ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
Π.Μ.Σ. ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΘΕΜΑ: ***Εφαρμογή μοντέλων πρόβλεψης με χρήση του εργαλείου της R για την πρόβλεψη εμφάνισης καρδιακών προβλημάτων σε ασθενείς.***

ΜΠΑΤΣΙΑΚΟΣ ΓΕΩΡΓΙΟΣ Μ.Ε.1909
ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΣΕΠΤΕΜΒΡΙΟΣ 2021

Περίληψη

Σύμφωνα με στατιστικές μελέτες και τον οργανισμό των κέντρων ελέγχου και πρόληψης νοσημάτων CDC (Centers for Disease Control and Prevention), που αποτελεί τον μεγαλύτερο οργανισμό δημόσιας υγείας των Ηνωμένων Πολιτειών της Αμερικής, οι καρδιακές ασθένειες (heart diseases) αποτελούν ένα από τα κυριότερα αίτια θανάτου στην Αμερική αλλά και στον υπόλοιπο κόσμο. Συγκεκριμένα περισσότεροι από εξακόσιες χιλιάδες άνθρωποι χάνουν την ζωή τους λόγω κάποιας μορφής καρδιακής παθήσεως ετησίως, ένα ποσοστό που ανάγεται σε περίπου το ¼ των θανάτων συνολικά.

Ο όρος των καρδιακών ασθενειών μπορεί να αναφέρεται σε διάφορες μορφές παθήσεων με τον κυριότερο αυτών να είναι η στεφανιαία νόσος (Coronary Artery Disease). Η στεφανιαία νόσος είναι η πιο συνηθισμένη καρδιακή ασθένεια που συναντάται σε μεγάλο μέρος των καρδιακά νοσούντων. Συγκεκριμένα το 2017 υπήρξαν περίπου 365 χιλιάδες θάνατοι που οφείλονταν στην συγκεκριμένη πάθηση ενώ περίπου το 7% των ανθρώπων ηλικίας άνω των 20 εμφανίζουν την εν λόγω ασθένεια.

Ο κύριος τρόπος αντιμετώπισης αυτών των παθήσεων είναι η πρόληψη και η πρόγνωση της ώστε να αντιμετωπισθεί προτού κυρίως εμφανιστεί. Σε περίπτωση παθήσεως η κύρια αντιμετώπιση αφορά πάλι σε αλλαγή συνηθειών και τρόπου ζωής.

Όπως είναι λοιπόν εμφανές μια τέτοια πάθηση με μεγάλο βαθμό βαρύτητας είναι άκρως σημαντικό να μπορεί να προβλεφθεί σύμφωνα με συμπτώματα και στοιχεία που μπορεί να σχετίζονται με την εμφάνιση της.

Μιλώντας για πρόγνωση αναφερόμαστε σε πρόβλεψη των πιθανοτήτων εμφάνισης της ασθένειας σύμφωνα και με άλλα δεδομένα. Στην σημερινή λοιπόν πραγματικότητα, όπου τα δεδομένα αποτελούν πλέον τον πυρήνα των διαδικασιών σε όλο το φάσμα της καθημερινότητας, η πιο χρήσιμη και βάσιμη μέθοδος για την πρόβλεψη είναι η εξόρυξη δεδομένων με την εφαρμογή των μεθόδων μηχανικής μάθησης (machine learning). Μέσω της μηχανικής μάθησης έχουμε πρόσβαση σε μια πληθώρα αλγορίθμων που προσδίδουν ακρίβεια και ευελιξία για την ανάλυση των δεδομένων και την πρόβλεψη μέσω αυτών.

Στην παρούσα εργασία θα προσπαθήσουμε μέσω της εξόρυξης και της διερεύνησης των δεδομένων, της οπτικοποίησης αυτών και της εφαρμογής αλγορίθμων μηχανικής μάθησης σε ένα συγκεκριμένο σύνολο(dataset) να κατανοήσουμε την σχετικότητα των μεταβλητών με την παρουσία καρδιακών παθήσεων και να χαρακτηρίσουμε την σχετικότητα της εμφάνισης αυτών με την παρουσία άλλων συμπτωμάτων και την ασφάλεια πρόβλεψης της καρδιακής πάθησης μέσω αυτών.

Λέξεις Κλειδιά: Heart Disease, CHD, Μηχανική Μάθηση, Αλγόριθμοι, Data exploration, data analysis, data visualization, data mining.

Abstract

According to statistical analysis provided from the organization of CDC (Centers for Disease Control and Prevention), which is the most important organization of public health in the United States of America heart disease is one of, if not the most critical causes of death not only in America but in the whole world. Specifically more than 600.000 of people are affected and end up losing their life from some kind of heart disease. That is about 7 percent and almost the ¼ of the total annual deaths in America and the rest of the world.

The terminology given as heart disease refers to different types of conditions with the most crucial and common amongst them being the Coronary Artery Disease. This type of disease is the most usual and commonly found in the samples of heart disease patients. In particular, in the year 2017 there was about 365 thousands of deaths caused by CAD whereas around 7 percent of the people over 20 years old have this specific disease.

The main way of dealing with that kind of conditions are prediction and forecasting of them via the symptoms or other relevant characteristics so that it is avoided. In case it is not timely predicted the only countermeasures are the change of habits and way of life.

As easy as is to see then this type of disease is crucial and most important to be predicted via the symptoms and the characteristics which are relevant, so that more lives are saved.

Regarding forecasting we are referencing the prediction of the possibility of people having CAD or in general heart diseases in correlation with other characteristics. In today's world and reality, where data are in the center of processes in every aspect of the world, the most reliable and accountable method for prediction is data mining using machine learning algorithms. Through machine learning we have access in a great variety of algorithms which provide accuracy and flexibility for analyzing data and predicting possible outcomes.

In the aforementioned paper we will try using methods of data mining and exploratory analysis and by visualizing the data to apprehend the correlation of the variables with the appearance of heart disease, we will understand the connection of other symptoms with them and how accurate we can predict a heart disease through the correlated variables.

Keywords: Heart Disease, CHD, Machine Learning, Algorithms, Data exploration, Data analysis, Data visualization, Data mining.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Φιλιππάκη Μιχαήλ για την συνεχή του καθοδήγηση, για την διαρκή του συμμετοχή παρέχοντας τις πολύτιμες συμβουλές και παρατηρήσεις του και την υπομονή και επιμονή του από την αρχή έως την διεκπεραίωση της διπλωματικής μου εργασίας.

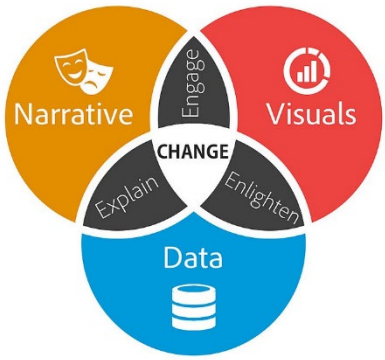
Επίσης θα ήθελα να ευχαριστήσω τη ΔΡ. Μαρία Ελένη Πούλου για την πολύτιμη βοήθεια της στην επίβλεψη της διπλωματικής.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου εκ βαθέων στους γονείς μου για την ασύγκριτη τους συμπαράσταση και καθοδήγηση καθ' όλη την σταδιοδρομία μου μέχρι σήμερα καθώς αποτελούν παράδειγμα προς μίμηση από την εκκίνηση αυτής.

Πίνακας Περιεχομένων

Εισαγωγή

Στην σημερινή εποχή όλος ο κόσμος κατακλύζεται από την διάχυση και συνεχή διάδοση των δεδομένων. Τα δεδομένα αυτά χρίζουν ανεκτίμητης αξίας σε μια σύγχρονη κοινωνία μέσα και από την απίστευτα ραγδαία ανάπτυξη της τεχνολογίας και των μεθοδολογιών. Η ανάλυση των δεδομένων πέραν του ευρέως φάσματος που επιδέχεται εφαρμογής πλέον χρίζει και απαραίτητη σε σύγχρονους οργανισμούς για να επιβιώσουν και να εξελιχθούν. Μέσα από την εξερεύνηση και την ανάλυση τους μια επιχείρηση για παράδειγμα μπορεί να μειώσει το ρίσκο και τον επιχειρηματικό της κίνδυνο, να αυξήσει τα κέρδη της και να βοηθήσει στην λήψη επιχειρηματικών αποφάσεων. Αυτό εφαρμόζεται σε διάφορους τομείς και για διάφορους λόγους. Στον τομέα της υγείας συγκεκριμένα, πέρα από την δυνατότητα βελτίωσης των διαδικασιών προσδίδει και την ευκαιρία της πρόβλεψης και της πρόγνωσης.



Εικόνα 0.1

Όπως είναι φυσικό, η πρόγνωση στον τομέα της υγείας αποτελεί ένα ζήτημα ανυπολόγιστης αξίας καθώς μέσω αυτής επιτυγχάνεται η διάσωση της ανθρώπινης ζωής. Ειδικότερα για παθήσεις που ο μόνος τρόπος αντιμετώπισης τους είναι η πρόγνωση και η πρόβλεψη τους, η δυνατότητα αυτή καταλαμβάνει κεντρικό και κύριο χαρακτήρα. Στην περίπτωση των καρδιακών παθήσεων που αποτελούν μια από τις πιο σημαντικές παθήσεις ως προς τα ποσοστά θνησιμότητας, η πρόβλεψη και πρόγνωση τους καθίσταται ως αναγκαία, κρίσιμη και επιτακτική.

Η πρόβλεψη λοιπόν με την χρησιμοποίηση των δεδομένων επιτυγχάνεται μέσω αλγορίθμων μηχανικής μάθησης. Οι εν λόγω αλγόριθμοι αποτελούν μαθηματικές εξισώσεις εκ των οποίων αποζητούμε όχι μόνο την ακρίβεια αλλά και την ταχύτητα των υπολογισμών ώστε να έχουμε τα επιθυμητά αποτελέσματα.

Οι προαναφερθείσες μαθηματικές εξισώσεις αποτελούν την βάση των αλγορίθμων και κατηγοριοποιούνται σύμφωνα με την εφαρμογή τους για την εξαγωγή των αποτελεσμάτων. Μέσω αυτών έχει δημιουργηθεί ο πιο σύγχρονος τρόπος ανάλυσης

των δεδομένων που είναι η εξόρυξη δεδομένων (Data Mining) τους οποίους θα χρησιμοποιήσουμε και εν συνεχεία θα αξιολογήσουμε στην εν λόγω εργασία.

Δομή Διπλωματικής Εργασίας

Η εκπόνηση της εν προκειμένω εργασίας πραγματοποιήθηκε σε 6 στάδια τα οποία αποτελούν και τα κεφάλαια της.

- Στο 1ο κεφάλαιο έχουμε την ανάλυση της γενικότερης έννοιας της εξόρυξης δεδομένων και της μηχανικής μάθησης καθώς και των συσχετιζόμενων διαδικασιών.
- Στο 2ο κεφάλαιο γίνεται μια περαιτέρω ανάλυση των συγκεκριμένων αλγορίθμων μηχανικής μάθησης εκ των οποίων κάποιοι θα εφαρμοστούν στην συνέχεια, καθώς και το μαθηματικό υπόβαθρο αυτών.
- Στο 3ο κεφάλαιο θα κάνουμε μια τεχνική και γενική ανασκόπηση στο προγραμματιστικό περιβάλλον της R, το εργαλείο που χρησιμοποιήθηκε για την πρακτική εφαρμογή όλων των διαδικασιών και των αλγορίθμων.
- Στο 4ο κεφάλαιο ξεκινάει ουσιαστικά και το πρακτικό μέρος της εργασίας όπου αναλύεται η επιλογή των δεδομένων, γίνεται η προεργασία αυτών καθώς και η διερευνητική ανάλυση τους.
- Στο 5ο κεφάλαιο διενεργείται η εφαρμογή των επιλεγμένων αλγορίθμων, η πρόβλεψη με τη χρήση αυτών και μια τεχνική διαβεβαίωσης της ακρίβειας των αποτελεσμάτων.
- Στο 6ο και τελευταίο κεφάλαιο έχουμε τα αποτελέσματα από την πρακτική εφαρμογή των αλγορίθμων, την έκβαση συμπερασμάτων από τις διαδικασίες που εφαρμόστηκαν καθώς και προτάσεις για περαιτέρω έρευνα.

Συνεισφορά Εργασίας

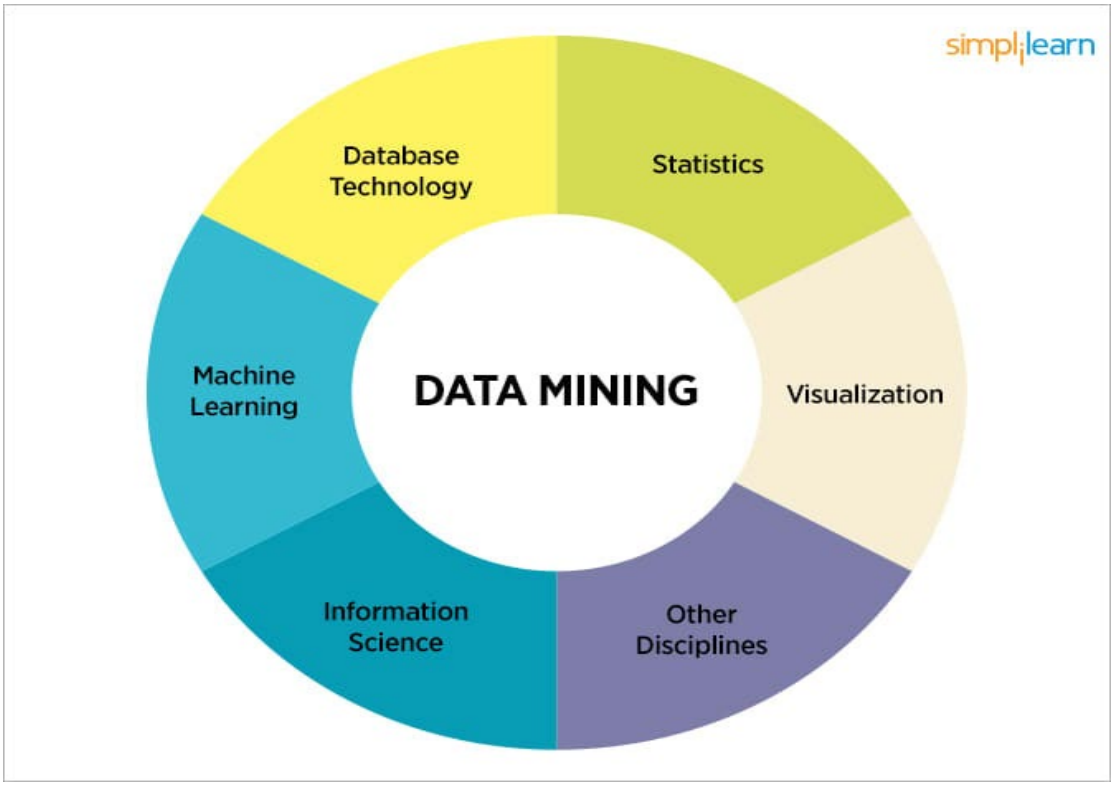
Αν και υπάρχουν εργασίες που ήδη έχουν διερευνήσει το εν λόγω θέμα με την συγκεκριμένη εργασία στοχεύουμε στην δημιουργία μιας βάσης έρευνας σε έναν τομέα παθήσεων όπως είναι αυτός των καρδιακών ασθενειών όπου η χρησιμοποίηση των δεδομένων που είναι διαθέσιμα, λόγω και της προσπάθειας ψηφιοποίησης των διαδικασιών στον τομέα της Ιατρικής αλλά και όλης της σύγχρονης πραγματικότητας, μπορεί να φανεί άκρως σημαντική και να αλλάξει τον τρόπο αντιμετώπισης αυτών. Έτσι θα θέσουμε τη βάση για περαιτέρω ανάλυση και μελλοντική έρευνα.

ΜΕΡΟΣ Α’: ΘΕΩΡΗΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

Κεφάλαιο 1 *Εξόρυξη Δεδομένων Και Μηχανική Μάθηση*

1.1 Μια εισαγωγή στο Data Mining

Η εξόρυξη δεδομένων στοχεύει στην ανακάλυψη χρήσιμων προτύπων δεδομένων από τεράστιες ποσότητες δεδομένων. Σύμφωνα με το Παγκόσμιο Ινστιτούτο McKinsey (MGI) στις περισσότερες επιχειρήσεις των Ηνωμένων Πολιτειών της Αμερικής που ξεπερνούσαν τους 1000 υπαλλήλους σε αριθμό υπάρχουν περίπου 200 terabytes αποθηκευμένων δεδομένων. Το μέγεθος των δεδομένων που χρησιμοποιούνται και προσπελαύνονται καθημερινά αυξάνεται με έναν ταχύτατο ρυθμό στην σύγχρονη πραγματικότητα. Τα δεδομένα αυτά και η συλλογή αλλά και η ανάλυση τους προσδίδουν ευκαιρίες στις επιχειρήσεις για μείωση κόστους αλλά και βελτίωσης των διαδικασιών τους. Σύμφωνα με έρευνες του MIT η εκλογή του προέδρου Obama στην Αμερική το 2012 επετεύχθη με την σωστή εφαρμογή μεθόδων εξόρυξης δεδομένων για την διερεύνηση δεδομένων όπως το ποιος ήταν πιθανότερο να τον ψηφίσει αλλά και την πρόβλεψη των αποτελεσμάτων από περιφέρεια σε περιφέρεια. Η πρόβλεψη μάλιστα που έκαναν με τη χρήση αυτών των μεθόδων ήταν περί το 56,4% των ψήφων, ενώ το τελικό πραγματικό αποτέλεσμα ήταν 56,6%. Όπως φαίνεται μια τόσο ισχυρή δυνατότητα πρόβλεψης επέτρεψε στο επιτελείο του προέδρου να διαχωρίσει και να καταμερίσει τους πιο αναγκαίους πόρους με μεγαλύτερη ακρίβεια και αποδοτικότητα. Σε ένα ακόμα παράδειγμα, στην West Coast Bank of America επί αρκετά χρόνια οι πελάτες που καλούσαν στο τηλεφωνικό κέντρο άκουγαν την ίδια διαφήμιση και προωθητική ενέργεια που θα άκουγε οποιοσδήποτε καλούσε εκείνη την στιγμή ανεξαρτήτως προτιμήσεων και προσωπικών απαιτήσεων. Η τράπεζα όμως θέλοντας να είναι όσο το δυνατόν πιο κοντά στα θέλω του κάθε πελάτη προχώρησε στην ανάλυση των προφίλ τους ξεχωριστά με σκοπό μια πιο στοχευμένη και συνυφασμένη με τα ενδιαφέροντα του εκάστοτε πελάτη προωθητική ενέργεια. Μια τέτοια ενέργεια αντικατοπτρίζει πλήρως μια ενέργεια εξόρυξης δεδομένων με σκοπό την επιλογή της κατάλληλης στρατηγικής marketing για κάθε περίπτωση. *Τι είναι λοιπόν η εξόρυξη δεδομένων;* (Ye, 2017)



Εικόνα 1.1 Έννοιες Data Mining

1.2 Ορισμός Εξόρυξης Δεδομένων

Με απλά λόγια, η εξόρυξη δεδομένων ορίζεται ως μια διαδικασία που χρησιμοποιείται για την εξαγωγή χρήσιμων δεδομένων από ένα μεγαλύτερο σύνολο τυχόν ακατέργαστων δεδομένων. Υπονοεί την ανάλυση των προτύπων δεδομένων σε μεγάλες παρτίδες δεδομένων χρησιμοποιώντας ένα ή περισσότερα λογισμικά. Η εξόρυξη δεδομένων έχει εφαρμογές σε πολλούς τομείς, όπως η επιστήμη και η έρευνα. Ως εφαρμογή εξόρυξης δεδομένων, οι επιχειρήσεις μπορούν να μάθουν περισσότερα για τους πελάτες τους και να αναπτύξουν πιο αποτελεσματικές στρατηγικές που σχετίζονται με διάφορες επιχειρηματικές λειτουργίες και με τη σειρά τους να αξιοποιήσουν τους πόρους με έναν πιο βέλτιστο και διορατικό τρόπο. Αυτό βοηθά τις επιχειρήσεις να είναι πιο κοντά στον στόχο τους και να λαμβάνουν καλύτερες αποφάσεις. Η εξόρυξη δεδομένων περιλαμβάνει αποτελεσματική συλλογή και αποθήκευση δεδομένων καθώς και επεξεργασία υπολογιστή. Για την τμηματοποίηση των δεδομένων και την αξιολόγηση της πιθανότητας μελλοντικών γεγονότων, η εξόρυξη δεδομένων χρησιμοποιεί εξελιγμένους μαθηματικούς αλγόριθμους. Η εξόρυξη δεδομένων είναι επίσης γνωστή ως ανακάλυψη γνώσης στα δεδομένα (Knowledge Discovery in Databases - KDD). (ΠΑΠΑΘΑΝΑΣΙΟΥ, 2019)

1.3 Παράγοντες εξέλιξης των δεδομένων

Η απότομη ανάπτυξη της ανάγκης για την εφαρμογή μεθόδων εξόρυξης δεδομένων συνάδει με ένα εύρος σχετικών παραγόντων που συνέβαλαν σε αυτή.

i. Η αποθήκευση των δεδομένων σε αποθήκες δεδομένων με σκοπό την εύκολη πρόσβαση από οποιοδήποτε μέλος του οργανισμού
ii. Η διαθεσιμότητα του Ίντερνετ και Ίντρανετ για την προσπέλαση των δεδομένων
iii. Η ανταγωνιστικότητα στην παγκόσμια αγορά για την αύξηση του μεριδίου αυτής στο πλαίσιο της παγκόσμιας οικονομίας
iv. Η ανάπτυξη διαφόρων προγραμμάτων για την εφαρμογή μεθόδων
v. Η τεράστια ανάπτυξη της υπολογιστικής αλλά και αποθηκευτικής ισχύος με την ραγδαία εξέλιξη της τεχνολογίας

Η αναφορά του McKinsey προέβλεψε την έλλειψη του ανθρώπινου δυναμικού με τις κατάλληλες ικανότητες για την εκμετάλλευση των δυνατοτήτων των μεγάλων δεδομένων και συνεπώς και την τεράστια έξαρση των διαθέσιμων θέσεων εργασίας σε ρόλους που αφορούν στην διαχείριση των δεδομένων. (ΠΑΠΑΘΑΝΑΣΙΟΥ, 2019)

1.4 Ροή ανάλυσης των δεδομένων

Η διαδικασία της ανακάλυψης της γνώσης στα δεδομένα (KDD) αποτελείται από 5 στάδια, ένα εκ των οποίων είναι και η εξόρυξη δεδομένων. Από κάθε στάδιο απορρέει ένα αποτέλεσμα το οποίο οδηγεί στο επόμενο στάδιο.

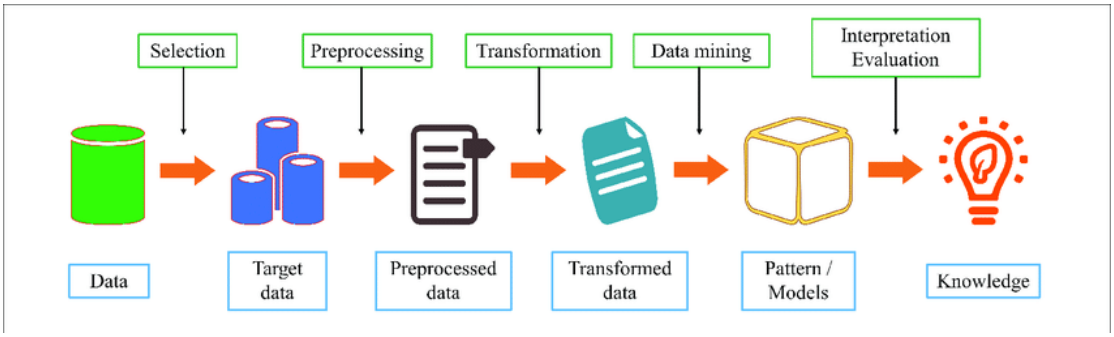
Στο *1ο στάδιο* έχουμε την συγκέντρωση των δεδομένων και των διαχωρισμό τους σε 2 μέρη εκ των οποίων στο ένα θα πραγματοποιηθεί η εξόρυξη. (train and test data)

Στο *2ο στάδιο* έχουμε την προεπεξεργασία των δεδομένων όπου γίνεται ο καθαρισμός τους και η προετοιμασία τους για να μπορούν να χρησιμοποιηθούν ορθά και αξιόπιστα δεδομένα.

Στο *3ο στάδιο* έχουμε την επεξεργασία των υπαρχόντων δεδομένων που λαμβάνουμε από το προηγούμενο στάδιο και η μετατροπή τους με σκοπό την εξόρυξη στο επόμενο βήμα. Επιπλέον επιλέγουμε και τον ή τους αλγορίθμους που θα χρησιμοποιήσουμε για την εξόρυξη.

