ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ.: ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ
ΚΑΤΕΥΘΥΝΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πρόβλεψη στεφανιαίας νόσου με χρήση τεχνικών μηχανικής μάθησης

Νικόλαος Μπούνας
Αριθμός Μητρώου: ΜΕ2028

Επιβλέπων Καθηγητής:
Ηλίας Μαγκλογιάννης, Καθηγητής

ΠΕΙΡΑΙΑΣ

ΙΟΥΝΙΟΣ 2022

Περίληψη

Ο τομέας της υγείας έχει επηρεαστεί σε μεγάλο βαθμό από την τεχνολογία προς το καλύτερο, καθώς με την ραγδαία ανάπτυξη τής εξελίσσεται και η ίδια. Όπως κάθε σύγχρονο τεχνολογικό μέσο, έτσι και τα ιατρικά μηχανήματα και εργαλεία παράγωγουν μεγάλο βαθμό δεδομένων. Τα τελευταία χρόνια ο όγκος των δεδομένων που παράγονται καθημερινά σε όλο τον κόσμο είναι δύσκολα διαχειρίσιμος και συνεχώς αναζητούνται οι καλύτερες λύσεις για την βελτίωση της διαχείρισης τους. Πολλές φορές, μέσα από αυτά μπορούν δημιουργηθούν εύλογα και χρήσιμα συμπεράσματα για διάφορα θέματα της καθημερινότητας όπως στην προκειμένη περίπτωση η υγεία. Μέσω της καταχώρησης ιστορικού των ιατρικών δεδομένων για διάφορα συμπτώματα που αφορούν ασθενείς προκύπτουν συμπεράσματα μέσω των οποίων προτείνονται πρόνοιες και προλήψεις για τους ίδιους τους ασθενείς αλλά και για μελλοντικούς με παρόμοιο ιατρικό ιστορικό ή συμπτώματα. Η παρούσα διπλωματική έχει ως στόχο τη δημιουργία μοντέλων πρόβλεψης και με τη βοήθεια ενός συνόλου δεδομένων από ανθρώπους που υποβλήθηκαν σε εξετάσεις που αφορούν την καρδιά τους, να προβλεφθεί αν τα επόμενα 10 χρόνια ενδεχομένως θα εμφανίσουν συμπτώματα στεφανιαίας νόσου. Η στεφανιαία νόσος είναι μια πάθηση όπου οι αρτηρίες της καρδιάς δεν μπορούν να παραδώσουν αρκετό αίμα πλούσιο σε οξυγόνο στην καρδιά. Η συγκεκριμένη πάθηση μπορεί να προκαλέσει σοβαρότατα προβλήματα στην υγεία ενός ανθρώπου και να τον οδηγήσει μέχρι και στον θάνατο. Έτσι, έχοντας το σύνολο δεδομένων και με τη βοήθεια ορισμένων μοντέλων μηχανικής μάθησης, αλλά και άλλων τεχνικών μηχανικής μάθησης έγινε προσπάθεια να βρεθεί το καλύτερο από αυτά που να προβλέπει, σύμφωνα με τα δεδομένα κάθε ασθενούς, με την καλύτερη ακρίβεια την εμφάνιση της στεφανιαίας νόσου μέσα στα επόμενα 10 χρόνια. Επίσης, μέσω της εκτίμησης ρίσκου και με τη βοήθεια αλγορίθμων που απονέμουν πόντους σε ορισμένα συμπτώματα του ασθενή, γίνεται μια αντίστοιχη πρόβλεψη υπολογίζοντας το ποσοστό της πιθανότητας εμφάνισης της στεφανιαίας νόσου μέσα από διαφορετικές συνθήκες. Με τη βοήθεια των παραπάνω, έγινε προσπάθεια υλοποίησης μιας απλής εφαρμογής μέσω της οποίας ένας ασθενής, καταχωρώντας τα συμπτώματα και τα δεδομένα του, μπορεί να ενημερωθεί για το μέλλον σε ό,τι αφορά το καρδιαγγειοακό πρόβλημα της στεφανιαίας νόσου. Στο τέλος, αναλύονται τα συμπεράσματα που προέκυψαν στην πειραματική διαδικασία.

Abstract

The health sector has been influenced by technology, as it is evolving rapidly. Like any modern technological tool, medical devices and tools produce a great deal of data. In recent years the daily volume of data produced around the world has become difficult to manage and the best solutions are constantly being sought to improve their management. Many times, through them, reasonable and useful conclusions can be drawn on various issues of everyday life, as in the case of health. Through the recording of the history of medical data for various symptoms that concern patients, conclusions are drawn through which provisions and precautions are proposed for the patients themselves, but also for future ones with a similar medical history or symptoms. The purpose of this thesis is to create model predictions through a dataset from people who have undergone examinations concerning their heart, if in the next 10 years they will possibly show symptoms of coronary heart disease. Coronary heart disease is a condition in which the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. This condition can cause serious health problems and even lead to death. Thus, having the dataset and with the help of some machine learning models, but also machine learning techniques, an attempt was made to find the best one that predicts, according to the data of each patient, with the best accuracy the occurrence of coronary heart disease in next 10 years. Also, through risk assessment and with the help of algorithms that award points to certain symptoms of the patient, a corresponding prediction is made by calculating the percentage of the probability of occurrence of coronary heart disease through different conditions. With the help of the above, an attempt was made to implement a simple application through which a patient, by registering his/her symptoms and data, can be informed about the future in terms of the cardiovascular problem of coronary heart disease. In the end, the conclusions that emerged in the experimental process are analyzed.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα καθηγήτη μου, κύριο Ηλία Μαγκλογιάννη για την πολύτιμη βοήθεια του καθ'όλη την διάρκεια της συγγραφής της παρούσας διπλωματικής εργασίας, όπως και τους κοντινούς μου ανθρώπους που πίστεψαν σε μένα και με στήριξαν μέχρι την διεκπεραίωση του Μεταπτυχιακού Προγράμματος Σπουδών.

Πίνακας περιεχομένων

Κεφάλαιο 1 Εισαγωγή 10
1.1 Στεφανιαία νόσος 10
1.2 Στόχος διπλωματικής εργασίας 11
1.3 Δομή διπλωματικής εργασίας 11
Κεφάλαιο 2 Βιβλιογραφική επισκόπηση 12
2.1 Σχετικές έρευνες 12
2.2 Επιστήμη των δεδομένων 13
2.3 Ανάλυση δεδομένων και ιατρική 14
2.4 Μηχανική μάθηση 16
2.5 Τύποι μηχανικής μάθησης 18
2.5.1 Εποπτευόμενη μάθηση 18
2.5.2 Μάθηση χωρίς επίβλεψη 18
2.5.3 Ημι-εποπτευόμενη μάθηση 19
2.5.4 Ενισχυτική μάθηση 19
2.6 Αλγόριθμοι και είδη αλγορίθμων 20
2.6.1 Αλγόριθμος δέντρων απόφασης 20
2.6.2 Αλγόριθμος τυχαίων δασών 22
2.6.3 Αλγόριθμος Κ κοντινότερων γειτόνων (KNN) 23
2.6.4 Αλγόριθμος κατά Μπέιζ (Gaussian) 24
2.6.5 Αλγόριθμος μηχανών διανυσμάτων υποστήριξης 25
2.6.6 Αλγόριθμος λογιστικής παλινδρόμησης 26
2.6.7 Αλγόριθμος τεχνητών νευρωνικών δικτύων 26
2.6.8 Αλγόριθμος Gradient Boosting 27
2.6.9 Αλγόριθμος AdaBoost 28
2.6.10 Αλγόριθμος Extreme Gradient Boosting 28
2.6.11 Αλγόριθμος Light Gradient Descent 29
2.6.12 Αλγόριθμος Stochastic Gradient Descent 30
2.7 Τρόποι εφαρμογής των μοντέλων σε δεδομένα 30
2.7.1 Train Test Split 30
2.7.2 Cross Validation 31
2.8 Μετρικές αξιολόγησης 31

2.8.1 Confusion Matrix 31
2.8.2 Ακρίβεια κατηγοριοποίησης 32
2.8.3 Precision 32
2.8.4 Recall 32
2.8.5 F1 Score 32
2.8.6 Sensitivity and Specificity 33
2.8.7 Roc Curve 33
2.8.8 AUC 33
2.9 Επιλογή χαρακτηριστικών 34
2.9.1 Τεχνικές χωρίς επίβλεψη 34
2.9.2 Εποπτευόμενες τεχνικές 34
2.9.2.1 Μέθοδοι φιλτραρίσματος 34
2.9.2.2 Wrapper Methods 35
2.9.2.3 Embedded methods 36
2.10 Ανισόρροπα δεδομένα 36
Κεφάλαιο 3 Προβλεπτικά μοντέλα στεφανιαίας νόσου με χρήση αλγορίθμων κατηγοριοποίησης 38
3.1 Εισαγωγή 38
3.2 Διερεύνηση συνόλου δεδομένων 38
3.3 Προ-επεξεργασία δεδομένων 39
3.4 Πειραματική διαδικασία με χρήση κατηγοριοποίησης 43
3.4.1 1ο Πείραμα: Απλή κατηγοριοποίηση 43
3.4.2 Κατηγοριοποίηση με χρήση Feature Selection 45
3.4.3 2ο Πείραμα: Chi Square και Correlation Matrix 45
3.4.4 3ο Πείραμα: Mutual Information και Correlation Matrix 47
3.4.5 4ο Πείραμα: Chi Square και Correlation Matrix με χρήση Smote 48
3.4.6 5ο Πείραμα: Mutual Information και Correlation Matrix με χρήση Smote 49
Κεφάλαιο 4 Εκτίμηση ρίσκου 50
4.1 Εισαγωγή 50
4.2 Εκτίμηση ρίσκου: Framingham Score 51
4.3 Εκτίμηση ρίσκου: TIMI Score 54
4.4 Εκτίμηση ρίσκου: Grace Score 55
4.5 Εκτίμηση ρίσκου και Feature Selection 58
Κεφάλαιο 5 Εφαρμογή πρόβλεψης στεφανιαίας νόσου 59
5.1 Εισαγωγή 59

5.2 Εφαρμογή 59
Κεφάλαιο 6 Συμπεράσματα 62
Πίνακας ορολογίας 63
Συντμήσεις – Αρτικόλεξα – Ακρωνύμια 64
Παράρτημα Ι Βασικός κώδικας μοντέλων μηχανικής μάθησης 65
Βιβλιογραφία 68

Λίστα πινάκων

Πίνακας 1. Σύγκριση μοντέλων απλής κατηγοριοποίησης 44
Πίνακας 2. Κατηγορικά δεδομένα βάσει P-value 45
Πίνακας 3.Σύγκριση μοντέλων με Chi Square και Correlation Matrix 46
Πίνακας 4. Σύγκριση μοντέλων με Mutual Information και Correlation Matrix 48
Πίνακας 5. Σύγκριση μοντέλων Chi Square και Correlation Matrix με SMOTE 49
Πίνακας 6. Σύγκριση μοντέλων με Mutual Info και Correlation Matrix με SMOTE 49
Πίνακας 7. Framingham Score - Ηλικία 52
Πίνακας 8.Framingham Score - Χοληστερόλη 52
Πίνακας 9. Framingham Score - Συστολική πίεση 52
Πίνακας 10.Framingham Score - Διαστολική πίεση 52
Πίνακας 11.Framingham Score - Διαβήτης 52
Πίνακας 12. Framingham Score - Καπνιστής 52
Πίνακας 13.Framingham Score - Τελικό σκορ 53
Πίνακας 15. Ηλικία και πόντοι Grace Score 56
Πίνακας 16.Ανακοπή και πόντοι Grace Score 56
Πίνακας 17. Οξύ έμφραγμα μυοκαρδίου και πόντοι Grace Score 56
Πίνακας 18. Heart Rate και πόντοι Grace Score 56
Πίνακας 19. Συστολική πίεση και πόντοι Grace Score 57
Πίνακας 20. Προβληματικό καρδιογράφημα και πόντοι Grace Score 57
Πίνακας 21. Κρεατίνη και πόντοι Grace Score 57
Πίνακας 22. Ένζυμα-δείκτες και πόντοι Grace Score 57
Πίνακας 23. Διαδερμική επαναγγείωση και πόντοι Grace Score 57
Πίνακας 24. Grace Score για θάνατο εντός του νοσοκομείου 58
Πίνακας 25. Grace score για θάνατο έξι μήνες από το εξιτήριο 58

Λίστα εικόνων

Εικόνα 1. Γράφημα ανάλυσης δεδομένων 14
Εικόνα 2. Οι τέσσερις τύποι των Data Analytics 15
Εικόνα 3. Αλγόριθμος δέντρων απόφασης 22
Εικόνα 4. Αλγόριθμος τυχαίων δασών 23
Εικόνα 5. Αλγόριθμος KNN 24
Εικόνα 6. Αλγόριθμος SVM 25
Εικόνα 7. Αλγόριθμος τεχνητών νευρωνικών δικτύων 27
Εικόνα 8. Διερεύνηση συνόλου δεδομένων 38
Εικόνα 9. Τύποι μεταβλητών συνόλου δεδομένων 40
Εικόνα 10. Δεδομένα με μηδενικές τιμές 40
Εικόνα 11. Γραφήματα ακραίων τιμών 42
Εικόνα 12. Πίνακας συσχετίσεων 43
Εικόνα 13. Confusion Matrix 44
Εικόνα 14. Πίνακας συσχετίσεων Feature Selection 46
Εικόνα 15. Γράφημα Mi Score 47
Εικόνα 16. Ιστόγραμμα ισορροπημένων δεδομένων 48
Εικόνα 17. Εκτίμηση ρίσκου για ηλικιακές ομάδες γυναικών και αντρών 54
Εικόνα 18. Interface πρόβλεψης στεφανιαίας νόσου με Tkinter 59
Εικόνα 19. Αποτέλεσμα νευρωνικού δικτύου 60
Εικόνα 20. Αποτέλεσμα Framingham Score 61

Κεφάλαιο 1 Εισαγωγή

1.1 Στεφανιαία νόσος

Οι καρδιακές παθήσεις είναι μια γενική φράση για μια ποικιλία καταστάσεων που επηρεάζουν τη δομή της καρδιάς και τον τρόπο λειτουργίας της. Η στεφανιαία νόσος (ή ισχαιμική καρδιοπάθεια) είναι ένας τύπος καρδιακής νόσου όπου οι αρτηρίες της καρδιάς δεν μπορούν να παραδώσουν αρκετό αίμα πλούσιο σε οξυγόνο στην καρδιά. Προκαλείται συχνά από τη χοληστερόλη, μια κηρώδη ουσία που συσσωρεύεται στο εσωτερικό της επένδυσης των στεφανιαίων αρτηριών σχηματίζοντας πλάκα, ένα γεγονός που ονομάζεται αθηροσκλήρωση και προκαλεί ισχαιμία του μυοκαρδίου. Το αθηρωματικό υλικό είναι ένα μαλακό, λιπώδες υλικό το οποίο δημιουργείται στην εσωτερική επιφάνεια των αρτηριών από την αλληλεπίδραση με τα στοιχεία του αίματος (κύτταρα και παράγοντες πήξης) και τα λίπη που μεταφέρονται με το αίμα.

Είναι η κύρια αιτία θανάτου στις Ηνωμένες Πολιτείες, καθώς περίπου 18 εκατομμύρια Αμερικανοί ενήλικες πάσχουν από αυτήν με αποτέλεσμα να είναι η πιο κοινή καρδιακή νόσος, σύμφωνα με τα Κέντρα Ελέγχου και Πρόληψης Νοσημάτων [1], ενώ το 2001 ήταν η αιτία για το 33% των θανάτων σε παγκόσμιο επίπεδο. Για τους περισσότερους ανθρώπους, μπορεί να προληφθεί με έναν υγιεινό τρόπο ζωής, φιλικότερο ως προς την υγεία της καρδιάς. Τα συμπτώματα της μπορεί να είναι διαφορετικά, αναλόγως τον ανθρώπινο οργανισμό, ακόμη και αν έχουν παρόμοια ή και ίδια μορφή στεφανιαία νόσο. Ωστόσο, πολλοί άνθρωποι θεωρούνται ασυμπτωματικοί και δεν γνωρίζουν ότι πάσχουν από αυτήν [2], με αποτέλεσμα ένας ασθενής να διαπιστώσει το πρόβλημα είτε με κάποια καρδιακή προσβολή είτε με άλλη καρδιακή επιπλοκή όπως το έμφραγμα. Αυτό μπορεί να καταστήσει δύσκολη τη διάγνωση της στεφανιαίας νόσου πριν παρουσιαστεί πρόβλημα και αυτός είναι και ο λόγος που η πρόληψη της είναι τόσο σημαντική. Τα πιο κοινά συμπτώματα καρδιακού επεισοδίου μπορεί να είναι:

- **Στηθάγχη** ή ο πόνος στο στήθος που προκαλείται από την καρδιά μπορεί να αισθάνεται σαν πίεση, συμπίεση, δυσπεψία, κάψιμο ή σφίξιμο και μερικές φορές σχετίζεται με τη σωματική δραστηριότητα. Ο πόνος ή η ενόχληση συνήθως ξεκινά πίσω από το στέρνο, αλλά μπορεί επίσης να εμφανιστεί στα χέρια, τους ώμους, τη γνάθο, το λαιμό ή την πλάτη.
- **Κρύος ιδρώτας**
- **Ζάλη**
- **Λιποθυμικά επεισόδια**
- **Ναυτία ή αίσθημα δυσπεψίας**
- **Πονόλαιμος**
- **Δύσπνοια**
- **Διαταραχές ύπνου**
- **Αδυναμία**

Φυσικά, τα παραπάνω μπορούν να διαγνωσθούν με σχετική ευκολία λόγω της μεγάλης εξέλιξης της ιατρικής και της μεγάλης βοήθειας που προσφέρει σήμερα η τεχνολογία. Ένας πάροχος υγειονομικής περίθαλψης μπορεί με σχετική ευκολία να υποβάλει έναν πιθανό ασθενή σε εξετάσεις για να ελέγξει τα επίπεδα χοληστερόλης, τριγλυκεριδίων,

ζάχαρης, λιποπρωτεϊνών ή πρωτεϊνών που είναι σημάδι φλεγμονής. Άλλες εξετάσεις που μπορεί να υποβληθεί ο ασθενής είναι :

- **Ηλεκτροκαρδιογράφημα**
- **Σάρωση στεφανιαίου ασβεστίου**
- **Τεστ άγχους**
- **Μαγνητική τομογραφία**
- **Σάρωση καρδιακής τομογραφίας εκπομπής ποζιτρονίων (PET)**
- **Στεφανιογραφία ή στεφανιογράφημα**
- **Στεφανιαία αξονική τομογραφία**

Στις παραπάνω εξετάσεις μπορεί να υποβληθεί οποιοσδήποτε άνθρωπος οποιαδήποτε στιγμή της ζωής του επιθυμεί. Όμως, επειδή όπως ειπώθηκε, οι άνθρωποι ευαισθητοποιούνται για εξετάσεις καρδιακών νόσων μόνο σε περίπτωση συμπτωμάτων, υπάρχουν σύγχρονοι τρόποι με τους οποίους μπορεί έγκαιρα να προβλεφθεί η πιθανότητα εμφάνισης τους.

1.2 Στόχος διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι με τη βοήθεια διαφόρων εργαλείων μηχανικής μάθησης, της αναλυτικής δεδομένων και της εκτίμησης ρίσκου να προβλεφθεί με όσο το δυνατόν ασφαλέστερο και αξιόπιστο τρόπο η περίπτωση, ένας ασθενής σύμφωνα με τα συμπτώματα, το ιστορικό αλλά και δεδομένα που αφορούν τον εαυτό του, αν πρόκειται να εμφανίσει στεφανιαία νόσο μέσα στα επόμενα 10 χρόνια. Ο ασθενής μπορεί να ενημερωθεί, μέσω της μερίδας ανθρώπων που συμμετείχαν στις εξετάσεις πάνω στις οποίες βασίστηκε το σύνολο δεδομένων της εργασίας και σύμφωνα με τους προβλεπτικούς δείκτες να κάνει την δική του εκτίμηση για τον εαυτό του αλλά και για ανθρώπους του στενού του κύκλου. Επίσης, με την κατασκευή μιας απλής εφαρμογής στην οποία μπορεί ο οποιοσδήποτε να εισάγει τα ανάλογα συμπτώματα ή αριθμούς που αφορούν στατιστικά της καρδιάς και του αίματος του (μέσω καρδιακό ρυθμό, ποσοστό χοληστερόλης κλπ.) μπορεί να γίνει μια εξίσου εκτίμηση εμφάνισης των συμπτωμάτων. Φυσικά, τα αποτελέσματα των παραπάνω τεχνικών, θα αποτελέσουν εκτιμήσεις κατά προσέγγιση που μπορούν εύκολα να μην θεωρηθούν έγκυρες.

1.3 Δομή διπλωματικής εργασίας

Στο πρώτο κεφάλαιο παρέχονται συνοπτικές πληροφορίες σχετικά με τη στεφανιαία νόσο, όπως τα συμπτώματα της και οι σύγχρονοι τρόποι διάγνωσης της. Επίσης, γίνεται αναφορά στον στόχο της εν λόγω διπλωματικής εργασίας. Στο δεύτερο κεφάλαιο γίνονται αναφορές σε σχετικές έρευνες που αφορούν την στεφανιαία νόσο, αναλύονται διάφορες έννοιες της επιστήμης των δεδομένων και της ανάλυσης δεδομένων σε συνδυασμό με την ιατρική, ενώ παράλληλα περιγράφεται η ευρύτερη έννοια της μηχανικής μάθησης (machine learning). Ακόμα, αναφέρονται όλοι οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν για την διερεύνηση του συνόλου δεδομένων. Το τρίτο κεφάλαιο περιλαμβάνει το σύνολο δεδομένων της εργασίας, στο οποίο εφαρμόστηκε η πειραματική διαδικασία και δημιουργήθηκαν τα προβλεπτικά μοντέλα, σκοπός των οποίων είναι η πρόβλεψη συμπτωμάτων στεφανιαίας νόσου σε πιθανούς ασθενείς μέσα στα επόμενα

10 χρόνια. Πραγματοποιήθηκε προ-επεξεργασία δεδομένων και σύμφωνα με αυτήν εφαρμόστηκαν οι αλγόριθμοι μηχανικής μάθησης που περιγράφηκαν στο κεφάλαιο 2. Στο τέταρτο κεφάλαιο τα προβλεπτικά μοντέλα επεκτείνονται με την εφαρμογή τεχνικών εκτίμησης ρίσκου (risk stratification), αλλά και συνδυασμού εκτίμησης ρίσκου και επιλογής χαρακτηριστικών, ενώ έγινε προσπάθεια δημιουργίας μιας εφαρμογής εκτίμησης ρίσκου πιο φιλική προς τον χρήστη. Στο πέμπτο κεφάλαιο, αναλύεται μια απλή ιατρική εφαρμογή που δημιουργήθηκε στα πλαίσια της διπλωματικής, η οποία έχει τη δυνατότητα πρόβλεψης της νόσου είτε μέσω τεχνητών νευρωνικών δικτύων είτε μέσω του Framingham Score. Τέλος, στο κεφάλαιο 6 αναφέρονται τα συμπεράσματα, μελλοντικές έρευνες που μπορεί να προκύψουν βάσει αυτών, καθώς και ο επίλογος της εργασίας.

Κεφάλαιο 2 Βιβλιογραφική επισκόπηση

2.1 Σχετικές έρευνες

Σχετικά με τη στεφανιαία νόσο έχουν πραγματοποιηθεί κατά καιρούς διάφορες έρευνες με σκοπό την ασφαλέστερη κατανόηση των συνθηκών υπό τις οποίες κάποιος άνθρωπος μπορεί να του συμβεί κάποιο καρδιοαγγειακό επεισόδιο. Λόγω αυτού υπάρχουν πιθανότητες να εμφανίσει τα επόμενα χρόνια συμπτώματα στεφανιαίας νόσου ή ακόμα και να οδηγηθεί στο θάνατο. Στην έρευνα που πραγματοποιήθηκε Western Electric Company στην πόλη του Σικάγο στις ΗΠΑ [3] το 1969, συμμετείχαν εργαζόμενοι της οι οποίοι στην ηλικία 44 έως και 55 ετών εμφάνισαν στηθάγχη (πάθηση της καρδιάς με έντονο πόνο στο στήθος) ή επεισόδιο εμφράγματος. Σε αυτούς δόθηκε ένας τυχαίος αύξοντας αριθμός, έτσι ώστε να είναι γνωστό ποιοι είναι οι εργαζόμενοι που θα συνεισφέρουν στο συμπέρασμα της έρευνας, μιας και στην εταιρία θα μπορούσαν να υπάρξουν αποχωρήσεις ή νέες αφίξεις εργαζομένων. Φαίνεται πως κάποια χρόνια αργότερα, σημαντικό ποσοστό ανθρώπων εμφάνισαν συμπτώματα στεφανιαίας νόσου, που όμως ήταν μεγαλύτεροι σε ηλικία σε σχέση με ανθρώπους δεν εμφάνισαν τέτοια συμπτώματα. Σε όλους τους συμμετέχοντες πραγματοποιήθηκαν περαιτέρω έρευνες που αφορούν περισσότερα προβλήματα που θα μπορούσαν να προκαλέσουν καρδιοαγγειακά επεισόδια, όπως το οικογενειακό ιστορικό υγείας, δυσφορία και πόνος στο στήθος, η κατάσταση της χολής, το ύψος και το βάρος του ασθενούς, οι σφυγμοί και καρδιακή πίεση, το κάπνισμα κλπ. Τελικώς, μετά από περίπου 4 χρόνια, φαίνεται πως 88 από τους 1989 εργαζομένους, εμφάνισαν συμπτώματα στεφανιαίας νόσου και τα ασφαλή συμπεράσματα ήταν πως η στεφανιαία νόσος συνδέεται άμεσα με την ηλικία στην οποία έφυγε από τη ζωή ο πατέρας του ασθενούς (άρα κληρονομικότητα), με επεισόδια «κομμένης» ανάσας, με το πεπτικό έλκος, την καρδιακή πίεση, το κάπνισμα και την ποσότητα κατανάλωσης του καφέ.

Μια άλλη έρευνα που δημοσιεύθηκε το 2010 σε συνεργασία από πανεπιστήμιο της Ολλανδίας [4], είχε στόχο να γίνει περισσότερο κατανοητό αν υπάρχει διάκριση των συμπτωμάτων της στεφανιαίας νόσου μεταξύ των δύο φύλων. Είναι ευρέως γνωστό ότι τα καρδιαγγειακά προβλήματα παρατηρούνται περισσότερο στους άντρες παρά στις γυναίκες, όμως στην έρευνα αυτή υποστηρίζεται ότι οι γυναίκες με διαβήτη, υψηλούς δείκτες λιποπρωτεΐνης και υψηλά επίπεδα τριγλυκεριδίων, έχουν περισσότερες πιθανότητες από τους άντρες. Όπως και στην προηγούμενη έρευνα, συγκρίνει διάφορα

καρδιαγγειακά συμπτώματα αλλά αυτή τη φορά ανάμεσα στα δύο φύλα. Καταλήγει στο συμπέρασμα ότι δεν υπάρχει διάφορα ανάμεσα σε άντρες και γυναίκες σχετικά με τη στεφανιαία νόσο και ότι οι περισσότερες έρευνες αναζητούν τελικά τις καλύτερες μεθόδους πρόληψης, καθώς σύμφωνα με τις ηλικιακές ομάδες και τα συμπτώματα των φύλων, μπορεί να βοηθηθούν να προλάβουν τέτοια επεισόδια.

Σε μια ακόμη έρευνα σχετική, μάλιστα, με το θέμα που πρόκειται να αναλυθεί στην παρούσα εργασία, αναφέρεται στην ομάδα του Framingham Heart Study η οποία αναζήτησε εξίσου διαφορές ανάμεσα στα δύο φύλα, την πάροδο του χρόνου που μπορεί να οδηγήσει σε τέτοιες παθήσεις, αλλά και τη διαφορά μεταξύ των περιοχών οι οποίες δείχνουν πληθυσμιακές διαφορές στους βιολογικούς, συμπεριφορικούς και περιβαλλοντικούς παράγοντες και επηρεάζουν την καρδιαγγειακή υγεία [5]. Η έρευνα έδειξε ότι σε 20 χρόνια παρακολούθησης για ορισμένες ηλικιακές ομάδες δεν επηρεάζει το τελικό αποτέλεσμα το φύλο του ασθενούς, καθώς η παχυσαρκία, η υπερχοληστερολαιμία και η υψηλή αρτηριακή πίεση ήταν σημαντικά χαμηλότερα σε ποσοστά. Το ίδιο, ακριβώς, έδειξε και η συχνότητα του καπνίσματος, καθώς και η αρτηριακή πίεση. Τα παραπάνω, είχαν διαφορετική επιρροή τα πρώτα 10 χρόνια της έρευνας, και διαφορετική τα επόμενα 10, καθώς τα επίπεδα των παραπάνω ήταν διαφορετικά.

