ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»

ΕΙΔΙΚΕΥΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Ιατρικών Δεδομένων

Αγγελίδης Χρήστος

Α.Μ: 1801

Επιβλέπων: Ηλίας Μαγκλογιάννης, Καθηγητής

ΠΕΙΡΑΙΑΣ
ΦΕΒΡΟΥΑΡΙΟΣ 2021

Περίληψη

Η παρούσα εργασία ασχολείται με την μελέτη και εξαγωγή συμπερασμάτων από δεδομένα νεφροπαθών ασθενών. Χρησιμοποιώντας Περιγραφική Στατιστική διερευνώνται πτυχές της ασθένειας, όπως ο συχνότερος τρόπος θανάτου των ασθενών, τα φάρμακα που συνταγογραφούνται, σε ποιες ηλικίες υπάρχει μεγαλύτερη θνησιμότητα κ.α.

Ακόμη χρησιμοποιούνται τεχνικές μηχανικής μάθησης ώστε να προβλεφθεί ο χρόνος ζωής των ασθενών, η αιτία θανάτου, καθώς και η οικονομική επιβάρυνση που προκαλεί η συγκεκριμένη ασθένεια στους ασθενείς, χρησιμοποιώντας χαρακτηριστικά που εξήχθησαν μέσω τεχνικών επιλογής χαρακτηριστικών (Feature Selection).

Επιπρόσθετα γίνεται ανάλυση θνησιμότητας ώστε να διαπιστωθεί ο χρόνος ζωής των ασθενών ανάλογα με το στάδιο που βρίσκονται, είτε σε αιμοκάθαρση είτε σε μεταμόσχευση. Τέλος χρησιμοποιώντας Causal Analysis διερευνήθηκε αν το φάρμακο ‘Trombyl’, ανάλογα με την δοσολογία του επιμηκύνει τη διάρκεια ζωής των ασθενών.

Abstract

The current Thesis is dealing with the study and analysis of data of patients suffering from Chronic Kidney Disease. Using Descriptive Statistics, aspects of the disease are investigated, such as the most common cause of death, the prescribed drugs and at what ages there is a higher mortality, etc.

Machine learning techniques are also used to predict patients' life expectancy, cause of death and the financial burden of the disease on patients, using features that were refined through Feature Selection techniques.

In addition, a Survival analysis is performed to determine the life expectancy of patients depending on the stage they are in, either on dialysis or transplant. Finally, using Causal Analysis, it was investigated whether the drug ‘Trombyl’, depending on its dosage, prolongs the life of patients.

Ευχαριστίες

Για την υλοποίηση της διπλωματικής εργασίας θα ήθελα να ευχαριστήσω θερμά τον κ. Μαγκλογιάννη Ηλία, ο οποίος είναι καθηγητής του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς, τόσο για την επίβλεψή της εργασίας, όσο και για την πολύτιμη βοήθειά του καθ' όλη την διάρκεια της υλοποίησης της εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τον στενά οικογενειακό και φιλικό μου κύκλο για την στήριξη που μου πρόσφεραν σε όλη την εκπαιδευτική μου πορεία μέχρι σήμερα και ελπίζω να συνεχίσουν να το κάνουν.

Πίνακας περιεχομένων

Λίστα Πινάκων

Λίστα Εικόνων

1. Εισαγωγή

Η χρόνια νεφρική ανεπάρκεια αναγνωρίζεται ως ένα μείζον πρόβλημα υγείας που ταλαιπωρεί την παγκόσμια κοινότητα υγείας. Η τεχνολογική εξέλιξη των τελευταίων χρόνων έχει προκαλέσει ραγδαία αύξηση του συνόλου των διαθέσιμων δεδομένων καθώς και ανάγκη για αξιοποίηση αυτών για την εξαγωγή αποτελεσμάτων. Οι αλγόριθμοι Μηχανικής Μάθησης σε συνδυασμό με τη χρήση τεχνικών Feature Selection προσφέρουν τη δυνατότητα εξαγωγής συμπερασμάτων για τις παραμέτρους που επηρεάζουν την ασθένεια καθώς και την πρόβλεψη για την εξέλιξη της ασθένειας μέσα από μοτίβα στα δεδομένα των ασθενών. Η πρόβλεψη του χρόνου ζωής των ασθενών δίνει τη δυνατότητα για έγκυρη διαβούλευση και έγκυρη θεραπεία. Ακόμη βοηθά στην έγκυρη αναγνώριση και διόρθωση αντιστρέψιμων παραγόντων. Η πρόβλεψη της αιτίας θανάτου των ασθενών βοηθάει στην εύρεση των μεταβλητών που επηρεάζουν την πορεία του ασθενή προς τον εκάστοτε θάνατο και δίνει την δυνατότητα τροποποίησης των διαδικασιών ώστε να αποφευχθεί. Το κόστος θεραπείας είναι ένα πρόβλημα που έχουν να αντιμετωπίσουν ασθενείς με μακροχρόνιες ασθένειες. Η πρόβλεψη του για μακροχρόνιες ασθένειες σε συνδυασμό με την επιλογή κατάλληλων μεταβλητών που το επηρεάζει προσφέρει την δυνατότητα προσαρμογής των διαδικασιών ώστε να μειωθεί. Ακόμη η γνώση αυτή βοηθά στον καλύτερο προγραμματισμό από την πλευρά των ασθενών και εύρεση λιγότερο δαπανηρών θεραπειών σε συνδυασμό πάντα με την προτροπή του γιατρού.

Χρησιμοποιώντας Περιγραφική Στατιστική εξάγονται χρήσιμα συμπεράσματα που χρησιμοποιούνται αργότερα για να λυθούν τα υπόλοιπα προβλήματα που τέθηκαν στην παρούσα εργασία.

Παράλληλα, με τη χρήση Ανάλυσης Επιβίωσης σε χρόνιες ασθένειες δίνεται η δυνατότητα της μελέτης του ρυθμού επιβίωσης των ασθενών με χρόνια ασθένεια.

Η Ανάλυση Επιβίωσης αναφέρεται στην ανάλυση δεδομένων που αφορούν στο χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν. Σε όλη την διάρκεια ο ασθενής παρακολουθείται. Η Ανάλυση Επιβίωσης χρησιμοποιείται σε μελέτες όπως οι μελέτες Καρκίνου και απαντάει σε ερωτήματα όπως ποια είναι η επίδραση χαρακτηριστικών όπως φαρμάκων στην επιβίωση των ασθενών ή πόσοι ασθενείς επιβίωσαν σε ένα χρονικό διάστημα κ.α. Ακόμη μελετώνται παράμετροι που επηρεάζουν το ποσοστό θνησιμότητας των ασθενών. Με αυτόν τον τρόπο οι ασθενείς γνωρίζουν το χρόνο ζωής τους και οι επιστήμονες με την χρήση μεταβλητών, όπως φάρμακα ή αλλαγή κατάστασης από Αιμοκάθαρση σε μεταμόσχευση ή και αντίστροφα, προσπαθούν να παρατείνουν τον χρόνο ζωής όσο το δυνατόν περισσότερο.

Υπάρχουν αρκετές τεχνικές για Ανάλυση επιβίωσης. Στην παρούσα εργασία θα μελετηθεί η τεχνική Kaplan-Meier. Η συγκεκριμένη τεχνική χρησιμοποιείται στην πρόβλεψη του ποσοστού θανάτου ασθενών από συγκεκριμένη ασθένεια καθώς και η επιρροή χαρακτηριστικών όπως το φύλο στην επιβίωση των ασθενών.

Η χρήση Αιτιολογικής Ανάλυσης χρησιμοποιείται ευρέως στην εξέταση επιρροής αιτιών όπως η χορήγηση φαρμάκων στην βελτίωση της κατάστασης των ασθενών με χρόνια ασθένεια. Ακόμη εξετάζεται η επιρροή χαρακτηριστικών όπως το φύλο ή η λήψη θεραπείας στην βελτίωση της κατάστασης των ασθενών. Σύμφωνα με τα δεδομένα της παρούσας εργασίας το πιο συχνά συνταγογραφούμενο φάρμακο είναι το “Trombyl”. Χρησιμοποιώντας Αιτιολογική Ανάλυση ελέγχεται κατά πόσο το συγκεκριμένο φάρμακο παρατείνει την διάρκεια ζωής των ασθενών σε ηλικίες 31-85.

Ακόμη χρησιμοποιούνται χαρακτηριστικά όπως η ηλικία και η διάγνωση στην επιρροή του αποτελέσματος. Υπάρχουν πολλές τεχνικές Αιτιολογικής Ανάλυσης. Στην παρούσα εργασία θα μελετηθούν 3, η τεχνική Ordinary Least Squares(OLS), η τεχνική Nearest Neighbour Matching, η τεχνική weighting. Θα γίνει χρήση Propensity score καθώς και τεχνική trimming για βελτίωση της απόδοσης των αλγορίθμων.

Όλα τα συμπεράσματα που θα προκύψουν βοηθούν στην καλύτερη κατανόηση της νεφρικής ανεπάρκειας δίνοντας απάντηση στα ερωτήματα που τέθηκαν και θα δώσουν την δυνατότητα καλύτερης αντιμετώπισής της.

Τα επόμενα κεφάλαια είναι οργανωμένα ως εξής: Στο δεύτερο κεφάλαιο παρουσιάζεται το τεχνολογικό υπόβαθρο καθώς και σχετικές εργασίες με το αντικείμενο της διπλωματικής και η μεθοδολογία αυτών. Στο τρίτο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο των τεχνικών που χρησιμοποιήθηκαν. Στο τέταρτο κεφάλαιο παρουσιάζεται η μεθοδολογία που ακολουθήθηκε για την απάντηση των ερευνητικά ερωτημάτων. Στο πέμπτο παρουσιάζονται τα πειράματα που έγιναν καθώς και η αξιολόγηση των πειραματισμών. Στο έκτο κεφάλαιο αναφέρονται τα συμπεράσματα της έρευνας. Τέλος στο έβδομο κεφάλαιο αναφέρονται προβληματισμοί για το μέλλον.

2. Υπόβαθρο

2.1. Μεγάλα Δεδομένα

Η ραγδαία ανάπτυξη του διαδικτύου ήταν ο κύριος παράγοντας που δημιούργησε μια συνεχόμενη ροή δεδομένων με αποτέλεσμα ο όγκος των δεδομένων να αυξηθεί κατακόρυφα και έτσι να δημιουργηθεί η έννοια των μεγάλων δεδομένων. Βεβαίως πριν από αυτό υπήρχαν και άλλοι παράγοντες που επηρέασαν την αύξηση του όγκου των δεδομένων. Το μεγαλύτερο μέρος των μεγάλων δεδομένων που παράγονται προέρχεται από τρεις κύριες πηγές: κοινωνικά δεδομένα, δεδομένα μηχανών και δεδομένα συναλλαγών. Τα κοινωνικά δεδομένα προέρχονται από τα Likes, Tweets & Retweets, σχόλια, μεταφορτώσεις βίντεο και γενικά μέσα που μεταφορτώνονται και κοινοποιούνται μέσω των αγαπημένων πλατφορμών κοινωνικών μέσων. Αυτό το είδος δεδομένων παρέχει ανεκτίμητες πληροφορίες σχετικά με τη συμπεριφορά και το συναίσθημα των καταναλωτών. Τα δεδομένα των μηχανών ορίζονται ως πληροφορίες που παράγονται από βιομηχανικό εξοπλισμό, αισθητήρες που είναι εγκατεστημένοι σε μηχανήματα, ακόμη και αρχεία καταγραφής ιστού που παρακολουθούν τη συμπεριφορά των χρηστών. Αυτός ο τύπος δεδομένων αναμένεται να αυξηθεί εκθετικά καθώς οι έξυπνες συσκευές αυξάνονται όλο και περισσότερο. Αισθητήρες όπως ιατρικές συσκευές, έξυπνοι μετρητές, κάμερες δρόμου, δορυφόροι, παιχνίδια και το ταχύτατα αναπτυσσόμενο θα προσφέρουν υψηλή ταχύτητα, αξία, όγκο και ποικιλία δεδομένων στο εγγύς μέλλον. Τα δεδομένα συναλλαγών δημιουργούνται από όλες τις καθημερινές συναλλαγές που πραγματοποιούνται τόσο διαδικτυακά όσο και εκτός σύνδεσης. Τιμολόγια, εντολές πληρωμής, αρχεία αποθήκευσης, αποδείξεις παράδοσης - όλα χαρακτηρίζονται ως δεδομένα συναλλαγών, αλλά μόνο τα δεδομένα είναι σχεδόν χωρίς νόημα και οι περισσότεροι οργανισμοί αγωνίζονται να κατανοήσουν τα δεδομένα που δημιουργούν και πώς μπορούν να χρησιμοποιηθούν σωστά. Η αρχή όλων αυτών έγινε με την αρχή της ψηφιοποίησης. Η ανάγκη για ψηφιοποίηση όπως βιβλιοθηκών δημιούργησε μεγάλο όγκο δεδομένων. Ακόμη η δημιουργήθηκε αύξηση των πληροφοριών στον τομέα της επιστήμης και στον επιχειρηματικό τομέα. Λόγω της εισροής πληροφοριών οι περισσότεροι οργανισμοί άρχισαν να σχεδιάζουν, να αναπτύσσουν και να εφαρμόζουν κεντρικά

υπολογιστικά συστήματα που τους επέτρεπαν να αυτοματοποιήσουν τα συστήματά τους απογραφής τους. Επιπλέον η ανάγκη εταιρειών να χρησιμοποιήσουν τα δεδομένα για να εξάγουν πληροφορίες που θα βοηθούσαν στην λήψη σωστότερων και πιο επικερδών αποφάσεων δημιούργησε την ανάγκη μεγαλύτερης αποθήκευσης των δεδομένων και ανάπτυξη τρόπων επίτευξης του σκοπού αυτού. Πειράματα όπως αυτά στο CERN παράγουν ένα μεγάλο όγκο δεδομένων εδώ και αρκετές δεκαετίες. Τα τελευταία χρόνια η ραγδαία ανάπτυξη των τεχνολογιών όπως έξυπνα τηλέφωνα, ταμπλέτες και ασύρματες συνδέσεις Wi-Fi, παράγονται μεγάλοι όγκοι δεδομένων με ξέφρενους ρυθμούς. Τα μεγάλα δεδομένα συναντώνται παντού στην σημερινή εποχή. Επιπρόσθετα η ανάπτυξη των υπολογιστικών δυνατοτήτων και το χαμηλό κόστος αυτών καθώς και η ανάπτυξη μεγάλος όγκος την πληροφορίας και η συνεχόμενη αύξηση του μεγέθους των δεδομένων καθιστά αναγκαία την χρήση τεχνικών επεξεργασίας μεγάλων δεδομένων.

Ειδικότερα στον τομέα της Ιατρικής η ευρεία χρήση έξυπνων συσκευών αύξησε κατακόρυφα τον όγκο των δεδομένων βοηθώντας και στην αποτελεσματικότερη λειτουργία του κλάδου. Χρησιμοποιώντας τον μεγάλο όγκο των δεδομένων δημιουργήθηκαν συστήματα τα οποία μπορούν μέσα από μοτίβα που εξάγουν από τα δεδομένα να προτείνουν ενέργειες σύμφωνα με τις απαιτήσεις του ασθενή. Σύμφωνα με τον (Sahoo, 2019) δημιουργήθηκε ένα σύστημα συστάσεων για την υγεία το οποίο χρησιμοποιεί την συλλογή μεγάλου όγκου δεδομένων για να εκπαιδευτεί και να μπορεί να προτείνει ενέργειες. Το σύστημα χρησιμοποιεί έναν συνδυασμό από αλγορίθμους Μηχανικής Μάθησης όπως Δέντρα Απόφασης, κατηγοριοποιητές Bayes και αλγορίθμους βαθιάς μάθησης όπως τα Νευρωνικά Δίκτυα για να κατηγοριοποιήσει τα δεδομένα που λαμβάνει από τους χρήστες του συστήματος. Έπειτα αποθηκεύει τα αποτελέσματα σε μια βάση. Χρησιμοποιεί τεχνικές επιλογής χαρακτηριστικών για να εξάγει τις κατάλληλες μεταβλητές που θα χρησιμοποιηθούν στην εκπαίδευση των αλγορίθμων Μηχανικής Μάθησης. Προσθέτως χρησιμοποιώντας τεχνικές filtering όπως η τεχνική content-based filtering εστιάζει στην ανάλυση των χαρακτηριστικών των προϊόντων για τη δημιουργία προβλέψεων. Μια άλλη τεχνική filtering που χρησιμοποιείται είναι η τεχνική Collaborative-based filtering η οποία χρησιμοποιεί αξιολογήσεις βάσει χρηστών για να βρει ομοιότητα μεταξύ αντικειμένων. Μετά τη συλλογή όλων των αξιολογήσεων των χρηστών, το σύστημα συγκρίνει αυτές τις αξιολογήσεις με άλλους χρήστες με τη βοήθεια ενός πίνακα βοηθητικών προγραμμάτων και προτείνει στοιχεία με κορυφαία βαθμολογία στον χρήστη. Χρησιμοποιώντας τα δεδομένα που εξήχθησαν από την παραπάνω διαδικασία συστήνει κατάλληλες θεραπείες στους ασθενείς. Οι κατηγορίες στις οποίες προτείνει ενέργειες είναι:

- **Διατροφικά δεδομένα**: Η δημιουργία συστάσεων για την αύξηση της διατροφής. Ο γιατρός μπορεί να αλλάξει τις διατροφικές συνήθειες, έτσι ώστε οι ασθενείς να λαμβάνουν σωστή διατροφή, ώστε να αναρρώσει από ασθένεια ή ασθένεια. Οι συστάσεις θα μπορούσαν να είναι ισορροπημένο φαγητό, υποκατάστατα είδη διατροφής, λιγότερο πικάντικα γεύματα ή προσθήκες σε μια διατροφή.
- **Σωματική άσκηση**: Προτεινόμενες ενέργειες για το είδος της γιόγκα και τη σωματική άσκηση που πρέπει να κάνουν οι ασθενείς για γρήγορη ανάρρωση βάσει των απαιτήσεων των χρηστών. Οι απαιτήσεις του χρήστη μπορεί να περιλαμβάνουν τοποθεσία, που σχετίζεται με ασθένειες, καιρό κ.λπ.
- **Διάγνωση**: Δημιουργία συστάσεων για τη διάγνωση ασθενών από το γιατρό με βάση τα συμπτώματα που εμφανίζονται σε παρόμοιες περιπτώσεις.