Στο *4ο στάδιο* πραγματοποιείται η εφαρμογή των επιλεγμένων αλγορίθμων και η εξόρυξη των δεδομένων. Ταυτόχρονα γίνεται και πιθανή οπτικοποίηση των εξαγόμενων αποτελεσμάτων από τον κάθε αλγόριθμο.

Στο *5ο και τελευταίο στάδιο* έχουμε την κριτική ανάλυση των αποτελεσμάτων που απορρέουν από την εφαρμογή των τεχνικών της εξόρυξης καθώς και τα συμπεράσματα στα οποία καταλήγουμε, καθώς και πιθανές προτάσεις για περαιτέρω έρευνα μελλοντικά.



Εικόνα 1.2 Data Mining Flow

Οι τεχνικές εξόρυξης δεδομένων (*Data Mining*) και μηχανικής μάθησης (*Machine Learning*) είναι ευρέως διαδεδομένες στην σημερινή πραγματικότητα όπου υπάρχει τεράστιο πλήθος δεδομένων από τα οποία μπορούμε να λάβουμε σημαντικές πληροφορίες και εφαρμόζονται σε όλους τους τομείς από τον επιχειρηματικό κλάδο μέχρι και την Ιατρική όπου μπορεί να έχει πολύ σημαντικά αποτελέσματα. (Kantardzic, 2019)

1.5 Μέθοδοι εξόρυξης των δεδομένων

Οι μέθοδοι εξόρυξης των δεδομένων χωρίζονται σε κάποιες βασικές κατηγορίες σύμφωνα με τους στόχους τους οποίους προσπαθούν να επιτύχουν.

Περιγραφή (Description)

Κάποιες περιπτώσεις ανάλυσης των δεδομένων αποσκοπούν στην περιγραφή και την παρατήρηση δεδομένων που επαναλαμβάνονται και δημιουργούν μοτίβα και συνήθειες αλλά και στην ανάλυση και εξήγηση αυτών. Αυτές οι περιπτώσεις ανάλυσης κατατάσσονται στην κατηγορία της περιγραφής. Τα μοντέλα σε τέτοιες περιπτώσεις πρέπει να είναι όσο πιο κατανοητά και ξεκάθαρα γίνεται.

Εκτίμηση (Estimation)

Στην εκτίμηση ουσιαστικά προσπαθούμε να υπολογίσουμε την αξία μιας κεντρικής μεταβλητής που έχουμε επιλέξει σύμφωνα με διάφορες άλλες ποσοτικές ή κατηγορικές μεταβλητές. Τα μοντέλα δημιουργούνται έτσι ώστε να μας παρέχουν την αξία τόσο της μεταβλητής – στόχου όσο και των μεταβλητών πρόβλεψης (predictors) και στην συνέχεια οποιαδήποτε εκτίμηση κάποιας νέας παρατήρησης γίνεται με την χρήση αυτών των predictors.

Ταξινόμηση (Classification)

Η ομαδοποίηση είναι μια παρόμοια μέθοδος με την εκτίμηση με την μόνη διαφορά πως η μεταβλητή – στόχος είναι κατηγορική και όχι αριθμητική. Εδώ η κατηγορική μεταβλητή χωρίζεται σε κάποιες ομάδες – κατηγορίες, όπως για παράδειγμα για το εισόδημα μπορούμε να δημιουργήσουμε ομάδες όπως υψηλό/μεσαίο/χαμηλό εισόδημα και να τα διαχωρίσουμε σε αυτές ορίζοντας τα όρια της κάθε μιας. Έτσι ύστερα το μοντέλο εξόρυξης εξετάζει μεγάλα πλήθη δεδομένων με βάση αυτές τις κατηγορίες και προσδίδει αποτελέσματα σύμφωνα με τη σχετικότητα μεταβλητών με αυτή την μεταβλητή – στόχο και ύστερα κατηγοριοποιεί κάθε νέο δεδομένο σε αυτές τις κατηγορίες σύμφωνα με τα κριτήρια που έχει δημιουργήσει.

Πρόβλεψη (Prediction)

Η πρόβλεψη είναι μια μέθοδος παρόμοια με τις προηγούμενες δύο που αναφέρθηκαν. Η μόνη και κύρια διαφορά τους είναι πως τα αποτελέσματα της συγκεκριμένης μεθόδου σχετίζονται με το μέλλον. Όπως και η πρόβλεψη έτσι και οι μέθοδοι της εκτίμησης και της ταξινόμησης μπορούν υπό προϋποθέσεις να χρησιμοποιηθούν για πρόβλεψη. Εδώ εμπεριέχονται όλες οι κλασικές στατιστικές μέθοδοι όπως η γραμμική παλινδρόμηση (linear regression), συσχέτιση (correlation), αλλά και μέθοδοι εξόρυξης γνώσης και δεδομένων όπως KNN, δέντρα αποφάσεων (decision trees) και νευρωνικά δίκτυα (neural networks).

Συσταδοποίηση (Clustering)

Η συσταδοποίηση αναφέρεται ουσιαστικά στην ομαδοποίηση των παρατηρήσεων και των δεδομένων σε κλάσεις (clusters). Οι κλάσεις είναι ομάδες που περιέχουν παρόμοια δεδομένα, δεδομένα δηλαδή που μοιράζονται κάποια κοινά χαρακτηριστικά. Η κύρια διαφορά με την ταξινόμηση είναι πως στην συσταδοποίηση δεν υπάρχει κάποια μεταβλητή – στόχος με την οποία επιτυγχάνεται κάποια πρόβλεψη. Στην προκειμένη περίπτωση η μέθοδος προσπαθεί να χωρίσει ολόκληρα τα δεδομένα σε ομοιογενείς υποομάδες όπου μεγιστοποιείται η ομοιογένεια των δεδομένων και ελαχιστοποιείται η σχετικότητα με δεδομένα εκτός αυτών. Πολλές φορές η συσταδοποίηση χρησιμοποιείται ως εισαγωγική μέθοδος για την ανάλυση δεδομένων όπου αργότερα εφαρμόζεται και άλλη μέθοδος εξόρυξης όπως τα νευρωνικά δίκτυα.

Συσχέτιση (Association)

Η μέθοδος της συσχέτισης για την εξόρυξη δεδομένων αποσκοπεί στην εύρεση όλων εκείνων των χαρακτηριστικών των δεδομένων που σχετίζονται μεταξύ τους. Τα κοινά σημεία δηλαδή μεταξύ των παρατηρήσεων. Έτσι δημιουργούνται κανόνες οι οποίοι χαρακτηρίζουν τη σχέση μεταξύ των χαρακτηριστικών των μεταβλητών οι οποίοι είναι γνωστοί και ως κανόνες σχετικότητας (association rules). Οι πιο γνωστοί και συχνά εμφανιζόμενοι αλγόριθμοι αυτής της μεθόδου είναι οι A-priori και Generalized Rule Induction (GRI). (ΠΑΠΑΘΑΝΑΣΙΟΥ, 2019)

Κεφάλαιο 2 *Προεπεξεργασία Δεδομένων & Μοντέλα Μηχανικής Μάθησης*

2.1 Προ επεξεργασία δεδομένων (Data Preprocessing)

Η προ επεξεργασία των δεδομένων είναι μια διαδικασία που συμβαίνει πριν οποιαδήποτε ενέργεια εξόρυξης των δεδομένων. Αυτό συμβαίνει γιατί όταν ο οποιοσδήποτε αναλυτής παραλαμβάνει τα δεδομένα από μια βάση δεδομένων αυτά βρίσκονται σε μια μορφή ημιτελή και ακατέργαστη που μπορεί να περιέχει μέσα «θόρυβο», δεδομένα δηλαδή που να είναι είτε αχρείαστα είτε μη αντιπροσωπευτικά για το σύνολο. Τέτοια δεδομένα μπορεί να είναι κενά (nulls), ακραίες τιμές (outliers), μη λογικές τιμές, ή και τιμές που είναι αχρείαστες. Όλα αυτά λοιπόν θα πρέπει να φιλτραριστούν και να διαμορφωθούν ούτως ώστε να έχουμε αποτελέσματα στα επόμενα στάδια που θα είναι αντιπροσωπευτικά και κοντά στην πραγματικότητα. Αυτό γίνεται με την διαμόρφωση και τον καθαρισμό των δεδομένων (data transformation and data cleaning). (Kantardzic, 2019)

2.1.1 Καθαρισμός των δεδομένων (data cleaning)

Κατά τον καθαρισμό των δεδομένων ελέγχουμε τα δεδομένα για κάποια χαρακτηριστικά που μπορεί να εμφανίζουν. Όπως προαναφέραμε συνήθως ελέγχουμε για κενές τιμές, για ακραίες τιμές, για μη λογικές τιμές ή και για τιμές που μπορεί να μην είναι χρήσιμες γενικότερα.

- Κενές τιμές (Null)

  Τα κενά είναι το πιο συνηθισμένο χαρακτηριστικό που καλείται να αντιμετωπίσει κάποιος που διαχειρίζεται δεδομένα. Όσο και αν προχωράνε και εξελίσσονται οι μέθοδοι που χρησιμοποιούμε και τα μέσα που τα εφαρμόζουμε δεν παύουν, ιδιαίτερα τα δεδομένα που έρχονται σε μεγάλο όγκο, να χρειάζονται προετοιμασία ώστε να έρθουν πάντα στην επιθυμητή μορφή. Στα μεγάλα δεδομένα είναι μεγάλη η πιθανότητα να υπάρχουν κενές τιμές. Ο τρόπος που διαχειριζόμαστε αυτά τα δεδομένα ποικίλει. Το πιο εύκολο που θα μπορούσε να γίνει είναι απλά να αφαιρέσουμε τα συγκεκριμένα δεδομένα από την ανάλυση και το υπόλοιπο σύνολο. Κάτι τέτοιο όμως μπορεί να αποφέρει τα αντίθετα αποτελέσματα από αυτά που προσπαθούμε να επιτύχουμε καθώς μπορεί η παράλειψη αυτή στα δεδομένα μπορεί να είναι συστηματική και να παρουσιάζει μια περιοδικότητα που να μπορεί να εκτιμηθεί και να μελετηθεί. Έτσι λοιπόν βλέπουμε πως δεν μπορούμε απλώς να αγνοήσουμε τέτοιες περιπτώσεις. Έτσι οι μελετητές έχουν καταλήξει στο συμπέρασμα πως το καλύτερο είναι να αντικατασταθούν αυτά τα κενά χρησιμοποιώντας διάφορα κριτήρια. Τα πιο συνηθισμένα κριτήρια είναι:

- ✓ Αντικατάσταση των κενών με τον μέσο όρο της μεταβλητής
- ✓ Αντικατάσταση με μια τυχαία τιμή μέσα από το εύρος της μεταβλητής
- ✓ Αντικατάσταση με μια σταθερά που ορίζει ο αναλυτής
- ✓ Αντικατάσταση με τιμές υπολογισμένες σύμφωνα με άλλα χαρακτηριστικά της μεταβλητής

Από τα παραπάνω η πιο αντιπροσωπευτική είναι η τελευταία, αλλά είναι και η πιο χρονοβόρα καθώς και η πιο δύσκολη περίπτωση να επιτευχθεί καθώς χρειάζεται αρκετά μεγαλύτερη ανάλυση.

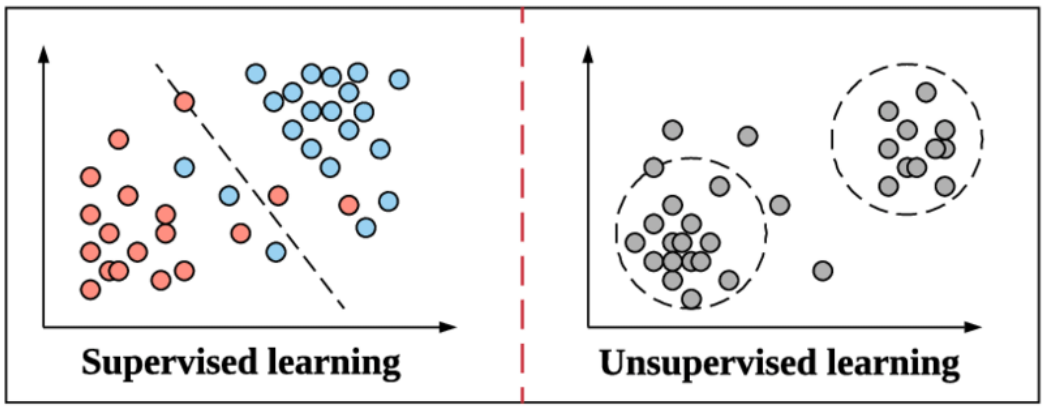
- Ακραίες τιμές (Outliers)
  Τα outliers είναι τιμές αρκετά ακραίες που εν τέλει επηρεάζουν την περιοδικότητα και την τάση των δεδομένων. Τέτοιες τιμές είναι σημαντικό να αναγνωρίζονται καθώς μπορεί να αποτελούν λάθη που πρέπει να υποστούν επεξεργασία. Ακόμα όμως και αν δεν αποτελούν λάθη οι ακραίες τιμές επηρεάζουν αρνητικά τις μεθόδους που εφαρμόζονται αργότερα και καλό είναι να αντιμετωπίζονται. Αυτό γίνεται είτε γραφικά, μέσω οπτικοποίησης με χρήση συγκεκριμένων γραφημάτων, είτε αριθμητικά, με χρήση κατάλληλων μεθόδων. Γραφικά μπορεί να επιτευχθεί είτε με χρήση ιστογράμματος (histogram) είτε με θηκόγραμμα (boxplot). Αριθμητικά χρησιμοποιούνται μέθοδοι όπως η Z-score και η IQR (interquartile range).
- Άλλες περιπτώσεις
  Κατά την προεπεξεργασία μπορεί να χρειαστεί να κάνουμε και κάποιες άλλες ενέργειες σε άλλες περιπτώσεις δεδομένων. Τέτοιες διεργασίες μπορεί να είναι η αφαίρεση μη λογικών η επαναλαμβανόμενων τιμών, η ομαλοποίηση των δεδομένων, η μετατροπή αριθμητικών δεδομένων σε κατηγορικά και αντίστροφα κ.α., σύμφωνα πάντα με το τι εξυπηρετεί την ανάλυση που προσπαθούμε να επιτύχουμε. (Kantardzic, 2019)

2.2 Μοντέλα μηχανικής μάθησης κατά την εξόρυξη δεδομένων

2.2.1 Υποκατηγορίες μοντέλων

Η εξόρυξη των δεδομένων χωρίζεται σε δύο βασικές υποκατηγορίες εργασιών που εφαρμόζονται, την επιβλεπόμενη (supervised learning) και την μη επιβλεπόμενη μάθηση (non-supervised learning). Και στις 2 περιπτώσεις έχουμε τα δεδομένα εισαγωγής (input data) που είναι τα αρχικά δεδομένα έπειτα από την προεπεξεργασία και τα εξαγόμενα δεδομένα (output data) που αφορούν τα τελικά μας αποτελέσματα. Η κύρια διαφορά μεταξύ των δύο είναι πως στην επιβλεπόμενη μάθηση χρησιμοποιούμε μια βασική αλήθεια, γνωρίζουμε δηλαδή εκ των προτέρων ποια είναι τα αποτελέσματα στα οποία πρέπει να καταλήξουμε. Έτσι ο στόχος είναι να καταλήξουμε σε μια μέθοδο όπου δοθέντος ενός δείγματος δεδομένων και κάποιων επιθυμητών αποτελεσμάτων μπορεί να μας δώσει με την μεγαλύτερη δυνατή ακρίβεια

την σχετικότητα των δεδομένων εισαγωγής και αποτελεσμάτων. Στην περίπτωση της μη επιβλεπόμενης μάθησης



Εικόνα 2.1 Μορφές Machine Learning

2.2.1.1 Επιβλεπόμενη μάθηση (supervised learning)

Στην κατηγορία του supervised learning χρησιμοποιούνται συνήθως οι μέθοδοι του classification και του regression, όταν θέλουμε την αντιστοίχιση των δεδομένων εισαγωγής με τα αποτελέσματα είτε με κάποιες ετικέτες για τα δεδομένα είτε με μια συνεχή μεταβλητή εξόδου. Οι πιο συνηθισμένοι αλγόριθμοι των μεθόδων που χρησιμοποιούνται περιλαμβάνουν την λογιστική παλινδρόμηση (logistic regression), την μέθοδο Naive Bayes, την μέθοδο support vector machines (SVM), την random forest και τα νευρωνικά δίκτυα (neural networks). Και στις δύο περιπτώσεις του classification και του regression ο κύριος στόχος είναι η ανακάλυψη συσχετίσεων και δομών στα δεδομένα εισαγωγής που μας βοηθούν να έχουμε σωστά και επιθυμητά αποτελέσματα. Τα αποτελέσματα αυτά βέβαια παρότι θεωρούνται σωστά μέσα από τον αλγόριθμο δεν σημαίνει πως πάντα θα ανταποκρίνονται σε πραγματικές καταστάσεις, καθώς οι αλγόριθμοι βασίζονται στο training data set εξολοκλήρου. Έτσι μπορεί να υπάρξει «θόρυβος» στα δεδομένα των αποτελεσμάτων ο οποίος σαφώς θα μειώσει και την αποτελεσματικότητα του μοντέλου.

Κατά την εφαρμογή της επιβλεπόμενης μάθησης τα κύρια σημεία που πρέπει να επικεντρωθούμε είναι η πολυπλοκότητα του μοντέλου και η ελαχιστοποίηση των σφαλμάτων. Η *πολυπλοκότητα* του μοντέλου αναφέρεται στην πολυπλοκότητα της συνάρτησης που προσπαθούμε να κάνουμε να μάθει. Το κατάλληλο επίπεδο πολυπλοκότητας καθορίζεται συνήθως από το training μέρος του dataset. Στις περιπτώσεις που έχουμε μικρό όγκο δεδομένων ή δεδομένα τα οποία δεν είναι σωστά κατανεμημένα και δομημένα θα πρέπει να επιλέξουμε κάποιο μοντέλο χαμηλής πολυπλοκότητας. Αυτό γιατί σε περίπτωση που επιλέξουμε ένα πιο πολύπλοκο μοντέλο τότε θα έχουμε overfitting. Το overfitting αναφέρεται στην περίπτωση όπου ουσιαστικά το μοντέλο υπερκαλύπτει τις ανάγκες των δεδομένων και δεν μπορεί να γενικεύσει τα αποτελέσματα για τα υπόλοιπα δεδομένα. Στην πραγματικότητα δηλαδή καταλήγει σε αποτελέσματα μέσω της χρήσης των training δεδομένων, αλλά τα

αποτελέσματα δεν μπορούν να μας υποδείξουν χαρακτηριστικά όπως η τάση για τα υπόλοιπα δεδομένα. Η διαφορά-διακύμανση στην ελαχιστοποίηση των σφαλμάτων (bias-variance tradeoff) σχετίζεται άμεσα με τη γενίκευση των μοντέλων. Σε οποιοδήποτε μοντέλο υπάρχει ισορροπία μεταξύ της διαφοράς (bias) και της διακύμανσης (variance). Με την διαφορά αναφερόμαστε στα σταθερά σφάλματα (constant errors) και με τη διακύμανση στην διαφορά ουσιαστικά που μπορεί να έχει το σφάλμα από το ένα training dataset στο άλλο. Έτσι αν έχουμε υψηλή διαφορά και χαμηλή διακύμανση θα έχουμε για παράδειγμα ένα ποσοστό λάθους 15% ενώ αν έχουμε χαμηλή διαφορά και υψηλή διακύμανση θα έχουμε ένα περιθώριο λάθους από 10% έως 50%. Συνήθως αυτά τα δύο κινούνται και καθαυτό τον τρόπο δηλαδή αυξάνεται το ένα και μειώνεται το άλλο. Κατά την εξέταση των δεδομένων μπορούμε να δούμε αν θα βρισκόμαστε σε αυτή την κατηγορία και κατά πόσο θα πρέπει να τείνουμε προς μία κατεύθυνση ανάλογα με τα αποτελέσματα που επιδιώκουμε. Συνήθως στα μικρότερα dataset έχουμε μοντέλα χαμηλότερης διακύμανσης ενώ στα μεγαλύτερα μοντέλα υψηλότερης διακύμανσης. (Larose and Larose, 2014)

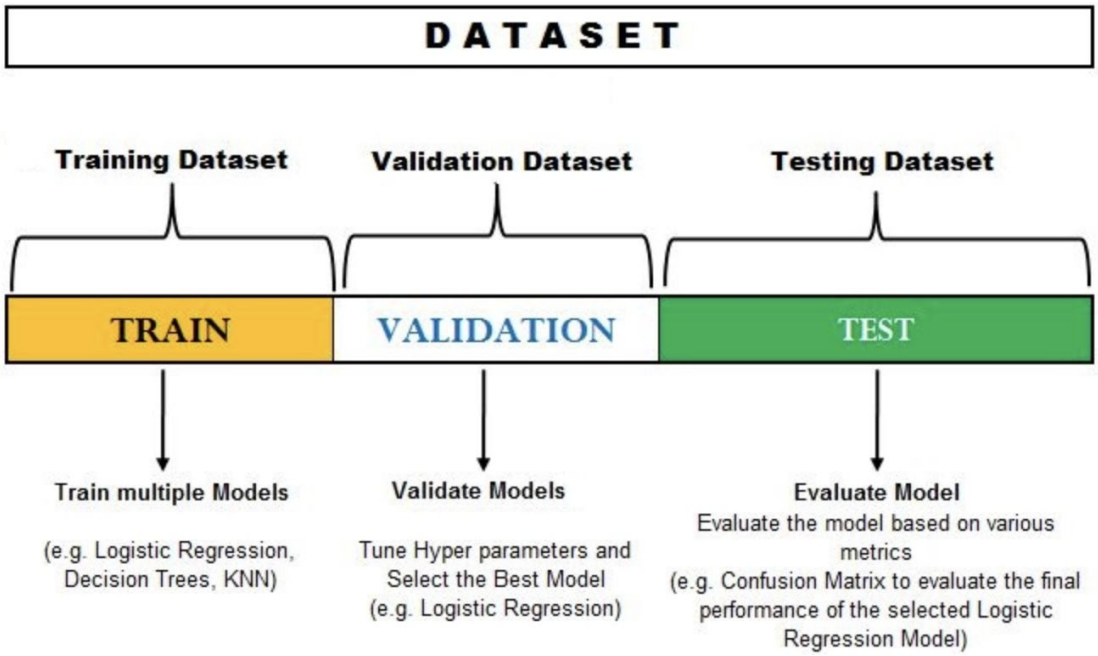
2.2.2.2 Μη επιβλεπόμενη μάθηση (unsupervised learning)

Οι συνηθέστεροι αλγόριθμοι που συναντάμε σε περιπτώσεις unsupervised learning είναι η συσταδοποίηση (clustering), η μάθηση χαρακτηριστικών (feature learning) και η εκτίμηση της πυκνότητας (density estimation). Σε όλες αυτές τις περιπτώσεις προσπαθούμε να εκτιμήσουμε την φυσική δομή των δεδομένων μας χωρίς να προσδίδουμε συγκεκριμένα χαρακτηριστικά και ετικέτες και για αυτό τον λόγο δεν μπορούμε να έχουμε και ακριβή εκτίμηση της επίδοσης των αλγορίθμων μη επιβλεπόμενης μάθησης. Κάποιες συνηθισμένες περιστάσεις που μπορεί να χρειαστεί να εφαρμόσουμε μη επιβλεπόμενη μάθηση είναι η διερευνητική ανάλυση (exploratory data analysis) και η μείωση της διαστατικότητας (dimensionality reduction). Στην πρώτη περίπτωση χρησιμεύει καθώς μπορεί να προσδώσει άμεσα την δομή των δεδομένων. Σε περιστάσεις όπου δεν είναι εφικτό να κατανοήσουμε την τάση των δεδομένων η μη επιβλεπόμενη μάθηση μπορεί να μας παρέχει γνώση για διάφορες υποθέσεις. Επίσης μπορούμε να επιτύχουμε την μείωση των χαρακτηριστικών και των στηλών των δεδομένων με το dimensionality reduction για περιπτώσεις όπου δεν θέλουμε τεράστιο πλήθος δεδομένων. (Larose and Larose, 2014)

2.2.2 Διαχωρισμός δεδομένων

Για την εφαρμογή των μεθόδων και των μοντέλων απαιτείται ο διαχωρισμός των δεδομένων σε δύο μέρη. Αυτά είναι το training set και το test set. Το training set εμπεριέχει τα δεδομένα τα οποία θα χρησιμοποιήσουμε ώστε να εκπαιδεύσουμε ουσιαστικά τους αλγορίθμους μας ώστε να ανακαλύψουμε τα διάφορα μοτίβα και τις τάσεις τους. Συνήθως το κομμάτι αυτό αποτελεί το 70% - 80% του αρχικού dataset. Το test set είναι το υπόλοιπο μέρος του αρχικού dataset το οποίο το κρατάμε και το χρησιμοποιούμε μετά την εφαρμογή των μοντέλων για την επικύρωση της ακρίβειας

των αποτελεσμάτων μας και για τη γενίκευση αυτών στο υπόλοιπο φάσμα των σχετικών δεδομένων. Όπως βλέπουμε λοιπόν ο διαχωρισμός αυτός είναι απαραίτητος προτού εφαρμοστεί οποιαδήποτε μέθοδος εξόρυξης.



