου απαιτείται σημαντικός αριθμός εξειδικευμένου ανθρώπινου δυναμικού για την ανάθεση και την επαλήθευση των αρχικών ετικετών και μεγαλύτερη ελευθερία αναγνώρισης και εκμετάλλευσης προηγούμενων μη ανιχνευμένων προτύπων. που μπορεί να μην έχουν προσέξει οι «ειδικοί» (Harrington, 2012).

Από την άλλη η μέθοδος αυτή έχει και κάποια μειονεκτήματα. Το πιο συνηθισμένο μειονέκτημα των τεχνικών αυτών είναι ότι χρειάζεται μεγαλύτερο χρόνο για να φτάσουν σε υψηλές επιδόσεις ενώ απαιτείται και ένας ιδιαίτερα μεγάλος όγκος διαθέσιμων δεδομένων. Επίσης ως προαπαιτούμενα τίθενται ιδιαίτερα αυξημένες απαιτήσεις σχετικά με τον υπολογισμό και την αποθήκευση των δεδομένων. Συγχρόνως υπάρχει και συγκριτικά αρκετά μεγάλη ευαισθησία των τεχνουργημάτων

αλλά και ανωμαλίες στα παρεχόμενα δεδομένα εκπαίδευσης. Τα δεδομένα αυτά μπορεί να είναι λανθασμένα η άσχετα αλλά παρόλα αυτά ο αλγόριθμος της μάθησης χωρίς επίβλεψη τους αποδίδει εσφαλμένα υψηλή σημασία επηρεάζοντας το αποτέλεσμα. (Goodfellow, Bengio & Courville, 2016).

2.5. Μετασχηματισμός

Εξαιτίας του τεράστιου όγκου δεδομένων που έχουν να διαχειριστούν οι διάφοροι οργανισμοί σήμερα χρειάζεται να εξασφαλίσουν ότι ανάλυση της ποσότητας αυτής των δεδομένων θα γίνει σωστά χωρίς λάθη με ακρίβεια και αξιοπιστία. Για το λόγο αυτό προσπαθούμε η ποιότητα των δεδομένων που έχουν να είναι αξιόπιστη και έγκυρη. Στην ιδανική περίπτωση, προτιμάται η διεξαγωγή ανάλυσης σε καθαρά δεδομένα που δεν έχουν άσχετες τιμές, αλλά στην πραγματική ζωή, αυτού του είδους τα δεδομένα είναι σπάνια διαθέσιμα (Goodfellow, Bengio & Courville, 2016).

Οι επιστήμονες δεδομένων προκειμένου να διασφαλίσουν ότι η ποιότητα των δεδομένων που διαθέτουν είναι αξιόπιστη ότι βρίσκονται στην σωστή μορφή και ικανοποιούν όλους τους όρους και τις προϋποθέσεις χρειάζονται να επεξεργάζονται τα δεδομένα προς ανάλυση πριν τις βασικές διαδικασίες. Είναι χαρακτηριστικό πως πολλές έρευνες με συνεντεύξεις τέτοιων επιστημόνων καταδεικνύουν ότι ένα πολύ μεγάλο μέρος των διαδικασιών ανάλυσης δεδομένων καλύπτονται από διαδικασίες τροποποίησης των δεδομένων ώστε να είναι στην σωστή μορφή. (Jordan & Mitchell, 2015).

Ο μετασχηματισμός και η προ επεξεργασία των δεδομένων πριν από την κυρία ανάλυση είναι ζωτικής σημασίας προκειμένου τα αποτελέσματα της ανάλυσης να είναι ποιοτικά και αξιόπιστα. Είναι μία κεριά διαδικασία που σε μεγάλο βαθμό εξασφαλίζει την επιτυχία της ανάλυσης των δεδομένων. Ο μετασχηματισμός αυτός και λαμβάνει κυρίως υπόψιν χαρακτηριστικά των δεδομένων όπως την συλλογή και τη δομή του στις στατιστικές του ιδιότητες. Οι επιστήμονες των δεδομένων χρειάζεται και συνήθως διαθέτουν την απαραίτητη εμπειρία και γνώση προκειμένου η προεπεξεργασία και ο μετασχηματισμός να πραγματοποιηθούν κατάλληλα. (Mohri, Rostamizadeh & Talwalkar, 2018).

Στη σύγχρονη αγορά υπάρχει μία μεγάλη πληθώρα διαφορετικών εργαλείων και τεχνικών προκειμένου ο επιστήμονας δεδομένων να τα επεξεργαστεί και να τα μετασχηματίσει με αποτελεσματικότητα. Αξίζει να σημειωθεί ότι τα πιο ευρέως χρησιμοποιούμενα πλαίσια και γλώσσες όπως για παράδειγμα το R5 ή το scikit-learn 4 περιλαμβάνουν και κάποιες επιπρόσθετες μεθόδους μετασχηματισμού παράλληλα με τα ήδη υπάρχοντα εργαλεία.

Παρόλα αυτά αυτός ο μεγάλος αριθμός διαθέσιμων εργαλείων μετασχηματισμού μπορεί να δυσκολέψει ιδιαίτερα τους επιστήμονες δεδομένων προκειμένου να προβούν στην κατάλληλη επιλογή και στην σωστή διαδικασία προεπεξεργασίας

αλλά και στη συνέχεια να παρουσιάσουν τις αλλαγές αυτές με διαδραστικό τρόπο. Είναι μία ιδιαίτερα χρονοβόρα και δύσκολη διαδικασία η οποία παρόλα αυτά θα μπορούσε να γίνει πιο λειτουργική και εύκολη. Μία μέθοδος είναι η αξιοποίηση μιας διαδραστικής γραφικής διεπαφής χρήστη GUI με σκοπό τον αποτελεσματικό μετασχηματισμό δεδομένων. (Goodfellow, Bengio & Courville, 2016).

2.6. Συσταδοποίηση

Ως συσταδοποίηση η αλγόριθμος μάθησης χωρίς επίβλεψη ονομάζεται εκείνος ο αλγόριθμος που αναλαμβάνει να διαχωρίζει τα δεδομένα όταν ο όγκος τους είναι πολύ μεγάλος σε ομάδες δεδομένων. Αυτό συμβαίνει όταν πιο συγκεκριμένα δεν υπάρχουν ετικέτες που να διευκολύνουν τη διαδικασία επεξεργασίας των δεδομένων. Ο αλγόριθμος της συσταδοποίησης είναι από τους πιο ευρέως χρησιμοποιούμενος στον χώρο της μηχανικής μάθησης. (Harrington, 2012).

Ο αλγόριθμος συσταδοποίησης η ομαδοποίησης αναλαμβάνει να ταξινομήσει όλα τα σημεία δεδομένων μέσα σε ένα σύμπλεγμα. Καθένα από αυτά τα συμπλέγματα περιέχει ένα σύνολο δεδομένων. Μία σημαντική παράμετρος είναι πως μέσα στο ίδιο σύμπλεγμα τα σημεία των δεδομένων θα πρέπει να έχουν παρόμοια χαρακτηριστικά ενώ τα δεδομένα σε διαφορετικά συμπλέγματα θα πρέπει να έχουν πολύ διαφορετικά και ανόμοια χαρακτηριστικά. Η μέθοδος της συσταδοποίησης έχει αξιοποιηθεί σε πολλές ερευνητικές εργασίες με μερικούς από τους πιο γνωστούς αλγόριθμους να είναι τα k-means k-midoids ή partitioning around midoids and hierarchical. (Marsland, 2011).

Οι κύριες διαδικασίες αυτών των αλγορίθμων ταξινομούνται ως εξής: Ο αλγόριθμος ομαδοποίησης K-means είναι ένας απλός και πιο γενικός αλγόριθμος ομαδοποίησης που χρησιμοποιείται κυρίως για την ταξινόμηση του δεδομένου δεδομένων που δεν έχει ετικέτα. Αυτός ο αλγόριθμος στοχεύει κυρίως στην εύρεση όμοιων συστάδων που αντιπροσωπεύονται από τη μεταβλητή k (Mitchell, 2006).

Η παραπάνω αλγοριθμικές ομαδοποιήσεις παρουσιάζουν διαφορετικά χαρακτηριστικά και ιδιότητες. Για παράδειγμα ο πρώτος αλγόριθμος, ο αλγόριθμος k θεωρείται ο πιο απλός από όλους. Χρησιμοποιείται με στόχο την ταξινόμηση των δεδομένων χωρίς ετικέτα καθώς και στην εξεύρεση συστάδων. Είναι σημαντικό όμως οι συστάδες αυτές να αντιπροσωπεύονται και να οριοθετούνται από την μεταβλητή K. Για το χαρακτηρισμό του συμπλέγματος αξιοποιείται το μέσο orcentroid. Στο σημείο αυτό πρέπει να τονιστεί ότι κέντρο ονομάζεται ένα σημείο δεδομένων στο σύμπλεγμα. Το σημείο αυτό των δεδομένων Σε πολλές περιπτώσεις δεν είναι απαραίτητο μέλος των δεδομένων που τίθενται προς επεξεργασία. (Mitchell, 2006).

Με τον τρόπο αυτόν το σύνολο των δεδομένων διαιρείται σε κάπα αριθμό συστάδων. Έτσι τα σημεία δεδομένων υπεισέρχονται και ανήκουν στα κατάλληλα συμπλέγματα..

Μετά την παραπάνω διαίρεση υπολογίζεται η Ευκλείδεια απόσταση. Η απόσταση αυτή αφορά τόσο τα σημεία δεδομένων αλλά και την απόσταση από το κέντρο μέσα στο σύμπλεγμα. Αξίζει να σημειωθεί πως τα σημεία αυτά με βάση την απόσταση από το κέντρο εκχωρούνται σε αυτό. Όταν δεν υπάρχει διαθέσιμο σημείο δεδομένων για εκχώρηση, λαμβάνεται υπόψη η προκαταρκτική ομαδοποίηση.

Σε αυτήν την περίπτωση, τα νέα κεντροειδή «c» υπολογίζονται εκ νέου, επομένως η νέα επανάληψη συνεχίζεται έως ότου τα κεντροειδή «c» σταματήσουν να αλλάζουν τη θέση τους. Στον αλγόριθμο της συσταδοποίησης λειτουργεί ως μεσοειδής στο εκάστοτε σύμπλεγμα πού βρίσκεται κεντρικά το σημείο δεδομένων. Μάλιστα εντός του συμπλέγματος αυτού η απόσταση μεταξύ των σημείων θα πρέπει να είναι η ελάχιστη δυνατή. Έτσι γίνεται αντιληπτό πως αναφορικά με την αντιπροσωπεία των υπόλοιπων σημείων μπορεί να αξιοποιηθεί το thismedoid εντός του συμπλέγματος. (Harrington, 2012).

Η κύρια ιδέα του PAM είναι να υπολογίσει πρώτα το κύριο σημείο δεδομένων ως medoid σε ένα συγκεκριμένο σύμπλεγμα, να ομαδοποιήσει το σύνολο των medoids και στη συνέχεια κάθε σημείο δεδομένων εκχωρείται στο πλησιέστερο medoid σε ένα δεδομένο σύμπλεγμα. Build και swapface είναι ίδιο φάσεις του αλγόριθμου αυτού. Κατά τη διάρκεια της πρώτης φάσης σκοπός είναι να επιλεγεί το medoid εκείνο το οποίο έχει τη χαμηλότερη μέση ανομοιότητα αναφορικά με το υπόλοιπο δεδομένων. Το medoid αυτό στη συνέχεια ορίζεται ως σημείο δεδομένων. Αντίθετα κατά τη διάρκεια της swapface γίνεται αξιολόγηση σε όλα τα σημεία δεδομένων με βάση το σύνολο των k medoids. Έτσι μέσα από τις διαδικασίες αυτές προκύπτει ένα καινούργιο medoid. Το medoid αυτό δημιουργείται καθώς πραγματοποιείται ανταλλαγή δεδομένων ανάμεσα στο παλιό medoid και στα σημεία δεδομένων ενός καινούργιου μη medoid. (Jordan & Mitchell, 2015).

Τέλος έχουμε τον αλγόριθμο hierarchical ή hierarchical cluster analysis. Ο συγκεκριμένος αλγόριθμος ομαδοποίησης είναι ένας ειδικός τύπος και αποσκοπεί στην ανάλυση των συστάδων με έναν ιεραρχικό τρόπο. Πιο συγκεκριμένα Ο σκοπός είναι μη επισημασμένα σημεία δεδομένων με παρόμοια χαρακτηριστικά να ομαδοποιηθούν σε έναν συγκεκριμένο αριθμό συστάδων. Η δομή που χρησιμοποιείται για την διαδικασία αυτή θυμίζει ένα δέντρο. Έτσι στο τέλος δημιουργείται ένας αριθμός συστάδων με διακριτά μεταξύ τους χαρακτηριστικά. Τέλος πρέπει να σημειωθεί πώς μέσα σε ένα σύμπλεγμα τα χαρακτηριστικά των δεδομένων είναι παρόμοια ή και ταυτόσημα σε σχέση με τα υπόλοιπα συμπλέγματα και τα σημεία δεδομένων που αυτά περιλαμβάνουν.

Αυτός ο αλγόριθμος έχει την δομή ενός δέντρου που ονομάζεται δενδρογραμμάριο (Langley, 1996).

Αυτός ο αλγόριθμος έχει την δομή ενός δέντρου που ονομάζεται δενδρογραμμάριο. Ο αλγόριθμος αυτός στηρίζεται στην ιεραρχία. Συγχρόνως ο αλγόριθμος αυτός αποδίδει δύο διαφορετικούς τύπους οι οποίοι είναι η Σωρευτική Ιεραρχική Ομαδοποίηση ή Agnes και η Διαίρεση Ιεραρχικής Ομαδοποίησης ή Diana. (Divisive Analysis) (Marsland, 2011).

3. Αλγόριθμοι Κατηγοριοποίησης

3.1. Βασικά Χαρακτηριστικά

Με βάση το αν γνωρίζουμε ετικέτες για τα δεδομένα, οι αλγόριθμοι μπορούν να ταξινομηθούν σε δύο μεγάλες κατηγορίες – εποπτευόμενης μάθησης και μάθησης χωρίς επίβλεψη. Οι αλγόριθμοι εποπτευόμενης μάθησης χρησιμοποιούνται όταν είναι γνωστές οι ετικέτες των δεδομένων που πρόκειται να γίνουν (Jain, Topchy, Law & Buhmann, 2004).

Ο εντοπισμός της ανεπιθύμητης αλληλογραφίας αποτελεί καίριο παράδειγμα κατά το οποίο η εποπτευόμενη μάθηση μπορεί να χρησιμοποιηθεί για ταξινόμηση. Μια ακόμη εφαρμογή της εποπτευόμενης μάθησης είναι η πρόβλεψη μιας αριθμητικής τιμής στην παλινδρόμηση (Murtagh & Contreras, 2012).

Ο στόχος της εποπτευόμενης μάθησης είναι να πραγματοποιηθεί η χαρτογράφηση από την είσοδο στην έξοδο των οποίων οι σωστές τιμές παρέχονται από τον επόπτη. Η μάθηση χωρίς επίβλεψη χρησιμοποιείται σε προβλήματα ταξινόμησης όπου οι ετικέτες για τα δεδομένα δεν είναι γνωστές. Στόχος της εποπτευόμενης εκμάθησης είναι να βρει κανονικότητες ή να κάνεις συσχετίσεις των δεδομένων εισόδου, χωρίς ρητή ανάγκη επόπτη (Narmadha, alias Balamurugan, Sundar & Priya, 2016).

3.2. Λειτουργία, Πλεονεκτήματα και Μειονεκτήματα

Αυτή η ενότητα παρέχει μια σύντομη επισκόπηση ορισμένων από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους κατηγοριοποίησης με πλεονεκτήματα και μειονεκτήματα για τους αλγόριθμους

3.2.1. Naive Bayes

Ο Naive Bayes είναι ένας από τους απλούστερους αλγόριθμους μάθησης. Το πιο ενδιαφέρον είναι ότι σε ορισμένες περιπτώσεις μπορεί να ξεπεράσει τους περισσότερους εξελιγμένους αλγόριθμους μάθησης. Ο Naive Bayes χρησιμοποιεί εκτίμηση μέγιστης πιθανότητας για να ταξινομήσει νέα παραδείγματα. Βασίζεται στο θεώρημα του Bayes το οποίο λέει,

$P(A|B) = P(B|A) \times P(A) / P(B)$

Όπου το P(A), το P(B) είναι οι πιθανότητες του A και του B, και το P(A|B) και το P(B|A) είναι πιθανότητες υπό όρους του A, δεδομένου του B και του B, δεδομένου του A αντίστοιχα. Όταν εκπαιδεύονται οι Naive Bayes η πιθανότητα εύρεσης ενός παραδείγματος κάθε τάξης στον πληθυσμό δείγματος υπολογίζεται ως η

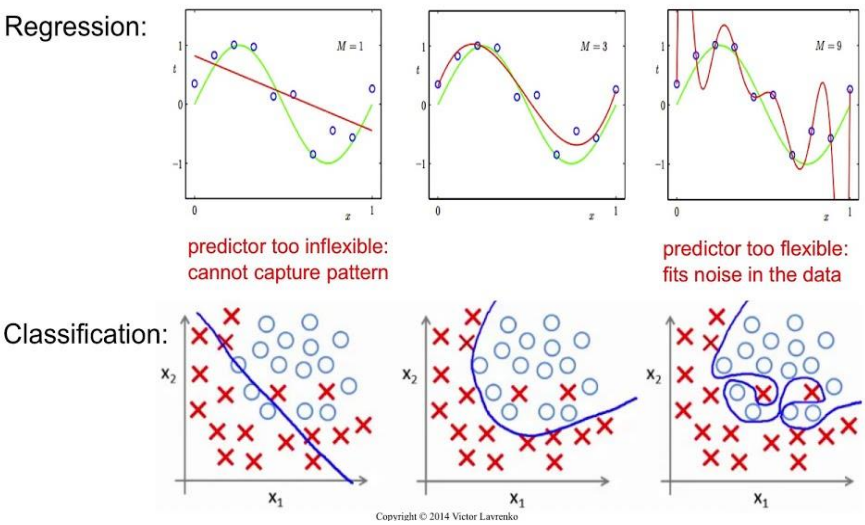
προηγούμενη πιθανότητα για αυτήν την τάξη. Ακόμη βρίσκει την πιθανότητα για τις περιπτώσεις x δεδομένης της κλάσης c. Με την υπόθεση ότι τα χαρακτηριστικά του x είναι ανεξάρτητα, γίνεται απλώς γινόμενο των πιθανοτήτων κάθε μεμονωμένου χαρακτηριστικού (Jain, Topchy, Law & Buhmann, 2004).

Ως εκ τούτου, το θεώρημα Bayes, όταν εφαρμόζεται στο πρόβλημα ταξινόμησης χρησιμοποιώντας το Naive Bayes είναι:

$P(Ci\ |x) = P(x|Ci) \times P(C) / P(x)$

Η πιθανότητα $P(Ci\ |x)$ αναφέρεται ως μεταγενέστερη πιθανότητα.

Μια κλάση Ci επιλέγεται εάν $P(Ci\ |x) = \max k\ \{P(Ck|x)\}$, δηλαδή η κλάση με την υψηλότερη οπίσθια πιθανότητα. Ένα σαφές πλεονέκτημα της χρήσης του Naive Bayes είναι ότι είναι γρήγορο στην εκπαίδευση και γρήγορο στην ταξινόμηση των δεδομένων (Fahad, Alshatri, Tari, Alamri, Khalil, Zomaya & Bouras, 2014).



Εικόνα 1 - Παράδειγμα υπερπροσαρμογής και υποσυναρμολόγησης

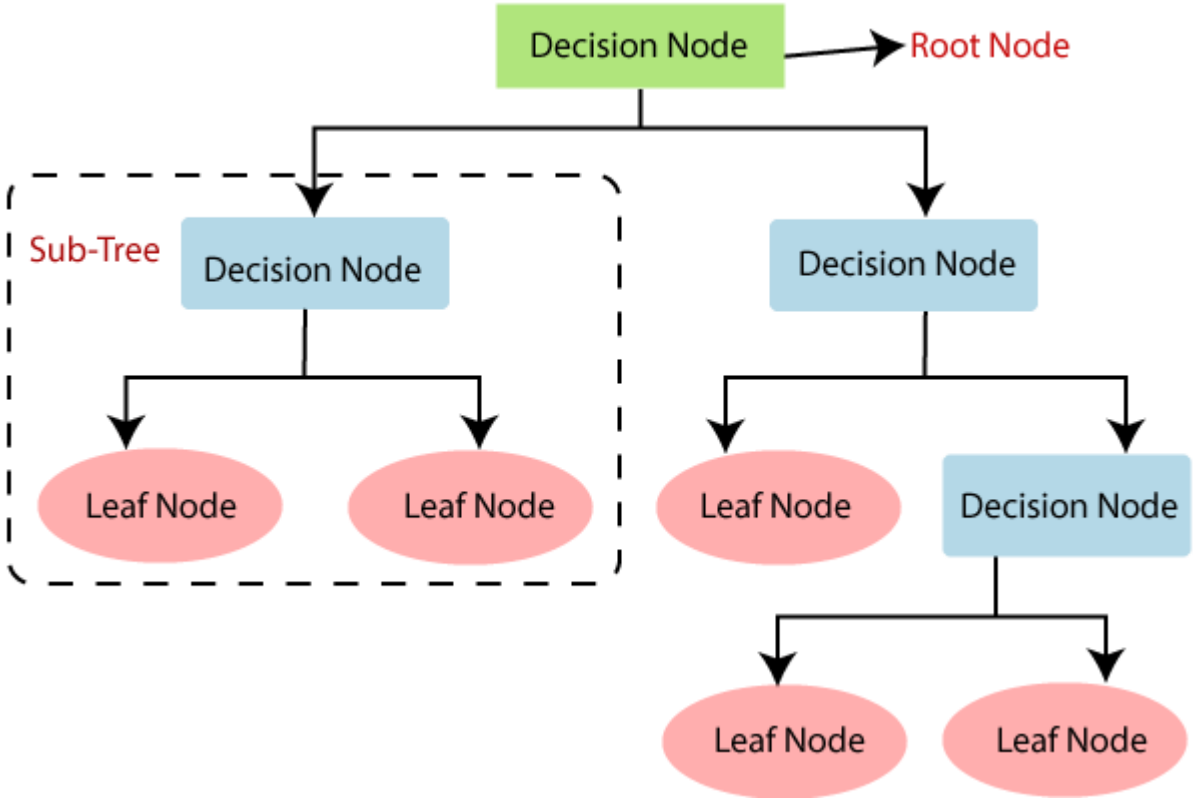
3.2.2. Δέντρα απόφασης

Τα δέντρα αποφάσεων είναι ιεραρχικά μοντέλα, στα οποία κάθε βήμα αποτελεί μια απλή συνάρτηση ελέγχου κατωφλίου ονομαστικής τιμής ενός χαρακτηριστικού έναντι μιας σταθερής τιμής κατωφλίου (Benabdellah, Benghabrit & Bouhaddou, 2019).