2.2 Επιστήμη των δεδομένων

Η επιστήμη δεδομένων είναι το πεδίο εφαρμογής προηγμένων τεχνικών ανάλυσης και επιστημονικών αρχών για την εξαγωγή πολύτιμων πληροφοριών από δεδομένα για τη λήψη επιχειρηματικών αποφάσεων, τον στρατηγικό σχεδιασμό και άλλες πολλές χρήσεις. Είναι ολοένα και πιο χρήσιμη ως επιστήμη για τις επιχειρήσεις διότι οι πληροφορίες που δημιουργεί, βοηθούν τους οργανισμούς να αυξήσουν τη λειτουργική τους αποτελεσματικότητα, να εντοπίσουν νέες επιχειρηματικές ευκαιρίες και να οδηγηθούν σε ανταγωνιστικά πλεονεκτήματα έναντι των επιχειρηματικών αντιπάλων[6].

Η επιστήμη δεδομένων ενσωματώνει διάφορους κλάδους όπως για παράδειγμα, την μηχανική δεδομένων (data engineering), την προετοιμασία δεδομένων (data preparation), την εξόρυξη δεδομένων (data mining) , την προγνωστική ανάλυση (predictive analytics), την μηχανική μάθηση (machine learning) και την δεδομένων (data visualization), καθώς και την στατιστική, τα μαθηματικά και τον προγραμματισμό. Γίνεται κυρίως από ειδικευμένους επιστήμονες δεδομένων (Data Scientists), αν και μπορούν να συμμετέχουν και αναλυτές δεδομένων χαμηλότερου επιπέδου. Επιπλέον, πολλοί οργανισμοί βασίζονται πλέον εν μέρει σε επιστήμονες δεδομένων, μια ομάδα που μπορεί να περιλαμβάνει επαγγελματίες επιχειρηματικής ευφυΐας (Business Intelligence), επιχειρησιακούς αναλυτές, επιχειρηματικούς χρήστες με γνώσεις δεδομένων, μηχανικούς δεδομένων και άλλους εργαζόμενους που δεν έχουν επίσημο υπόβαθρο επιστήμης δεδομένων. Ένας επιστήμονας δεδομένων εντοπίζει σημαντικές ερωτήσεις, συλλέγει σχετικά δεδομένα από διάφορες πηγές, αποθηκεύει και οργανώνει δεδομένα, αποκρυπτογραφεί χρήσιμες πληροφορίες και, τέλος, τις μεταφράζει σε επιχειρηματικές λύσεις και κοινοποιεί τα ευρήματα για να επηρεάσει θετικά την επιχείρηση. Εκτός από τη δημιουργία πολύπλοκων ποσοτικών αλγορίθμων και τη σύνθεση μεγάλου όγκου πληροφοριών, οι επιστήμονες δεδομένων διαθέτουν επίσης εμπειρία σε επικοινωνιακές

και ηγετικές δεξιότητες, οι οποίες είναι απαραίτητες για την επίτευξη μετρήσιμων και απτών αποτελεσμάτων σε διάφορους επιχειρηματικούς φορείς[7].

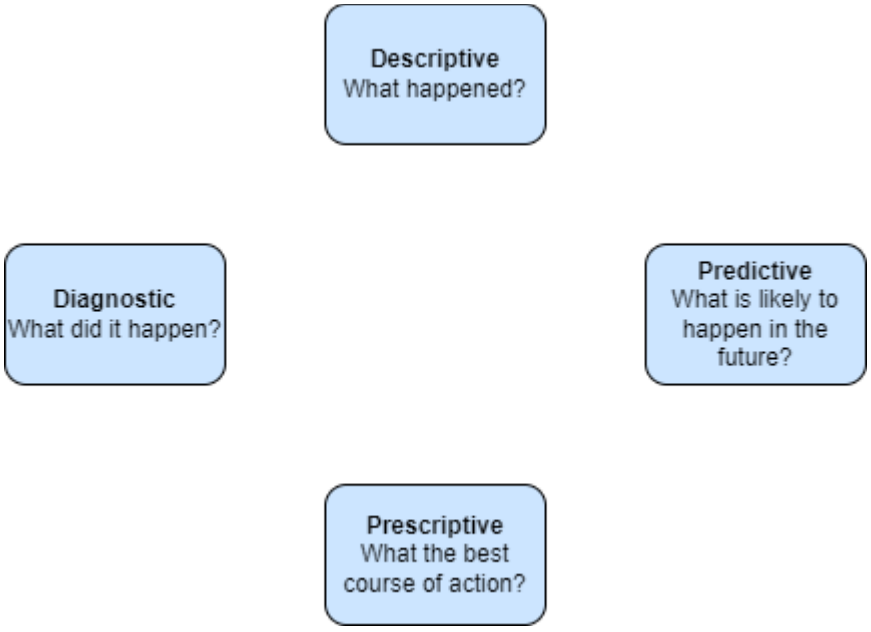
Εικόνα 1. Γράφημα ανάλυσης δεδομένων

2.3 Ανάλυση δεδομένων και ιατρική

Η ανάλυση δεδομένων είναι η διαδικασία ανάλυσης ακατέργαστων δεδομένων με σκοπό την εξαγωγή ουσιαστικών, αξιόπιστων πληροφοριών. Στη συνέχεια, αυτές οι πληροφορίες χρησιμοποιούνται για την ενημέρωση και τη λήψη έξυπνων επιχειρηματικών αποφάσεων. Έτσι, ένας αναλυτής δεδομένων (Data Analyst) θα εξάγει ακατέργαστα δεδομένα, θα τα οργανώσει και στη συνέχεια θα τα αναλύσει, μετατρέποντάς τα από ακατανόητους αριθμούς σε συνεκτικές, κατανοητές πληροφορίες. Έχοντας ερμηνεύσει τα δεδομένα, ο αναλυτής δεδομένων θα διαβιβάσει τα ευρήματά του με τη μορφή προτάσεων ή συστάσεων σχετικά με τα επόμενα βήματα της εταιρείας. Η παραπάνω περιγραφή συμπίπτει σε αρκετό βαθμό με την επιστήμη των δεδομένων, όμως υπάρχουν διαφορές. Μια βασική διαφορά μεταξύ των επιστημόνων δεδομένων και των αναλυτών δεδομένων έγκειται στο τι κάνουν με τα δεδομένα και στα αποτελέσματα που επιτυγχάνουν. Ένας αναλυτής δεδομένων θα προσπαθήσει να απαντήσει σε συγκεκριμένες ερωτήσεις ή να αντιμετωπίσει συγκεκριμένες προκλήσεις που έχουν ήδη εντοπιστεί και είναι γνωστές στην επιχείρηση. Για να γίνει αυτό, εξετάζουν μεγάλα σύνολα δεδομένων με στόχο τον εντοπισμό τάσεων και προτύπων. Στη συνέχεια «οπτικοποιούν» τα ευρήματά τους με τη μορφή γραφημάτων και πινάκων. Αυτές οι οπτικοποιήσεις μοιράζονται με τους βασικούς ενδιαφερόμενους και χρησιμοποιούνται για τη λήψη στρατηγικών αποφάσεων με βάση τα δεδομένα και τις πληροφορίες. Ένας επιστήμονας δεδομένων, από την άλλη πλευρά, εξετάζει ποιες ερωτήσεις θα έπρεπε ή θα μπορούσε να κάνει η επιχείρηση. Σχεδιάζουν νέες διαδικασίες για τη μοντελοποίηση δεδομένων, γράφουν αλγόριθμους, επινοούν μοντέλα πρόβλεψης και εκτελούν προσαρμοσμένες αναλύσεις. Εν ολίγοις, οι αναλυτές δεδομένων αντιμετωπίζουν και λύνουν διακριτές ερωτήσεις σχετικά με δεδομένα, συχνά κατόπιν αιτήματος, αποκαλύπτοντας ιδέες που μπορούν να ληφθούν από άλλους ενδιαφερόμενους, ενώ οι επιστήμονες δεδομένων κατασκευάζουν συστήματα για την αυτοματοποίηση και τη βελτιστοποίηση της συνολικής λειτουργίας της επιχείρησης [8].

Η ανάλυση δεδομένων χωρίζεται σε τέσσερις κύριους [Εικόνα 2] τύπους ανάλυσης δεδομένων[9]:

- **Περιγραφική ανάλυση δεδομένων (Descriptive analytics):** Πρόκειται για έναν απλό τύπο ανάλυσης που εξετάζει τι έχει συμβεί στο παρελθόν και οι δύο κύριες τεχνικές του είναι η συγκέντρωση δεδομένων (data aggregation) και η εξόρυξη δεδομένων (data mining).
- **Διαγνωστική ανάλυση δεδομένων (Diagnostic analytics):** Ο συγκεκριμένος τύπος ανάλυσης διερευνά το «γιατί». Κατά την εκτέλεση διαγνωστικών αναλύσεων, οι αναλυτές δεδομένων επιδιώκουν πρώτα να εντοπίσουν ανωμαλίες μέσα στα δεδομένα, δηλαδή οτιδήποτε δεν μπορεί να εξηγηθεί από τα δεδομένα που έχουν μπροστά τους.
- **Προγνωστική ανάλυση δεδομένων (Predictive analytics):** Αυτός ο τύπος ανάλυσης προσπαθεί να προβλέψει τι είναι πιθανό να συμβεί στο μέλλον. Οι αναλυτές δεδομένων αρχίζουν να καταλήγουν σε πρακτικές, βασισμένες σε δεδομένα πληροφορίες όπως ιστορικά δεδομένα και μέσω αυτών εταιρεία να σχεδιάσει τα επόμενα βήματά της.
- **Κανονιστική ανάλυση δεδομένων (Prescriptive analytics):** Αυτός ο τύπος ανάλυσης συμβουλεύει για τις ενέργειες και τις αποφάσεις που πρέπει να ληφθούν. Με άλλα λόγια, τα πώς μπορείτε να ωφεληθεί κάποιος από τα αποτελέσματα που έχουν προβλεφθεί, όπως για παράδειγμα η συνταγογράφηση ενός ιατρού.



Εικόνα 2. Οι τέσσερις τύποι των Data Analytics

Παράλληλα, ο κλάδος της υγειονομικής περίθαλψης υπήρξε ένας από τους σημαντικότερους ωφελούμενους από την εμφάνιση της επιστήμης δεδομένων. Χάρη στους επιστήμονες δεδομένων, τα ιατρικά διαγνωστικά γίνονται πιο αποτελεσματικά, η

ιατρική θεραπεία πιο εξατομικευμένη και η ιατρική έρευνα βασίζεται περισσότερο στα δεδομένα. Καθιερωμένες φαρμακευτικές εταιρείες και ερευνητικά κέντρα βασίζονται σε έργα επιστήμης δεδομένων για την ανάλυση τεράστιων ποσοτήτων δεδομένων στην επιδίωξη νέων γνώσεων.

Επιπρόσθετα, η υγειονομική περίθαλψη είναι ένας σημαντικός τομέας για την προγνωστική ανάλυση (predictive analytics), καθώς είναι ένα από τα πιο δημοφιλή θέματα στις αναλύσεις υγείας. Ένα μοντέλο πρόβλεψης χρησιμοποιεί ιστορικά δεδομένα, μαθαίνει από αυτά, βρίσκει μοτίβα, δημιουργεί ακριβείς προβλέψεις από αυτό, βρίσκει διάφορους συσχετισμούς, συσχετισμούς συμπτωμάτων, συνήθειες, και ασθένειες και στη συνέχεια κάνει ουσιαστικές προβλέψεις.

Η προγνωστική αναλυτική διαδραματίζει σημαντικό ρόλο στη βελτίωση της φροντίδας των ασθενών, στη διαχείριση χρόνιων ασθενειών και στην αύξηση της αποτελεσματικότητας των αλυσίδων εφοδιασμού και της φαρμακευτικής επιμελητείας. Η διαχείριση της υγείας του πληθυσμού γίνεται όλο και πιο δημοφιλές θέμα στην προγνωστική ανάλυση. Είναι μια προσέγγιση που βασίζεται σε δεδομένα και εστιάζει στην πρόληψη ασθενειών που είναι συνήθως διαδεδομένες στην κοινωνία. Με την επιστήμη των δεδομένων, τα νοσοκομεία μπορούν να προβλέψουν την επιδείνωση της υγείας του ασθενούς και να παρέχουν προληπτικά μέτρα και να ξεκινήσουν έγκαιρη θεραπεία που θα βοηθήσει στη μείωση του κινδύνου περαιτέρω επιδείνωσης της υγείας των ασθενών. Επιπλέον, η προγνωστική αναλυτική διαδραματίζει σημαντικό ρόλο στην παρακολούθηση της εφοδιαστικής προσφοράς νοσοκομείων και φαρμακευτικών τμημάτων.

2.4 Μηχανική μάθηση

Η μηχανική μάθηση (Machine Learning) ή σε συντομογραφία (ML) είναι μια υποκατηγορία της τεχνολογίας της τεχνητής νοημοσύνης (Artificial Intelligence). Αν και η μηχανική μάθηση είναι ένας τομέας στην επιστήμη των υπολογιστών, διαφέρει από τις παραδοσιακές υπολογιστικές προσεγγίσεις. Στην κόσμο της πληροφορικής, ένας αλγόριθμος είναι ένα σύνολο από προγραμματισμένες εντολές που χρησιμοποιούνται για τον υπολογισμό ή την επίλυση προβλημάτων με τον καλύτερο και τον πιο επιθυμητό τρόπο. Αντίθετα, οι αλγόριθμοι μηχανικής μάθησης επιτρέπουν στους υπολογιστές να εκπαιδεύονται όταν εισάγονται δεδομένα και να χρησιμοποιούν στατιστική ανάλυση προκειμένου να εξάγουν τιμές μέσως των οποίων προκύπτουν συμπεράσματα. Εξαιτίας αυτού, η μηχανική μάθηση διευκολύνει τους υπολογιστές στη δημιουργία μοντέλων από δείγματα δεδομένων, προκειμένου να αυτοματοποιηθούν οι διαδικασίες λήψης αποφάσεων με βάση τις εισροές δεδομένων[10].

Δεν μπορεί να θεωρηθεί ότι ένα άτομο εφηύρε τη μηχανική μάθηση, καθώς πολλά άτομα συνέβαλαν στην ανάπτυξή του. Παρόλ’ αυτά, ο Άρθουρ Σάμουελ, επιστήμονας υπολογιστών στην IBM και πρωτοπόρος στην τεχνητή νοημοσύνη και τα παιχνίδια υπολογιστών, επινόησε τον όρο «Μηχανική Μάθηση» το 1952. Τότε ήταν που σχεδίασε ένα πρόγραμμα υπολογιστή για να παίζει πούλια. Όσο περισσότερο έπαιζε το πρόγραμμα το παιχνίδι, τόσο περισσότερα μάθαινε από την εμπειρία του, χάρη σε έναν

αλγόριθμο minimax για τη μελέτη των κινήσεων για να καταλήξει σε στρατηγικές νίκης. Έχουν υπάρξει πολλές πρωτοβουλίες μηχανικής εκμάθησης μέχρι σήμερα, που βοήθησαν την μηχανική μάθηση να εξελιχθεί σε μεγάλο βαθμό από τη δεκαετία του '50. Απογειώθηκε μέχρι τα τέλη της δεκαετίας του 1990 όταν η IBM ανέπτυξε τον υπερυπολογιστή της Deep Blue. Από τότε, πολλοί επιστήμονες και ερευνητές άρχισαν να αναπτύσσουν νέα προγράμματα και αλγόριθμους. Ωστόσο, όλα βασίζονται στους πρώτους αλγόριθμους μηχανικής μάθησης του Άρθουρ Σάμουελ[11].

Τα τελευταία χρόνια, η ευρεία διαθεσιμότητα ισχυρού υλικού και υπολογιστικού νέφους είχε ως αποτέλεσμα την ευρύτερη υιοθέτηση της μηχανικής μάθησης σε διαφορετικούς τομείς της ανθρώπινης ζωής, όπως για παράδειγμα η χρήση του για συστάσεις στα μέσα κοινωνικής δικτύωσης, το digital marketing, η υιοθέτησή του για αυτοματοποίηση διεργασιών στα εργοστάσια, καθώς και σε μεγάλο βαθμό στην αναγνώριση εικόνων και τον τομέα της ιατρικής. Η μηχανική μάθηση, επίσης, είναι σημαντική επειδή δίνει στις επιχειρήσεις μια άποψη για τις τάσεις στη συμπεριφορά των πελατών και τα επιχειρησιακά πρότυπα των επιχειρήσεων, καθώς και υποστηρίζει την ανάπτυξη νέων προϊόντων. Πολλές από τις κορυφαίες εταιρείες του σήμερα, όπως το Facebook, η Google και η Uber, χρησιμοποιούν κατά κόρον τη μηχανική μάθηση στις καθημερινές εργασίες τους. Πιο συγκεκριμένα, το Facebook χρησιμοποιεί μηχανική εκμάθηση για να εξατομικεύσει τον τρόπο με τον οποίο παρέχεται η ροή κάθε μέλους. Εάν ένα μέλος σταματά συχνά για να διαβάζει τις αναρτήσεις μιας συγκεκριμένης ομάδας, η μηχανή προτάσεων θα αρχίσει να εμφανίζει περισσότερη από τη δραστηριότητα αυτής της ομάδας νωρίτερα στη ροή.

Ωστόσο, η χρήση μηχανικής μάθησης σε λειτουργίες υγειονομικής περίθαλψης μπορεί να είναι εξαιρετικά επωφελής για μια εταιρεία. Η μηχανική μάθηση δημιουργήθηκε για να αντιμετωπίζει μεγάλα σύνολα δεδομένων και τα αρχεία χρειάζονται ενδελεχή ανάλυση και οργάνωση. Επιπλέον, ενώ ένας επαγγελματίας υγείας και ένας αλγόριθμος μηχανικής μάθησης πιθανότατα θα καταλήξουν στο ίδιο συμπέρασμα με βάση το ίδιο σύνολο δεδομένων, η χρήση μηχανικής μάθησης θα έχει τα αποτελέσματα πολύ πιο γρήγορα, επιτρέποντας την έναρξη της θεραπείας νωρίτερα.

Ένα άλλο σημείο για τη χρήση τεχνικών μηχανικής μάθησης στην υγειονομική περίθαλψη είναι η εξάλειψη της ανθρώπινης συμμετοχής σε κάποιο βαθμό, γεγονός που μειώνει την πιθανότητα ανθρώπινου λάθους. Αυτό αφορά ιδιαίτερα τις εργασίες αυτοματισμού διαδικασιών, καθώς η κουραστική εργασία ρουτίνας είναι εκεί που οι άνθρωποι κάνουν τα περισσότερα σφάλματα [12]. Η υγειονομική περίθαλψη είναι ένας κλάδος που συμβαδίζει επίσης με την εποχή. Με τον όγκο των δεδομένων που παράγονται για κάθε ασθενή, οι αλγόριθμοι μηχανικής μάθησης στην υγειονομική περίθαλψη έχουν μεγάλες δυνατότητες. Επομένως, δεν είναι περίεργο ότι υπάρχουν πολλές επιτυχημένες εφαρμογές μηχανικής εκμάθησης στον τομέα της υγείας αυτή τη στιγμή.

2.5 Τύποι μηχανικής μάθησης

Η κλασική μηχανική μάθηση κατηγοριοποιείται συχνά από το πώς ένας αλγόριθμος μαθαίνει να γίνεται πιο ακριβής στις προβλέψεις του. Υπάρχουν τέσσερις βασικές προσεγγίσεις: η εποπτευόμενη μάθηση (supervised machine learning), η μάθηση χωρίς επίβλεψη (unsupervised machine learning), η ημι-εποπτευόμενη μάθηση (semi-supervised learning) και η ενισχυτική μάθηση (reinforcement learning) [13]. Οι τύποι αλγορίθμου δεδομένων που επιλέγουν να χρησιμοποιήσουν οι επιστήμονες εξαρτάται από το είδος των δεδομένων που θέλουν να προβλέψουν. Παρακάτω, θα γίνει περιγραφή των τεσσάρων αυτών κατηγοριών.

2.5.1 Εποπτευόμενη μάθηση

Σε αυτόν τον τύπο μηχανικής μάθησης, οι Data Scientists παρέχουν αλγορίθμους με δεδομένα εκπαίδευσης που είναι labeled (τα δεδομένα που είναι ξεκάθαρη η κλάση τους) και ορίζουν τις μεταβλητές που θέλουν να αξιολογήσει ο αλγόριθμος για συσχετίσεις. Καθορίζονται τόσο η είσοδος όσο και η έξοδος του αλγορίθμου. Πιο συγκεκριμένα, απαιτεί την εκπαίδευση του αλγορίθμου τόσο με labeled data όσο και με επιθυμητές εξόδους. Οι αλγόριθμοι εποπτευόμενης μάθησης είναι καλοί για τις ακόλουθες εργασίες:

- **Δυαδική ταξινόμηση**: Διαίρεση δεδομένων σε δύο κατηγορίες.
- **Ταξινόμηση πολλαπλών τάξεων**: Επιλογή μεταξύ περισσότερων από δύο τύπων απαντήσεων.
- **Μοντελοποίηση**: παλινδρόμησης: Πρόβλεψη συνεχών τιμών.
- **Ensembling**: Συνδυασμός των προβλέψεων πολλαπλών μοντέλων μηχανικής μάθησης για την παραγωγή ακριβούς πρόβλεψης.

2.5.2 Μάθηση χωρίς επίβλεψη

Αυτός ο τύπος μηχανικής μάθησης περιλαμβάνει αλγορίθμους που εκπαιδεύονται σε δεδομένα χωρίς ετικέτα (unlabeled). Ο αλγόριθμος σαρώνει μέσα από σύνολα δεδομένων αναζητώντας οποιαδήποτε ουσιαστική σύνδεση. Τα δεδομένα στα οποία εκπαιδεύονται οι αλγόριθμοι, καθώς και οι προβλέψεις ή οι συστάσεις που παράγουν είναι προκαθορισμένα. Ειδικότερα, δεν απαιτούν επισήμανση δεδομένων (δεδομένα χωρίς ετικέτα). Διαπερνούν τα δεδομένα χωρίς ετικέτα για να αναζητήσουν μοτίβα που μπορούν να χρησιμοποιηθούν για την ομαδοποίηση σημείων δεδομένων σε υποσύνολα. Οι περισσότεροι τύποι βαθιάς μηχανικής μάθησης (deep machine learning), συμπεριλαμβανομένων των νευρωνικών δικτύων, είναι αλγόριθμοι χωρίς επίβλεψη. Θεωρούνται αποτελεσματικοί για τις ακόλουθες εργασίες:

- **Ομαδοποίηση**: Διαχωρισμός του συνόλου δεδομένων σε ομάδες με βάση την ομοιότητα. Διασημότερος αλγόριθμος είναι K-means.
- **Ανίχνευση ανωμαλιών**: Προσδιορισμός ασυνήθιστων σημείων δεδομένων σε ένα σύνολο δεδομένων.
- **Εξόρυξη συσχέτισης**: Προσδιορισμός συνόλων στοιχείων σε ένα σύνολο δεδομένων που εμφανίζονται συχνά μαζί.
- **Μείωση διαστάσεων**: Μείωση του αριθμού των μεταβλητών σε ένα σύνολο δεδομένων.

2.5.3 Ημι-εποπτευόμενη μάθηση

Αυτή η προσέγγιση στη μηχανική μάθηση περιλαμβάνει έναν συνδυασμό των δύο προηγούμενων τύπων. Οι Data Scientists μπορεί να τροφοδοτούν έναν αλγόριθμο που φέρει ως επί το πλείστων δεδομένα εκπαίδευσης, αλλά το μοντέλο είναι ελεύθερο να εξερευνήσει τα δεδομένα μόνο του και να αναπτύξει τη δική του κατανόηση του συνόλου δεδομένων. Τροφοδοτεί μια μικρή ποσότητα δεδομένων με ετικέτα εκπαίδευσης σε έναν αλγόριθμο. Από αυτό, ο αλγόριθμος μαθαίνει τις διαστάσεις του συνόλου δεδομένων, τις οποίες μπορεί στη συνέχεια να εφαρμόσει σε νέα, χωρίς ετικέτα δεδομένα. Η απόδοση των αλγορίθμων συνήθως βελτιώνεται όταν εκπαιδεύονται σε σύνολα δεδομένων με ετικέτα. Αλλά η επισήμανση δεδομένων μπορεί να είναι χρονοβόρα και δαπανηρή. Η ημι-εποπτευόμενη μάθηση χτυπά μια μέση λύση μεταξύ της απόδοσης της εποπτευόμενης μάθησης και της αποτελεσματικότητας της μάθησης χωρίς επίβλεψη. Ορισμένοι τομείς όπου χρησιμοποιείται ημι-εποπτευόμενη μάθηση περιλαμβάνουν:

- **Μηχανική μετάφραση**: Διδασκαλία αλγορίθμων για τη μετάφραση γλώσσας που βασίζονται σε λιγότερο από ένα πλήρες λεξικό λέξεων.
- **Ανίχνευση απάτης**: Εντοπισμός περιπτώσεων απάτης όταν έχετε μόνο μερικά θετικά παραδείγματα.
- **Δεδομένα επισήμανσης**: Οι αλγόριθμοι που έχουν εκπαιδευτεί σε μικρά σύνολα δεδομένων μπορούν να μάθουν να εφαρμόζουν αυτόματα ετικέτες δεδομένων σε μεγαλύτερα σύνολα.

2.5.4 Ενισχυτική μάθηση

Οι επιστήμονες δεδομένων χρησιμοποιούν συνήθως την ενισχυτική μάθηση για να διδάξουν σε μια μηχανή πως να ολοκληρώσει μια διαδικασία πολλαπλών βημάτων για την οποία υπάρχουν σαφώς καθορισμένοι κανόνες. Συγκεκριμένα, προγραμματίζουν έναν αλγόριθμο για την ολοκλήρωση μιας εργασίας και του δίνουν θετικά ή αρνητικά στοιχεία καθώς επεξεργάζεται πώς να ολοκληρώσει μια εργασία. Αλλά ως επί το πλείστων, ο αλγόριθμος αποφασίζει μόνος του ποια βήματα θα κάνει στην πορεία. Η ενισχυτική μάθηση λειτουργεί προγραμματίζοντας έναν αλγόριθμο με έναν ξεχωριστό

στόχο και ένα προδιαγεγραμμένο σύνολο κανόνων για την επίτευξη αυτού του στόχου. Οι Data Scientists προγραμματίζουν, επίσης, τον αλγόριθμο να αναζητά θετικές ανταμοιβές, τις οποίες λαμβάνει όταν εκτελεί μια ενέργεια που είναι ευεργετική για τον τελικό στόχο και να αποφεύγει τις τιμωρίες, τις οποίες λαμβάνει όταν εκτελεί μια ενέργεια που τον απομακρύνει από τον τελικό του στόχο στόχος. Η ενισχυτική μάθηση χρησιμοποιείται συχνά σε τομείς όπως:

- **Ρομποτική**: Τα ρομπότ μπορούν να μάθουν να εκτελούν εργασίες στον φυσικό κόσμο χρησιμοποιώντας αυτήν την τεχνική.
- **Βιντεοπαιχνίδια**: Η ενισχυτική μάθηση έχει χρησιμοποιηθεί για να διδάξει τα ρομπότ να παίζουν διάφορα βιντεοπαιχνίδια.
- **Διαχείριση πόρων**: Δεδομένων των πεπερασμένων πόρων και ενός καθορισμένου στόχου, η ενισχυτική μάθηση μπορεί να βοηθήσει τις επιχειρήσεις να σχεδιάσουν τον τρόπο κατανομής των πόρων.

2.6 Αλγόριθμοι και είδη αλγορίθμων

Στη μηχανική μάθηση υπάρχουν πολλοί αλγόριθμοι που έχουν ως στόχο την κατηγοριοποίηση ή ταξινόμηση (classification) των δεδομένων, ενώ άλλοι εξειδικεύονται στην παλινδρόμηση (regression). Στην συγκεκριμένη έρευνα που διεξάγεται στο Κεφάλαιο 3 το πρόβλημα είναι γραμμικό (δυαδικό ή binary) με στόχο την κατηγοριοποίηση, γι’ αυτό και θα πραγματοποιηθεί η περιγραφή των αλγορίθμων κατηγοριοποίησης που χρησιμοποιήθηκαν.