- **Θεραπεία / Φάρμακα**: Δημιουργία συστάσεων σχετικά με διαφορετικούς τύπους φαρμάκων για μια συγκεκριμένη ασθένεια ή συγκεκριμένη θεραπεία για τον ασθενή.

Το συγκεκριμένο σύστημα μπορεί να χρησιμοποιηθεί από γιατρούς ή ερευνητές ή ασθενείς καθώς και φαρμακοποιούς. Παρέχει τη δυνατότητα καταγραφής των ζωτικών ενδείξεων του ασθενή όπως οι χτύποι της καρδιάς με αποτέλεσμα τη μείωση των ιατρικών τεστ. Χρησιμοποιώντας τον μεγάλο όγκο των δεδομένων το σύστημα εκπαιδεύτηκε ώστε να διαθέτει ακρίβεια της τάξης του 63%, ποσοστό που είναι αρκετά ακριβές για βοηθητική χρήση στη λήψη αποφάσεων. Η ανάλυση Μεγάλων Δεδομένων δημιουργεί ευκαιρίες για την ανάπτυξη τέτοιων συστημάτων καθώς και προκλήσεις για περαιτέρω ανάπτυξη. Επιπρόσθετα, η ανάλυση μεγάλων δεδομένων μπορεί να βοηθήσει στην προληπτική ιατρική. Σύμφωνα με (Razzak, 2020) η χρήση των ιατρικών δεδομένων μεγάλου όγκου σε συνδυασμό με αλγορίθμους Μηχανικής Μάθησης χρησιμοποιούνται στην πρόβλεψη ασθενειών. Ακόμη η χρήση αλγορίθμων Clustering με την χρήση των μεγάλων δεδομένων βοηθά στην κατηγοριοποίηση των ασθενειών βασισμένοι σε μοτίβα που προκύπτουν από τα δεδομένα. Επιπλέον χρησιμοποιώντας Association Analysis προκύπτουν συνδέσεις από τα δεδομένα όπως για παράδειγμα η σχέση εγκυμοσύνης και διαβήτη.

2.2. Μηχανική Μάθηση

Έχουν γίνει αρκετές έρευνες που χρησιμοποιούν ιατρικά δεδομένα και προσπαθούν μέσα από αυτά να εξάγουν σημαντικές πληροφορίες. Όπως παρουσιάζεται στο (Akbilgic, 2019), χρησιμοποιώντας δημογραφικά και κλινικά δεδομένα έγινε σύγκριση του αλγορίθμου Random Forest με άλλους αλγορίθμους Μηχανικής Μάθησης οι οποίοι ήταν ο K- Nearest Neighbours, Logistic Regression, Νευρωνικά Δίκτυα καθώς και Support Vector Machines. (Dey, 2016) Στόχος ήταν η ανάπτυξη ενός προβλεπτικού μοντέλου χρησιμοποιώντας αλγόριθμους Μηχανικής Μάθησης. Πιο συγκεκριμένα χρησιμοποιήθηκε ο αλγόριθμος Random Forest, ο οποίος και προσαρμόστηκε στα δεδομένα. Οι ασθενείς που χρησιμοποιήθηκαν ήταν ασθενείς που μπήκαν σε κατάσταση αιμοκάθαρσης σε περίοδο 30,60 και 90 ημερών. Ακόμη οι ασθενείς χωρίστηκαν ανάλογα με τον ταχυδρομικό κώδικα της περιοχής τους σε εννιά κατηγορίες. Χρησιμοποιήθηκαν 49 μοντέλα Random Forest που περιείχαν 500 δέντρα απόφασης το καθένα. Σαν μεταβλητή πρόβλεψης χρησιμοποιήθηκε μια τιμή που αναπαριστούσε το χρόνο από την αρχή της αιμοκάθαρσης μέχρι το θάνατο των ασθενών. Αφού δημιουργήθηκαν τα μοντέλα πρόβλεψης και, για να εκτιμηθεί πόσο καλά οι προβλεπόμενες πιθανότητες αντικατοπτρίζουν το πραγματικό ποσοστό επιβίωσης, χωρίστηκαν όλες οι προβλεπόμενες τιμές πιθανότητας από τα τελικά μοντέλα σε 5 συστάδες με βάση τα αυξανόμενα διαστήματα του προβλεπόμενου κινδύνου: 0 έως 5, 5 έως 35, 35 έως 65, 65 έως 95 και 95 έως 100 εκατοστημόρια. Για κάθε ομάδα κινδύνου, υπολογίστηκε το ποσοστό επιβίωσης και συγκρίθηκε γραφικά με το παρατηρούμενο ποσοστό επιβίωσης. Σύμφωνα με τα αποτελέσματα της έρευνας ο αλγόριθμος Random Forest αποδείχθηκε η καλύτερη επιλογή ως προς την ακρίβεια ταξινόμησης όταν τα δεδομένα χωρίστηκαν με βάση τον ταχυδρομικό κώδικα της περιοχής τους. Το μοντέλο που αναπτύχθηκε έχει χρησιμοποιώντας C-statistics(95% διάστημα εμπιστοσύνης) σαν μέτρο ακρίβειας, είχε ακρίβεια 0.7185 (0.6994–0.7377) για την πρόβλεψη ρίσκου θνησιμότητας μέσα σε 30 μέρες, 0.7446 (0.7346–0.7546) για την πρόβλεψη ρίσκου θνησιμότητας μέσα σε 90 μέρες, 0.7504 (0.7425–0.7583) για την πρόβλεψη ρίσκου θνησιμότητας μέσα σε 180 μέρες και 0.7488 (0.7421–0.7554) για την πρόβλεψη ρίσκου θνησιμότητας μέσα σε 365 μέρες. Αν και το μοντέλο που αναπτύχθηκε φαίνεται να είναι ακριβές δεν παύει να έχει και αδυναμίες. Επειδή στα δεδομένα που

χρησιμοποιήθηκαν τα δείγματα ήταν άντρες ίσως το μοντέλο δεν μπορεί να γενικεύει και σε άλλες περιπτώσεις όπως για γυναίκες. Ακόμη η έλλειψη κλινικών πληροφοριών όπως οροθετικότητα ηπατίτιδας C.39 περιόριζε την ακόμη καλύτερη απόδοση του μοντέλου. Όλοι αυτοί οι περιορισμοί εμποδίζουν την περαιτέρω γενίκευση του μοντέλου. Σύμφωνα με (Khalilia, 2011) ο αλγόριθμος Random Forest παρουσιάζει καλή απόδοση σε imbalanced ιατρικά δεδομένα (Chen, 2004). Ο αλγόριθμος χρησιμοποιήθηκε για την πρόβλεψη οκτώ κατηγοριών ασθενειών. (Healthcare Cost and Utilization Project. "Overview of the nationwide inpatient sample (NIS), 2020) Χρησιμοποιήθηκαν ακόμη ο αλγόριθμος SVM και μέθοδοι bagging and boosting. (Chandrahasan, 2011) Χρησιμοποιώντας σαν μέτρο σύγκρισης την καμπύλη ROC (Bradley, 1997) αποδείχθηκε ότι ο αλγόριθμος Random Forest έχει την καλύτερη απόδοση. Ακόμη εξετάστηκε και η απόδοση των αλγορίθμων σε sampling και non-sampling ταξινόμηση. Τα αποτελέσματα έδειξαν ότι η απόδοση των αλγορίθμων με sampling είναι καλύτερη. Από τα αποτελέσματα παρατηρήθηκε ότι χρησιμοποιώντας τυχαίο sub-sampling οι αλγόριθμοι έχουν καλύτερα αποτελέσματα σε imbalanced δεδομένα. Κάποιοι περιορισμοί υπήρξαν και σε αυτή την εργασία. Για παράδειγμα δεν υπήρχε αναγνώριση της μοναδικότητας των εγγραφών οπότε εγγραφές χρησιμοποιήθηκαν πολλές φορές.

Σύμφωνα με (Luo, 2019)) χρησιμοποιώντας αλγόριθμους Μηχανικής Μάθησης, μπορεί να προβλεφθεί ποιοι ασθενείς είναι σε μεγάλο οικονομικό ρίσκο όσον αφορά το κόστος θεραπείας. Χρησιμοποιήθηκαν αλγόριθμοι όπως Random Forest, Logistic Regression and XGBoost. (Ichikawa, 2016) Η σύγκριση των μεθόδων Μηχανικής Μάθησης και η εύρεση των τελικών χαρακτηριστικών για τη δημιουργία προγνωστικών μοντέλων έγινε βασισμένη σε τρεις μετρικές, η περιοχή κάτω από τη χαρακτηριστική καμπύλη (AUC), ευαισθησία (SEN) και ειδικότητα (SPE). (Rajkomar, 2018), (Spathis, 2019) Αυτή η μελέτη χρησιμοποίησε δεδομένα ιατρικής ασφάλισης από μια μεγάλη πόλη στη δυτική Κίνα για την περίοδο από την 1η Ιανουαρίου 2011 έως τις 31 Δεκεμβρίου 2013. Εξήχθησαν δεδομένα από τα αρχεία εισαγωγής ασθενών με νεφρική ανεπάρκεια, των οποίων οι κωδικοί ICD-10 για οποιαδήποτε από τις τρεις πρώτες διαγνώσεις ήταν J44.1 (οξεία έξαρση διαφόρων μορφών χρόνιας αποφρακτικής πνευμονοπάθειας) ή J44.90 (διάφορες μορφές χρόνιας αποφρακτικής πνευμονοπάθειας). Σύμφωνα με (Luo, 2019) χρησιμοποιήθηκαν τα δεδομένα για να βρεθεί ποια είναι η τιμή του υψηλού και χαμηλού ρίσκου των ασθενών. Το αθροιστικό κόστος για τους ασθενείς με νεφρική ανεπάρκεια το έτος 2011-2012 ήταν 7 εκατομμύρια, με τα κορυφαία 5, 10 και 20 εκατοστημόρια να αντιπροσωπεύουν το 35,2 τοις εκατό, το 51,8 τοις εκατό, και το 72,1 τοις εκατό του συνολικού κόστους, αντίστοιχα. Χρησιμοποιήθηκε η αρχή Pareto 80/20 (επίσης γνωστή ως ο κανόνας 80/20) που προτάθηκε από τον Ιταλό καθηγητή πολιτικής οικονομίας του 19ου αιώνα Vilfredo Pareto. (Sanders, 1987) Σύμφωνα με αυτή την μέθοδο υπολογίζεται ότι το κορυφαίο 20 τοις εκατό του πληθυσμού κάθε χώρας αντιπροσωπεύει περίπου το 80 τοις εκατό του συνολικού εισοδήματός της. Σε αυτή τη μελέτη, παρατηρήθηκε ότι το 20% των ασθενών με νεφρική ανεπάρκεια κατανάλωσε σχεδόν το 80% των ιατρικών πόρων. Σύμφωνα με τα αποτελέσματα, όπως αναφέρεται και (Luo, 2019), ο αλγόριθμος XGboost είχε την καλύτερη απόδοση.

Χρησιμοποιώντας ιατρικά δεδομένα ασθενών και συγκεκριμένα μετρήσεις όπως αρτηριακή πίεση, έμετος, GFR, κρεατινίνη, διαβήτης, καρδιαγγειακός παράγοντας, αναιμία, ουρία, πρωτεΐνη ούρων, ηλικία και φύλο γίνεται η πρόβλεψη αν ο ασθενής πάσχει από χρόνια νεφρική ανεπάρκεια. (Khan, 2010) Η επιλογή των κατάλληλων μεταβλητών έγινε με την χρήση chi-squared test. Το τεστ chi-square είναι μια δοκιμή στατιστικής υπόθεσης στην οποία η κατανομή δειγματοληψίας της στατιστικής δοκιμής είναι μια κατανομή chi-square όταν η μηδενική υπόθεση είναι αληθινή ή οποιαδήποτε στην οποία αυτό είναι ασυμπτωτικά αληθές, που σημαίνει ότι η κατανομή δειγματοληψίας (εάν η

μηδενική υπόθεση είναι αληθινή) μπορεί να γίνει με προσέγγιση της κατανομής chi-square κάνοντας το μέγεθος του δείγματος αρκετά μεγάλο. Τα αποτελέσματα των δοκιμών έδειξαν ότι ο αλγόριθμος Naïve Bayes είχε την καλύτερη ακρίβεια στην πρόβλεψη.

2.3. Επιλογή Χαρακτηριστικών

Η χρήση Feature Selection τεχνικών είναι μια σημαντική διαδικασία που βοηθά στην καλύτερη απόδοση των αλγορίθμων Μηχανικής Μάθησης. Χρησιμοποιώντας τεχνικές Feature Selection επιλέγονται οι καταλληλότερες μεταβλητές ώστε να βελτιωθεί η απόδοση των αλγορίθμων σύμφωνα πάντα με τα υπάρχοντα δεδομένα. Οι μέθοδοι Feature Selection μπορούν να κατηγοριοποιηθούν με πολλούς τρόπους. Η πιο συνηθισμένη είναι η κατηγοριοποίηση σε Filters, Wrappers και Embedded. ενσωματωμένες και υβριδικές μεθόδους. Η προαναφερθείσα ταξινόμηση προϋποθέτει ανεξαρτησία χαρακτηριστικών ή σχεδόν ανεξαρτησία.

Το πρώτο σύνολο μεθόδων Feature Selection είναι οι μέθοδοι Filter, οι οποίοι επιλέγουν χαρακτηριστικά βάσει ενός μέτρου απόδοσης ανεξάρτητα από τον αλγόριθμο μοντελοποίησης δεδομένων που χρησιμοποιείται. Μόνο αφού βρεθούν τα καλύτερα χαρακτηριστικά, οι αλγόριθμοι μοντελοποίησης μπορούν να τις χρησιμοποιήσουν. Οι μέθοδοι Filter μπορούν να ταξινομήσουν μεμονωμένα χαρακτηριστικά ή να αξιολογήσουν ολόκληρα υποσύνολα χαρακτηριστικών. Τα μέτρα για την επιλογή των καλύτερων χαρακτηριστικών μέσω των μεθόδων Filter είναι: η πληροφορία, η απόσταση, η συνέπεια, η ομοιότητα και στατιστικά μέτρα. Οι μέθοδοι Filter ταξινομούνται σε δύο μεγάλες κατηγορίες. Η πρώτη κατηγορία είναι τα Univariate Filters, τα οποία αξιολογούν (και συνήθως ταξινομούν) ένα μόνο χαρακτηριστικό. Η δεύτερη κατηγορία είναι τα Multivariate Filters, τα οποία αξιολογούν ένα ολόκληρο υποσύνολο χαρακτηριστικών. (Jović, 2015)

Το δεύτερο σύνολο μεθόδων Feature Selection είναι οι μέθοδοι Wrappers. Οι μέθοδοι Wrappers θεωρούν υποσύνολα χαρακτηριστικών από την ποιότητα της απόδοσης σε έναν αλγόριθμο μοντελοποίησης. Έτσι, για εργασίες ταξινόμησης, μια μέθοδος Wrapper θα αξιολογεί τα υποσύνολα με βάση την απόδοση ενός ταξινομητή, ενώ για το clustering, μια μέθοδος Wrapper θα αξιολογεί τα υποσύνολα με βάση την απόδοση ενός αλγορίθμου clustering. Η αξιολόγηση επαναλαμβάνεται για κάθε υποσύνολο και η δημιουργία υποομάδων εξαρτάται από τη στρατηγική αναζήτησης. (Jović, 2015)

Το τρίτο σύνολο μεθόδων Feature Selection είναι οι μέθοδοι Embedded. Οι μέθοδοι Embedded εκτελούν επιλογή χαρακτηριστικών κατά την εκτέλεση του αλγορίθμου μοντελοποίησης. Αυτές οι μέθοδοι ενσωματώνονται έτσι στον αλγόριθμο είτε ως κανονική είτε ως εκτεταμένη λειτουργικότητά αυτού. Ορισμένες μέθοδοι Embedded εκτελούν στάθμιση χαρακτηριστικών με βάση μοντέλα κανονικοποίησης με αντικειμενικές λειτουργίες που ελαχιστοποιούν τα λάθη τοποθέτησης και στο μέσο χρονικό διάστημα αναγκάζουν τους συντελεστές χαρακτηριστικών να είναι μικροί ή να είναι ακριβείς μηδέν. Τέτοιες μέθοδοι χρησιμοποιούν αλγορίθμους Μηχανικής Μάθησης όπως για παράδειγμα ο αλγόριθμος Lasso. (Jović, 2015)

Σύμφωνα με (Baranidharan, 2019) ο αλγόριθμος Extra Tree Classifier είναι μια αξιόπιστη λύση όσον αφορά το Feature Selection. Χρησιμοποιώντας ιατρικά δεδομένα προσπαθεί να προβλέψει την καρδιακή προσβολή. Από τα αποτελέσματα που προέκυψαν παρατηρήθηκε ότι χρησιμοποιώντας Extra Tree Classifier σαν τεχνική Feature Selection και χρησιμοποιώντας μόνο τις καταλληλότερες

μεταβλητές που προέκυψαν από την διαδικασία αυτή, οι αλγόριθμοι είχαν καλύτερη απόδοση από ότι χρησιμοποιώντας όλες τις διαθέσιμες μεταβλητές. Η χρήση Feature Selection αύξησε την απόδοση των αλγορίθμων κατά 10%-15%. Σύμφωνα με (Kaushik, 2019) μία ακόμη τεχνική για Feature Selection σε ιατρικά δεδομένα είναι η τεχνική Lasso. Στην συγκεκριμένη εργασία (Kaushik, 2019) έγινε σύγκριση τεχνικών Feature Selection A-priori method, information gain, CFS, LASSO, ridge regression, and PCA. Τα δεδομένα που χρησιμοποιήθηκαν ήταν δεδομένα ασφάλισης των ασθενών. Σύμφωνα με τα αποτελέσματα οι αλγόριθμοι είχαν την καλύτερη απόδοση χρησιμοποιώντας τις μεταβλητές που προέκυψαν χρησιμοποιώντας την τεχνική LASSO.