Τα βήματα στην ιεραρχία ονομάζονται κόμβοι απόφασης και μια δοκιμή υλοποιείται με τη μορφή μιας συνάρτησης στα χαρακτηριστικά x ενός παραδείγματος, με διακριτά αποτελέσματα να αναπαρίστανται ως κλάδοι. Όσον αφορά το πρόγραμμα ενός υπολογιστή, αυτός ο αλγόριθμος δημιουργεί ένα σύνολο κανόνων που μπορούν να ερμηνευθούν ως ένθετη δομή IF-ELSE. Οι αλγόριθμοι εκμάθησης δένδρων αποφάσεων χρησιμοποιούνται για την εξαγωγή δέντρων αποφάσεων. Τα ID3, C4.5 είναι μερικά παραδείγματα αυτών των αλγορίθμων. Ο ID3 είναι ο απλούστερος από αυτούς τους αλγόριθμους (Fahad, Alshatri, Tari, Alamri, Khalil, Zomaya & Bouras, 2014).

Στην περίπτωση των δέντρων απόφασης, η εκπαίδευση είναι να επιλέξει κανείς χαρακτηριστικά που παρέχουν τις περισσότερες πληροφορίες σχετικά με το σύνολο εκπαίδευσης. Στη συνέχεια κατασκευάζει δέντρο χρησιμοποιώντας προσέγγιση από πάνω προς τα κάτω.

Άλλοι προηγμένοι αλγόριθμοι όπως ο RIPPER [C +95] βασίζονται στην ίδια προσέγγιση και στη συνέχεια χρησιμοποιούν κλάδεμα για να μειώσουν το σφάλμα εκπαίδευσης. Τα δέντρα αποφάσεων χρησιμοποιούν μια προσέγγιση «λευκού κουτιού», όπου η εσωτερική λήψη αποφάσεων και η δομή του δέντρου είναι ορατά στον χρήστη. Αυτό καθιστά επίσης εύκολη την οπτικοποίηση και την ερμηνεία των δέντρων αποφάσεων (Murtagh & Contreras, 2012).



Εικόνα 2 - Decision tree classification algorithm

3.2.3. Perceptrons Multi-Layer (MLP)

Τα Perceptrons Multi-Layer (MLP) αποτελούν έναν τύπο μοντέλων τεχνητών νευρωνικών δικτύων και χρησιμοποιούνται από τις αρχές της δεκαετίας του '80. Σε αυτό το μοντέλο, κάθε χαρακτηριστικό και έξοδος αποτελούν κόμβους χαρακτηριστικών για κάθε στρώμα και συνδέονται με το ανώτερο στρώμα χρησιμοποιώντας βάρη ή συνάψεις.

Ένα παράδειγμα απλού perceptron έχει δύο στρώσεις:

Είσοδοι x1, x2. . .

Τα xk είναι χαρακτηριστικά και το x0 = +1 είναι ένα στοιχείο μεροληψίας επιτρέποντας στον χρήστη να ρυθμίσει με ακρίβεια την έξοδο μετατοπίζοντας τη συνάρτηση εξόδου. Τα α, β είναι πίνακες βαρών στις συνάψεις στο πρώτο προς το κρυφό στρώμα και το κρυφό στρώμα προς την έξοδο, αντίστοιχα (Murtagh & Contreras, 2012).

Η έξοδος του perceptron μπορεί να αναπαρασταθεί μαθηματικά ως:

y = Xn j=0 bj (X k i=1 (aijxi + a0))

Κατά την εκπαίδευση ενός perceptron, ο αλγόριθμος εκπαίδευσης θα προσπαθήσει να βρει τα κατάλληλα συνδετικά βάρη. Μπορούν να κατασκευαστούν πολλαπλά στρώματα perceptrons εφαρμόζοντας ένα κρυφό στρώμα κόμβων μεταξύ των χαρακτηριστικών και της εξόδου, με αυτόν τον τρόπο μπορεί κανείς να εφαρμόσει μη γραμμικές συναρτήσεις εξόδου.

Ο backpropagation είναι ένας από τους πιο χρησιμοποιούμενους αλγόριθμους για την εκπαίδευση MLP. Λειτουργεί υπολογίζοντας τις συσχετίσεις σφαλμάτων σε κάθε έξοδο και τις χρησιμοποιεί για να υπολογίσει τους όρους σφάλματος στα προηγούμενα επίπεδα και ούτω καθεξής (Benabdellah, Benghabrit & Bouhaddou, 2019).

Οι όροι σφάλματος χρησιμοποιούνται στη συνέχεια για την προσαρμογή των βαρών των μεμονωμένων συνάψεων.

Η συνάρτηση σφάλματος σε αυτήν την περίπτωση ορίζεται ως:

E(a, b|X) = 1 2 X t (l t − y t) 2

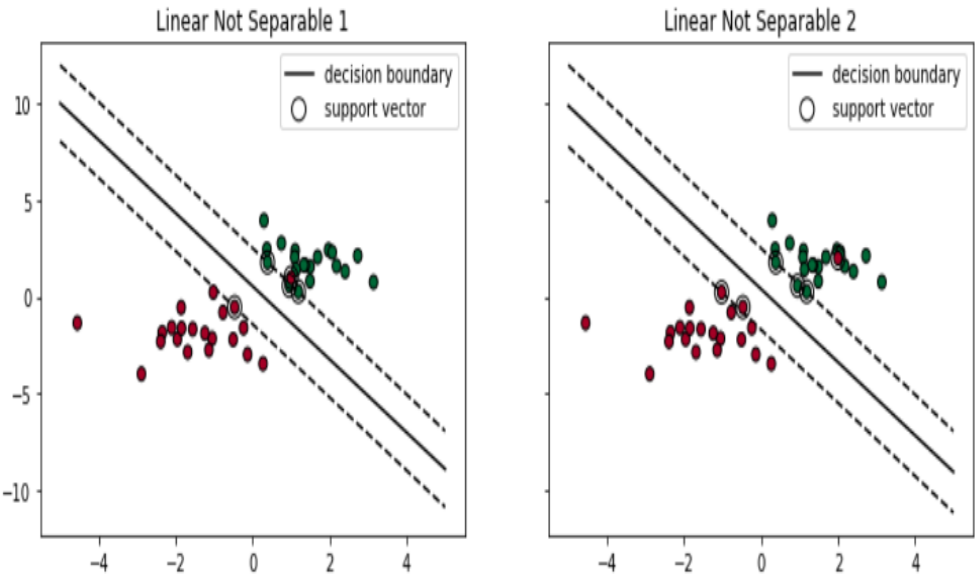
Η διαβάθμιση αυτής της συνάρτησης σφάλματος υπολογίζεται κατά την εκ νέου διάδοση και τα βάρη ενημερώνονται μόλις η συνάρτηση διαβάθμισης φτάσει στο τοπικό ελάχιστο. Μόλις εκπαιδευτούν, οι MLP είναι σε θέση να ταξινομούν γρήγορα τα δεδομένα (Fahad, Alshatri, Tari, Alamri, Khalil, Zomaya & Bouras, 2014).

Ένα μειονέκτημα είναι ότι το δίκτυο πρέπει να επανεκπαιδευτεί πλήρως όταν πρόκειται να προστεθούν νέα δεδομένα εκπαίδευσης. Η επιλογή των χαρακτηριστικών έχει επίσης βαθύ αντίκτυπο στην απόδοση των MLP (Murtagh & Contreras, 2012).

3.2.4. Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support-vector machine - SVM) υπάρχουν εδώ και πολύ καιρό, αλλά η έρευνα για αυτές απέκτησε ιδιαίτερη ώθηση από τότε που ο Vapnik αξιολόγησε αυτές τις μεθόδους στο βιβλίο του για τη στατιστική θεωρία μάθησης (Fahad, Alshatri, Tari, Alamri, Khalil, Zomaya & Bouras, 2014).

Το SVM ανήκει στην κατηγορία των γραμμικών ταξινομητών. Σε πιο ψηλές διαστάσεις, το SVM προσπαθεί να διαιρέσει τον χώρο χαρακτηριστικών χρησιμοποιώντας υπερεπίπεδα απόφασης. Όμως το SVM επιλέγει το επίπεδο με τη μέγιστη απόσταση από τα διανύσματα υποστήριξης ενώ ενδέχεται να υπάρχουν πολλά επίπεδα που διαιρούν τον χώρο χαρακτηριστικών. Τα παραδείγματα με τη μικρότερη απόσταση από το υπερεπίπεδο απόφασης ονομάζονται διανύσματα υποστήριξης (Narmadha, alias Balamurugan, Sundar & Priya, 2016).



Εικόνα 3 - SVM with linear decision boundary

Η συνάρτηση γραμμικής διάκρισης που χρησιμοποιείται σε αυτήν την περίπτωση μπορεί να εκφραστεί ως:

g(x) = wTx + w0

Όπου το w0 υποδηλώνει μια προκατάληψη και το διάνυσμα w, που ονομάζεται διάνυσμα βάρους, είναι η απόσταση των αντίστοιχων υπερεπιπέδων που διέρχονται από διανύσματα στήριξης από την αρχή. Το υπερεπίπεδο h2 είναι στην πραγματικότητα αποτέλεσμα δύο υπερεπίπεδων, που ορίζονται από τα αντίστοιχα διανύσματα υποστήριξης δύο τάξεων (Fahad, Alshatri, Tari, Alamri, Khalil, Zomaya & Bouras, 2014).

Έστω τα h21 και h22 που αντιπροσωπεύουν αυτά τα υπερεπίπεδα έτσι ώστε:

- ✓ h21 : wTx + b = 1 όταν η ετικέτα είναι +1

- ✓ h22 : wTx + b = −1 όταν η ετικέτα είναι -1

Παρόλα αυτά ο χώρος χαρακτηριστικών μπορεί να μην είναι καθόλου γραμμικά διαχωρισμένος. Σε αυτές τις περιπτώσεις το SVM χρησιμοποιεί το κόλπο του πυρήνα για να επιτύχει γραμμικά διαχωρίσιμο χώρο πυρήνα. Μετά τον

μετασχηματισμό, χρησιμοποιείται ο κανονικός αλγόριθμος SVM για ταξινόμηση. Αυτό κάνει τη γραμμική διάκριση που χρησιμοποιείται της φόρμας,

$y = w^T \phi(x) + w_0$

όπου ϕ υποδηλώνει τη συνάρτηση που χρησιμοποιείται για τη μετατροπή του μη γραμμικού χώρου χαρακτηριστικών σε γραμμικό

4. Αλγόριθμοι Συσταδοποίησης

4.1. Βασικά Χαρακτηριστικά

Σύμφωνα με τους Zubin και Joseph, η ανάλυσης ή ομαδοποίηση συστάδων έχει ως στόχο την ομαδοποίηση μιας συλλογής αντικειμένων με τέτοιο τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (*σύμπλεγμα*) να μοιάζουν περισσότερο μεταξύ τους παρά αντικείμενα σε άλλες ομάδες (*συστάδες*).

Οι δημοφιλείς έννοιες συμπλέγματος περιλαμβάνουν ομάδες με μικρές αποστάσεις μεταξύ των μελών της συστάδας, μεγάλες περιοχές χώρου δεδομένων, διαστήματα ή μοναδικές στατιστικές κατανομές. Άρα η ομαδοποίηση μπορεί να θεωρηθεί πρόβλημα βελτιστοποίησης πολλαπλών στόχων (Benabdellah, Benghabrit & Bouhaddou, 2019).

Η ανάλυση συμπλέγματος ως ενέργεια δεν είναι μια αυτόματη εργασία, αλλά μια επαναληπτική διαδικασία ανακάλυψης γνώσης ή πολλαπλών στόχων διαδραστική βελτιστοποίηση που περιλαμβάνει δοκιμές και αποτυχίες. Μερικές φορές, τα δεδομένα προεπεξεργασίας και οι παράμετροι μοντελοποίησης πρέπει να αλλάξουν έως ότου η έξοδος επιτύχει τις επιθυμητές ιδιότητες.

4.2. Λειτουργία, Πλεονεκτήματα και Μειονεκτήματα

Αυτή η ενότητα παρέχει μια σύντομη επισκόπηση ορισμένων από τους πιο συχνά χρησιμοποιούμενους αλγόριθμους clustering με πλεονεκτήματα και μειονεκτήματα για αυτούς.

4.2.1. DBSCAN

Το DBSCAN σημαίνει χωρική ομαδοποίηση εφαρμογών με θόρυβο με βάση την πυκνότητα και είναι ένας προτεινόμενος αλγόριθμος ομαδοποίησης δεδομένων από τους Hans-Peter Kriegel, Martin Ester, Jorg Sander και Xiaowei Xu το 1996.

Πρόκειται για έναν μη παραμετρικός αλγόριθμος ομαδοποίησης που βασίζεται στην πυκνότητα: δεδομένου ενός συνόλου σημείων σε κάποιο χώρο, συγκεντρώνει σημεία που είναι σφιχτά συσκευασμένα μεταξύ τους, χαρακτηρίζοντάς τα ως ακραία σημεία που βρίσκονται μόνα τους σε περιοχές χαμηλής πυκνότητας (Benabdellah, Benghabrit & Bouhaddou, 2019).

Ο αλγόριθμος που εκτελείται για το DBSCAN έχει έναν στόχο και αυτός στοχεύει να ορίσει πυκνές περιοχές που μπορούν να υπολογιστούν από τον αριθμό των αντικειμένων κοντά σε ένα δεδομένο σημείο. Το DBSCAN απαιτεί δύο σημαντικές παραμέτρους: epsilon ("eps") και ελάχιστους πόντους ("MinPts"). Η παράμετρος MinPts είναι ο ελάχιστος αριθμός γειτόνων στην ακτίνα "eps". Η παράμετρος eps ορίζει την ακτίνα γειτονιάς γύρω από ένα σημείο x.

Κάθε σημείο x στο σύνολο δεδομένων επισημαίνεται ως κεντρικό σημείο, με πλήθος γειτόνων μεγαλύτερο ή ίσο με MinPts. Το x είναι συνοριακό σημείο αν ο αριθμός των γειτόνων του είναι μικρότερος από MinPts, αλλά ανήκει σε κάποια γειτονιά έψιλον του πυρήνα του z (Benabdellah, Benghabrit & Bouhaddou, 2019). Το παρακάτω σχήμα δείχνει τους διαφορετικούς τύπους σημείων χρησιμοποιώντας MinPts= 4 (πυρήνας, οριακά και ακραία σημεία). Εδώ το x είναι ένα σημείο πυρήνα επειδή (γείτονες έψιλον(x) = 4), το y είναι οριακό σημείο επειδή (γείτονες έψιλον(υ) < MinPts), επομένως ανήκει στο σημείο πυρήνα x (έψιλον) γειτονιά. Το Z είναι επιτέλους ένα επίπεδο θορύβου.

Ο αλγόριθμος της ομαδοποίησης με βάση την πυκνότητα λειτουργεί όπως παρακάτω, όπου αυτό ξεκινά με ένα αυθαίρετο σημείο εκκίνησης που δεν έχει επισκεφτεί. Το επιχείρημα είναι αλλιώς γνωστό ως θόρυβος. Αυτό το σημείο μπορεί να βρεθεί σε ένα επαρκούς μεγέθους έψιλον-περιβάλλον ενός διαφορετικού σημείου και να γίνει μέρος ενός συμπλέγματος.

Εάν ένα σημείο βρεθεί ότι είναι ένα πυκνό μέρος ενός συμπλέγματος, είναι επίσης μέρος αυτού του συμπλέγματος στη γειτονιά του έψιλον (Boley, Gini, Gross, Han, Hastings, Karypis & Moore, 1999).

Επομένως, περιλαμβάνονται όλα τα σημεία που βρίσκονται εντός της γειτονιάς, όπως και η δική τους γειτονιά όταν είναι πυκνά. Αυτή η διαδικασία συνεχίζεται μέχρι να καθοριστεί πλήρως το σύμπλεγμα που σχετίζεται με την πυκνότητα. Στη συνέχεια, ένα νέο μη επισκέψιμο σημείο ανακτάται και αναλύεται, με αποτέλεσμα να ανιχνεύεται ένα περαιτέρω σύμπλεγμα ή θόρυβος (Boley, Gini, Gross, Han, Hastings, Karypis & Moore, 1999).

```
DBSCAN(D, eps, MinPts) {
   C = 0
   for each point P in dataset D {
      if P is visited
         continue next point
      mark P as visited
      NeighborPts = regionQuery(P, eps)
      if sizeof(NeighborPts) < MinPts
         mark P as NOISE
      else {
         C = next cluster
         expandCluster(P, NeighborPts, C, eps, MinPts)
      }
   }
}

expandCluster(P, NeighborPts, C, eps, MinPts) {
   add P to cluster C
   for each point P' in NeighborPts {
      if P' is not visited {
         mark P' as visited
         NeighborPts' = regionQuery(P', eps)
         if sizeof(NeighborPts') >= MinPts
            NeighborPts = NeighborPts joined with NeighborPts'
      }
      if P' is not yet member of any cluster
         add P' to cluster C
   }
}

regionQuery(P, eps)
   return all points within P's eps-neighborhood (including P)
```

Εικόνα 4 - The DBSCAN pseudo-algorithm

4.2.2. Μονή σύνδεση

Η ομαδοποίηση μιας σύνδεσης είναι μία από τις πολλές μεθόδους ιεραρχικής ομαδοποίησης στη στατιστική. Βασίζεται στην ομαδοποίηση συστάδων με τρόπο από κάτω προς τα πάνω (συσσωρευτική ομαδοποίηση), συνδυάζοντας δύο συστάδες σε κάθε βήμα που περιέχουν το πλησιέστερο ζεύγος στοιχείων που δεν αποτελούν ακόμη μέρος του ίδιου συμπλέγματος το ένα με το άλλο (Boley, Gini, Gross, Han, Hastings, Karypis & Moore, 1999).

Ένα αρνητικό στοιχείο της προσέγγισης είναι ότι παράγει μεγάλα, λεπτά συμπλέγματα όπου τα γειτονικά στοιχεία του ίδιου συμπλέγματος έχουν μικρές αποστάσεις, αλλά τα στοιχεία στα αντίθετα άκρα ίσως είναι πολύ πιο μακριά το ένα από το άλλο.

Αυτό μπορεί να οδηγήσει σε δυσκολίες στον ορισμό της κλάσης που θα μπορούσε να υποδιαιρέσει τα δεδομένα. Αυτός ο συμμετέχων βρίσκεται μέσα σε ένα δικό του σύμπλεγμα στην αρχή της διαδικασίας συσσωρευτικής ομαδοποίησης (Zu Eissen & Stein, 2002).

Τα συμπλέγματα ομαδοποιούνται αργότερα σε μεγαλύτερα συμπλέγματα διαδοχικά έως ότου όλα τα συστατικά καταλήξουν στο ίδιο σύμπλεγμα. Οι δύο ομάδες που διαιρούνται με τη μικρότερη απόσταση συνδυάζονται σε κάθε στάδιο. Ο ορισμός της «μικρότερης απόστασης» είναι αυτός που διακρίνει τις διαφορετικές μεθόδους συσσωμάτωσης (Abbasi & Younis, 2007).

Στην ομαδοποίηση ενός συνδέσμου, η απόσταση μεταξύ δύο συστάδων καθορίζεται από ένα μόνο ζεύγος στοιχείων, δηλαδή από εκείνα τα δύο στοιχεία που είναι πιο κοντά το ένα στο άλλο (ένα σε κάθε συστάδα).

Η πιο γρήγορη από αυτές τις συνδέσεις που παραμένουν σε οποιοδήποτε σημείο προκαλεί τα δύο συμπλέγματα των οποίων τα στοιχεία εμπλέκονται στη σύντηξη μεταξύ τους. Η διαδικασία είναι επίσης γνωστή ως η πλησιέστερη ομαδοποίηση γειτόνων (Jiang, Pang, Wu & Kuang, 2012). Το γινόμενο της ομαδοποίησης μπορεί να απεικονιστεί ως δενδρόγραμμα που δείχνει τη σειρά σύντηξης συστάδων και την απόσταση στην οποία συνέβη κάθε σύντηξη. Μαθηματικά, η συνάρτηση σύνδεσης – η απόσταση D(X,Y) μεταξύ των συστάδων X και Y – περιγράφεται από την έκφραση:

(X, Y) = ελάχ. $x \in X, y \in Y$ (x, y)

όπου X και Y είναι οποιαδήποτε δύο σύνολα στοιχείων που θεωρούνται συμπλέγματα και το d(x,y) υποδηλώνει την απόσταση μεταξύ των στοιχείων x και y.

4.2.3. K-means

Η ομαδοποίηση K-means, όπως ορίστηκε από τον James MacQueen το 1967, είναι μια μέθοδος κβαντοποίησης διανυσμάτων που είναι κοινή για την ανάλυση συστάδων στην εξόρυξη δεδομένων, αρχικά από την επεξεργασία σήματος. Ο στόχος της ομαδοποίησης των k-means είναι να διαιρεθούν n παρατηρήσεις σε k συστάδες. Σε αυτό, κάθε παρατήρηση ανήκει στο σύμπλεγμα με τον πλησιέστερο μέσο όρο, που χρησιμεύει ως πρωτότυπο συμπλέγματος (Zu Eissen & Stein, 2002).

Το K-Means ελαχιστοποιεί τις αποστάσεις μέσα στο σύμπλεγμα (τετράγωνες Ευκλείδειες αποστάσεις), αλλά όχι τις κανονικές Ευκλείδειες αποστάσεις, το οποίο θα ήταν το πιο δύσκολο πρόβλημα για το πρόβλημα του Weber: Ο μέσος όρος βελτιστοποιεί τα τετράγωνα σφάλματα, ενώ οι Ευκλείδειες αποστάσεις ελαχιστοποιούνται μόνο από τη γεωμετρική διάμεσο. Για παράδειγμα, καλύτερες ευκλείδειες λύσεις μπορούν να βρεθούν χρησιμοποιώντας k-διάμεσους και k-medoids (Jiang, Pang, Wu & Kuang, 2012).

Το πρόβλημα είναι υπολογιστικά πολύπλοκο (NP-hard), ωστόσο, αποτελεσματικοί ευρετικοί αλγόριθμοι συγκλίνουν εύκολα σε ένα τοπικό βέλτιστο. Συνήθως αυτά είναι κοντά στον αλγόριθμο προσδοκίας-μεγιστοποίησης για μείγματα κατανομής Gauss μέσω μιας επαναληπτικής προσέγγισης βελτιστοποίησης που χρησιμοποιείται τόσο από τη μοντελοποίηση k-means όσο και από τη μοντελοποίηση μειγμάτων Gauss (Narmadha, alias Balamurugan, Sundar & Priya, 2016).