2.6.1 Αλγόριθμος δέντρων απόφασης

Ο αλγόριθμος δέντρων απόφασης (Decision Tree) αποτελεί μια σημαντική μέθοδο ταξινόμησης (Classification), καθώς χρησιμοποιείται σε διάφορες επιστήμες και τομείς. Ανήκει στην οικογένεια των αλγορίθμων εποπτευόμενης μάθησης και έχει το χαρακτηριστικό ότι μπορεί να επιλύσει θέματα που αφορούν προβλήματα ταξινόμησης (ή αλλιώς κατηγοριοποίησης), αλλά και θέματα παλινδρόμησης (Regression). Το ίδιο το όνομα του υποδηλώνει ότι κατά την εφαρμογή του δημιουργεί ένα διάγραμμα ροής το οποίο μοιάζει με δέντρο και μέσω αυτού πραγματοποιεί τις προβλέψεις που προκύπτουν από μια σειρά διαχωρισμών, βασιζόμενες σε δοθείσες συνθήκες [14]. Αποτελείται από τα παρακάτω χαρακτηριστικά:

- Κόμβος ρίζας (Root Node) – Είναι ο κόμβος που θεωρείται «πατέρας» όλων των κόμβων, καθώς υπάρχει στην αρχή κάθε δέντρου απόφασης. Μετά από αυτόν τον κόμβο, ξεκινά ο διαχωρισμός των δεδομένων, σύμφωνα με τα χαρακτηριστικά του εκάστοτε συνόλου δεδομένων.
- Κόμβοι απόφασης (Decision Nodes) – Αποτελούν τους κόμβους που λαμβάνονται μετά τον διαχωρισμό από τον κόμβο ρίζας.
- Κόμβοι φύλλων (Leaf Nodes) – Είναι οι τελικοί κόμβοι, καθώς δεν διαδέχονται περαιτέρω διαχωρισμό.

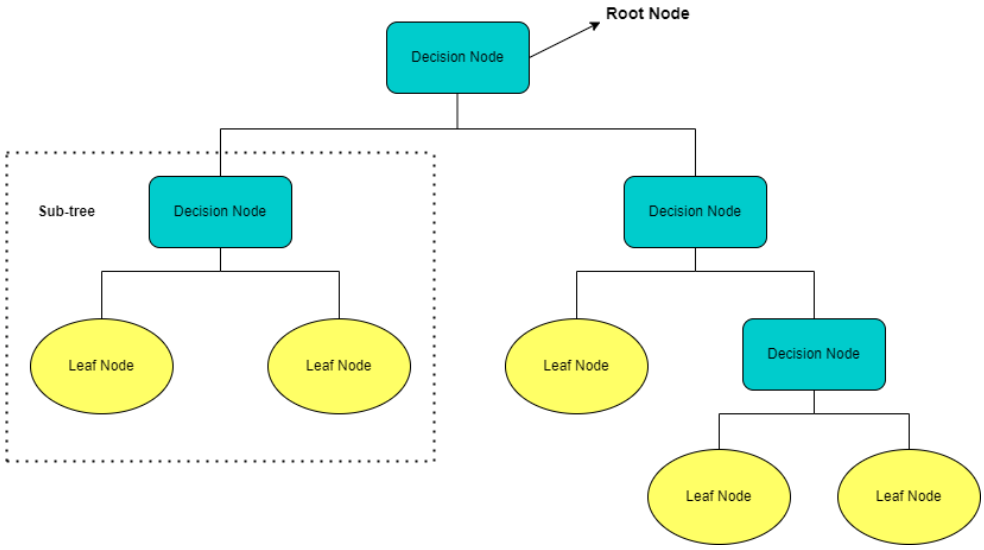
Υπάρχουν και οι περιπτώσεις των υποδέντρων (Sub-trees) που αποτελούν ένα μικρότερο τμήμα ενός δέντρου το οποίο μπορεί και να αναπαρασταθεί και μόνο του. Η λογική του αλγορίθμου είναι πως η διαμέριση των κόμβων του δέντρου πραγματοποιείται με αναδρομικό τρόπο, από πάνω προς τα κάτω και με βάση τα χαρακτηριστικά του συνόλου δεδομένων. Επίσης, αξιοσημείωτο είναι το γεγονός ότι μιμείται σε μεγάλο βαθμό τον τρόπο που σκέφτονται οι άνθρωποι και με αυτόν τον τρόπο επεξεργάζεται τις πληροφορίες και λαμβάνει αποφάσεις. Είναι, άλλωστε, ο λόγος που είναι περισσότερο οικείος στους ανθρώπους που τον μελετούν.

Τα δέντρα αποφάσεων κατασκευάζονται χρησιμοποιώντας μια μέθοδο που στον κόσμο της πληροφορικής και τον μαθηματικών ονομάζεται διαίρει και βασίλευε (Divide and Conquer). Κάθε κόμβος που ακολουθεί τον ριζικό κόμβο χωρίζεται σε πολλούς κόμβους. Τα δέντρα απόφασης διαχωρίζουν το χώρο δεδομένων σε πυκνές και αραιές περιοχές. Ο αλγόριθμος χωρίζει το δέντρο μέχρι τα δεδομένα να είναι αρκετά ομοιογενή. Στο τέλος της εκπαίδευσης, επιστρέφεται ένα δέντρο αποφάσεων που μπορεί να χρησιμοποιηθεί για να γίνουν βέλτιστες κατηγοριοποιημένες προβλέψεις [15]. Παράλληλα, ένας σημαντικός όρος στην ανάπτυξη αυτού του αλγορίθμου είναι η εντροπία. Μπορεί να θεωρηθεί ως το μέτρο της αβεβαιότητας ενός συνόλου δεδομένων και η τιμή του περιγράφει τον βαθμό τυχαιότητας ενός συγκεκριμένου κόμβου. Όσο μεγαλύτερη είναι η εντροπία, τόσο μεγαλύτερη θα είναι η τυχαιότητα στο σύνολο δεδομένων. Κατά την κατασκευή ενός δέντρου αποφάσεων, θα προτιμάται χαμηλότερη εντροπία. Η εντροπία μπορεί περαιτέρω να χρησιμοποιηθεί για τον προσδιορισμό του ριζικού κόμβου του δέντρου αποφάσεων και του αριθμού των διαχωρισμών που πρόκειται να γίνουν. Η έκφραση για τον υπολογισμό της εντροπίας ενός δέντρου απόφασης είναι η παρακάτω:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

Μια άλλη μέτρηση που χρησιμοποιείται για παρόμοιο σκοπό είναι ο δείκτης Gini. Χρησιμοποιεί τη μέθοδο Gini για τη δημιουργία σημείων διαχωρισμού. Το κέρδος πληροφοριών είναι η μέτρηση που χρησιμοποιείται γενικά για τη μέτρηση της αβεβαιότητας στο σύνολο δεδομένων. Το κέρδος πληροφοριών στα δέντρα απόφασης περιγράφεται γενικά από τους τύπους:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \quad (2)$$



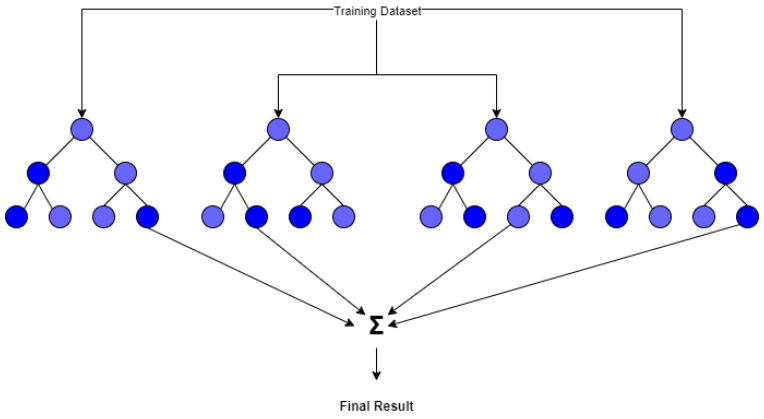
Εικόνα 3. Αλγόριθμος δέντρων απόφασης

2.6.2 Αλγόριθμος τυχαίων δασών

Ο αλγόριθμος των τυχαίων δασών (Random Forest) ανήκει στην κατηγορία των αλγορίθμων εποπτευόμενης μηχανικής μάθησης (Supervised Machine Learning). Ο χαρακτηρισμός του ως δάσος, προσδίδεται στο γεγονός ότι αποτελείται από ένα σύνολο δέντρων απόφασης (Decision Trees), από τα οποία ο συνδυασμός τους πάνω στο σύνολο εκπαίδευσης προσφέρει ένα ακριβέστερο και σταθερό αποτέλεσμα. Όπως και στα δέντρα απόφασης, έτσι κι αυτός ο αλγόριθμος ενδείκνυται για προβλήματα ταξινόμησης (ισχύει κυρίως για κατηγορικές μεταβλητές), όσο και για προβλήματα παλινδρόμησης (ισχύει κυρίως για συνεχείς μεταβλητές). Χαρακτηριστικό του είναι η τυχαιότητα που δίνεται στο μοντέλο, κατά τον πολλαπλασιασμό των δέντρων σε υποσύνολα δεδομένων εκπαίδευσης με αντικατάσταση, γνωστή και ως τεχνική «bagging». Σκοπός του δεν είναι μόνο η εύρεση του σημαντικότερου χαρακτηριστικού κατά την διάσπαση κάθε κόμβου όπως συμβαίνει στον αλγόριθμο των δέντρων απόφασης, αλλά και η εύρεση του καλύτερου χαρακτηριστικού ανάμεσα σε τυχαία υποσύνολα χαρακτηριστικών, γεγονός που οδηγεί σε πιο αξιόπιστα και καλύτερα αποτελέσματα [16]. Συγκεκριμένα, τα βήματα για την εκτέλεση του είναι τα εξης:

- Στα τυχαία δάση (Random Forest) λαμβάνονται n αριθμοί τυχαίων εγγραφών από ένα σύνολο δεδομένων που περιέχει k αριθμό εγγραφών.
- Για κάθε υποσύνολο, κατασκευάζονται μεμονωμένα δέντρα απόφασης
- Κάθε δέντρο απόφασης θα δημιουργήσει ένα αποτέλεσμα.

- Το τελικό αποτέλεσμα εξέρχεται με βάση την ψηφοφορία πλειοψηφίας ανάμεσα στα υποσύνολα των τυχαίων δασών, όσον αφορά την ταξινόμηση και στον μέσο όρο, όσον αφορά την παλινδρόμηση αντίστοιχα.



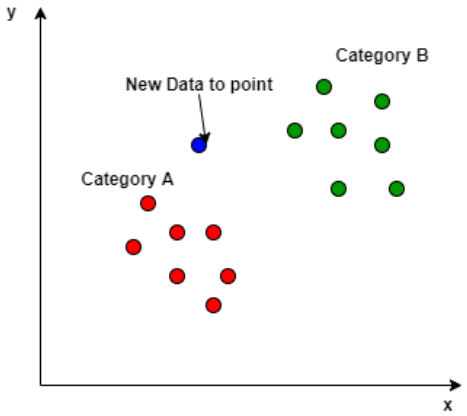
Εικόνα 4. Αλγόριθμος τυχαίων δασών

2.6.3 Αλγόριθμος Κ κοντινότερων γειτόνων (KNN)

Ο αλγόριθμος k κοντινότερων γειτόνων (k nearest neighbours) είναι ένας μη παραμετρικός αλγόριθμος, που σημαίνει η δομή του μοντέλου καθορίζεται από το σύνολο δεδομένων. Είναι, άλλωστε, και ο λόγος για τον οποίο χρησιμοποιείται ευρέως και στην καθημερινή ζωή, καθώς λόγω του παραπάνω δεν βασίζεται σε θεωρητικές μαθηματικές υποθέσεις. Επίσης, ανήκει στους γνωστούς και ως «Lazy» αλγορίθμους, που σημαίνει ότι δεν χρειάζεται εκμάθηση ή εκπαίδευση όλων των δεδομένων που χρησιμοποιούνται κατά τη φάση της πρόβλεψης και όλα τα δεδομένα χρησιμοποιούνται για τη φάση του «testing». Αυτό, έχει ως αποτέλεσμα τα «training» δεδομένα να εκπαιδεύονται πιο γρήγορα και η πρόβλεψη να είναι πιο αργή και πιο δαπανηρή, άρα δημιουργείται ανάγκη για περισσότερο χρόνο και μνήμη. Στην συγκεκριμένη ανάγκη, η χειρότερη περίπτωση παρατηρείται όταν ο KNN χρειάζεται περισσότερο χρόνο για να σαρώσει όλα τα δεδομένα και η σάρωση τους θα απαιτήσει περισσότερη μνήμη για την αποθήκευση δεδομένων εκπαίδευσης [17].

Στον αλγόριθμο KNN, το K είναι ο αριθμός των πλησιέστερων γειτόνων. Ο αριθμός των γειτόνων είναι ο βασικός αποφασιστικός παράγοντας. Το K είναι γενικά ένας περιττός αριθμός εάν ο αριθμός των κλάσεων είναι 2, όταν δηλαδή μιλάμε για ένα δυαδικό πρόβλημα ταξινόμησης. Στην περίπτωση μικρού αριθμού γειτόνων, ο θόρυβος θα έχει μεγαλύτερη επίδραση στο αποτέλεσμα και ένας μεγάλος αριθμός γειτόνων τον καθιστά υπολογιστικά ακριβό, οπότε είναι πολύ απαραίτητη η ακριβής επιλογή των γειτόνων που θα επιλεχτούν για την εφαρμογή του αλγορίθμου. Επίσης, ο KNN αποδίδει καλύτερα με μικρότερο αριθμό χαρακτηριστικών. Όταν ο αριθμός των χαρακτηριστικών αυξάνεται, απαιτεί περισσότερα δεδομένα και πόρους. Η αύξηση της διάστασης οδηγεί επίσης στο

πρόβλημα του «overfitting». Για την αποφυγής του, τα απαραίτητα δεδομένα θα πρέπει να αυξάνονται εκθετικά καθώς αυξάνεται ο αριθμός των διαστάσεων.



Εικόνα 5. Αλγόριθμος KNN

2.6.4 Αλγόριθμος κατά Μπέιζ (Gaussian)

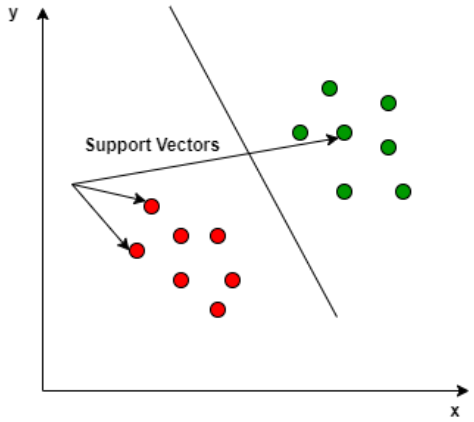
Ο μπεϊζιανός αλγόριθμος (Naive Bayes) είναι ένα βασικό και αποτελεσματικό μοντέλο ταξινόμησης στη μηχανική μάθηση που αντλεί επιρροή από το θεώρημα Bayes και χρησιμοποιείται για προβλήματα ταξινόμησης δυαδικών (binary) και πολλαπλών κλάσεων. Το θεώρημα Bayes είναι ένας τύπος που προσφέρει μια υπό όρους πιθανότητα να συμβεί ένα γεγονός A, δεδομένου ότι ένα άλλο γεγονός B έχει συμβεί προηγουμένως [18]. Ο μαθηματικός τύπος του είναι ο εξής: $P(H/E) = P(E/H) * P(H)/P(E)$, όπου:

- Το H και το E είναι δύο γεγονότα
- Το P(H|E) είναι η πιθανότητα του γεγονότος A με την προϋπόθεση ότι το γεγονός B έχει ήδη συμβεί.
- P(E|H) είναι η πιθανότητα του γεγονότος B, εφόσον το γεγονός A έχει ήδη συμβεί.
- Το P(E) είναι η ανεξάρτητη πιθανότητα του E
- Το P(H) είναι η ανεξάρτητη πιθανότητα του H

Η μέθοδος Naive Bayes κάνει την υπόθεση ότι οι προγνωστικοί παράγοντες συμβάλλουν εξίσου και ανεξάρτητα στην επιλογή της κλάσης εξόδου, του αποτελέσματος δηλαδή του μοντέλου. Αν και η υπόθεση του μοντέλου Naive Bayes ότι όλοι οι προγνωστικοί παράγοντες είναι ανεξάρτητοι ο ένας από τον άλλο είναι ανέφικτη σε πραγματικές συνθήκες, αυτή η υπόθεση παράγει ένα ικανοποιητικό αποτέλεσμα στην πλειονότητα των περιπτώσεων. Παράλληλα, χρησιμοποιείται συχνά για την κατηγοριοποίηση κειμένου, καθώς η διάσταση των δεδομένων είναι συχνά αρκετά μεγάλη. Στα πλαίσια της παρούσας έρευνας θα χρησιμοποιηθεί ο GaussianNB, ο οποίος χρησιμοποιείται όταν οι τιμές πρόβλεψης είναι συνεχείς και αναμένεται να ακολουθήσουν μια κατανομή Gauss. Μπορούν να χρησιμοποιηθούν άλλες συναρτήσεις για την εκτίμηση της κατανομής των δεδομένων, αλλά η Gaussian (ή η Κανονική κατανομή) είναι η πιο εύκολη στην κατανόηση, επειδή χρειάζεται μόνο να υπολογιστεί η μέση τιμή και η τυπική απόκλιση από τα δεδομένα εκπαίδευσης.

2.6.5 Αλγόριθμος μηχανών διανυσμάτων υποστήριξης

Ο αλγόριθμος Support Vector Machines έχει ως στόχο την να αναζητήσει ένα υπερεπίπεδο σε ένα χώρο N-διάστασης, όπου N ο αριθμός των χαρακτηριστικών που ταξινομεί ευδιάκριτα τα σημεία δεδομένων. Για να διαχωριστούν οι δύο κατηγορίες σημείων δεδομένων, υπάρχουν πολλά πιθανά υπερεπίπεδα που θα μπορούσαν να επιλεγούν. Στόχος του αλγόριθμου είναι να βρεθεί ένα επίπεδο που να έχει το μέγιστο περιθώριο, δηλαδή τη μέγιστη απόσταση μεταξύ των δεδομένων και των δύο κατηγοριών του SVM. Η μεγιστοποίηση της απόστασης περιθωρίου παρέχει κάποια ενίσχυση, έτσι ώστε τα μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη σιγουριά και να μην υπάρχει ο φόβος να «χαθούν» σημεία. Τα υπερεπίπεδα είναι όρια απόφασης που ορίζονται από τον αλγόριθμό και που βοηθούν στην ταξινόμηση των δεδομένων με σωστό τρόπο. Τα δεδομένα που εμπίπτουν σε κάθε πλευρά του υπερεπίπεδου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Επίσης, η διάσταση του υπερεπίπεδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 2, τότε το υπερεπίπεδο είναι απλώς μια γραμμή. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 3, τότε το υπερεπίπεδο γίνεται δισδιάστατο επίπεδο. Τα διανύσματα του αλγορίθμου SVM είναι σημεία δεδομένων που βρίσκονται πιο κοντά στο υπερεπίπεδο και επηρεάζουν τη θέση και τον προσανατολισμό του υπερεπίπεδου. Χρησιμοποιώντας αυτά τα διανύσματα υποστήριξης, μεγιστοποιείται το περιθώριο του ταξινομητή. Σκοπός του SVM είναι η δημιουργία ενός όσο το δυνατόν καλύτερου δρόμου, ο οποίος να περικλείεται από τις δύο γραμμές. Η απόσταση ανάμεσα από τις διακεκομμένες γραμμές που περιλαμβάνει τα δεδομένα ονομάζεται περιθώριο (margin) και η διαδικασία αυτή ονομάζεται maximal margin classification. Επιπλέον, τα σημεία που βρίσκονται πάνω στις διακεκομμένες γραμμές ονομάζονται support vectors, καθώς έχουν το ρόλο του υποστηρικτή για το μέγιστο περιθώριο του υπερεπίπεδου [19].



Εικόνα 6. Αλγόριθμος SVM

2.6.6 Αλγόριθμος λογιστικής παλινδρόμησης

Ο αλγόριθμος λογιστικής παλινδρόμησης είναι μια τεχνική ταξινόμησης που δανείστηκε η μηχανική μάθηση από τον τομέα της στατιστικής. Αποτελεί μια στατιστική μέθοδο για την ανάλυση ενός συνόλου δεδομένων στο οποίο υπάρχουν μία ή περισσότερες ανεξάρτητες μεταβλητές που καθορίζουν ένα αποτέλεσμα. Σκοπός του είναι να βρεθεί το καλύτερο μοντέλο προσαρμογής για να περιγράψει τη σχέση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής. Χρησιμοποιείται κυρίως σε προβλέψεις δυαδικών προβλημάτων ταξινόμησης αλλά μπορεί να επεκταθεί και να ταξινομηθεί περαιτέρω σε τρεις διαφορετικούς τύπους που αναφέρονται παρακάτω:

- Binomial: Όπου όπως αναφέρθηκε η μεταβλητή στόχος (label) μπορεί να έχει μόνο δύο πιθανούς τύπους.
- Multinomial: Όπου η μεταβλητή στόχος (label) έχει τρεις ή περισσότερους πιθανούς τύπους, οι οποίοι μπορεί να μην έχουν καμία ποσοτική σημασία.
- Ordinal: Όπου οι μεταβλητές-στόχοι έχουν ταξινομημένες κατηγορίες.

Στην λογιστική παλινδρόμηση για να χαρτογραφηθούν οι προβλεπόμενες τιμές σε πιθανότητες, χρησιμοποιείται σιγμοειδής συνάρτηση. Αυτή η συνάρτηση αντιστοιχίζει οποιαδήποτε πραγματική τιμή σε μια άλλη τιμή μεταξύ 0 και 1. Αυτή η συνάρτηση έχει μια μη αρνητική παράγωγο σε κάθε σημείο και ακριβώς ένα σημείο καμπής.

Επίσης, η συνάρτηση κόστους που αφορά τον αλγόριθμο Logistic Regression είναι ένας μαθηματικός τύπος που χρησιμοποιείται για να ποσοτικοποιήσει το σφάλμα μεταξύ των προβλεπόμενων τιμών και των αναμενόμενων τιμών. Με απλά λόγια, μια συνάρτηση κόστους είναι ένα μέτρο του πόσο λάθος είναι το μοντέλο όσον αφορά την ικανότητά του να εκτιμά τη σχέση μεταξύ x και y. Η τιμή που επιστρέφεται από τη συνάρτηση κόστους αναφέρεται ως κόστος ή απώλεια ή απλά, σφάλμα [20] . Για την λογιστική παλινδρόμηση, η συνάρτηση κόστους δίνεται από την εξίσωση:

$$Cost(h\theta(X), Y(actual) = -\log\big(h\theta(X)\big)\, if\ y = 1$$

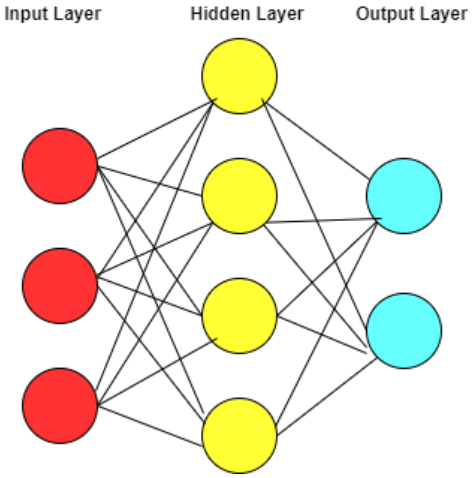
$$or - \log\,(1 - h\theta(x)\ if\ y = 0 \quad (3)$$

2.6.7 Αλγόριθμος τεχνητών νευρωνικών δικτύων

Τα τεχνητά νευρωνικά δίκτυα είναι μοντέλο ταξινόμησης το οποίο λειτουργεί όπως η δομή του ανθρώπινου εγκεφάλου. Δεν είναι απαραίτητα μια ακριβής αντιγραφή του εγκεφάλου, καθώς υπάρχουν πολλά που δεν είναι ακόμη γνωστά για τον εγκέφαλο και τον τρόπο λειτουργίας του, αλλά έχει χρησιμεύσει ως έμπνευση σε πολλούς επιστημονικούς τομείς λόγω της ικανότητάς του να αναπτύσσει νοημοσύνη. Αν και σήμερα το Perceptron αναγνωρίζεται ευρέως ως αλγόριθμος, αρχικά προοριζόταν ως μηχανή αναγνώρισης εικόνας. Πήρε το όνομά του από την εκτέλεση της ανθρώπινης λειτουργίας της αντίληψης , της θέασης και της αναγνώρισης εικόνων.

Το Multilayer Perceptron είναι ένα νευρωνικό δίκτυο όπου η αντιστοίχιση μεταξύ εισόδων και εξόδων είναι μη γραμμική. Ένα Multilayer Perceptron μπορεί να χρησιμοποιήσει οποιαδήποτε αυθαίρετη συνάρτηση ενεργοποίησης. Το Multilayer Perceptron εμπίπτει

στην κατηγορία των αλγορίθμων «feedforward» , επειδή οι είσοδοι συνδυάζονται με τα αρχικά βάρη σε ένα σταθμισμένο άθροισμα και υπόκεινται στη συνάρτηση ενεργοποίησης. Κάθε επίπεδο τροφοδοτεί το επόμενο με το αποτέλεσμα του υπολογισμού του, την εσωτερική του αναπαράσταση των δεδομένων. Αυτό περνάει από τα κρυφά επίπεδα μέχρι το επίπεδο εξόδου. Εάν ο αλγόριθμος υπολόγιζε μόνο τα σταθμισμένα αθροίσματα σε κάθε νευρώνα, διέδιδε τα αποτελέσματα στο επίπεδο εξόδου και σταματούσε εκεί, δεν θα μπορούσε να μάθει τα βάρη που ελαχιστοποιούν τη συνάρτηση κόστους. Επίσης, αν υπολόγιζε μόνο μία επανάληψη, δεν θα υπήρχε πραγματική μάθηση. Η backpropagation είναι ο μηχανισμός εκμάθησης που επιτρέπει στο Multilayer Perceptron να προσαρμόζει επαναληπτικά τα βάρη στο δίκτυο, με στόχο την ελαχιστοποίηση της συνάρτησης κόστους. Η συνάρτηση που συνδυάζει εισόδους και βάρη σε έναν νευρώνα, για παράδειγμα το σταθμισμένο άθροισμα, και η συνάρτηση κατωφλίου, για παράδειγμα η ReLU, πρέπει να είναι αυστηρά διαφοροποιήσιμες. Αυτές οι συναρτήσεις πρέπει να έχουν μια οριοθετημένη παράγωγο. Σε κάθε επανάληψη, αφού τα σταθμισμένα αθροίσματα προωθηθούν σε όλα τα επίπεδα, η κλίση του μέσου τετραγώνου σφάλματος υπολογίζεται σε όλα τα ζεύγη εισόδου και εξόδου. Στη συνέχεια, για να το διαδοθεί ξανά, τα βάρη του πρώτου κρυφού στρώματος ενημερώνονται με την τιμή της κλίσης. Έτσι, διαδίδονται τα βάρη πίσω στην αφετηρία του νευρωνικού δικτύου. Αυτή η διαδικασία συνεχίζεται έως ότου συγκλίνει η κλίση για κάθε ζεύγος εισόδου-εξόδου, που σημαίνει ότι η νέα κλίση που υπολογίστηκε δεν έχει αλλάξει περισσότερο από ένα καθορισμένο όριο σύγκλισης, σε σύγκριση με την προηγούμενη επανάληψη [21].