2.4. Ανάλυση Επιβίωσης

Η ανάλυση επιβίωσης αναφέρεται στην ανάλυση δεδομένων που αφορούν στο χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν. Το συμβάν αυτό συνήθως είναι ο θάνατος. Μπορεί όμως να είναι και η ανάρρωση ή η εμφάνιση κάποιου συμπτώματος. Ο χρόνος όμως μέχρι το συμβάν είναι συνήθως λογοκριμένος με την έννοια ότι δεν είναι γνωστή η πραγματική του τιμή για κάποιους ασθενείς. Το μόνο που είναι γνωστό για τους λογοκριμένους χρόνους είναι ότι οι χρόνοι αυτοί είναι μεγαλύτεροι από την τιμή που έχει καταγραφεί. Στην Ανάλυση Επιβίωσης χρησιμοποιούνται ειδικές τεχνικές όπως η τεχνική Kaplan Meier. Ένας από τους λόγους για τους οποίους η ανάλυση επιβίωσης απαιτεί «ειδικές» τεχνικές είναι η δυνατότητα να μην παρατηρείται το γεγονός για ορισμένα άτομα. Άτομα που παίρνουν μέρος σε μια μελέτη μπορεί να εγκαταλείψουν τη μελέτη, ή μετά από κάποια άλλη περίπτωση. Μια άλλη πιθανότητα είναι να υπάρχει εκεί ένα χρονικό σημείο στο οποίο η μελέτη τελειώνει και, επομένως, σε κάποιο άτομο δεν έχει ακόμη συμβεί το συμβάν που μελετάται. Αυτές οι ελλιπείς παρατηρήσεις δεν μπορούν να αγνοηθούν, αλλά πρέπει να αντιμετωπιστούν διαφορετικά. Αυτό ονομάζεται λογοκρισία. Ένα άλλο χαρακτηριστικό των δεδομένων επιβίωσης είναι ότι οι δυσκολίες είναι συχνά ασύμμετρες και έτσι απλές τεχνικές που βασίζονται στην κανονική κατανομή δεν μπορούν να χρησιμοποιηθούν άμεσα. Οι στόχοι της ανάλυσης επιβίωσης περιλαμβάνουν την ανάλυση των προτύπων των χρόνων των γεγονότων, τη σύγκριση των κατανομών των χρόνων επιβίωσης σε διαφορετικές ομάδες ατόμων και την εξέταση του κατά πόσο και πόσοι παράγοντες επηρεάζουν τον κίνδυνο ενός συμβάντος.

Όσον αφορά την λογοκρισία(Censoring) υπάρχουν δυο είδη λογοκρισίας: right censoring, left censoring. Right censoring εμφανίζεται όταν ένα άτομο παρακολουθείται από μια χρονική στιγμή μέχρι κάποια άλλη στιγμή αργότερα και δεν έχει παρουσιάσει το συμβάν που μελετάται, έτσι ώστε το μόνο που είναι γνωστό είναι ότι το συμβάν δεν έχει σημειωθεί μέσα στο χρονοδιάγραμμα λογοκρισίας. Αυτό μπορεί να συμβεί, για παράδειγμα, εάν ένα άτομο εγκαταλείψει μια μελέτη πριν συμβεί το παρατηρούμενο γεγονός. Συνήθως οι μελέτες τερματίζονται σε κάποιο καθορισμένο χρόνο και στο τέλος της μελέτης σε ορισμένα άτομα δεν έχει παρατηρηθεί το συμβάν. Το δεύτερο είδος είναι το left censoring. Left censoring είναι το είδος στο οποίο ένα άτομο είναι γνωστό ότι βιώσε το συμβάν πριν από μια συγκεκριμένη ώρα, αλλά αυτό θα μπορούσε να είναι οποιαδήποτε στιγμή πριν από το χρόνο λογοκρισίας. Είναι επίσης δυνατό να υπάρχει λογοκρισία διαστήματος όπου είναι γνωστό ότι ένα άτομο βίωσε μόνο το συμβάν μεταξύ δύο χρονικών στιγμών, αλλά η ακριβής ώρα του συμβάντος δεν είναι γνωστή. (Kartsonaki, 2016)

Ένα άλλο σενάριο είναι το Truncation. Το Truncation είναι κάτι που συμβαίνει κατόπιν σχεδιασμού. Το Left Truncation είναι ο συνηθέστερος τύπος Truncation, όπου τα άτομα εισέρχονται στη μελέτη

αφού έχουν βιώσει ένα συμβάν (το οποίο δεν είναι το ίδιο με το συμβάν που μελετάται). (Kartsonaki, 2016)

Η Ανάλυση Επιβίωσης συναντάται συχνά στην βιβλιογραφία σε προβλήματα ασθενών στο τελικό στάδιο νεφρικής ανεπάρκειας. Σύμφωνα με (Urrutia, 2015) οι δύο συχνά χρησιμοποιούμενες τεχνικές ανάλυσης επιβίωσης είναι η τεχνική Kaplan-Meier (Goel, 2010) και η τεχνική Weibull Distribution. (Carroll, 2003). Οι δυο αυτές τεχνικές συγκρίνονται σύμφωνα με τα αποτελέσματα που προκύπτουν χρησιμοποιούμενες σε δεδομένα. Χρησιμοποιώντας ως μετρητή το p-value έγινε σύγκριση των δυο τεχνικών. Σύμφωνα με (Urrutia, 2015) οι γυναίκες ασθενείς έχουν μεγαλύτερο κίνδυνο θανάτου σε σύγκριση με τους άνδρες , και οι ασθενείς με ηλικίες άνω των 52 ετών έχουν μεγαλύτερο κίνδυνο θανάτου σε σύγκριση με ασθενείς με ηλικίες κάτω των 52 ετών. Ακόμη το ποσοστό των ασθενών που η αιτία θανάτου ήταν καρδιακή προσβολή ήταν μεγαλύτερο από αυτό των ασθενών που απεβίωσαν από πνευμονική συμφόρηση. Επιπρόσθετα άτομα μεγαλύτερα από 52 ετών είχαν μεγαλύτερο ποσοστό θνησιμότητας. Οι δυο τεχνικές δεν είχαν σημαντικές διαφορές στα αποτελέσματα με γνώμονα την τιμή του p-value.

2.5. Αιτιολογική Ανάλυση

Η χρήση μεθόδων Αιτιολογικής Ανάλυσης χρησιμοποιείται ευρέως στον τομέα της ιατρικής. Χρησιμοποιώντας μεθόδους Αιτιολογικής Ανάλυσης εξετάζεται η επίδραση μιας αιτίας στο αποτέλεσμα. Ειδικότερα στον τομέα της Επιδημιολογίας χρησιμοποιώντας μεθόδους Αιτιολογικής Ανάλυσης ελέγχεται τι επίδραση θα έχει η λήψη μιας θεραπείας σε μια ασθένεια. Επίσης μέσω των τεχνικών για Αιτιολογική Ανάλυση μπορεί να ελέγχει ποια θεραπεία είναι πιο καλή στην καταπολέμηση μιας ασθένειας. Η βιβλιογραφία παραθέτει αρκετές τεχνικές Αιτιολογικής Ανάλυσης όπως οι τεχνικές Matching (Stuart, 2010) (Matschinger, 2020) και Weighting (Tan, 2006). Ο (Matschinger, 2020) συγκρίνει την λειτουργία των παραπάνω αλγορίθμων και χρησιμοποιώντας Propensity Score (Chatton, 2020) προσπαθεί να βελτιώσει τα αποτελέσματα. Χρησιμοποιώντας δεδομένα παρατηρεί ασθενείς με διαβήτη συμφορητική καρδιακή ανεπάρκεια, αρτηριομυκητίαση, στεφανιαία νόσο ή υπέρταση ενός γερμανικού ταμείου ασθενείας και συγκρίνει τα αποτελέσματα των ασθενών που έλαβαν την συνηθισμένη θεραπεία με τους ασθενείς που δεν την έλαβαν αλλά τους προσφέρθηκε ατομική τηλεφωνική καθοδήγηση για τη βελτίωση της συμπεριφοράς στην υγεία και την επιβράδυνση της εξέλιξης της νόσου. Στα μοντέλα χρησιμοποιήθηκαν σαν επιπλέον μεταβλητές επιρροής το φύλο, η ηλικία, η επαγγελματική κατάσταση, το πρόγραμμα διαχείρισης ασθενειών, το καθεστώς ασφάλισης υγείας, το επίπεδο φροντίδας, η ομοσπονδιακή πολιτεία διαμονής, οι βασικές τιμές των υπηρεσιών και οι δαπάνες υγειονομικής περίθαλψης, καθώς και τα 31 συστατικά στοιχεία του δείκτη συν νοσηρότητας Elix-hauser (Quan, 2005). Σύμφωνα με την (Hill, 2011) υπάρχει δυνατότητα χειρισμού και συνεχών μεταβλητών. Χρησιμοποιώντας δεδομένα για το χρόνο παραμονής παιδιών ηλικίας 3 ετών σε κέντρο παρακολούθησης και αποτροπής ασθενειών (Centers for Disease Control and Prevention) μελετήθηκε κατά πόσο αυξήθηκε ο δείκτης IQ των παιδιών. Χρησιμοποιώντας Bayesian Additive Regression Trees (Chipman, Bayesian ensemble learning. In Advances in neural information processing systems, 2007) και σύμφωνα με τα αποτελέσματα μόνο τα παιδιά που παρέμειναν παραπάνω από 200 μέρες μέσα στο κέντρο έδειξαν να επηρεάζονται θετικά. Ακόμη προκύπτει ότι τα παιδιά που είχαν από την αρχή υψηλό δείκτη IQ είχαν τις λιγότερες μέρες παραμονής στο κέντρο. Αυτό συνέβη διότι οι γονείς πιστεύουν ότι είναι σε αρκετά καλή κατάσταση ώστε να μην χρειάζονται συνεχή παρακολούθηση. Ο αλγόριθμος Bayesian

Additive Regression Trees μπορεί και αποτυπώνει μια μη γραμμική σχέση μεταξύ των ημερών συμμετοχής των παιδιών στο κέντρο και της αύξησης του δείκτη IQ. Ακόμη χρησιμοποιήθηκε και το μοντέλο γραμμικής παλινδρόμησης. Σύμφωνα με τα αποτελέσματα το μοντέλο Γραμμικής Παλινδρόμησης δεν μπορεί να αποτυπώσει την μη-γραμμικότητα μεταξύ των ημερών συμμετοχής των παιδιών στο κέντρο και της αύξησης του δείκτη IQ χρησιμοποιώντας τετραγωνικούς όρους.

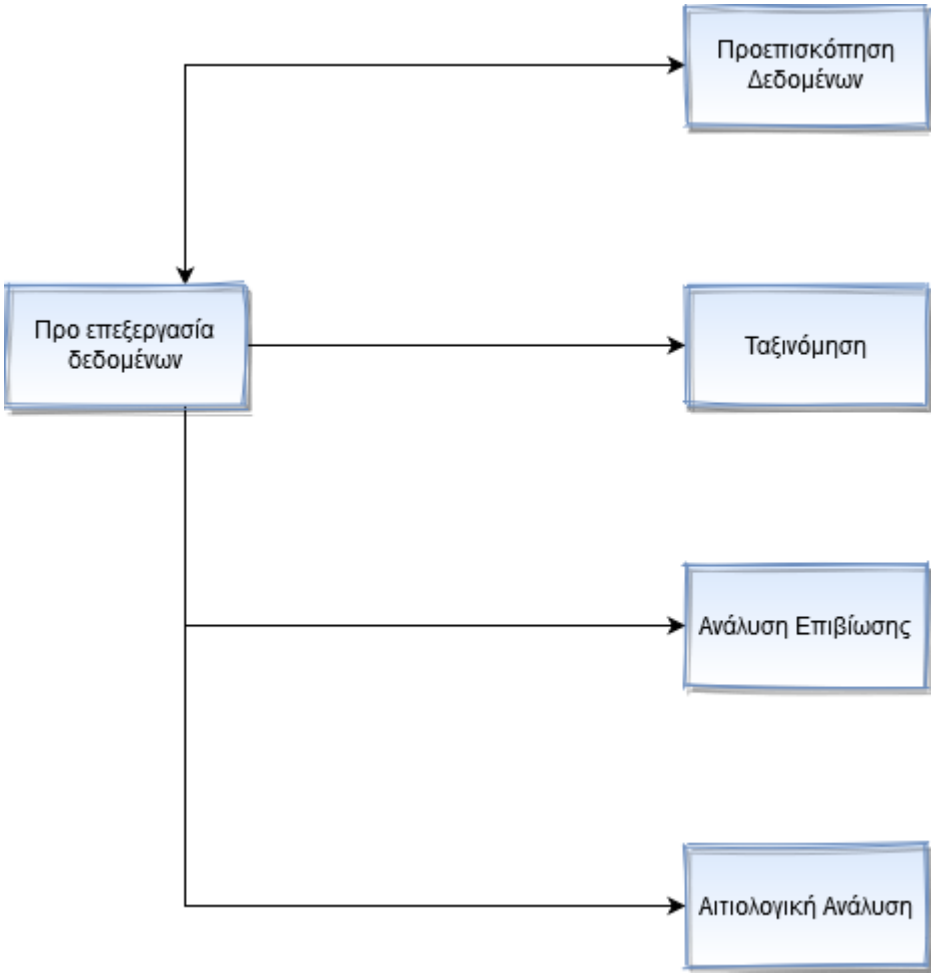
Επιπρόσθετα στην βιβλιογραφία παρατίθενται και ο χειρισμός κατηγορικών μεταβλητών όσον αφορά την Αιτιολογική Ανάλυση. Σύμφωνα με (Stuart, 2010) υπάρχουν τεχνικές που χρησιμοποιούν κατηγορικά δεδομένα σαν παράγοντα επιρροής. Η χρήση μεθόδων Matching μπορούν να διαχειριστούν τέτοιου είδους δεδομένα. Μετατρέποντας τις κατηγορηματικές μεταβλητές σε μια σειρά δυαδικών δεικτών, και χρησιμοποιώντας μέτρα απόστασης υπολογίζεται η επιρροή αυτών των μεταβλητών. Ακόμη (Cai, 2018) αναπτύχθηκε μια διαδικασία η οποία περιλαμβάνει δύο στάδια: Το πρώτο στάδιο είναι χαρτογραφεί την αιτία σε μια κρυφή μεταβλητή και το δεύτερο στάδιο δημιουργεί το αποτέλεσμα από την κρυφή αναπαράσταση. Σύμφωνα με (Cai, 2018) αναπτύχθηκε ένα low-cardinality κρυφό μοντέλο συμπαγούς αναπαράστασης και χρησιμοποιώντας συνθετικά δεδομένα έγινε ο έλεγχος για την ποιότητα των αποτελεσμάτων. Ακόμη χρησιμοποιώντας κατηγορικά δεδομένα που περιείχαν καταγραφές 108 γεφυρών. Υπήρχαν 4 ζεύγη αιτίας-αποτελέσματος, τα οποία είναι: 1) Στεγασμένο (Χειροτεχνία, Αναδυόμενα, Ώριμα, Μοντέρνα) → Span (Long, Medium, Short), 2) Material (Steel, Iron, Wood) → Span (Long, Medium, Short); 3) Υλικό (χάλυβας, σίδερο, ξύλο) → λωρίδες (1, 2, 4, 6). 4) Σκοπός (Walk, Aqueduct, RR, Highway) → τύπος (Ξύλο, Suspen, Simple-T, Arch, Cantilev, CONT-T). Από τα αποτελέσματα που προέκυψαν το μοντέλο αναγνώρισε και τα τέσσερα ζεύγη αιτίας αποτελέσματος. Επίσης από τα αποτελέσματα φάνηκε ότι το προτεινόμενο μοντέλο είχε καλά αποτελέσματα ακόμη και για μικρό σύνολο δεδομένων. Επιπρόσθετα χρησιμοποιήθηκε ακόμη ένα σύνολο δεδομένων που περιείχε 4177 δείγματα και κάθε δείγμα έχει 4 διαφορετικές ιδιότητες. Περιείχε τρία ζεύγη αιτίων-αποτελεσμάτων, Σεξ → {Μήκος, Διάμετρος, Ύψος}. Το σεξ ιδιοκτησίας έχει τρεις αξίες, αρσενικό, θηλυκό και βρέφος. Το μήκος, η διάμετρος και το ύψος μετρούνται σε mm και αντιμετωπίζονται ως διακριτές τιμές. Σε αυτό το σύνολο δεδομένων, το μοντέλο καθόρισε με επιτυχία την αιτιώδη κατεύθυνση και για τα τρία ζεύγη.