Ο αλγόριθμος έχει μια χαλαρή σχέση με τον ταξινομητή k-πλησιέστερου γείτονα, μια κοινή τεχνική μηχανικής μάθησης ταξινόμησης που συχνά συγχέεται με το k-means με το όνομα. Η εφαρμογή του ταξινομητή 1-πλησιέστερου γείτονα στα κέντρα συμπλέγματος που λαμβάνονται από k-means ταξινομεί τα νέα δεδομένα σε υπάρχοντα cluster. Αυτός είναι γνωστός ως αλγόριθμος Rocchio ή ο πλησιέστερος κεντροειδής ταξινομητής (Abbasi & Younis, 2007).

Ο αλγόριθμος K-means ξεκινά με την πρώτη ομάδα τυχαία στην εξόρυξη δεδομένων επιλεγμένων κεντροειδών, που λειτουργούν ως σημεία εκκίνησης για κάθε σύμπλεγμα και στη συνέχεια εκτελούν επαναληπτικούς υπολογισμούς για τη βελτιστοποίηση των θέσεων του κέντρου. Αυτό αποτρέπει το σχηματισμό και τη βελτιστοποίηση συμπλεγμάτων. Εάν είτε τα κεντροειδή σταθεροποιηθούν, οι τιμές τους παραμένουν αμετάβλητες επειδή η ομαδοποίηση ήταν αποτελεσματική ή επετεύχθη ο αριθμός των επαναλήψεων που καθορίστηκαν (Jiang, Pang, Wu & Kuang, 2012).

Input:
D= {t1, t2, ... Tn } // Set of elements
K // Number of desired clusters
Output:
K // Set of clusters
K-Means algorithm:
Assign initial values for *m1, m2,.... mk*
repeat
assign each item tj to the clusters which has the closest mean;
calculate new mean for each cluster;
until convergence criteria is met;

Εικόνα 5 - The K-Means pseudo-algorithm

4.2.4. Βελτιστοποίηση σμήνους σωματιδίων

Στην υπολογιστική επιστήμη, η βελτιστοποίηση σμήνους σωματιδίων (PSO) αποτελεί μια προσέγγιση που βελτιστοποιεί υπολογιστικά ένα πρόβλημα επιχειρώντας να αναπτύξει επαναληπτικά μια υποψήφια λύση σε σχέση με ένα ορισμένο μέτρο ποιότητας (Jiang, Pang, Wu & Kuang, 2012).

Επιλύει ένα πρόβλημα έχοντας έναν πληθυσμό υποψήφιων λύσεων, εδώ μεταγλωττισμένα σωματίδια, και μετακινώντας αυτά τα σωματίδια πάνω από τη θέση και την ταχύτητα του σωματιδίου στο χώρο αναζήτησης με βάση απλούς μαθηματικούς τύπους. Η κίνηση κάθε σωματιδίου καθορίζεται από την τοπικά πιο γνωστή θέση του, αλλά κατευθύνεται επίσης προς τις πιο γνωστές θέσεις χώρου αναζήτησης, οι οποίες τροποποιούνται καθώς άλλα σωματίδια βρίσκουν καλύτερες θέσεις. Αυτό θα πρέπει να ωθήσει το σμήνος προς τις καλύτερες λύσεις (Rodriguez, Comin, Casanova, Bruno, Amancio, Costa & Rodrigues, 2019).

Το PSO προοριζόταν αρχικά να μοντελοποιήσει την κοινωνική συμπεριφορά ως μια στυλιζαρισμένη απεικόνιση της δραστηριότητας του οργανισμού σε ένα κοπάδι πουλιών ή ένα κοπάδι ψαριών. Ο αλγόριθμος απλοποιήθηκε και παρατηρήθηκε βελτιστοποίηση. Η Poli πραγματοποιεί λεπτομερή μελέτη των εφαρμογών PSO (Jiang, Pang, Wu & Kuang, 2012).

Το PSO είναι μεταευρετικό, καθώς κάνει ελάχιστες ή καθόλου υποθέσεις σχετικά με τη βελτιστοποίηση του προβλήματος και μπορεί να αναζητήσει τεράστιους χώρους υποψήφιων λύσεων (Narmadha, alias Balamurugan, Sundar & Priya, 2016).

Μια απλή έκδοση του αλγορίθμου PSO λειτουργεί έχοντας έναν πληθυσμό υποψήφιων λύσεων (που ονομάζεται σμήνος) (που ονομάζεται σωματίδια). Σύμφωνα με μερικούς απλούς τύπους, αυτά τα σωματίδια μετακινούνται στον χώρο αναζήτησης. Η κίνηση των σωματιδίων καθοδηγείται από τη δική τους πιο γνωστή θέση στον χώρο αναζήτησης, καθώς και από την πιο γνωστή θέση ολόκληρου του σμήνους. Όταν βρεθούν βελτιωμένες θέσεις, αυτές θα έρθουν στη συνέχεια να καθοδηγήσουν τις κινήσεις του σμήνους (Rodriguez, Comin, Casanova, Bruno, Amancio, Costa & Rodrigues, 2019).

Η διαδικασία επαναλαμβάνεται και αναμένεται ότι θα μπορέσει τελικά να βρεθεί μια ικανοποιητική λύση αλλά όχι εγγυημένη. πάνω από τη θέση και την ταχύτητα του σωματιδίου στο χώρο αναζήτησης με βάση απλούς μαθηματικούς τύπους. Η κίνηση κάθε σωματιδίου καθορίζεται από την τοπικά πιο γνωστή θέση του, αλλά κατευθύνεται επίσης προς τις πιο γνωστές θέσεις χώρου αναζήτησης, οι οποίες τροποποιούνται καθώς άλλα σωματίδια βρίσκουν καλύτερες θέσεις. Αυτό θα πρέπει να ωθήσει το σμήνος προς τις καλύτερες λύσεις (Narmadha, alias Balamurugan, Sundar & Priya, 2016).

```
Initialize population
for t = 1: maximum generation
    Initialize global and local best particles (p_i and p_g)
    for i = 1: population size
        for d = 1: dimension
            v_{i,d}(t+1)=w(t)v_{i,d}(t)+c_1 r_1(p_i - x_{i,d}(t))+c_2 r_2(p_g - x_{i,d}(t));
            if v_{i,d}(t+1)>v_max then v_{i,d}(t+1)=v_max;
               else if v_{i,d}(t+1)<v_min then v_{i,d}(t+1) = v_min;
            end
            x_{i,d}(t+1)=x_{i,d}(t)+v_{i,d}(t+1);
            if x_{i,d}(t+1)>x_max then x_{i,d}(t+1)=x_max;
               else if x_{i,d}(t+1)<x_min then x_{i,d}(t+1) = x_min;
            end
        end
        if f(x_{i,d}(t))<f(p_i(t)) then p_i(t)=x_{i,d}(t);
        end
    end
    f(p_g(t))<min_i(f(p_i(t)));
    w(t)= (t_max - t)/t_max (w_max - w_min)+w_min;
end
```

Εικόνα 6 - The PSO pseudo-algorithm

5. Η Μηχανική Μάθηση στην Ιατρική

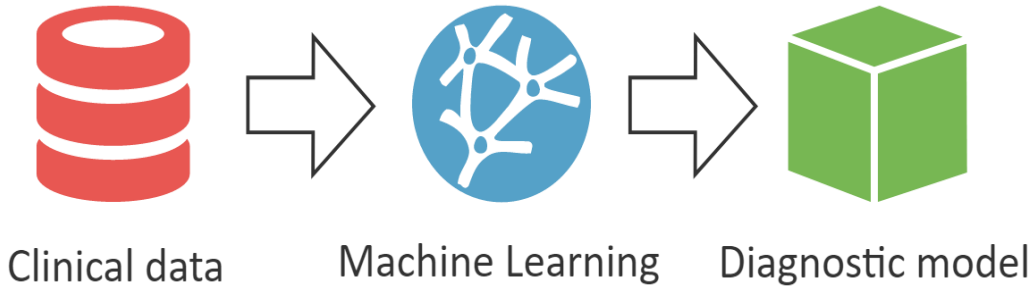
Η μηχανική μάθηση συνιστά τον κλάδο ο οποίος επικεντρώνεται στον τρόπο με τον οποίο οι υπολογιστές μαθαίνουν από δεδομένα. Ειδικότερα, προκύπτει στη διασταύρωση της στατιστικής, η οποία εστιάζει στο να μάθει σχέσεις από δεδομένα, αλλά και στη διασταύρωση της επιστήμης των υπολογιστών. Η σύνδεση των μαθηματικών και της επιστήμης υπολογιστών καθοδηγείται από τις υπολογιστικές προκλήσεις για δημιουργία στατιστικών μοντέλων από πολύ μεγάλα σύνολα δεδομένων (Rajkomar, Dean & Kohane, 2019).

Οι υπολογιστές χρησιμοποιούνται για την εκτέλεση ενός μεγάλου φάσματος σύνθετων εργασιών με γνώμονα την αύξηση της υπολογιστικής ισχύος, της αποθήκευσης, της μνήμης και της δημιουργίας μεγάλου όγκου δεδομένων. Η μηχανική μάθηση (Machine Learning - ML) αφορά τόσο στον ακαδημαϊκό κλάδο όσο και στη συλλογή τεχνικών που βοηθούν στους υπολογιστές να εκτελούν σύνθετες διεργασίες (Rajula, Verlato, Manchia, Antonucci & Fanos, 2020).

Οι μέθοδοι μάθησης που αναπτύχθηκαν για αυτούς τους κλάδους προσφέρουν τεράστιες δυνατότητες για την ενίσχυση της ιατρικής έρευνας και της κλινικής περίθαλψης, ειδικά καθώς οι πάροχοι χρησιμοποιούν όλο και περισσότερο ηλεκτρονικά αρχεία υγείας. Οι τομείς που ωφελούνται από την εφαρμογή των τεχνικών ML στον ιατρικό τομέα είναι η διάγνωση και η πρόβλεψη των αποτελεσμάτων. Αυτό περιλαμβάνει τη δυνατότητα εντοπισμού υψηλού κινδύνου για ιατρικά επείγοντα περιστατικά όπως υποτροπή ή μετάβαση σε άλλη κατάσταση ασθένειας (Cabitza, Rasoini & Gensini, 2017).

Οι αλγόριθμοι ML αξιοποιούνται επίσης για την ταξινόμηση του καρκίνου του δέρματος χρησιμοποιώντας εικόνες με συγκρίσιμη ακρίβεια με έναν εκπαιδευμένο δερματολόγο και για την πρόβλεψη της εξέλιξης από τον προδιαβήτη στον διαβήτη τύπου 2 χρησιμοποιώντας τακτικά συλλεγόμενα δεδομένα ηλεκτρονικών αρχείων υγείας (Vayena, Blasimme & Cohen, 2018).

Η μηχανική μάθηση χρησιμοποιείται όλο και περισσότερο σε συνδυασμό με την Επεξεργασία Φυσικής Γλώσσας (NLP) για να κατανοήσει τα μη δομημένα δεδομένα κειμένου. Συνδυάζοντας την ML με τις τεχνικές NLP, οι ερευνητές μπόρεσαν να αντλήσουν νέες γνώσεις από σχόλια από αναφορές κλινικών περιστατικών, δραστηριότητα στα μέσα κοινωνικής δικτύωσης, σχόλια απόδοσης γιατρών και αναφορές ασθενών μετά από επιτυχημένες θεραπείες καρκίνου. Οι πληροφορίες που δημιουργούνται αυτόματα από μη δομημένα δεδομένα θα μπορούσαν να είναι εξαιρετικά χρήσιμες όχι μόνο για την απόκτηση εικόνας σχετικά με την ποιότητα, την ασφάλεια και την απόδοση, αλλά και για την έγκαιρη διάγνωση (Rajula, Verlato, Manchia, Antonucci & Fanos, 2020).



Εικόνα 7 - Μοντέλο διάγνωσης με χρήση machine learning

Η μηχανική μάθηση διαδραματίζει επιπλέον κρίσιμο ρόλο στην ανάπτυξη συστημάτων μάθησης υγειονομικής περίθαλψης. Τα συστήματα μάθησης υγειονομικής περίθαλψης περιλαμβάνουν περιβάλλοντα που ευθυγραμμίζουν την πληροφορική, την επιστήμη, τα κίνητρα και την προσπάθεια για συνεχή βελτίωση και καινοτομία (Rajula, Verlato, Manchia, Antonucci & Fanos, 2020).

Τα συστήματα αυτά θα μπορούσαν να εκτελεστούν σε οποιαδήποτε κλίμακα, από πρακτικές μικρών ομάδων έως μεγάλους εθνικούς παρόχους, θα συνδυάζουν διάφορες πηγές δεδομένων με σύνθετους αλγόριθμους ML. Έτσι θα προκύψουν γνώσεις που βασίζονται σε δεδομένα για τη βελτιστοποίηση της δημόσιας υγείας, της βιοϊατρικής έρευνας και της βελτίωσης ποιότητας υγειονομικής περίθαλψης (Darcy, Louie & Roberts, 2016).

Οι τεχνικές μηχανικής μάθησης βασίζονται σε αλγόριθμους – σύνολα μαθηματικών διαδικασιών που περιγράφουν τις σχέσεις μεταξύ μεταβλητών. Αν και οι αλγόριθμοι λειτουργούν με διαφορετικούς τρόπους ανάλογα με τον τύπο τους, υπάρχουν αξιοσημείωτα κοινά στοιχεία στον τρόπο με τον οποίο αναπτύσσονται. Αν και οι πολυπλοκότητες των αλγορίθμων ML μπορεί να φαίνονται εσωτερικές, συχνά μοιάζουν περισσότερο από μια λεπτή ομοιότητα με τις συμβατικές στατιστικές αναλύσεις (Darcy, Louie & Roberts, 2016).

Ο στόχος των στατιστικών μεθόδων είναι το συμπέρασμα, δηλαδή να καταλήξουν σε συμπεράσματα σχετικά με τους πληθυσμούς ή να αντλήσουν επιστημονικές γνώσεις από δεδομένα που συλλέγονται από ένα αντιπροσωπευτικό δείγμα αυτού του πληθυσμού. Αν και πολλές στατιστικές τεχνικές, όπως η γραμμική και η λογιστική παλινδρόμηση, είναι ικανές να δημιουργήσουν προβλέψεις για νέα δεδομένα, το κίνητρο της χρήσης τους ως στατιστικής μεθοδολογίας είναι η

εξαγωγή συμπερασμάτων σχετικά με τις σχέσεις μεταξύ των μεταβλητών (Vokinger, Feuerriegel & Kesselheim, 2021).

Αντίθετα, στον τομέα της ML, το κύριο μέλημα είναι μια ακριβής πρόβλεψη, το «τι» παρά το «πώς». Για παράδειγμα, στην αναγνώριση εικόνας, η σχέση μεταξύ των επιμέρους χαρακτηριστικών (pixel) και του αποτελέσματος είναι μικρής σημασίας εάν η πρόβλεψη είναι ακριβής (Rajkomar, Dean & Kohane, 2019).

Ευτυχώς για τον ιατρικό τομέα, πολλές σχέσεις ενδιαφέροντος είναι εύλογα απλές, όπως αυτές μεταξύ του δείκτη μάζας σώματος και του κινδύνου διαβήτη ή της χρήσης καπνού από καρκίνο του πνεύμονα. Εξαιτίας αυτού, η αλληλεπίδρασή τους μπορεί συχνά να εξηγηθεί αρκετά καλά χρησιμοποιώντας σχετικά απλά μοντέλα (Rajula, Verlato, Manchia, Antonucci & Fanos, 2020).

Σε πολλές δημοφιλείς εφαρμογές της ML, όπως η βελτιστοποίηση της πλοήγησης, η μετάφραση εγγράφων και ο εντοπισμός αντικειμένων στα βίντεο, η κατανόηση της σχέσης μεταξύ των χαρακτηριστικών και των αποτελεσμάτων είναι λιγότερο σημαντική. Δεδομένης αυτής της βασικής διαφοράς, μπορεί να είναι χρήσιμο για τους ερευνητές να θεωρήσουν ότι οι αλγόριθμοι υπάρχουν σε ένα συνεχές μεταξύ αυτών των αλγορίθμων που είναι εύκολα ερμηνεύσιμοι (Auditable Algorithms) και εκείνων που δεν είναι (Black Boxes) (Deo, 2015).

Η εποπτευόμενη ML αναφέρεται σε τεχνικές στις οποίες ένα μοντέλο εκπαιδεύεται σε μια σειρά εισροών (ή χαρακτηριστικών) που σχετίζονται με ένα γνωστό αποτέλεσμα. Στην ιατρική, αυτό μπορεί να αντιπροσωπεύει την εκπαίδευση ενός μοντέλου για να συσχετίσει τα χαρακτηριστικά ενός ατόμου (π.χ. ύψος, βάρος, κατάσταση καπνίσματος) με ένα συγκεκριμένο αποτέλεσμα (π.χ. εμφάνιση διαβήτη εντός πέντε ετών). Μόλις ο αλγόριθμος εκπαιδευτεί επιτυχώς, θα είναι σε θέση να κάνει προβλέψεις αποτελεσμάτων όταν εφαρμόζεται σε νέα δεδομένα. Οι προβλέψεις που γίνονται από μοντέλα που έχουν εκπαιδευτεί χρησιμοποιώντας εποπτευόμενη μάθηση μπορεί να είναι είτε διακριτές (π.χ. θετικές ή αρνητικές, καλοήθεις ή κακοήθεις) είτε συνεχείς (π.χ. βαθμολογία από 0 έως 100) (Sidey-Gibbons & Sidey-Gibbons, 2019).

Ένα μοντέλο που παράγει διακριτές κατηγορίες (μερικές φορές αναφέρονται ως κλάσεις) αναφέρεται ως αλγόριθμος ταξινόμησης. Παραδείγματα αλγορίθμων ταξινόμησης περιλαμβάνουν αυτούς που προβλέπουν εάν ένας όγκος είναι καλοήθης ή κακοήθης ή για να διαπιστωθεί εάν τα σχόλια που γράφτηκαν από έναν ασθενή μεταφέρουν θετικό ή αρνητικό συναίσθημα (Vayena, Blasimme & Cohen, 2018).

Ένα μοντέλο που επιστρέφει μια πρόβλεψη μιας συνεχούς τιμής είναι γνωστό ως αλγόριθμος παλινδρόμησης. Η χρήση του όρου παλινδρόμηση στο ML διαφέρει από τη χρήση του στη στατιστική, όπου η παλινδρόμηση χρησιμοποιείται συχνά για να αναφέρεται τόσο σε δυαδικά αποτελέσματα (δηλαδή, λογιστική παλινδρόμηση) όσο και σε συνεχή αποτελέσματα (δηλαδή, γραμμική παλινδρόμηση). Στην ML, ένας αλγόριθμος που αναφέρεται ως αλγόριθμος παλινδρόμησης μπορεί να

χρησιμοποιηθεί για να προβλέψει το προσδόκιμο ζωής ενός ατόμου ή την ανεκτή δόση χημειοθεραπείας (Sidey-Gibbons & Sidey-Gibbons, 2019).

Μόλις ένα σύνολο δεδομένων οργανωθεί σε χαρακτηριστικά και αποτελέσματα, ένας αλγόριθμος ML μπορεί να εφαρμοστεί σε αυτό. Ο αλγόριθμος βελτιώνεται επαναληπτικά για να μειωθεί το σφάλμα πρόβλεψης χρησιμοποιώντας μια τεχνική βελτιστοποίησης.

Μόλις ολοκληρωθεί η εκπαίδευση, ο αλγόριθμος εφαρμόζεται στα χαρακτηριστικά του συνόλου δεδομένων δοκιμής χωρίς τα σχετικά αποτελέσματα (Cabitza, Rasoini & Gensini, 2017). Οι προβλέψεις από τον αλγόριθμο συγκρίνονται με τα γνωστά αποτελέσματα του συνόλου δεδομένων δοκιμής για να καθοριστεί η απόδοση του μοντέλου. Αυτό είναι απαραίτητο για να αυξηθεί η πιθανότητα ο αλγόριθμος να γενικευτεί καλά σε νέα δεδομένα.

Οι μη εποπτευόμενες τεχνικές είναι επομένως διερευνητικές και χρησιμοποιούνται για την εύρεση απροσδιόριστων προτύπων ή συμπλεγμάτων που εμφανίζονται μέσα σε σύνολα δεδομένων. Οι τεχνικές αυτές αναφέρονται συχνά ως τεχνικές μείωσης των διαστάσεων και περιλαμβάνουν διαδικασίες όπως η ανάλυση κύριου συστατικού, η λανθάνουσα ανάλυση Dirichlet και η t-Distributed Stochastic Neighbor Embedding (t-SNE) (Rajula, Verlato, Manchia, Antonucci & Fanos, 2020).

Με τη συμπίεση των πληροφοριών σε ένα σύνολο δεδομένων σε λιγότερα χαρακτηριστικά ή διαστάσεις, μπορεί να αποφευχθούν ζητήματα όπως η πολλαπλή συγγραμμικότητα ή το υψηλό υπολογιστικό κόστος.

Με παρόμοιο τρόπο με τους αλγόριθμους εποπτευόμενης μάθησης που περιγράφηκαν προηγουμένως, έχουν επίσης πολλές ομοιότητες με τις στατιστικές τεχνικές που θα είναι γνωστές στους ιατρικούς ερευνητές. Οι τεχνικές μάθησης χωρίς επίβλεψη χρησιμοποιούν παρόμοιους αλγόριθμους που χρησιμοποιούνται για ομαδοποίηση και μείωση διαστάσεων στα παραδοσιακά στατιστικά στοιχεία (Cabitza, Rasoini & Gensini, 2017). Όσοι είναι εξοικειωμένοι με την ανάλυση βασικών στοιχείων και την ανάλυση παραγόντων θα είναι ήδη εξοικειωμένοι με πολλές από τις τεχνικές που χρησιμοποιούνται στη μάθηση χωρίς επίβλεψη.

Η εποπτευόμενη μάθηση ξεκινά με στόχο την πρόβλεψη μιας γνωστής παραγωγής ή στόχου. Σε διαγωνισμούς μηχανικής μάθησης, όπου μεμονωμένοι συμμετέχοντες κρίνονται για την απόδοσή τους σε κοινά σύνολα δεδομένων, τα επαναλαμβανόμενα εποπτευόμενα μαθησιακά προβλήματα περιλαμβάνουν την αναγνώριση χειρόγραφου (όπως η αναγνώριση χειρόγραφων ψηφίων), την ταξινόμηση εικόνων αντικειμένων (π.χ. είναι γάτα ή σκύλος;) και ταξινόμηση εγγράφων (π.χ. είναι μια κλινική δοκιμή για καρδιακή ανεπάρκεια ή μια οικονομική έκθεση;) (Vayena, Blasimme & Cohen, 2018).