Εικόνα 2.2 Διαχωρισμός Dataset

2.2.3 Ανάλυση Μεθόδων Μηχανικής Μάθησης

Στην συνέχεια θα αναλύσουμε το θεωρητικό υπόβαθρο των αλγορίθμων εκ των οποίων θα επιλέξουμε κάποιους να εφαρμόσουμε κατά την εξόρυξη δεδομένων στο δικό μας dataset που αφορά στα άτομα με καρδιακές παθήσεις. Οι αλγόριθμοι που θα εξερευνήσουμε είναι η λογιστική παλινδρόμηση (Logistic Regression), τα δέντρα αποφάσεων (Decision Trees), ο αλγόριθμος Naive Bayes, ο αλγόριθμος Random Forest και ο αλγόριθμος K-Nearest Neighbors (KNN). Θα δούμε τη θεωρητική και μαθηματική τους προσέγγιση καθώς και το πως εφαρμόζεται ο κάθε αλγόριθμος πρακτικά μέσα από το περιβάλλον της R με το οποίο θα υλοποιήσουμε το πρακτικό κομμάτι της εργασίας.

2.2.3.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Περιγραφή μεθόδου

Τα μοντέλα παλινδρόμησης καταγράφουν πώς μια ή περισσότερες μεταβλητές στόχοι ποικίλλουν με μία ή περισσότερες μεταβλητές χαρακτηριστικών. Μπορούν να χρησιμοποιηθούν για την πρόβλεψη των τιμών του μεταβλητές στόχου χρησιμοποιώντας τις τιμές των μεταβλητών χαρακτηριστικών.

Η παλινδρόμηση είναι μια στατιστική τεχνική που προηγήθηκε της μηχανικής μάθησης, αλλά χρησιμοποιείται επίσης ως μέρος πολλών αναλύσεων μηχανικής μάθησης. Η παλινδρόμηση δημιουργεί ένα μοντέλο που εξηγεί πώς ένα σύνολο ανεξάρτητων

μεταβλητών συμβάλλει σε μια δυαδική εξαρτώμενη μεταβλητή (αποτέλεσμα). (Ye, 2017)

Η *παλινδρόμηση* ορίζει μια εξίσωση με μεταβλητές ή χαρακτηριστικά που πιστεύεται ότι επηρεάζουν μια διακύμανση και προσπαθεί να προβλέψει τον καλύτερο συντελεστή για κάθε μεταβλητή για ένα σύνολο περιπτώσεων με γνωστό αποτέλεσμα. Εάν οι συντελεστές είναι στατιστικά σημαντικοί, η ίδια εξίσωση μπορεί να χρησιμοποιηθεί για την πρόβλεψη της αναταραχής για μια μεταβλητή με άγνωστο αποτέλεσμα.

Υπάρχουν 2 είδη λογιστικής παλινδρόμησης, η δυαδική και η πολυωνυμική.

- Η ***δυαδική*** είναι χρησιμότερη όταν επιθυμούμε να υπολογίσουμε την πιθανότητα ενός συμβάντος για μια κατηγορική μεταβλητή με δύο αποτελέσματα. Συνήθως σε τέτοιες μεταβλητές έχουμε περιπτώσεις ναι ή όχι.
- Η ***πολυωνυμική*** μπορεί να χρησιμοποιηθεί για την ταξινόμηση των παρατηρήσεων σε ομάδες με βάση μια κατηγορική σειρά μεταβλητών για την πρόβλεψη της συμπεριφοράς. Για παράδειγμα, αν πραγματοποιούσαμε μια έρευνα στην οποία οι συμμετέχοντες καλούνται να επιλέξουν ένα από τα πολλά ανταγωνιστικά προϊόντα ως το αγαπημένο τους θα μπορούσαμε να δημιουργήσουμε ένα προφίλ ατόμων που είναι πιο πιθανό να ενδιαφέρονται για κάποιο συγκεκριμένο προϊόν.

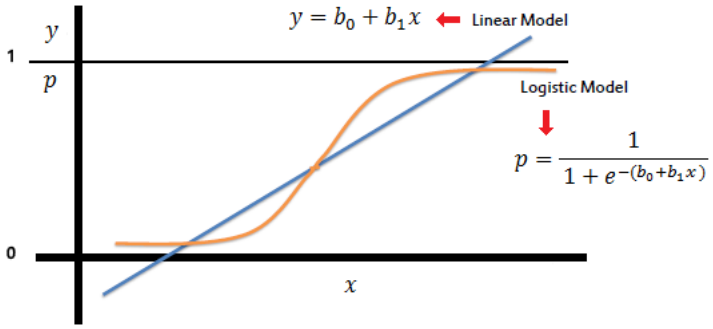
<u>Μαθηματικό υπόβαθρο</u>

- *Linear Model*

$$y = b_0 + b_1 x$$

- *Logistic Model*

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$



Εικόνα 2.3 Γραμμική και Λογιστική Παλινδρόμηση

Σύνοψη μεθόδου

Τα κύρια πλεονεκτήματα της είναι:

- Ευκολία στην εφαρμογή και την κατανοήση
- Λειτουργεί σε παραπάνω από μια κλάσεις
- Δεν επηρεάζεται από τις κατανομές των κλάσεων

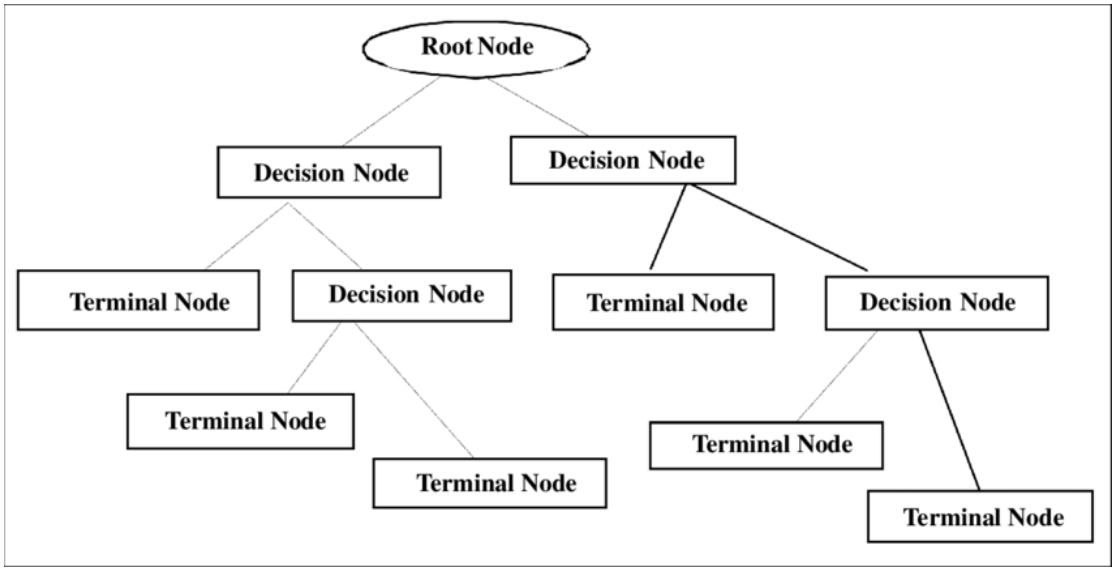
Τα κύρια μειονεκτήματα της είναι:

- Μπορεί να οδηγήσει σε overfitting αν οι παρατηρήσεις είναι λίγες σε αριθμό
- Βασίζεται στη γραμμικότητα και υποθέτει ότι υπάρχει
- Μπορεί να χρησιμοποιηθεί μόνο για διακριτές μεταβλητές

2.2.3.2 Δέντρα Αποφάσεων (Decision Trees)

Περιγραφή μεθόδου

Η μέθοδος των δέντρων αποφάσεων αποτελεί μια μέθοδο κατηγοριοποίησης (classification). Αποτελείται από την κατασκευή ενός τέτοιου δέντρου (διαγράμματος), που συντελείται από έναν αριθμό κόμβων απόφασης (decision nodes) που συνδέονται από κάποια κλαδιά (decision branches) μέχρι που ολοκληρώνονται σε κάποιους συγκεκριμένους κόμβους – φύλλα (leaf nodes). Ξεκινώντας από τον αρχικό κόμβο (root node) που κατά κανόνα τοποθετείται στην κορυφή του διαγράμματος του δέντρου, σε κάθε κόμβο απόφασης εξετάζονται τα χαρακτηριστικά των δεδομένων και κάθε πιθανό αποτέλεσμα δημιουργεί ένα κλαδί το οποίο με τη σειρά του καταλήγει είτε σε έναν νέο κόμβο απόφασης είτε σε έναν τελικό κόμβο – φύλλο. Με αυτόν τον τρόπο δημιουργείται ένα απλό δέντρο απόφασης (decision tree). (Ye, 2017)



Εικόνα 2.4 Μορφή Δέντρου Απόφασης

Με τα δέντρα αποφάσεων μπορούμε να κατηγοριοποιήσουμε και να προβλέψουμε κάποιες περιπτώσεις σύμφωνα με ένα σύνολο κανόνων αποφάσεων που ονομάζονται decision rules. Η βασική μεθοδολογία και οπτική που χρησιμοποιεί ο συγκεκριμένος αλγόριθμος είναι πως τα δεδομένα θα διαχωριστούν και θα οριστούν σε κατηγορίες όπου τα δεδομένα θα έχουν παρόμοια χαρακτηριστικά των μεταβλητών που θέλουμε να προβλέψουμε.

Για να χρησιμοποιηθούν τα δέντρα αποφάσεων όμως θα πρέπει να πληρούνται κάποια κριτήρια. Συγκεκριμένα:

I. Αρχικά όντας ένας αλγόριθμος επιβλεπόμενης μάθησης θα πρέπει να υπάρχουν δεδομένα που μπορούν να κατηγοριοποιηθούν. Θα πρέπει λοιπόν να έχουμε ορίσει ένα training set που θα περιέχει τις τιμές της μεταβλητής στόχου.
II. Επίσης θα πρέπει να υπάρχουν μεταβλητές στο training set που να έχουν μεγάλη έκταση αλλά και ποικιλομορφία έτσι ώστε να μπορεί ο αλγόριθμος να μην είναι προκατειλημμένος. Ο αλγόριθμος μαθαίνει χρησιμοποιώντας παραδείγματα οπότε αν τα παραδείγματα δεν είναι τα κατάλληλα τότε και τα αποτελέσματα θα είναι προβληματικά.
III. Τα χαρακτηριστικά για την μεταβλητή στόχο προς την ομαδοποίηση θα πρέπει να είναι διακριτά. Για παράδειγμα δεν μπορούμε να εφαρμόσουμε τον αλγόριθμο για συνεχείς μεταβλητές. Πρέπει οι τιμές να είναι διακριτές ώστε ο αλγόριθμος να μπορεί να εντοπίσει αν κάποιο δεδομένο ανήκει σε μια κλάση ή όχι.

Μαθηματικό υπόβαθρο

Ουσιαστικά ο συγκεκριμένος αλγόριθμος βασίζεται σε έναν μαθηματικό τύπο μέσω του οποίου και της επανάληψης αυτού επιτυγχάνεται η κατασκευή του δέντρου απόφασης. Πιο συγκεκριμένα:

$$Info(D) = -\sum^{m} p_i log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j)$$

Το info(D) αποκαλείται και εντροπία του D και δείχνει ουσιαστικά μια πληροφορία που εισέρχεται σε ένα σύνολο (συστάδα) για την ιδιότητα D και το pi αναφέρεται στην πιθανότητα κάποιων δεδομένων να ανήκουν σε μία συστάδα με την συγκεκριμένη ιδιότητα.

Η πληροφορία που λαμβάνεται δίνεται από τον απλό τύπο:

$$Gain(A) = Info(D) - Info_A(D)$$

Πολλές φορές κατά τη δημιουργία ενός δέντρου απόφασης μπορεί να έχουμε κάποια κλαδιά στα οποία να έχουμε κάποιες ανωμαλίες στο training dataset λόγω θορύβου ή και outliers. Αυτή η ασυνέπεια λύνεται με διάφορους τρόπους με τον πιο συνηθισμένο να είναι το λεγόμενο κλάδεμα (pruning). Με αυτό τον τρόπο το δέντρο απαλλάσσεται από τις περιττές πληροφορίες και γίνεται μικρότερο, απλούστερο και πολύ πιο εύκολα κατανοήσιμο. (Ye, 2017)

Σύνοψη μεθόδου

Για την πρακτική εφαρμογή της μεθόδου στην R χρησιμοποιούμε την βιβλιοθήκη rpart μέσω της οποίας μπορούμε με πολύ απλές εντολές να δημιουργήσουμε το δέντρο που θέλουμε καθώς και να το οπτικοποιήσουμε.

Τα κύρια πλεονεκτήματα της είναι:

- Είναι απλή στην εφαρμογή αλλά και την κατανόηση αυτής και των αποτελεσμάτων της.
- Διαχειρίζεται και κατηγορικά και αριθμητικά δεδομένα με την ίδια επιτυχία.
- Δεν χρειάζεται παραμετροποιήσεις.
- Είναι πολύ πιο ευέλικτη μέθοδος ως προς την προεπεξεργασία των δεδομένων.
- Είναι αρκετά γρήγορο και δεν χρειάζεται υψηλούς υπολογιστικούς πόρους.

Τα κύρια μειονεκτήματα της είναι:

- Μπορεί να επηρεαστεί από το overfitting που προαναφέραμε.
- Είναι αρκετά ασταθή καθώς μικρές αλλαγές μπορεί να επιφέρουν μεγάλες διαφορές αποτελεσμάτων.
- Δεν είναι εντελώς αντικειμενικά και τείνουν να ευνοούν σύνολα που είναι πιο πιθανό να προβλεφθούν.

2.2.3.3 Αλγόριθμος Naive Bayes

Περιγραφή μεθόδου

Ο αλγόριθμος αποτελεί μια τεχνική classification και βασίζεται στο θεώρημα του Bayes υποθέτοντας πως υπάρχει ανεξαρτησία μεταξύ των παραγόντων πρόβλεψης. Ο κατηγοριοποιητής (classifier) Bayes υποθέτει ουσιαστικά πως η ύπαρξη οποιουδήποτε χαρακτηριστικού σε μια κλάση είναι ανεξάρτητη από την ύπαρξη άλλων χαρακτηριστικών στην ίδια κλάση. Το θεώρημα του Bayes είναι μια πολύ σημαντική μέθοδος και παρόλο που είναι μια απλή μέθοδος πολλές φορές υπερβαίνει σε αποτελεσματικότητα αρκετούς πιο περίπλοκους αλγορίθμους. Το θεώρημα αυτό παρέχει την δυνατότητα να υπολογιστεί μιας πιθανότητας ενός χαρακτηριστικού δεδομένου της ύπαρξης ενός άλλου. Ανήκει στα Μπευσιανά δίκτυα (Bayesian Networks) που είναι διάφορες μέθοδοι κατηγοριοποίησης και αποτελεί μια σημαντική

μέθοδο κατηγοριοποίησης καθώς παρέχει μια πολύ ουσιώδη βάση και μπορεί να εξηγήσει περίπλοκες σχέσεις μεταξύ των δεδομένων. Ο αλγόριθμος χρησιμοποιεί μοντέλα που προσδίδουν ετικέτες σε διάφορα στιγμιότυπα προβλημάτων που χωρίζονται σε τομείς που λέγονται vectors και έχουν τιμές των χαρακτηριστικών. Δεν χρησιμοποιείται ένας μοναδικός αλγόριθμος αλλά μια ομάδα που εκπαιδεύει τους κατηγοριοποιητές. (Ye, 2017)

Μαθηματικό υπόβαθρο

Όπως προαναφέραμε λοιπόν ο αλγόριθμος βασίζεται στο θεώρημα του Bayes που στηρίζεται στην στατιστική επιστήμη και χρησιμοποιεί την υπό συνθήκη πιθανότητα ενός δεδομένου Α να ισχύει δεδομένου να ισχύει ένα άλλο ενδεχόμενο Β. η πιθανότητα αυτή δίνεται από τον τύπο:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Όπου P(A) και P(B) είναι οι πιθανότητες των ενδεχομένων Α και Β αντίστοιχα ενώ P(A|B) είναι η πιθανότητα να ισχύει το ενδεχόμενο Α δεδομένου ότι ισχύει το Β και P(B|A) είναι η πιθανότητα να ισχύει το ενδεχόμενο Β δεδομένου ότι ισχύει το Α. Έτσι με τη χρήση του παραπάνω τύπου ο αλγόριθμος καταλήγει στην κατηγοριοποίηση με τη χρήση του:

$$p(C_k|x) = \frac{p(C_k) * p(x|C_k)}{p(x)}$$

Όπου Ck είναι οι κλάσεις που θα δημιουργηθούν από την κατηγοριοποίηση με k τον αριθμό τους ενώ το x είναι ο τομέας (vector) που περιέχει τις ανεξάρτητες μεταβλητές $x_1,\ldots,x_n$.

Σύνοψη μεθόδου

Για την εφαρμογή της μεθόδου στην R χρησιμοποιήθηκε η ομώνυμη βιβλιοθήκη της naïve bayes.

Τα πλεονεκτήματα της μεθόδου είναι:

- Με την προϋπόθεση πως ισχύει η ανεξαρτησία των predictors έχει πολύ καλύτερα αποτελέσματα σε σύγκριση με άλλα μοντέλα κατηγοριοποίησης.
- Χρειάζεται μικρό μέρος training data.
- Είναι αρκετά απλός και εύκολος στη χρήση και την κατανόηση.
- Χρησιμοποιείται για δυωνυμικά και μη προβλήματα.
- Μπορεί να χρησιμοποιηθεί και για συνεχής και για διακριτές μεταβλητές χωρίς να επηρεάζεται σημαντικά από μικρές διακυμάνσεις.

Τα μειονεκτήματα της μεθόδου είναι:

Η ανάγκη για την ύπαρξη της προϋπόθεσης της ανεξαρτησίας. Το θεώρημα λαμβάνει ως δεδομένο πως όλα τα χαρακτηριστικά είναι ανεξάρτητα κάτι που δεν ισχύει πάντα σε προβλήματα της πραγματικότητας.

Αν υπάρχει κάποια κατηγορία μιας κατηγορικής μεταβλητής στο test dataset που δεν παρατηρήθηκε στο training dataset τότε το μοντέλο δεν θα μπορέσει να καταλήξει σε πρόβλεψη. Αυτό αναφέρεται και ως zero frequency και λύνεται με τεχνικές smoothing των δεδομένων.

2.2.3.4 Αλγόριθμος Random Forest

Περιγραφή μεθόδου

Η μέθοδος των random forest ουσιαστικά είναι πολύ κοντά με αυτή των decision trees καθώς κατά κύριο λόγο είναι η δημιουργία πολλών decision trees ταυτόχρονα. Ο αλγόριθμος δημιουργεί διάφορα decision trees τα οποία συμβάλλουν όλα στην τελική πρόβλεψη. Ουσιαστικά δημιουργεί ασυσχέτιστα δέντρα μεταξύ τους και εν τέλει παίρνει την ψήφο για την κατηγοριοποίηση από το κάθε δέντρο. Τα δέντρα λειτουργούν λοιπόν σαν μια ομάδα και παρόλο που είναι θεωρητικά ασυσχέτιστα καταλήγουν σε ένα αρκετά ισχυρό αποτέλεσμα. Η κύρια σκέψη πίσω από αυτή τη μέθοδο είναι ότι η ομάδα των δέντρων αυτών θα υπερισχύσει του κάθε ξεχωριστού δέντρου απόφασης. Αν κάποιο ή κάποια δέντρα έχουν κάποια λάθος πρόβλεψη είναι πιθανό πολλά άλλα να έχουν σωστή πρόβλεψη και εκεί στηρίζεται ο συγκεκριμένος αλγόριθμος, στην δύναμη των πολλών.

Για να εφαρμοστεί ο συγκεκριμένος αλγόριθμος θα πρέπει όμως να έχουμε χαρακτηριστικά που να έχουν χαρακτήρα και δυνατότητα πρόβλεψης. Επίσης θα πρέπει το κάθε δέντρο καθώς και η πρόβλεψη του να είναι ασυσχέτιστη με τα υπόλοιπα ή τουλάχιστον να έχει χαμηλό βαθμό συσχέτισης. Τα δεδομένα που θα επιλέξουμε και οι παράμετροι θα επηρεάσουν σε μεγάλο βαθμό την ύπαρξη της συσχέτισης και ως εκ τούτου τα αποτελέσματα του αλγορίθμου. (Ye, 2017)

Μαθηματικό υπόβαθρο

Στην ουσία η μαθηματική βάση του αλγορίθμου είναι η ίδια με των decision trees η οποία επαναλαμβάνεται κάποιες φορές και χρησιμοποιεί τον δείκτη gini για να επιλέξει τις προβλέψεις από κάθε δέντρο που δίνεται από τον τύπο:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Σύνοψη μεθόδου

Για την εφαρμογή της μεθόδου στην R χρησιμοποιήθηκε η βιβλιοθήκη randomforest.

Τα πλεονεκτήματα της συγκεκριμένης μεθόδου είναι:

- Μπορεί να εφαρμοστεί και σε προβλήματα ομαδοποίησης και σε προβλήματα παλινδρόμησης.
- Μπορεί να διαχειριστεί μεγάλο όγκο δεδομένων.
- Μεγιστοποιεί την ακρίβεια του μοντέλου και αποφεύγει το overfitting.

Τα μειονεκτήματα της συγκεκριμένης μεθόδου είναι:

- Πως παρόλο που μπορεί να εφαρμοστεί και σε προβλήματα classification και regression δεν έχει καλή εφαρμογή στην περίπτωση του regression.

2.2.3.5 K-Nearest Neighbors (KNN)

Περιγραφή μεθόδου

Η μέθοδος των k nearest neighbors είναι μια κατά κύριο λόγο μέθοδος classification. Έχουμε λοιπόν μια κατηγορική μεταβλητή στόχο η οποία χωρίζεται σε κατηγορίες. Στην συνέχεια ο αλγόριθμος θα ελέγξει ένα μεγάλο μέρος δεδομένων και βασιζόμενο στα χαρακτηριστικά τους θα διαχωρίσει τα δεδομένα στις κατηγορίες που έχει διαχωριστεί η μεταβλητή στόχος. Συγκεκριμένα η ροή που έχει μια μέθοδος classification είναι πως πρώτα ελέγχει τα δεδομένα (πρόβλεψης και μεταβλητής στόχου), στη συνέχεια ΄΄μαθαίνει΄΄ με βάση διάφορα χαρακτηριστικά και εν συνεχεία εφαρμόζει τους κανόνες που δημιούργησε από τα προηγούμενα δεδομένα και κατηγοριοποιεί τα νέα δεδομένα.