Συμπερασματικά υπάρχουν ποικίλες τεχνικές για την επίλυση των προβλημάτων που τέθηκαν στην παρούσα εργασία. Αρχικά με την χρήση αλγορίθμων Μηχανικής Μάθησης θα γίνει προσπάθεια να προβλεφθεί ο χρόνος ζωής των ασθενών σε μεταμόσχευση. Θα χρησιμοποιηθούν τεχνικές Feature Selection καθώς και τεχνικές συσχέτισης ώστε να βρεθούν οι καταλληλότερες μεταβλητές ώστε να επιτευχθεί η καλύτερη απόδοση των αλγορίθμων σύμφωνα πάντα με τα δεδομένα που εξετάζονται. Συνδυάζοντας τα αποτελέσματα των τεχνικών Feature Selection με των τεχνικών συσχέτισης θα γίνει μια πιο ορθή επιλογή των καταλληλότερων μεταβλητών ώστε οι αλγόριθμοι Μηχανικής Μάθησης να έχουν καλύτερη απόδοση. Ακόμη δοκιμάζοντας διάφορους αλγορίθμους Μηχανικής Μάθησης και αλλάζοντας τις παραμέτρους του καθένα θα γίνει σύγκριση όσον αφορά τη απόδοση τους στην πρόβλεψη του χρόνου ζωής των ασθενών, στην πρόβλεψη της αιτίας θανάτου των ασθενών καθώς και της πρόβλεψης του κόστους της θεραπείας. Χρησιμοποιώντας Ανάλυση Επιβίωσης ερευνάται το ποσοστό επιβίωσης των ασθενών σε μεταμόσχευση και αιμοκάθαρση σε περίοδο 12 χρόνων από την τελευταία έξοδο από το νοσοκομείο. Χρησιμοποιείται η τεχνική Kaplan-Meier προσαρμόζοντάς την στα δεδομένα. Τέλος χρησιμοποιούνται τεχνικές Αιτιολογικής Ανάλυσης για να φανεί αν το φάρμακο 'Trombyl' είχε επιπτώσεις στην διάρκεια ζωής των ασθενών. Χρησιμοποιούνται εκτιμητές Matching και Weighting καθώς και ο εκτιμητής Ordinary Least Square.Παρατηρείται αν οι παραπάνω τεχνικές

έχουν μεγάλη απόκλιση η μια από την άλλη. Επιπλέον χρησιμοποιώντας Propensity Score καθώς και τεχνικές Trimming γίνεται προσπάθεια για βελτίωση των παραπάνω τεχνικών.

Η παρούσα εργασία χρησιμοποιεί επιστημονικές μεθόδους, διαδικασίες, αλγορίθμους προκειμένου να εξάγει γνώση και χρήσιμες πληροφορίες από ιατρικά δεδομένα ώστε να κατανοηθεί καλύτερα ο τρόπος επιρροής της Νεφρικής Ανεπάρκειας στην υγεία του ασθενή. Επιπλέον προκύπτει γνώση που βοηθά στην καλύτερη αντιμετώπιση της ασθένειας αλλά και στην πιο στοχευμένη θεραπεία γνωρίζοντας την πιθανή αιτία θανάτου. Η γνώση του πιθανού κόστους της θεραπείας θα ωφελήσει τους ασθενείς στην καλύτερη διαχείριση των οικονομικών τους ώστε να ανταπεξέλθουν στην θεραπεία.

3. Μεθοδολογία



Εικόνα 1 Ροή της μεθοδολογίας

Η μεθοδολογία που ακολουθήθηκε στην παρούσα εργασία είναι η εξής:

- Προ επεξεργασία δειγμάτων σε δεδομένα πολλών διαστάσεων

  Για κάθε πρόβλημα ταξινόμησης χρησιμοποιήθηκαν διαφορετικές μεταβλητές. Οι μεταβλητές αυτές προέκυψαν χρησιμοποιώντας τεχνικές Feature Selection καθώς και κανόνες συσχέτισης ώστε να επιλεχθούν οι πιο συσχετισμένες μεταβλητές με τις μεταβλητές για τις οποίες έγινε η πρόβλεψη. Ακόμη για της ανάγκες της ταξινόμησης δημιουργήθηκε μια νέα μεταβλητή. Στο πρόβλημα πρόβλεψης τις διάρκειας ζωής των ασθενών σε μεταμόσχευση δημιουργήθηκε μια νέα μεταβλητή “Transplant_Days”, η οποία περιέχει τα έτη ζωής των ασθενών από τη μέρα της μεταμόσχευσης μέχρι το θάνατό τους. Τα έτη ζωής που προκύπτουν είναι διακριτές τιμές και κυμαίνονται από 5 ως 40. Έτσι δημιουργήθηκαν 5 κλάσεις οι οποίες περιέχουν περίπου ίδιο αριθμό δειγμάτων. Οι κλάσεις είναι [4-8] έτη, (8-15] έτη, (15-16] έτη, (16-21] έτη, (21-40] έτη. Πριν χρησιμοποιηθούν τα δεδομένα έγινε μια επεξεργασία αυτών. Έγιναν ενέργειες για να έρθουν τα δεδομένα σε κατάλληλη μορφή για την περαιτέρω χρήση τους. Αρχικά παρατηρήθηκε ότι κάποιες εγγραφές στα σύνολα

δεδομένων ‘demo_large’ και ‘lab_values_large’ και συγκεκριμένα στη μεταβλητή ‘lan’ υπήρχαν εγγραφές που παρόλο που ανήκαν στην ίδια κατηγορία είχαν λίγο διαφορετική ονομασία, όπως παραδείγματος χάριν υπήρχαν τιμές 1.0 και 1, οι οποίες είναι ίδιες, αλλά κατά την επεξεργασία φαίνονταν σαν δύο διαφορετικές, το οποίο δυσκόλευε στην διαχείρισή τους, οπότε και τροποποιήθηκαν ώστε να είναι ίδιες. Έπειτα έγινε συγχώνευση των δεδομένων βασισμένη στον μοναδικό κωδικό του κάθε ασθενή “lopnr”. Κατά την συγχώνευση προέκυψαν διπλότυπες τιμές, οι οποίες και διαγράφηκαν. Ακόμη για την χρήση της κατάστασης των ασθενών (αιμοκάθαρση, μεταμόσχευση) λόγω χαμένων τιμών έγινε συμπλήρωση των χαμένων τιμών τυχαία με βάση τις τιμές που υπήρχαν και το πλήθος των δειγμάτων σε κάθε κατάσταση.

- Προεπισκόπηση Δεδομένων.
  Στο συγκεκριμένο στάδιο γίνεται μια προεπισκόπηση των δεδομένων ώστε να υπάρξει μια αρχική εικόνα αυτών. Υπολογίστηκε η μέση τιμή, ποσοστό καθώς και αθροιστικά αποτελέσματα τα οποία αναπαρίστανται με την χρήση ραβδογραμμάτων. Με την χρήση των μεθόδων αυτών προκύπτουν συμπεράσματα όπως πιο φάρμακο χρησιμοποιείται περισσότερο ή τι ποσοστό ανδρών και γυναικών πάσχουν από την ασθένεια κ.α. τα οποία χρησιμοποιήθηκαν για να υπάρξει μια καλύτερη εικόνα των δεδομένων ώστε να χρησιμοποιηθούν πιο αποτελεσματικά στα προβλήματα Ταξινόμησης, Ανάλυσης Επιβίωσης και Αιτιολογικής Ανάλυσης και τα οποία παρατίθενται σε επόμενο κεφάλαιο αναλυτικά. Η εύρεση του πιο συχνά συνταγογραφούμενο φαρμάκου βοήθησε στην χρήση του ώστε να γίνει Αιτιολογική Ανάλυση βάσει αυτού. Βρίσκοντας το μέσο κόστος θεραπείας ανά ασθενή έγινε σύγκριση με άλλες χρόνιες ασθένειες για να διαπιστωθεί αν η νεφρική ανεπάρκεια είναι δαπανηρή.

- Ταξινόμηση
  Χρησιμοποιήθηκαν οι αλγόριθμοι Random Forest, Decision Tree, k-Nearest Neighbors, XGboost, Naïve Bayes. Ο αλγόριθμοι αυτοί έχουν χρησιμοποιηθεί με επιτυχία σε παρόμοιες εργασίες όπως αναφέρεται σε προηγούμενο κεφάλαιο και τα αποτελέσματά τους είναι ακριβή και γρήγορα σε σχέση με άλλους αλγορίθμους.

  Ο K- Nearest Neighbours (KNN) είναι ένας απλός αλγόριθμος μηχανικής μάθησης που αποθηκεύει όλες τις διαθέσιμες περιπτώσεις και ταξινομεί νέες περιπτώσεις με βάση ένα μέτρο ομοιότητας (π.χ. συνάρτηση απόστασης). Κατά την εφαρμογή του KNN, το πρώτο βήμα είναι η μετατροπή σημείων δεδομένων σε διανύσματα χαρακτηριστικών ή στην μαθηματική τους τιμή. Ο αλγόριθμος στη συνέχεια συνεχίζει με την εύρεση της απόστασης μεταξύ των μαθηματικών τιμών αυτών των σημείων. Ο πιο συνηθισμένος τρόπος να βρεθεί αυτή η απόσταση είναι η ευκλείδεια απόσταση η οποία δίνεται από την εξίσωση (1) παρακάτω.

$$d(q,p) = d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \quad (1)$$

  Ο KNN χρησιμοποιεί την εξίσωση (1) για τον υπολογισμό της απόστασης μεταξύ κάθε νέας τιμής που εισάγεται στο σύνολό των δεδομένων με τα δεδομένα που έχουν ήδη κατηγοριοποιηθεί. (Zhang Z. , 2016) Η νέα τιμή προστίθεται στην κλάση στην οποία είναι η τιμή με την οποία είναι πιο κοντά. (Guo, 2003)

Ο αλγόριθμος μηχανικής μάθησης XGBoost είναι μια εφαρμογή του αλγορίθμου Gradient Boosted Decision Trees(GBM) ο οποίος σχεδιάστηκε για μεγαλύτερη ταχύτητα και απόδοση. Για την κατανόηση του συγκεκριμένου αλγορίθμου πρέπει αρχικά να γίνει μια περιγραφή του τρόπου λειτουργίας του αλγορίθμου Gradient Boosted Decision Tree.

Το πρώτο βήμα στο αλγόριθμο Gradient Boosted Decision Tree(GBM) είναι να υπολογιστεί η μέση τιμή της εξαρτημένης μεταβλητής. Χρησιμοποιώντας τη μέση τιμή της εξαρτημένης μεταβλητής σαν αρχικό φύλλο γίνεται προσέγγιση της σωστής λύσης στα στάδια της διαδικασίας. Στο δεύτερο βήμα γίνεται υπολογισμός των υπολοίπων(residuals). Για κάθε δείγμα, αν θεωρηθούν σαν Y τα υπόλοιπα σαν R η πραγματική τιμή και σαν P η προβλεπόμενη τιμή τότε ο υπολογισμός των υπολοίπων γίνεται σύμφωνα με την εξίσωση 2.

$$Y = R - P \tag{2}$$

Έπειτα γίνεται κατασκευή του δέντρου απόφασης. Η κατασκευή του δέντρου απόφασης γίνεται για την πρόβλεψη των υπολοίπων. Δηλαδή κάθε φύλλο στο δέντρο απόφασης θα περιέχει την πρόβλεψη ως προς την τιμή των υπολοίπων και όχι της εξαρτημένης τιμής. Στην περίπτωση που υπάρχουν περισσότερα υπόλοιπα από ότι φύλλα, κάποια υπόλοιπα θα καταλήξουν μέσα στο ίδιο φύλλο. Όταν συμβαίνει αυτό, υπολογίζεται ο μέσος όρος των υπολοίπων και τοποθετείται μέσα στο φύλλο. Στο επόμενο βήμα προβλέπεται η εξαρτημένη μεταβλητή(κλάση) χρησιμοποιώντας όλα τα δημιουργημένα δέντρα. Κάθε δείγμα περνά μέσα από τους κόμβους απόφασης του νέου δέντρου. Το υπόλοιπο στο εν λόγω φύλλο χρησιμοποιείται για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής. Για να αποφευχθεί το over fitting, εισάγεται ένας ρυθμός εκμάθησης. Όταν γίνει μια πρόβλεψη, κάθε υπόλοιπο πολλαπλασιάζεται με τον ρυθμό εκμάθησης. Αυτό ωθεί στην χρησιμοποίηση περισσότερων δέντρων απόφασης, το καθένα κάνοντας ένα μικρό βήμα προς την τελική λύση. Έπειτα γίνεται υπολογισμός νέων υπολοίπων. Υπολογίζεται ένα νέο σύνολο υπολοίπων αφαιρώντας τις πραγματικές τιμές της εξαρτημένης μεταβλητής από τις προβλέψεις που έγιναν στο προηγούμενο βήμα. Τα υπόλοιπα στη συνέχεια θα χρησιμοποιηθούν για τα φύλλα του επόμενου δέντρου απόφασης όπως περιγράφεται στο προηγούμενο βήμα. Έπειτα επαναλαμβάνονται τα παραπάνω βήματα από την αρχή. Τέλος χρησιμοποιούνται όλα τα δέντρα απόφασης για να γίνει μια τελική πρόβλεψη.

Ουσιαστικά ο αλγόριθμος XGBoost είναι η υλοποίηση των παραπάνω βημάτων. Αυτό που τον κάνει να διαφέρει από τους υπόλοιπους αλγορίθμους Gradient Boosted Decision Tree(GBM) είναι το κλάδεμα δέντρων. Σε αντίθεση με το GBM, όπου το κλάδεμα των δέντρων σταματάει όταν εμφανίζεται αρνητική απώλεια, ο XGBoost αυξάνει το δέντρο μέχρι το max_depth και στη συνέχεια κλαδεύει προς τα πίσω μέχρι να βελτιωθεί η loss function δηλαδή μα μειωθεί κάτω από ένα όριο. Επιπλέον, ο XGBoost έχει σχεδιαστεί για να χειρίζεται εσωτερικά τις χαμένες τιμές. Οι χαμένες τιμές αντιμετωπίζονται με τέτοιο τρόπο ώστε εάν υπάρχει κάποιο μοτίβο, καταγράφεται από το μοντέλο. Τέλος το μεγαλύτερο πλεονέκτημα του XGBoost είναι η κανονικοποίηση. Ο αλγόριθμος GBM δεν χρησιμοποιεί κανονικοποίηση κατά τη διαδικασία πρόβλεψης.

Ο αλγόριθμος Random Forest είναι μια μέθοδος μηχανικής μάθησης που λειτουργεί με την κατασκευή πολλαπλών δέντρων απόφασης. Η τελική απόφαση γίνεται με βάση την πλειοψηφία των δένδρων και επιλέγεται από το Random Forest. Ο αλγόριθμος αυτός

ουσιαστικά δουλεύει όπως ο αλγόριθμος του δέντρου απόφασης, ο τρόπος λειτουργίας του οποίου δίνεται παραπάνω. Η καλή λειτουργία του αλγορίθμου προϋποθέτει την καλή δημιουργία των δέντρων απόφασης. Επιπρόσθετα, η χρήση του αλγορίθμου Random Forest αποτρέπει το over fitting και μπορεί να χειριστεί τις χαμένες τιμές σε ένα σύνολο δεδομένων. (Synced, 2017).

Υπάρχουν δύο στάδια στον αλγόριθμο Random Forest, το πρώτο είναι δημιουργία του τυχαίου δάσους, και το δεύτερο είναι η πρόβλεψη από τον αλγόριθμο που δημιουργήθηκε στο πρώτο στάδιο.

Στο πρώτο στάδιο ο αλγόριθμος δημιουργεί πολλά δέντρα απόφασης από τα δεδομένα εκπαίδευσης και έπειτα επιλέγει την καλύτερη πρόβλεψη ανάμεσα σε όλες τις προβλέψεις που έγιναν από τα επιμέρους δέντρα απόφασης. Η δημιουργία των επιμέρους δέντρων γίνεται με τα παρακάτω βήματα:

- Επιλέγονται τυχαία "k" χαρακτηριστικά από τα συνολικά "m" όπου k << m
- Μεταξύ των χαρακτηριστικών "k", υπολογίζεται ο κόμβος "d" χρησιμοποιώντας την καλύτερη διαίρεση των δεδομένων.
- Διαχωρισμός του κόμβου σε υποκόμβους χρησιμοποιώντας τον καλύτερο διαχωρισμό.
- Επανάληψη των παραπάνω βημάτων έως τον αριθμό "i" κόμβων.
- Γίνεται κατασκευή του δάσους επαναλαμβάνοντας όλα τα παραπάνω βήματα "n" φορές για την δημιουργία "n" δέντρων.