Συγκεκριμένα, όλες αυτές είναι εργασίες που μπορεί να κάνει καλά ένα εκπαιδευμένο άτομο και έτσι ο υπολογιστής συχνά προσπαθεί να προσεγγίσει την ανθρώπινη απόδοση. Η εποπτευόμενη μάθηση εστιάζει στην ταξινόμηση, η οποία

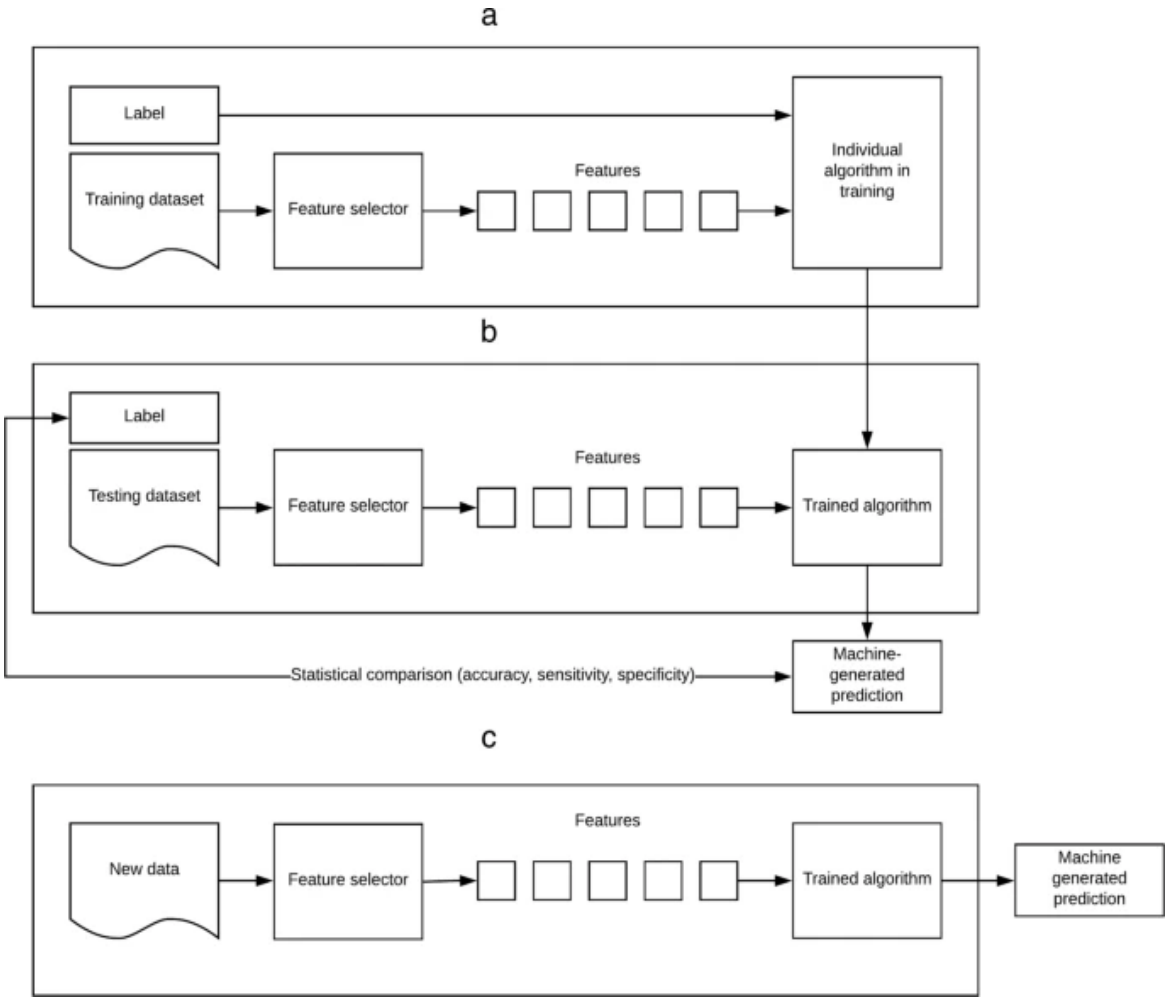
περιλαμβάνει την επιλογή μεταξύ των υποομάδων για την καλύτερη περιγραφή μιας νέας παρουσίας δεδομένων και την πρόβλεψη, η οποία περιλαμβάνει την εκτίμηση μιας άγνωστης παραμέτρου (όπως η θερμοκρασία στο Σαν Φρανσίσκο αύριο το απόγευμα) (Cleophas, Zwinderman & Cleophas-Allers, 2013).

Η εποπτευόμενη μάθηση χρησιμοποιείται για την εκτίμηση του κινδύνου. Το Framingham Risk Score3 για στεφανιαία νόσο (CHD) είναι η πιο συχνά χρησιμοποιούμενη περίπτωση εποπτευόμενης μάθησης στην ιατρική. Τέτοια μοντέλα κινδύνου περιλαμβάνουν την καθοδηγητική αντιθρομβωτική θεραπεία στην κολπική μαρμαρυγή και την εμφύτευση αυτοματοποιημένων εμφυτεύσιμων απινιδωτών στην υπερτροφική μυοκαρδιοπάθεια (Vayena, Blasimme & Cohen, 2018). Κατά τη μοντελοποίηση του κινδύνου, ο υπολογιστής κάνει περισσότερα από την απλή προσέγγιση των δεξιοτήτων του γιατρού, αλλά βρίσκει νέες σχέσεις που δεν είναι άμεσα εμφανείς στα ανθρώπινα όντα.

Απεναντίας, στη μάθηση χωρίς επίβλεψη, δεν υπάρχουν αποτελέσματα για πρόβλεψη. Αντίθετα, προσπαθούμε να βρούμε φυσικά μοτίβα ή ομαδοποιήσεις μέσα στα δεδομένα. Αυτό είναι εγγενώς ένα πιο δύσκολο έργο για να κριθεί και συχνά η αξία τέτοιων ομάδων που μαθαίνονται μέσω της μάθησης χωρίς επίβλεψη αξιολογείται από την απόδοσή του σε επόμενες εποπτευόμενες μαθησιακές εργασίες (δηλαδή είναι χρήσιμα αυτά τα νέα πρότυπα με κάποιο τρόπο) (Cleophas, Zwinderman & Cleophas-Allers, 2013).

Ας σκεφτούμε πώς θα μπορούσε κανείς να εφαρμόσει τη μάθηση χωρίς επίβλεψη στην καρδιακή νόσο προς αυτή την κατεύθυνση, λαμβάνοντας μια ετερογενή κατάσταση όπως η μυοκαρδίτιδα. Μπορεί κανείς να ξεκινήσει με μια μεγάλη ομάδα φαινομενικά παρόμοιων ατόμων με ανεξήγητη οξεία συστολική καρδιακή ανεπάρκεια. Στη συνέχεια μπορεί κανείς να πραγματοποιήσει βιοψίες μυοκαρδίου σε αυτά, και να χαρακτηρίσει την κυτταρική σύνθεση κάθε δείγματος με μια τεχνική όπως η ανοσοχρώση (Cleophas & Zwinderman, 2015).

Για παράδειγμα, μπορεί να γίνει καταμέτρηση από Τ λεμφοκυττάρων, ουδετερόφιλων, μακροφάγων, ηωσινόφιλων, κ.λπ. Στη συνέχεια, ενδέχεται να γίνει μια εξέταση για το εάν υπάρχουν επαναλαμβανόμενα μοτίβα κυτταρικής σύνθεσης, τα οποία προτείνουν μηχανισμούς και θεραπείες καθοδήγησης. Μια παρόμοια προσέγγιση, αν και εστιασμένη στη γονιδιωματική, οδήγησε στον εντοπισμό ενός ηωσινοφιλικού υποτύπου άσθματος, ο οποίος ανταποκρίνεται μοναδικά σε μια νέα θεραπεία που στοχεύει την κυτοκίνη IL-138 που εκκρίνεται από ηωσινόφιλα (Gui & Chan, 2017).



Εικόνα 8 - Επισκόπηση της εποπτευόμενης μάθησης: α. Εκπαίδευση β. Επικύρωση γ. Εφαρμογή αλγορίθμου σε νέα δεδομένα

Με βάση τα παραπάνω παραδείγματα είναι προφανές ότι η μηχανική μάθηση – τόσο υπό επίβλεψη όσο και χωρίς επίβλεψη – μπορεί να εφαρμοστεί σε σύνολα κλινικών δεδομένων με σκοπό την ανάπτυξη ισχυρών μοντέλων κινδύνου και τον επαναπροσδιορισμό των κατηγοριών ασθενών (Van Calster & Wynants, 2019).

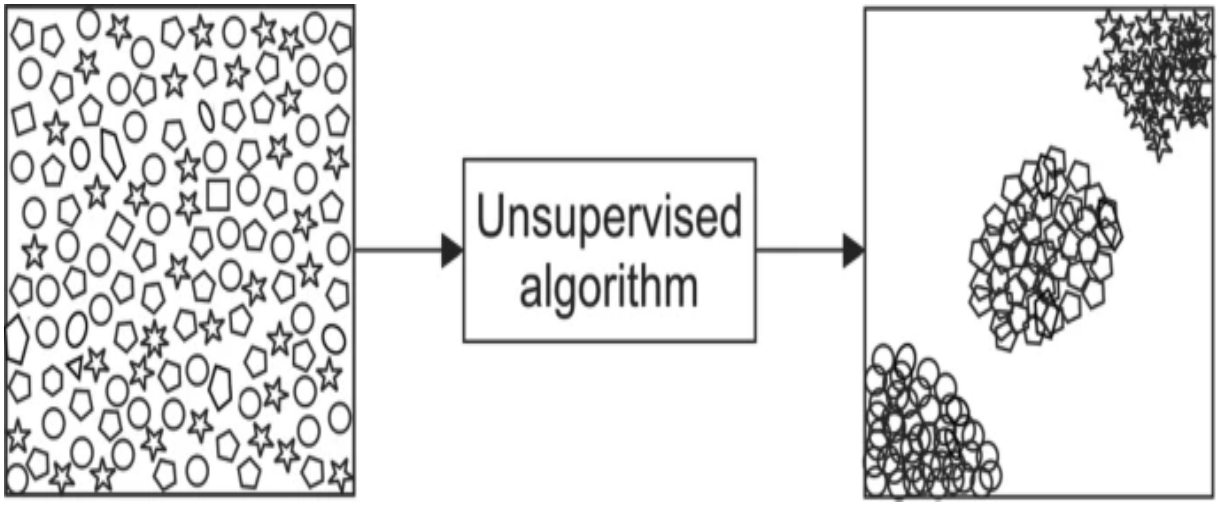
Δεδομένου του περιορισμένου κλινικού αποτυπώματος της μηχανικής μάθησης, υπάρχουν ορισμένα εμπόδια στη μετάφραση. Κάποια από αυτά συνδέονται με ρεαλιστικά ζητήματα που σχετίζονται με τον ιατρικό κλάδο, συμπεριλαμβανομένης της αποζημίωσης και της ευθύνης (Verma, Murray, Greiner, Cohen, Shojania, Ghassemi & Mamdani, 2021).

Για να ενσωματωθεί η μηχανική μάθηση σε τομείς όπου δεν μπορεί να υποσχεθεί τόσο υψηλή ακρίβεια όσο αυτή ενός ανθρώπινου ειδικού, πρέπει να υπάρχουν τρόποι για τους γιατρούς να αλληλεπιδρούν με συστήματα υπολογιστών για να διατηρήσουν την ακρίβεια και ωστόσο να αυξήσουν την απόδοση και να μειώσουν το κόστος (Van Calster & Wynants, 2019). Θα χρειαστεί ένα νέο μοντέλο

αποζημίωσης για μια τέτοια ολοκληρωμένη προσέγγιση ανθρώπου και μηχανής (Vokinger, Feuerriegel & Kesselheim, 2021).

Μια άλλη πρόκληση είναι εάν θα χορηγηθεί κλινική ένδειξη FDA σε ένα φάρμακο για μια υποομάδα ασθενών που έχει οριστεί με τρόπο που δεν σχετίζεται με τον μηχανισμό δράσης αυτού του φαρμάκου. Παρά το γεγονός ότι μπορούμε να στοχεύσουμε έναν συγκεκριμένο αναστολέα κινάσης σε καρκινοπαθείς με ενεργοποιητική μετάλλαξη οδηγού στην ίδια κινάση, δεν είναι ακόμη κατανοητό πώς γίνεται να δικαιολογήσουμε την αντιστοίχιση των κατηγοριών HFpEF με έναν συγκεκριμένο τύπο φαρμάκου (Darcy, Louie & Roberts, 2016).

Εμπειρική απόδειξη δυσανάλογου θεραπευτικού οφέλους σε μια κατηγορία σε σχέση με μια άλλη θα ήταν απαραίτητη, αλλά δεν είναι σίγουρο ότι θα ήταν επαρκής. Αυτή η αδυναμία μπορεί να δικαιολογηθεί με αντιστοίχιση μιας υποομάδας ασθενών με ένα φάρμακο σε βιολογική βάση που θα αντιπροσωπεύει μια εγγενή πρόκληση για την αναταξινόμηση των πιο πολύπλοκων ασθενειών, καθώς αυτές συνήθως δεν μπορούν να οριστούν μόνο από τη γενετική ή έναν προφανή βιοδείκτη που συνδέεται με τον θεραπευτικό μηχανισμό του φαρμάκου (Verma, Murray, Greiner, Cohen, Shojania, Ghassemi & Mamdani, 2021). Οι κλινικές δοκιμές θα μπορούσαν να τροφοδοτηθούν επαρκώς για όλες τις προκαθορισμένες υποομάδες, γεγονός που θα ήταν μια καλή λύση.



Εικόνα 9 - Μια οπτική απεικόνιση μιας τεχνικής μείωσης διαστάσεων χωρίς επίβλεψη

Ορισμένες δυσκολίες στην υιοθέτηση της μηχανικής μάθησης στην ιατρική μπορεί επίσης να σχετίζονται με πραγματικές στατιστικές προκλήσεις στη μάθηση. Προκύπτει από τα έως τώρα δεδομένα ότι θα χρειαστούν νέα πληροφοριακά χαρακτηριστικά για τη δημιουργία βελτιωμένων μοντέλων στην ιατρική, ιδιαίτερα

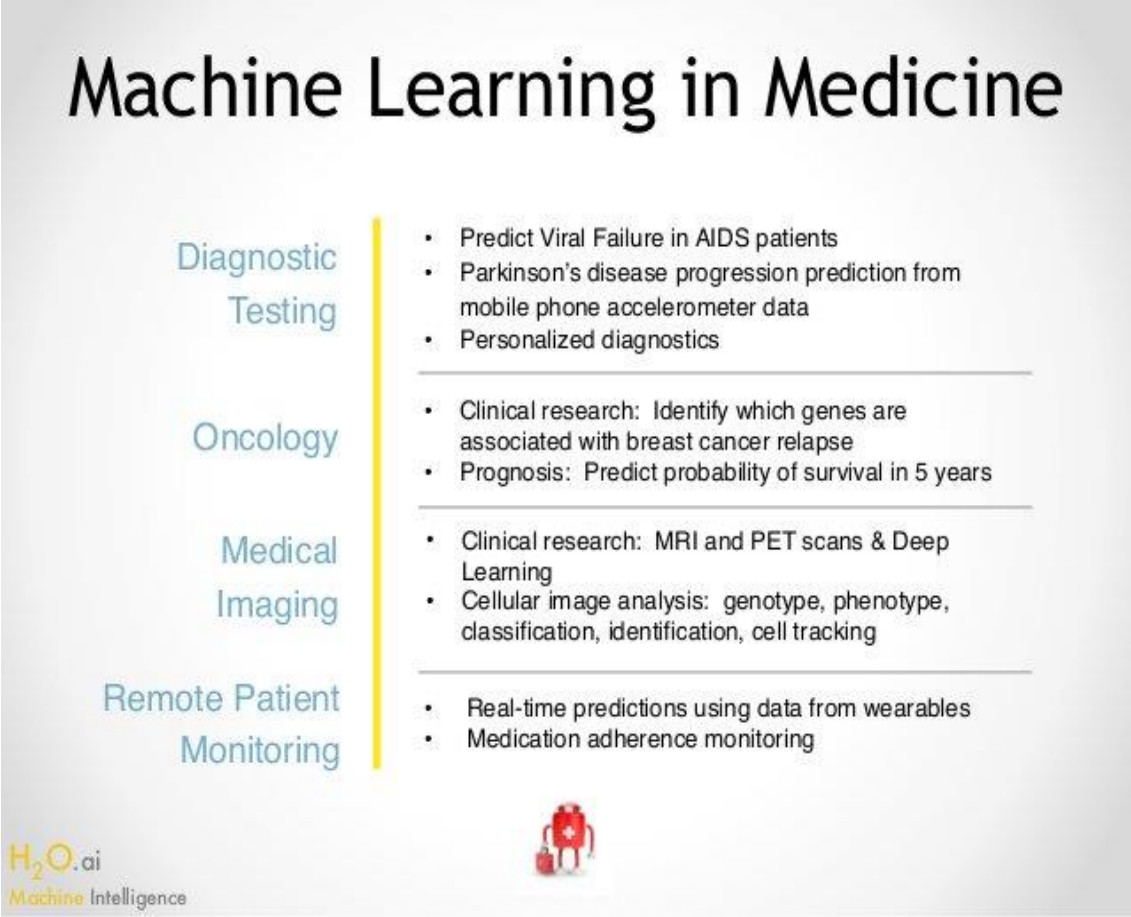
σε καταστάσεις μάθησης όπου ο υπολογιστής δεν προσεγγίζει απλώς την απόδοση του γιατρού (Darcy, Louie & Roberts, 2016).

Για τις καρδιαγγειακές παθήσεις θα είναι δύσκολο να βρεθούν μεγάλες αμερόληπτες πηγές φαινοτυπικών δεδομένων με επαρκή πληροφόρηση για τον χαρακτηρισμό της διαδικασίας της νόσου. Στη μελέτη για τους ασθενείς με HFpEF, χρησιμοποιήσαν ηχοκαρδιογραφικά δεδομένα (Van Calster & Wynants, 2019).

Ένα ακόμη συμπέρασμα που προκύπτει και είναι τεχνικό, σχετίζεται με την αλληλεπίδραση μορφών μάθησης χωρίς επίβλεψη και επίβλεψη. Η βαθιά μάθηση, με στοιβαγμένα επίπεδα αναπαραστάσεων αντικειμένων ολοένα και υψηλότερης τάξης, έχει κατακλύσει τον κόσμο της μηχανικής μάθησης (Vokinger, Feuerriegel & Kesselheim, 2021).

Η βαθιά εκμάθηση χρησιμοποιεί μάθηση χωρίς επίβλεψη για να βρει πρώτα ισχυρά χαρακτηριστικά, τα οποία στη συνέχεια μπορούν να βελτιωθούν και τελικά να χρησιμοποιηθούν ως προγνωστικοί παράγοντες σε ένα τελικό εποπτευόμενο μοντέλο (Verma, Murray, Greiner, Cohen, Shojania, Ghassemi & Mamdani, 2021).

Σε μια αναπαράσταση βαθιάς μάθησης της ανθρώπινης νόσου, τα κατώτερα στρώματα θα μπορούσαν να αντιπροσωπεύουν κλινικές μετρήσεις (όπως δεδομένα ΗΚΓ ή βιοδείκτες πρωτεΐνης), τα ενδιάμεσα στρώματα θα μπορούσαν να αντιπροσωπεύουν ανώμαλα μονοπάτια (που μπορεί ταυτόχρονα να επηρεάσουν πολλούς βιοδείκτες) και τα ανώτερα στρώματα θα μπορούσαν να αντιπροσωπεύουν υποκατηγορίες ασθενειών (που προκύπτουν από τις μεταβλητές συνεισφορές μιας ή περισσότερων ανώμαλων μονοπατιών) (Verma, Murray, Greiner, Cohen, Shojania, Ghassemi & Mamdani, 2021).



Εικόνα 10 - Εφαρμογές στην ιατρική

Τέτοιες υποκατηγορίες θα έκαναν περισσότερα από τη στρωματοποίηση βάσει κινδύνου και στην πραγματικότητα θα αντανακλούσαν τον κυρίαρχο μηχανισμό της νόσου (Harrison & Sidey-Gibbons, 2021).

Αυτό εγείρει ένα ερώτημα σχετικά με την υποκείμενη παθοφυσιολογική βάση της περίπλοκης νόσου σε κάθε δεδομένο άτομο: είναι αραιά κωδικοποιημένη σε ένα περιορισμένο σύνολο ανώμαλων οδών, οι οποίες θα μπορούσαν να ανακτηθούν με μια διαδικασία μάθησης χωρίς επίβλεψη (αν και με τα σωστά χαρακτηριστικά που συλλέγονται και ένα αρκετά μεγάλο δείγμα μέγεθος), ή είναι μια διάχυτη, πολυπαραγοντική διαδικασία με εκατοντάδες μικρούς καθοριστικούς παράγοντες που συνδυάζονται με εξαιρετικά μεταβλητό τρόπο σε διαφορετικά άτομα (Verma, Murray, Greiner, Cohen, Shojania, Ghassemi & Mamdani, 2021). Στην τελευταία περίπτωση, η έννοια της «ιατρικής ακριβείας» είναι απίθανο να είναι πολύ χρήσιμη.

Ωστόσο, στην προηγούμενη κατάσταση, η μη εποπτευόμενη και ίσως η βαθιά μάθηση θα μπορούσε να πραγματοποιήσει πραγματικά τον άπιαστο στόχο της επαναταξινόμησης των ασθενών σύμφωνα με πιο ομοιογενείς υποομάδες, με κοινή παθοφυσιολογία και τη δυνατότητα κοινής ανταπόκρισης στη θεραπεία (Vokinger, Feuerriegel & Kesselheim, 2021).