Ο αλγόριθμος KNN είναι από τους πιο συνηθισμένους στις μεθόδους classification. Η λογική του αλγορίθμου είναι πως υποθέτει ότι το δείγμα test θα έχει τις ίδιες ιδιότητες με το training δείγμα που βρίσκεται πιο κοντά σε αυτό. Ο KNN ανήκει στην κατηγορία lazy learners όπου σε αντίθεση με τα SVM και άλλα μοντέλα οι εκτιμήσεις γίνονται αφού συγκεντρωθούν όλα τα δεδομένα και όχι άμεσα. Το να εκπαιδευθεί ένας αλγόριθμος KNN χρειάζεται μόνο τον καθορισμό του k. Ο KNN χρησιμοποιεί τα k κοντινότερα πρότυπα χαρακτηριστικών για τη σύγκριση και τη κατηγοριοποίηση. Απλά αποθηκεύει και θυμάται όλα τα δείγματα από το training set και ύστερα συγκρίνει το test δείγμα με αυτά. Για αυτό και πολλές φορές αναφέρεται και ως αλγόριθμος μνημονικής μάθησης ή μάθησης στιγμιότυπων (memory based learning / instance based). Ο εν λόγω αλγόριθμος δεν χρειάζεται πολλούς πόρους για την εφαρμογή του καθώς μόνο αποθηκεύει τα δεδομένα και ύστερα τα συγκρίνει αλλά η κατηγοριοποίηση των δεδομένων στη συνέχεια χρειάζεται μεγαλύτερη υπολογιστική ισχύ. Πιο συγκεκριμένα ο αλγόριθμος καλείται να υπολογίσει όλες τις αποστάσεις των χαρακτηριστικών του τεστ με όλα τα δεδομένα που έχουμε, κάτι που σε μεγάλα δεδομένα είναι απαιτητικό. (Ye, 2017)

Μαθηματικό υπόβαθρο

Όπως προαναφέραμε λοιπόν ο αλγόριθμος κατηγοριοποιεί τα δεδομένα χρησιμοποιώντας την απόσταση ως κριτήριο για να εντοπίσει το πλησιέστερο training δείγμα στο χώρο των χαρακτηριστικών που έχουν επιλεχθεί. Έτσι για να εφαρμοσθεί ο αλγόριθμος θα πρέπει πρώτα να έχουν οριστεί οι κατηγορίες που θέλουμε μέσω του training δείγματος. Κάθε πλειάδα (tuple) του δείγματος έχει έναν n αριθμό χαρακτηριστικών. Έτσι κάθε φορά που ο αλγόριθμος δέχεται μια άγνωστη πλειάδα ένας KNN κατηγοριοποιητής αναζητά τα μοτίβα τα οποία είναι κοντινότερα στην άγνωστη αυτή πλειάδα. Οι k πλειάδες που βρίσκονται πιο κοντά είναι οι k-nearest neighbors της άγνωστης πλειάδας. Το πόσο κοντά βρίσκεται μια πλειάδα σε μια άλλη καθορίζεται μέσω μετρικών αποστάσεων όπως η ευκλείδια απόσταση. Ανάλογα τα δεδομένα μας μπορεί να χρησιμοποιηθεί και κάποιος άλλος τύπος για την απόσταση με τον πιο συνηθισμένο όμως να είναι αυτός της ευκλείδιας, ιδιαίτερα σε περιπτώσεις συνεχών μεταβλητών.

Μερικοί τύποι αποστάσεων:

i. Euclidean: $d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$
ii. Manhattan: $d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$
iii. Minkowski: $d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$

Σύνοψη μεθόδου

Τα πλεονεκτήματα της συγκεκριμένης μεθόδου είναι:

- Απλός και εύκολα κατανοητός αλγόριθμος όσον αφορά στην εφαρμογή και στην ερμηνεία
- Υψηλή αποτελεσματικότητα
- Εφαρμογή για σύνθετες και πολύπλοκες περιπτώσεις

Τα μειονεκτήματα της συγκεκριμένης μεθόδου είναι:

- Ανάγκη του αλγορίθμου να κρατάει στη μνήμη του όλα τα δεδομένα
- Δυσκολότερη επιλογή του k και της μεθόδου υπολογισμού της απόστασης

Κεφάλαιο 3 *Η R Και Το Περιβάλλον Του RStudio*

Η παρούσα εργασία πραγματοποιήθηκε με την χρησιμοποίηση της γλώσσας R και του περιβάλλοντος του RStudio. Στη συνέχεια θα έχουμε μια σύντομη αναφορά στα συγκεκριμένα εργαλεία.

3.1 Ιστορική αναδρομή της R

Η γλώσσα προγραμματισμού R κατασκευάστηκε από τους Ross Ihaka και Robert Gentleman, του πανεπιστημίου του Auckland στη Νέα Ζηλανδία. Υποστηρίζεται από τα R Core Team και R Foundation For Statistical Computing. Βασίστηκε στην στατιστική γλώσσα προγραμματισμού S από την TIBCO Software Inc και εμπνεύστηκε από την διάλεκτο Scheme. Η εμπορική ονομασία της ήταν S-PLUS και κυκλοφόρησε το 1988. Μεγάλο μέρος του κώδικα της χρησιμοποιείται μέχρι και σήμερα αναλλοίωτος μέσα στην R. Ο βασικός λόγος που δημιουργήθηκε η γλώσσα της R ήταν η ύπαρξη της ως ελεύθερης διανομής για ακαδημαϊκή χρήση (freeware).

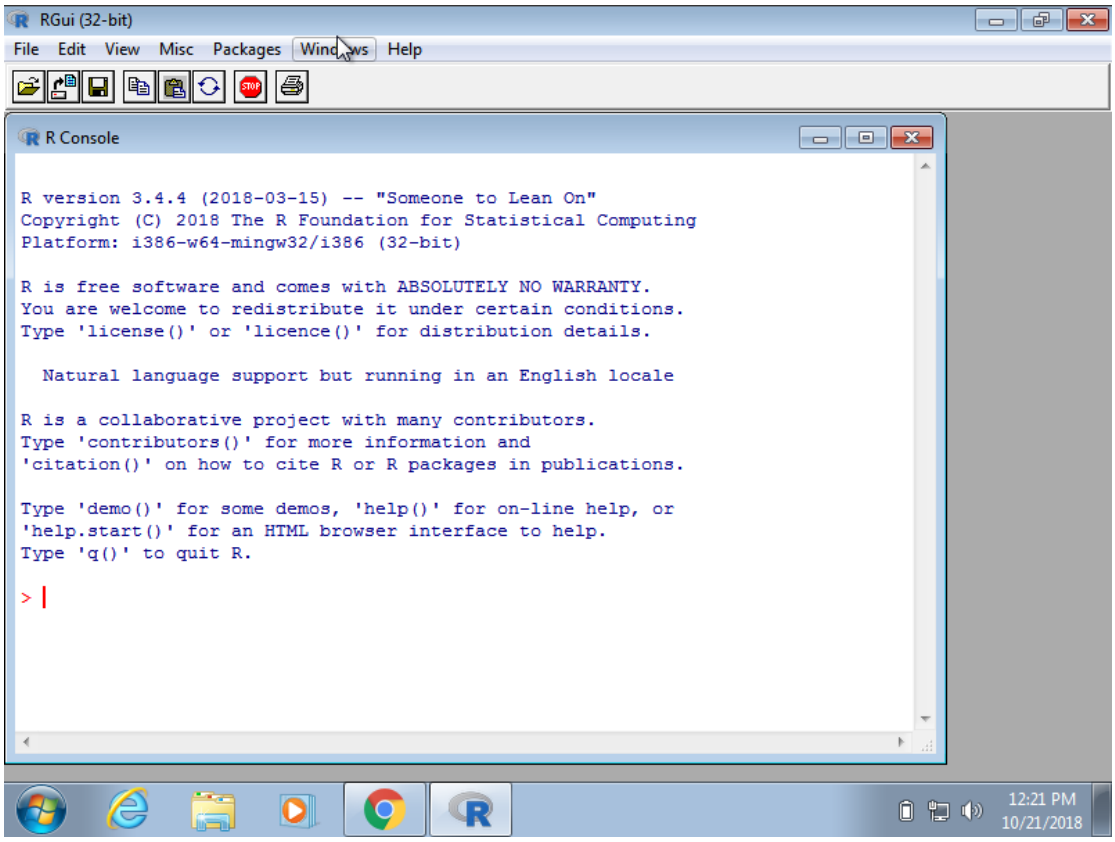
Εικόνα 3.1 R Logo

Η R θα μπορούσε να συγκριθεί με άλλα προγράμματα όπως το SAS και το SPSS. Το πρώτο είναι ένα από τα σημαντικότερα εργαλεία στον κόσμο των δεδομένων και χρησιμοποιείται από μια πληθώρα επιχειρήσεων ενώ το δεύτερο είναι ένα πολύ διαδεδομένο πρόγραμμα που επικεντρώνεται σε λειτουργίες στατιστικές. Κάποιοι από τους κύριους λόγους που η γλώσσα της R είναι κοινώς αποδεκτή και δημοφιλής είναι πως αρχικά προσδίδει την δυνατότητα για επεξεργασία των δεδομένων αλλά και στατιστική ανάλυση τους αλλά επιπλέον είναι μια εύκολη γλώσσα στην εκμάθηση και την κατανόηση. Επίσης είναι συμβατή με τα πιο διαδεδομένα λειτουργικά συστήματα όπως Windows, Linux και Mac OS καθώς και παρέχει πάρα πολλές δυνατότητες με την πρόσβαση σε πακέτα που δημιουργούνται από τους χρήστες που ονομάζονται βιβλιοθήκες (libraries) και τέλος είναι ένα δωρεάν software. Η τρέχουσα έκδοση του R (Οκτώβριος 2021) είναι η 4.1 και παρέχεται σε όλα τα προαναφερθέντα λειτουργικά συστήματα. Αποτελεί την πλέον διαδεδομένη μέθοδο ανάλυσης και εξόρυξης δεδομένων και εντάσσεται μέσα στις πλέον δημοφιλείς γλώσσες και βρίσκεται στις πρώτες 20 θέσεις προτίμησης για στατιστικό προγραμματισμό.(https://www.tiobe.com/tiobe-index/)

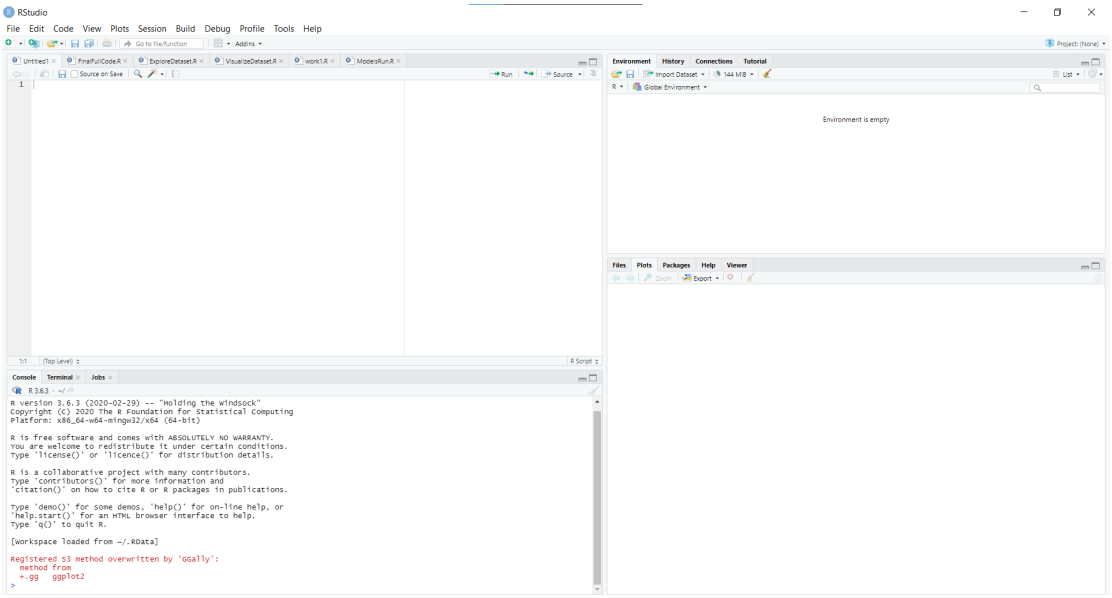
3.2 Περιβάλλον

Υπάρχουν διάφορες υλοποιήσεις της R. Κυρίως είναι γραμμένη σε C και Fortran. Το βασικό περιβάλλον της R είναι δημιουργημένο ως ένα πολύ απλό GUI(Graphical Users

Interface). Στην αρχή της R οι χρήστες χρησιμοποιούσαν το βασικό αυτό περιβάλλον που κυρίως εκτελούσε εντολές μέσω ενός command line αλλά με την πάροδο του χρόνου προτίμησαν την χρήση κάποιου ενσωματωμένου περιβάλλοντος IDE. Το IDE είναι ουσιαστικά ένα πιο φιλικό προς τον χρήστη περιβάλλον κάποιας γλώσσας προγραμματισμού και του παρέχει εύκολα και άμεσα πρόσβαση σε συνηθισμένες λειτουργίες και εργαλεία της γλώσσας. Λόγω της εκτεταμένης χρήσης της γλώσσας δημιουργήθηκαν διάφορα software IDE για την ευκολότερη χρήση της R. Το πιο γνωστό IDE για την R είναι το RStudio. (ΠΑΠΑΘΑΝΑΣΙΟΥ, 2019)



Εικόνα 3.2 R Console



Εικόνα 3.3 R General Environment View

Η R ως βασικό πακέτο παρέχει την δυνατότητα διάφορων μαθηματικών υπολογισμών αλλά και μετατροπών των δεδομένων. Επιπλέον εμπεριέχει κάποιες βασικές στατιστικές συναρτήσεις για υπολογισμό των μέτρων θέσης και διασποράς. Ακόμα έχει κάποιες βασικές λειτουργίες γραφημάτων. Τέλος δίνει την δυνατότητα ανάγνωσης δεδομένων από διάφορους τύπους αρχείων της excel, csv, txt κλπ. και πλέον έχει την δυνατότητα συνδεσιμότητας με άλλα εργαλεία για την οπτικοποίηση των αποτελεσμάτων και των γραφημάτων της είναι το PowerBI της Microsoft.

3.3 Βιβλιοθήκες (Libraries)

Το R προσφέρει της δυνατότητες για εξόρυξη δεδομένων με την βοήθεια των τεχνικών που περιεγράφηκαν προηγουμένων μέσα από την ένταξη κάποιων πακέτων – βιβλιοθηκών (libraries) τα οποία συνεχώς αναβαθμίζονται. Τα πακέτα αυτά φορτώνονται στην R μέσω απλών εντολών και παρέχουν της επιπλέον δυνατότητες. Τέτοια πακέτα μπορεί να προσδίδουν πολύπλοκους μαθηματικούς τύπους, συναρτήσεις, οπτικά διαγράμματα και αλγορίθμους.

ΜΕΡΟΣ Β’: ΠΡΑΚΤΙΚΗ ΑΝΑΛΥΣΗ

Κεφάλαιο 4 *Data Preprocess & Data Visualization*

Στο παρόν κεφάλαιο λοιπόν περνάμε στην εφαρμογή όσων είδαμε πριν σε θεωρητικό επίπεδο πάνω στο dataset που αφορά της καρδιακές παθήσεις. Αρχικά θα δούμε διερευνητικά τα δεδομένα με τα οποία θα ασχοληθούμε και εν συνεχεία θα προχωρήσουμε στην οπτικοποίηση της για περαιτέρω κατανόηση.

4.1 Σύντομη αναφορά της καρδιακές παθήσεις (heart diseases)

General Heart Diseases

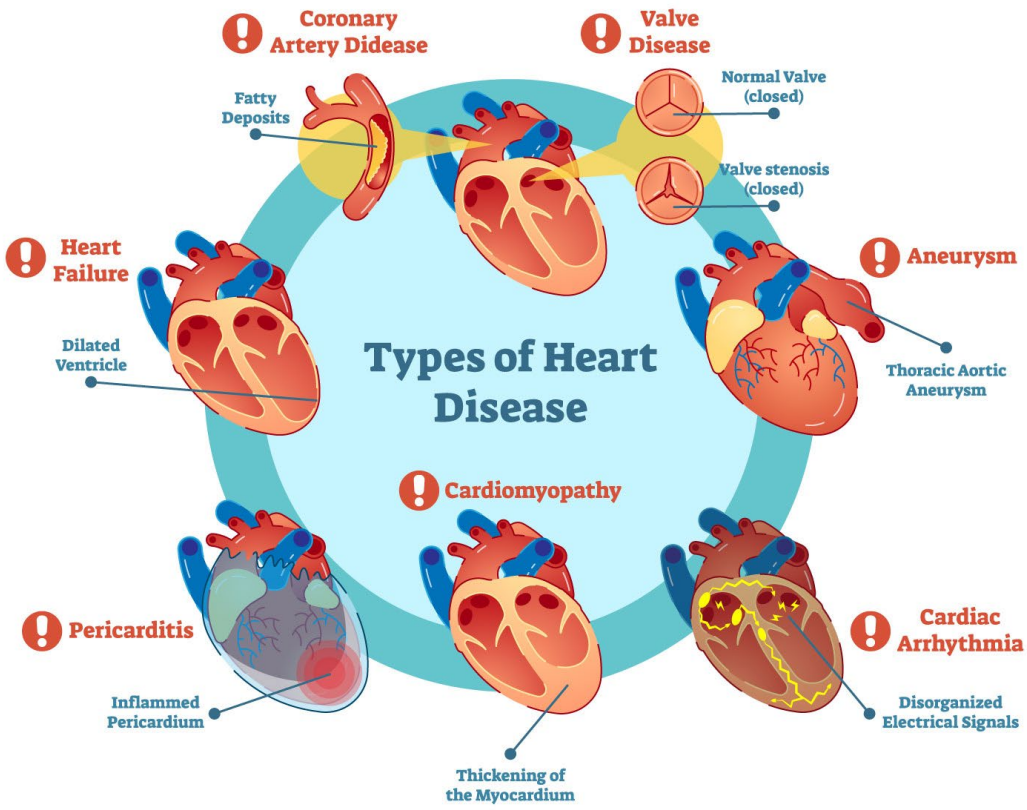
Προτού της προχωρήσουμε στην ανάλυση και επεξεργασία των δεδομένων της καλό θα ήταν να δούμε κάποια βασικά στοιχεία για της καρδιακές παθήσεις και την στεφανιαία νόσο.

Αναφέραμε πως οι καρδιακές παθήσεις είναι από τα σημαντικότερα αίτια θανάτου στον κόσμο σε καθημερινή βάση. Σύμφωνα και με τον παγκόσμιο οργανισμό υγείας (World Health Organization) αποτελεί την 1^{η} αιτία θανάτου με 17,9 εκατομμύρια θανάτους ετησίως, κάτι που μεταφράζεται στο 32% περίπου των ετήσιων αιτίων θανάτων. Συνήθως ενδείξεις που μπορεί να προκαλέσουν κάποια καρδιακά προβλήματα είναι η κακή διατροφή, η αποφυγή τακτικής φυσικής άσκησης, το κάπνισμα, η κατάχρηση αλκοόλ και άλλα. Αυτά τα δεδομένα μπορεί να κάνουν την εμφάνιση της στον κάθε άνθρωπο σε διάφορες μορφές, είτε ως αύξηση της γλυκόζης του αίματος, αύξηση της πίεσης του αίματος, αύξηση των λιπιδίων, παχυσαρκία και άλλα. Αυτοί οι παράγοντες ρίσκου που συσχετίζονται με της καρδιακές παθήσεις μπορούν να μετρηθούν σε διάφορα κέντρα υγείας και νοσοκομεία με σκοπό την πρόληψη εμφάνισης των συσχετιζόμενων παθήσεων. Για την πρόληψη αυτών των παθήσεων υπάρχουν μέτρα που μπορούν να ληφθούν, της η μείωση του καπνίσματος, η ελαχιστοποίηση της κατανάλωσης αλάτων σε συνδυασμό με την κατανάλωση φρούτων και λαχανικών και τακτική άσκηση, μέτρα τα οποία μπορούν να οδηγήσουν στην πρόληψη πρώιμων θανάτων.

CHD

Η καρδιά αποτελεί έναν μυ ο οποίος βοηθάει στην άντληση του αίματος στον υπόλοιπο οργανισμό. Της είναι φυσικό λοιπόν χρειάζεται διαρκή τροφοδοσία αίματος για να μπορεί να λειτουργεί. Παρόλα αυτά ο της ο μυς της καρδιάς δεν τροφοδοτείται από το αίμα το οποίο αντλεί αλλά έχει δικές του αρτηρίες από της οποίες θρέφεται. Αυτές οι αρτηρίες ονομάζονται στεφανιαίες(coronary). Στεφανιαίος μυς είναι μια εναλλακτική ονομασία για τον μυ της καρδιάς του ανθρώπου. Η καρδιακή πάθηση της στεφανιαίας νόσου (coronary artery or heart disease) εμφανίζεται όταν υπάρχει συμπλοκή της ροής του αίματος στον οργανισμό είτε αυτή είναι μερική είτε ολική. Αναφερόμαστε στην συγκεκριμένη νόσο καθώς αποτελεί το νούμερο ένα αίτιο εμφάνισης καρδιακών προσβολών, καρδιακής αστάθειας και στηθάγχης. Επιπλέον εμφανίζεται περισσότερο σε μεγαλύτερης ηλικίας ανθρώπους που ξεπερνούν συνήθως το πεντηκοστό έτος της

ηλικίας της, απειλεί σε μεγαλύτερο μέρος άνδρες και μπορεί να είναι κληρονομική. Ο πιο συνηθισμένος λόγος εμφάνισης της πάθησης είναι η σκλήρυνση των αρτηριών. Η καρδιακή αστάθεια και η καρδιακή προσβολή προκαλούν αρρυθμίες και μπορεί να κάνουν τον μυ της καρδιάς να σταματήσει. Κάτι τέτοιο μπορεί να αποβεί μοιραίο αν δεν αντιμετωπιστεί άμεσα. Επιπλέον οι ζημιές που προκαλεί μια καρδιακή προσβολή, ακόμα και αν αντιμετωπιστούν εγκαίρως είναι πιθανό να καταστήσουν μόνιμες βλάβες στον παθών οργανισμό.



Εικόνα 4.1 Είδη Καρδιακών Ασθενειών

Η πιο συνηθισμένη αιτία εμφάνισης της στεφανιαίας νόσου είναι η σκλήρυνση των αρτηριών(atherosclerosis). Ουσιαστικά τα λιπίδια, η χοληστερόλη και άλλα λίπη σταδιακά οδηγούν στην διάφραξη των αρτηριών. Αυτό οδηγεί είτε σε σμίκρυνση της διόδου της αρτηρίας είτε σε ξαφνική ρήξη που οδηγεί σε θρόμβο ο οποίος μπλοκάρει πλήρως την ροή της αρτηρίας και οδηγεί σε καρδιακή προσβολή(έμφραγμα). Άλλα αίτια εμφάνισης τα οποία είναι λιγότερο συχνά μπορεί να είναι διάφορων ειδών της σπασμοί από χρήση ναρκωτικών ουσιών, όπου η αρτηρία ξαφνικά κλείνει εντελώς και σε περίπτωση που παραμείνει κλειστή για αρκετή ώρα μπορεί να οδηγήσει σε καρδιακή προσβολή. (World Health Organization, 2021)

Terminology

a. Αθηροσκλήρωση(Atherosclerosis): η συσσώρευση ουσιών στα τοιχώματα των αρτηριών που μπορούν να παρεμποδίσουν την ροή του αίματος και να δημιουργήσουν θρόμβους.
b. Στηθάγχη(Angina): πόνος στην περιοχή του στήθους λόγω μείωσης της ροής του αίματος της καρδιακές αρτηρίες.
c. Σταθερή και ασταθης στηθάγχη(Stable & Unstable angina):η σταθερή αφορά σε περιπτώσεις που σχετίζονται με το οξυγόνο(της η άσκηση) και η ασταθής μπορεί να συμβεί σε οποιαδήποτε κατάσταση.
d. Τυπική και άτυπη στηθάγχη(Typical & atypical angina):η τυπική μορφή αναφέρεται σε πόνους που μπορεί να σχετίζονται με αδιαθεσία στη περιοχή του στήθους. Υπάρχουν και περιπτώσεις που εμφανίζουν προβλήματα αναπνοής και ναυτίες όπου μιλάμε για την άτυπη μορφή.
e. Θρόμβος(Thrombus): μάζα αίματος η οποία αποκτά στερεή μορφή και δυσχεραίνει την φυσιολογική ροή του αίματος. Αν της ο θρόμβος αποκολληθεί από τα τοιχώματα και προχωρήσει σε άλλα μέρη του σώματος μέσω των αρτηριών μιλάμε για έμβολο.
f. Ηλεκτροκαρδιογράφημα(Electrocardiogram): ηλεκτρονικό γράφημα κυματοειδούς μορφής που καταγράφει της καρδιακούς παλμούς.
g. Μαγνητική(Nuclear stress test): παρατήρηση της ροής του αίματος κατά την ανάπαυση και την άσκηση με τη χρήση της ραδιενεργού υγρού που εισάγεται στον ασθενή.
h. Ασυμπτωματική ασθένεια(Asymptomatic disease): μια ασθένεια όπου ο ασθενής παρατηρεί πολύ λίγα έως και καθόλου συμπτώματα.
i. Υπερτροφία(hypertrophy): μεγαλύτερο του φυσιολογικού μέγεθος των τοιχομάτων της καρδιάς που την κάνει λιγότερο ελαστική και προκαλεί δυσλειτουργίες.

4.2 Διερεύνηση των δεδομένων

Ξεκινώντας την διερεύνηση των δεδομένων που θα χρησιμοποιήσουμε θα δούμε το dataset που χρησιμοποιήθηκε για την περαιτέρω ανάλυση καθώς και της ιδιότητες που εμφανίζει.

Στην εν λόγω εργασία επιλέχθηκε το πλήθος των δεδομένων Heart Disease UCI το οποίο λήφθηκε από το www.kaggle.com (https://www.kaggle.com/ronitf/heart-disease-uci) και προέρχεται από το **UCI Machine Learning Repository** που αποτελεί μια από της μεγαλύτερες αποθήκες δεδομένων για ανάλυση και χρησιμοποιείται από της αναλυτές παγκοσμίως.