Στο επόμενο στάδιο, με τον αλγόριθμο Random Forest που δημιουργήθηκε, θα γίνει η πρόβλεψη. Σε ένα δέντρο αποφάσεων μια νέα τιμή πηγαίνει από πάνω προς τα κάτω μέχρι να ταξινομηθεί σε έναν κόμβο φύλλων. Στον αλγόριθμο Random Forest, κάθε νέο σημείο δεδομένων περνά από την ίδια διαδικασία, αλλά τώρα επισκέπτεται όλα τα διαφορετικά δέντρα στο σύνολο, τα οποία αναπτύχθηκαν χρησιμοποιώντας τυχαία δείγματα τόσο των δεδομένων εκπαίδευσης όσο και των χαρακτηριστικών. (Liu, 2012) Ανάλογα με την εργασία, οι λειτουργίες που χρησιμοποιούνται για τη συγκέντρωση θα διαφέρουν. Για προβλήματα ταξινόμησης, χρησιμοποιεί τη λειτουργία ή την πιο συχνή τάξη που προβλέπεται από τα μεμονωμένα δέντρα (επίσης γνωστή ως πλειοψηφία), ενώ για εργασίες παλινδρόμησης, χρησιμοποιεί τη μέση πρόβλεψη κάθε δέντρου.

Τα δέντρα απόφασης είναι ένας από τους πιο δημοφιλείς αλγορίθμους που χρησιμοποιούνται στη Μηχανική Μάθηση. Είναι μια μη παραμετρική εποπτευόμενη μέθοδος μηχανικής μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Τα δέντρα αποφάσεων μαθαίνουν από τα δεδομένα και προσπαθούν να προσεγγίσουν μια ημιτονοειδής καμπύλη με ένα σύνολο κανόνων if-then-else. Όσο πιο πολύ μεγαλώνει το δέντρο, τόσο πιο σύνθετοι είναι οι κανόνες απόφασης και πιο προσαρμοσμένο το μοντέλο.

Το δέντρο απόφασης δημιουργεί μοντέλα ταξινόμησης ή παλινδρόμησης με τη μορφή δομής δέντρου. Διαιρεί ένα σύνολο δεδομένων σε μικρότερα και μικρότερα υποσύνολα έως ότου όλες οι μεταβλητές στόχοι εμπίπτουν σε μία κατηγορία, ενώ παράλληλα αναπτύσσεται σταδιακά ένα συσχετισμένο δέντρο απόφασης. Το τελικό αποτέλεσμα είναι ένα δέντρο με κόμβους απόφασης και κόμβους φύλλων. Ένας κόμβος απόφασης έχει δύο ή περισσότερα

κλαδιά. Ο κόμβος φύλλων αντιπροσωπεύει μια ταξινόμηση ή μια απόφαση. Ο κόμβος απόφασης σε ένα δέντρο με το μεγαλύτερο κέρδος πληροφορίας ονομάζεται κόμβος ρίζας. Τα δέντρα αποφάσεων μπορούν να χειριστούν τόσο τα κατηγορικά όσο και τα αριθμητικά δεδομένα.

Υπάρχουν αρκετά βήματα για την κατασκευή ενός δέντρου απόφασης.

Αρχικά είναι το Splitting. Η διαδικασία διαίρεσης του συνόλου δεδομένων σε υποσύνολα. Οι διαχωρισμοί σχηματίζονται σε μια συγκεκριμένη μεταβλητή. Έπειτα είναι το pruning. Η συγκεκριμένη διαδικασία είναι η αποκοπή κάποιων κλαδιών του δέντρου. Το κλάδεμα (pruning) είναι η διαδικασία μείωσης του μεγέθους του δέντρου με την εναλλαγή ορισμένων κόμβων κλαδιών του δέντρου σε κόμβους φύλλων και την αφαίρεση των κόμβων φύλλων κάτω από το αρχικό κλαδί. Το κλάδεμα είναι χρήσιμο επειδή τα δέντρα ταξινόμησης μπορούν να προσαρμοστούν καλά στα δεδομένα εκπαίδευσης, αλλά μπορεί να κάνουν κακή δουλειά στην ταξινόμηση νέων τιμών. Ένα απλό δέντρο απόφασης συχνά αποφεύγει το over-fitting. Έπειτα υπάρχει η επιλογή δέντρου. Η διαδικασία εύρεσης του μικρότερου δέντρου που ταιριάζει στα δεδομένα. Συνήθως αυτό είναι το δέντρο που δίνει το χαμηλότερο cross-validated error.

Υπάρχουν κάποιοι βασικοί συντελεστές σύμφωνα με τους οποίους ένα δέντρο απόφασης δημιουργείται.

Αρχικά είναι η εντροπία. Ένα δέντρο απόφασης είναι "χτισμένο" από την αρχή σε έναν κόμβο ρίζας (root node) και περιλαμβάνει τη διαίρεση των δεδομένων σε υποσύνολα που περιέχουν περιπτώσεις με παρόμοιες τιμές (ομοιογενείς). Συνήθως χρησιμοποιείται η εντροπία για τον υπολογισμό της ομοιογένειας ενός δείγματος. Αν το δείγμα είναι εντελώς ομοιογενές, η εντροπία είναι μηδέν.

Η εξίσωση 3 υπολογίζει την εντροπία. Όπου το p_i είναι η συχνότερη πιθανότητα ενός στοιχείου / κλάσης «i» στα δεδομένα.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (3)$$

όπου c είναι ο αριθμός των μοναδικών διαφόρων τιμών της κλάσης και p_i είναι η πιθανότητα της εκάστοτε τιμής i της κλάσης.

Έπειτα είναι το κέρδος πληροφορίας ο τύπος του οποίου δίνεται στην εξίσωση 4.

$$Gain(S, A) = Entropy(S) \sum_{v \in Values} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (4)$$

όπου A είναι η ανεξάρτητη μεταβλητή με τιμές $Values$, v είναι οι δυνατές τιμές του A, S_v είναι το πλήθος των εγγραφών με $A = v$ και S είναι το πλήθος όλων των εγγραφών. (Kingsford, 2008) Αυτό το τυχαίο δείγμα χαρακτηριστικών οδηγεί στη δημιουργία πολλαπλών ασυσχέτιστων δέντρων απόφασης. Το κέρδος της πληροφορίας βασίζεται στη

μείωση της εντροπίας μετά την κατάτμηση ενός συνόλου δεδομένων σε ένα χαρακτηριστικό. Η κατασκευή ενός δέντρου αποφάσεων αφορά αποκλειστικά την εύρεση ενός χαρακτηριστικού που επιστρέφει το υψηλότερο κέρδος πληροφορίας (δηλ. Τα πιο ομοιογενή κλαδιά). Υπάρχουν κάποια βήματα για την υλοποίηση αυτής της διαδικασίας. Πρώτα γίνεται υπολογισμός της εντροπίας του κάθε στόχου (μεταβλητής). Στην συνέχεια το σύνολο δεδομένων χωρίζεται σε διαφορετικά χαρακτηριστικά. Έπειτα υπολογίζεται η εντροπία για κάθε κλαδί. Ύστερα, προστίθενται αναλόγως, για να υπολογιστεί η συνολική εντροπία. Η εντροπία που προκύπτει αφαιρείται από την εντροπία πριν από τη διάσπαση. Το αποτέλεσμα είναι η αύξηση της πληροφορίας ή η μείωση της εντροπίας. Τέλος επιλέγεται το χαρακτηριστικό με το μεγαλύτερο κέρδος πληροφορίας ως κόμβος απόφασης, διαιρείται το σύνολο δεδομένων από τα κλαδιά του και επαναλαμβάνεται η ίδια διαδικασία σε κάθε κλαδί

Ο Naive Bayes είναι ένας αλγόριθμος μηχανικής μάθησης, βασισμένος στο θεώρημα του Bayes, που μπορεί να χρησιμοποιηθεί σε μια μεγάλη ποικιλία εργασιών ταξινόμησης, όπως η ταξινόμηση κειμένου, φιλτραρίσματος spam mail καθώς και ανάλυση συναισθήματος από κείμενο. Το όνομα “αφελές” (Naïve) χρησιμοποιείται επειδή ο αλγόριθμος υποθέτει ότι τα χαρακτηριστικά που εμπλέκονται στο μοντέλο είναι ανεξάρτητα το ένα από το άλλο. Αυτό σημαίνει ότι αλλάζοντας την τιμή ενός χαρακτηριστικού, δεν επηρεάζεται άμεσα η τιμή οποιουδήποτε από τα άλλα χαρακτηριστικά που χρησιμοποιούνται στον αλγόριθμο.

Το Θεώρημα του Bayes υπολογίζει την πιθανότητα να συμβεί κάποιο γεγονός, δεδομένης της πιθανότητας ενός άλλου γεγονότος που έχει ήδη συμβεί. Η εξίσωση (5) περιγράφει το θεώρημα Bayes που αναφέρθηκε παραπάνω.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{5}$$

Όπου τα A,B είναι σύνολα. Η γενικευμένη μορφή του παραπάνω τύπου για περισσότερα από 2 σύνολα δίνεται στην εξίσωση (6).

$$P(y/X) = \frac{P(X/y)P(y)}{P(X)} \tag{6}$$

Το y είναι η εξαρτημένη κλάση και το $X = (x_1, \dots, x_n)$ είναι ο ανεξάρτητος παράγοντας μεγέθους n. Κάνοντας χρήση της “αφελής” υπόθεσης ότι όλα τα σύνολα είναι ανεξάρτητα μεταξύ τους προκύπτει η εξίσωση (7).

$$P(y/X) = \frac{P(x_1/y)P(x_2/y) \dots P(x_n/y)P(y)}{P(x_1)P(x_2) \dots P(x_n)} \tag{7}$$

Χρησιμοποιώντας την εξίσωση (7) γίνεται η πρόβλεψη της εκάστοτε εξαρτημένης μεταβλητής λαμβάνοντας υπόψη τις ανεξάρτητες μεταβλητές.

Όταν οι προβλέψεις παίρνουν μια συνεχή τιμή και δεν είναι διακριτές, γίνεται η υπόθεση ότι αυτές οι τιμές λαμβάνονται από μια Gaussian κατανομή. Δεδομένου ότι ο τρόπος με τον οποίο οι τιμές που υπάρχουν στα δεδομένα αλλάζει, ο τύπος για την υπό συνθήκη πιθανότητα αλλάζει και δίνεται από την εξίσωση 8.

$$P(X/y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{\left(-\frac{(x_i-\mu_i)^2}{2\sigma_y^2}\right)} \quad (8)$$

Όπου μ_i είναι η μέση τιμή και σ_y η διακύμανση (Cichosz, 2015).

Χρησιμοποιήθηκαν δυο τεχνικές Feature Selection ο αλγόριθμος Extra tree Classifier και η κανονικοποίηση Lasso για να αποφασιστούν οι καταλληλότερες μεταβλητές που θα χρησιμοποιηθούν στης προβλέψεις. Ο ταξινομητής Extra Tree είναι ένας αλγόριθμος μηχανικής μάθησης, ο οποίος συγκεντρώνει τα αποτελέσματα πολλών ασυσχέτιστων δέντρων απόφασης που συλλέγονται από ένα «δάσος» για την εξαγωγή του αποτελέσματος ταξινόμησης. Σαν ιδέα, είναι παρόμοιο με έναν αλγόριθμό Random Forest ταξινομητή, ωστόσο διαφέρει στον τρόπο κατασκευής των δέντρων απόφασης.

Κάθε δέντρο απόφασης μέσα στο Extra Tree Forest κατασκευάζεται από τα αρχικά δεδομένα εκπαίδευσης. Στη συνέχεια, σε κάθε κόμβο των δέντρων απόφασης, κάθε δέντρο διαθέτει ένα τυχαίο δείγμα χαρακτηριστικών k από το σύνολο των χαρακτηριστικών από το οποίο πρέπει να επιλέξει το καλύτερο χαρακτηριστικό για το διαχωρισμό των δεδομένων βάσει ορισμένων μαθηματικών κριτηρίων. Αυτό το τυχαίο δείγμα χαρακτηριστικών οδηγεί στη δημιουργία πολλαπλών ασυσχέτιστων δέντρων απόφασης. Για την επιλογή χαρακτηριστικών χρησιμοποιώντας την παραπάνω διαδικασία, κατά την κατασκευή του δάσους, για κάθε χαρακτηριστικό, υπολογίζεται η κανονικοποιημένη συνολική μείωση των μαθηματικών κριτηρίων που χρησιμοποιήθηκαν στην απόφαση για την διαίρεση κάθε χαρακτηριστικού. Ένα από τα κριτήρια συνήθως είναι η εντροπία η οποία υπολογίζεται από την εξίσωση (3) και ένα δεύτερο το κέρδος πληροφορίας το οποίο υπολογίζεται από την εξίσωση (4).

Για την επιλογή χαρακτηριστικών (Feature Selection) χρησιμοποιώντας την παραπάνω διαδικασία, υπολογίζεται το κέρδος πληροφορίας κάθε μεταβλητής. Για να γίνει η επιλογή των χαρακτηριστικών, κάθε χαρακτηριστικό ταξινομείται σε φθίνουσα σειρά σύμφωνα με το κέρδος πληροφορίας που δίνει και ο χρήστης επιλέγει τα πρώτα k χαρακτηριστικά. Στην παρούσα εργασία ο τύπος που χρησιμοποιήθηκε για τον υπολογισμό του k είναι: $k = \sqrt{number\ of\ features}$.

Η κανονικοποίηση Lasso ανήκει στην ομάδα των Embedded methods. Στις ενσωματωμένες τεχνικές (Embedded methods), ο αλγόριθμος επιλογής χαρακτηριστικών ενσωματώνεται ως μέρος του αλγορίθμου εκμάθησης. Η μέθοδος LASSO εισάγει έναν περιορισμό στο άθροισμα των απόλυτων τιμών των παραμέτρων του μοντέλου, το άθροισμα πρέπει να είναι μικρότερο από μια σταθερή τιμή (άνω όριο). Για να γίνει αυτή η μέθοδος εφαρμόζεται μια διαδικασία κανονικοποίησης όπου τιμωρεί τους συντελεστές των μεταβλητών παλινδρόμησης κάνοντας μερικές από αυτές μηδέν. Κατά τη διαδικασία επιλογής των χαρακτηριστικών, οι μεταβλητές που εξακολουθούν να έχουν μη μηδενικό συντελεστή μετά τη διαδικασία κανονικοποίησης επιλέγονται ως μέρος του μοντέλου. Ο στόχος αυτής της διαδικασίας είναι να ελαχιστοποιήσει το σφάλμα πρόβλεψης. Στην πράξη, η παράμετρος συντονισμού λ, που ελέγχει την ισχύ της ποινής, έχει μεγάλη σημασία. Όταν οι παράμετροι είναι αρκετά μεγάλοι, τότε οι συντελεστές αναγκάζονται να είναι μηδέν, με αυτόν τον τρόπο μπορεί να μειωθεί η διάσταση. Όσο μεγαλύτερη είναι η παράμετρος λ, τόσο περισσότερος αριθμός συντελεστών

μειώνεται ξανά στο μηδέν. Από την άλλη πλευρά, εάν το λ = 0 έχουμε παλινδρόμηση OLS (Ordinary Least Sqaure).

Συνοψίζοντας, Ο αλγόριθμος Lasso δουλεύει με τον εξής τρόπο: Εάν το χαρακτηριστικό είναι ασυσχέτιστο, ο αλγόριθμος Lasso το τιμωρεί με συντελεστή 0. Ως εκ τούτου, τα χαρακτηριστικά με συντελεστή = 0 καταργούνται και τα υπόλοιπα λαμβάνονται υπόψη.

- Ανάλυση Επιβίωσης

Στο στάδιο αυτό μελετήθηκε το ποσοστό των ασθενών που επέζησαν σε διάρκεια 12 χρόνων από την τελευταία έξοδο τους από το νοσοκομείο. Για τις ανάγκες του προβλήματος διατηρήθηκαν οι εγγραφές από την τελευταία έξοδος από το νοσοκομείο και δημιουργήθηκε μια νέα μεταβλητή 'Dialysis_Days' στην οποία αποθηκεύτηκε η διάρκεια ζωής των ασθενών από την τελευταία έξοδο από το νοσοκομείο μέχρι το θάνατό τους. Ακόμη δημιουργήθηκε μια νέα μεταβλητή 'Agehosp' η οποία περιέχει την ηλικία των ασθενών κατά την τελευταία έξοδο τους από το νοσοκομείο. Τα δείγματα χωρίστηκαν σε αυτά σε κατάσταση αιμοκάθαρσης και σε αυτά σε κατάσταση μεταμόσχευσης ώστε να φανεί σε πια κατάσταση οι ασθενείς ζουν περισσότερο. Επιπλέον απομονωνόταν δείγματα με ηλικία μικρότερη των 65 ετών προέκυψε τι ποσοστό των δειγμάτων σε αιμοκάθαρση και μεταμόσχευση επέζησαν σε διάρκεια 12 χρόνων. Τέλος τα δείγματα χωρίστηκαν με βάση το φύλο τους ώστε να φανεί τι ποσοστό ανδρών και γυναικών επέζησαν σε διάρκεια 12 χρόνων. Χρησιμοποιήθηκε η μέθοδος Kaplan Meier για την εξαγωγή των αποτελεσμάτων.