Εικόνα 7. Αλγόριθμος τεχνητών νευρωνικών δικτύων

2.6.8 Αλγόριθμος Gradient Boosting

Ο αλγόριθμος Gradient Boosting είναι ένας από τους πιο ισχυρούς αλγόριθμους στον τομέα της μηχανικής μάθησης. Σκοπός του είναι να μειώσει όσο δυνατόν περισσότερο το σφάλμα πρόβλεψης. Τα σφάλματα στους αλγόριθμους μηχανικής μάθησης ταξινομούνται σε δύο κατηγορίες, σε σφάλματα μεροληψίας και σε σφάλματα διακύμανσης. Παρόλο που ο Gradient Boosting είναι ένας από τους αλγόριθμους ενίσχυσης, χρησιμοποιείται για την ελαχιστοποίηση του σφάλματος μεροληψίας του μοντέλου. Ο βασικός εκτιμητής για τον αλγόριθμο Gradient Boost είναι σταθερός και μπορεί να χρησιμοποιηθεί για την πρόβλεψη όχι μόνο της μεταβλητής συνεχούς στόχου (ως Regressor) αλλά και για την κατηγορική μεταβλητή στόχου (ως ταξινομητή). Όταν χρησιμοποιείται για πρόβλεψη

θεμάτων παλινδρόμησης, η συνάρτηση κόστους είναι το μέσο τετραγωνικό σφάλμα (MSE), ενώ όταν χρησιμοποιείται ως ταξινομητής, τότε η συνάρτηση κόστους είναι απώλεια καταγραφής (Log loss). Η δομή του περιλαμβάνει τρία στοιχεία: μια συνάρτηση απώλειας που πρέπει να βελτιστοποιηθεί, ένας αδύναμο learner με την ιδιότητα της πρόβλεψης και ένα μοντέλο για την προσθήκη αδύναμων learners για την ελαχιστοποίηση της συνάρτησης απώλειας [22].

2.6.9 Αλγόριθμος AdaBoost

Ο αλγόριθμος AdaBoost είναι μια πολύ δημοφιλής τεχνική ενίσχυσης που στοχεύει στο συνδυασμό πολλών αδύναμων ταξινομητών για τη δημιουργία ενός ισχυρού ταξινομητή. Ο AdaBoost συντάχθηκε, αρχικά, από τους Yoav Freund και Robert Schapire. Είναι ένας μεμονωμένος ταξινομητής ο οποίος δεν έχει την ικανότητα να προβλέπει με ακρίβεια την κλάση ενός αντικειμένου, όμως όταν ομαδοποιεί πολλούς αδύναμους ταξινομητές με τον καθένα να μαθαίνει σταδιακά από τα εσφαλμένα ταξινομημένα αντικείμενα των άλλων προηγούμενων στη σειρά ταξινομητών, τότε δημιουργείται ένα πολύ ισχυρό μοντέλο. Μπορεί να χρησιμοποιήσει οποιονδήποτε ταξινομητή επιθυμεί, όμως κατά κύριο λόγο χρησιμοποιεί τα δέντρα απόφασης (Decision Trees). Ένας αδύναμος ταξινομητής είναι αυτός που αποδίδει καλύτερα από την τυχαία εικασία, αλλά εξακολουθεί να έχει κακή απόδοση στον προσδιορισμό κλάσεων σε αντικείμενα. Ο AdaBoost μπορεί να εφαρμοστεί πάνω από οποιονδήποτε ταξινομητή για να μάθει από τα μειονεκτήματά του και να προτείνει ένα πιο ακριβές μοντέλο. Συγκεκριμένα, ο αλγόριθμος ακολουθεί τα παρακάτω βήματα: Αρχικά, Ένας αδύναμος ταξινομητής δημιουργείται πάνω από τα δεδομένα εκπαίδευσης με βάση τα σταθμισμένα δείγματα. Τα βάρη των δειγμάτων είναι πολύ σημαντικά για να πραγματοποιηθεί σωστή ταξινόμηση. Έπειτα, δημιουργείται ένα δέντρο απόφασης για κάθε μεταβλητή και διαφαίνεται το πόσο καλά ταξινομούνται τα δεδομένα στην κλάση. Επίσης, αποδίδεται μεγαλύτερο βάρος στα λανθασμένα ταξινομημένα δείγματα, ώστε να ταξινομηθούν σωστά στον επόμενο ταξινομητή που θα ακολουθήσει. Το βάρος αποδίδεται, επίσης, σε κάθε ταξινομητή με βάση την ακρίβεια του ταξινομητή. Όσο μεγαλύτερη η ακρίβεια, τόσο μεγαλύτερο το βάρος στις μεταβλητές. Ο αλγόριθμος επαναλαμβάνεται μέχρι να ταξινομηθούν σωστά όλα τα δεδομένα ή να επιτευχθεί το μέγιστο επίπεδο επανάληψης, καθώς αυτό μπορεί να οριστεί από το χρήστη [23].

2.6.10 Αλγόριθμος Extreme Gradient Boosting

Ο αλγόριθμος XGBoost είναι η σύντομη μορφή της λέξης Extreme Gradient Boosting. Είναι μια παραλληλισμένη και προσεκτικά βελτιστοποιημένη έκδοση του αλγόριθμου Gradient Boosting. Ο παραλληλισμός της όλης διαδικασίας, βελτιώνει σημαντικά τον χρόνο εκπαίδευσης του αλγορίθμου. Αντί να εκπαιδεύσουμε το καλύτερο δυνατό μοντέλο στα δεδομένα (όπως στις παραδοσιακές μεθόδους), εκπαιδεύουμε χιλιάδες μοντέλα σε διάφορα υποσύνολα του συνόλου δεδομένων εκπαίδευσης και στη συνέχεια ψηφίζεται το μοντέλο με την καλύτερη απόδοση. Για πολλές περιπτώσεις, ο XGBoost είναι καλύτερος από τους συνηθισμένους αλγόριθμους Gradient Boosting. Μερικά σημαντικά χαρακτηριστικά του XGBoost είναι:

Παραλληλισμός: Το μοντέλο υλοποιείται για εκπαίδευση με πολλαπλούς πυρήνες CPU.

Τακτοποίηση: Το XGBoost περιλαμβάνει διαφορετικές ποινές τακτοποίησης για την αποφυγή overfitting. Οι ομαλοποιήσεις των ποινών παράγουν επιτυχημένη εκπαίδευση, ώστε το μοντέλο να μπορεί να γενικευτεί επαρκώς.

Μη γραμμικότητα: Το XGBoost μπορεί να εντοπίσει και να μάθει από μη γραμμικά μοτίβα δεδομένων.

Cross-validation: Δεν χρειάζεται απαραίτητα τεχνική Cross Validation για την πρόβλεψη της ακρίβειας, καθώς το μοντέλο παρέχει από μόνο του.

Επεκτασιμότητα: Το XGBoost μπορεί να εκτελείται κατανεμημένο χάρη σε κατανεμημένους διακομιστές και συμπλέγματα όπως το Hadoop και το Spark, ώστε να μπορείτε να επεξεργάζεστε τεράστιες ποσότητες δεδομένων. Είναι επίσης διαθέσιμο για πολλές γλώσσες προγραμματισμού όπως C++, JAVA, Python και Julia.

Η παρακάτω συνάρτηση θα χρησιμεύσει ως μέτρο σφάλματος για να μειωθεί η απώλεια και να διατηρηθεί η απόδοση με την πάροδο του χρόνου. Η ακολουθία συγκλίνει στο ελάχιστο της συνάρτησης. Η συγκεκριμένη σημείωση ορίζει τη συνάρτηση σφάλματος που εφαρμόζεται κατά την αξιολόγηση ενός Gradient Boost Regressor [24].

$$f(x,\theta) = \sum l(F(X_i,\theta), y_i)) \quad (4)$$

2.6.11 Αλγόριθμος Light Gradient Descent

Ο αλγόριθμος LGMBoost αναφέρεται κι αυτός στην κατηγορία αλγορίθμων μηχανικής μάθησης συνόλου που μπορούν να χρησιμοποιηθούν για προβλήματα μοντελοποίησης πρόβλεψης ταξινόμησης ή παλινδρόμησης. Τα σύνολα κατασκευάζονται από μοντέλα δέντρων αποφάσεων, όπως ακριβώς ο Gradient Boosting. Τα δέντρα προστίθενται ένα κάθε φορά στο σύνολο και στοχεύουν την διόρθωση των σφαλμάτων πρόβλεψης που έγιναν από προηγούμενα μοντέλα.

Τα μοντέλα προσαρμόζονται χρησιμοποιώντας οποιαδήποτε αυθαίρετη συνάρτηση διαφοροποιήσιμης απώλειας και αλγόριθμο βελτιστοποίησης gradient descent. Αυτό δίνει στην τεχνική το όνομά της, « ενίσχυση κλίσης », καθώς η κλίση απώλειας ελαχιστοποιείται, καθώς το μοντέλο είναι κατάλληλο, σαν ένα νευρωνικό δίκτυο. Ο LGMBoost έχει σχεδιαστεί για να είναι αποτελεσματικός και ίσως πιο αποτελεσματικός από άλλες υλοποιήσεις αλγορίθμων boosting. Ο LGMBoost περιγράφηκε από τους Guolin Ke, et al. στην εργασία του 2017 με τίτλο " LightGBM: A Highly Efficient Gradient Boosting Decision Tree." Η υλοποίηση εισάγει δύο βασικές ιδέες: GOSS και EFB. Η δειγματοληψία μίας όψης βάσει διαβάθμισης, ή εν συντομία GOSS, είναι μια τροποποίηση στη μέθοδο ενίσχυσης διαβάθμισης που εστιάζει την προσοχή σε εκείνα τα παραδείγματα εκπαίδευσης που οδηγούν σε μεγαλύτερη διαβάθμιση, επιταχύνοντας με τη σειρά της την εκμάθηση και μειώνοντας την υπολογιστική πολυπλοκότητα της μεθόδου. Με το GOSS, αποκλείεται ένα σημαντικό ποσοστό παρουσιών δεδομένων με μικρές διαβαθμίσεις και χρησιμοποιούμε μόνο τα υπόλοιπα για να εκτιμήσουμε το κέρδος πληροφοριών. Αποδεικνύεται ότι, δεδομένου ότι τα στιγμιότυπα δεδομένων με μεγαλύτερες διαβαθμίσεις παίζουν πιο σημαντικό ρόλο στον υπολογισμό του κέρδους πληροφοριών, το GOSS μπορεί να λάβει αρκετά ακριβή εκτίμηση του κέρδους

πληροφοριών με πολύ μικρότερο μέγεθος δεδομένων. Η αποκλειστική ομαδοποίηση λειτουργιών, ή EFB για συντομία, είναι μια προσέγγιση για τη ομαδοποίηση αραιών (κυρίως μηδενικών) αμοιβαία αποκλειστικών χαρακτηριστικών, όπως κατηγορικές μεταβλητές εισόδους που έχουν κωδικοποιηθεί μία φορά. Ως εκ τούτου, είναι ένας τύπος αυτόματης επιλογής χαρακτηριστικών. Μαζί, αυτές οι δύο αλλαγές μπορούν να επιταχύνουν τον χρόνο εκπαίδευσης του αλγορίθμου έως και 20x. Ως εκ τούτου, ο LGMBoost μπορεί να θεωρηθεί μια παραλλαγή του αλγορίθμου δέντρων απόφασης ενίσχυσης κλίσης (GBDT) με την προσθήκη GOSS και EFB [25].

2.6.12 Αλγόριθμος Stochastic Gradient Descent

Ο αλγόριθμος Stochastic Gradient Descent (SGD) υπολογίζει τη διαβάθμιση χρησιμοποιώντας ένα μόνο δείγμα. Η ντεγκραντέ κάθοδος mini-batch παίρνει τελικά το καλύτερο και των δύο κόσμων και εκτελεί μια ενημέρωση για κάθε μίνι-παρτίδα n παραδειγμάτων εκπαίδευσης. Όπως μπορεί να διαπιστωθεί το SGD επιτρέπει τη mini-batch (διαδικτυακή/εκτός πυρήνα) μάθηση. Επομένως, είναι λογικό να χρησιμοποιείτε το SGD για προβλήματα μεγάλης κλίμακας όπου θεωρείται πολύ αποτελεσματικό. Το ελάχιστο της συνάρτησης κόστους της Logistic Regression δεν μπορεί να υπολογιστεί άμεσα, επομένως ελαχιστοποιείται μέσω Stochastic Gradient Descent, γνωστό και ως Online Gradient Descent. Σε αυτή τη διαδικασία κατεβαίνει κατά μήκος της συνάρτησης κόστους προς το ελάχιστο της (παρακαλούμε ρίξτε μια ματιά στο παραπάνω διάγραμμα) για κάθε παρατήρηση εκπαίδευσης. Ένας άλλος λόγος για να χρησιμοποιηθεί ο ταξινομητής SGD είναι ότι το SVM ή η λογιστική παλινδρόμηση δεν θα λειτουργήσουν εάν δεν μπορεί να διατηρηθεί η εγγραφή στη μνήμη RAM. Ωστόσο, ο SGD Classifier συνεχίζει να λειτουργεί [26].

2.7 Τρόποι εφαρμογής των μοντέλων σε δεδομένα

Για την εφαρμογή των παραπάνω μοντέλων-αλγορίθμων σε οποιαδήποτε δεδομένα χρειάζεται κάποια επεξεργασία, έτσι ώστε να μπορούν να διαχωριστούν και οι αλγόριθμοι να αποδώσουν. Στην μηχανική μάθηση οι πιο γνωστοί τρόποι είναι με χρήση train_test_split που είναι μέθοδος της βιβλιοθήκης sklearn, και η χρήση Cross-Validation [27].

2.7.1 Train Test Split

Η χρήση της μεθόδου train_test_split είναι μια διαδικασία επικύρωσης ενός μοντέλου που επιτρέπει την προσομοίωση για την καλύτερη απόδοση ενός μοντέλου πάνω σε ένα σύνολο δεδομένων. Συγκεκριμένα, τα δεδομένα πρέπει να είναι διατεταγμένα με το σωστό τρόπο, ώστε να είναι αποδεκτά. Συνήθως αυτός ο τρόπος είναι ο διαχωρισμός των δεδομένων σε Features και Target. Αυτό επιτυγχάνεται με χρήση ξεχωριστών μεταβλητών. Έπειτα, τα δεδομένα διαχωρίζονται σε train και test. Το ποσοστό του διαχωρίσμου για το test set είναι στην ευχέρια του χρήστη και μπορεί να χρειαστούν διάφορα πειράματα για να επιτευχθεί το καλύτερο αποτέλεσμα. Έτσι, λοιπόν, αν οι μεταβλητές έχουν οριστεί ως X (Features) και Y (Target), τότε στην εφαρμογή train και test θα οριστούν νέες μεταβλητές που θα ονομαστούν για παράδειγμα X_train, y_train,

X_test, Y_test. Έτσι, τα δεδομένα είναι έτοιμα για αξιολόγηση από κάποιο προβλεπτικό μοντέλο.

2.7.2 Cross Validation

Η μέθοδος Cross Validation είναι εξίσου μια στατιστική μέθοδος που χρησιμοποιείται για την εκτίμηση και την απόδοση των μοντέλων μηχανικής μάθησης. Χρησιμοποιείται συνήθως στην μηχανική μάθηση για τη σύγκριση και την επιλογή του κατάλληλου μοντέλου πρόβλεψης πάνω σε ένα σύνολο δεδομένων. Η διαδικασία έχει μια μοναδική παράμετρο που ονομάζεται k και αναφέρεται στον αριθμό των ομάδων στις οποίες πρόκειται να χωριστεί ένα δείγμα δεδομένων. Συχνά αναφέρεται με την ονομασία k-fold cross-validation. Όταν επιλέγεται μια συγκεκριμένη τιμή για το k, μπορεί να χρησιμοποιηθεί στη θέση του k στην αναφορά στο μοντέλο, όπως το k=10, που σημαίνει ότι για την πραγματοποίηση της αξιολόγησης από ένα μοντέλο, το σύνολο δεδομένων έχει χωριστεί 10 φορές cross validation. Η τιμή k πρέπει να επιλεγεί προσεκτικά και μπορεί να χρειαστούν πολλαπλά πειράματα για την επιλογή του κατάλληλου αριθμού που να πετυχαίνει την καλύτερη απόδοση στα προβλεπτικά μοντέλα. Στο τέλος χρησιμοποιούνται τα αποτελέσματα για να βγει ένας μέσος όρος ακρίβειας για το κάθε μοντέλο, σύμφωνα πάντα με τον αριθμό k που έχει αρχικοποιηθεί.

2.8 Μετρικές αξιολόγησης

Η επιλογή της σωστής μέτρησης είναι ζωτικής σημασίας κατά την αξιολόγηση μοντέλων μηχανικής μάθησης (ML). Διάφορες μετρήσεις προτείνονται για την αξιολόγηση μοντέλων ML σε διαφορετικές εφαρμογές και μπορεί να είναι χρήσιμη μια σύνοψη δημοφιλών μετρήσεων για καλύτερη κατανόηση κάθε μέτρησης και των εφαρμογών για τις οποίες μπορούν να χρησιμοποιηθούν. Σε ορισμένες εφαρμογές, η εξέταση μιας μεμονωμένης μέτρησης μπορεί να μην δώσει την πλήρη εικόνα του προβλήματος που πρόκειται να επιλυθεί, γι’ αυτό πολλές φορές μπορεί να είναι δόκιμη η χρήση τους σε υποσύνολα του συνόλο δεδομένων.

2.8.1 Confusion Matrix

Μία από τις βασικές έννοιες στην απόδοση ταξινόμησης είναι ο confusion matrix [29], ο οποίος είναι μια απεικόνιση σε πίνακα των προβλέψεων του μοντέλου, έναντι των συνολικών τιμών του στόχου πρόβλεψης. Κάθε γραμμή του confusion matrix αντιπροσωπεύει τα στιγμιότυπα σε στόχο πρόβλεψης και κάθε στήλη αντιπροσωπεύει τα στιγμιότυπα σε μια πραγματική κλάση. Συγκεκριμένα, ο confusion matrix περιλαμβάνει τις παρακάτω τιμές:

TN: True Negative υποδεικνύει τον αριθμό των αρνητικών παραδειγμάτων που ταξινομήθηκαν σωστά.

TP: True Positive υποδεικνύει τον αριθμό των θετικών παραδειγμάτων που ταξινομήθηκαν σωστά.

FP: False Positive υποδεικνύει τον αριθμό των πραγματικών αρνητικών παραδειγμάτων που ταξινομήθηκαν λανθασμένα.

FN: False Negative υποδεικνύει τον αριθμό των πραγματικών αρνητικών παραδειγμάτων που ταξινομήθηκαν λανθασμένα.

2.8.2 Ακρίβεια κατηγοριοποίησης

Η ακρίβεια ταξινόμησης [29] είναι ίσως η απλούστερη μέτρηση όσον αφορά και τον υπολογισμό της, καθώς ορίζεται ως ο αριθμός των σωστών προβλέψεων διαιρεμένος με τον συνολικό αριθμό των προβλέψεων, πολλαπλασιαζόμενος επί 100. Έτσι, αν για παράδειγμα σε ένα σύνολο δεδομένων, από τα 100 δείγματα, τα 13 έχουν προβλεφθεί σωστά, το αποτέλεσμα για την ακρίβεια ταξινόμησης θα είναι:
Ακρίβεια ταξινόμησης = 13/100= 13%

2.8.3 Precision

Το precision [30] ενός αλγορίθμου αντιπροσωπεύεται ως η αναλογία των σωστά ταξινομημένων ασθενών με τη νόσο προς το σύνολο των ασθενών που προβλέπεται ότι θα πάσχουν από τη νόσο.

Precision = True_Positive / (True_Positive + False_Positive),

2.8.4 Recall

Σε πολλές περιπτώσεις η ακρίβεια ταξινόμησης δεν αρκεί για να εκτιμηθεί ο δείκτης απόδοσης ενός μοντέλου. Ένα από αυτά τα σενάρια είναι όταν η κατανομή της μεταβλητής στόχου δεν είναι ισορροπημένη (στην περίπτωση της δυαδικής ταξινόμησης η μια κλάση είναι πολύ μεγαλύτερη από την άλλη). Σε αυτήν την περίπτωση ένα υψηλό ποσοστό ακριβείας δεν δείχνει να ικανοποιεί μια επικείμενη πρόβλεψη επειδή το μοντέλο δεν μαθαίνει τίποτα. Η recall [31] είναι μια άλλη σημαντική μέτρηση, η οποία ορίζεται ως το κλάσμα των δειγμάτων από μια κατηγορία που προβλέπονται σωστά από το μοντέλο. Πιο επίσημα:

Recall = True_Positive/ (True_Positive+ False_Negative),

Όπου κι εδώ ισχύουν τα παραπάνω και η false negative είναι ένα αποτέλεσμα όπου το μοντέλο προβλέπει λανθασμένα την αρνητική κλάση.

2.8.5 F1 Score

Υπάρχουν πολλές περιπτώσεις στις οποίες τόσο η ανάκληση όσο και η ακρίβεια είναι σημαντικές. Επομένως, ήταν αναγκαίο με κάποιο τρόπο να υπάρξει μια μετρική που να συσχετίζει τις δύο παραπάνω μετρικές. Μια δημοφιλής μέτρηση που συνδυάζει ακρίβεια και ανάκληση ονομάζεται F1-score [32] , η οποία είναι ο αρμονικός μέσος όρος ακρίβειας και ανάκλησης που ορίζεται ως:

F1 Score = 2*Ακρίβεια*Ανάκληση/(Ακρίβεια+Ανάκληση)

2.8.6 Sensitivity and Specificity

Το Sensitivity και το Specificity [33] είναι δύο άλλες δημοφιλείς μετρήσεις που χρησιμοποιούνται κυρίως σε πεδία που σχετίζονται με την ιατρική και τη βιολογία και ορίζονται ως:
Sensitivity = recall = TP/(TP+FN)
Specificity = TN/(TN+FP)

Επειδή, λοιπόν, το σύνολο δεδομένων που πρόκειται να διερευνηθεί αφορά ιατρικά δεδομένα και εφόσον το sensitivity συμπίπτει με την μετρική recall, θα δοθεί ιδιαίτερη βαρύτητα στην πειραματική διαδικασία σε ό,τι αφορά το recall. Αυτό, θα χρησιμεύσει στις περιπτώσεις που ένα προβλεπτικό μοντέλο επιφέρει μεγάλο ποσοστό ακρίβειας και πρέπει να διερευνηθεί αν αυτό το ποσοστό είναι πραγματικό ή έχει γίνει κατανομή των δεδομένων σε μια κλάση από τις 2 της μεταβλητής-στόχου σε μεγαλύτερο βαθμό.

2.8.7 Roc Curve

Η χαρακτηριστική καμπύλη ROC [34] είναι η γραφική παράσταση που δείχνει την απόδοση ενός δυαδικού ταξινομητή ως συνάρτηση του ορίου αποκοπής του. Ουσιαστικά δείχνει το πραγματικό θετικό ποσοστό (TPR) έναντι του ψευδώς θετικού ποσοστού (FPR) για διάφορες τιμές κατωφλίου. Πολλά από τα μοντέλα ταξινόμησης είναι πιθανολογικά, δηλαδή προβλέπουν την πιθανότητα για παράδειγμα ένα δείγμα να είναι ασθενής. Στη συνέχεια συγκρίνουν αυτή την πιθανότητα εξόδου με κάποιο όριο αποκοπής και αν είναι μεγαλύτερο από το όριο προβλέπουν την ετικέτα του ως ασθενή, διαφορετικά ως υγιή. Για παράδειγμα, το μοντέλο μπορεί να προβλέψει τις παρακάτω πιθανότητες για 4 δείγματα εικόνων: [0,45, 0,6, 0,7, 0,3] . Στη συνέχεια, ανάλογα με τις παρακάτω τιμές κατωφλίου, λαμβάνονται διαφορετικές ετικέτες:
cut-off= 0,5: predicted-labels= [0,1,1,0] (προεπιλεγμένο όριο)
cut-off = 0,2: predicted-labels= [1,1,1,1]
cut-off = 0,8: προβλεπόμενο- ετικέτες= [0,0,0,0]
Μεταβάλλοντας τις τιμές κατωφλίου, θα λαμβάνονται εντελώς διαφορετικές ετικέτες. Καθένα από αυτά τα σενάρια θα είχε ως αποτέλεσμα διαφορετική ακρίβεια και ανάκληση (καθώς και TPR, FPR).
Η καμπύλη ROC ουσιαστικά ανακαλύπτει το TPR και το FPR για διάφορες τιμές κατωφλίου και σχεδιάζει το TPR έναντι του FPR.

2.8.8 AUC

Η περιοχή κάτω από την καμπύλη (AUC) [35], είναι ένα συγκεντρωτικό μέτρο της απόδοσης ενός δυαδικού ταξινομητή σε όλες τις πιθανές τιμές κατωφλίου (και επομένως είναι αμετάβλητο κατωφλίου). Η AUC υπολογίζει την περιοχή κάτω από την καμπύλη ROC και επομένως είναι μεταξύ 0 και 1. Ένας τρόπος ερμηνείας της AUC είναι η πιθανότητα το μοντέλο να κατατάσσει ένα τυχαίο θετικό παράδειγμα υψηλότερα από ένα τυχαίο αρνητικό παράδειγμα. Σε υψηλό επίπεδο, όσο υψηλότερη είναι η AUC ενός μοντέλου τόσο καλύτερο είναι. Αλλά μερικές φορές το ανεξάρτητο μέτρο κατωφλίου δεν είναι αυτό που θέλετε, π.χ. μπορεί να ενδιαφέρει τον χρήστη η ανάκληση του μοντέλου

σας και να απαιτεί να είναι υψηλότερο από 99% (ενώ έχει λογική ακρίβεια ή FPR). Σε αυτήν την περίπτωση, μπορεί να θέλει να συντονίσει το όριο του μοντέλου, έτσι ώστε να ανταποκρίνεται στην ελάχιστη απαίτησή του σε αυτές τις μετρήσεις.

2.9 Επιλογή χαρακτηριστικών

Στη μηχανική μάθηση, όσο περισσότερες είναι οι μεταβλητές σε ένα μοντέλο, τόσο αυξάνεται η συνολική πολυπλοκότητα του μοντέλου. Συνήθως, οι μεταβλητές ενός συνόλου δεδομένων δεν είναι όλες χρήσιμες για την δημιουργία του μοντέλου και στη συνέχεια την πρόβλεψη. Η προσθήκη μεταβλητών που μπορεί να μην είναι σημαντικές μπορεί να μειώσει τη συνολική ακρίβεια ενός ταξινομητή. Ο στόχος της επιλογής χαρακτηριστικών (Feature Selection) [36] στη μηχανική μάθηση είναι να βρει το καλύτερο σύνολο χαρακτηριστικών που επιτρέπει σε κάποιον να επιτύχει το καλύτερο δυνατό αποτέλεσμα μέσω ενός μοντέλου για την διεξαγωγή πρόβλεψης και γνώσης.

Οι τεχνικές για την επιλογή χαρακτηριστικών στη μηχανική εκμάθηση μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες, από τις οποίες θα περιγραφούν οι τεχνικές που θα χρησιμοποιηθούν παρακάτω στην πειραματική διαδικασία:

2.9.1 Τεχνικές χωρίς επίβλεψη

Αυτές οι τεχνικές αναφέρονται στις μεθόδους που δεν χρειάζονται την μεταβλητή στόχο (target) για την επιλογή χαρακτηριστικών και συγκεκριμένα μπορούν να χρησιμοποιηθούν για unlabeled data [37].

2.9.2 Εποπτευόμενες τεχνικές

Μπορούν να χρησιμοποιηθούν για labeled data και για τον προσδιορισμό των σχετικών χαρακτηριστικών για την αύξηση της αποτελεσματικότητας των εποπτευόμενων μοντέλων [38], όπως η ταξινόμηση και η παλινδρόμηση και μπορούν να ταξινομηθούν ως εξής:

2.9.2.1 Μέθοδοι φιλτραρίσματος

Οι μέθοδοι φιλτραρίσματος συλλέγουν τις εγγενείς ιδιότητες των χαρακτηριστικών που μετρούνται μέσω μονομεταβλητών στατιστικών και απορρίπτονται με βάση τη σχέση τους με την έξοδο ή τον τρόπο συσχέτισης τους με την έξοδο. Αυτές οι μέθοδοι είναι ταχύτερες και λιγότερο υπολογιστικά δαπανηρές. Η ενασχόληση με δεδομένα υψηλών διαστάσεων, είναι υπολογιστικά φθηνότερο να χρησιμοποιείτε μεθόδους φιλτραρίσματος. Στην έρευνα που πραγματοποιήθηκε, οι filter methods που χρησιμοποιήθηκαν είναι οι παρακάτω:

- **Chi-square**

Η μέθοδος chi-square [39] χρησιμοποιείται για κατηγορικά χαρακτηριστικά σε ένα σύνολο δεδομένων. Υπολογίζεται το chi-square μεταξύ κάθε χαρακτηριστικού και του

στόχου και επιλέγεται ο επιθυμητός αριθμός χαρακτηριστικών με τις καλύτερες βαθμολογίες chi-square. Προκειμένου να εφαρμοστεί σωστά το chi-square και για να ελεγχθεί η σχέση μεταξύ των διαφόρων χαρακτηριστικών του συνόλου δεδομένων και της μεταβλητής στόχου (target), πρέπει να πληρούνται οι προϋποθέσεις κατά τις οποίες οι μεταβλητές πρέπει να είναι κατηγορικές και οφείλουν να δειγματίζονται ανεξάρτητα.