6. Υλοποιητικό μέρος

6.1. Εισαγωγή – Φιλοσοφία Υλοποίησης

Στα πλαίσια της παρούσας διπλωματικής εργασίας στόχος μας ήταν η εφαρμογή αλγορίθμων classification πάνω σε ιατρικά δεδομένα, και πιο συγκεκριμένα πάνω σε δεδομένα τα οποία σχετίζονται με την ύπαρξη εγκεφαλικού επεισοδίου. Οι αλγόριθμοι οι οποίοι χρησιμοποιήθηκαν είναι οι:

- Random Forests
- Extra Trees
- kNN
- SVM

6.2. Γλώσσα – Χρησιμοποιούμενα Εργαλεία και Βιβλιοθήκες

Το υλοποιητικό μέρος της διπλωματικής εργασίας υλοποιήθηκε σε γλώσσα Python. Το περιβάλλον το οποίο χρησιμοποιήθηκε για την συγγραφή του κώδικα είναι το Anaconda Spyder. Σημαντικές βιβλιοθήκες οι οποίες μας βοήθησαν στην ολοκλήρωση της πειραματικής διαδικασίας είναι οι:

- pandas
- numpy
- sklearn
- seaborn
- matplotlib

6.3. Ανάλυση Dataset

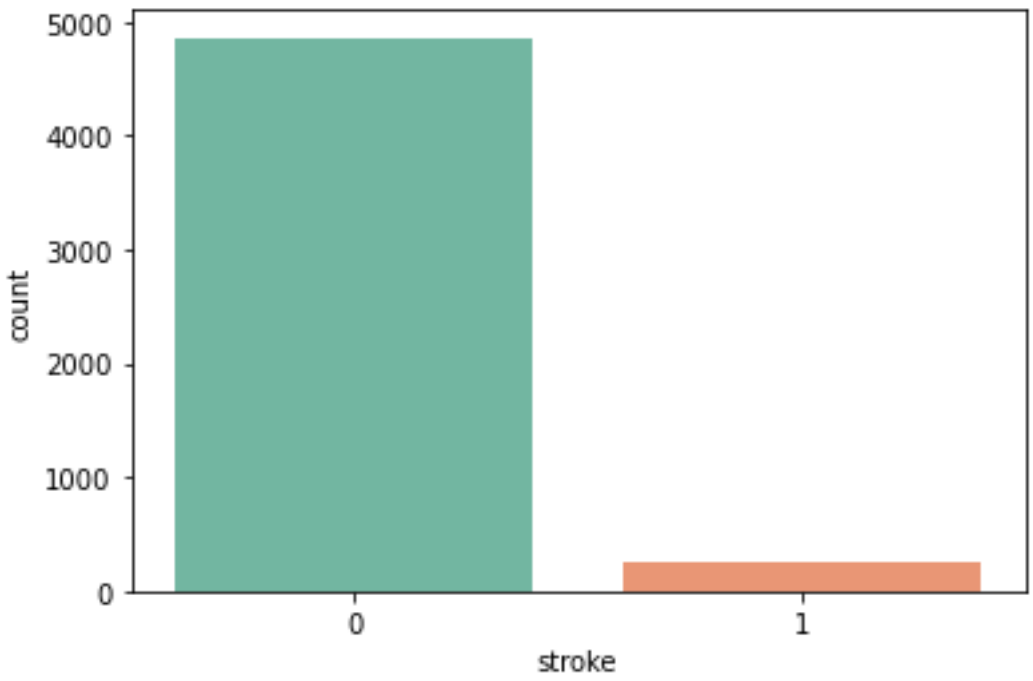
Για την πραγματοποίηση της πειραματικής διαδικασίας, χρησιμοποιήθηκε το αρχείο healthcare-dataset-stroke-data.csv το οποίο περιέχει 5110 εγγραφές με τα παρακάτω πεδία:

- id: το id της εγγραφής
- gender: το φύλο
- hypertension: η ύπαρξη υπέρτασης ή όχι (με τιμές 1 ή 0 αντίστοιχα)
- heart_disease: η ύπαρξη καρδιακής ασθένειας ή όχι (με τιμές 1 ή 0 αντίστοιχα)

- ever_married: που περιγράφει το εάν είναι παντρεμένος ή όχι
- work_type: ο τύπος εργασίας
- residence_type: το περιβάλλον στο οποίο ζει (urban ή rural)
- avg_glucose_level: το μέσο επίπεδο γλυκόζης
- bmi: η τιμή του δείκτη bmi
- smoking_status: το κατά πόσο συχνά καπνίζει
- stroke: η ύπαρξη εγκεφαλικού επεισοδίου ή όχι (με 1 ή 0 τιμές αντίστοιχα)

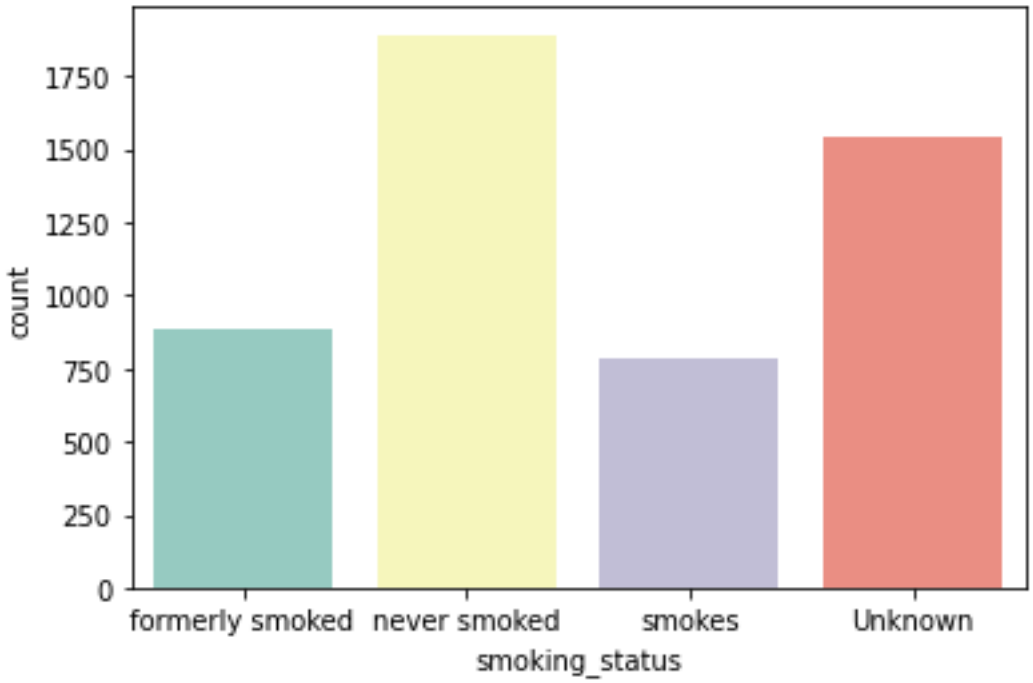
Για την καλύτερη οπτικοποίηση του χρησιμοποιούμενου dataset, παρουσιάζουμε την κατανομή του συνόλου των εγγραφών ως προς διάφορα χαρακτηριστικά του.

Κατανομή του πλήθους των εγγραφών ως προς την ύπαρξη εγκεφαλικού επεισοδίου:

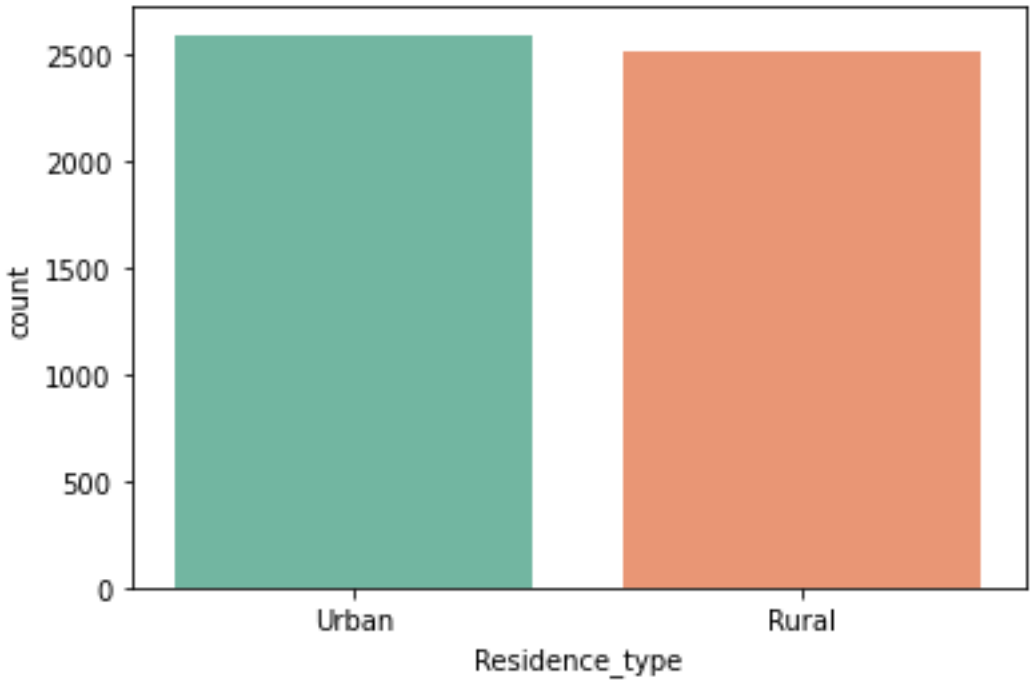


Εικόνα 11 - Κατανομή πλήθους εγγραφών ως προς την ύπαρξη εγκεφαλικού επεισοδίου

Κατανομή του πλήθους των εγγραφών ως προς τις συνήθειες του καπνίσματος:

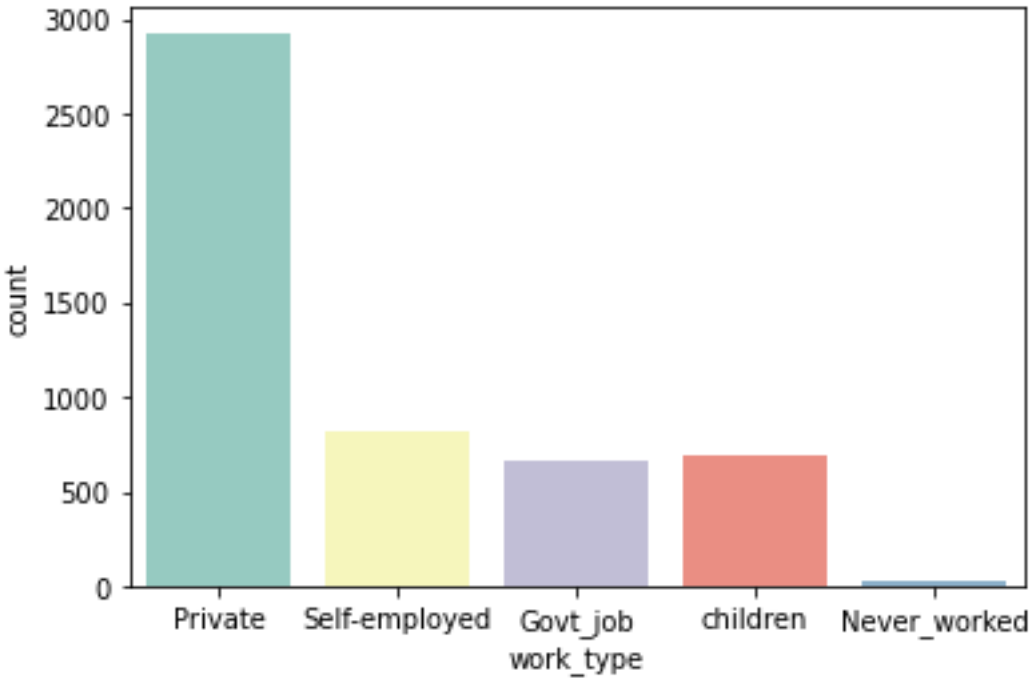


Εικόνα 12 - Κατανομή πλήθους εγγραφών ως προς τις συνήθειες του καπνίσματος



Εικόνα 13 - Κατανομή πλήθους εγγραφών ως προς τον τόπο της κατοικίας

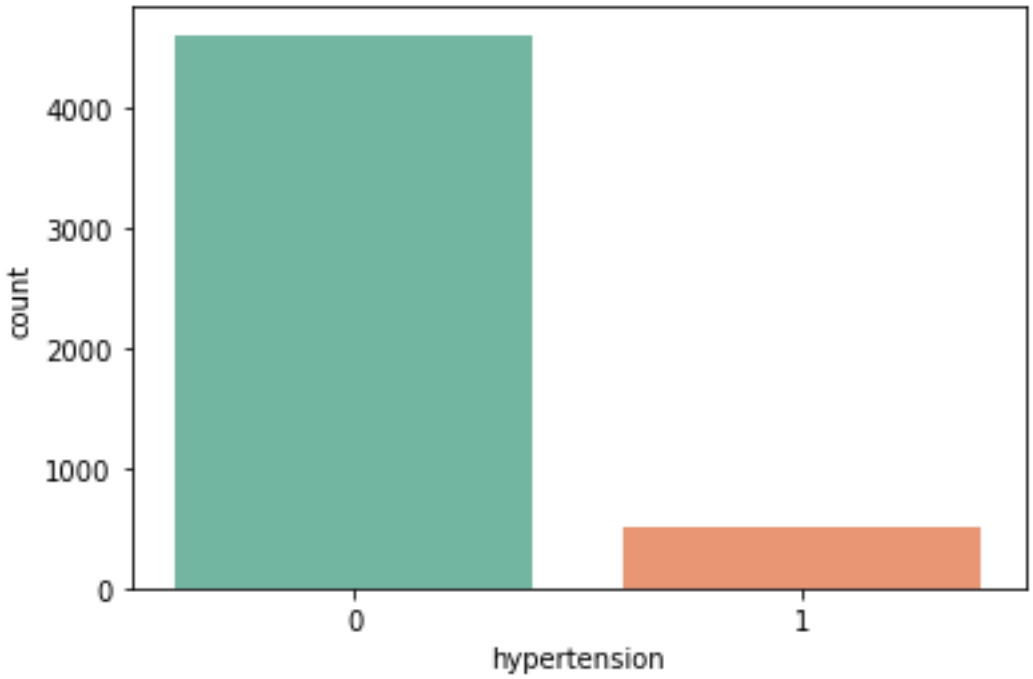
Κατανομή του πλήθους των εγγραφών ως προς τον τόπο της κατοικίας:



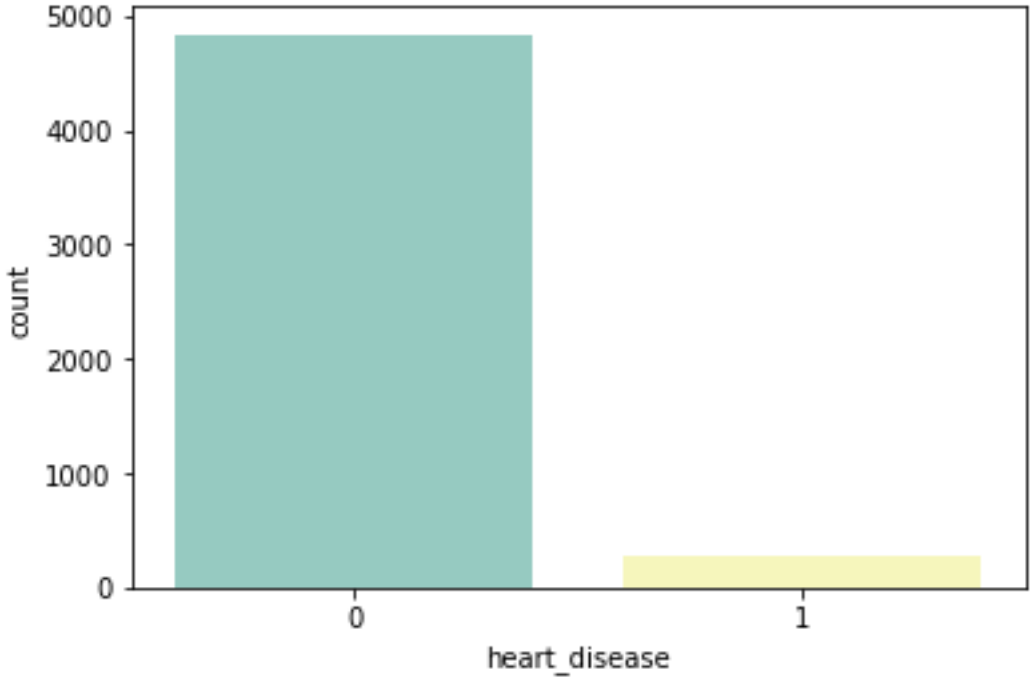
Εικόνα 14 - Κατανομή πλήθους εγγραφών ως προς τον τύπο εργασίας

Κατανομή του πλήθους των εγγραφών ως προς τον τύπο εργασίας:

Κατανομή του πλήθους των εγγραφών ως προς την ύπαρξη υπέρτασης:

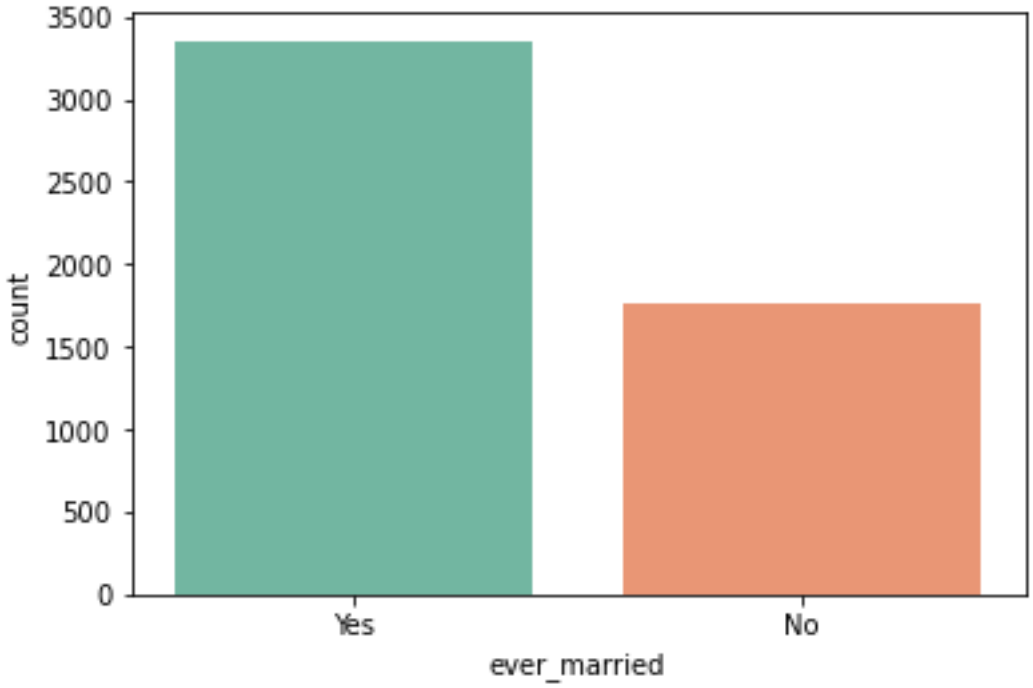


Εικόνα 15 - Κατανομή πλήθους εγγραφών ως προς την ύπαρξη υπέρτασης



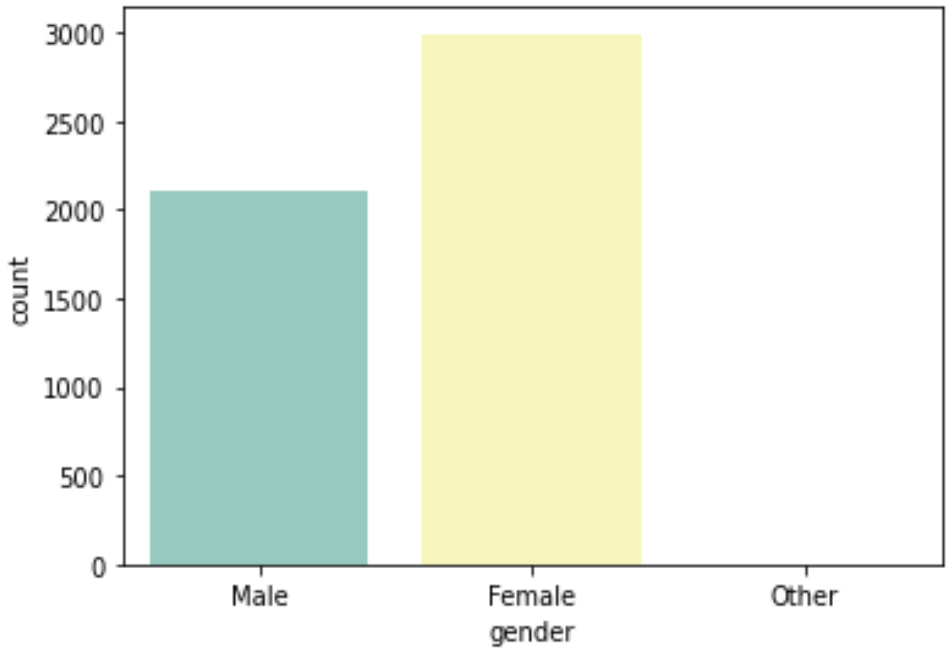
Εικόνα 16 - Κατανομή πλήθους εγγραφών ως προς την ύπαρξη καρδιακής ασθένειας

Κατανομή του πλήθους των εγγραφών ως προς την ύπαρξη καρδιακής ασθένειας:



Εικόνα 17 - Κατανομή πλήθους εγγραφών ως προς το εάν είναι έγγαμος ή όχι

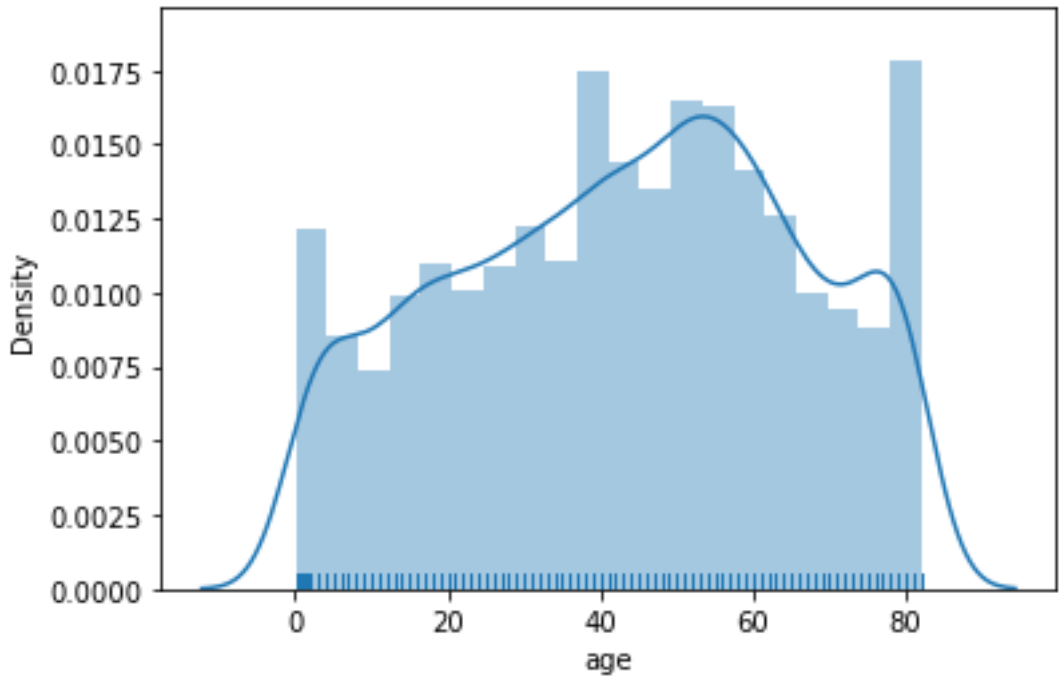
Κατανομή του πλήθους των εγγραφών ως προς το εάν είναι έγγαμος ή όχι:



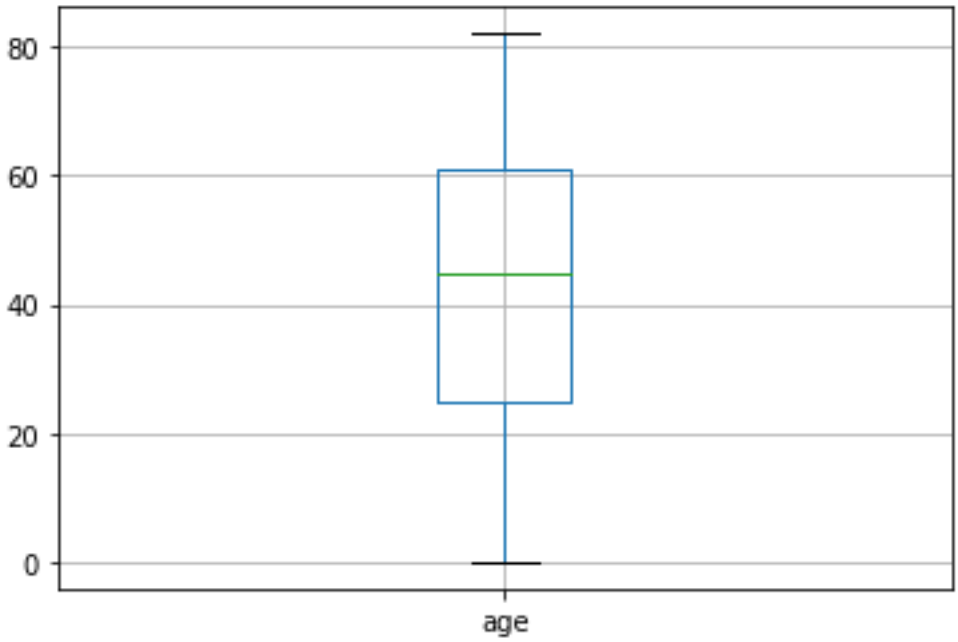
Εικόνα 18 - Κατανομή πλήθους εγγραφών ως προς το φύλο

Κατανομή του πλήθους των εγγραφών ως προς το φύλο:

Το density graph για την ηλικία:



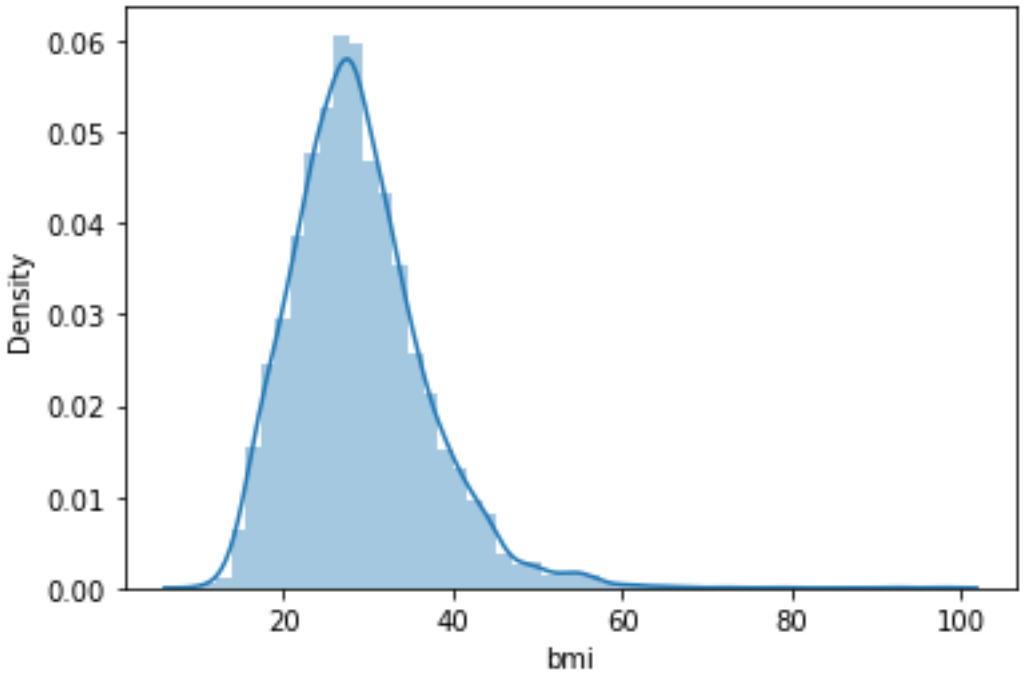
Εικόνα 19 - Density graph για την ηλικία



Εικόνα 20 - Box plot για την ηλικία

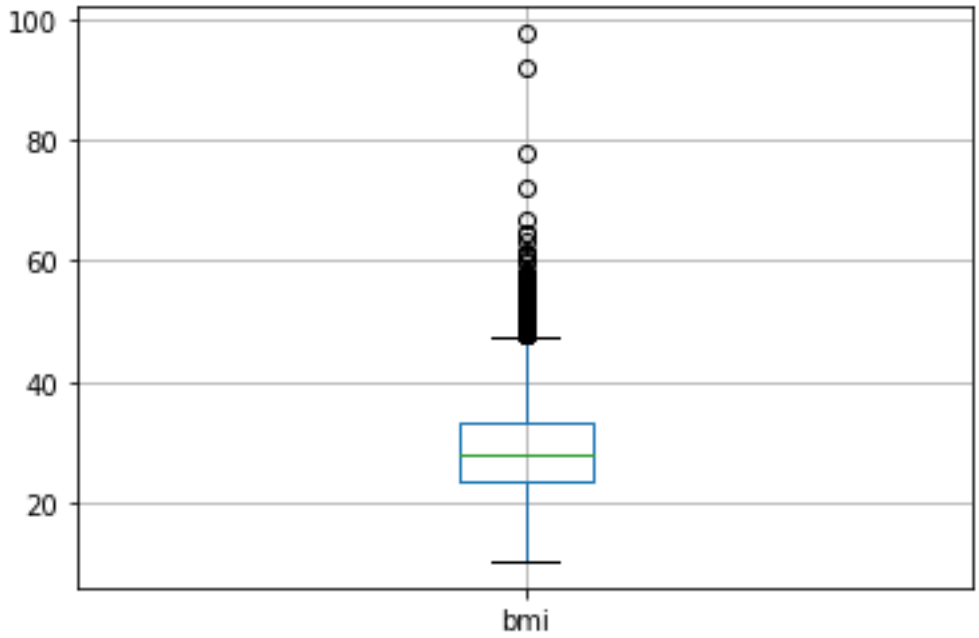
Και το αντίστοιχο box plot για την ηλικία:

Το density graph για το bmi:

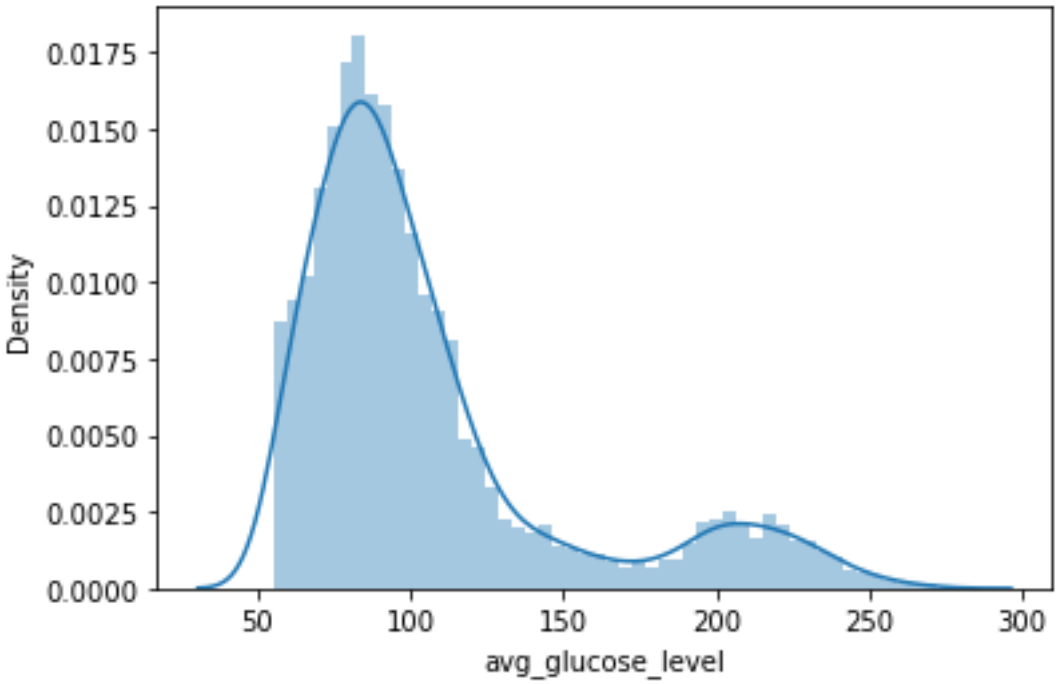


Εικόνα 21 - Density graph για το bmi

Και το αντίστοιχο box plot για το bmi:



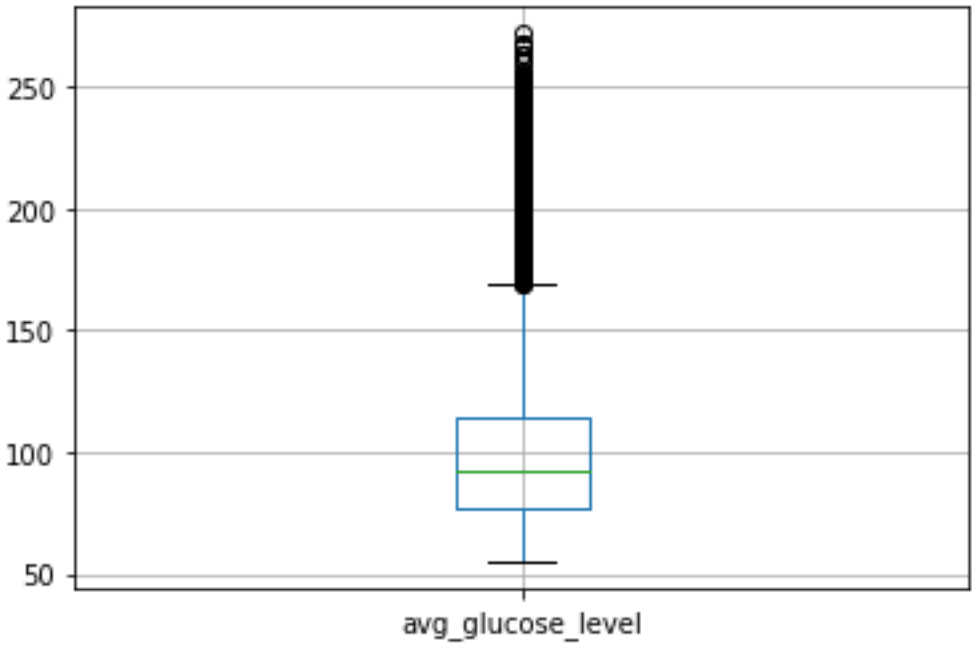
Εικόνα 22 - Box plot για το bmi



Εικόνα 23 - Density graph για το avg_glucose_level

Τέλος, το density graph για το avg_glucose_level:

Και το αντίστοιχο box plot για το avg_glucose_level:



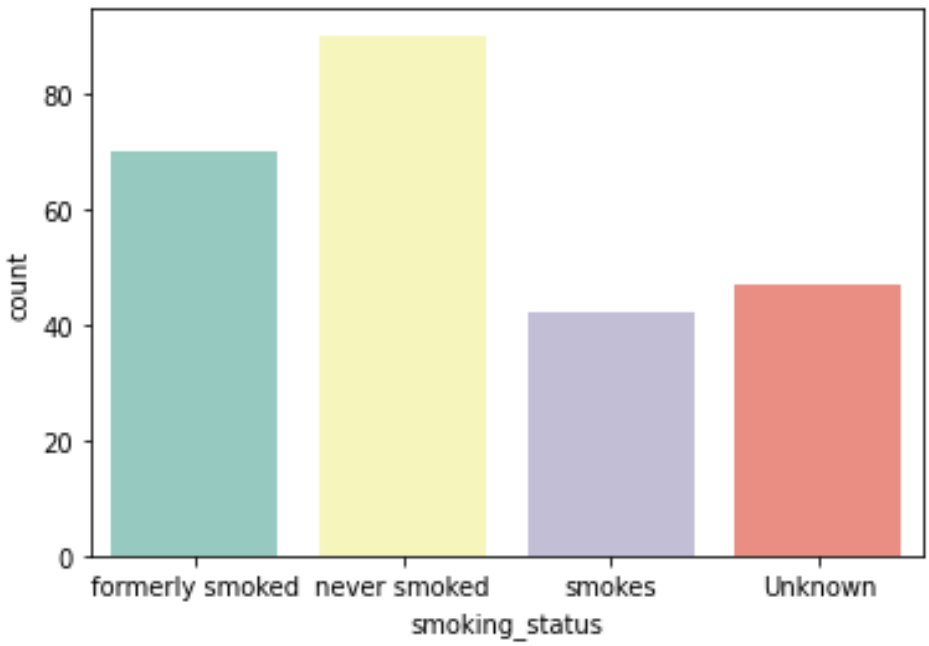
Εικόνα 24 - Box plot για το avg_glucose_level

Στην συνέχεια, διαχωρίσαμε το dataset σε 2 επιμέρους datasets:

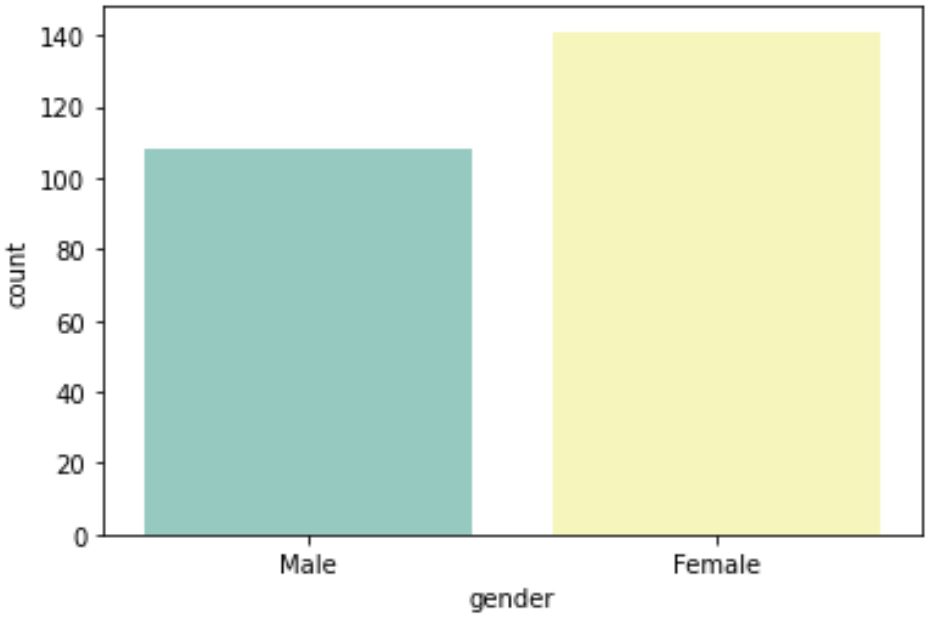
- σε αυτό που περιέχει όλες τις περιπτώσεις που έχουν υποστεί εγκεφαλικό επεισόδιο
- σε αυτό που περιέχει όλες τις περιπτώσεις που δεν έχουν υποστεί εγκεφαλικό επεισόδιο

Ο παραπάνω διαχωρισμός ήταν σημαντικός, προκειμένου να μπορέσουμε να εξάγουμε καλύτερα κάποια συμπεράσματα σχετικά με την κατανομή χαρακτηριστικών στους δύο τύπους υποσυνόλων δεδομένων.

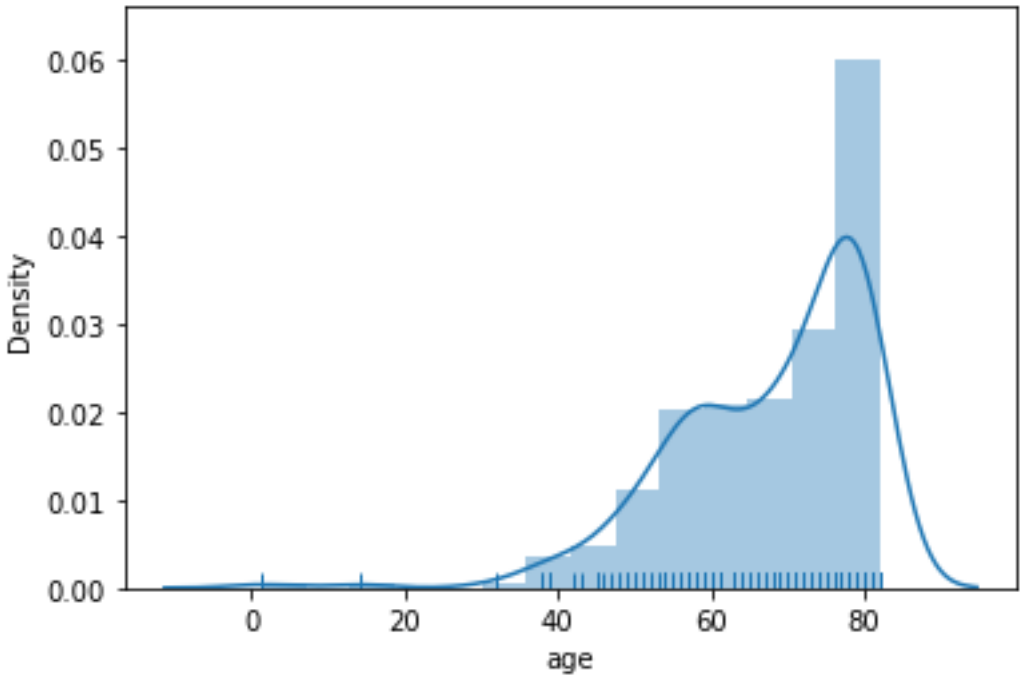
Παραθέτουμε ενδεικτικά γραφήματα, όμοια με πριν άλλα μόνο για όσους έχουν υποστεί εγκεφαλικό επεισόδιο:



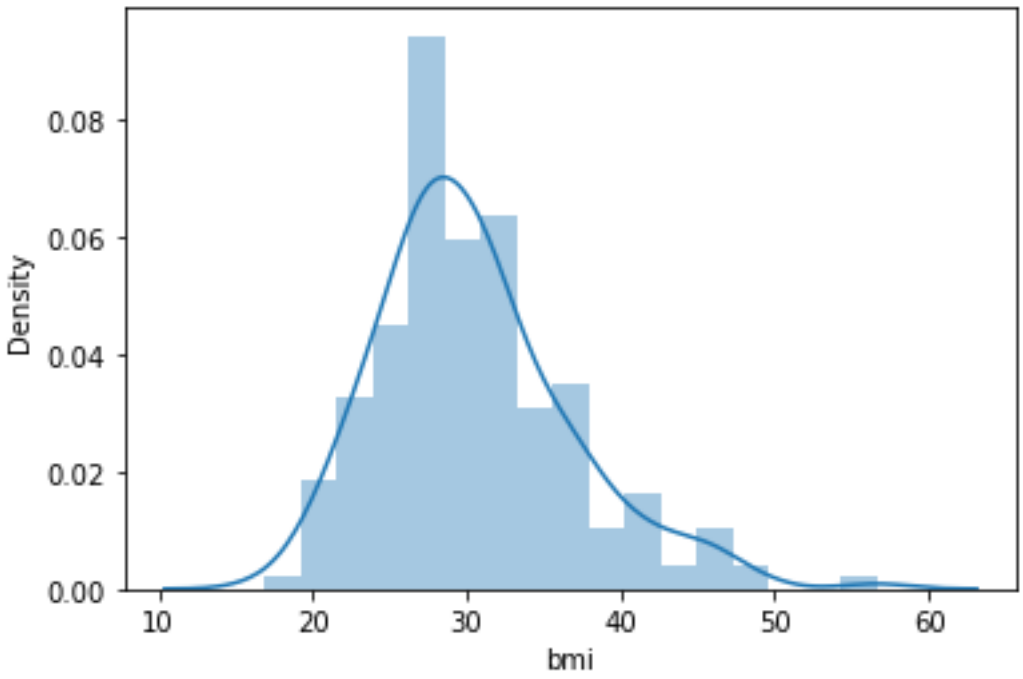
Εικόνα 25 - Γράφημα για όσους έχουν υποστεί εγκεφαλικό επεισόδιο



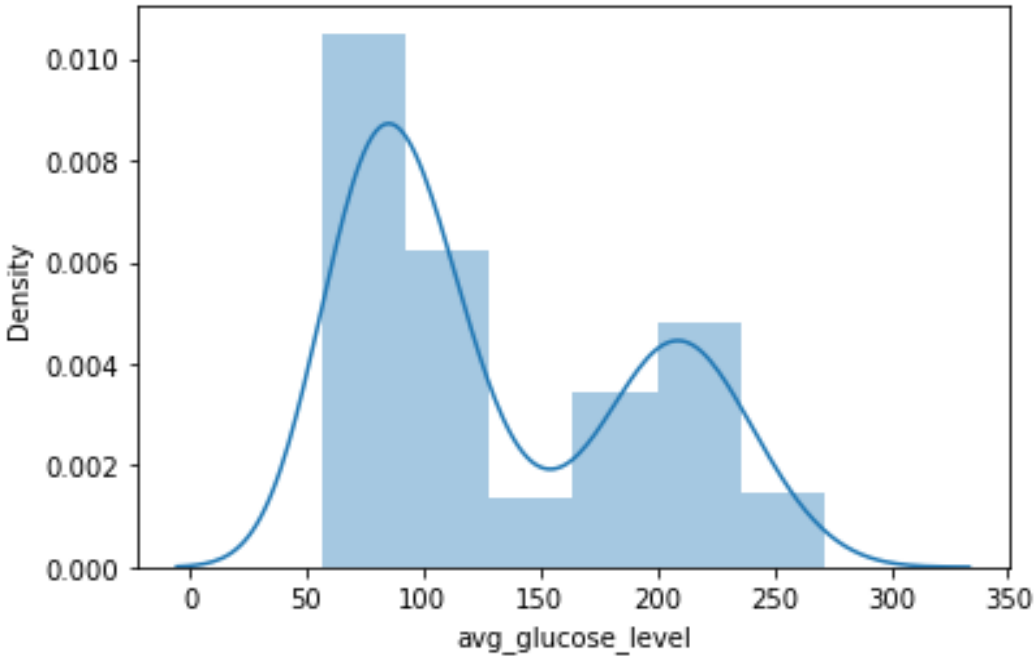
Εικόνα 26 - Γράφημα για όσους έχουν υποστεί εγκεφαλικό επεισόδιο