Η χρήση του εκτιμητή Kaplan-Meier έγινε διότι μπορεί να υπολογίσει το ποσοστό επιβίωσης ακόμα και αν ο ασθενής φύγει ή πεθάνει κατά την διάρκεια του πειράματος. Αυτό δίνει την δυνατότητα ακριβέστερων αποτελεσμάτων ακόμη και με censored Data. Ο εκτιμητής Kaplan-Meier είναι μία από τις επιλογές που χρησιμοποιείται για τη μελέτη ενός υποσυνόλου των υποκειμένων που ζουν για ορισμένο χρονικό διάστημα μετά από μια υποβαλλόμενη θεραπεία. Σε κλινικές δοκιμές, η επίδραση μιας παρέμβασης αξιολογείται μετρώντας τον αριθμό των υποκειμένων που επιβίωσαν ή πέθαναν μετά από αυτή την παρέμβαση σε μια χρονική περίοδο. Η παρέμβαση αυτή μπορεί να είναι μια θεραπεία ή ένα σύνολο από ενέργειες. Ο χρόνος αρχίζει από την εκκίνηση της παρέμβασης έως την εμφάνιση ενός δεδομένου συμβάντος. Τα αποτελέσματα της ανάλυσης επιβίωσης μπορούν να επηρεαστούν από συμβάντα όπως η έλλειψη συνεργασίας από τους ασθενείς ή η αποχώρηση των ασθενών από τη μελέτη πριν αυτή ολοκληρωθεί, η μη βίωση του γεγονότος που μελετάται όπως για παράδειγμα ασθενείς στους οποίους χορηγείται μια θεραπεία αλλά δεν έχει τα επιθυμητά αποτελέσματα. Αυτές οι καταστάσεις ονομάζονται censored situations. Ο εκτιμητής Kaplan - Meier είναι ένας τρόπος υπολογισμού της συνάρτησης επιβίωσης με την πάροδο του χρόνου. Η καμπύλη επιβίωσης μπορεί να δημιουργηθεί με διάφορες καταστάσεις. Περιλαμβάνει τον υπολογισμό των πιθανοτήτων εμφάνισης του συμβάντος σε συγκεκριμένο χρονικό σημείο και τον πολλαπλασιασμό αυτών των διαδοχικών πιθανοτήτων με τυχόν προηγούμενες υπολογισμένες πιθανότητες για να ληφθεί η τελική εκτίμηση. Ο εκτιμητής για την μέθοδο αυτή δίνεται από τον τύπο:

$$\widehat{S(t)} = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right), \qquad (9)$$

Όπου t_i είναι ο χρόνος κατά τον οποίο έγινε τουλάχιστον ένα γεγονός, d_i είναι ο αριθμός των συμβάντων που έγιναν(π.χ. θάνατοι) μέχρι την χρονική περίοδο t_i, και n_i είναι τα άτομα που έχουν επιβιώσει (δεν τους έχει συμβεί ακόμα κάποιο γεγονός) μέχρι το χρόνο t_i (Andersen, 2014)

Ακόμη κατά τα πειράματα υπήρχαν και Censored Data. Τα censored data είναι μεταβλητές ενός προβλήματος για τις οποίες δεν υπάρχουν διαθέσιμες όλες οι πληροφορίες σχετικά με τον χρόνο εκδήλωσης του γεγονότος που μελετάται ή η έλλειψη πληροφορίας σχετικά με την κατάσταση του υποκειμένου που μελετάται λόγω απώλειας παρακολούθησης ή μη εκδήλωσης του γεγονότος που μελετάται, πριν την λήξη της μελέτης.

Υπάρχουν κάποιες κατηγορίες censored δεδομένων που θα αναφερθούν στην παρούσα εργασία. Αρχικά υπάρχουν τα *Right censored* δεδομένα τα οποία είναι δεδομένα στα οποία το συμβάν που μελετάται δεν συμβαίνει κατά την διάρκεια της μελέτης. Για παράδειγμα αν μελετάται η θνησιμότητα των ατόμων σε μια χρονική περίοδο τότε τα άτομα που έζησαν περισσότερο από την περίοδο της μελέτης θεωρούνται Right censored δεδομένα.

Έπειτα υπάρχουν τα *Left censored* δεδομένα. Η συγκεκριμένη κατηγορία εμφανίζεται όταν τα δεδομένα της μελέτης εμφανίσουν το συμβάν πριν τη λήξη της μελέτης αλλά δεν είναι γνωστό πότε συνέβη το συμβάν

Τέλος υπάρχουν τα *interval censored* δεδομένα. Η συγκεκριμένη κατηγορία εμφανίζεται όταν δεν υπάρχει συνεχείς χρονικές στιγμές της μελέτης. Για παράδειγμα μπορεί η παρακολούθηση των ατόμων που παίρνουν μέρος στην μελέτη να γίνεται ανά μήνα, χρόνο ή και περισσότερο.

Στην παρούσα εργασία τα δεδομένα που χρησιμοποιήθηκαν παρουσιάζουν τις 2 από τις τρεις προαναφερθείσες κατηγορίες. Υπάρχουν ασθενείς που έζησαν και μετά το πέρας των 12 χρόνων κάτι που υποδεικνύει Right censoring στα δεδομένα. Ακόμα η περίοδος παρακολούθησης των ασθενών δεν ήταν συνεχής κάτι που υποδεικνύει interval censoring στα δεδομένα. Η μέθοδος Kaplan Meier που χρησιμοποιήθηκε διαχειρίζεται τα censored δεδομένα δίνοντας τις ίδιες προοπτικές επιβίωσης με τα non-censored δεδομένα. Αυτό επιτυγχάνεται με τον υπολογισμό των πιθανοτήτων εμφάνισης ενός συμβάντος σε μια συγκεκριμένη χρονική στιγμή και τον πολλαπλασιασμό αυτών των διαδοχικών πιθανοτήτων με τυχόν προηγούμενες υπολογισμένες πιθανότητες για να πάρει την τελική εκτίμηση.

Η Ανάλυση Επιβίωσης χρησιμοποιείται για να παραχθούν συμπεράσματα σχετικά με την επιβίωση ασθενών όταν προστίθεται ένας παράγοντας όπως η λήψη μις θεραπείας σε μια χρονική περίοδο. Η χρήση των τεχνικών Ανάλυσης Επιβίωσης χρησιμοποιούνται ευρέως σε κλινικές δοκιμές. Χρησιμοποιώντας τεχνικές όπως ο εκτιμητής Kaplan-Meier το αποτέλεσμα μιας θεραπείας μετριέται με τον αριθμό των ασθενών που επιβίωσαν μετά την λήψη της θεραπείας. (Goel, 2010) Επιπλέον σύμφωνα με την εργασία (Cole, 2010) η Ανάλυση Επιβίωσης χρησιμοποιείται για την εύρεση του χρόνου ζωής από την διάγνωση μιας ασθένειας μέχρι το θάνατο του ασθενή. Χρησιμοποιώντας παραδείγματα ασθένειας όπως το AIDS, χρησιμοποιεί τεχνικές Ανάλυση Επιβίωσης για να εξετάσει από τι επηρεάζονται οι τεχνικές Ανάλυσης Επιβίωσης. Επιπρόσθετα χρησιμοποιώντας Ανάλυση Επιβίωσης σύμφωνα με την εργασία (İlkerEtikan*, 2018) συγκρίνονται ασθενείς με διαφορετικά

χαρακτηριστικά όπως το φύλο , η ηλικία, ως προς την διάρκεια ζωής τους. Αυτό προσφέρει ένα εύρος συμπερασμάτων ως προς την επιρροή μιας ασθένειας στις διάφορες κατηγορίες ασθενών.

- Αιτιολογική Ανάλυση

Χρησιμοποιώντας αποτελέσματα από την Περιγραφική Στατιστική όπως το πιο διαδεδομένο φάρμακο που χορηγείται έγινε μια Αιτιολογική Ανάλυση για να εξεταστεί αν το συγκεκριμένο φάρμακο επηρεάζει την διάρκεια ζωής των ασθενών.

Η αιτιότητα περιγράφει ιδέες σχετικά με τη φύση των σχέσεων αιτίας και αποτελέσματος. Η αιτία είναι ένα γεγονός που παράγει ή δημιουργεί ένα αποτέλεσμα. Η αιτιώδης επίδραση (causal effect) της λήψης μιας θεραπείας για ένα αντικείμενο Α είναι η σύγκριση των πιθανών αποτελεσμάτων του με την λήψη της θεραπείας και χωρίς αυτήν. Έστω το αποτέλεσμα Υ. Τότε η αιτιώδης επίδραση μιας θεραπείας στο αντικείμενο Α ορίζεται ως:

$$Y(Θεραπεία) - Y(Όχι\ θεραπεία). \quad (10)$$

Έστω ότι υπάρχει ένα $Y(0)$ που υποδηλώνει το πιθανό αποτέλεσμα ενός υποκειμένου με την απουσία θεραπείας και $Y(1)$ υποδηλώνει το αποτέλεσμα ενός υποκειμένου με την λήψη της θεραπείας. Έστω D υποδηλώνει την κατάσταση της θεραπείας. Το $D = 1$ υποδεικνύει λήψη της θεραπείας και $D = 0$ υποδεικνύει μη λήψη της θεραπείας και το Χ είναι ένα διάνυσμα με τα χαρακτηριστικά που επηρεάζουν το αποτέλεσμα όπως για παράδειγμα η ηλικία του ασθενή.

Για κάθε υποκείμενο $i, i = 1,2, \dots n$ το παρατηρούμενο αποτέλεσμα δίνεται από τον τύπο:

$$Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1). \quad (11)$$

Το σύνολο των παρατηρούμενων $(Y_i, D_i, X_i), i = 1,2, \dots, n$ είναι τα στοιχεία που χρησιμοποιούνται για τον υπολογισμό. Στο κείμενο που ακολουθεί περιγράφονται τέσσερις διαφορετικοί εκτιμητές της επίδρασης μιας μεταβλητής (αίτιο), φάρμακο Α στην μεταβλητή (αιτιατό), θνησιμότητα.

Αρχικά δημιουργήθηκε μια νέα μεταβλητή 'Age' η οποία περιέχει τις ηλικίες του δείγματος κατά την αγορά του φαρμάκου. Για τις ανάγκες της Ανάλυσης κρατήθηκαν μόνο τα δείγματα στα οποία χορηγήθηκε σύμφωνα με την Περιγραφική Στατιστική το πιο διαδεδομένο φάρμακο 'Trombyl'. Χρησιμοποιήθηκαν σαν μεταβλητές επιρροής η ηλικία, η πρώτη διάγνωση καθώς και το φύλο. Χρησιμοποιήθηκαν τεχνικές όπως Ordinary Least Squares, Matching και Weighting για την εξαγωγή των αποτελεσμάτων. Χρησιμοποιήθηκε και Propensity score ώστε να βοηθήσει στην καλύτερη λειτουργία των παραπάνω τεχνικών.

Η πιθανότητα λήψης της θεραπείας, γνωστή και ως Propensity Score, παίζει πολύ ιδιαίτερο ρόλο στην εκτίμηση των επιδράσεων της θεραπείας. Χωρίς βλάβη της γενικότητας μπορεί να θεωρηθεί ότι τα αποτελέσματα Υ(0), Υ(1) είναι ανεξάρτητα από την εκχώρηση θεραπείας:

$$(Y(0), Y(1)) \perp D|X. \quad (12)$$

Με άλλα λόγια για την υπό προϋποθέσεις πιθανότητα Propensity Score $p(X) = P(D = 1|X)$ η λήψη της θεραπείας είναι ουσιαστικά τόσο καλή όσο και τυχαία (Wong). Αυτό

σημαίνει ότι για τα άτομα που μοιράζονται το ίδιο Propensity Score, η διαφορά μεταξύ των ατόμων που λαμβάνουν την θεραπεία και των ατόμων τα οποία δεν λαμβάνουν την θεραπεία αναγνωρίζει ένα μέσο όρο θεραπείας, δηλαδή $E[Y(1) - Y(0)|p(X)]$. Επομένως αντί να γίνει συσχέτιση στα covariate διανύσματα X, γίνεται συσχέτιση στα Propensity Score $p(X)$ των διανυσμάτων αυτών, και υπάρχει ακόμα και μια έγκυρη εκτίμηση του συνολικού μέσου αποτελέσματος της θεραπείας.

Το Propensity Score, μπορεί να εκτιμηθεί με βάση τα δεδομένα για τις παρατηρήσιμες μεταβλητές D και X.

Ο παρακάτω αλγόριθμος που παρουσιάζεται στην αναφορά[1] (Imbens, 2015) ασχολείται με την επιλογή μεταβλητών για την εκτίμηση του Propensity score.

- Στο πρώτο βήμα γίνεται η επιλογή ενός συνόλου συντελεστών X_B που θα επηρεάζουν το αποτέλεσμα και χρήση του για την πρόβλεψη της μεταβλητής μέσω λογιστικής παλινδρόμησης.
- Στο δεύτερο βήμα Χρήση μιας επιπλέον μεταβλητής $X \neq X_B$ και επαναχρησιμοποίηση λογιστικής παλινδρόμησης. Υπολογισμός του test likelihood-ratio, το οποίο αξιολογεί την καλή εφαρμογή στατιστικών μοντέλων με βάση την αναλογία των πιθανοτήτων. Αν για τα δεδομένα ισχύει η μηδενική υπόθεση δηλαδή ο περιορισμός που έχει τεθεί, τότε οι πιθανότητες των δυο μοντέλων δεν πρέπει να διαφέρουν πολύ.
- Στο τρίτο βήμα γίνεται επανάληψη του δεύτερου βήματος για κάθε μεταβλητή $X \neq X_B$. Εάν η τιμή του στατιστικού κριτηρίου είναι μεγαλύτερη από κάποια τιμή C_{lin}, η μεταβλητή αυτή συμπεριλαμβάνεται στο βασικό σύνολο ομοιοτήτων και γίνεται επανεκτέλεση του δεύτερου βήματος στο νέο σύνολο. Διαφορετικά απορρίπτεται διότι θεωρείται μη σημαντική.
- Στο τέταρτο βήμα γίνεται επανάληψη του δεύτερου και τρίτου βήματος ώστε να αποφασιστεί ποιες από τις αλληλεπιδράσεις και τους τετραγωνικούς όρους θα συμπεριληφθούν στην επιλογή του propensity Score. Χρησιμοποιώντας το σύνολο των συντελεστών του πρώτου βήματος επιλέγεται ποιοι από τους τετραγωνικούς όρους ή οι συνδυασμοί των συντελεστών θα χρησιμοποιηθούν στην επιλογή του propensity score.

Τα μοντέλα γραμμικής παλινδρόμησης έχουν πολλές εφαρμογές στην πραγματική ζωή. Στην οικονομετρία, η μέθοδος Ordinary Least Squares(OLS) χρησιμοποιείται ευρέως για την εκτίμηση των παραμέτρων ενός μοντέλου γραμμικής παλινδρόμησης. Για την εγκυρότητα των εκτιμήσεων OLS, υπάρχουν υποθέσεις που γίνονται κατά τη λειτουργία μοντέλων γραμμικής παλινδρόμησης.

1. Το μοντέλο γραμμικής παλινδρόμησης είναι "γραμμικό σε παραμέτρους".
2. Υπάρχει τυχαία δειγματοληψία παρατηρήσεων.
3. Ο υπό προϋποθέσεις μέσος όρος πρέπει να είναι μηδέν.
4. Δεν υπάρχει multi-collinearity (ή perfect collinearity).

[1] Imbens, G. & Rubin, D. (2015). Causal inference in statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

5. Σφαιρικά σφάλματα: Υπάρχει ομοσκεδαστικότητα και δεν υπάρχει αυτοσυσχέτιση.

Αυτές οι υποθέσεις είναι εξαιρετικά σημαντικές επειδή η παραβίαση οποιασδήποτε από αυτές τις παραδοχές θα έκανε τις εκτιμήσεις της OLS αναξιόπιστες και ανακριβείς. Συγκεκριμένα, μια παραβίαση θα οδηγούσε σε λανθασμένες εκτιμήσεις του OLS ή η διακύμανση των εκτιμήσεων OLS θα ήταν αναξιόπιστη, οδηγώντας σε διαστήματα εμπιστοσύνης που είναι υπερβολικά ευρεία ή πολύ στενά.

Ο τύπος που περιγράφει την μέθοδο αυτή είναι ο ακόλουθος:

$$Y_i = a + \beta D_i + \gamma'(X_i - \bar{X}) + \delta' D_i(X_i - \bar{X}) + \varepsilon, \quad (13)$$

Όπου $\beta = E[Y_i(1) - Y_i(0)]$ και $\gamma' = E[Y_i(1) - Y_i(0)|D_i = 1]$. Με Ε συμβολίζεται το υπό συνθήκη ενδεχόμενο.

<u>Ιδιότητες του Ordinary Least Squares(OLS) εκτιμητή</u>

<u>Ιδιότητα 1: Γραμμικότητα</u>

Η γραμμική ιδιότητα του εκτιμητή OLS σημαίνει ότι ο OLS ανήκει στην κατηγορία εκτιμητών, που η εκτίμηση τους για την εξαρτημένη μεταβλητή είναι γραμμική συσχέτιση των ανεξάρτητων μεταβλητών.

<u>Ιδιότητα 2 : Αμεροληψία</u>

Ο OLS εκτιμητής προσφέρει αμεροληψία όσον αφορά την εκτίμηση της εξαρτημένης μεταβλητής. Έτσι και η πρόβλεψη στην εξαρτημένης μεταβλητής θα είναι αμερόληπτη.