- **Συντελεστής συσχέτισης**

Η συσχέτιση είναι ένα μέτρο της γραμμικής σχέσης 2 ή περισσότερων μεταβλητών. Μέσω της συσχέτισης, μπορούμε να προβλέψουμε τη μία μεταβλητή από την άλλη. Η λογική πίσω από τη χρήση συσχέτισης για την επιλογή χαρακτηριστικών είναι ότι οι καλές μεταβλητές συσχετίζονται σε μεγάλο βαθμό με τον στόχο. Επιπλέον, οι μεταβλητές θα πρέπει να συσχετίζονται με τον στόχο αλλά να μην συσχετίζονται μεταξύ τους. Εάν δύο μεταβλητές συσχετίζονται, μπορούμε να προβλέψουμε τη μία από την άλλη. Επομένως, εάν δύο χαρακτηριστικά συσχετίζονται, το μοντέλο χρειάζεται πραγματικά μόνο ένα από αυτά, καθώς το δεύτερο δεν προσθέτει πρόσθετες πληροφορίες. Στην περίπτωση της έρευνας και του συνόλου δεδομένων που χρησιμοποιήθηκε, η ιδανική μέθοδος συσχέτισης είναι η kendall’s, καθώς ενδείκνυται για κατηγορικά δεδομένα [40].

- **Information gain**

Η μέθοδος Information gain υπολογίζει τη μείωση της εντροπίας από τον μετασχηματισμό ενός συνόλου δεδομένων. Μπορεί να χρησιμοποιηθεί για την επιλογή χαρακτηριστικών αξιολογώντας το κέρδος πληροφοριών κάθε μεταβλητής στο πλαίσιο της μεταβλητής στόχου [41]. Η τεχνική feature selection που χρησιμοποιεί η Information Gain είναι γνωστή με το δείκτη Mutual Information ή σε συντομογραφία Mi. Σύμφωνα με την παραπάνω πηγή, όσο μεγαλύτερό το Mi, τόσο μεγαλύτερη είναι η συσχέτιση του χαρακτηριστικού με τη μεταβλητή-στόχο.

2.9.2.2 Wrapper Methods

Οι μέθοδοι αυτές αναζητούν των χώρο όλων των πιθανών υποσυνόλων χαρακτηριστικών, αξιολογώντας την ποιότητά τους, μαθαίνοντας και αξιολογώντας έναν ταξινομητή με αυτό το υποσύνολο χαρακτηριστικών. Η διαδικασία επιλογής χαρακτηριστικών βασίζεται σε έναν συγκεκριμένο αλγόριθμο μηχανικής μάθησης που προσπαθούμε να χωρέσουμε σε ένα σύνολο δεδομένων. Ιδιαίτερες τεχνικές αυτών των μεθόδων αποτελούν οι [42]:

- **Forward Feature Selection**

Αποτελεί μια επαναληπτική μέθοδος η οποία ξεκινάμε χωρίς να μην έχει κανένα χαρακτηριστικό στο μοντέλο. Σε κάθε επανάληψη, συνεχίζει να προσθέτει το χαρακτηριστικό που βελτιώνει καλύτερα το μοντέλο έως ότου η προσθήκη μιας νέας μεταβλητής δεν βελτιώσει περαιτέρω την απόδοση του μοντέλου.

- **Backward Feature Elimination**

Η τεχνική αυτή ξεκινάει με όλα τα χαρακτηριστικά και αφαιρεί το λιγότερο σημαντικό χαρακτηριστικό σε κάθε επανάληψη βελτιώνοντας την απόδοση του μοντέλου. Επαναλαμβάνει τη διαδικασία μέχρι να μην παρατηρηθεί περαιτέρω βελτίωση στην αφαίρεση των χαρακτηριστικών.

- **Recursive Feature Elimination**

Είναι ένας άπληστος αλγόριθμος βελτιστοποίησης που στοχεύει στην εύρεση του υποσυνόλου χαρακτηριστικών με την καλύτερη απόδοση. Δημιουργεί επανειλημμένα μοντέλα και κρατά στην άκρη την καλύτερη ή τη χειρότερη απόδοση σε κάθε επανάληψη. Ακόμα, κατασκευάζει το επόμενο μοντέλο με τα εναπομείναντα χαρακτηριστικά μέχρι να εξαντληθούν όλα. Στη συνέχεια, ταξινομεί τα χαρακτηριστικά με βάση τη σειρά εξάλειψής τους.

2.9.2.3 Embedded methods

Αυτές οι μέθοδοι περιλαμβάνουν τα οφέλη τόσο των μεθόδων περιτύλιξης όσο και των μεθόδων φίλτρου, συμπεριλαμβάνοντας αλληλεπιδράσεις χαρακτηριστικών αλλά και διατηρώντας λογικό υπολογιστικό κόστος. Οι embedded methods [43] είναι επαναληπτικές με την έννοια ότι φροντίζουν για κάθε επανάληψη της διαδικασίας εκπαίδευσης του μοντέλου και εξάγουν προσεκτικά εκείνα τα χαρακτηριστικά που συμβάλλουν περισσότερο στην εκπαίδευση για μια συγκεκριμένη επανάληψη. Μερικά από τα πιο δημοφιλή παραδείγματα αυτών των μεθόδων είναι η παλινδρόμηση LASSO και RIDGE που έχουν ενσωματωμένες λειτουργίες τιμωρίας για τη μείωση της υπερπροσαρμογής. Η παλινδρόμηση Lasso εκτελεί κανονικοποίηση L1 που προσθέτει ποινή ισοδύναμη με την απόλυτη τιμή του μεγέθους των συντελεστών. Η παλινδρόμηση Ridge εκτελεί κανονικοποίηση L2 που προσθέτει ποινή ισοδύναμη με το τετράγωνο του μεγέθους των συντελεστών. Άλλα παραδείγματα ενσωματωμένων μεθόδων (Embedded methods) είναι τα κανονικοποιημένα δέντρα, ο μεμετικός αλγόριθμος και ο τυχαίος πολυωνυμικός.

2.10 Ανισόρροπα δεδομένα

Τα ανισόρροπα προβλήματα κατηγοριοποίησης [44], όπως για παράδειγμα οι υποψήφιοι ασθενείς μιας νόσου, αποτελούν σημαντική πρόκληση για τα μοντέλα μηχανικής εκμάθησης. Όταν η μεταβλητή στόχος, όπως η εμφάνιση στεφανιαίας νόσου στα επόμενα 10 χρόνια στις νεαρές ηλικίες, αποτελεί ένα αρκετά μικρό ποσοστό του συνόλου δεδομένων, που μπορεί να είναι δύσκολο για το μοντέλο να τα αναγνωρίσει και ως αποτέλεσμα μπορεί να προβλέψει υπερβολικά την πλειοψηφική κλάση. Σε αυτές τις περιπτώσεις φαίνεται να είναι απόλυτα επιτυχημένο και με υψηλή ακρίβεια πρόβλεψης. Στην πραγματικότητα, όμως, αυτό συμβαίνει λόγο της ανισορροπίας των δεδομένων στην μεταβλητή στόχο, στις περιπτώσεις που το πρόβλημα είναι δυαδικό.

Οι πιο διαδεδομένες τεχνικές βελτίωσης για την καλύτερη απόδοση των μοντέλων μηχανικής μάθησης όταν υπάρχει η παραπάνω ανισορροπία, αποτελούν η υποδειγματοληψία (undersampling) και η υπερδειγματοληψία (oversampling). Η υποδειγματοληψία ουσιαστικά απορρίπτει δεδομένα για την πλειοψηφική κλάση μέχρις ότου ισορροπήσει με την κατηγορία μειοψηφίας. Το ίδιο ακριβώς συμβαίνει και με την υπερδειγματοληψία στην οποία τα δεδομένα απορρίπονται για την μειοψηφική κλάση μέχρις ότου ισορροπήση με την κατηγορία πλειοψηφίας.

Όσον αφορά την υπερδειγματοληψία, η δημοφιλέστερη βιβλιοθήκη που προσφέρει η Python αποτελεί η SMOTE, η οποία σημαίνει συνθετική τεχνική υπερδειγματοληψίας μειονοτήτων. Όπως υποδηλώνει το όνομα, αυτό παίρνει την κατηγορία μειοψηφίας και προσθέτει νέα παραδείγματα στο σύνολο δεδομένων έως ότου η ποσότητα των δύο κατηγοριών εξισωθεί. Ουσιαστικά, δημιουργεί νέα συνθετικά δεδομένα που περιέχουν εύλογες τιμές που είναι κοντά στον «χώρο χαρακτηριστικών» της κατηγορίας μειοψηφίας χρησιμοποιώντας την αύξηση δεδομένων και έτσι καταφέρνει να φέρει σε ισορροπία τις δύο κλάσεις της μεταβλητής στόχου.

Αντίστοιχα, η υποδειγματοληψία είναι μια τεχνική για την εξισορρόπηση ανομοιόμορφων συνόλων δεδομένων διατηρώντας όλα τα δεδομένα στην κατηγορία μειοψηφίας και μειώνοντας το μέγεθος της κλάσης πλειοψηφίας. Είναι μία από τις πολλές τεχνικές που μπορούν να χρησιμοποιήσουν οι επιστήμονες δεδομένων για να εξάγουν πιο ακριβείς πληροφορίες από αρχικά μη ισορροπημένα σύνολα δεδομένων. Αν και έχει μειονεκτήματα, όπως η απώλεια δυνητικά σημαντικών πληροφοριών, παραμένει μια κοινή και σημαντική δεξιότητα για τους επιστήμονες δεδομένων.

Ένας καλός τρόπος για να φανεί η ανισορροπία μεταξύ των δεδομένων αλλά και έπειτα από την εφαρμογή υπερδειγματοληψίας ή υποδειγματοληψίας η ισορροπία μεταξύ των δεδομένων, είναι η χρήση του Confusion Matrix που αναφέρθηκε νωρίτερα. Χάρει στην παροχή των TP, TN, FP, FN, μπορεί να φανεί εύκολα η κατανομή των δεδομένων στην μεταβλητή στόχο. Έτσι, όταν τελικά τα μοντέλα προβλέπουν με μεγαλύτερη ακρίβεια και πιο ικανοποιητικά ποσοστά μετρικών αξιολόγησης, τότε η υπερδειγματοληψία ή η υποδειγματοληψία φαίνονται επιτυχημένες.

Κεφάλαιο 3 Προβλεπτικά μοντέλα στεφανιαίας νόσου με χρήση αλγορίθμων κατηγοριοποίησης

3.1 Εισαγωγή

Το Framingham Heart Study [45] είναι μια έρευνα που διενεργείται από μια ομάδα ανθρώπων που ξεκίνησε το 1948 στις ΗΠΑ. Εκεί, εξετάζονταν άνθρωποι που κατοικούσαν στην πόλη Φράμινγκχαμ στη Μασαχουσέτη, έτσι ώστε να τους παρακολουθήσει και να προβλέψει το ενδεχόμενο να αναπτύξουν πρόβλημα με τη καρδία τους στα επόμενα δέκα χρόνια και συγκεκριμένα να εμφανίσουν στεφανιαία νόσο. Η προσεκτική παρακολούθηση του πληθυσμού της μελέτης αυτής οδήγησε στον εντοπισμό σημαντικών παραγόντων κινδύνου, παράγοντες σύμφωνα με τους οποίους μπορούν να βγουν ορθά συμπεράσματα για την καρδιολογική υγεία ασθενών. Παρακάτω, θα γίνει ανάλυση των δεδομένων από το σύνολο που βρίσκεται στο Kaggle (https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression?select=framingham.csv).

3.2 Διερεύνηση συνόλου δεδομένων

Μελετώντας το σύνολο δεδομένων, παρατηρείται ότι περιέχει 4238 παρατηρήσεις, οι οποίες περιγράφονται από 15 διαφορετικά χαρακτηριστικά και 1 μεταβλητή που αποτελεί τον στόχο που πρέπει να προβλεφθεί, αν δηλαδή ένας ασθενής πρόκειται να εμφανίσει στεφανιαία νόσο ή όχι μέσα στα επόμενα 10 χρόνια.

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4233	1	50	1.0	1	1.0	0.0	0	1	0	313.0	179.0	92.0	25.97	66.0	86.0	1
4234	1	51	3.0	1	43.0	0.0	0	0	0	207.0	126.5	80.0	19.71	65.0	68.0	0
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0	22.00	84.0	86.0	0
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0	19.16	86.0	NaN	0
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80.0	107.0	0

Εικόνα 8. Διερεύνηση συνόλου δεδομένων

Τα χαρακτηριστικά αυτά περιέχουν πληροφορίες ατομικών δεδομένων, ιστορικών δεδομένων και δημογραφικών δεδομένων και είναι τα παρακάτω:

- **Sex**: Φύλο ασθενούς (Κατηγορικό χαρακτηριστικό)
- **Education**: Επίπεδο εκπαίδευσης ασθενούς (Κατηγορικό χαρακτηριστικό)
- **Age**: Ηλικία ασθενή (Συνεχές χαρακτηριστικό)
- **Current Smoker**: Αν ο ασθενούς είναι καπνιστής ή όχι (Κατηγορικό χαρακτηριστικό)

- **Cigs Per Day**: Αριθμός τσιγάρων που καπνίζει ένας ασθενής κατά μέσο όρο σε μία μέρα (Συνεχές χαρακτηριστικό)
- **BP Meds**: Αν ο ασθενής λαμβάνει χάπια που αφορούν την αρτηριακή του πίεση ή όχι (Κατηγορικό χαρακτηριστικό)
- **Prevalent Stroke**: Αν ο ασθενής έχει περάσει κάποιο εγκεφαλικό επεισόδιο ή όχι (Κατηγορικό χαρακτηριστικό)
- **Prevalent Hyp**: Αν ο ασθενής είναι υπερτασικός ή όχι (Κατηγορικό χαρακτηριστικό)
- **Diabetes**: Αν ο ασθενής έχει διαβήτη ή όχι (Κατηγορικό χαρακτηριστικό)
- **Tot Chol**: Επίπεδο χοληστερόλης ασθενούς (Συνεχές χαρακτηριστικό)
- **Sys BP**: Συστολική πίεση ασθενούς (Συνεχές χαρακτηριστικό)
- **Dia BP**: Διαστολική πίεση ασθενούς (Συνεχές χαρακτηριστικό)
- **BMI**: Σωματική μάζα ασθενούς (Συνεχές χαρακτηριστικό)
- **Heart Rate**: Ρυθμός καρδιακού παλμού ασθενούς (Συνεχές χαρακτηριστικό)
- **Glucose**: Επίπεδο γλυκόζης ασθενούς (Συνεχές χαρακτηριστικό)
- **TenYearCHD**: Αν ο ασθενής πρόκειται να εμφανίσει στεφανιαία νόσο ή όχι στα επόμενα 10 χρόνια (Δυαδικό: “1”, που σημαίνει “Ναι”, “0” που σημαίνει “Όχι”)

Όπως φαίνεται παραπάνω, κάθε χαρακτηριστικό έχει αντιστοιχιστεί σε μια κατηγορία. Συγκεκριμένα, τα χαρακτηριστικά των οποίων οι τιμές τους μπορούν να βρίσκονται ανάμεσα σε ένα εύρος τιμών ονομάζονται συνεχή, ενώ κατηγορικά χαρακτηριστικά ονομάζονται αυτά που οι τιμές τους είναι διακριτές ανάμεσα σε ένα πεπερασμένο πλήθος τιμών.

3.3 Προ-επεξεργασία δεδομένων

Η προ-επεξεργασία δεδομένων είναι ένα αναπόσπαστο βήμα στη Μηχανική Μάθηση, καθώς η ποιότητα των δεδομένων και οι χρήσιμες πληροφορίες που μπορούν να προκύψουν από αυτήν επηρεάζουν άμεσα την ικανότητα των μοντέλων να μαθαίνουν. Επομένως, είναι εξαιρετικά σημαντική πριν τροφοδοτηθούν τα μοντέλα.

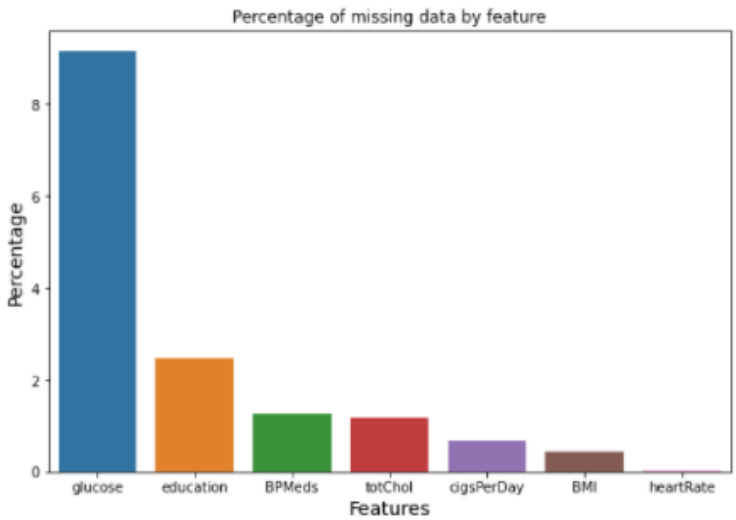
Στην περίπτωση του συνόλου δεδομένων που αφορά τη στεφανιαία νόσο, έχουν παρατεθεί παραπάνω τα χαρακτηριστικά του. Μια σημαντική διαδικασία που θα πρέπει να ακολουθηθεί είναι η αναζήτηση μηδενικών τιμών στο σύνολο δεδομένων, τιμών οι οποίες ακραίες και μπορεί να προκαλούν «θόρυβο», καθώς και εύρεση χαρακτηριστικών που η παρουσία τους στο σύνολο δεδομένων να μην είναι απαραίτητη και σχετική για την επιθυμητή πρόβλεψη. Ξεκινώντας, λοιπόν, και με τη βοήθεια της βιβλιοθήκης pandas και ενός dataframe, λαμβάνονται οι πληροφορίες των τιμών των χαρακτηριστικών με την εντολή info() και έχουμε το παρακάτω αποτέλεσμα:

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   male             4238 non-null   int64
 1   age              4238 non-null   int64
 2   education        4133 non-null   float64
 3   currentSmoker    4238 non-null   int64
 4   cigsPerDay       4209 non-null   float64
 5   BPMeds           4185 non-null   float64
 6   prevalentStroke  4238 non-null   int64
 7   prevalentHyp     4238 non-null   int64
 8   diabetes         4238 non-null   int64
 9   totChol          4188 non-null   float64
 10  sysBP            4238 non-null   float64
 11  diaBP            4238 non-null   float64
 12  BMI              4219 non-null   float64
 13  heartRate        4237 non-null   float64
 14  glucose          3850 non-null   float64
 15  TenYearCHD       4238 non-null   int64
dtypes: float64(9), int64(7)
memory usage: 529.9 KB
```

Εικόνα 9. Τύποι μεταβλητών συνόλου δεδομένων

Γνωρίζοντας από πριν ότι οι συνολικές παρατηρήσεις του συνόλου δεδομένων είναι 4238, παρατηρείται ότι τα χαρακτηριστικά education, cigsPerDay, BPmeds, totChol, BMI, heartrate και glucose περιέχουν τιμές οι οποίες είναι μηδενικές, γεγονός που μπορεί να οδηγήσει σε λάθος αποτελέσματα στην πρόβλεψη. Η συγκεκριμένη παρατήρηση μπορεί να φανεί και με την εντολή isnull, η οποία επιστρέφει τα χαρακτηριστικά που περιέχουν μηδενικές τιμές, καθώς επίσης και πόσες παρατηρήσεις είναι αυτές με τις μηδενικές τιμές, όπως φαίνεται στις παρακάτω φωτογραφίες:

```
df.isnull().sum()