Εικόνα 27 - Γράφημα για όσους έχουν υποστεί εγκεφαλικό επεισόδιο

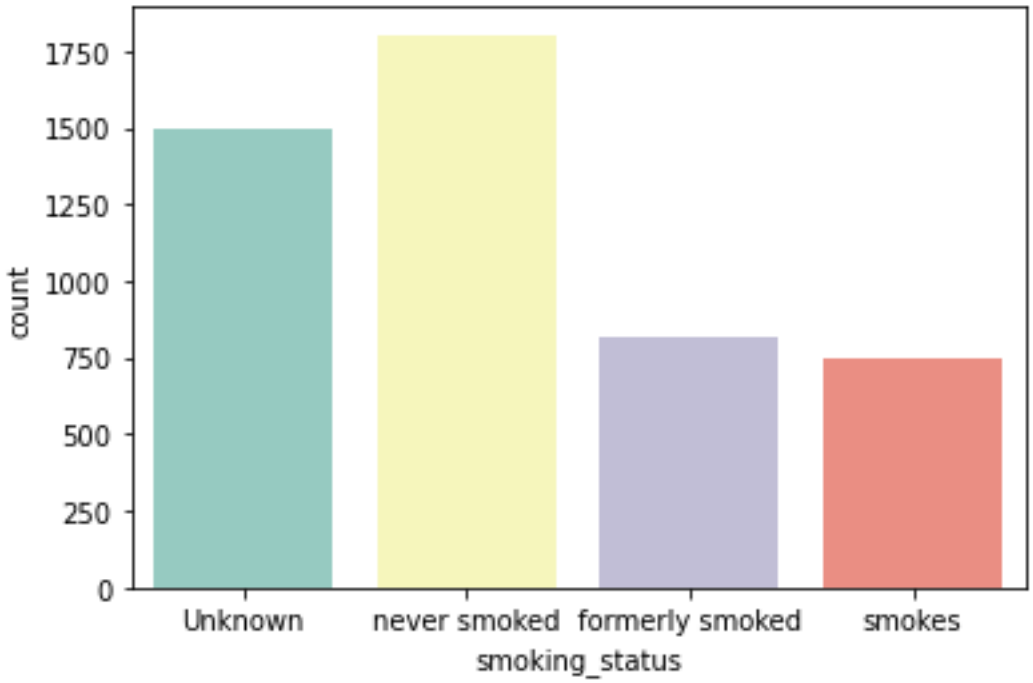


Εικόνα 28 - Γράφημα για όσους έχουν υποστεί εγκεφαλικό επεισόδιο

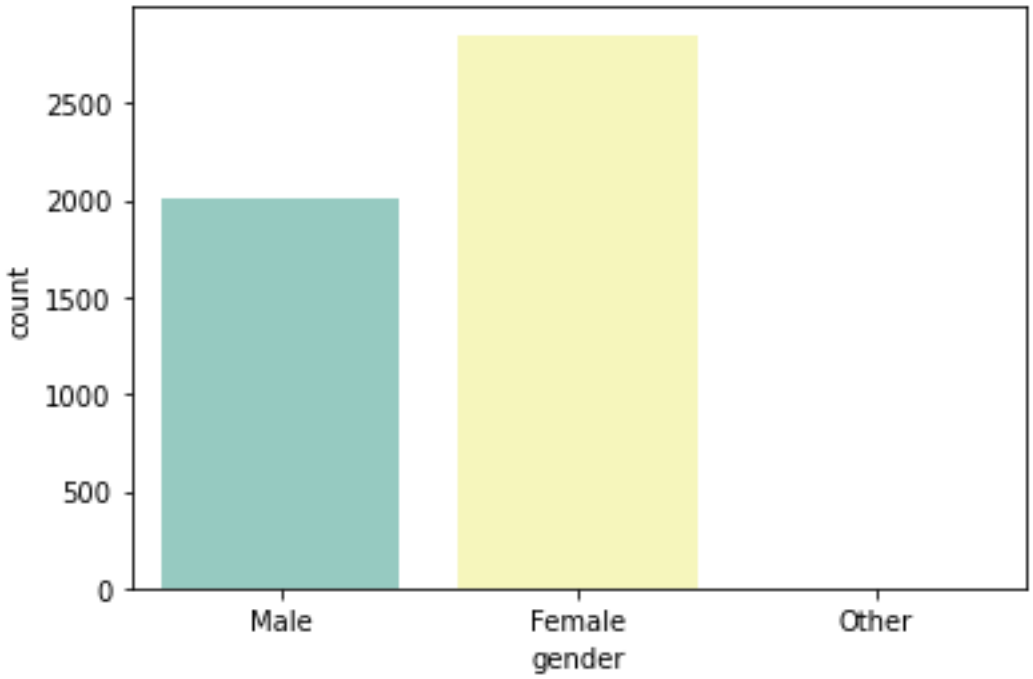


Εικόνα 29 - Γράφημα για όσους έχουν υποστεί εγκεφαλικό επεισόδιο

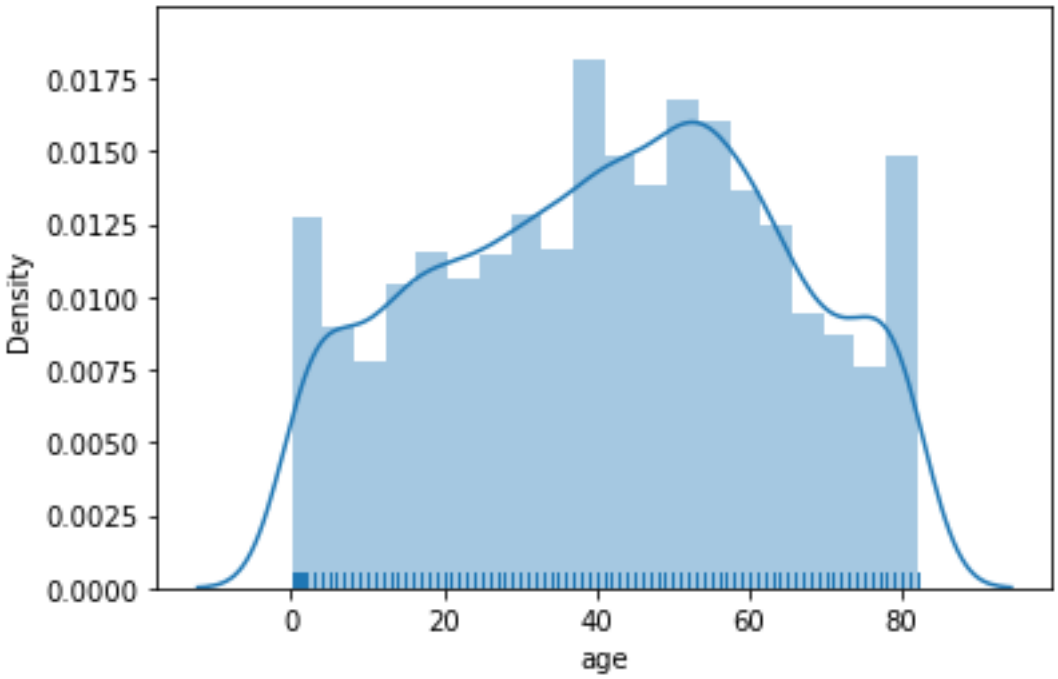
Και τέλος, παραθέτουμε αντίστοιχα γραφήματα για εκείνες τις περιπτώσεις, οι οποίες δεν έχουν υποστεί εγκεφαλικό επεισόδιο:



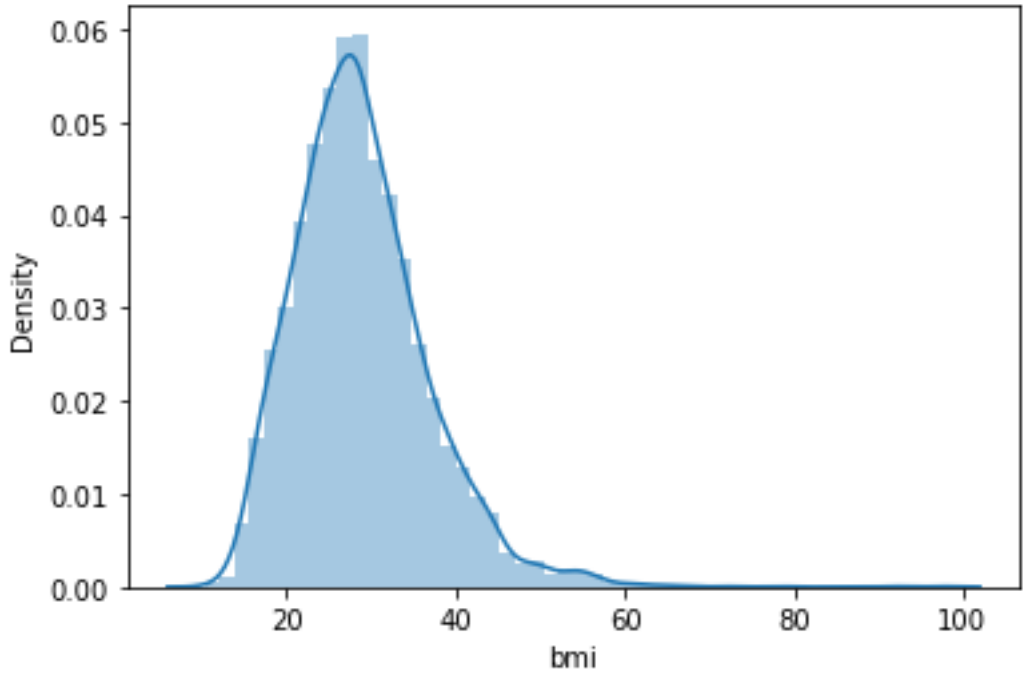
Εικόνα 30 - Γράφημα για όσους δεν έχουν υποστεί εγκεφαλικό επεισόδιο



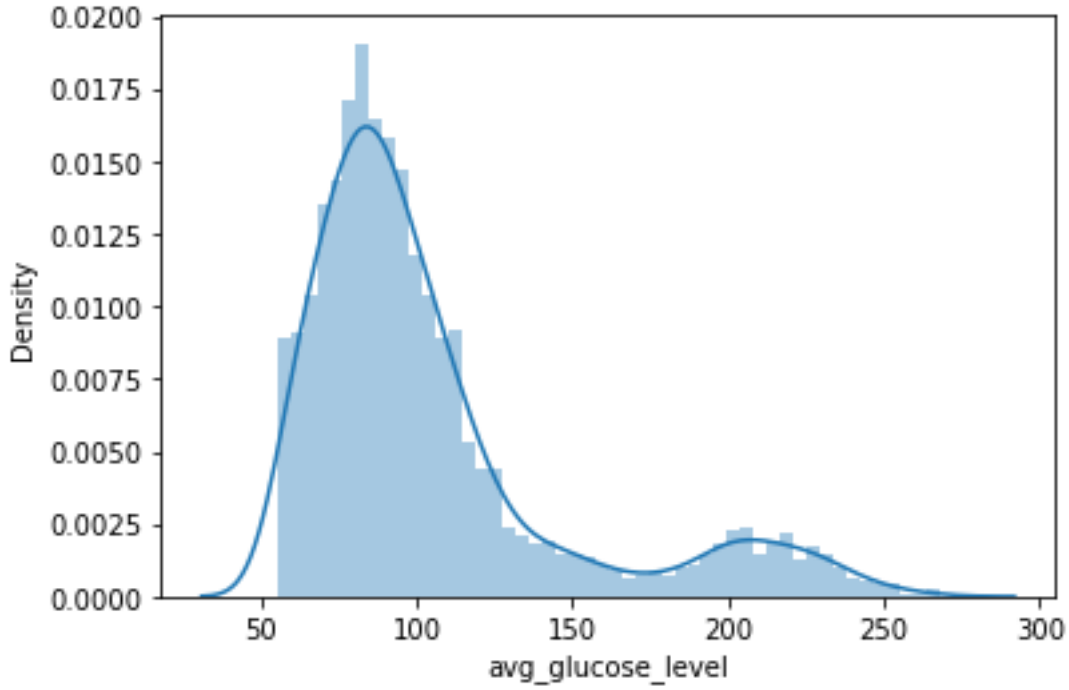
Εικόνα 31 - Γράφημα για όσους δεν έχουν υποστεί εγκεφαλικό επεισόδιο



Εικόνα 32 - Γράφημα για όσους δεν έχουν υποστεί εγκεφαλικό επεισόδιο



Εικόνα 33 - Γράφημα για όσους δεν έχουν υποστεί εγκεφαλικό επεισόδιο



Εικόνα 34 - Γράφημα για όσους δεν έχουν υποστεί εγκεφαλικό επεισόδιο

6.4. Προεπεξεργασία Δεδομένων

Προκειμένου να ενισχύσουμε την αξιοπιστία της πειραματικής μας διαδικασίας και να δημιουργήσουμε εισόδους αποδεκτές στους αλγορίθμους κατηγοριοποίησης εφαρμόσαμε τις παρακάτω τεχνικές:

- συμπλήρωση των τιμών του bmi που έλειπαν με τον μέσο όρο των τιμών της στήλης του bmi. Με τον τρόπο αυτό δώσαμε λύση στο πρόβλημα των ελλιπών τιμών της συγκεκριμένης στήλης
- Χρήση της τεχνικής Label Encoding για την μετατροπή των στηλών που περιείχαν αλφαριθμητικές τιμές σε αριθμητικές. Οι στήλες αυτές είναι οι: gender, ever_married, work_type, residence_type,smoking_status
- Χρήση της βιβλιοθήκης imblearn και πιο συγκεκριμένα της κλάσης SMOTE που περιέχει η συγκεκριμένη βιβλιοθήκη για να δώσουμε λύση στο πρόβλημα της ανομοιόμορφης κατανομής μεταξύ των 2 κλάσεων εξόδου

6.5. Πειραματικά Αποτελέσματα

Όσο αφορά τα αποτελέσματα της πειραματικής διαδικασίας, παρουσιάζουμε τον παρακάτω πίνακα, ο οποίος δείχνει την τιμή του Precision για κάθε έναν από τους 4 αλγορίθμους που εφαρμόσαμε:

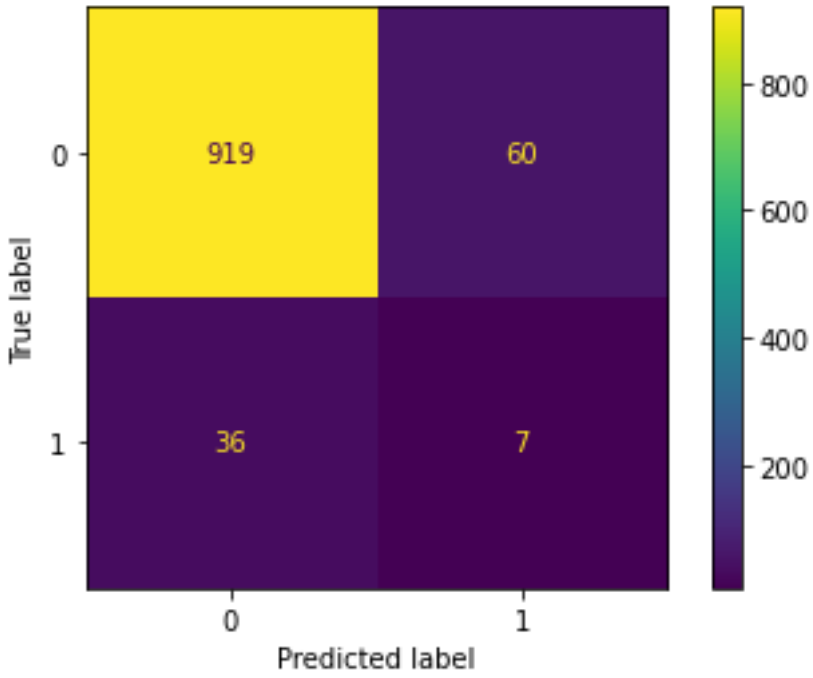
Αλγόριθμος	**Presicion**
Random Forests	0.91
Extra Trees	0.91
kNN	0.78
SVM	0.69

Πίνακας 1 - Αλγόριθμοι και precision

Από τα παραπάνω μπορούμε να διαπιστώσουμε ότι οι αλγόριθμοι Random Forests και Extra Trees παρουσιάζουν καλύτερα αποτελέσματα σε σχέση τους αλγόριθμους kNN και SVM (με τον kNN να παρουσιάζει ελαφρώς υψηλότερο precision σε σχέση με το μοντέλο SVM).

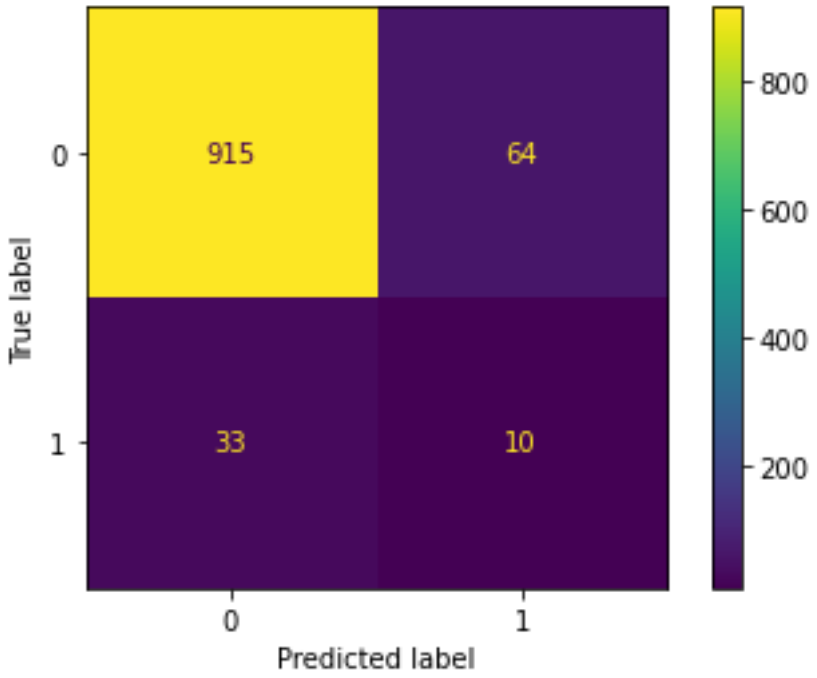
Τέλος, παρουσιάζουμε τα αντίστοιχα confusion matrixes, τα οποία προέκυψαν από την εκτέλεση των 4 παραπάνω αλγορίθμων.

Το confusion matrix για τον Random Forests:

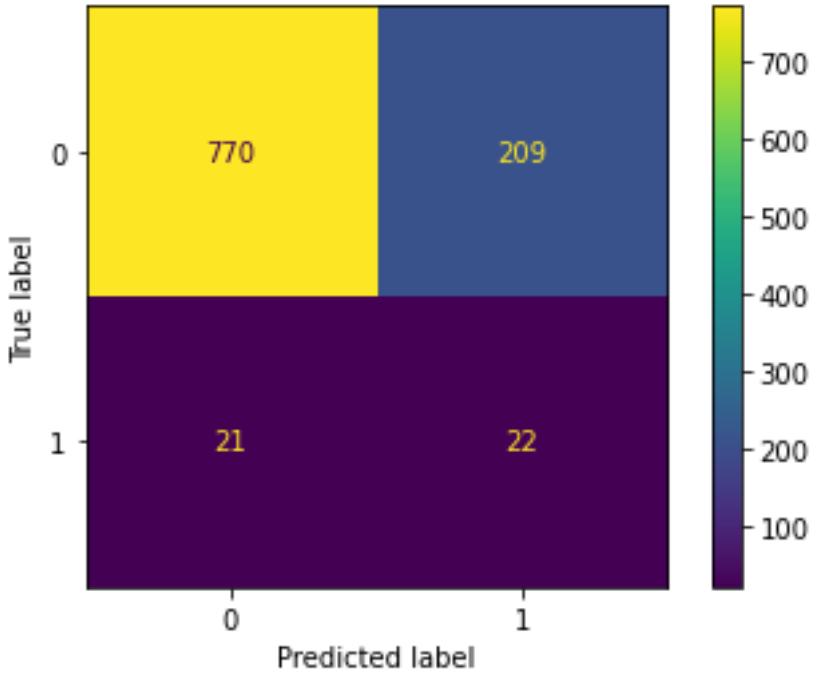


Εικόνα 35 - Confusion matrix για Random Forests

Το confusion matrix για τον Extra Trees:

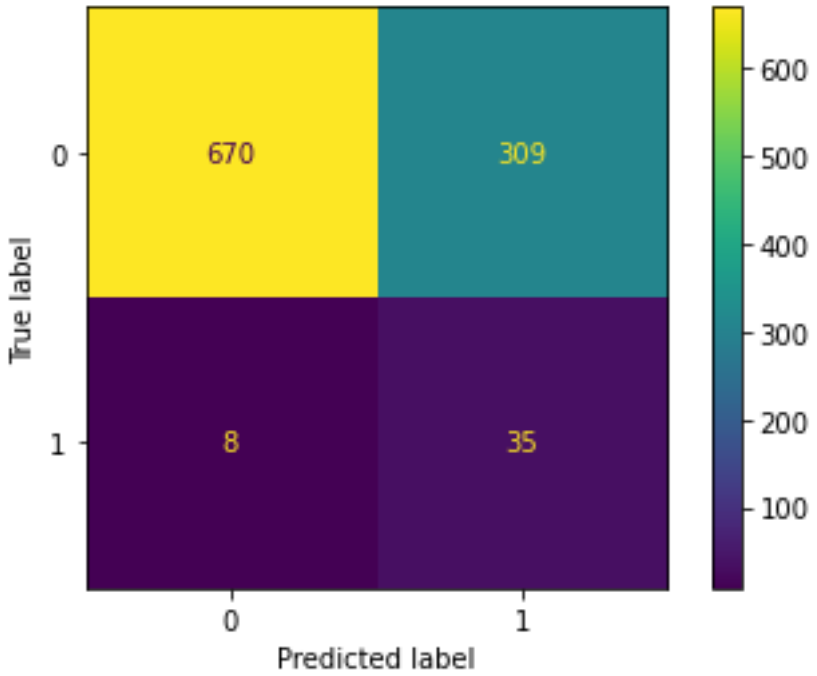


Εικόνα 36 - Confusion matrix για τον Extra Trees



Εικόνα 37 - Confusion matrix για τον kNN

Το confusion matrix για τον kNN:



Εικόνα 38 - Confusion matrix για το μοντέλο SVM

Και τέλος, το confusion matrix για το μοντέλο SVM:

6.6. Συμπεράσματα – Προτάσεις

Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας, αφού πραγματοποιήσαμε την παρουσίαση του κατάλληλου θεωρητικού υποβάθρου, μελετήσαμε την πρακτική εφαρμογή αλγορίθμων που πραγματοποιούν classification πάνω σε ιατρικά δεδομένα και πιο συγκεκριμένα πάνω σε δεδομένα που σχετίζονται με την ύπαρξη εγκεφαλικού επεισοδίου.