<u>Ιδιότητα 3:Ασυμπτωτική αμεροληψία</u>

Αυτή η ιδιότητα του OLS λέει ότι καθώς αυξάνεται το μέγεθος του δείγματος, ο εκτιμητής γίνεται πιο αμερόληπτος.

<u>Ιδιότητα 4: Συνέπεια</u>

Ένας εκτιμητής λέγεται ότι είναι συνεπής αν η τιμή του προσεγγίζει την πραγματική τιμή παραμέτρου (πληθυσμού) όσο αυξάνεται το μέγεθος του δείγματος. Ένας εκτιμητής είναι συνεπής αν ικανοποιεί δύο προϋποθέσεις:

1. Είναι αμερόληπτος
2. Η διακύμανσή του συγκλίνει στο 0 καθώς αυξάνεται το μέγεθος του δείγματος.

Οι δύο αυτές προϋποθέσεις ισχύουν για τον OLS εκτιμητή και, ως εκ τούτου, είναι συνεπής εκτιμητής. Για να είναι χρήσιμος ο εκτιμητής, η συνέπεια είναι η ελάχιστη βασική απαίτηση. (Rahman, 2018)

Ο εκτιμητής Matching αξιολογεί τα αποτελέσματα μιας θεραπείας συγκρίνοντας τα αποτελέσματα για τα άτομα που υποβλήθηκαν σε θεραπεία με αυτά των ατόμων που δεν υποβλήθηκαν στη θεραπεία. Τα άτομα που έχουν λάβει τη θεραπεία ‘ζευγαρώνουν’ με άτομα που δεν έχουν λάβει την θεραπεία χρησιμοποιώντας μια συνάρτηση ομοιότητας, η

οποία επιλέγει ζευγάρια με παρόμοια χαρακτηριστικά. Ο Matching εκτιμητής χρησιμοποιεί ζεύγη από τα σύνολα τον ασθενών που έλαβαν την θεραπεία και αυτών που δεν την έλαβαν προσαρμόζοντας απευθείας τις μεταβλητές που τους επηρεάζουν. Πιο συγκεκριμένα χρησιμοποιεί τον τύπο:

$$m(i) = argmin\left|\left|X_j - X_i\right|\right|, \quad (14)$$

Όπου το $\left|\left|X_j - X_i\right|\right|$ είναι κάποιο μέτρο ομοιότητας όπως η ευκλείδεια απόσταση. Τα X_j, X_i είναι διανύσματα που περιέχουν τα χαρακτηριστικά του ατόμου που λαμβάνει την θεραπεία με του ατόμου που δεν τη λαμβάνει.

Στην παρούσα εργασία χρησιμοποιήθηκε ο Matching εκτιμητής Nearest Neighborhood Matching. Το άτομο από την ομάδα που δεν έχει λάβει τη θεραπεία επιλέγεται ως αντιστοιχισμένος συνεργάτης για ένα άτομο που έχει λάβει τη θεραπεία σύμφωνα με ένα μέτρο ομοιότητας. Υπάρχουν δύο τρόποι χρησιμοποίησης του εκτιμητή. Ο πρώτος είναι ότι κάθε άτομο από το σύνολο των ατόμων που δεν έλαβαν τη θεραπεία μπορεί να αντιστοιχηθεί με περισσότερα από ένα άτομο από το σύνολο των ατόμων που έχουν λάβει τη θεραπεία. Ο δεύτερος τρόπος είναι όταν το κάθε άτομο από το σύνολο των ατόμων που δεν έχουν λάβει τη θεραπεία αντιστοιχηθεί μόνο με ένα άτομο από το σύνολο των ατόμων που έχουν λάβει τη θεραπεία. Η διαδικασία αντιστοίχισης επαναλαμβάνεται ωσότου υπάρξει το βέλτιστο αποτέλεσμα (Becker, 2002).

Ο εκτιμητής Weighting δουλεύει με τον παρακάτω τρόπο.

Αρχικά, ισχύουν οι ακόλουθες δύο εξισώσεις:

$$E[DYp(X)] = E[Y(1)] \quad (15)$$

$$E[(1-D)Y1-p(X)] = E[Y(0)]. \quad (16)$$

Η διαφορά μεταξύ αυτών των δύο μέσων είναι συνεπώς ένας έγκυρος εκτιμητής του μέσου αποτελέσματος θεραπείας $E\ [Y\ (1)\ -Y\ (0)]$. Ο συγκεκριμένος εκτιμητής υπολογίζει την αντίστροφη πιθανότητα Propensity score $\frac{1}{p(X)}$ η οποία αντιπροσωπεύει την πιθανότητα να έχει χορηγηθεί θεραπεία στον ασθενή και την πιθανότητα $\frac{1}{1-p(X)}$ να μην έχει χορηγηθεί θεραπεία στον ασθενή.

Παρόλο που το πραγματικό Propensity Score είναι σπάνια γνωστό στην πράξη, είναι γνωστό ποια άτομα είχαν πράγματι λάβει την θεραπεία μαζί με των αριθμό των παραγόντων που επηρεάζουν το αποτέλεσμα (Matschinger, 2020). Έτσι μπορεί να χρησιμοποιηθεί ο παρακάτω τροποποιημένος τύπος:

$$\mathrm{E}[\mathrm{Y}(1) - \mathrm{Y}(0)] = \left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{D_i}}{\hat{\mathrm{p}}(\mathrm{X_i})}\right)^{-1} \left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\mathrm{D_i Y_i}}{\hat{\mathrm{p}}(\mathrm{X_i})}\right) - \left(\frac{1-\mathrm{D_i}}{1-\hat{\mathrm{p}}(\mathrm{X_i})}\right)^{-1} \left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{(1-\mathrm{D_i})\mathrm{Y_i}}{\hat{\mathrm{p}}(\mathrm{X_i})}\right). \quad (17)$$

4. Πειραματική μελέτη

4.1. Περιγραφή μεταβλητών

Στην παρούσα εργασία χρησιμοποιήθηκαν δεδομένα ασθενών με νεφρική ανεπάρκεια. Στην οξεία μορφή της η νεφρική ανεπάρκεια οδηγεί στη παύση λειτουργίας των νεφρών. Τα νεφρά απομακρύνουν τα τοξικά προϊόντα του μεταβολισμού και βοηθούν στην ισορροπία νερού και ηλεκτρολυτών στον οργανισμό. Όταν τα νεφρά σταματήσουν να λειτουργούν, τα άχρηστα προϊόντα του μεταβολισμού, τα υγρά και οι ηλεκτρολύτες συσσωρεύονται στο σώμα. Αυτό μπορεί να προκαλέσει προβλήματα, που μπορεί να οδηγήσουν μέχρι και τον θάνατο. Η χρόνια νεφρική ανεπάρκεια είναι μια ασθένεια κατά την οποία τα νεφρά δεν επιτελούν σωστά το έργο τους. Έτσι συμβαίνει συσσώρευση των άχρηστων ουσιών του μεταβολισμού στον οργανισμό.

Τα δεδομένα που χρησιμοποιούνται, είναι ιατρικά δεδομένα τα οποία περιέχουν μεταξύ άλλων και εργαστηριακά δεδομένα ατόμων που κατοικούν ή έχουν πρόσβαση στην υγειονομική περίθαλψη στην περιοχή της Στοκχόλμης. Τα άτομα αυτά υποβλήθηκαν σε αξιολογήσεις κρεατινίνης μεταξύ των ετών 2006-11. Το πλήθος των ασθενών που χρησιμοποιήθηκαν ήταν 500.000. Οι εργαστηριακές εξετάσεις περιλαμβάνουν τις μετρήσεις-αποτελέσματα των εργαστηριακών ελέγχων. Ωστόσο, υπάρχουν στο σύνολο των δεδομένων οι πίνακες με τα δημογραφικά στοιχεία των ασθενών, με τα χορηγούμενα φάρμακα, με την κατάσταση των ασθενών αναφορικά με το στάδιο στο οποίο βρίσκονται, με το αν είναι ζωντανοί ή όχι, καθώς και πληροφορίες με την εισαγωγή τους στο νοσοκομείο.

Στην παρούσα εργασία χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python. Η Python διαθέτει βιβλιοθήκες για περιγραφική στατιστική όπως είναι η βιβλιοθήκη Pandas και η βιβλιοθήκη Numpy οι οποίες χρησιμοποιήθηκαν στην παρούσα εργασία. Ακόμη η βιβλιοθήκη Matplotlib διαθέτει μια μεγάλη ποικιλία από γραφήματα κάποια από τα οποία χρησιμοποιήθηκαν στην παρούσα εργασία. Επιπρόσθετα, η βιβλιοθήκη scikit learn η οποία διαθέτει μεγάλη ποικιλία από αλγορίθμους Μηχανικής Μάθησης κα δίνει την δυνατότητα επιλογής των κατάλληλων παραμέτρων ώστε να επιτευχθεί η καλύτερη προσαρμογή τους στα δεδομένα. Επιπλέον διαθέτει την βιβλιοθήκη lifelines που προσφέρει πολλές δυνατότητες για Ανάλυση θνησιμότητας. Τέλος διαθέτει την βιβλιοθήκη causalinference που προσφέρει πολλές δυνατότητες για Causal Analysis.

Ακόμη, στην ενότητα αυτή, παρατίθενται οι πίνακες που περιγράφουν τις τιμές που παίρνουν οι μεταβλητές που χρησιμοποιήθηκαν στην παρούσα εργασία. Η χρήση αυτής της ενότητας γίνεται για την πιο εύκολη κατανόηση των αποτελεσμάτων που παρατίθενται στην εργασία αυτή. Στην παρούσα εργασία χρησιμοποιείται ο κατάλογος ταξινόμησης των νόσων από τον παγκόσμιο οργανισμό υγείας για την εύρεση των διαφόρων ασθενειών και θεραπειών που εμφανίζονται στο σύνολο δεδομένων. (World Health Organization, 2018) Στον Πίνακα 1 περιέχονται όλα τα ονόματα των εργαστηριακών τεστ που υπάρχουν στο σύνολο των δεδομένων. Στον Πίνακα 2 χρησιμοποιούμενων μεταβλητών περιέχονται όλες οι μεταβλητές που χρησιμοποιήθηκαν στο κεφάλαιο 3. Ο Πίνακας 3 περιέχει κωδικοποιημένες τους δήμους στους οποίους έγινε η πρώτη μέτρηση της κρεατινίνης του κάθε ασθενή. Ο Πίνακας 4 περιέχει δυο μεταβλητές ναι ή όχι ανάλογα με το αν ο εκάστοτε ασθενής έκανε την μέτρηση εντός ή εκτός νοσοκομείου. Στον Πίνακα 5 περιέχονται τέσσερις μεταβλητές που ορίζουν σε ποιο στάδιο της ασθένειας βρίσκεται ο εκάστοτε ασθενής. Ο Πίνακας 6 περιέχει δυο κατηγορίες που περιγράφουν αν η εκάστοτε μέτρηση έγινε μέσα ή εκτός νοσοκομείου. Ο Πίνακας 7 περιέχει δυο

τιμές 0 και 1 που αντιπροσωπεύουν το ενδεχόμενο η εκάστοτε μέτρηση να έγινε στην περιοχή της Στοκχόλμης ή όχι. Ο Πίνακας 8 περιέχει δυο τιμές 0 και 1 που αντιπροσωπεύουν το φύλο του ασθενή. Ο Πίνακας 9 περιέχει κωδικούς που αντιπροσωπεύουν την περιοχή κατοικίας των ασθενών. Ο Πίνακας 10 περιέχει κωδικούς που αντιπροσωπεύουν την περιοχή που διαμένει ο κάθε ασθενής. Πιο συγκεκριμένα, τα δύο πρώτα ψηφία δείχνουν τον νομό, τα επόμενα δυο το δήμο μέσα στο νομό, και τα υπόλοιπα δυο δείχνουν τη συνοικία. Τέλος ο Πίνακας 11 περιέχει κωδικούς που αντιπροσωπεύουν το χώρο στον οποίο ο εκάστοτε ασθενής νοσηλεύεται.

Value/Code	Name	Acronym in CSV file	Description	Measurement Unit
Creatinine	Creatinine	screa	The clinical attribute of creatinine	ΞΌmol/l
Cystatin-c	Cystatin-c	cysc	The clinical attribute of Cystatin-c	NA
Serum albumin	Serum albumin	alb	The clinical attribute of Serum albumin	g/l
Haemoglobin	Haemoglobin	hb	The clinical attribute of Haemoglobin	g/l
Glucose	Glucose	glucose	The clinical attribute of Glucose	mmol/l
C-reactive protein	C-reactive protein	crp	The clinical attribute of C-reactive protein	mg/l
Parathyroid	Parathyroid	pth	The clinical attribute of Parathyroid	pg/ml
Dipstick albuminuria	Dipstick albuminuria	dpr	The clinical attribute of Dipstick albuminuria	dipstick
Albumin-to-creatinine ratio	Albumin-to-creatinine ratio	acr	The clinical attribute of Albumin-to-creatinine ratio	mg/mmol
Potassium	Potassium	potassium	The clinical attribute of Potassium	mmol/l
Cholesterol	Cholesterol	chol	The clinical attribute of Cholesterol	mmol/l
Sodium	Sodium	sodium	The clinical attribute of Sodium	mmol/l

Thyroid-Stimulating Hormone	Thyroid-Stimulating Hormone	tsh	The clinical attribute of Thyroid-Stimulating Hormone	mU/l
Low density lipoprotein cholesterol	Low density lipoprotein cholesterol	ldlc	The clinical attribute of Low density lipoprotein cholesterol	mmol/l
High Sensitivity C-reactive protein	High Sensitivity C-reactive protein	crp_hs	The clinical attribute of high Sensitivity C-reactive protein	mg/l
Glycated Haemoglobin	Glycated Haemoglobin	hba1c	The clinical attribute of glycated Haemoglobin	%
Calcium	Calcium	calcium	The clinical attribute of calcium	mmol/l
Protein kinase	Protein kinase	pk	The clinical attribute of protein kinase	NA
Free T4	Free T4	t4_free	The clinical attribute of free T4	pmol/l
Phosphate	Phosphate	phosphate	The clinical attribute of phosphate	mmol/l
Urine albumin	Urine albumin	ualb	The clinical attribute of urine albumin	mg/l
Bicarbonate	Bicarbonate	bicarbonate	The clinical attribute of bicarbonate	mmol/l
Free T3	Free T3	t3_free	The clinical attribute of free T3	pmol/l
T4	T4	t4	The clinical attribute of t4	nmol/l
Carbon dioxide	Carbon dioxide	co2	The clinical attribute of carbon dioxide	mmol/l
T3	T3	t3	The clinical attribute of t3	nmol/l

high-density lipoprotein cholesterol	high-density lipoprotein cholesterol	hdlc	The clinical attribute of high-density lipoprotein cholesterol	mmol/l
LDL or HDL cholesterol	LDL or HDL cholesterol	ldlhdlc	The clinical attribute of LDL or HDL cholesterol	%
triglycerides	triglycerides	trig	The clinical attribute of triglycerides	mmol/l

Πίνακας 1 Εργαστηριακές εξετάσεις

Name/Code	**Description**
Kommun	(The municipality of residency at date of first creatinine measurement)
Diag (1,2,3,4,5,6,7,8,9,10)	(The ICD-10 diagnostic codes)
Opk (1,2,3,4,5,6,7,8,9,10)	(The NOMESCO surgical procedure codes)
Female	(female or not)
rightOP	(The answer whether it was an out-patient measurement or not)
rightIP	(The answer whether it was an out-patient measurement or not)
vtype	(The aggregation of rightIP and rightOP)
Ulorsak	(The cause that led to the specific citizen's death)
ev	(The renal-replacement therapy event)
analys	(The kind of the performed laboratory test)
Atc	(The ATC code of the medication)
Spkod (1,2,3,4,5,6,7,8,9,10)	(The prescriber's speciality code)
Forddd	(The recommended daily drug dosage)
Resident	(The binary indicator of residency in the County of Stockholm for each measurement)

resultat	(The value of the performed test)
patkost	(The total cost of the medication, covered by the citizen)
datum	(The date of renal-replacement therapy event)
dodsdat	(The date that a specific citizen died)
utdat	The date of discharge from Hospital
fodelsedatum	The date that the citizen was born
edatum	The date that the medication was purchased
Ikf	The county/municipality/assembly code of residency relative to each measurement
Lan_d	The county of residency at date of first creatinine measurement
idx_dt	(The date that the Creatinine was firstly calculated)
usaett	(The code of discharge to modality)
akut	The answer whether hospitalization was pre-planned or not
Spkod(1,2,3)	The prescriber's speciality code
Lan	The county of residency relative to each measurement

Πίνακας 2 Περιγραφή χρησιμοποιούμενων μεταβλητών

Value/Code	Name	
0100-9999	0100-9999	- the numbers 1-2 indicate the county - the numbers 3-4 indicate the municipality (within the county).
0198	0198	Stockholmers with a protected address
2998	2998	foreigner with protected address
3399	3399	foreign patient (incl. Foreign Swedish)
3499	3499	not registered with a personal identification number
9999	9999	unknown
OB99	OB99	unknown (in some cases).