Το συγκεκριμένο dataset αποτελείται από 14 στήλες(columns) και 303 παρατηρήσεις(rows). Το εν λόγω dataset είναι υποσύνολο του αρχικού που

αποτελούνταν από 76 χαρακτηριστικά(attributes) εκ των οποίων κρατήθηκαν οι σχετικότερες με το θέμα που καλείται να αναλύσει η παρούσα εργασία.

Η λίστα με τα attributes που περιέχει το πλήθος των δεδομένων που θα χρησιμοποιήσουμε είναι:

i. *age:* Η ηλικία των ατόμων που απαρτίζουν τον πληθυσμό. Παρουσιάζεται σε χρόνια και αποτελείται από ηλικίες ….
ii. *sex:* Το φύλο των ατόμων. Εντός του συνόλου δεδομένων υπάρχουν 2 τιμές 0 και 1. Η τιμή 0 αντιστοιχεί στο γυναικείο φύλο ενώ η τιμή 1 στο ανδρικό.
iii. *cp (chest pain type):* η συγκεκριμένη μεταβλητή δείχνει το είδος του πόνου στο στήθος. Δεν υπάρχει κάποια συγκεκριμένη μέθοδος κατηγοριοποίησης της εν λόγω μεταβλητής. Οι κατηγορίες είναι δοσμένες με αριθμούς 0 έως 3. Με 0 συμβολίζεται η ασυμπτωματική ομάδα, με 1 αυτή του μη τυπικού πόνου, με 2 πόνος που δεν σχετίζεται με στηθάγχη και με 3 αυτή της τυπικής στηθάγχης.
iv. *trestbps(resting blood pressure):* Η πίεση του αίματος σε συνδυασμό με την ποσότητα υδραργύρου που υπάρχει σε αυτό μετρημένη σε mg.
v. *chol(cholesterol):* Τα επίπεδα χοληστερόλης του οργανισμού των ατόμων στο αίμα.
vi. *fbs(fasting blood sugar):* Μια μέτρηση των σακχάρων του αίματος που υποδεικνύει αν τα επίπεδα αυτών είναι μεγαλύτερα η μικρότερα από 120 mg/dl. Εδώ οι παρατηρήσεις χαρακτηρίζονται με 0 αν δεν ξεπερνούν το όριο ενώ με 1 αν το ξεπερνούν.
vii. *restecg(resting electrocardiographic results):* Τα αποτελέσματα της ηλεκτρικού καρδιογραφήματος. Η μεταβλητή λαμβάνει τιμές από 0 έως 2 όπου 0 είναι πιθανή αριστερή υπερτροφία, 1 φυσιολογικά αποτελέσματα και 2 αφύσικες ενδείξεις του μυοκαρδίου στο διάστημα ST ή στο T wave.
viii. *thalach(maximum heart rate):* Οι μέγιστοι καρδιακοί παλμοί κατά τη διάρκεια του stress test.
ix. *exang(exercise induced angina):* Η εμφάνιση στηθάγχης στον ασθενή κατά τη διάρκεια της άσκησης. Ως 0 χαρακτηρίζεται η απουσία στηθάγχης ενώ με 1 η εμφάνιση της.
x. *oldpeak(ST depression induced by exercise relative to rest):* Μείωση του διαστήματος ST κατά τη διάρκεια της άσκησης σε σχέση με τα επίπεδα αυτού πριν από αυτή.
xi. *slope(slope of ST segment):* κλίση του διαστήματος ST κατά τη διάρκεια απαιτητικής άσκησης. Η μεταβλητή λαμβάνει τιμές από 0 έως 2 όπου 0 είναι καθοδική κλίση, 1 είναι σταθερή και 2 ανοδική.
xii. *ca(Number of major vessels-flourosopy):* Ο αριθμός των κύριων αιμοφόρων αγγείων που έχουν χρωματιστεί από το ραδιενεργό υγρό.
xiii. *Thal(blood flow results):* Τα αποτελέσματα από την παρατήρηση της ροής του αίματος με τη χρήση ραδιενεργού υγρού.
xiv. *Target:* Η εμφάνιση καρδιακού νοσήματος στο άτομο. Η τιμή 0 δείχνει παρουσία καρδιακού νοσήματος ενώ η τιμή 1 μη εμφάνιση κάποιας πάθησης.

Τα δεδομένα που έχουμε επιλέξει έχουν συλλεχθεί από διαφορετικά νοσοκομεία των Ηνωμένων Πολιτειών της Αμερικής και συγκεκριμένα της πολιτείας του Cleveland. Περιέχει περίπου 300 παρατηρήσεις ασθενών με τα χαρακτηριστικά της για κάθε μια από της προαναφερθείσες μεταβλητές.

4.2 Προεπεξεργασία των δεδομένων

Λαμβάνοντας τα δεδομένα της και προτού ξεκινήσουμε οποιαδήποτε ενέργεια επεξεργασίας των δεδομένων χρησιμοποιώντας της εντολές summary και glimpse της R μπορούμε να πάρουμε μια πρώτη εικόνα των δεδομένων που περιέχονται στο dataset που έχουμε επιλέξει.

```
> summary(datawork)
      age            sex                      cp          trestbps          chol          fbs          restecg
 Min.   :29.00   female: 95   asymptomatic angina:141   Min.   : 94.0   Min.   :126.0   <=120:253   Length:296
 1st Qu.:48.00   male  :201   atypical angina    : 49   1st Qu.:120.0   1st Qu.:211.0   >120 : 43   Class :character
 Median :56.00                non-anginal        : 83   Median :130.0   Median :242.5               Mode  :character
 Mean   :54.52                typical angina     : 23   Mean   :131.6   Mean   :247.2
 3rd Qu.:61.00                                          3rd Qu.:140.0   3rd Qu.:275.2
 Max.   :77.00                                          Max.   :200.0   Max.   :564.0
    thalach       exang       oldpeak             slope     ca                 thal        target
 Min.   : 71.0   no :199   Min.   :0.000   downsloping: 21   0:173   fixed defect     : 18   Length:296
 1st Qu.:133.0   yes: 97   1st Qu.:0.000   flat       :137   1: 65   normal           :163   Class :character
 Median :152.5             Median :0.800   upsloping  :138   2: 38   reversable defect:115   Mode  :character
 Mean   :149.6             Mean   :1.059                      3: 20
 3rd Qu.:166.0             3rd Qu.:1.650                      4:  0
 Max.   :202.0             Max.   :6.200
> |
```

Εικόνα 4.2 Γενική εικόνα μεταβλητών

```
> glimpse(datawork)
Rows: 296
Columns: 14
$ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 50, 58, 66, 43, 69, 59, 44, 42, 61, 40, 71,~
$ sex      <fct> male, male, female, male, female, male, female, male, male, male, male, female, male, male, female, fem~
$ cp       <fct> typical angina, non-anginal, atypical angina, atypical angina, asymptomatic angina, asymptomatic angina~
$ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 130, 110, 150, 120, 120, 150, 150, 140, 135~
$ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 266, 211, 283, 219, 340, 226, 247, 239, 234~
$ fbs      <fct> >120, <=120, <=120, <=120, <=120, <=120, <=120, <=120, <=120, <=120, <=120, >120, <=120, <=120, <=120, <~
$ restecg  <chr> "hypertrophy", "normal", "hypertrophy", "normal", "normal", "normal", "hypertrophy", "normal", "normal"~
$ thalach  <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 171, 144, 162, 158, 172, 114, 171, 151, 161~
$ exang    <fct> no, no, no, no, yes, no, no, no, no, no, no, no, yes, no, no, no, no, no, no, no, yes, no, yes, yes, yes~
$ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0.6, 1.8, 1.0, 1.6, 0.0, 2.6, 1.5, 1.8, 0.5~
$ slope    <fct> downsloping, downsloping, upsloping, upsloping, upsloping, upsloping, flat, flat, upsloping, upsloping, upsloping,~
$ ca       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 1, 0, 1, 0, 0, 1, 1~
$ thal     <fct> fixed defect, normal, normal, normal, normal, fixed defect, normal, reversable defect, reversable defec~
$ target   <chr> "no disease", "no disease", "no disease", "no disease", "no disease", "no disease", "no disease", "no d~
```

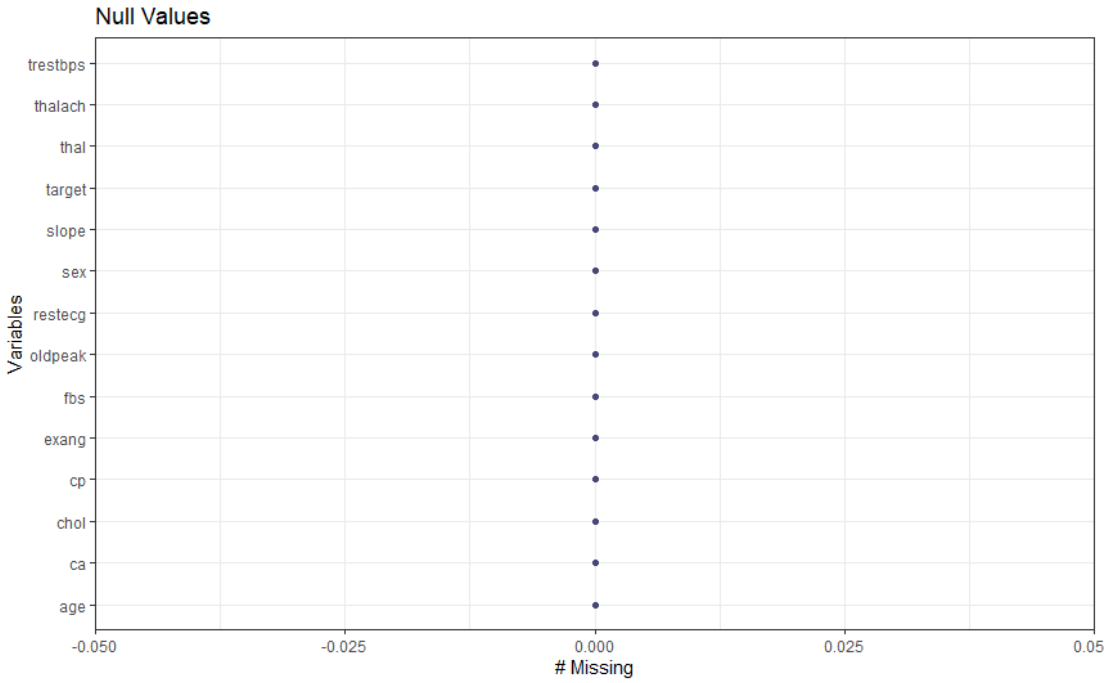
Εικόνα 4.3 Εμφάνιση πρώτων στοιχείων μεταβλητών

Με μια γρήγορη πρώτη ματιά στα αποτελέσματα που προκύπτουν από τις παραπάνω εντολές μπορούμε να έχουμε μια πρώτη εικόνα της κάθε μεταβλητής του δείγματος. Συγκεκριμένα, τα σημαντικότερα που παρατηρούμε από την εντολή summary, πέραν της έκτασης των δεδομένων όπως μέσος όρος κ.α., είναι πως στο δείγμα μας περιέχονται περισσότεροι άνδρες από γυναίκες(95 γυναίκες/201 άνδρες) και πως ο μέσος όρος των ατόμων του δείγματος είναι περίπου τα 54 έτη.

Επίσης από την εντολή glimpse μπορούμε να πάρουμε μια γενική εικόνα των inputs κάθε μεταβλητής.

Προχωρώντας και ξεκινώντας να κάνουμε μια ανάλυση και κατανόηση των δεδομένων με τα οποία θα δουλέψουμε, ύστερα από την αρχική κατανόηση των μεταβλητών και σε τι αναφέρονται, το πρώτο που ελέγχουμε είναι για κενές τιμές εντός του δείγματος. Οι κενές τιμές, σε περίπτωση ύπαρξης τους θέλουμε να φιλτραριστούν εκτός του δείγματος μας καθώς μπορεί να μας οδηγήσουν σε λάθος αποτελέσματα.

Για να ελέγξουμε όμως τις κενές τιμές και να προχωρήσουμε σε οποιαδήποτε διαχείριση των δεδομένων μας, χρειάστηκε να κάνουμε κάποιες ενέργειες ώστε να φέρουμε τα δεδομένα μας σε μορφή αξιοποιήσιμη και εύχρηστη από το περιβάλλον της R. Έτσι αρχικά μετονομάσαμε την στήλη age, καθώς εμφανιζόταν με λάθος όνομα και μετατρέψαμε τα στοιχεία των δεδομένων μας που ήταν όλα αριθμητικά σε κατηγορικά, όπου ήταν αναγκαίο, ώστε να είναι περισσότερο κατανοήσιμα. Τέλος μετατρέψαμε τις μεταβλητές σε factors, ώστε να είναι αξιοποιήσιμα από τις εντολές της R. (βλ. Μέρος Α



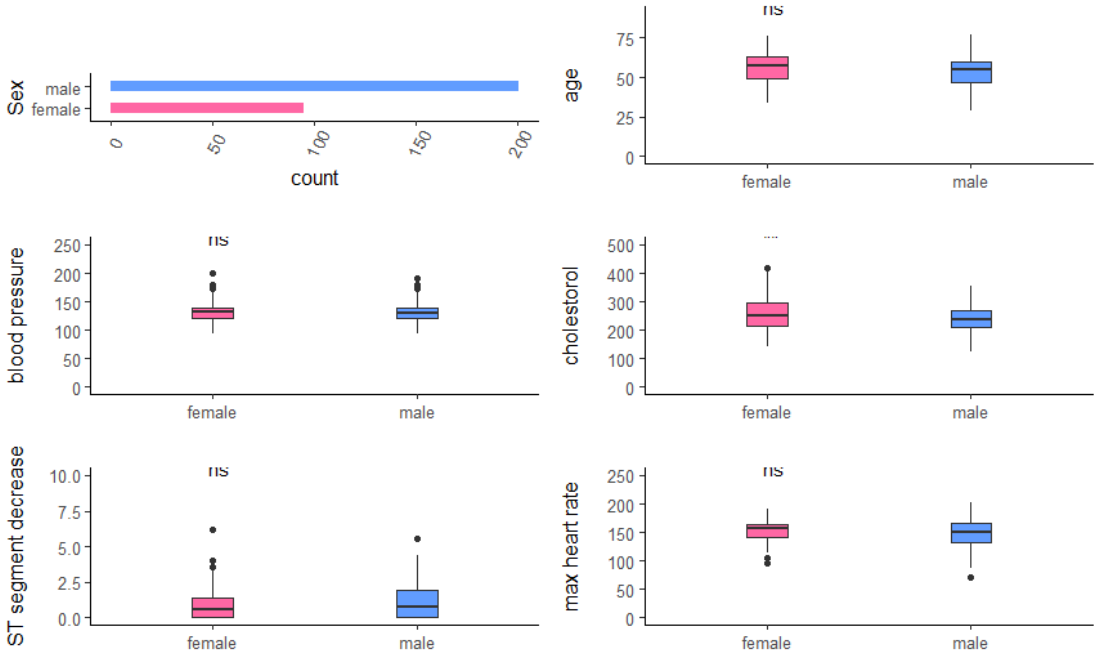
Διάγραμμα 4.1: Κενές τιμές(nulls)

Όπως παρατηρούμε λοιπόν δεν υπάρχουν κενές τιμές στο σύνολο των δεδομένων μας για καμία μεταβλητή.

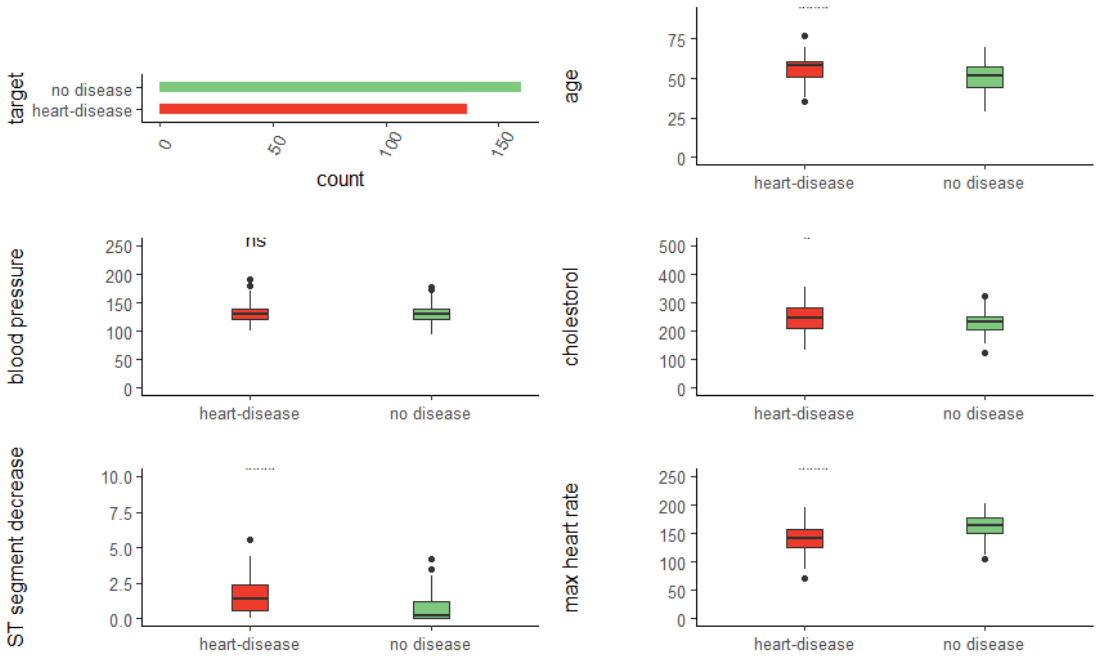
Στη συνέχεια θα πρέπει να ελέγξουμε για τυχόν ακραίες ή μη-αποδεκτές τιμές. Στην προκειμένη περίπτωση, παρατηρείται η λάθος δοσμένη τιμή στην μεταβλητή thal που ισούται με 0, όπου αρχικά δεν υπήρχε ως δοθείσα τιμή, αλλά λόγω του ότι το συγκεκριμένο dataset είναι ελαφρώς προεπεξεργασμένο, έχουν αντικατασταθεί τιμές που προυπήρχαν ως κενά με την τιμή 0. Ομοίως έχει συμβεί και στην μεταβλητή ca για την τιμή 4. Έτσι προχωρώντας σε οποιαδήποτε περαιτέρω ανάλυση, αν συμπεριληφθούν οι προαναφερθείσες τιμές σίγουρα θα οδηγήσουν σε εσφαλμένα(biased) αποτελέσματα. Έτσι με μία απλή εντολή τα φιλτράρουμε από το συγκεκριμένο πλήθος των δεδομένων. (*βλ. Κεφάλαιο 7*)

Γενική εικόνα διακύμανσης ποσοτικών μεταβλητών δείγματος

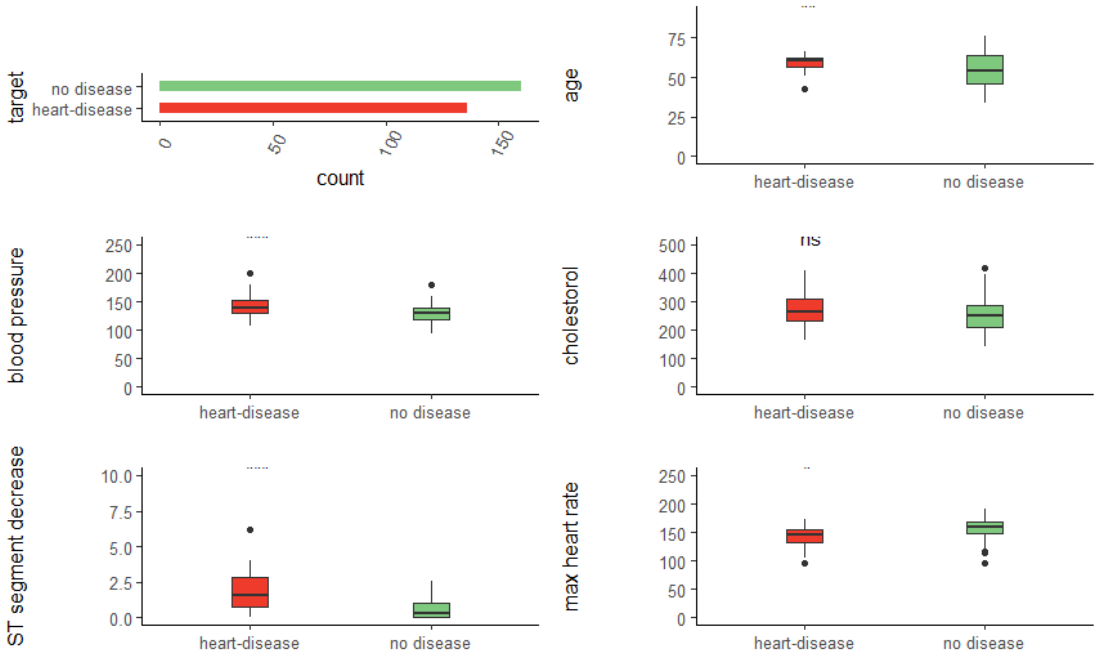
Αρχικά ελέγξαμε την γενικότερη εικόνα των ποσοτικών μεταβλητών του δείγματος για τα δύο φύλα. Συγκεκριμένα:



Διάγραμμα 4.2: Γενική εικόνα ποσοτικών μεταβλητών βάσει φύλου



Διάγραμμα 4.3: Γενική εικόνα ποσοτικών μεταβλητών ανδρών

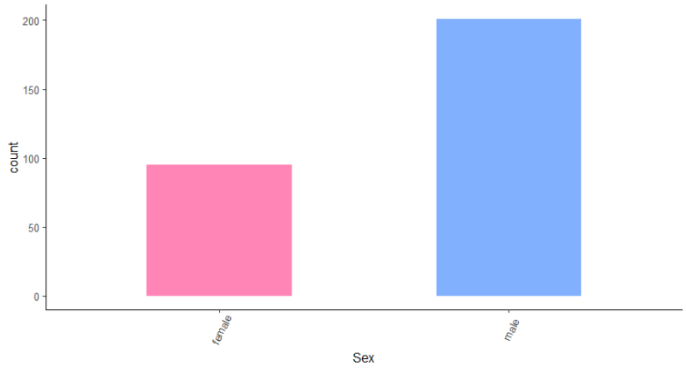


Διάγραμμα 4.4: Γενική εικόνα ποσοτικών μεταβλητών γυναικών

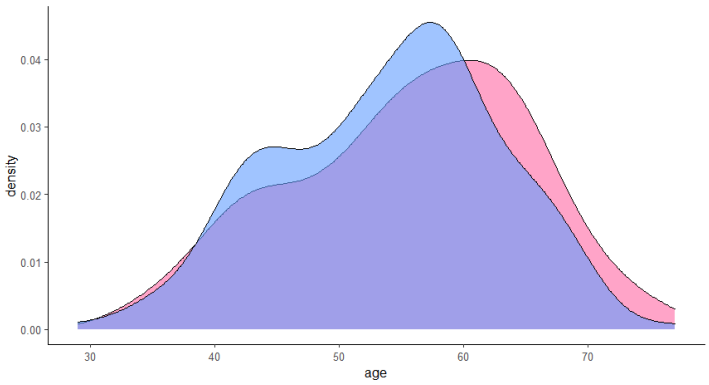
Αρχικά κάνουμε μια πρώτη απεικόνιση των ποσοτικών μεταβλητών του δείγματος. Η γενικότερη τάση δείχνει πως το δείγμα μας παρουσιάζει μια αριστερή ασυμμετρία.

Συνεχίζοντας στην οπτική ανάλυση του δείγματος θα προσπαθήσουμε να παρατηρήσουμε τις διακυμάνσεις των μεταβλητών αναφορικά με τις μεταβλητές της ηλικίας, του φύλου και της μεταβλητής – στόχου(εμφάνιση ασθένειας).