male                0
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
TenYearCHD          0
dtype: int64
```

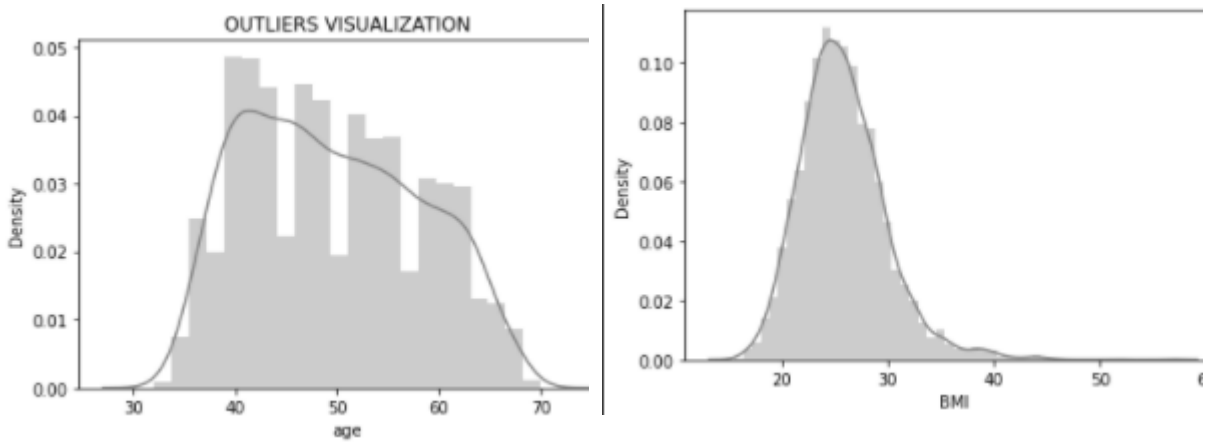


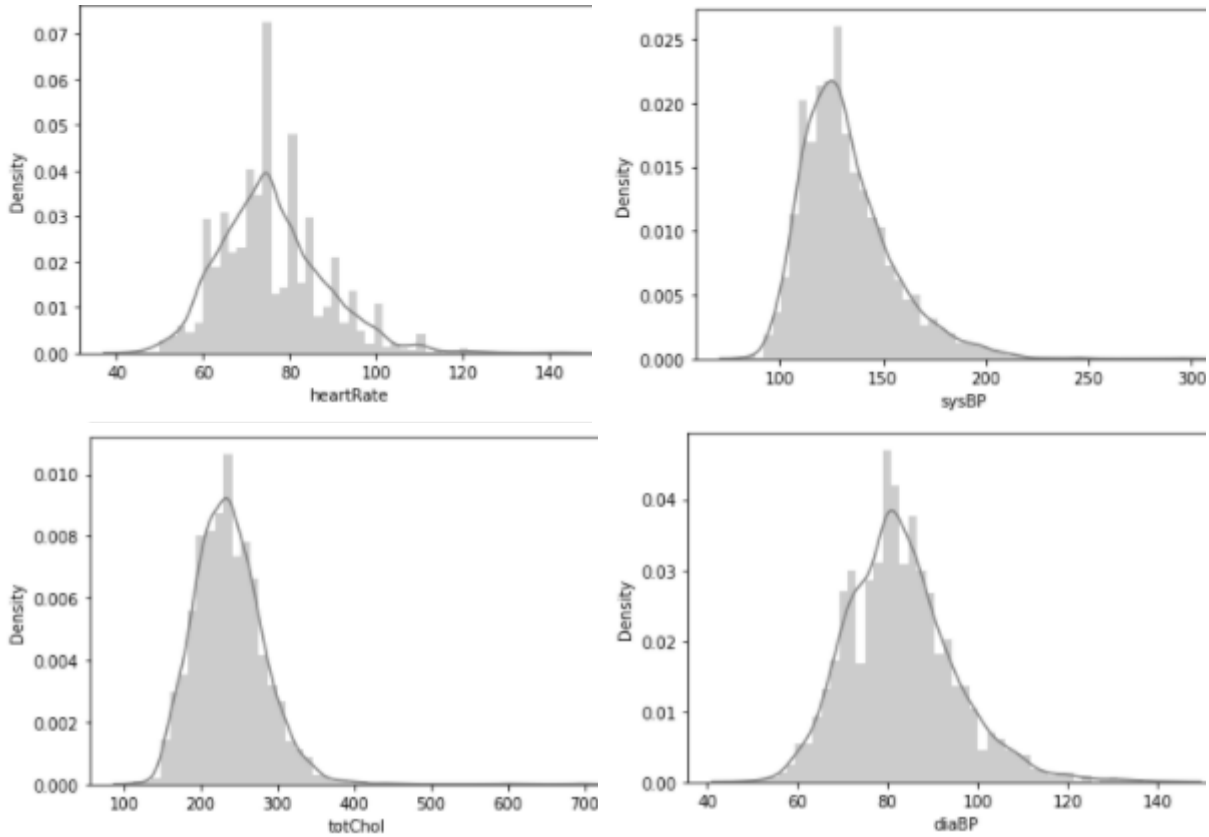
Εικόνα 10. Δεδομένα με μηδενικές τιμές

Αριστερά παρατίθεται το σύνολο των μηδενικών τιμών ανά χαρακτηριστικό, ενώ δεξιά [Εικόνα 10] φαίνεται στο διάγραμμα το ποσοστό των μηδενικών τιμών στις παρατηρήσεις στα εν λόγω χαρακτηριστικά. Το χαρακτηριστικό της γλυκόζης (glucose) έχει τις περισσότερες ελλιπείς τιμές και ίσως είναι το μοναδικό χαρακτηριστικό που θα μπορούσε να προκαλέσει ανεπιθύμητα αποτελέσματα στην πρόβλεψη.

Έτσι, λοιπόν, αρχικά δημιουργείται ο ισχυρισμός ότι το χαρακτηριστικό education που περιγράφει την εκπαιδευτική κατάρτιση κάθε ασθενούς δεν επηρεάζει το τελικό αποτέλεσμα της εμφάνισης της στεφανιαίας νόσου μέσα στα επόμενα 10 χρόνια και έτσι αποφασίστηκε να διαγραφεί ολοκληρωτικά από το σύνολο δεδομένων. Επίσης, τα υπόλοιπα χαρακτηριστικά που αναφέρθηκαν τα οποία περιέχουν μηδενικές τιμές στις παρατηρήσεις, αποφασίστηκε οι συγκεκριμένες παρατηρήσεις να διαγραφούν εξίσου από το σύνολο δεδομένων, με στόχο να παραμείνουν μόνο παρατηρήσεις που να μην είναι μηδενικές. Τελικά, στα δεδομένα θα παραμείνουν 3749 παρατηρήσεις και 15 χαρακτηριστικά.

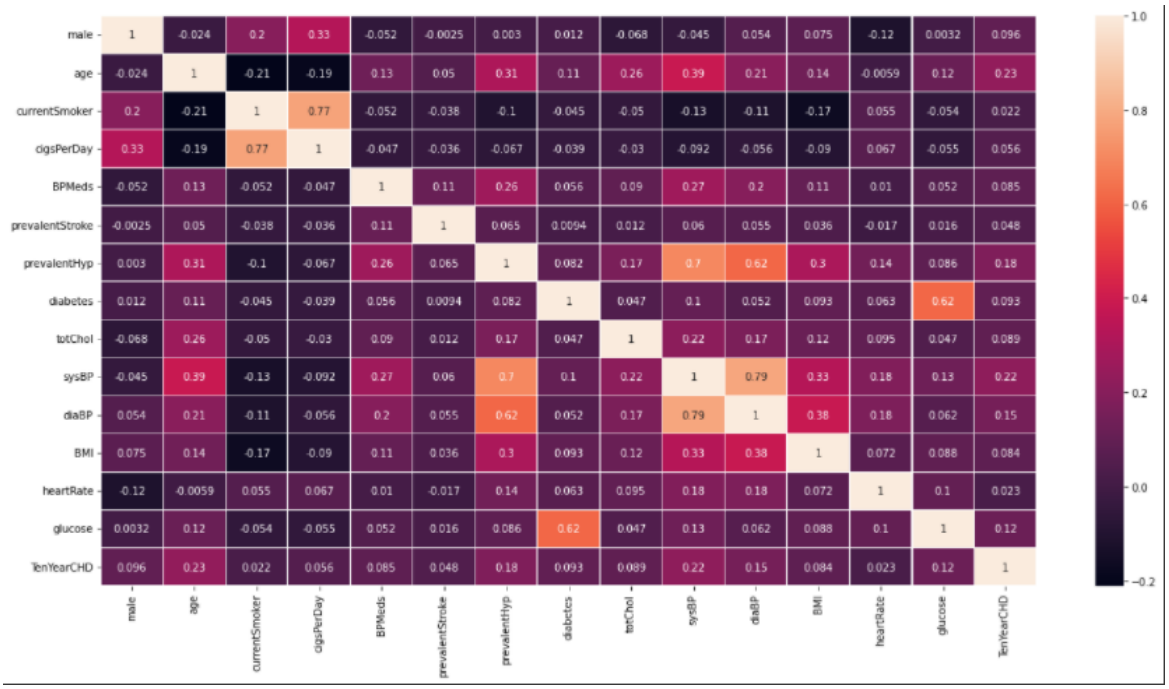
Επιπρόσθετα, όσον αφορά τις ακραίες τιμές που υπάρχουν στις παρατηρήσεις είναι τα σημεία που διαφέρουν σημαντικά σε σχέση με τις υπόλοιπες παρατηρήσεις. Ο τρόπος οπτικοποίησης της παραπάνω έννοιας μπορεί να αποφανθεί στην [Εικόνα 11], όπου τα χαρακτηριστικά που είναι συνεχή φαίνεται ποια είναι η ακραία τους τιμή. Διατυπώθηκε προηγουμένως ότι οι ακραίες τιμές μπορεί να προκαλούν «θόρυβο». Για παράδειγμα ένας ασθενής που μπορεί να έχει πολύ υψηλά ποσοστά γλυκόζης στο αίμα του είναι πιθανό να συμπαρασύρει και τους υπόλοιπους ασθενής στην πρόβλεψη. Οι τιμές να είναι σωστές (για παράδειγμα αν κάποιου ασθενούς ο ρυθμός των παλμών της καρδιάς του είναι υψηλός όταν παρατηρήθηκε), οι συγκεκριμένες τιμές δεν θα θεωρηθούν θόρυβος και δεν θα διαγραφούν από το σύνολο δεδομένων.





Εικόνα 11. Γραφήματα ακραίων τιμών

Μία ακόμα αξιόλογη περίπτωση έρευνας των τιμών των παρατηρήσεων στα χαρακτηριστικά των δεδομένων είναι οι συσχετίσεις μεταξύ τους. Η συγκεκριμένη έρευνα μπορεί να πραγματοποιηθεί με τη βοήθεια του πίνακα συσχετίσεων. Ο πίνακας αυτός χαρακτηρίζεται από έναν συντελεστή, η μετρική του οποίου πληροφορεί τον χρήστη για συσχέτιση δύο ποσοτικών μεταβλητών. Ο συντελεστής μπορεί να πάρει τιμές στο διάστημα [−1, 1] και σε περίπτωση που η τιμή είναι αρνητική, τότε η γραμμική σχέση είναι αρνητική, διαφορετικά, είναι θετική. Όσο πιο κοντά είναι η τιμή στο 1 ή στο -1, σημαίνει ότι η σχέση αυτή είναι ισχυρότερη, ως προς τη γραμμικότητα, είτε είναι θετική είτε είναι αρνητική. Ο παρακάτω πίνακας συσχετίσεων, με τη βοήθεια της βιβλιοθήκης seaborn, δίνει τη δυνατότητα κατανόησης της συσχέτισης των χαρακτηριστικών σε σχέση με την μεταβλητή στόχο, που είναι και ο στόχος της έρευνας. Γίνεται σαφές ότι η μεταβλητή (χαρακτηριστικό) με την μεγαλύτερη συσχέτιση ως προς την στεφανιαία νόσο είναι η ηλικία των ασθενών. Εκτός από την τιμή, αυτό φαίνεται και από το χρώμα που παίρνει το κουτάκι, σύμφωνα με τις τιμές που αναγράφονται δεξιά.



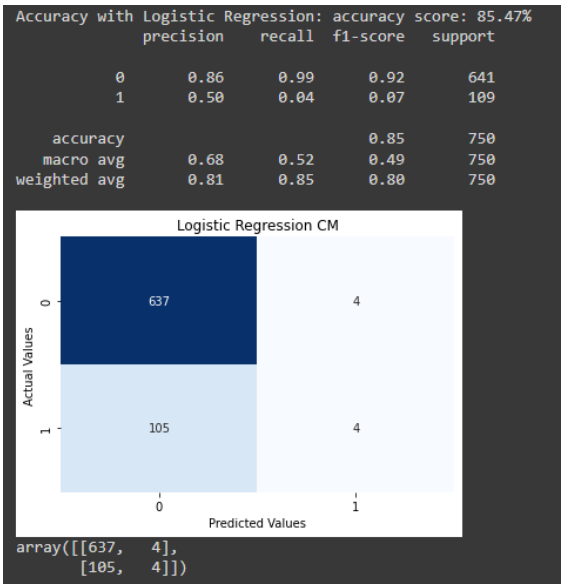
Εικόνα 12. Πίνακας συσχετίσεων

3.4 Πειραματική διαδικασία με χρήση κατηγοριοποίησης

3.4.1 1ο Πείραμα: Απλή κατηγοριοποίηση

Έχοντας κάνει τις απαραίτητες ενέργειες για την διερεύνηση και την προ-επεξεργασία των δεδομένων, πλέον μπορεί να γίνει εφαρμογή των μοντέλων-αλγορίθμων μηχανικής μάθησης που αναφέρθηκαν και περιεγράφηκαν στο Κεφάλαιο 2. Για να την πραγματοποίηση αυτού, αρχικά χωρίστηκαν τα δεδομένα σε δύο διαφορετικές μεταβλητές, όπου η μεταβλητή X περιέχει όλες τις κολώνες (columns) εκτός της κολώνας στόχου (TenYearCHD), η οποία ανατέθηκε στην μεταβλητή Y. Έπειτα, με χρήση της μεθόδου train_test_split της βιβλιοθήκης sklearn, χωρίζονται τα δεδομένα της μεταβλητής X σε train και test, με ποσοστό των test δεδομένων στο 20%.

Με τις απαραίτητες ενέργειες και τον κατάλληλο κώδικα, δημιουργήθηκε μια μέθοδος η οποία δίνει ως αποτέλεσμα τον confusion matrix του εκάστοτε αλγορίθμου και που τρέχει εκείνη τη στιγμή. Ένα παράδειγμα, λοιπόν, του αποτελέσματος του confusion matrix είναι η εφαρμογή του αλγορίθμου Logistic Regression, το οποίο φαίνεται στην Εικόνα 13:



Εικόνα 13. Confusion Matrix

Συγκεκριμένα, περιγράφονται οι σημαντικές μετρικές που αναφέρθηκαν στο Κεφαλαίο 2. Είναι σημαντικό το ποσοστό του 85% της ακρίβειας (Accuracy) που λαμβάνεται από αυτόν τον αλγόριθμο. Κάνοντας, τις απαραίτητες ενέργειες και εφαρμογές όλων των αλγορίθμων με τις κατάλληλες προϋποθέσεις του καθενός, προκύπτει ο παρακάτω πίνακας με τα αποτελέσματα όλων, ταξινομημένα κατά τη μετρική recall:

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	Roc_Auc
DecisionTreeClassifier	**22**	**556**	**85**	**87**	**0.771**	**0.206**	**0.202**	**0.204**	**0.545**
GaussianNB	18	595	46	91	0.817	0.281	0.165	0.208	0.582
LGBMClassifier	7	625	16	102	0.843	0.304	0.064	0.106	0.571
AdaBoostClassifier	7	629	12	102	0.848	0.368	0.064	0.109	0.543
XGBClassifier	5	632	9	104	0.849	0.357	0.046	0.081	0.542
GradientBoostingClassifier	5	631	10	104	0.848	0.333	0.046	0.081	0.534
LogisticRegression	4	637	4	105	0.855	0.500	0.037	0.068	0.531
RandomForestClassifier	4	633	8	105	0.849	0.333	0.037	0.066	0.522
KNeighborsClassifier	2	636	5	107	0.851	0.286	0.018	0.034	0.515
MLPClassifier	1	639	2	108	0.853	0.333	0.009	0.018	0.504
SVMClassification	0	638	3	109	0.851	0.000	0.000	0.000	0.501
SGDClassifier	0	632	0	118	0.843	0.0	0.0	0.0	0.5

Πίνακας 1. Σύγκριση μοντέλων απλής κατηγοριοποίησης

Είναι κατανοητό ότι ο αλγόριθμος με την καλύτερη ακρίβεια είναι ο DecisionTreeClassifier, όμως στα πλαίσια ενός ιατρικού προβλήματος όπως αυτό, η σημαντικότερη μετρική είναι το Recall, καθώς όπως έχει ήδη αναφερθεί, είναι η μετρική του μοντέλου που προβλέπει σωστά τα δείγματα, οπότε ο ιδανικός αλγόριθμός είναι ο DecisionTree. Το ποσοστό 20% recall, δεν μπορεί παρολ' αυτά να θεωρηθεί αξιόπιστο. Η λύση αυτού του προβλήματος θα συζητηθεί παρακάτω.

3.4.2 Κατηγοριοποίηση με χρήση Feature Selection

Στο κεφάλαιο 2 αναφέρθηκαν διάφορες τεχνικές κατηγοριοποίησης με χρήση feature selection. Κάποιες από αυτές χρησιμοποιήθηκαν στην πειραματική διαδικασία, με στόχο την επιλογή των καλύτερων χαρακτηριστικών του συνόλου δεδομένων που προσφέρουν το καλύτερο δυνατό αποτέλεσμα ως προς την ακρίβεια (accuracy), αλλά κυρίως στον δείκτη Recall. Όπως σημειώθηκε και προηγουμένως, το σύνολο δεδομένων περιέχει τόσο κατηγορικές μεταβλητές, όσο και συνεχείς μεταβλητές. Οι τεχνικές feature selection είναι διαφορετικές γι' αυτές τις 2 κατηγορίες μεταβλητών, οπότε προσπάθησε να εφαρμοστεί συνδυαστικά κάθε τεχνική, έτσι ώστε να έχουμε τα καλύτερα χαρακτηριστικά κατηγορικών και συνεχών μεταβλητών και στη συνέχεια να εφαρμοστούν οι γνωστές τεχνικές κατηγοριοποίησης. Στην εκάστοτε επιλογή τεχνικής οφείλουν να συνυπολογίζονται τόσο οι μεταβλητές του input του συνόλου δεδομένων (οι μεταβλητές δηλαδή που αποτελούν την πληροφορία που χρειάζεται ο αλγόριθμος για να εξάγει την επιθυμητή απάντηση), όσο και η μεταβλητή του output του συνόλου δεδομένων (η μεταβλητή που περιγράφει την απάντηση των μοντέλων που εφαρμόζονται στο σύνολο δεδομένων. Οι τεχνικές που επιλέγονται είναι διαφορετικές αναλόγως το είδος των μεταβλητών που είναι στο input και της μεταβλητής του output, αν είναι δηλαδή συνεχείς, κατηγορικές ή ένας συνδυασμός τους. Αποφασίστηκε να πραγματοποιηθούν πειραματικές διαδικασίες στην κατηγοριοποίηση με χρήση chi square, mutual information και correlation matrix και χρήση της kendall's συνδυαστικά.

3.4.3 2ο Πείραμα: Chi Square και Correlation Matrix

Στο συγκεκριμένο πείραμα πραγματοποιήθηκε συνδυασμός των τεχνικών chi square και mutual information. Συγκεκριμένα, η chi square εφαρμόστηκε στις συνεχείς μεταβλητές, δηλαδή στις male, prevalentStroke, currentSmoker, prevalentHyp και diabetes και σύμφωνα με τη θεωρία του κεφαλαίου 2 και τον δείκτη p-value, φαίνεται πως το αποτέλεσμα με αύξουσα αρίθμηση είναι το εξής:

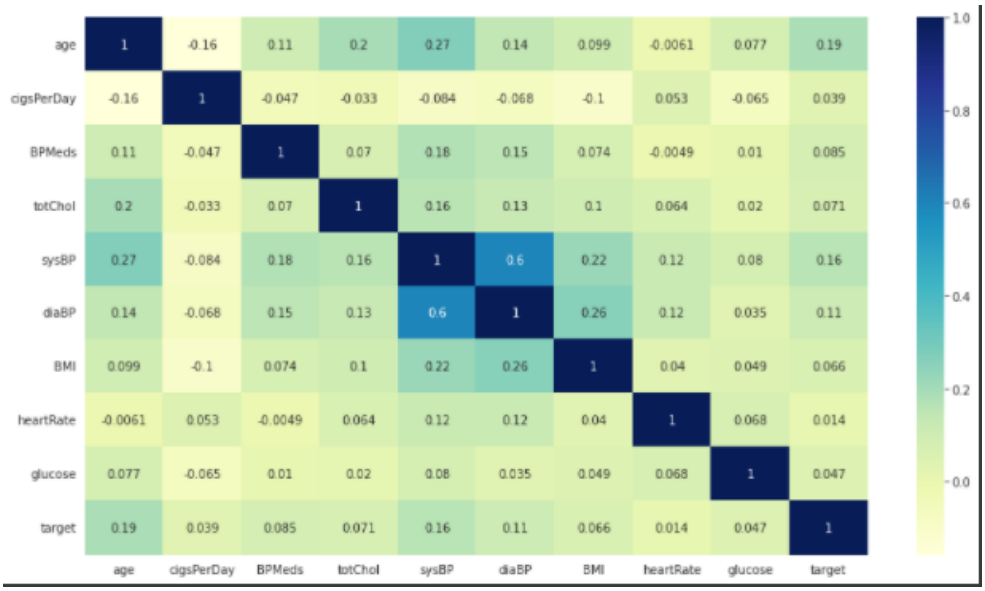
feature	p-value
prevalentHyp	1.077578e-19
diabetes	1.825666e-08
male	1.182306e-05
prevalentStroke	3.608060e-03
currentSmoker	3.414540e-01

Πίνακας 2. Κατηγορικά δεδομένα βάσει P-value

Από τη θεωρία είναι γνωστό ότι όσο μικρότερο είναι το p-value, τόσο σημαντικότερο θεωρείται για την συνεισφορά του στην ακρίβεια πρόβλεψης. Έπειτα από αρκετές πειραματικές διαδικασίες, αποδείχθηκε από τις 5 κατηγορικές μεταβλητές, το ιδανικότερο αποτέλεσμα επιτυγχάνεται επιλέγοντας τις 3 πρώτες με το χαμηλότερο p-value, δηλαδή τις prevelentHyp, diabetes και male.

Έπειτα, στις συνεχείς μεταβλητές εφαρμόστηκε Correlation Matrix με χρήση Kendall's, καθώς με τη βοήθεια του μπορεί να παρατηρηθεί η συσχέτιση μεταξύ των κατηγορικών μεταβλητών (age, cigsPerDay, BPMeds, totChol, sysBP, diaBP, BMI, heartRate,

glucose), σε σχέση με την μεταβλητή στόχο (TenYearCHD). Οι μεταβλητές με τη μεγαλύτερη συσχέτιση είναι κι αυτές που τελικά θα χρησιμοποιηθούν συνδυαστικά με τις μεταβλητές που κρατήθηκαν μέσω της Chi Square, ώστε στο τέλος να εφαρμοστεί η κατηγοριοποίηση τόσο στις συνεχείς, όσο και τις κατηγορικές. Το αποτέλεσμα του Kendall's Correlation Matrix είναι το εξής:



Εικόνα 14. Πίνακας συσχετίσεων Feature Selection

Από την παραπάνω εικόνα, προκύπτει πως η ηλικία (age), η sysBP και η diaBP είναι οι μεταβλητές που συσχετίζονται περισσότερο με την μεταβλητή στόχο. Μάλιστα, έπειτα από πολλές πειραματικές διαδικασίες προέκυψε ότι πράγματι αποφέρουν το καλύτερο δυνατό αποτέλεσμα σε ό,τι αφορά τους γνωστούς μετρικούς δείκτες, όπως φαίνεται παρακάτω στον πίνακα:

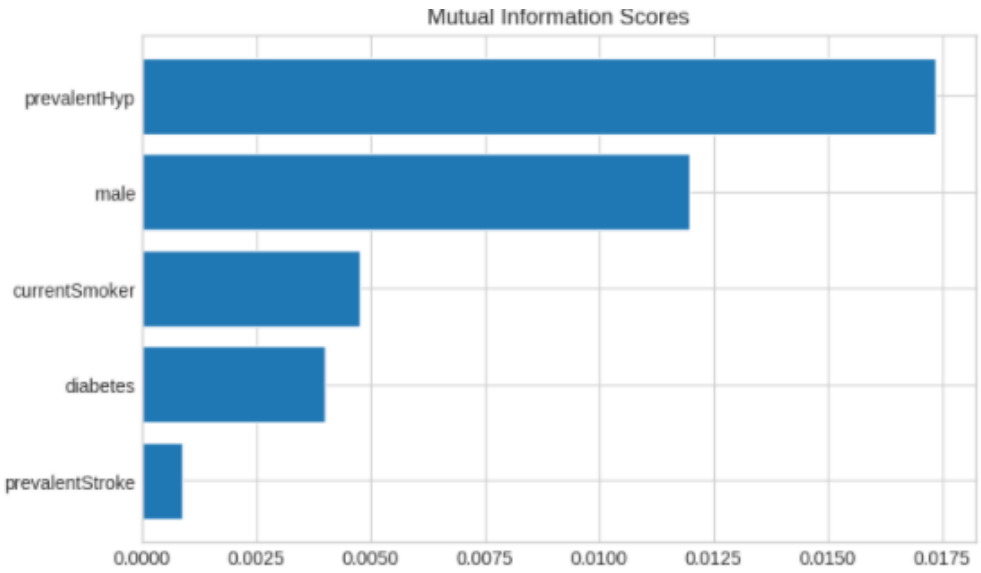
Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	Roc_Auc
DecisionTreeClassifier	**27**	**549**	**92**	**82**	**0.768**	**0.227**	**0.248**	**0.237**	**0.552**
GaussianNB	24	584	57	85	0.811	0.296	0.220	0.253	0.566
LGBMClassifier	14	621	20	95	0.847	0.412	0.128	0.196	0.549
RandomForestClassifier	14	625	16	95	0.852	0.467	0.128	0.201	0.552
GradientBoostingClassifier	7	629	12	102	0.848	0.368	0.064	0.109	0.523
XGBClassifier	6	632	9	103	0.851	0.400	0.055	0.097	0.521
AdaBoostClassifier	3	631	10	106	0.845	0.231	0.028	0.049	0.506
LogisticRegression	1	637	4	108	0.851	0.200	0.009	0.018	0.501
KNeighborsClassifier	1	638	3	108	0.852	0.250	0.009	0.018	0.502
MLPClassifier	1	639	2	108	0.853	0.333	0.009	0.018	0.503
SVMClassification	0	641	0	109	0.855	0.000	0.000	0.000	0.500
SGDClassifier	0	632	0	118	0.843	0.0	0.0	0.0	0.5

Πίνακας 3.Σύγκριση μοντέλων με Chi Square και Correlation Matrix

Παρατηρείται μια αισθητή βελτίωση στον δείκτη Recall που σημαίνει ότι η απόδοση ως προς την σωστή πρόβλεψη της στεφανιαίας νόσου έχει ανέβει σε αρκετό ποσοστό, που όμως ακόμα δεν θεωρείται τόσο αξιόπιστο.

3.4.4 3ο Πείραμα: Mutual Information και Correlation Matrix

Έχοντας κρατήσει από το 2ο πείραμα τις συνεχείς μεταβλητές που απέφεραν τα καλύτερα αποτελέσματα μέσω του Correlation Matrix, θα εφαρμοστεί η τεχνική mutual information στις κατηγορικές μεταβλητές. Συγκεκριμένα, κάνοντας αυτή την τεχνική έχουμε το παρακάτω διάγραμμα όπου φαίνεται το Mi Score κάθε συνεχούς μεταβλητής, με φθίνουσα αρίθμηση:



Εικόνα 15. Γράφημα Mi Score

Είναι εμφανές ότι οι prevelantHyp και male έχουν τη μεγαλύτερη σημαντικότητα για την επιλογή τους στην εφαρμογή της κατηγοριοποίησης, καθώς σύμφωνα με τη θεωρία του υποκεφαλαίου 2.9.2.1, όσο μεγαλύτερο είναι το Mi, τόσο μεγαλύτερη συσχέτιση έχει το χαρακτηστικό με τη μεταβλητή στόχο. Υποψήφια μεταβλητή αποτελεί και η currentSmoker. Έπειτα από πειραματικές διαδικασίες, τα καλύτερα αποτελέσματα φαίνεται πως παρουσιάστηκαν με τη χρήση και των 3 αυτών μεταβλητών και το αποτέλεσμα της κατηγοριοποίησης με τη χρήση των μοντέλων είναι τα παρακάτω:

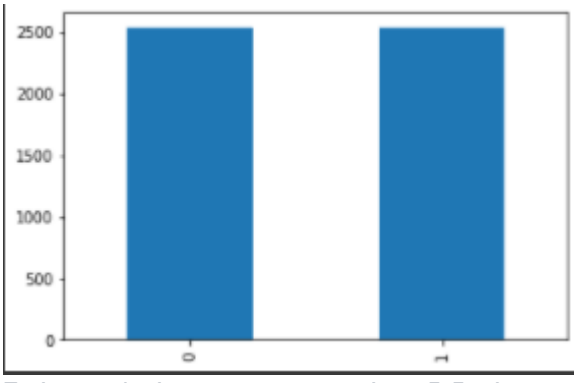
Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	Roc_Auc
DecisionTreeClassifier	**30**	**547**	**94**	**79**	**0.769**	**0.242**	**0.275**	**0.258**	**0.560**
GaussianNB	27	587	54	82	0.819	0.333	0.248	0.284	0.582
SGDClassifier	17	610	31	92	0.836	0.354	0.156	0.217	0.554
LGBMClassifier	17	621	20	92	0.851	0.459	0.156	0.233	0.554
RandomForestClassifier	16	617	24	93	0.844	0.400	0.147	0.215	0.552
AdaBoostClassifier	8	630	11	101	0.851	0.421	0.073	0.125	0.543
GradientBoostingClassifier	8	621	20	101	0.839	0.286	0.073	0.117	0.528
XGBClassifier	7	631	10	102	0.851	0.412	0.064	0.111	0.527
MLPClassifier	3	636	5	106	0.852	0.375	0.028	0.051	0.515
LogisticRegression	1	637	4	108	0.851	0.200	0.009	0.018	0.501
KNeighborsClassifier	1	638	3	108	0.852	0.250	0.009	0.018	0.502
SVMClassification	0	641	0	109	0.855	0.000	0.000	0.000	0.503

Πίνακας 4. Σύγκριση μοντέλων με Mutual Information και Correlation Matrix

Φαίνεται ότι παρόλο που μειώθηκαν τα χαρακτηριστικά δεν έγινε κάποια σπουδαία αλλαγή των σημαντικών αριθμών που αφορούν το Recall. Παρακάτω, θα πραγματοποιηθούν σημαντικότερες κινήσεις για την βελτίωση τους.

3.4.5 4ο Πείραμα: Chi Square και Correlation Matrix με χρήση Smote

Στα επόμενα δύο πειράματα θα χρησιμοποιηθούν οι μέθοδοι Feature Selection με τις οποίες πραγματοποιηθήκαν τα σενάρια 2 και 3, συνδυαστικά με την τεχνική Smote που συζητήθηκε στο κεφάλαιο 2. Στην διερεύνηση του συνόλου δεδομένων, αλλά και με την εφαρμογή της κατηγοριοποίησης στα προηγούμενα σενάρια ήταν εμφανές ότι η μεταβλητή στόχος περιέχει ανισόρροπα δεδομένα. Τα περισσότερα δεδομένα είναι «μαζεμένα» στην μία κλάση από τις δύο κι αυτό έχει ως αποτέλεσμα τα μοντέλα να προβλέπουν με ακρίβεια, όχι όμως με την αξιοπιστία που θα έπρεπε. Αυτό φάνηκε από τις μετρικές αξιολόγησης (πχ το recall) που ήταν χαμηλές (<60%). Έτσι, έχοντας κάνει τα ίδια βήματα με τις προηγούμενες διαδικασίες, αφού σπάσει το σύνολο δεδομένων σε train και test, εφαρμόστηκε η τεχνική Smote (υπερδειγματοληψία) και έτσι ώστε να επιτευχθεί την ισορροπία που χρειάζεται, έχοντας ως βάσει την κλάση με τις περισσότερες παρατηρήσεις. Η μεταβλητή στόχος πλέον τροποποιήθηκε ως εξής:



Εικόνα 16. Ιστόγραμμα ισορροπημένων δεδομένων

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	Roc_Auc
MLPClassifier	**79**	**358**	**283**	**30**	**0.583**	**0.218**	**0.725**	**0.335**	**0.642**
SVMClassification	62	431	210	47	0.657	0.228	0.569	0.325	0.621
LogisticRegression	56	428	213	53	0.645	0.208	0.514	0.296	0.591
KNeighborsClassifier	56	421	220	53	0.636	0.203	0.514	0.291	0.585
AdaBoostClassifier	55	443	198	54	0.664	0.217	0.505	0.304	0.598
GaussianNB	55	482	159	54	0.716	0.257	0.505	0.341	0.628
XGBClassifier	49	473	168	60	0.696	0.226	0.450	0.301	0.594
GradientBoostingClassifier	49	478	163	60	0.703	0.231	0.450	0.305	0.598
SGDClassifier	37	547	94	72	0.779	0.282	0.339	0.308	0.596
LGBMClassifier	36	546	95	73	0.776	0.275	0.330	0.300	0.591
RandomForestClassifier	35	543	98	74	0.771	0.263	0.321	0.289	0.584
DecisionTreeClassifier	33	497	144	76	0.707	0.186	0.303	0.231	0.539

Πίνακας 5. Σύγκριση μοντέλων Chi Square και Correlation Matrix με SMOTE

Όπως αναφέρθηκε στην βιβλιογραφία στο κεφάλαιο 2, η τεχνική SMOTE της βιβλιοθήκης sklearn είναι η λύση στην ισορροπία των δύο κλάσεων της μεταβλητής στόχου. Επιλέγοντας oversampling επιτεύχθηκε η απόλυτη ισορροπία των δύο κλάσεων, γεγονός που μπορεί να έριξε αισθητά το accuracy, όμως ανέβασε κατακόρυφα το Recall. Το recall σε προβλεπτικά μοντέλα ιατρικών συνόλων δεδομένων είναι η σημαντικότερη μετρική, καθώς αναφέρεται στους πραγματικά ασθενείς ανθρώπους του συνόλου δεδομένων. Λαμβάνοντας υπόψιν το Accuracy απλώς υπολογίζει τους κατανεμημένους από τα μοντέλα ως «θετικούς», υπάρχει μεγάλη πιθανότητα να έχει κατανέμει λάθος κάποιον ασθενή. Είναι, όμως, προτιμότερο ο αλγόριθμος να προβλέψει λανθασμένα έναν ασθενή ως ασθενή (να μην εμφανίσει τελικά συμπτώματα στεφανιαίας νόσου), από το να μην προβλέψει σωστά τον ασθενή ως ασθενή και τελικά να εμφανίσει συμπτώματα στεφανιαίας νόσου. Έτσι, στο παραπάνω πείραμα το καλύτερο αποτέλεσμα το προσφέρουν τα τεχνητά νευρωνικά δίκτυα.

3.4.6 5ο Πείραμα: Mutual Information και Correlation Matrix με χρήση Smote

Model	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	Roc_Auc
MLPClassifier	**74**	**345**	**296**	**35**	**0.559**	**0.200**	**0.679**	**0.309**	**0.609**
LogisticRegression	62	423	218	47	0.647	0.221	0.569	0.319	0.614
SVMClassification	62	428	213	47	0.653	0.225	0.569	0.323	0.618
AdaBoostClassifier	56	435	206	53	0.655	0.214	0.514	0.302	0.596
GaussianNB	55	464	177	54	0.692	0.237	0.505	0.323	0.614
KNeighborsClassifier	54	413	228	55	0.623	0.191	0.495	0.276	0.570
XGBClassifier	51	457	184	58	0.677	0.217	0.468	0.297	0.590
GradientBoostingClassifier	51	460	181	58	0.681	0.220	0.468	0.299	0.593
SGDClassifier	41	527	114	68	0.757	0.265	0.376	0.311	0.599
LGBMClassifier	33	537	104	76	0.760	0.241	0.303	0.268	0.570
DecisionTreeClassifier	31	498	143	78	0.705	0.178	0.284	0.219	0.584
RandomForestClassifier	28	544	97	81	0.763	0.224	0.257	0.239	0.553

Πίνακας 6. Σύγκριση μοντέλων με Mutual Info και Correlation Matrix με SMOTE

Χρησιμοποιώντας ξανά την τεχνική SMOTE αλλά αυτή τη φορά συνδυαστικά με τις τεχνικές Mutual Information και Correlation Matrix με kendall's, παρατηρήθηκε ότι τα τεχνικά νευρωνικά δίκτυα είναι ξανά το καλύτερο μοντέλο για πρόβλεψη της στεφανιαίας νόσου, όμως υστερεί σε σχέση με το πείραμα 4.

Έπειτά από 5 πειράματα εφαρμογής κατηγοριοποίησης και μετά από χρήση Feature Selection και κάνοντας τα δεδομένα ισορροπημένα, φαίνεται πως το πείραμα 4 προσφέρει το ιδανικό μοντέλο πρόβλεψης με χρήση Chi Square και Correlation Matrix με Kendall's και το καλύτερο μοντέλο αποτελούν μακράν τα τεχνητά νευρωνικά δίκτυα, καθώς μπορούν με ποσοστό **72.5%** να προβλέψουν έναν μελλοντικό ασθενή.

Κεφάλαιο 4 Εκτίμηση ρίσκου

4.1 Εισαγωγή

Η εκτίμηση ρίσκου, γνωστή και ως Risk Stratification, ορίζεται ως η διαδικασία ανάθεσης μιας συγκεκριμένης κατάστασης κινδύνου σε ασθενείς για θέματα που αφορούν την υγεία τους, διαδικασία η οποία πραγματοποιείται σε συνεχή διαστήματα της ζωής τους [45]. Οι καταστάσεις κινδύνου που αναθέτονται στους ασθενείς βασίζονται σε δεδομένα που αντικατοπτρίζουν ζωτικούς δείκτες υγείας, τον τρόπο ζωής, συμπτώματα παρελθουσών ασθενειών και το ιατρικό ιστορικό του καθενός. Οι ερευνητές και οι κλινικοί γιατροί μπορούν να αντιστοιχίσουν τον κίνδυνο που διατρέχει κάθε ασθενής με τα επίπεδα φροντίδας που πρέπει να λαμβάνει ο καθένας, εξατομικευμένα σχέδια θεραπείας που αντιστοιχούν, αναλόγως το επίπεδο του προβλήματος, να ακολουθήσουν προσεγγίσεις περίθαλψης που βασίζονται στην αξία και να αντιμετωπίσουν τις προκλήσεις διαχείρισης της υγείας του πληθυσμού σε γενικότερο επίπεδο.