Η ανάλυση και η εξαγωγή συμπερασμάτων με βάση ιατρικά δεδομένα χρησιμοποιώντας μεθόδους μηχανικής μάθησης αποτελεί μια από τις μεγαλύτερες προκλήσεις της εποχής μας, καθώς οι χρησιμοποιούμενες τεχνικές του Machine Learning αναμένεται να αποτελέσουν ένα χρήσιμο εργαλείο σε διάφορους επιστημονικούς κλάδους, ανάμεσα στους οποίους είναι και η ιατρική επιστήμη.

Όσο αφορά το dataset το οποίο χρησιμοποιήθηκε, οι 5110 εγγραφές που περιείχε ήταν ικανές για να μπορούμε να εξάγουμε χρήσιμα συμπεράσματα. Βέβαια, σαν μελλοντική κατεύθυνση θα μπορούσε να είναι η χρήση μεγαλύτερων dataset με γνώμονα τις 2 παρακάτω παραμέτρους:

- το πλήθος των εγγραφών
- το πλήθος των χαρακτηριστικών που λαμβάνονται υπόψη

Τα χαρακτηριστικά που περιείχε το συγκεκριμένο dataset παρέχουν μια καλή εικόνα, θα μπορούσαν όμως να ληφθούν υπόψη και ακόμη πιο εξειδικευμένα, τα οποία θα μπορούσαν να εξαχθούν από ιατρικές βάσεις δεδομένα και πάντα σε συνεργασία με την ιατρική επιστημονική κοινότητα.

Ενδιαφέρον επίσης αποτελεί το γεγονός ότι υπήρχε ανισοκατανομή μεταξύ των 2 υποψήφιων κλάσεων (ύπαρξη εγκεφαλικού επεισοδίου ή όχι). Σαν μελλοντική επέκταση θα μπορούσε να αποτελέσει η εύρεση εναλλακτικών datasets, ακόμη και η προσπάθεια για επεξεργασία και πιο ισομοιρασμένη αναλογία του τρέχοντος dataset (πέραν της μεθόδου SMOTE που χρησιμοποιήθηκε).

Στην συνέχεια παραθέτουμε κάποια χρήσιμα συμπεράσματα που εξάγαμε από την ανάλυση του dataset:

- Μεγάλο ποσοστό εμφάνισης στις εγγραφές του dataset είχαν οι ηλικίες 40 και 80.
- Όσο αφορά τον δείκτη μάζας σώματος, παρουσιάζεται μεγαλύτερη πυκνότητα εμφάνισης σε τιμές κοντά στο 30.
- Για το μέσο επίπεδο γλυκόζης, τιμές κοντά στο 100 είναι οι πιο συχνά εμφανιζόμενες.
- Οι γυναίκες ήταν περισσότερες από τους άντρες.
- Είναι χαρακτηριστικό το γεγονός, ότι στην περίπτωση των ατόμων που έχουν είχαν υποστεί εγκεφαλικό, αρκετοί δεν είχαν καπνίσει ποτέ.

- Στην πλειονότητά τους, άτομα που είχαν υποστεί εγκεφαλικό επεισόδιο ήταν άτομα μεγάλης ηλικίας (60 – 80)

Πρόκειται για συμπεράσματα, τα οποία είμαστε σε θέση να εξάγουμε εξαιτίας της δυνατότητας που μας παρέχει η Python για οπτικοποίηση των δεδομένων. Πρόκειται για μια πολύ ισχυρή γλώσσα, η οποία διαθέτει αρκετές βιβλιοθήκες για επεξεργασία και οπτικοποίηση δεδομένων. Πιο συγκεκριμένα οι τύποι γραφημάτων που χρησιμοποιήθηκαν είναι:

- Ραβδογράμματα
- Box Plots
- Density Graphs

Σαν μελλοντική κατεύθυνση στην συγκεκριμένη περίπτωση δίνεται η χρήση επιπλέον διαγραμμάτων και διαφορετικών τύπων, έτσι ώστε να έχουμε μια πληρέστερη εικόνα για το χρησιμοποιούμενο ΄σύνολο δεδομένων.

Σημαντική επίσης ήταν η προσπάθεια που έγινε πάνω στο ζήτημα της προεπεξεργασία των δεδομένων. Πιο συγκεκριμένα:

- Για το πρόβλημα των ελλιπών τιμών του bmi, χρησιμοποιήθηκε η τεχνική του μέσου όρου. Εναλλακτικά θα μπορούσαν να χρησιμοποιηθούν και τεχνικές πρόβλεψης τιμής με μεθόδους όπως : νευρωνικά δίκτυα, linear regression, logistic regression, clustering εγγραφών. Θα είχε μεγάλο ενδιαφέρον να προχωρήσουμε σε συγκριτική αξιολόγηση έτσι ώστε εξετάσουμε εάν βελτιώνονται ή όχι τα πειραματικά αποτελέσματα των αλγορίθμων classification.
- Για την μετατροπή των αλφαριθμητικών τιμών σε αριθμητικές, χρησιμοποιήθηκε η τεχνική του Label encooding. Η συγκεκριμένη μέθοδος λειτούργησε με τρόπο αποδοτικό και χωρίς να εισάγει καθυστέρηση στην εκτέλεση της πειραματικής διαδικασίας. Σίγουρα θα μπορούσαν να χρησιμοποιηθούν και εναλλακτικές μέθοδοι όπως ή one hot encoding, έτσι ώστε να γίνουν οι απαραίτητες συγκρίσεις ειδικά όσο αφορά την χρονική επιβάρυνση.
- Τέλος, όσο αφορά το πρόβλημα της ανομοιόμορφης κατανομής χρησιμοποιήθηκε η κλάση SMOTE, η οποία ενίσχυσε αρκετά την πραγματοποίηση της πειραματικής διαδικασίας και την αξιοπιστία των αποτελεσμάτων. Μια μελλοντική βιβλιογραφική μελέτη για την εύρεση εναλλακτικών τεχνικών, οι οποίες να είναι πρακτικά εφαρμόσιμες με χρήση μιας γλώσσας όπως η Python, προτείνεται και στην συγκεκριμένη περίπτωση.

Πολύ σημαντικό επίσης ρόλο έπαιξε η γλώσσα Python με την χρήση των βιβλιοθηκών που διαθέτει στην πραγματοποίηση της πειραματικής διαδικασίας. Συνοψίζουμε τις παρακάτω σημαντικές βιβλιοθήκες:

- pandas, για την διαχείριση pandas dataframes
- numpay, για την διαχείριση numpy arrays
- sklern, για χρήση των αλγορίθμων μηχανικής μάθησης
- searborn για οπτικοποίηση
- matplotlib για οπτικοποίηση

Θα κλείσουμε την ενότητα των συμπερασμάτων, με τους αλγορίθμους που χρησιμοποιήθηκαν και τα πειραματικά τους αποτελέσματα. Πιο συγκεκριμένα χρησιμοποιήθηκαν οι αλγόριθμοι:

- Random Forests
- Extra Trees
- kNN
- SVM

Σαν μελλοντική κατεύθυνση δίνεται σαφώς και η μελέτη επιπλέον αλγορίθμων και τεχνικών classification, έτσι ώστε να υπάρχει πληρέστερη και πιο αξιόπιστη εικόνα. Στα πλαίσια της συγκεκριμένης διπλωματικής εργασίας, αποφασίσαμε να υλοποιήσουμε 4 από τις δημοφιλέστερες τεχνικές που πραγματοποιούν classification και που έχουν εφαρμοστεί ξανά στο παρελθόν σε ιατρικά δεδομένα με ικανοποιητικά αποτελέσματα.

Όσο αφορά τα αποτελέσματα της δικής μας πειραματικής διαδικασίας, οι 4 μέθοδοι συγκρίθηκαν ως προς το Precision και συμπεράναμε τα εξής:

- Στην περίπτωση των Random Forests και Extra Trees παρατηρήσαμε την βέλτιστη συμπεριφορά, με την τιμή του precision να είναι 0.91.
- Στην περίπτωση του αλγορίθμου kNN, παρατηρήθηκε σχετικά ικανοποιητική συμπεριφορά με την τιμή του precision να είναι 0.78
- Στην περίπτωση της μεθόδου SVM είχαμε τα χειρότερα αποτελέσματα, με το precision να παίρνει την τιμή 0.61.

Επίσης αξίζει να σημειωθεί ότι η εκτέλεση των πειραμάτων δεν ήταν χρονοβόρα διαδικασία, γεγονός που ενισχύει την αξιοπιστία των μοντέλων μας (ειδικά στην περίπτωση των Random Forests και Extra Trees). Μικρός χρόνος εκτέλεσης σε συνδυασμό με τιμή του precision που πλησιάζει την ιδανική τιμή του 1, αποτελεί το επιδιωκόμενο.

Τέλος, όσο αφορά τα confusion matrixes τα οποία παραθέσαμε, παρατηρούμε ότι η μεγαλύτερη δυσκολία βρίσκεται στην απόδοση τις ετικέτας 1 (εγκεφαλικό), γεγονός που πρέπει να παρακινήσει να συνεχίσουμε με:

- την εξέταση εναλλακτικών αλγορίθμων και προσεγγίσεων
- την χρήση πιο εξειδικευμένων datasets
- η χρήση των βέλτιστων παραμέτρων των χρησιμοποιούμενων μοντέλων
- η συνέχιση της συνεργασίας με την επιστημονική κοινότητα

Βιβλιογραφικές Αναφορές

1. Abbasi, A. A., & Younis, M. (2007). A survey on clustering algorithms for wireless sensor networks. *Computer communications*, *30*(14-15), 2826-2841.
2. Acierno, L. J. (1994). *The history of cardiology*. CRC Press.
3. Antman, E. M., & Loscalzo, J. (2016). Precision medicine in cardiology. *Nature Reviews Cardiology*, *13*(10), 591-602.
4. Ashley, E. A., & Niebauer, J. (2004). *Cardiology explained*.
5. Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, *3*, 19-48.
6. Benabdellah, A. C., Benghabrit, A., & Bouhaddou, I. (2019). A survey of clustering algorithms for an industrial context. *Procedia computer science*, *148*, 291-302.
7. Boley, D., Gini, M., Gross, R., Han, E. H. S., Hastings, K., Karypis, G., ... & Moore, J. (1999). Partitioning-based clustering for web document categorization. *Decision Support Systems*, *27*(3), 329-341.
8. Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.
9. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547-555.
10. Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, *318*(6), 517-518.
11. Cheng, T. O. (2004). The current state of cardiology in China. *International journal of cardiology*, *96*(3), 425-439.
12. Cleophas, T. J., & Zwinderman, A. H. (2015). *Machine Learning in Medicine-a Complete Overview* (pp. 17-24). Springer International Publishing.
13. Cleophas, T. J., Zwinderman, A. H., & Cleophas-Allers, H. I. (2013). *Machine learning in medicine* (Vol. 9). Dordrecht, The Netherlands:: Springer.
14. Constantine, G., Shan, K., Flamm, S. D., & Sivananthan, M. U. (2004). Role of MRI in clinical cardiology. *The Lancet*, *363*(9427), 2162-2171.
15. Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine learning and the profession of medicine. *Jama*, *315*(6), 551-552.
16. Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920-1930.
17. El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham.
18. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, *2*(3), 267-279.
19. Garcia, E. V., Faber, T. L., Cooke, C. D., Folks, R. D., Chen, J., & Santana, C. (2007). The increasing role of quantification in clinical nuclear cardiology: the Emory approach. *Journal of nuclear cardiology*, *14*(4), 420-432.

20. Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning.
21. Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning.
22. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Machine learning basics. *Deep learning*, *1*(7), 98-164.
23. Gui, C., & Chan, V. (2017). Machine learning in medicine. *University of Western Ontario Medical Journal*, *86*(2), 76-78.
24. Harrington, P. (2012). *Machine learning in action*. Simon and Schuster.
25. Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology*, *21*(1), 1-11.
26. Jain, A. K., Topchy, A., Law, M. H., & Buhmann, J. M. (2004, August). Landscape of clustering algorithms. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* (Vol. 1, pp. 260-263). IEEE.
27. Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*(1), 1503-1509.
28. Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., ... & Dudley, J. T. (2018). Artificial intelligence in cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668-2679.
29. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.
30. Langley, P. (1996). *Elements of machine learning*. Morgan Kaufmann.
31. Limbacher, M., Douglas, P. S., & Germano, G. (1998). Radiation safety in the practice of cardiology. *Journal of the American College of Cardiology*, *31*(4), 892-915.
32. Marr, C., & Bowen, M. (Eds.). (2011). *Cardiology of the Horse E-Book*. Elsevier Health Sciences.
33. Marsland, S. (2011). *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.
34. Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.
35. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.
36. Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
37. Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86-97.

38. Narmadha, D., alias Balamurugan, A., Sundar, G. N., & Priya, S. J. (2016). Survey of clustering algorithms for categorization of patient records in healthcare. *Indian Journal of Science and Technology*, *9*(8), 1-5.
39. Niederer, S. A., Lumens, J., & Trayanova, N. A. (2019). Computational models in cardiology. *Nature Reviews Cardiology*, *16*(2), 100-111.
40. Patton, J. A., Slomka, P. J., Germano, G., & Berman, D. S. (2007). Recent technologic advances in nuclear cardiology. *Journal of nuclear cardiology*, *14*(4), 501-513.
41. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.
42. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, *56*(9), 455.
43. Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.
44. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.
45. Rozanski, A. (2014). Behavioral cardiology: current advances and future directions. *Journal of the American College of Cardiology*, *64*(1), 100-110.
46. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
47. Shobha, G., & Rangaswamy, S. (2018). Machine learning. In *Handbook of statistics* (Vol. 38, pp. 197-228). Elsevier.
48. Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC medical research methodology*, *19*(1), 1-18.
49. Sra, S., Nowozin, S., & Wright, S. J. (Eds.). (2012). *Optimization for machine learning*. Mit Press.
50. Timmis, A., Townsend, N., Gale, C. P., Torbica, A., Lettino, M., Petersen, S. E., ... & Vardas, P. (2020). European Society of Cardiology: cardiovascular disease statistics 2019. *European heart journal*, *41*(1), 12-85.
51. Topol, E. J., & Teirstein, P. S. (2015). *Textbook of interventional cardiology E-Book*. Elsevier Health Sciences.
52. Van Calster, B., & Wynants, L. (2019). Machine learning in medicine. *New England Journal Of Medicine*, *380*(26), 2588-2588.
53. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, *15*(11), e1002689.
54. Verma, A. A., Murray, J., Greiner, R., Cohen, J. P., Shojania, K. G., Ghassemi, M., ... & Mamdani, M. (2021). Implementing machine learning in medicine. *CMAJ*, *193*(34), E1351-E1357.
55. Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications medicine*, *1*(1), 1-3.

56. Waljee, A. K., & Higgins, P. D. (2010). Machine learning in medicine: a primer for physicians. *Official journal of the American College of Gastroenterology| ACG*, *105*(6), 1224-1226.

57. Yusuf, S., Cairns, J., Camm, J., Fallen, E. L., & Gersh, B. J. (Eds.). (2003). *Evidence based cardiology*. London: BMJ Books.

58. Zaret, B. L., & Wackers, F. J. (1993). Nuclear cardiology. *New England Journal of Medicine*, *329*(11), 775-783.

59. Zhang, X. D. (2020). Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence* (pp. 223-440). Springer, Singapore.

60. Zu Eissen, S. M., & Stein, B. (2002, December). Analysis of clustering algorithms for web-based search. In *International Conference on Practical Aspects of Knowledge Management* (pp. 168-178). Springer, Berlin, Heidelberg.

Παράρτημα Α: Οπτικοποίηση δεδομένων

```
import pandas as pd
import matplotlib.pyplot as plt
import sklearn
import numpy as np
from pandas.plotting import scatter_matrix
from matplotlib import cm
import seaborn as sns
import math
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier



mydata = pd.read_csv('healthcare-dataset-stroke-data.csv')
res1 = mydata.describe()
res2 = mydata.describe(include="O")

sns.countplot(mydata['stroke'], palette = "Set2")
plt.show()

sns.countplot(mydata['smoking_status'], palette = "Set3")
plt.show()

sns.countplot(mydata['Residence_type'], palette = "Set2")
plt.show()

sns.countplot(mydata['work_type'], palette = "Set3")
plt.show()

sns.countplot(mydata['hypertension'], palette = "Set2")
plt.show()
sns.countplot(mydata['heart_disease'], palette = "Set3")
plt.show()

sns.countplot(mydata['ever_married'], palette = "Set2")
plt.show()


sns.countplot(mydata['gender'], palette = "Set3")
plt.show()

sns.distplot(mydata['age'], rug=True)
plt.show()
mydata.boxplot(column="age")
```

```
plt.show()


sns.distplot(mydata["bmi"])
plt.show()
mydata.boxplot(column="bmi")
plt.show()


sns.distplot(mydata["avg_glucose_level"])
plt.show()
mydata.boxplot(column="avg_glucose_level")
plt.show()

#################################################################


with_stroke=mydata[mydata["stroke"]==1]

sns.countplot(with_stroke['smoking_status'], palette = "Set3")
plt.show()

sns.countplot(with_stroke['Residence_type'], palette = "Set2")
plt.show()

sns.countplot(with_stroke['work_type'], palette = "Set3")
plt.show()

sns.countplot(with_stroke['hypertension'], palette = "Set2")
plt.show()
sns.countplot(with_stroke['heart_disease'], palette = "Set3")
plt.show()

sns.countplot(with_stroke['ever_married'], palette = "Set2")
plt.show()


sns.countplot(with_stroke['gender'], palette = "Set3")
plt.show()

sns.distplot(with_stroke['age'], rug=True)
plt.show()
with_stroke.boxplot(column="age")
plt.show()


sns.distplot(with_stroke["bmi"])
```

```
plt.show()
with_stroke.boxplot(column="bmi")
plt.show()


sns.distplot(with_stroke["avg_glucose_level"])
plt.show()
with_stroke.boxplot(column="avg_glucose_level")
plt.show()


non_stroke=mydata[mydata["stroke"]==0]

sns.countplot(non_stroke['smoking_status'], palette = "Set3")
plt.show()

sns.countplot(non_stroke['Residence_type'], palette = "Set2")
plt.show()

sns.countplot(non_stroke['work_type'], palette = "Set3")
plt.show()

sns.countplot(non_stroke['hypertension'], palette = "Set2")
plt.show()
sns.countplot(non_stroke['heart_disease'], palette = "Set3")
plt.show()

sns.countplot(non_stroke['ever_married'], palette = "Set2")
plt.show()


sns.countplot(non_stroke['gender'], palette = "Set3")
plt.show()

sns.distplot(non_stroke['age'], rug=True)
plt.show()
non_stroke.boxplot(column="age")
plt.show()


sns.distplot(non_stroke["bmi"])
plt.show()
non_stroke.boxplot(column="bmi")
plt.show()


sns.distplot(non_stroke["avg_glucose_level"])
```

```
plt.show()
non_stroke.boxplot(column="avg_glucose_level")
plt.show()
```

Παράρτημα Β: Εφαρμογή αλγορίθμων Classification

```
import pandas as pd
import matplotlib.pyplot as plt
import sklearn
import numpy as np
from pandas.plotting import scatter_matrix
from matplotlib import cm
import seaborn as sns
import math
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.metrics import plot_confusion_matrix
from imblearn.over_sampling import SMOTE
from sklearn import svm



mydata = pd.read_csv('healthcare-dataset-stroke-data.csv')

feature_names = ['gender', 'age', 'hypertension', 'heart_disease', 'ever_married','work_type','Residence_type','avg_glucose_level','bmi','smoking_status']
X = mydata[feature_names]

y = mydata['stroke']

le = LabelEncoder()

X['gender']= le.fit_transform(X['gender'])

X['ever_married']= le.fit_transform(X['ever_married'])

X['work_type']= le.fit_transform(X['work_type'])

X['Residence_type']= le.fit_transform(X['Residence_type'])

X['smoking_status']= le.fit_transform(X['smoking_status'])
```

```
sum_bmi=0

count_bmi=0


for x in X['bmi']:
   if  not math.isnan(x):
      sum_bmi=sum_bmi+x
      count_bmi=count_bmi+1

avg_bmi = sum_bmi/count_bmi



for i in range(len(X['bmi'])):
   if math.isnan(X['bmi'][i]):
      X['bmi'][i]=avg_bmi



X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)

smote=SMOTE()
x_train_smote, y_train_smote=smote.fit_resample(X_train, y_train)

#random forest
rf=RandomForestClassifier()

rf.fit(x_train_smote,y_train_smote)

y_pred_rf=rf.predict(X_test)

print("\nRandom Forests: \n")
accuracy = accuracy_score(y_test, y_pred_rf)

print('Accuracy: %.3f \n' % accuracy)

plot_confusion_matrix(rf, X_test, y_test)

plt.show()



#extratrees
et=ExtraTreesClassifier()

et.fit(x_train_smote,y_train_smote)
```

```
y_pred_et=et.predict(X_test)

print("\nExtra Trees: \n")

accuracy = accuracy_score(y_test, y_pred_et)

print('Accuracy: %.3f \n' % accuracy)

plot_confusion_matrix(et, X_test, y_test)

plt.show()



#knn classifier
knn=KNeighborsClassifier()

knn.fit(x_train_smote,y_train_smote)

y_pred_knn=knn.predict(X_test)

print("\nKNN: \n")

accuracy = accuracy_score(y_test, y_pred_knn)
print('Accuracy: %.3f \n' % accuracy)

plot_confusion_matrix(knn, X_test, y_test)

plt.show()

#svm classifier
clf = svm.SVC()

clf.fit(x_train_smote,y_train_smote)

y_pred_svm=clf.predict(X_test)

print("\nSVM: \n")

accuracy = accuracy_score(y_test, y_pred_svm)

print('Accuracy: %.3f \n' % accuracy)

plot_confusion_matrix(clf, X_test, y_test)

plt.show()
```