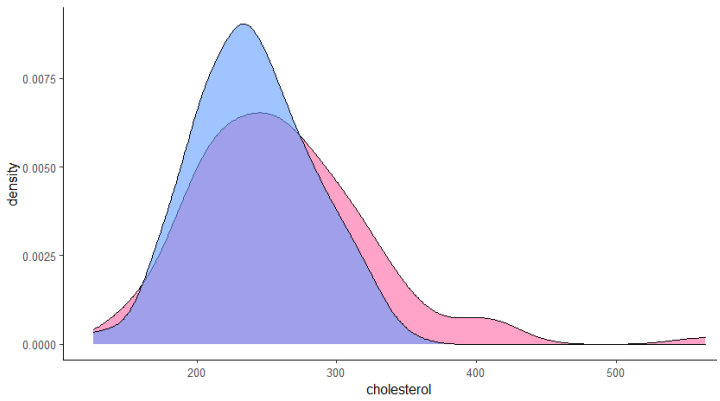
Παρουσίαση δεδομένων σύμφωνα με τη μεταβλητή του φύλου(sex):



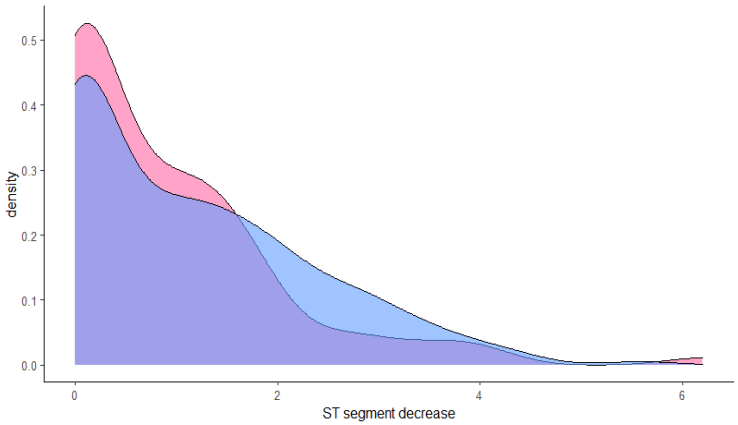
Διάγραμμα 4.5 Πλήθος Ανδρών & Γυναικών



Διάγραμμα 4.6 Διακύμανση ηλικίας

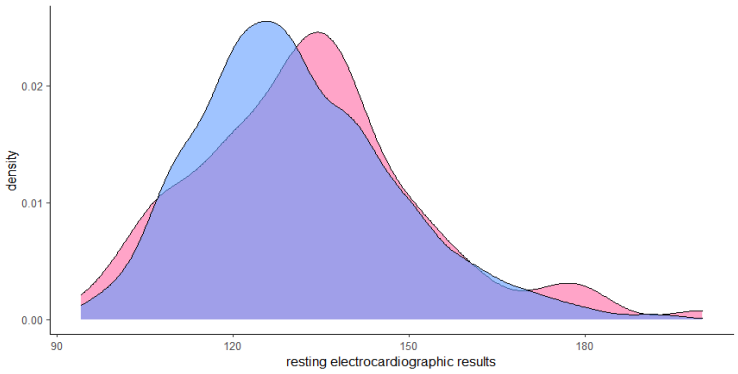


Διάγραμμα 4.7 Διακύμανση χοληστερόλης



Διάγραμμα 4.8 Διακύμανση διαστήματος ST

Διάγραμμα 4.9 Διακύμανση καρδιακών παλμών

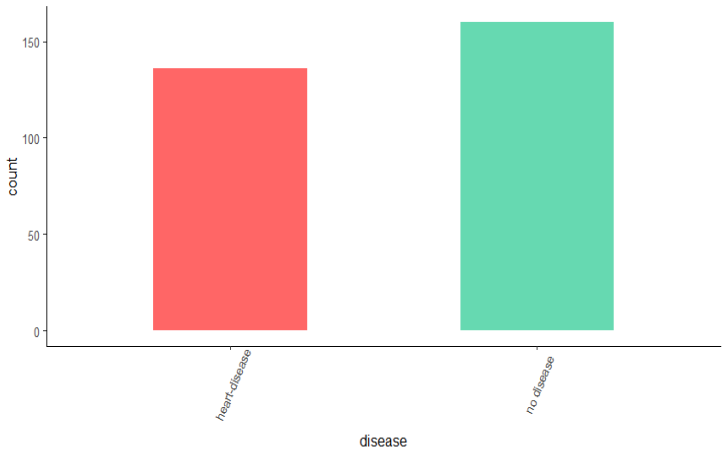


Διάγραμμα 4.10 Διακύμανση αποτελεσμάτων αξονικής

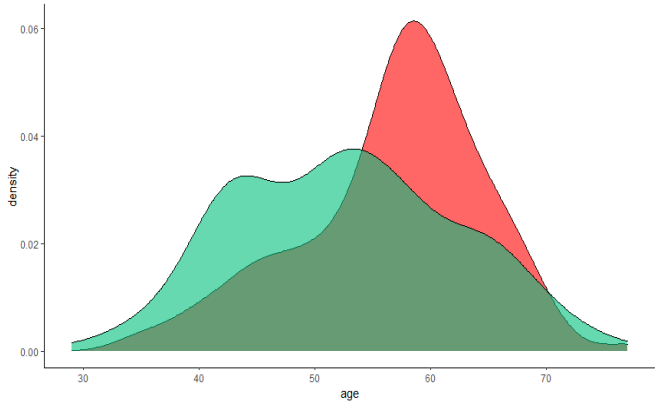
Ξεκινώντας θα προσπαθήσουμε να κατανοήσουμε τα δεδομένα μας με βάση την μεταβλητή του φύλου. Το πρώτο χαρακτηριστικό του δείγματος που παρατηρούμε είναι πως ο αριθμός των ανδρών στο δείγμα είναι σχεδόν διπλάσιος από αυτόν των γυναικών*(Διάγραμμα 2.1)*. Επιπλέον παρατηρούμε πως οι ηλικίες του δείγματος στο μεγαλύτερο ποσοστό είναι στην περιοχή 40 με 60 ετών*(Διάγραμμα 2.2)*. Στις υπόλοιπες μεταβλητές παρατηρούνται αρκετά κοινά μεταξύ των δύο φύλων χωρίς μεγάλες

διαφορές. Τα επίπεδα χοληστερόλης είναι σχετικά υψηλά στο δείγμα*(Διάγραμμα 2.3)* καθώς βλέπουμε πως ο μέσος όρος του δείγματος κινείται γύρω στο 200mg/dL. Τέλος παρατηρούμε πως ο μέγιστος αριθμός των καρδιακών παλμών των ανδρών κινείται σε χαμηλότερα επίπεδα σε σχέση με αυτόν των γυναικών.

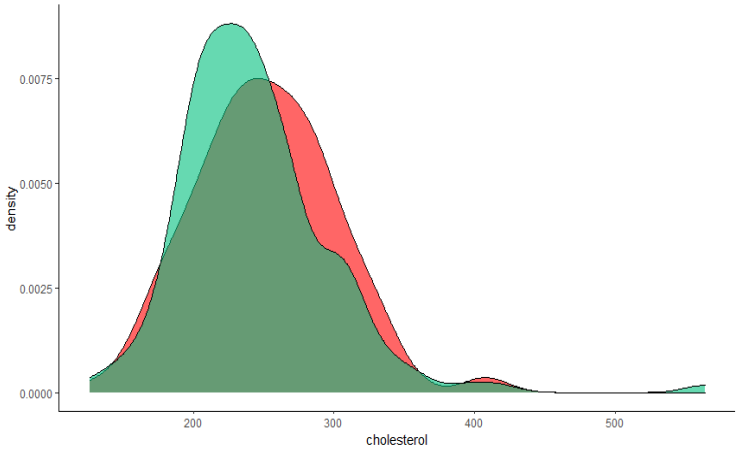
Παρουσίαση δεδομένων σύμφωνα με τη μεταβλητή της πάθησης(disease):



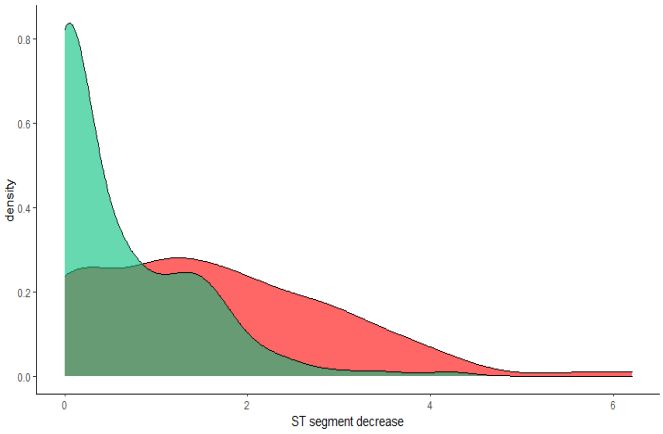
Διάγραμμα 4.11 Πλήθος νοσούντων ατόμων



Διάγραμμα 4.12 Διακύμανση ηλικίας

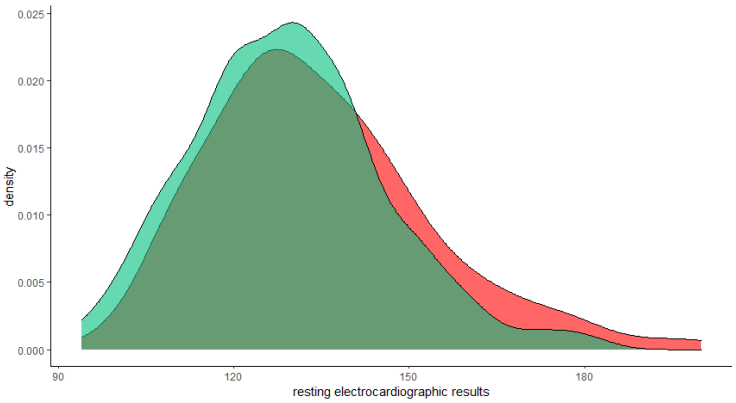


Διάγραμμα 4.13 Διακύμανση χοληστερόλης



Διάγραμμα 4.14 Διακύμανση διαστήματος ST

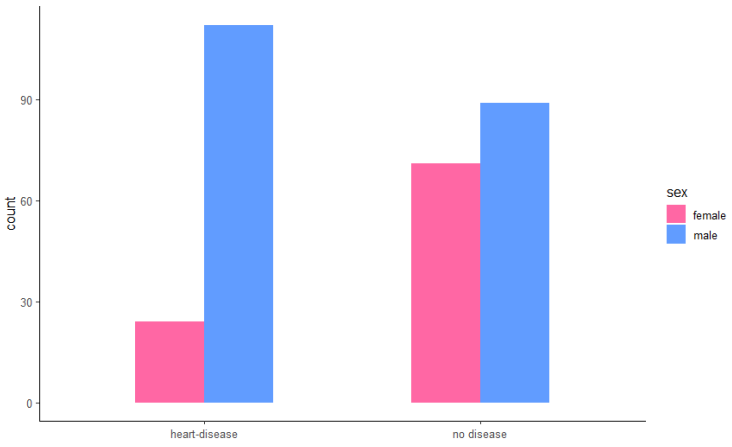
Διάγραμμα 4.15 Διακύμανση καρδιακών παλμών



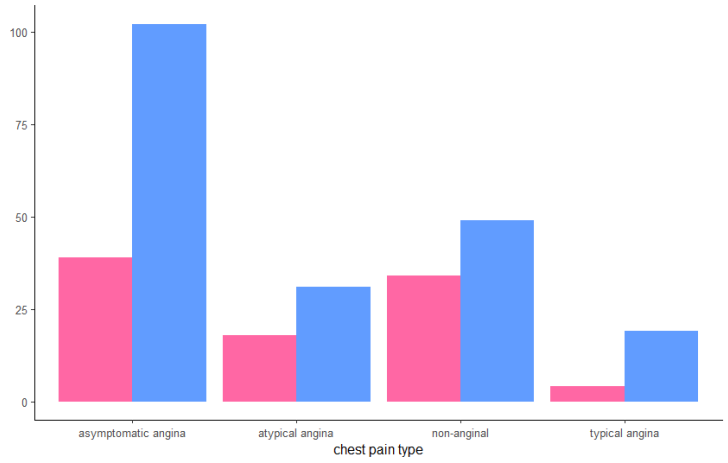
Διάγραμμα 4.16 Διακύμανση αποτελεσμάτων αξονικής

Συνεχίζοντας λοιπόν την οπτική μας ανάλυση, προσπαθήσαμε να κάνουμε μια πρώτη οπτική εκτίμηση των δεδομένων μας με βάσει την μεταβλητή – στόχο που έχουμε, δηλαδή την εμφάνιση καρδιακής ασθένειας*(disease)*. Σε γενική ανάλυση το δείγμα μας παρουσιάζει σχεδόν τον ίδιο αριθμό νοσούντων και μη-νοσούντων*(Διάγραμμα 2.7)*. Με μια πρώτη ματιά παρατηρούμε πως δεν υπάρχουν σαφείς οπτικές διαφορές συγκριτικά με την πρώτη μας εικόνα από την μεταβλητή του φύλου, όσον αφορά στις μεταβλητές της χοληστερόλης, του διαστήματος ST και των αποτελεσμάτων της αξονικής τομογραφίας. Ωστόσο, όσον αφορά στην ηλικία, παρατηρούμε μια αυξημένη εμφάνιση ασθενειών στις ηλικίες μεταξύ 50 και 65*(Διάγραμμα 2.8).* Επιπλέον σύμφωνα με το διάγραμμα των καρδιακών παλμών*(Διάγραμμα 2.11)* παρατηρούμε πως οι άνθρωποι που εμφανίζουν κάποια καρδιακή ασθένεια έχουν χαμηλότερο μέσο όρο ανώτατων τιμών παλμών σε σχέση με εκείνους που δεν εμφανίζουν ασθένεια και έχουν υψηλότερο μέσο όρο. Με μια σύντομη οπτική σύγκριση παρατηρούμε ένα κοινό μοτίβο με αυτό των παλμών των ανδρών*(Διάγραμμα 2.5)*. Έτσι, καθώς οπτικά δεν εμφανίζονται κάποιες άλλες ιδιαίτερες ενδείξεις, θα είχε ενδιαφέρον προχωρώντας να δούμε την συσχέτιση μεταξύ της εμφάνισης ασθένειας με το φύλο αλλά και την ηλικία των ατόμων του δείγματος.

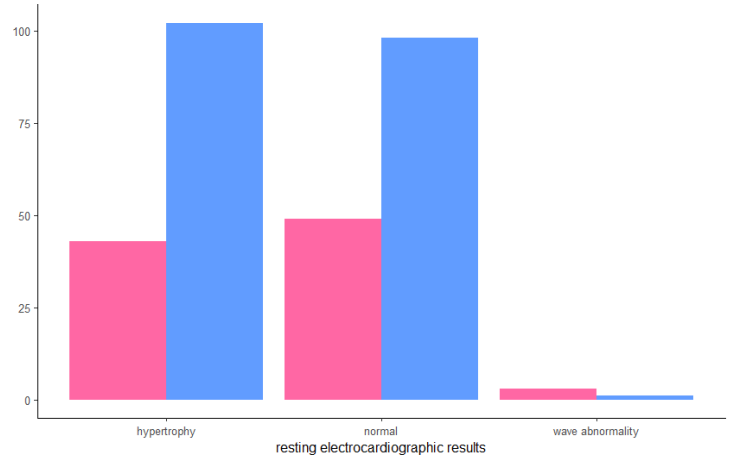
Παρουσίαση δεδομένων σύμφωνα με τις μεταβλητές της πάθησης(disease) και του φύλου(sex):



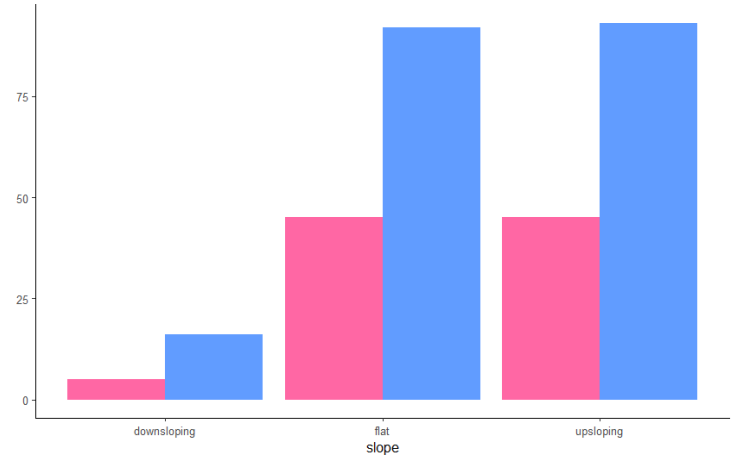
Διάγραμμα 4.17



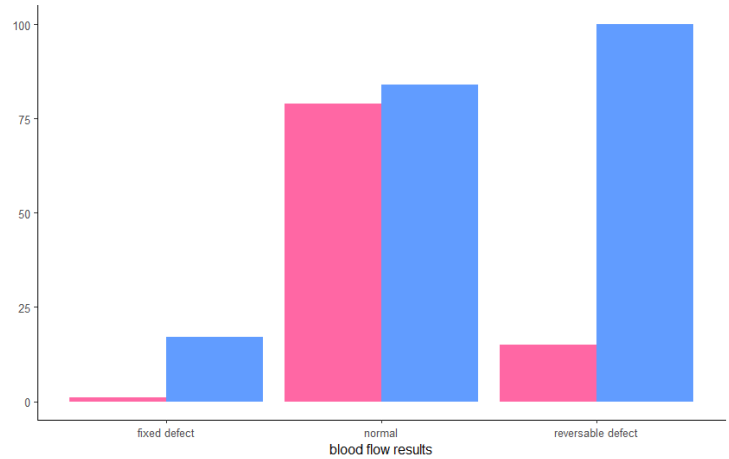
Διάγραμμα 4.18



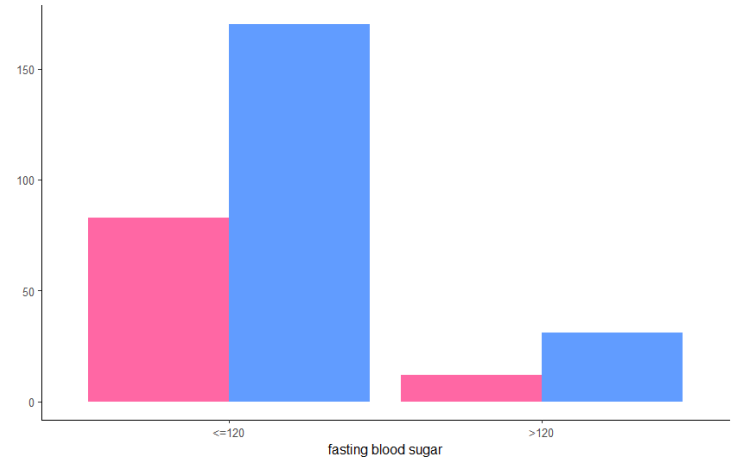
Διάγραμμα 4.19



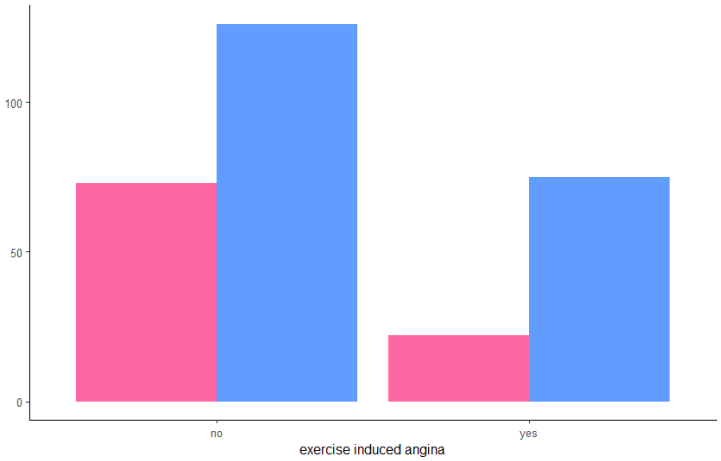
Διάγραμμα 4.20



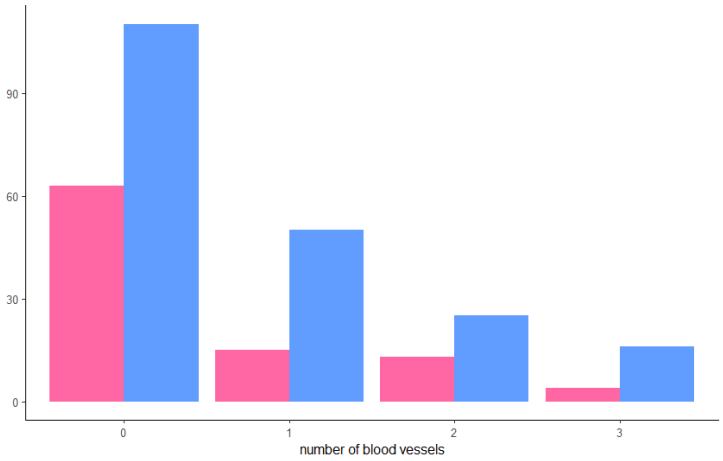
Διάγραμμα 4.21



Διάγραμμα 4.22



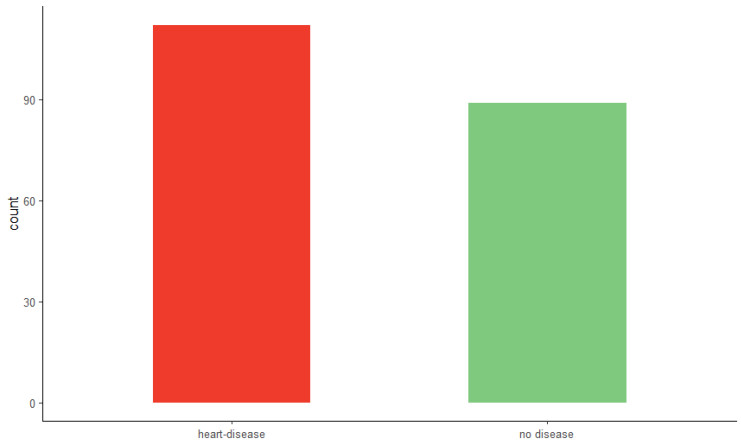
Διάγραμμα 4.23



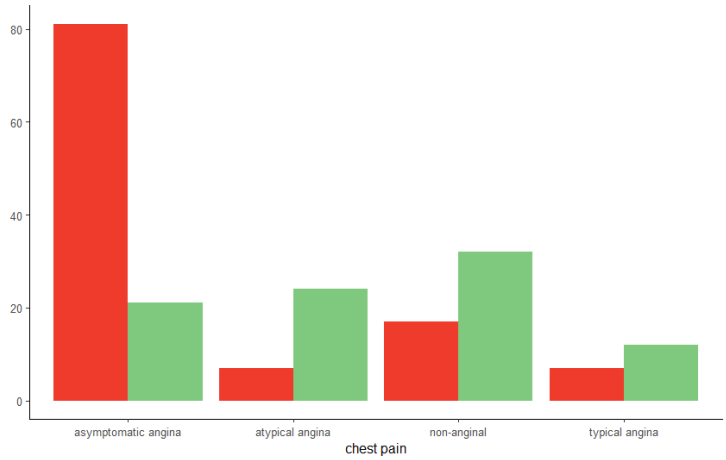
Διάγραμμα 4.24

Εν συνεχεία προσπαθούμε να εμβαθύνουμε λίγο στο δείγμα μας βλέποντας την εμφάνιση ασθενειών με βάση το φύλο. Έτσι παρατηρούμε ότι στο δείγμα μας εμφανίζεται μεγαλύτερος αριθμός ανδρών που παρουσιάζει κάποια μορφή καρδιακής ασθένειας συγκριτικά με τις γυναίκες*(Διάγραμμα 2.13)*. Επιπλέον για την μεταβλητή που αφορά στον τύπο της στηθάγχης(angina) παρατηρούμε πως οι άνδρες εμφανίζουν κατά βάση τον τύπο της ασυμπτωματικής στηθάγχης ενώ οι υπόλοιποι τύποι στηθάγχης δεν εμφανίζουν κάποια τάση μεταξύ των φύλων*(Διάγραμμα 2.14)*. Όσον αφορά στα αποτελέσματα των καρδιογραφημάτων παρατηρούμε ότι μεγαλύτερος αριθμός ατόμων εμφανίζει φυσιολογικά αποτελέσματα ή και υπερτροφία στον καρδιακό μυ ενώ ελάχιστος αριθμός ατόμων, ανεξαρτήτου φύλου εμφανίζει κάποια αφύσικη ένδειξη*(Διάγραμμα 2.15)*.