Ο γενικός στόχος της εκτίμησης ρίσκου είναι να εντοπιστούν οι ασθενείς που είναι πιο πιθανό να επωφεληθούν από τη διαχείριση φροντίδας για τη βελτίωση των αποτελεσμάτων των ασθενών και τη μείωση του ρίσκου που μπορεί να οδηγήσει τον ασθενή σε σοβαρό επεισόδιο που να αφορά την υγεία του, όπως κάποιο καρδιακό επεισόδιο ή ακόμα και στο θάνατο. Χρησιμοποιώντας εργαλεία ανάλυσης και αλγορίθμους που αναλύουν προγραμματιστές, επιστήμονες και αναλυτές δεδομένων, ανατίθεται σε κάθε ασθενή μια βαθμολογία ρίσκου η οποία αντιστοιχεί σε κάποιο επίπεδο σοβαρότητας της εν λόγω πιθανής ασθένειας. Οι ασθενείς με υψηλότερους βαθμούς ρίσκου θεωρούνται περισσότερο επιρρεπείς και οφείλουν να παρακολουθούν το πρόβλημα τους με συχνότερη περιοδικότητα, να υποστηρίζονται από την κοινωνία και να τηρούν τις συμβουλές φαρμακευτικής αγωγής ή μιας πρόσκλησης για εγγραφή σε ένα εκπαιδευτικό πρόγραμμα υποστήριξης ασθενών. Όσοι έχουν χαμηλότερες βαθμολογίες κινδύνου δεν σημαίνει ότι πρέπει να επαναπαύονται και ενδέχεται να επωφελούνται από ιατρικές υπηρεσίες και υποστήριξη όπως υπενθυμίσεις αυτοματοποιημένου ιατρικού ελέγχου ή μέσω τηλεϊατρικής. Αυτές οι στρατηγικές προληπτικής φροντίδας έχουν σκοπό να βοηθήσουν στη διατήρηση της υψηλότερης δυνατής κατάστασης υγείας του κάθε ασθενούς, αποφεύγοντας κρίσεις και δυσάρεστα και αναπάντεχα γεγονότα, μειώνοντας τις νοσηλεύσεις και βελτιώνοντας τη συνολική ποιότητα ζωής [46].

Ως αποτέλεσμα, οι γιατροί και οι πάροχοι ιατρικής διάγνωσης ενδέχεται να είναι σε θέση να μειώσουν τις ακριβές υπηρεσίες, να αυξήσουν την ικανοποίηση των ασθενών

και να βελτιώσουν τη συνολική υγεία των ασθενών τους. Η εκτίμηση ρίσκου παίζει σημαντικό ρόλο στη διαχείριση της υγείας του πληθυσμού, κατανοώντας τις ανάγκες των ασθενών σε διαφορετικές κατηγορίες κινδύνου, βελτιώνοντας τα αποτελέσματα για την υγεία.

4.2 Εκτίμηση ρίσκου: Framingham Score

Σύμφωνα με την παραπάνω εισαγωγή στην έννοια της εκτίμησης ρίσκου και συνδυαστικά με το γεγονός ότι μπορεί να βοηθήσει στην πρόβλεψη μιας ασθένειας ή νόσου, μπορεί να βοηθήσει και στην πρόβλεψη της στεφανιαίας νόσου, χρησιμοποιώντας το σύνολο δεδομένων του κεφαλαίου 3. Η ομάδα έρευνας από την πόλη Framingham στην πολιτεία της Μασαχουσέτης στις ΗΠΑ, που δραστηριοποιήθηκε με αυτό το σύνολο δεδομένων δημιούργησε έναν δικό της τρόπο για να μπορεί να προβλέψει την εμφάνιση της στεφανιαίας νόσου σε έναν ασθενή μέσα στα επόμενα 10 χρόνια [47]. Ο αλγόριθμος που δημιούργησαν, για κάθε σύμπτωμα ή γενικότερα κάποιο χαρακτηριστικό που έχει ένας ασθενής προσθέτει μια συγκεκριμένη κλίμακα πόντων το οποίο και ονόμασαν Framingham Score, σύμφωνα με την οποία στο τέλος δείχνει το ποσοστό πιθανότητας εμφάνισης της νόσου. Μάλιστα, ο αλγόριθμος είναι υλοποιημένος και μπορεί κάποιος χρήστης να βάλει τα συμπτώματα του και κάποια χαρακτηριστικά που αφορούν την υγεία του και κατευθείαν το σύστημα κάνει την πρόβλεψη. Τα χαρακτηριστικά αυτά αφορούν το φύλο, την ηλικία, αν ο ασθενής πάσχει από διαβήτη, αν είναι καπνιστής, την αρτηριακή του πίεση και το επίπεδο χοληστερόλης.

Αυτός ο αλγόριθμος διαμορφώθηκε και έτρεξε για το παραπάνω σύνολο δεδομένων, σύμφωνα με τα παρακάτω βήματα. Ανάλογα με τα χαρακτηριστικά του κάθε ασθενούς, πιστώνονται πόντοι οι οποίοι θα κρίνουν τελικό αποτέλεσμα της πρόβλεψης. Αυτοί οι πόντοι υπολογίζονται είτε σύμφωνα με το χαρακτηριστικό LDL-C (5 low-density lipoprotein cholesterol) είτε σύμφωνα με το χαρακτηριστικό της συνολική χοληστερόλης totChol. Στο σύνολο δεδομένων που έχει ήδη διερευνηθεί παραπάνω, υπάρχει το χαρακτηριστικό totChol οπότε και επιλέχθηκε αυτό. Κάθε βήμα περιλαμβάνει έναν πίνακα με τη λογική αλγορίθμου για κάθε σύμπτωμα-χαρακτηριστικό.

Βήμα 1

Σύμφωνα με την ηλικία των ασθενών που διερευνώνται, διαμορφώνεται η βαθμολόγια σύμφωνα με τον παρακάτω πίνακα. Να σημειωθεί ότι εξαιρούνται οι μικρές ηλικίες, καθώς ο κίνδυνος καρδιακών προβλημάτων είναι αρκετά μικρός:

Ηλικία	Chol Points
30-34	-1
35-39	0
40-44	1
45-49	2
50-54	3
55-59	4
60-64	5
65-69	6
70-74	7

Πίνακας 7. Framingham Score - Ηλικία

Βήμα 2

Σε αυτό το βήμα ελέγχεται το συνολικό επίπεδο χοληστερόλης του ασθενούς και οι πόνοι διαμορφώνονται σύμφωνα με τον παρακάτω πίνακα:

Mg/dl	Chol Points
<160	-3
160-199	0
200-239	1
240-279	2
>=280	3

Πίνακας 8.Framingham Score - Χοληστερόλη

Βήμα 3

Στη συνέχεια προστίθενται πόντοι ανάλογα με το επίπεδο πίεσης στο αίμα κάθε ασθενούς, μετρώντας και την συστολική και την διαστολική, σύμφωνα με τον πίνακα:

SysBP(mm Hg)	Chol Points
<120	0
120-129	0
130-139	1
140-159	2
>=160	3

Πίνακας 9. Framingham Score - Συστολική πίεση

DiaBP(mm Hg)	Chol Points
<80	0
80-84	0
85-89	1
90-99	2
>100	3

Πίνακας 10.Framingham Score - Διαστολική πίεση

Βήμα 4

Έπειτα, διερευνάται αν ο ασθενής πάσχει από διαβήτη και οι πόντοι διαμορφώνονται σύμφωνα με τον παρακάτω πίνακα:

Διαβήτης	
	Chol Points
Όχι	0
Ναι	2

Πίνακας 11.Framingham Score - Διαβήτης

Βήμα 5

Καπνιστής	
	Chol Points
Όχι	0
Ναι	2

Πίνακας 12. Framingham Score - Καπνιστής

Βήμα 6

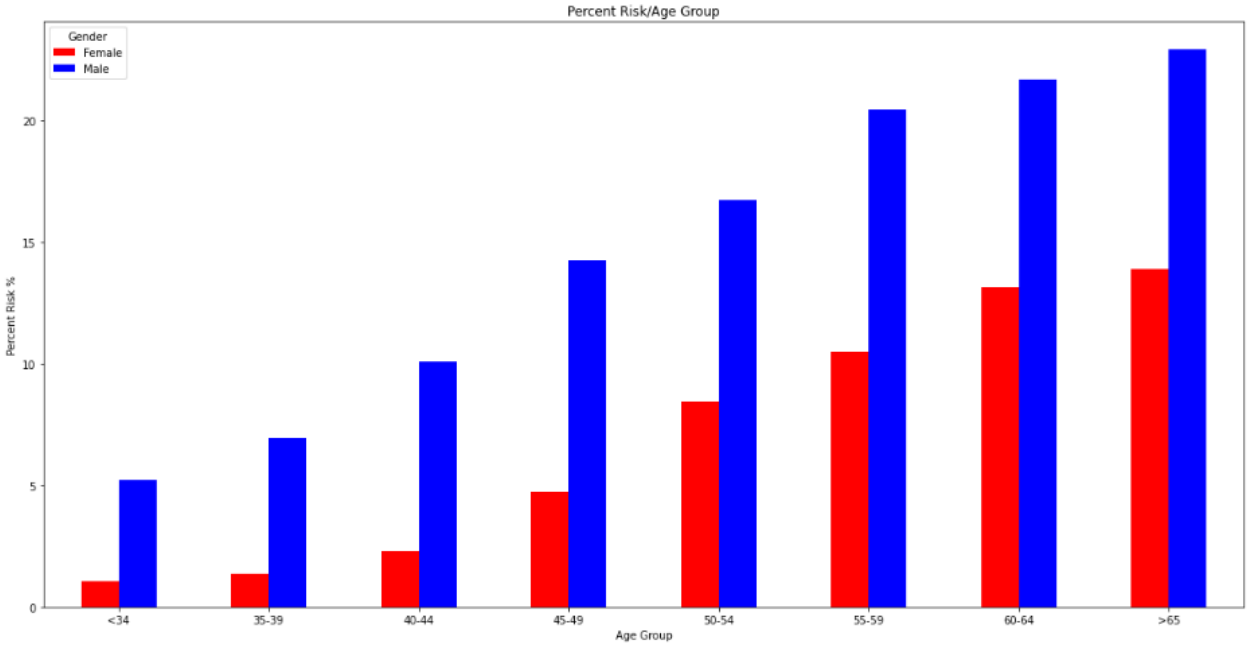
Σύμφωνα με τους πόντους που έχουν μαζευτεί συνολικά από τα παραπάνω βήματα υπολογίζονται για κάθε ασθενής οι συνολικοί πόντοι, οι οποίοι στο επόμενο βήμα κρίνουν το ποσοστό πιθανότητας εμφάνισης της στεφανιαίας νόσου τα επόμενα 10 χρόνια.

Βήμα 7

Chol Points	CHD Risk(10 years)
<-1	2%
0	3%
1	3%
2	4%
3	5%
4	7%
5	8%
6	10%
7	13%
8	16%
9	20%
10	25%
11	31%
12	37%
13	45%
>=14	>=53%

Πίνακας 13.Framingham Score - Τελικό σκορ

Σύμφωνα, λοιπόν, με τον παραπάνω αλγόριθμο, εφαρμόζοντας τον στο σύνολο δεδομένων του κεφαλαίου 3, βγαίνουν τα παρακάτω αποτέλεσμα, βάσει τον οποίων βγαίνουν ορισμένα συμπεράσματα ανάλογα με το φύλο και την ηλικία των πιθανών ασθενών.



Εικόνα 17. Εκτίμηση ρίσκου για ηλικιακές ομάδες γυναικών και αντρών

Για τη δημιουργία του παραπάνω γραφήματος δημιουργήθηκαν οι ηλικιακές ομάδες του αλγορίθμου με εξαίρεση τους ανθρώπους που είναι πάνω από 65 χρονών όπου και μπήκαν δύο ηλικιακές ομάδες (65-69, 70-74). Οι άντρες είναι πιο επιρρεπείς να εμφανίσουν στεφανιαία νόσο για όλα τα χρόνια της ζωής τους σε σχέση με τις γυναίκες. Μάλιστα, αυτό το ποσοστό όσο μεγαλώνουν, τόσο αυτό το ανεβαίνει με αποκορύφωμα τους άνω των 65 οι οποίοι έχουν ποσοστό ρίσκου κοντά στο 30%. Οι γυναίκες εξίσου όσο μεγαλώνουν υπάρχει κίνδυνος εμφάνισης του νοσήματος με κρίσιμες ηλικίες τις 50-59 ετών. Φαίνεται πως υπάρχει αρκετή διαφορά ανάμεσα στα δύο φύλα με τους άντρες να έχουν σε όλες τις ηλικιακές ομάδες σχεδόν διπλάσιο ποσοστό.

4.3 Εκτίμηση ρίσκου: TIMI Score

Εκτός από το Framingham Score που έχει υλοποιηθεί ως βαθμολογία για Risk Stratification από την ομάδα του Framingham υπάρχουν κι άλλα σκορ με τα οποία μπορούν να δημιουργηθούν ανάλογοι αλγόριθμοι. Ένα τέτοιο αποτελεί το TIMI Score [47]. Το TIMI Score βασίζεται σε διάφορους παράγοντες και χρησιμοποιείται για να εκτιμήσει πόσο πιθανό είναι κάποιος να έχει σοβαρές ή απειλητικές για τη ζωή καρδιακές συνέπειες. Η βαθμολογία αφορά τους κινδύνους για άτομα που έχουν διάφορους τύπους πόνου στο στήθος ή έχουν υποστεί καρδιακή προσβολή. Σκοπός και αυτού του σκορ είναι η πρόληψη και η φροντίδα των πιθανών ασθενών. Έτσι, λοιπόν, το TIMI Score αφορά ανθρώπους με τα παρακάτω χαρακτηριστικά:

- Άτομα άνω των 65 ετών
- Να έχει τουλάχιστον τρεις παράγοντες κινδύνου για στεφανιαία νόσο, όπως διαβήτη, υπέρταση, να είναι καπνιστής ή να έχει οικογενειακό ιστορικό καρδιακής νόσου

- Λήψη ασπιρίνης και γενικώς χαπιών σε διάστημα μιας εβδομάδας
- Σοβαρή ασταθή στηθάγχη το τελευταίο εικοσιτετράωρο
- Πρόβλημα που φαίνεται με το ηλεκτροκαρδιογράφημα
- Συγκεκριμένοι δείκτες στην καρδιακή υγεία

Σύμφωνα με τα παραπάνω, για κάθε σύμπτωμα σημαίνει 1 πόντος, και αυτό αποφέρει τα παρακάτω συμπεράσματα:

- Ένας βαθμός σημαίνει 5% κίνδυνο για θνησιμότητα που σχετίζεται με την καρδιά.
- Δύο βαθμοί σημαίνουν 8% κίνδυνο θνησιμότητας.
- Τρεις βαθμοί σημαίνουν % κίνδυνο θνησιμότητας.
- Τέσσερις βαθμοί σημαίνουν 20% κίνδυνο θνησιμότητας.
- Πέντε βαθμοί σημαίνουν 26% κίνδυνο θνησιμότητας.
- Έξι βαθμοί και πάνω σημαίνουν 41% κίνδυνο θνησιμότητας.

Το TIMI Score έχει βρεθεί ότι μετρά αποτελεσματικά τη λειτουργία των στεφανιαίων αιμοφόρων αγγείων που παρέχουν ροή αίματος στην καρδιά. Ωστόσο, ένα ελάττωμα στην κλίμακα TIMI είναι ότι ορισμένες κατηγορίες που περιλαμβάνονται στη βαθμολογία έχουν βρεθεί ότι έχουν μεγαλύτερο ιατρικό βάρος από άλλες.

Για το σύνολο δεδομένων της εργασίας φαίνεται ότι έχει εφαρμογή μόνο στην ηλικακή ομάδα των 65-69. Επίσης, θεωρήθηκε ότι υπερτασικός είναι ένας άνθρωπος με 140 συστολική πίεση και 90 διαστολική. Η συγκεκριμένη ηλικιακή ομάδα φαίνεται πως δεν είχε τα συμπτώματα που ζητάει ο αλγόριθμος, οπότε τα αποτελέσματα που λήφθηκαν δεν ήταν ικανοποιητικά και λογικά, καθώς η πιθανότητα κιμαινόταν μεταξύ του 1-2%, γεγονός απίθανο για μια τέτοια ηλικιακή ομάδα, αν συγκριθεί με το Framingham Score και για τα δύο φύλα.

4.4 Εκτίμηση ρίσκου: Grace Score

Το GRACE Score είναι ένα είδος υπολογισμού καρδιαγγειακών προβλημάτων ασθενών λίγο διαφορετικό σε σχέση με τα υπόλοιπα. Οι πόντοι του υπολογισμού του συγκεκριμένου σκορ δεν έχουν να κάνουν ούτε με το φύλο ούτε με την ηλικία του ασθενή αλλά με χαρακτηριστικά που αφορούν την υγεία του και κυρίως με καρδιακά επεισόδια που μπορεί να έχει περάσει ο ασθενής, όπως έμφραγμα, καρδιακή ανεπάρκεια κλπ [48]. Στα πλαίσια της διπλωματικής προσπάθησε να γίνει εφαρμογή του συγκεκριμένου αλγορίθμου, τα αποτελέσματα, όμως δεν είναι αντιπροσωπευτικά διότι για τον υπολογισμό χρειάζονται χαρακτηριστικά που δεν υπάρχουν στο σύνολο δεδομένων της εργασίας. Συγκεκριμένα, ο αλγόριθμος υπολογισμού διαμορφώνεται ως εξης:

Βήμα 1

Ηλικία	Πόντοι
<=29	0
30-39	0
40-49	18
50-59	36
60-69	55
70-79	73
80-89	91
>=90	100

Πίνακας 14. Ηλικία και πόντοι Grace Score

Βήμα 2

Ανακοπή καρδιάς	Πόντοι
Όχι	0
Ναι	24

Πίνακας 15.Ανακοπή και πόντοι Grace Score

Βήμα 3

Οξύ έμφραγμα μυοκαρδίου	Πόντοι
Όχι	0
Ναι	12

Πίνακας 16. Οξύ έμφραγμα μυοκαρδίου και πόντοι Grace Score

Βήμα 4

Heart Rate	Πόντοι
<=49.9	0
50-69.9	3
70-89.9	9
90-109.9	14
110-149.9	23
150-199.9	35
>=200	43

Πίνακας 17. Heart Rate και πόντοι Grace Score

Βήμα 5

Συστολική πίεση	Πόντοι
<=79.9	24
80-99.9	22
100-119.9	18
120-139.9	14
140-159.9	10
160-199.9	4
>=200	0

Πίνακας 18. Συστολική πίεση και πόντοι Grace Score

Βήμα 6

Προβληματικό καρδιογράφημα	Πόντοι
Όχι	0
Ναι	11

Πίνακας 19. Προβληματικό καρδιογράφημα και πόντοι Grace Score

Βήμα 7

Κρεατινή	Πόντοι
0-039	1
0.4-0.79	3
0.8-1.19	5
1.2-1.59	7
1.6-1.99	9
2-3.99	15
>=4	20

Πίνακας 20. Κρεατίνη και πόντοι Grace Score

Βήμα 8

Αυξημένα ένζυμα ή δείκτες	Πόντοι
Όχι	0
Ναι	15

Πίνακας 21. Ένζυμα-δείκτες και πόντοι Grace Score

Βήμα 9

Καμία διαδερμική επαναγγείωση	Πόντοι
Όχι	0
Ναι	14

Πίνακας 22. Διαδερμική επαναγγείωση και πόντοι Grace Score

Το σύνολο των πόντων που «μαζεύει» ένας ασθενής μπορούν να προβλέψουν αν αυτός ο ασθενής θα πεθάνει μέσα στο νοσοκομείο (εν ώρα νοσηλείας) ή 6 μήνες μετά το εξιτήριο από το νοσοκομείο[49], σύμφωνα με τον παρακάτω πίνακα:

Risk category	Grace Score	In hospital death
Low	<=108	<1
Intermediate	109-140	1-3
High	>140	3

Πίνακας 23. Grace Score για θάνατο εντός του νοσοκομείου

Risk category	Grace Score	Post discharge to 6 months death
Low	<=88	<3
Intermediate	89-118	3-8
High	118	>8

Πίνακας 24. Grace score για θάνατο έξι μήνες από το εξιτήριο

Ο συγκεκριμένος αλγόριθμος ενώ έχει αρκετό ενδιαφέρον, φαίνεται ότι κι αυτός δεν είναι ο ιδανικότερος για το σύνολο δεδομένων που μελετήθηκε, καθώς τα δεδομένα που ζητάει δεν υπάρχουν κατά μεγάλο βαθμό στα χαρακτηριστικά του. Ο ιδανικός αλγόριθμος γι’ αυτό το σύνολο δεδομένων αποτελεί σαφώς ο αλγόριθμος Framingham Score.

4.5 Εκτίμηση ρίσκου και Feature Selection

Μια απλή σκέψη για την εκτίμηση ρίσκου είναι τι θα γινόταν στην περίπτωση που εφαρμοζόταν συνδυαστικά με Feature Selection. Έχοντας εφαρμόσει Feature Selection στο κεφάλαιο 3 για να επιτευχθεί γρηγορότερο και πιο αξιόπιστο αποτέλεσμα, κρίνεται σοβαρότερο ερώτημα του τι θα γινόταν αν, κάνοντας την κατάλληλη επιλογή χαρακτηριστικών, εφαρμόζονταν τα παραπάνω είδη σκορ. Η απάντηση σε αυτό το ερώτημα έχει δύο πλευρές. Σίγουρα η επιλογή ή ακόμα και η μείωση χαρακτηριστικών θα επιφέρει βελτίωση στην απόδοση του αλγορίθμου. Από την άλλη, όμως, επιλέγοντας χαρακτηριστικά από το σύνολο δεδομένων, είτε σημαντικά είτε όχι τόσο σημαντικά, χρησιμοποιώντας τις ίδιες μεθόδους με το κεφάλαιο 3, το αποτέλεσμα μπορεί να μην είναι το επιθυμητό. Ο λόγος είναι ότι τα χαρακτηριστικά που αποφέρουν πόντους και σύμφωνα με τους πόντους γίνεται η εκτίμηση ρίσκου και ο υπολογισμός της πρόβλεψης. Έτσι, αν για παράδειγμα, ένας ασθενής έχει μαζέψει 31 πόντους και βρίσκεται οριακά σε μεσαίο επίπεδο ποσοστού εμφάνισης της στεφανιαίας νόσου, αν το αμέσως κατώτερο επίπεδο σταματάει στους 30 πόντους και ένα χαρακτηριστικό δεν θεωρείται για το σύνολο δεδομένων τόσο σημαντικό αλλά παρολ’ αυτά προσφέρει ένα πόντο στον υπολογισμό του ρίσκου είναι φανερό ότι ο αλγόριθμος θα υπολογίσει άλλο επίπεδο επικινδυνότητας. Αυτό σημαίνει ότι ακόμα και το Feature Selection να θεωρεί κάποιο χαρακτηριστικό λιγότερο σημαντικό, η εκτίμηση ρίσκου δεν μπορεί να εξαιρέσει κάποιο, λόγω των πόντων που μπορεί να προσφέρει.

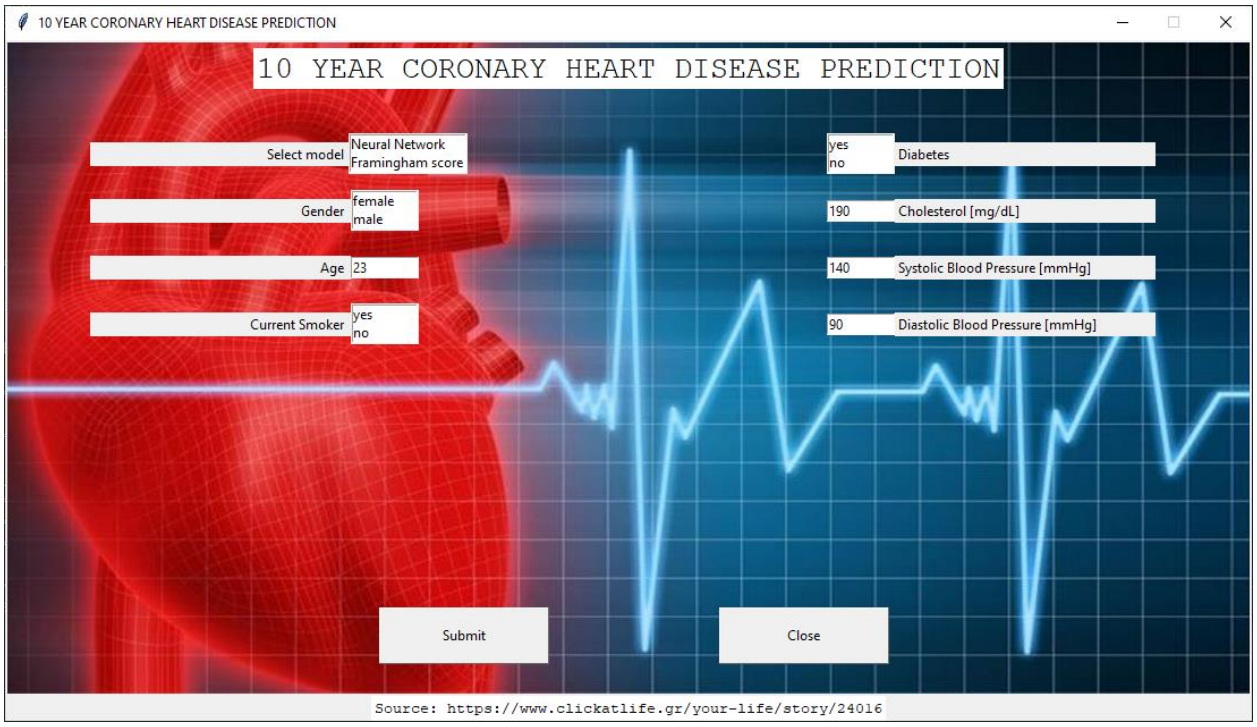
Κεφάλαιο 5 Εφαρμογή πρόβλεψης στεφανιαίας νόσου

5.1 Εισαγωγή

Τα προβλεπτικά μοντέλα του κεφαλαίου 3 και η ανάλυση των αλγορίθμων εκτίμησης ρίσκου του κεφαλαίου 4 πραγματοποιήθηκαν με τη βοήθεια της γλώσσας προγραμματισμού Python [50]. Η Python μέσα από τις πολλές της βιβλιοθήκες προσφέρει και το πακέτο Tkinter [51]. Το Tkinter είναι μια διεπαφή Python της βιβλιοθήκης Tcl/Tk GUI. Το πακέτο Tkinter είναι διαθέσιμο σε όλες τις πλατφόρμες υπολογιστών (Windows, Unix, MacOS κλπ). Προσφέρει ουσιαστικά μια serverless διεπαφή μεταξύ του backend και του frontend για απλές εφαρμογές που έχουν ως στόχο την φιλικότερη αλληλεπίδραση του χρήστη και του συστήματος. Η λογική του μοιάζει αρκετά με απλές γλώσσες frontend όπως η html, όμως περιέχει καθαρά την λογική με την οποία προγραμματίζει κανείς σε Python.

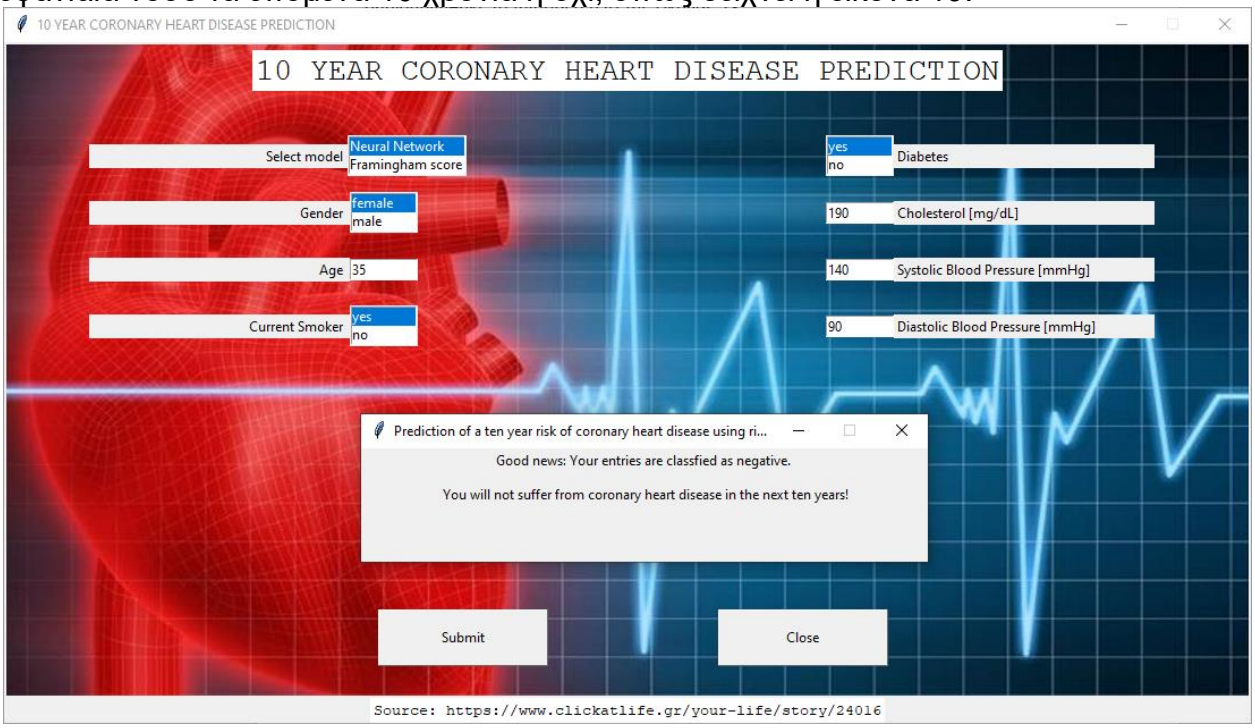
5.2 Εφαρμογή

Συνδυάζοντας, λοιπόν, τα παραπάνω επιλέχθηκε το καλύτερο προβλεπτικό μοντέλο μηχανικής μάθησης που ήταν στο 4ο πείραμα ο αλγόριθμός τεχνητών νευρωνικών δικτύων, ενώ από την εκτίμηση ρίσκου επιλέχθηκε το Framingham Score, το οποίο είναι το πιο συνδεδεμένο είδος εκτίμησης ρίσκου με το σύνολο δεδομένων που ερευνήθηκε. Επιπλέον, θεωρώντας ότι αυτά τα 2 πειράματα έχουν κοινά χαρακτηριστικά, κρατήθηκαν τα «Age, Male, TotChol, sysBP, diaBP, diabetes και currentSmoker». Έτσι, δημιουργήθηκε το Interface (UI) της εικόνας 18:



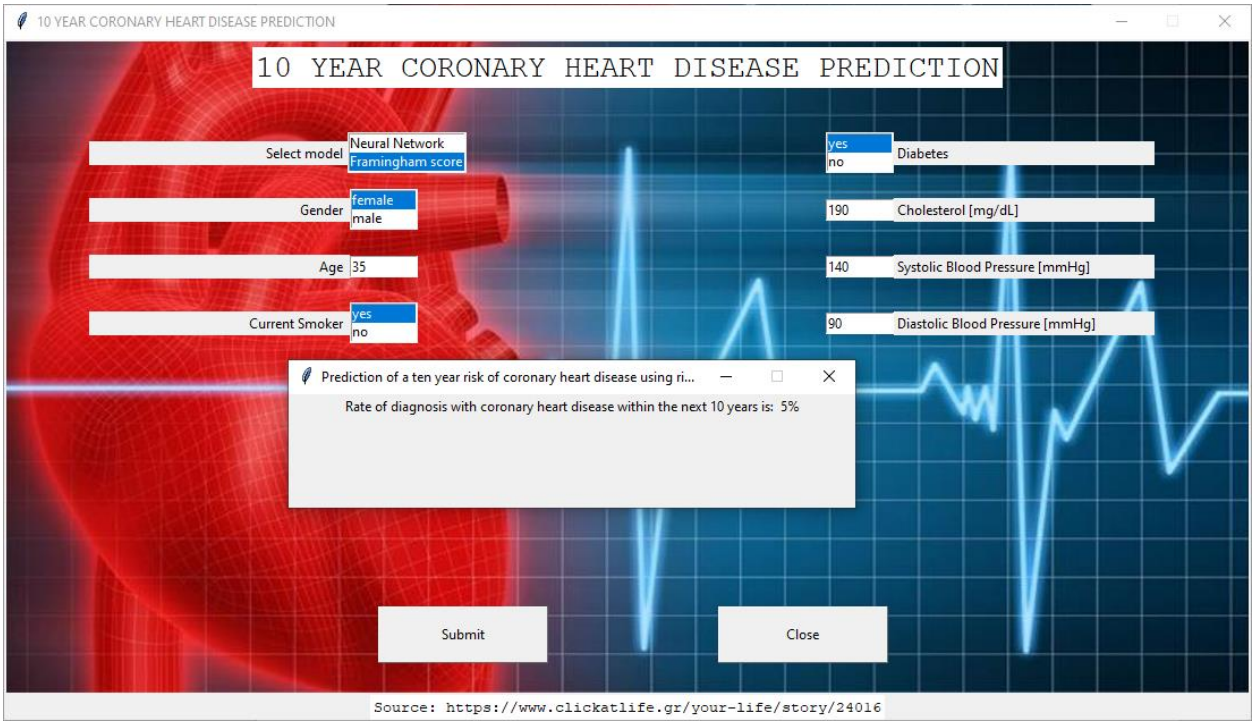
Εικόνα 18. Interface πρόβλεψης στεφανιαίας νόσου με Tkinter

Όπως φαίνεται, δίνεται η επιλογή στο χρήστη να επιλέξει μεταξύ των αλγορίθμων. Όταν επιλέξει τα τεχνητά νευρωνικά δίκτυα, τότε στο backend εκπαιδεύεται το μοντέλο των τεχνητών νευρωνικών δικτύων με το σύνολο δεδομένων του κεφαλαίου 3, προσαρμοσμένο στα χαρακτηριστικά της εφαρμογής και έχοντας υποστεί την ίδια προεπεξεργασία. Πραγματοποιείται εκπαίδευση με τη γνωστή μέθοδο train_test_split με μεγάλο ποσοστό στα δεδομένα εκπαίδευσης (κοντά στο 90-95%), διότι ο αλγόριθμος χρειάζεται μεγάλο ποσοστό γνώσης των δεδομένων εκπαίδευσης για να προβλέψει όσο το δυνατόν σωστότερα. Μετά την εκπαίδευση, γίνεται διάβασμα του input από την οθόνη και σύμφωνα με αυτό, το μοντέλο προσπαθεί να προβλέψει αν τελικά τα στοιχεία που εισήχθησαν θα ταξινομήσουν τον χρήστη σε αυτούς που πρόκειται να εμφανίσει στεφανιαία νόσο τα επόμενα 10 χρόνια ή όχι, όπως δείχνει η εικόνα 19:



Εικόνα 19. Αποτέλεσμα νευρωνικού δικτύου

Ο συγκεκριμένος ασθενής, σύμφωνα με τα τεχνητά νευρωνικά δίκτυα, φαίνεται πως δεν θα εμφανίσει στα επόμενα 10 χρόνια στεφανιαία νόσο. Αντίστοιχα, αν επιλεχθεί ο αλγόριθμος του Framingham Score, θα γίνει ο κατάλληλος υπολογισμός των πόντων και σύμφωνα με αυτούς θα γίνει γνωστό το ποσοστό εμφάνισης της νόσου για τον εν λόγω άνθρωπο στα επόμενα 10 χρόνια, όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 20. Αποτέλεσμα Framingham Score

Όπως φαίνεται, μετά την ολοκλήρωση του αλγορίθμου, εμφανίζεται Pop-up παράθυρο το οποίο αναφέρει ότι η πιθανότητα εμφάνισης της νόσου είναι 5%, ένα αρκετά μικρό ποσοστό που σχεδόν συμφωνεί με το προηγούμενο αποτέλεσμα του νευρωνικού δικτύου ότι ο εν λόγω ασθενής δεν θα εμφανίσει στεφανιαία νόσο.

Κεφάλαιο 6 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε το θέμα της στεφανιαίας νόσου. Με τη βοήθεια του συνόλου δεδομένων από μια ερευνητική ομάδα στην πόλη του Framingham και με τη βοήθεια διάφορων τεχνικών μηχανικής μάθησης αναλύθηκε η περίπτωση εμφάνισης στεφανιαίας νόσου εντός των επόμενων 10 χρόνων στους ανθρώπους που συμμετείχαν στην έρευνα. Χάρη σε αλγορίθμους εποπτευόμενης μηχανικής μάθησης δημιουργήθηκαν μοντέλα τα οποία με διάφορες μετρικές αξιολόγησης προσέφεραν το καλύτερο προβλεπτικό μοντέλο. Με χρήση επιλογής χαρακτηριστικών που θεωρήθηκαν σημαντικότερα και με μεγαλύτερη συσχέτιση με τη μεταβλήτη στόχο που περιγράφει η εμφάνιση ή όχι της νόσου, αλλά και με την τεχνική υπερδειγματοληψίας επιτεύχθηκε η καλύτερη ακρίβεια κατηγοριοποίηση. Όμως ακόμα πιο σημαντικό αποτελεί το γεγονός ότι το μοντέλο MLP απέφερε την καλύτερη μετρική “recall”, με την οποία ταξινομούνται οι πραγματικά μελλοντικοί ασθενείς από τη νόσο. Άλλωστε, όπως ειπώθηκε και παραπάνω, είναι προτιμότερο ένας αλγόριθμος να κατανείμει έναν άνθρωπο ως ασθενή και να μην είναι, παρά να τον κατανείμει ότι δεν είναι ασθενής και τελικά να είναι. Αυτή είναι μια πεποίθηση πολύ συνηθισμένη στον τομέα της ιατρικής, πόσω μάλλον στην μηχανική μάθηση, στην οποία σαφώς και υπάρχει η πιθανότητα λανθασμένης πρόβλεψης. Έπειτα, με τη βοήθεια τεχνικών εκτίμησης ρίσκου και την εφαρμογή διάφορων αλγορίθμων αποδείχθηκε ότι ο αποδοτικότερος για το σύνολο δεδομένων που μελετήθηκε είναι το Framingham Score, που είναι και αυτός που προτάθηκε από την ερευνητική ομάδα. Τα υπόλοιπα Scores σίγουρα αποτελούν σημαντικές τεχνικές πρόβλεψεις, όμως ζητούν άλλα δεδομένα σε σχέση με αυτά που παρέχει το συγκεκριμένο σύνολο. Τέλος, δημιουργήθηκε μια απλή εφαρμογή η οποία δίνει την επιλογή στο χρήστη να επιλέξει μεταξύ του μοντέλου μηχανικής μάθησης και του Framingham Score να καταχωρήσει τα χαρακτηριστικά και τα συμπτώματα του και βάσει αυτών αλλά και βάσει της εκπαίδευσης τους με τις τεχνικές μηχανικής μάθησης να προβλεφθεί, αν τελικώς ο ασθενής θα εμφανίσει μελλοντικά τη νόσο. Σε μελλοντικές έρευνες θα μπορούσε η συγκεκριμένη εφαρμογή να λειτουργήσει πάνω σε έναν Server, να υλοποιηθεί με σύγχρονες γλώσσες προγραμματισμού front-end, συνδυαστικά με Python στο back-end και επίσης να υλοποιθούν στο back-end περισσότερα μοντέλα μηχανικής μάθησης. Επίσης, ανεξαρτήτως του συνόλου δεδομένων, θα μπορούσε η εφαρμογή να περιέχει για κάθε είδους score που μελετάται, όλα τα χαρακτηριστικά που ζητάει ο εκάστοτε ο αλγόριθμος και ανάλογα το score να δημιουργείται η πρόβλεψη.

Πίνακας ορολογίας

Ξενόγλωσσος όρος	Ελληνικός Όρος
Data Science	Επιστήμη Δεδομένων
Data Analytics	Ανάλυση Δεδομένων
Supervised Machine Learning	Εποπτευόμενη Μάθηση
Unsupervised Machine Learning	Μάθηση χωρίς Επίβλεψη
Semi-supervised learning	Ημι-εποπτευόμενη Μάθηση
Reinforcement learning	Ενισχυτική Μάθηση
Classification	Κατηγοριοποίηση
Regression	Παλινδρόμηση
Decision Tree	Δέντρο Απόφασης
Random Forest	Τυχαίο Δάσος
K Nearest Neighbors	Κ κοντινότεροι γείτονες
Naive Bayes	Μπεϊζιανός
Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
Logistic Regression	Λογιστική παλινδρόμηση
Multilayer Perceptron	Τεχνητά Νευρωνικά Δίκτυα
Gradient Boosting	
Oversampling	Υπερδειγματοληψία
Undersampling	Υποδειγματοληψία
Risk Stratification	Εκτίμηση Ρίσκου
Ακρίβεια κατηγοριοποίησης	Classification Accuracy

Συντμήσεις – Αρτικόλεξα – Ακρωνύμια

K Nearest Neighbors	KNN
Support Vector Machine	SVM
Stochastic Gradient Descent	SGD
Cross Validation	CV
Area Under the Curve	AUC
Receiver Operating Characteristic	ROC
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
Mi	Mutual Information

Παράρτημα Ι Βασικός κώδικας μοντέλων μηχανικής μάθησης