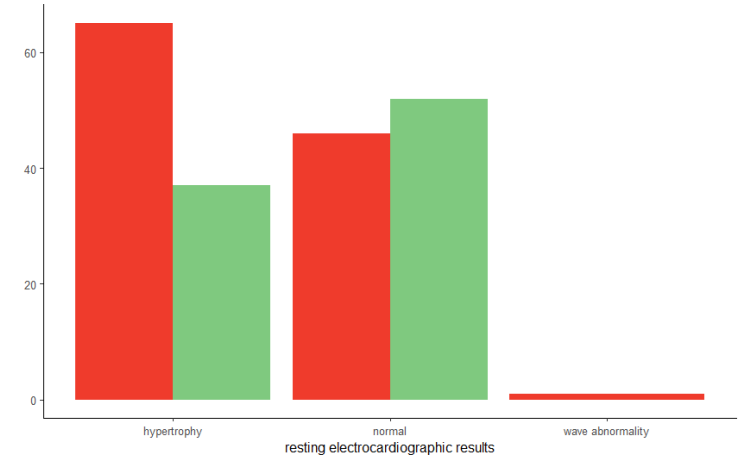
Παρουσίαση δεδομένων σύμφωνα με τη μεταβλητή της πάθησης(disease) μόνο για τους άνδρες(sex = male):



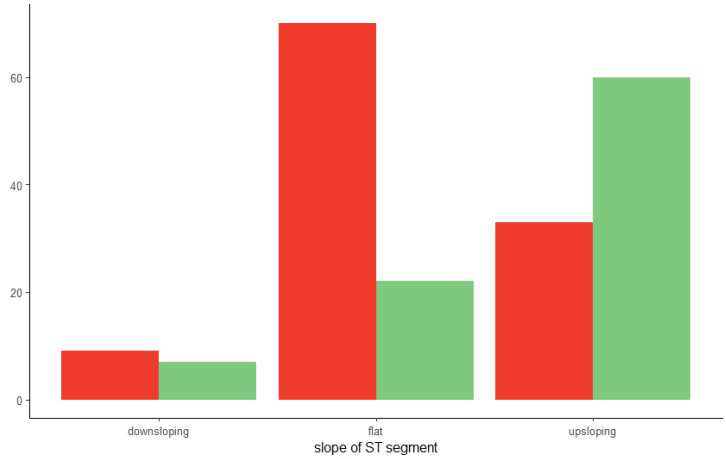
Διάγραμμα 4.25



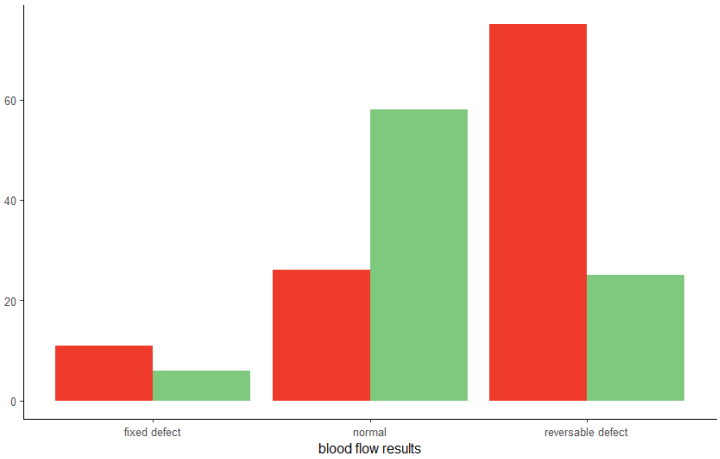
Διάγραμμα 4.26



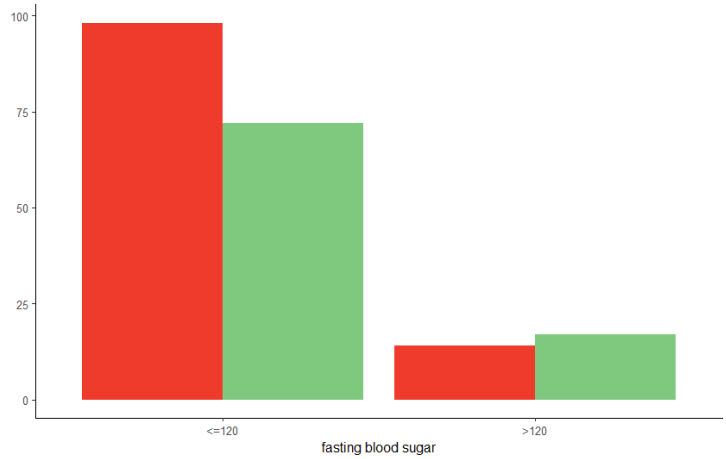
Διάγραμμα 4.27



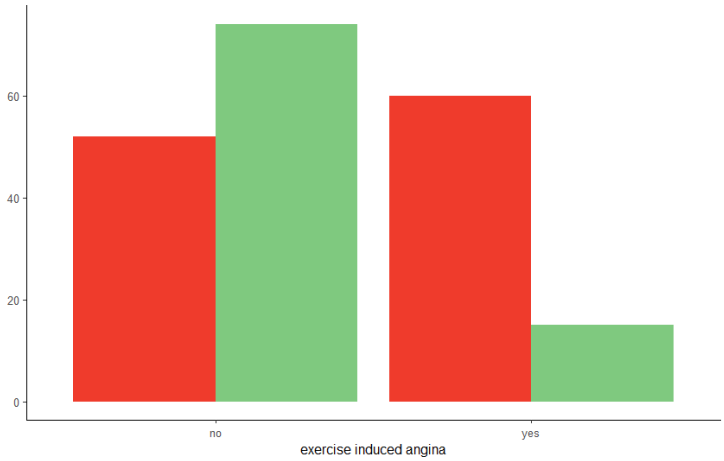
Διάγραμμα 4.28



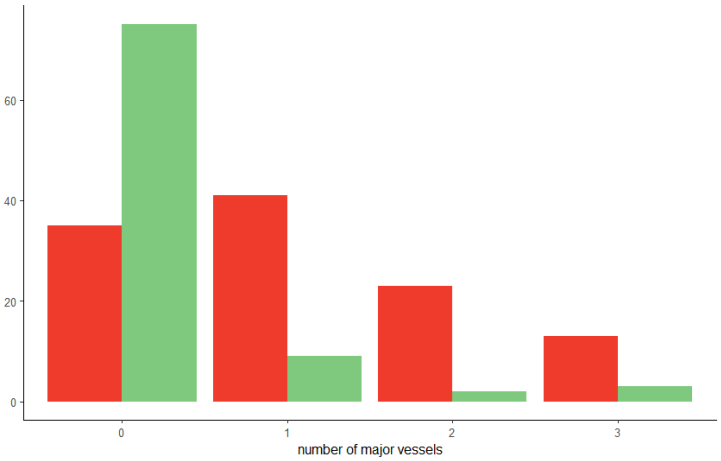
Διάγραμμα 4.29



Διάγραμμα 4.30



Διάγραμμα 4.31



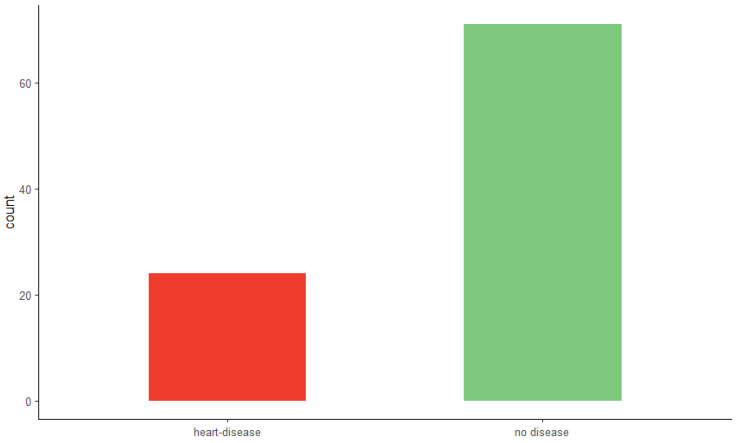
Διάγραμμα 4.32

Ακολούθως, καθώς η διαφορά μεταξύ ανδρών και γυναικών στο δείγμα παρουσιάζει ενδιαφέρον, διαχωρίσαμε τα δύο φύλα και παρατηρήσαμε τις μεταβλητές σε κάθε περίπτωση.

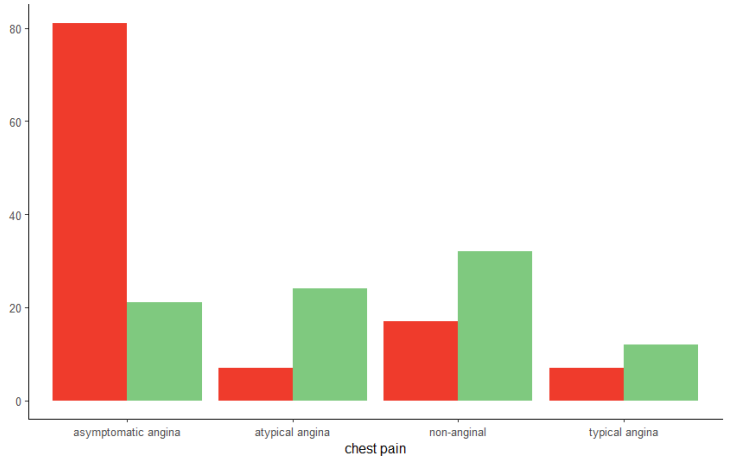
Ξεκινώντας με την περίπτωση των ανδρών και σύμφωνα με τα παραπάνω διαγράμματα παρατηρούμε πως μεγαλύτερος αριθμός ανδρών εμφανίζει κάποια καρδιακή πάθηση. Επιπλέον παρατηρούμε πως οι περισσότεροι που εμφανίζουν κάποιο είδος καρδιοπάθειας ανήκουν στην κατηγορία μη συμπτωματικής στηθάγχης και βλέπουμε πως παρουσιάζουν κάποιο είδος αναστρέψιμης βλάβης.

Έτσι με μια συνολική εικόνα παρατηρούμε πως οι άνδρες στο δείγμα μας είναι μεγαλύτερης ηλικίας, έχουν υψηλότερα επίπεδα χοληστερόλης και μικρότερο εύρος καρδιακών παλμών.

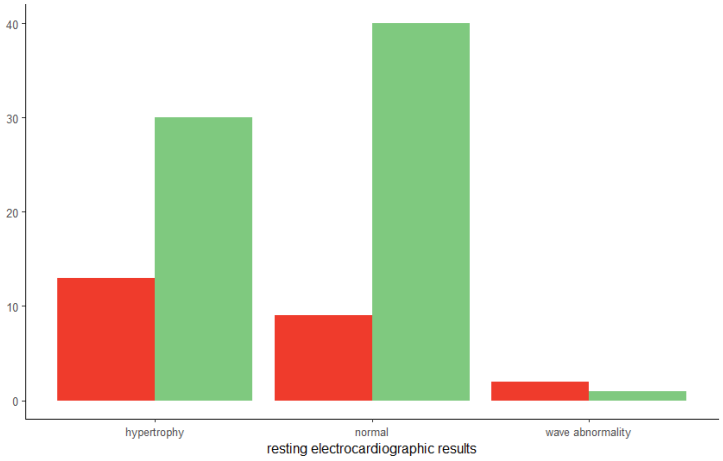
Παρουσίαση δεδομένων σύμφωνα με τη μεταβλητή της πάθησης(disease) μόνο για τους άνδρες(sex = female):



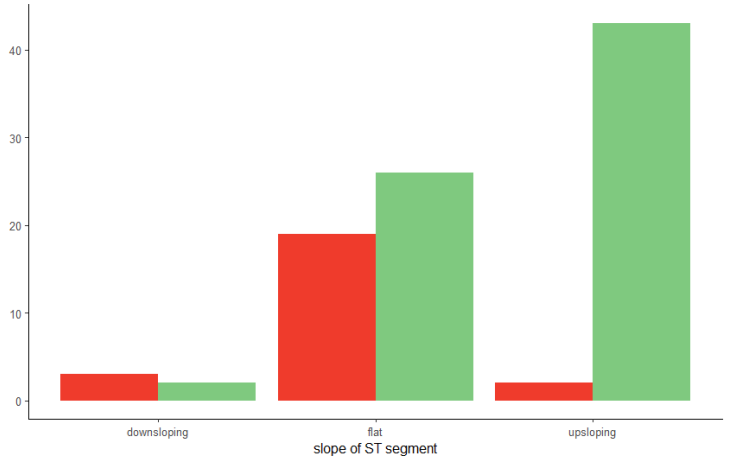
Διάγραμμα 4.33



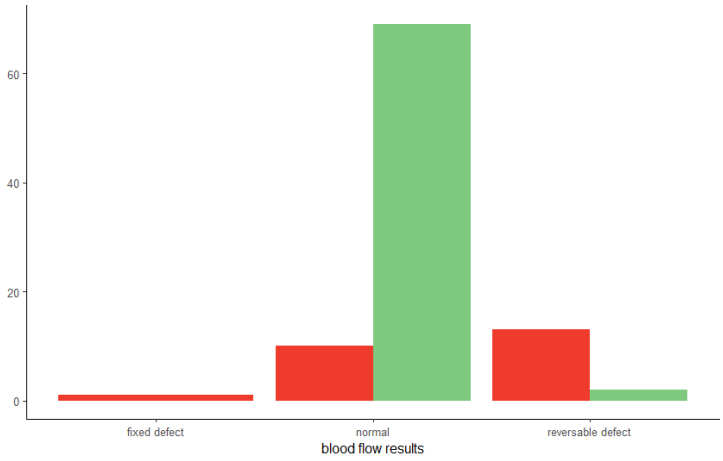
Διάγραμμα 4.34



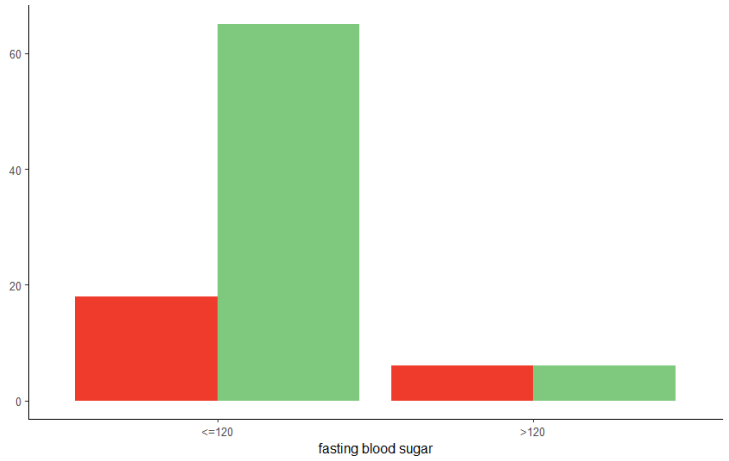
Διάγραμμα 4.35



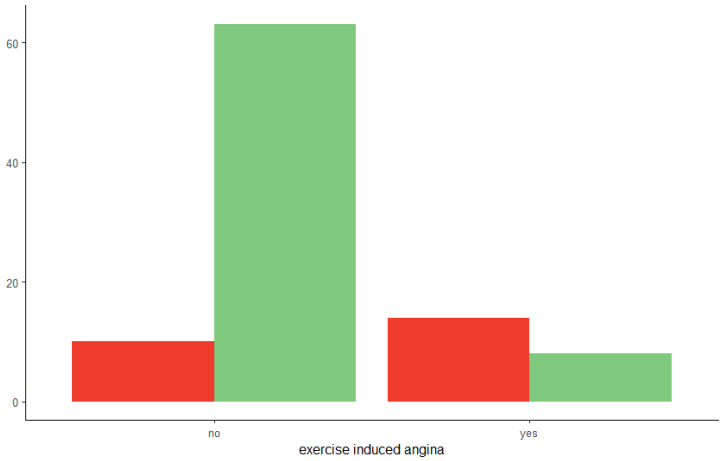
Διάγραμμα 4.36



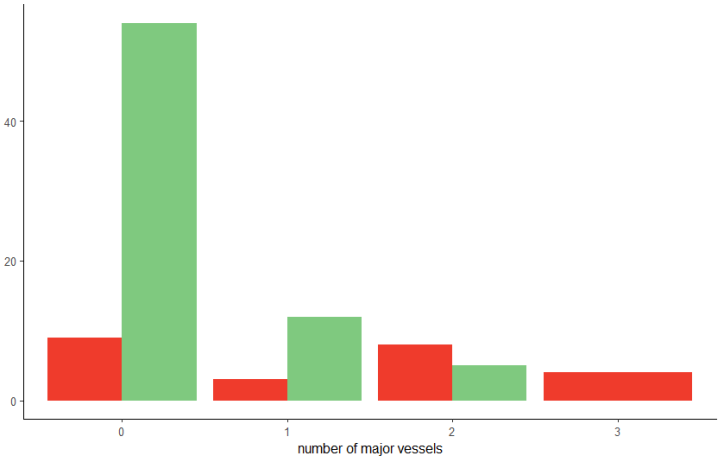
Διάγραμμα 4.37



Διάγραμμα 4.38



Διάγραμμα 4.39

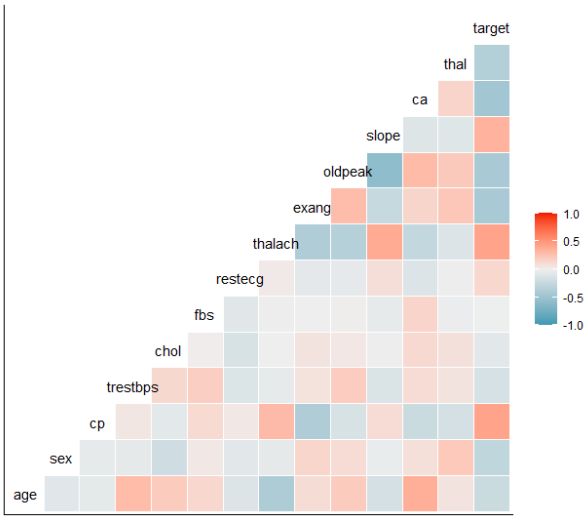


Διάγραμμα 4.40

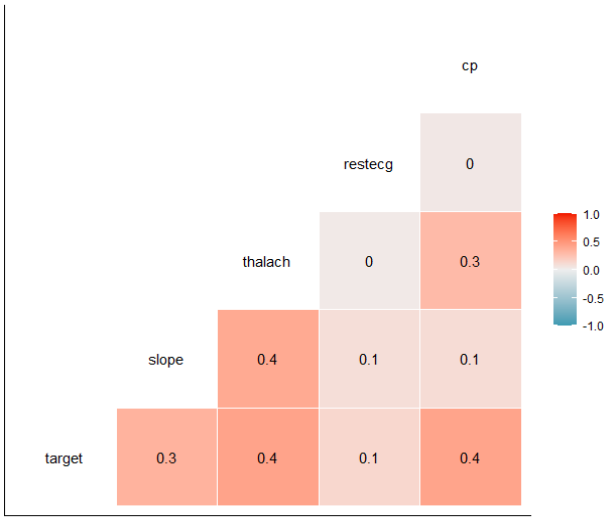
Συνεχίζοντας στο δείγμα μας με το μέρος των γυναικών παρατηρούμε πως οι περισσότερες δεν εμφανίζουν κάποια ασθένεια, ενώ όμοια με τους άνδρες, όσες εμφανίζουν ανήκουν στην κατηγορία μη συμπτωματικής στηθάγχης και παρουσιάζουν κάποιο είδος αναστρέψιμης βλάβης. Παρόλα αυτά, θα πρέπει να επισημάνουμε πως ο αριθμός των γυναικών στο δείγμα είναι σημαντικά μικρότερος.

Έτσι με μια συνολική εικόνα θα λέγαμε πως οι γυναίκες εμφανίζουν μεγαλύτερες ενδείξεις αρτηριακής πίεσης ενώ φαίνεται να έχουν όμοιο εύρος καρδιακών παλμών με αυτό των ανδρών.

Συσχετίσεις μεταβλητών(correlation plots)



Διάγραμμα 4.41



Διάγραμμα 4.42

Με τα παραπάνω διαγράμματα λαμβάνουμε και μια εικόνα των συσχετίσεων των μεταβλητών και των βαθμών συσχέτισης αυτών για περαιτέρω ανάλυση σε διαφορετικές μεταβλητές.

Κεφάλαιο 5 *Εφαρμογή Αλγορίθμων*

Σε συνέχεια λοιπόν της οπτικής ανάλυσης του δείγματος που επιλέξαμε θα εφαρμόσουμε κάποιους από τους αλγορίθμους που προαναφέραμε ώστε να προχωρήσουμε σε πρόβλεψη. Συγκεκριμένα θα εφαρμόσουμε τέσσερις αλγορίθμους ομαδοποίησης. Αυτοί θα είναι τα δέντρα αποφάσεων(decision trees), ο Naïve Bayes, ο random forest και η λογιστική παλινδρόμηση(logistic regression). Θα τρέξουμε τους συγκεκριμένους αλγορίθμους στο περιβάλλον της R, θα αναλύσουμε τα αποτελέσματα τους και εν τέλει θα προχωρήσουμε σε ένα μοντέλο πρόβλεψης, αναλύοντας την εφαρμοστικότητα του στα δεδομένα μας καθώς και την αποτελεσματικότητα του.

Ξεκινώντας επισημαίνουμε πως διαχωρίσαμε τα δεδομένα μας σε 75 τοις εκατό training και 25 τοις εκατό testing. Επιπλέον καθώς τα αποτελέσματα μας κρίνονται με τις μεταβλητές accuracy και kappa να αναφέρουμε πως η μεταβλητή kappa coefficient αναφέρεται στην ακρίβεια της διαχώρισης των δεδομένων(classification) και δέχεται τιμές από -1 έως 1. Οι τιμές που πλησιάζουν το 0 είναι οι βέλτιστες με τυχαία δεδομένα, οι αρνητικές τιμές δείχνουν πως η κατηγοριοποίηση είναι λίγο χειρότερη από τυχαία ενώ οι θετικές τιμές δείχνουν πως είναι λίγο καλύτερη από τυχαία. Ομοίως η μεταβλητή accuracy δείχνει το ποσοστό των επιτυχόντων κατηγοριοποιημένων κλάσεων. Η μέση τιμή για το εν λόγω ποσοστό είναι το 65%, ενώ τιμές μεγαλύτερες του 75% θεωρούνται αρκούντως καλές τιμές.

Έτσι εφαρμόζοντας τους αλγορίθμους έχουμε τα παρακάτω αποτελέσματα:

1. ***Δέντρο Απόφασης(decision tree)***

Για τον αλγόριθμο του δέντρου απόφασης βλέπουμε:

```
> model.tree
CART

222 samples
 13 predictor
  2 classes: 'heart-disease', 'no disease'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 200, 200, 199, 199, 200, 200, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.02941176  0.7249012  0.4460225
  0.03921569  0.7296443  0.4574207
  0.47058824  0.6126482  0.1879212

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.03921569.
```

Εικόνα 5.1

Από το δέντρο απόφασης βλέπουμε πως έχουμε θετικά αποτελέσματα, χωρίς όμως να λαμβάνουμε κάποια εξαιρετική ακρίβεια με τιμές 73% και 0.46 για τα accuracy και kappa coefficient αντίστοιχα.

2. *Naïve Bayes*

Για τον αλγόριθμο Naïve bayes βλέπουμε:

```
Naive Bayes

222 samples
 13 predictor
  2 classes: 'heart-disease', 'no disease'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 200, 200, 199, 199, 200, 200, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.8104743  0.6120596
   TRUE      0.8106719  0.6158715

Tuning parameter 'laplace' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = TRUE and adjust = 1.
```

Εικόνα 5.2

Τα αποτελέσματα του εν λόγω αλγορίθμου είναι άκρως ικανοποιητικά με ένα πολύ καλό ποσοστό ακρίβειας 81% και αρκετά θετικό kappa coefficient value.

3. *Random Forest*

Για τον αλγόριθμο Random Forest βλέπουμε:

```
Random Forest

222 samples
 13 predictor
  2 classes: 'heart-disease', 'no disease'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 200, 200, 199, 199, 200, 200, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7875494  0.5696777
  11    0.7741107  0.5425367
  21    0.7561265  0.5075259

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

Εικόνα 5.3

Ομοίως με τον αλγόριθμο Naïve Bayes έχουμε και εδώ ένα πολύ αντιπροσωπευτικό ποσοστό ακρίβειας 78% και πάλι θετικό kappa value.

4. Λογιστική Παλινδρόμηση (Logistic Regression)

Για τον αλγόριθμο της Λογιστικής Παλινδρόμησης βλέπουμε:

```
Generalized Linear Model

222 samples
 13 predictor
  2 classes: 'heart-disease', 'no disease'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 199, 200, 200, 200, 200, 199, ...
Resampling results:

  Accuracy   Kappa
  0.8337945  0.663182
```

Εικόνα 5.4

Και η λογιστική παλινδρόμηση ακολουθεί τους 2 προαναφερθέντες αλγορίθμους με ποσοστό ακρίβειας 83% περίπου και πολύ θετικό kappa value.

Με μια σύνοψη λοιπόν τον εφαρμοσμένων αλγορίθμων που προαναφέραμε έχουμε:

```
Models: LogisticReg, Tree, NaiveBayes, RandomForest
Number of resamples: 10

Kappa
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
LogisticReg  0.2786885 0.5658706 0.6440996 0.6445165 0.7896567 0.9090909    0
Tree         0.0000000 0.3774460 0.4944655 0.4574207 0.6118484 0.6333333    0
NaiveBayes   0.2204724 0.4813158 0.6239041 0.6158715 0.7919877 0.9090909    0
RandomForest 0.0000000 0.4357759 0.6332315 0.5696777 0.7740385 0.9090909    0

Accuracy
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
LogisticReg  0.6363636 0.7840909 0.8221344 0.8239130 0.8977273 0.9545455    0
Tree         0.5000000 0.6852767 0.7500000 0.7296443 0.8092885 0.8181818    0
NaiveBayes   0.5909091 0.7475296 0.8181818 0.8106719 0.8977273 0.9545455    0
RandomForest 0.5000000 0.7272727 0.8181818 0.7875494 0.8883399 0.9545455    0
```

Εικόνα 5.5

Τέλος για να δούμε τα αποτελέσματα της ανάλυσης και της πρόβλεψης μας τρέχουμε ένα confusion matrix για να συγκρίνουμε τα αποτελέσματα που λάβαμε από τις προαναφερθείσες εφαρμογές.

Για την λογιστική παλινδρόμηση βλέπουμε πως είναι πολύ κοντά τα αποτελέσματα με αυτά του cross validation. Όμοια αποτελέσματα θα έχουμε και για τα random forest, naïve bayes.

```
Confusion Matrix and Statistics

               Reference
Prediction      heart-disease no disease
  heart-disease            28          3
  no disease                6         37

               Accuracy : 0.8784
                 95% CI : (0.7816, 0.9429)
    No Information Rate : 0.5405
    P-Value [Acc > NIR] : 5.148e-10

                  Kappa : 0.7535

 Mcnemar's Test P-Value : 0.505

            Sensitivity : 0.8235
            Specificity : 0.9250
         Pos Pred Value : 0.9032
         Neg Pred Value : 0.8605
             Prevalence : 0.4595
         Detection Rate : 0.3784
   Detection Prevalence : 0.4189
      Balanced Accuracy : 0.8743

       'Positive' Class : heart-disease
```

Εικόνα 5.6

Αντίθετα για το decision tree παρατηρούμε μεγαλύτερη διαφορά που ήταν και αναμενόμενο σύμφωνα με τα προηγούμενα αποτελέσματα.

```
Confusion Matrix and Statistics

               Reference
Prediction      heart-disease no disease
  heart-disease            27          8
  no disease                7         32

               Accuracy : 0.7973
                 95% CI : (0.6878, 0.8819)
    No Information Rate : 0.5405
    P-Value [Acc > NIR] : 3.778e-06

                  Kappa : 0.5928

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7941
            Specificity : 0.8000
         Pos Pred Value : 0.7714
         Neg Pred Value : 0.8205
             Prevalence : 0.4595
         Detection Rate : 0.3649
   Detection Prevalence : 0.4730
      Balanced Accuracy : 0.7971

       'Positive' Class : heart-disease
```

Εικόνα 5.7

Κεφάλαιο 6 *Συμπεράσματα*

Πρωταρχικός στόχος της παρούσας εργασίας ήταν να δώσει την βάση και να δημιουργήσει τα θεμέλια μιας έρευνας χρησιμοποιώντας δεδομένα και μεθόδους εξόρυξης δεδομένων για την κατανόηση αλλά και την πρόβλεψη καρδιακών ασθενειών σε άτομα που εμφανίζουν συγκεκριμένα χαρακτηριστικά.

Ύστερα από την οπτική ανάλυση αλλά και την εφαρμογή των αλγορίθμων πρόβλεψης στο σύνολο των δεδομένων που επιλέξαμε να διαχειριστούμε, παρατηρήσαμε πως οι αλγόριθμοι εφαρμόζονται επιτυχώς και με αρκετά σημαντική ικανότητα πρόβλεψης.

Έτσι θα λέγαμε πως μέσω της παραπάνω εφαρμογής δίνεται η δυνατότητα της εκμάθησης των διαδικασιών και της ροής μιας διεργασίας που απαιτείται για την εφαρμογή τέτοιων μεθόδων αλλά παράλληλα δίνεται και η ικανότητα της περαιτέρω ανάπτυξης και εφαρμογής αλγορίθμων πρόβλεψης σε μεγαλύτερο εύρος δεδομένων και σε μεγαλύτερο βάθος διερεύνησης με αρκετά μεγάλη πιθανότητα επιτυχίας, όσον αφορά στα δεδομένα που σχετίζονται με τις καρδιακές παθήσεις.

Κεφάλαιο 7 *Παράρτημα*

ΚΩΔΙΚΑΣ R

7.1 Κώδικας Εισαγωγικός

```
###Libraries###

    library(tidyverse)  # For data cleaning, sorting, and visualization

    library(DataExplorer) # For Exploratory Data Analysis

    library(gridExtra) # To plot several plots in one figure

    library(ggpubr) # To prepare publication-ready plots

    library(GGally) # For correlations

    library(caTools) # For classification model

    library(rpart) # For classification model

    library(rattle) # Plot nicer descision trees

    library(randomForest) # For Random Forest model

    library(caret)

    library("rpart.plot")

    library(ggplot2)

    library(dplyr)

    library(naniar)


###Read Dataset###

    setwd("~/R/Heart Diseases Paper")

    dataset = read.csv('heart.csv')

    colnames(dataset)[colnames(dataset) == "o.Ωage"] <- "age"

    summary(dataset)


###Prepare Dataset###
```

```
datawork = dataset %>%
  mutate(
    sex = case_when(
      sex == 0 ~ "female",
      sex == 1 ~ "male"
    ),
    fbs = case_when(
      fbs == 0 ~ "<=120",
      fbs == 1 ~ ">120"
    ),
    exang = case_when(
      exang == 0 ~ "no",
      exang == 1 ~ "yes"
    ),
    cp = case_when(
      cp == 3 ~ "typical angina",
      cp == 2 ~ "non-anginal",
      cp == 1 ~ "atypical angina",
      cp == 0 ~ "asymptomatic angina"
    ),
    restecg = case_when(
      restecg == 0 ~ "hypertrophy",
      restecg == 1 ~ "normal",
      restecg == 2 ~ "wave abnormality"
    ),
    target = case_when(
      target == 1 ~ "no disease",
```

```
      target == 0 ~ "heart-disease"
    ),
    slope = case_when(
      slope == 2 ~ "upsloping",
      slope == 1 ~ "flat",
      slope == 0 ~ "downsloping"
    ),
    thal = case_when(
      thal == 1 ~ "fixed defect",
      thal == 2 ~ "normal",
      thal == 3 ~ "reversable defect"
    ),
    sex = as.factor(sex),
    fbs = as.factor(fbs),
    exang = as.factor(exang),
    cp = as.factor(cp),
    slope = as.factor(slope),
    ca = as.factor(ca),
    thal = as.factor(thal)
  )

## remove na and bad values ##
  datawork <- datawork %>%
    filter(thal != 0 & ca != 4)
  glimpse(datawork)
  summary(datawork)
```

```
nulls<-gg_miss_var(datawork) + theme_bw()
```

7.2 Κώδικας Διαγραμμάτων

```
##Variable Plots##


##1. General Density##

g1 <- plot_density(datawork, ggtheme = theme_classic2(),
geom_density_args = list("fill" = "blue", "alpha" = 0.6))

##2. By Sex##

s1 <- ggplot(datawork, aes(age))+theme_set(theme_classic2())+
geom_density(aes(fill=factor(sex)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="age")+scale_fill_manual(values = c("#ff67a4","#619cff"))+

  theme(legend.position='none')

s2 <- ggplot(datawork, aes(chol))+theme_set(theme_classic2())+
geom_density(aes(fill=factor(sex)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="cholesterol")+scale_fill_manual(values =
c("#ff67a4","#619cff"))+

  theme(legend.position='none')

s3 <- ggplot(datawork, aes(oldpeak))+theme_set(theme_classic2())+
geom_density(aes(fill=factor(sex)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",
```

```
      x="ST segment decrease")+scale_fill_manual(values =
c("#ff67a4","#619cff"))+

  theme(legend.position='none')

s4 <- ggplot(datawork, aes(thalach))+theme_set(theme_classic2())+
geom_density(aes(fill=factor(sex)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="max heart rate")+scale_fill_manual(values =
c("#ff67a4","#619cff"))+

  theme(legend.position='none')

s5 <- ggplot(datawork, aes(trestbps))+theme_set(theme_classic2())+
geom_density(aes(fill=factor(sex)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="resting electrocardiographic results")+scale_fill_manual(values
= c("#ff67a4","#619cff"))+

  theme(legend.position='none')
```

##3. By Disease##

```
cd <- ggplot(datawork, aes(target)) + geom_bar(width =
0.5,aes(fill=factor(target)), alpha=0.6) +
scale_fill_manual(values=c("#FF0000", "#00BF7D"))+

  labs(title="",

       subtitle="",

       caption="",x="disease") + theme_classic2()
+theme(legend.position='none')+

  theme(axis.text.x = element_text(angle=65, vjust=0.6))

t1 <- ggplot(datawork, aes(age))+theme_set(theme_classic())+
geom_density(aes(fill=factor(target)), alpha=0.6)+
```

```
  labs(title="",

       subtitle="",

       caption="",

       x="age") + scale_fill_manual(values=c("#FF0000",
"#00BF7D"))+

  theme(legend.position='none')

t2 <- ggplot(datawork, aes(chol))+theme_set(theme_classic())+
geom_density(aes(fill=factor(target)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="cholesterol") + scale_fill_manual(values=c("#FF0000",
"#00BF7D"))+

  theme(legend.position='none')

t3 <- ggplot(datawork, aes(oldpeak))+theme_set(theme_classic())+
geom_density(aes(fill=factor(target)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="ST segment decrease") +
scale_fill_manual(values=c("#FF0000", "#00BF7D"))+

  theme(legend.position='none')

t4 <- ggplot(datawork, aes(thalach))+theme_set(theme_classic())+
geom_density(aes(fill=factor(target)), alpha=0.6)+

  labs(title="",

       subtitle="",

       caption="",

       x="max heart rate") + scale_fill_manual(values=c("#FF0000",
"#00BF7D"))+

  theme(legend.position='none')
```

```
t5 <- ggplot(datawork, aes(trestbps))+theme_set(theme_classic())+
geom_density(aes(fill=factor(target)), alpha=0.6)+
  labs(title="",
       subtitle="",
       caption="",
       x="resting electrocardiographic results") +
scale_fill_manual(values=c("#FF0000", "#00BF7D"))+
  theme(legend.position='none')


c1 <- ggplot(datawork, aes(sex)) + geom_bar(width =
0.5,aes(fill=factor(sex)), alpha=0.8) +
scale_fill_manual(values=c("#ff67a4","#619cff"))+
  labs(title="",
       subtitle="",
       caption="",x="Sex") +
  theme_classic2() +
  theme(legend.position='none')+
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
##General boxplots##
cf1 <- ggplot(datawork, aes(sex)) + geom_bar(width =
0.5,aes(fill=factor(sex))) +
scale_fill_manual(values=c("#ff67a4","#619cff"))+
  labs(title="",
       subtitle="",
       caption="",x="Sex") + coord_flip() + theme_classic2()
+theme(legend.position='none')+
  theme(axis.text.x = element_text(angle=65, vjust=0.6))


b1 <- ggplot(datawork, aes(x= sex, y = age, fill = sex)) +
  geom_boxplot(width = 0.2) +
```

```
  labs(x="", y = "age") +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2()+
  theme(legend.position='none')

b2 <- ggplot(datawork, aes(x = sex, y = trestbps, fill = sex)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "blood pressure") +
  ylim(0,250) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2()+
  theme(legend.position='none')

b3 <- ggplot(datawork, aes(x = sex, y = chol, fill = sex)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "cholestorol") +
  ylim(0,500) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2()+
  theme(legend.position='none')

b4 <- ggplot(datawork, aes(x = sex, y = oldpeak, fill = sex)) +
  geom_boxplot(width = 0.2) +
```

```
  labs(x="", y = "ST segment decrease") +
  ylim(0,10) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2()+
  theme(legend.position='none')

b5 <- ggplot(datawork, aes(x = sex, y = thalach, fill = sex)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "max heart rate") +
  ylim(0,250) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2()+
  theme(legend.position='none')

suppressWarnings(ggarrange(cf1, b1, b2, b3, b4, b5,
                           ncol = 2, nrow = 3,
                           align = "v"))
#box males#

cm1 <- ggplot(datawork, aes(target)) + geom_bar(width =
0.5,aes(fill=factor(target))) +
scale_fill_manual(values=c("#ef3b2c","#7fc97f"))+
  labs(title="",
       subtitle="",
       caption="",x="target") + coord_flip() + theme_classic2()
+theme(legend.position='none')+
```

```
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
bm1 <- ggplot(datam, aes(x= target, y = age, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "age") +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bm2 <- ggplot(datam, aes(x = target, y = trestbps, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "blood pressure") +
  ylim(0,250) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bm3 <- ggplot(datam, aes(x = target, y = chol, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "cholestorol") +
  ylim(0,500) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')
```

```
bm4 <- ggplot(datam, aes(x = target, y = oldpeak, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "ST segment decrease") +
  ylim(0,10) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bm5 <- ggplot(datam, aes(x = target, y = thalach, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "max heart rate") +
  ylim(0,250) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

suppressWarnings(ggarrange(cm1,bm1, bm2, bm3, bm4, bm5,
                           ncol = 2, nrow = 3,
                           align = "v"))
#box females#
bfm1 <- ggplot(datafm, aes(x= target, y = age, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "age") +
  ylim(0, 90) +
```

```
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bfm2 <- ggplot(datafm, aes(x = target, y = trestbps, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "blood pressure") +
  ylim(0,250) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bfm3 <- ggplot(datafm, aes(x = target, y = chol, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "cholestorol") +
  ylim(0,500) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bfm4 <- ggplot(datafm, aes(x = target, y = oldpeak, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "ST segment decrease") +
  ylim(0,10) +
```

```
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

bfm5 <- ggplot(datafm, aes(x = target, y = thalach, fill = target)) +
  geom_boxplot(width = 0.2) +
  labs(x="", y = "max heart rate") +
  ylim(0,250) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2()+
  theme(legend.position='none')

suppressWarnings(ggarrange(cm1,bfm1, bfm2, bfm3, bfm4, bfm5,
                           ncol = 2, nrow = 3,
                           align = "v"))
###Visual per disease and sex###
gen1 <- ggplot(datawork, aes(x = target, fill = sex)) +
  geom_bar(width = 0.5, position = 'dodge') +
  labs(x = "",fill="sex") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='right')
# cp
gen2 <- ggplot(datawork, aes(cp, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
```

```
  labs(x = "chest pain type", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='none')

# restecg
gen3 <- ggplot(datawork, aes(restecg, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "resting electrocardiographic results", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='none')

# slope
gen4 <- ggplot(datawork, aes(slope, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "slope", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='none')

# thal
gen5 <- ggplot(datawork, aes(thal, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "blood flow results", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
```

```
  theme(legend.position='none')


# fbs
gen6 <- ggplot(datawork, aes(fbs, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "fasting blood sugar", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='none')


# exang
gen7 <- ggplot(datawork, aes(exang, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "exercise induced angina", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='none')


# ca
gen8 <- ggplot(datawork, aes(ca, group = sex, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "number of blood vessels", y = "") +
  scale_fill_manual(values = c("#ff67a4","#619cff"))+
  theme_classic2() +
  theme(legend.position='none')
##Male + Disease##
datam <- datawork %>% filter(sex == "male")
```

```
cm <- ggplot(datam, aes(target)) + geom_bar(width =
0.5,aes(fill=factor(target))) +  scale_fill_manual(values=c("#CC33FF",
"#66FF66"))+

  labs(title="Bar Chart",

      subtitle="Number of Male Observations",

      caption="Source:",x="Disease",

      fill="# Group") + theme_grey() +

  theme(axis.text.x = element_text(angle=65, vjust=0.6))

#boxplots#

bm1 <- ggplot(datam, aes(x= target, y = age, fill = target)) +

  geom_boxplot(width = 0.2) +

  ylim(0, 90) +

  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +

  scale_fill_manual(values = c("#386cb0","#fdb462"))+

  theme_classic2()

bm2 <- ggplot(datam, aes(x= target, y = trestbps, fill = target)) +

  geom_boxplot(width = 0.2) +

  ylim(0, 90) +

  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +

  scale_fill_manual(values = c("#386cb0","#fdb462"))+

  theme_classic2()

bm3 <- ggplot(datam, aes(x= target, y = chol, fill = target)) +

  geom_boxplot(width = 0.2) +

  ylim(0, 90) +

  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +

  scale_fill_manual(values = c("#386cb0","#fdb462"))+

  theme_classic2()

bm4 <- ggplot(datam, aes(x= target, y = oldpeak, fill = target)) +
```

```
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#386cb0","#fdb462"))+
  theme_classic2()
bm5 <- ggplot(datam, aes(x= target, y = thalach, fill = target)) +
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#386cb0","#fdb462"))+
  theme_classic2()
#Barplots#
##Men+disease##
brm1 <- ggplot(datam, aes(x = target, fill = target)) +
  geom_bar(width = 0.5, position = 'dodge') +
  labs(x = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brm2 <- ggplot(datam, aes(cp, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "chest pain", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brm3 <- ggplot(datam, aes(restecg, group = target, fill = target)) +
  geom_bar(position = "dodge") +
```

```
  labs(x = "resting electrocardiographic results", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brm4 <- ggplot(datam, aes(slope, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "slope of ST segment", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brm5 <- ggplot(datam, aes(thal, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "blood flow results", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brm6 <- ggplot(datam, aes(fbs, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "fasting blood sugar", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brm7 <- ggplot(datam, aes(exang, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "exercise induced angina", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
```

```
  theme(legend.position='none')
brm8 <- ggplot(datam, aes(ca, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "number of major vessels", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
##Female + Disease##
datafm <- datawork %>% filter(sex == "female")
cfm <- ggplot(datafm, aes(target)) + geom_bar(width =
0.5,aes(fill=factor(target))) +
scale_fill_manual(values=c("#386cb0","#fdb462"))+
  labs(title="Bar Chart",
       subtitle="Number of Male Observations",
       caption="Source:",x="Disease",
       fill="# Group") + theme_grey() +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
#boxplots#
bfm1 <- ggplot(datafm, aes(x= target, y = age, fill = target)) +
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#7fc97f","#ef3b2c"))+
  theme_classic2()
bfm2 <- ggplot(datafm, aes(x= target, y = trestbps, fill = target)) +
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
```

```
  scale_fill_manual(values = c("#7fc97f","#ef3b2c"))+
  theme_classic2()
bfm3 <- ggplot(datafm, aes(x= target, y = chol, fill = target)) +
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#7fc97f","#ef3b2c"))+
  theme_classic2()
bfm4 <- ggplot(datafm, aes(x= target, y = oldpeak, fill = target)) +
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#7fc97f","#ef3b2c"))+
  theme_classic2()
bfm5 <- ggplot(datafm, aes(x= target, y = thalach, fill = target)) +
  geom_boxplot(width = 0.2) +
  ylim(0, 90) +
  stat_compare_means(aes(label = ..p.signif..), method = "t.test") +
  scale_fill_manual(values = c("#7fc97f","#ef3b2c"))+
  theme_classic2()
#Barplots#
bfm1 <- ggplot(datafm, aes(x = target, fill = target)) +
  geom_bar(width = 0.5, position = 'dodge') +
  labs(x = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
```

```
brfm2 <- ggplot(datam, aes(cp, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "chest pain", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brfm3 <- ggplot(datafm, aes(restecg, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "resting electrocardiographic results", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brfm4 <- ggplot(datafm, aes(slope, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "slope of ST segment", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brfm5 <- ggplot(datafm, aes(thal, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "blood flow results", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brfm6 <- ggplot(datafm, aes(fbs, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "fasting blood sugar", y = "") +
```

```
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brfm7 <- ggplot(datafm, aes(exang, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "exercise induced angina", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')
brfm8 <- ggplot(datafm, aes(ca, group = target, fill = target)) +
  geom_bar(position = "dodge") +
  labs(x = "number of major vessels", y = "") +
  scale_fill_manual(values = c("#ef3b2c","#7fc97f"))+
  theme_classic2() +
  theme(legend.position='none')

#Correlations#
dw <- dataset %>% filter(
  thal != 0 & ca != 4
)
corr <- round(cor(dw), 1)
GGally::ggcorr(dw, geom = "circle")
select2 <- dw %>%
  dplyr::select(
    target,
    slope,
    thalach,
```

```
    restecg,
    cp
  )
ggcorr(select2, geom = "circle")
```

7.3 Κώδικας Αλγορίθμων

```
###Prediction Algorithms###
set.seed(10)
training_indeces <- createDataPartition(datawork$target, p = .75, list = FALSE)
datawork.train <- datawork[ training_indeces,]
datawork.test  <- datawork[-training_indeces,]
fitControl <- trainControl(method="cv", number=10)
#Logistic Regression#
set.seed(10)
model.lr <- train(target ~ .,
                  data = datawork.train,
                  method = "glm",
                  family=binomial(),
                  trControl = fitControl)
model.lr
#Decision Tree#
set.seed(10)
model.tree <- train(target ~ .,
                    data = datawork.train,
                    method = "rpart",
                    trControl = fitControl)
model.tree
```

```
        plot(model.tree)
        rpart.plot(model.tree$finalModel)
#Naive Bayes#
        set.seed(10)
        model.nb <- train(target ~ .,
                          data = datawork.train,
                          method = "naive_bayes",
                          trControl = fitControl)
        model.nb
        plot(model.nb)
#Random Forest#
        set.seed(10)
        model.rf <- train(target ~ .,
                          data = datawork.train,
                          method = "rf",
                          trControl = fitControl)
        model.rf
        plot(model.rf)

#Summary#
        model.all <- list(LogisticReg=model.lr, Tree=model.tree,
        NaiveBayes=model.nb, RandomForest=model.rf)
        results <- resamples(model.all)
        summary(results, metric=c("Kappa", "Accuracy"))

#Prediction#
        preds <- predict(model.lr, datawork.test)
```

```
confusionMatrix(preds, as.factor(datawork.test$target))


preds.tree <- predict(model.tree, datawork.test)
confusionMatrix(preds.tree, as.factor(datawork.test$target))
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

- ΒΙΒΛΙΑ
  i. Larose, D. and Larose, C., n.d. *Discovering knowledge in data*.
  ii. Hastie, T., Tibshirani, R. and Friedman, J., n.d. *The elements of statistical learning*.
  iii. Han, J., Kamber, M. and Pei, J., 2012. *Data mining concepts and techniques*. Amsterdam: Elsevier.
  iv. Field, A., Miles, J. and Field, Z., 2013. *Discovering statistics using R*. Los Angeles [u.a.]: SAGE.
  v. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2017. *An introduction to statistical learning*. New York: Springer.
  vi. Murphy, K., 2012. *Machine learning*. Cambridge, MA: MIT Press.
  vii. Ye, N., n.d. *Data mining*.
  viii. Kantardzic, M., n.d. *Data mining*.
  ix. Witten, I. and Frank, E., n.d. *Data Mining*. San Diego: Elsevier Science & Technology Books.

- ΑΡΘΡΑ
  i. Lavrac, N., n.d. *Machine Learning for Data Mining in Medicine*.
  ii. Khourdifi, Y. and Bahaj, M., 2018. *Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization*.
  iii. Ramalingam, V., Dandapath, A. and Raja, M., 2018. *Heart disease prediction using machine learning techniques : a survey*.
  iv. ΠΑΠΑΘΑΝΑΣΙΟΥ, Β., 2019. *Ανάλυση Κλινικών Δεδομένων με χρήση των Αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη του καρκίνου*.

- ΑΛΛΑ
  i. Who.int. 2022. *Home*. [online] Available at: <https://www.who.int/> [Accessed 2 September 2021].
  ii. HRT.org. 2022. *Health Organization (HRT.ORG) Trusted Health and Wellness Information*. [online] Available at: <https://www.hrt.org> [Accessed 2 September 2021].
  iii. Towards Data Science. 2022. *Towards Data Science*. [online] Available at: <https://towardsdatascience.com> [Accessed 2 September 2021].
  iv. GeeksforGeeks. 2022. *GeeksforGeeks | A computer science portal for geeks*. [online] Available at: <https://www.geeksforgeeks.org/> [Accessed 2 September 2021].
  v. Msdmanuals.com. 2022. *The Trusted Provider of Medical Information since 1899*. [online] Available at: <https://www.msdmanuals.com> [Accessed 2 September 2021].
  vi. Kaggle.com. 2022. *A detail description of the Heart Disease dataset*. [online] Available at: <https://www.kaggle.com/carlosdg/a-detail-description-of-the-heart-disease-dataset> [Accessed 2 September 2021].

vii. Kaggle.com. 2022. *Predicting Heart Disease Risk with Random Forest*. [online] Available at: <https://www.kaggle.com/wguesdon/predicting-heart-disease-risk-with-random-forest> [Accessed 2 September 2021].