```
    classifiers = {

        "LogisticRegression": LogisticRegression(C= 0.1, penalty= 'l2', random_state= 40,
solver='liblinear'),

        "LGBMClassifier": LGBMClassifier(n_estimators = 100, reg_alpha = 0.2,
reg_lambda = 0.1, random_state=10 ),

        "XGBClassifier": XGBClassifier(),

        "KNeighborsClassifier": KNeighborsClassifier(20),

        "DecisionTreeClassifier": DecisionTreeClassifier(),

        "RandomForestClassifier": RandomForestClassifier(n_estimators = 1000,
random_state = 1),

        "AdaBoostClassifier": AdaBoostClassifier(),

        "GradientBoostingClassifier": GradientBoostingClassifier(),

        "GaussianNB": GaussianNB(),

        "SVMClassification": SVC(C = 0.2, gamma = 0.01),

        "MLPClassifier": MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu',
solver='adam', max_iter=500, random_state=7),

        "SGDClassifier": SGDClassifier(loss='modified_huber', shuffle=True,
random_state= 1)

    }


    df_result = pd.DataFrame(columns=['model', 'tp', 'tn', 'fp', 'fn', 'correct', 'incorrect',

                                      'accuracy', 'precision', 'recall', 'f1', 'roc_auc','avg_pre'])


    for key in classifiers:


        print('*',key)


        start_time = time.time()
```

```
classifier = classifiers[key]
model = classifier.fit(X_train, y_train)
cv = StratifiedKFold(n_splits = 5, shuffle=True, random_state=42)
cv_scores = cross_val_score(model, X_test, y_test, cv=cv, scoring='accuracy')
y_pred = model.predict(X_test)
acc = accuracy_score(y_test, y_pred)*100


tn, fp, fn, tp = confusion_matrix(y_test, y_pred).ravel()
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred)
avg_precision = average_precision_score(y_test, y_pred)
row = {'model': key,
    'tp': tp,
    'tn': tn,
    'fp': fp,
    'fn': fn,
    'correct': tp+tn,
    'incorrect': fp+fn,
    'accuracy': round(accuracy,3),
    'precision': round(precision,3),
    'recall': round(recall,3),
    'f1': round(f1,3),
    'roc_auc': round(roc_auc,3),
```

```
        'avg_pre': round(avg_precision,3),
    }
    df_result = df_result.append(row, ignore_index=True)

#Using SMOTE to balance the training data
from imblearn.over_sampling import SMOTE
smote = SMOTE()
X_ros, y_ros = smote.fit_resample(X_train, y_train)
ros_chd_plot=y_ros.value_counts().plot(kind='bar')
plt.show()
```

Βιβλιογραφία

1. Rothstein, William G. The Coronary Heart Disease Pandemic in the Twentieth Century. CRC Press, 2017.

2. Marwick, Thomas H. “Coronary Heart Disease.” ESC CardioMed, edited by Frank Flachskampf, Oxford University Press, 2018, pp. 441–45, http://dx.doi.org/10.1093/med/9780198784906.003.0089.

3. Oglesby Paul, Mark H. Lepper, William H. Phelan, G. Wesley Dupertuis, Anne Macmillan, Harlley Mckean and Heekbok Park. “A Longitudinal Study of Coronary Heart Disease”, 1963, https://doi.org/10.1161/01.CIR.28.1.20

4. Maas, A. H. E. M., and Y. E. A. Appelman. “Gender Differences in Coronary Heart Disease.” Netherlands Heart Journal, no. 12, Springer Science and Business Media LLC, Nov. 2010, pp. 598–603. Crossref, doi:10.1007/s12471-010-0841-y.

5. Wawrzyniak, Andrew J. “Framingham Heart Study.” Encyclopedia of Behavioral Medicine, Springer International Publishing, 2020, pp. 892–95, http://dx.doi.org/10.1007/978-3-030-39903-0_802.

6. WHAT IS DATA SCIENCE?” Data Science, The MIT Press, 2018, http://dx.doi.org/10.7551/mitpress/11140.003.0005.

7. Doran, Derek. “Data Scientist.” Encyclopedia of Big Data, Springer International Publishing, 2022, pp. 332–35, http://dx.doi.org/10.1007/978-3-319-32010-6_61.

8. Diday, Edwin. “The State of the Art in Symbolic Data Analysis: Overview and Future.” Symbolic Data Analysis and the SODAS Software, John Wiley & Sons, Ltd, pp. 3–41, http://dx.doi.org/10.1002/9780470723562.ch1.

9. Riahi, Youssra, and Sara Riahi. “Big Data and Big Data Analytics: Concepts, Types and Technologies.” International Journal of Research and Engineering, no. 9, Marwah Infotech, Nov. 2018, pp. 524–28. Crossref, doi:10.21276/ijre.2018.5.9.5.

10. “Machine Learning, Statistics, and Data Analytics.” Machine Learning, The MIT Press, 2021, http://dx.doi.org/10.7551/mitpress/13811.003.0005.

11. Samuel, Arthur L. “Some Studies in Machine Learning Using the Game of Checkers. I.” Computer Games I, Springer New York, 1988, pp. 335–65, http://dx.doi.org/10.1007/978-1-4613-8716-9_14.

12. Kalaiselvi, K., and M. Deepika. “Machine Learning for Healthcare Diagnostics.” Learning and Analytics in Intelligent Systems, Springer International Publishing, 2020, pp. 91–105, http://dx.doi.org/10.1007/978-3-030-40850-3_5.

13. Oladipupo, Taiwo. “Types of Machine Learning Algorithms.” New Advances in Machine Learning, InTech, 2010, http://dx.doi.org/10.5772/9385.

14. Patel, Harsh H., and Purvi Prajapati. “Study and Analysis of Decision Tree Based Classification Algorithms.” International Journal of Computer Sciences and Engineering, no. 10, ISROSET: International Scientific Research Organization for Science, Engineering and Technology, Oct. 2018, pp. 74–78. Crossref, doi:10.26438/ijcse/v6i10.7478.

15. “A Survey on Decision Tree Algorithms of Classification in Data Mining.” International Journal of Science and Research (IJSR), no. 4, International Journal of Science and Research, Apr. 2016, pp. 2094–97. Crossref, doi:10.21275/v5i4.nov162954.

16. Sekhar, Pudi, and Sanjeeb Mohanty. “Classification and Assessment of Power System Static Security Using Decision Tree and Random Forest Classifiers.” International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, no. 3, Wiley, Aug. 2015, pp. 465–74. Crossref, doi:10.1002/jnm.2096.

17. Su, Yixin, and Sheng-Uei Guan. “Density and Distance Based KNN Approach to Classification.” International Journal of Applied Evolutionary Computation, no. 2, IGI Global, Apr. 2016, pp. 45–60. Crossref, doi:10.4018/ijaec.2016040103.

18. Johnson, Alicia A., et al. “Naive Bayes Classification.” Bayes Rules!, Chapman and Hall/CRC, 2022, pp. 355–72, http://dx.doi.org/10.1201/9780429288340-14.

19. “Supervised Learning-Classification Using Support Vector Machines.” Python® Machine Learning, John Wiley & Sons, Inc., 2019, pp. 177–203, http://dx.doi.org/10.1002/9781119557500.ch8.

20. “Logistic Regression and Text Classification.” Textual Information Access, John Wiley & Sons, Inc, 2013, pp. 59–84, http://dx.doi.org/10.1002/9781118562796.ch3.

21. Manaswi, Navin Kumar. “Multilayer Perceptron.” Deep Learning with Applications Using Python, Apress, 2018, pp. 45–56, http://dx.doi.org/10.1007/978-1-4842-3516-4_3.

22. Natekin, Alexey, and Alois Knoll. “Gradient Boosting Machines, a Tutorial.” Frontiers in Neurorobotics, Frontiers Media SA, 2013. Crossref, doi:10.3389/fnbot.2013.00021.

23. “Adaboost.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 19–20, http://dx.doi.org/10.1007/978-1-4899-7687-1_917.

24. Nokeri, Tshepo Chris. “Tree Modeling and Gradient Boosting with Scikit-Learn, XGBoost, PySpark, and H2O.” Data Science Solutions with Python, Apress, 2021, pp. 59–74, http://dx.doi.org/10.1007/978-1-4842-7762-1_6.

25. Korstanje, Joos. “Gradient Boosting with XGBoost and LightGBM.” Advanced Forecasting with Python, Apress, 2021, pp. 193–205, http://dx.doi.org/10.1007/978-1-4842-7150-6_15.

26. Theodoridis, Sergios. “Stochastic Gradient Descent.” Machine Learning, Elsevier, 2015, pp. 161–231, http://dx.doi.org/10.1016/b978-0-12-801522-3.00005-7.

27. “Cross-Validation.” Encyclopedia of Biometrics, Springer US, 2009, pp. 206–206, http://dx.doi.org/10.1007/978-0-387-73003-5_615.

28. Mohan Patro, and Manas Ranjan Patra. “A Novel Approach to Compute Confusion Matrix for Classification of N-Class Attributes with Feature Selection.” Transactions on Machine Learning and Artificial Intelligence, Scholar Publishing, Apr. 2015. Crossref, doi:10.14738/tmlai.32.1108.

29. “Accuracy.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 8–8, http://dx.doi.org/10.1007/978-1-4899-7687-1_3.

30. Ting, Kai Ming. “Precision.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 990–990, http://dx.doi.org/10.1007/978-1-4899-7687-1_658.

31. Ting, Kai Ming. “Recall.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 1056–1056, http://dx.doi.org/10.1007/978-1-4899-7687-1_702.

32. “F1-Measure.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 497–497, http://dx.doi.org/10.1007/978-1-4899-7687-1_298.

33. Ting, Kai Ming. “Sensitivity and Specificity.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 1152–1152, http://dx.doi.org/10.1007/978-1-4899-7687-1_758.

34. “ROC Curve.” Encyclopedia of Biometrics, Springer US, 2009, pp. 1131–1131, http://dx.doi.org/10.1007/978-0-387-73003-5_828.

35. “AUC.” Encyclopedia of Machine Learning and Data Mining, Springer US, 2017, pp. 75–75, http://dx.doi.org/10.1007/978-1-4899-7687-1_10025.

36. “Feature Engineering and Selection.” Machine Learning Refined, Cambridge University Press, 2020, pp. 237–72, http://dx.doi.org/10.1017/9781108690935.013.

37. “Unsupervised Feature Selection.” Computational Methods of Feature Selection, Chapman and Hall/CRC, 2007, pp. 35–56, http://dx.doi.org/10.1201/9781584888796-9.

38. Huang, Samuel H. “Supervised Feature Selection: A Tutorial.” Artificial Intelligence Research, no. 2, Sciedu Press, Apr. 2015. Crossref, doi:10.5430/air.v4n2p22.

39. “Chi-Square Test for Categorical Data.” Learning Statistics Using R, SAGE Publications, Inc., 2015, pp. 207–21, http://dx.doi.org/10.4135/9781506300160.n11.

40. Pham-Gia, Thu, and Vartan Choulakian. “Distribution of the Sample Correlation Matrix and Applications.” Open Journal of Statistics, no. 05, Scientific Research Publishing, Inc., 2014, pp. 330–44. Crossref, doi:10.4236/ojs.2014.45033.

41. Muller, M. E. “Information Gain.” Relational Knowledge Discovery, Cambridge University Press, pp. 92–120, http://dx.doi.org/10.1017/cbo9781139047869.006.

42. Novakovic Dj., Jasmina. “Improving the Accuracy of Classification Algorithms for Inductive Learning Rules Using Wrapper Methods.” Tehnika, no. 3, Centre for Evaluation in Education and Science (CEON/CEES), 2015, pp. 528–34. Crossref, doi:10.5937/tehnika1503528n.

43. Witten, Ian H., et al. “Embedded Machine Learning.” Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2011, pp. 531–38, http://dx.doi.org/10.1016/b978-0-12-374856-0.00015-8.

44. Komori, Osamu, and Shinto Eguchi. “Machine Learning Methods for Imbalanced Data.” Statistical Methods for Imbalanced Data in Ecological and Biological Studies, Springer Japan, 2019, pp. 45–55, http://dx.doi.org/10.1007/978-4-431-55570-4_5.

45. Graham, Ian, et al. Risk Stratification and Risk Assessment. Oxford University Press, 2015, http://dx.doi.org/10.1093/med/9780199656653.003.0005.

46. Farooq, Vasim, and Patrick W. Serruys. “Risk Stratification and Risk Scores.” ESC CardioMed, edited by William Wijns, Oxford University Press, 2018, pp. 1371–84, http://dx.doi.org/10.1093/med/9780198784906.003.0335.

47. T R Dawber, F E Moore, G V Mann. “Coronary heart disease in the Framingham study” Am J Public Health Nations Health, 2011, https://doi.org/10.2105/ajph.47.4_pt_2.4

48. Correia, J., et al. “Comparison of the GRACE Score, TIMI Score and a New Laboratorial Score to Predict Adverse Outcomes in Acute Coronary Syndrome.” European Heart Journal, no. Supplement_1, Oxford University Press (OUP), Oct. 2021. Crossref, doi:10.1093/eurheartj/ehab724.1117.

49. “Corrigendum to: 2020 ESC Guidelines for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation: The Task Force for the Management of

Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC).” European Heart Journal, no. 23, Oxford University Press (OUP), May 2021, pp. 2298–2298. Crossref, doi:10.1093/eurheartj/ehab285.

50. Severance, Charles R. Python for Everybody. 2016.

51. “Tkinter GUI.” Python by Example, Cambridge University Press, 2019, pp. 110–23, http://dx.doi.org/10.1017/9781108591942.019.