Πίνακας 3 Κωδικοί περιοχών

Value/Code	Name	Description
0	NO	The answer is no
1	YES	The answer is yes

Πίνακας 4 Η απάντηση αν ήταν μέτρηση έγινε εντός ή εκτός νοσοκομείου

Value/Code	Name	
TX	Transplant	The renal-replacement therapy was in the form of transplant
D	Dialysis initiation	The renal-replacement therapy was in the form of dialysis
RECOVERED	Recovered	The renal-replacement therapy was in the form of recovered
UNTRACED	Untraced	The renal-replacement therapy was in the form of untraced

Πίνακας 5 Στάδια των ασθενών

Value/Code	Name	Description
IH	IH	In-hospital measurement
OP	OP	Out-patient measurement
PC	PC	Primary care measurement

Πίνακας 6 Είδος μέτρων για τους ασθενείς

Value/Code	Name	Description
0	NO	The answer is no
1	YES	The answer is yes

Πίνακας 7 Δυαδικός δείκτης κατοικίας στην κομητεία της Στοκχόλμης για κάθε μέτρηση

Value/Code	Name	Description
0	Male	The gender is male
1	Female	The gender is female

Πίνακας 8 Δυαδικός δείκτης του φύλου

Value/Code	Name	
01	01	Stockholm County
03-25	03-25	other counties in Sweden
29	29	foreigner with protected address
33	33	Foreign Patient (including foreign Swedes)
34	34	not registered with a personal identification number
99	99	unknown
OB	OB	unknown (in some special cases).

Πίνακας 9 Κωδικός για τον νομό κατοικίας σε σχέση με κάθε μέτρηση

Value/Code	Name	
010000-999999	010000-999999	- the numbers 1-2 indicate the county - the numbers 3-4 indicate the municipality (within the county) - the numbers 5-6 indicate the parish (within the municipality).
019899	019899	Stockholmers with a protected address
299899	299899	foreigner with protected address
339999	339999	foreign patient (in some cases codes such as 333333)
349999	349999	not registered with a personal identification number
999999	999999	unknown
OB9999	OB9999	unknown (in some special cases).

Πίνακας 10 Κωδικός για νομό / δήμο / συγκέντρωσης κατοικίας σε σχέση με κάθε μέτρηση

Value/Code	Name	

0	0	To primary municipal care, SMHI (service residence with helin settlement)
1	1	To the home
2	2	To another clinic within the same institution
3	3	To another establishment
4	4	Dead, autopsy
5	5	Deceased, not autopsy
6	6	To the Red Cross (KS only)
7	7	Death
8	8	Change of patient class (KS only)
9	9	To another establishment in the same sector.

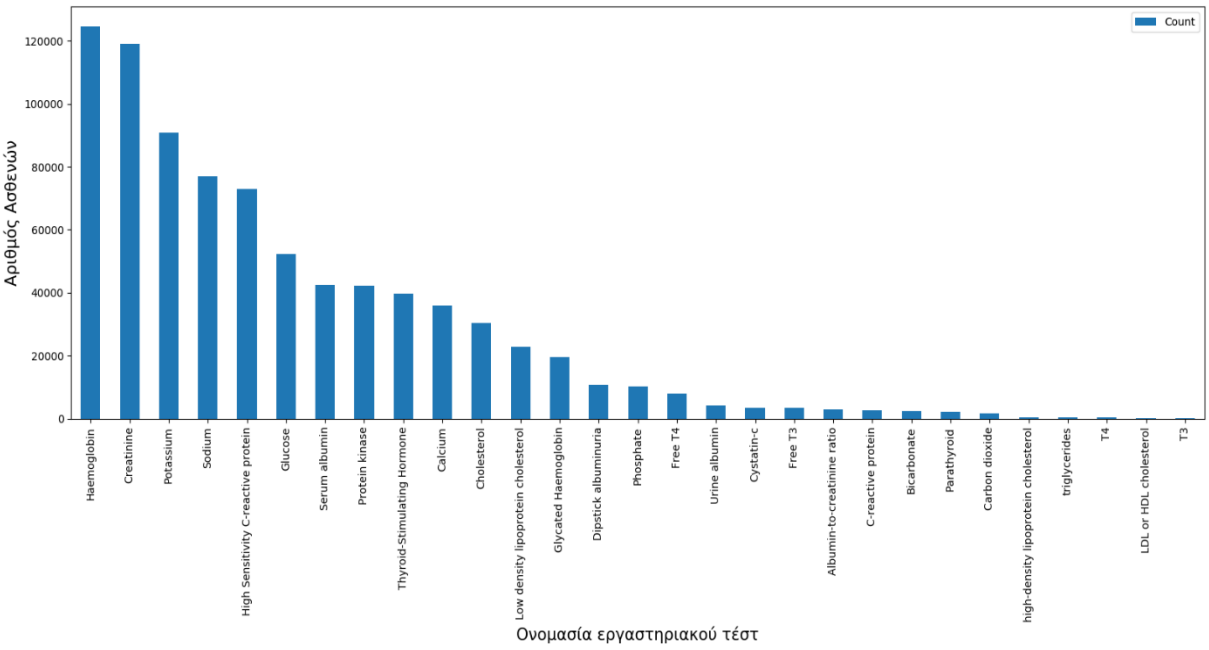
Πίνακας 11 Κωδικός για την περιγραφή του είδους θεραπείας

4.2. Αποτελέσματα Περιγραφικής Στατιστικής

4.2.1. Στατιστικά αποτελέσματα

Στο κεφάλαιο αυτό δίνεται μια πρώτη εικόνα των δεδομένων. Παρακάτω παρατίθενται τα αποτελέσματα τα οποία εξήχθησαν.

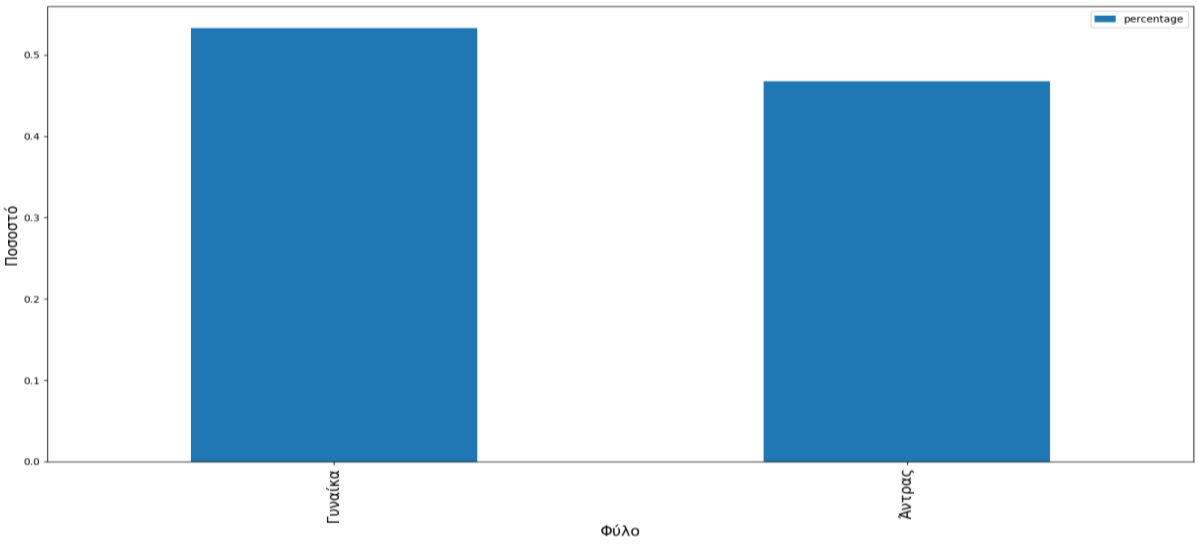
Στην Εικόνα 1 φαίνεται το ιστόγραμμα που δείχνει ποιο εργαστηριακό τεστ γίνεται σε διαφορετική περιοχή από τον τόπο διαμονής. Το πλήθος των δειγμάτων ήταν 824.716. Το πλήθος των ασθενών που έκαναν το εκάστοτε τεστ ήταν 23.693. Όπως φαίνεται και στο ιστόγραμμα για το “Hemoglobin Test” οι περισσότεροι ασθενείς αλλάζουν περιοχή.



Εικόνα 1 Ιστόγραμμα συχνοτήτων διεξαγωγής τύπου τεστ σε διαφορετική περιοχή από τη περιοχή πρώτης μέτρησης της κρεατινίνης

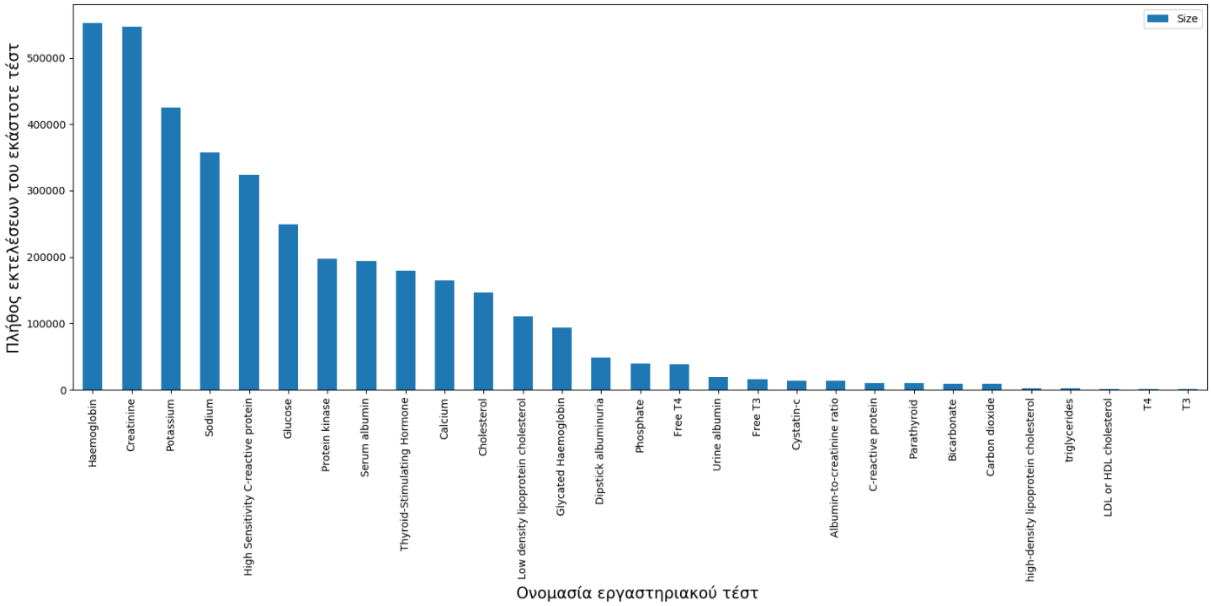
Εδώ αξίζει να σημειωθεί ότι οι περισσότεροι ασθενείς είτε μένουν μόνιμα στη Στοκχόλμη είτε πηγαίνουν εκεί για να κάνουν τις περισσότερες εξετάσεις.

Στην Εικόνα 2 φαίνεται το ποσοστό των ασθενών που είναι γυναίκες και άντρες. Σύμφωνα με το σχήμα το 53% των ασθενών είναι γυναίκες και το 46% είναι άντρες.



Εικόνα 2 Ιστόγραμμα συχνοτήτων για το ποσοστό αντρών και γυναικών

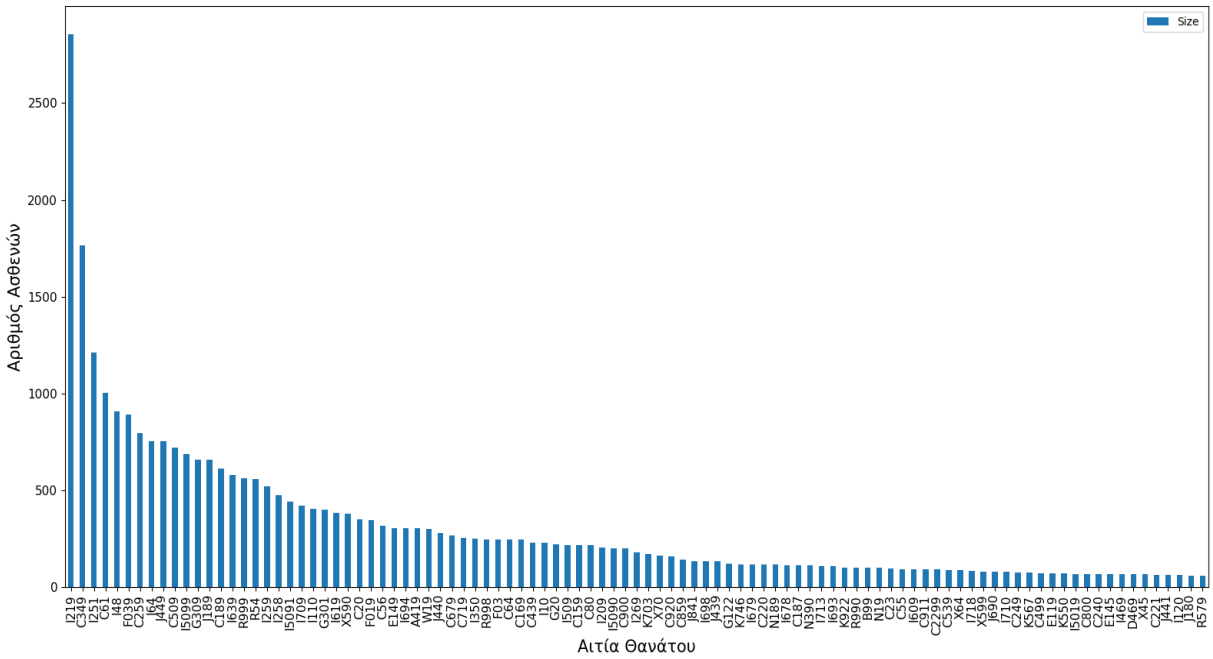
Στην Εικόνα 3 φαίνεται πιο εργαστηριακό τεστ γίνεται πιο συχνά στους ασθενείς. Το τεστ, το οποίο γίνεται πιο συχνά, είναι το “Hemoglobin Test”. Το πλήθος των δειγμάτων ήταν 3.777.881, τα οποία δημιουργήθηκαν από 500.000 ασθενείς. Αυτό δείχνει ότι στα αρχικά στάδια της νόσου οι ειδικοί θέλοντας να δουν αν τα συμπτώματα υποδεικνύουν νεφρική ανεπάρκεια προχωράνε σε περαιτέρω εξετάσεις όσον αφορά την μεταβολή των αιμοσφαιρίων στο αίμα κάτι που εμφανίζεται στα πρώτα στάδια της νεφρικής ανεπάρκειας.



Εικόνα 3 Ιστόγραμμα συχνοτήτων για το πιο συχνά εκτελέσιμο εργαστηριακό τεστ

Στην Εικόνα 4 φαίνονται οι αιτίες θανάτου και ο αριθμός των ασθενών που απεβίωσαν από την εκάστοτε αιτία. Το πλήθος των δειγμάτων ήταν 500.000. Οι χαμένες τιμές, δηλαδή οι άνθρωποι που επιβίωσαν, οι οποίες και διαγράφτηκαν ήταν 461.686. Οπότε το τελικό πλήθος δειγμάτων ήταν

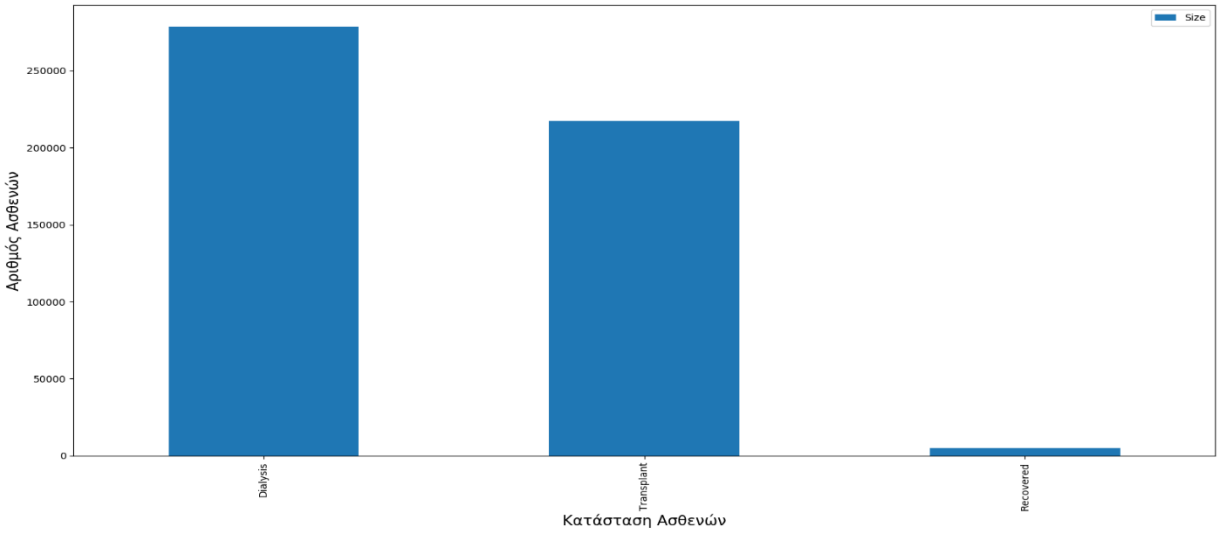
38.314. Όπως φαίνεται στο σχήμα ο κωδικός με τους περισσότερους θανάτους είναι ο I219, ο οποίος αντιστοιχίζεται σε “Οξύ έμφραγμα του μυοκαρδίου” [2]. Η δεύτερη πιο διαδεδομένη αιτία θανάτου είναι το “Κακοήθες νεόπλασμα στο αναπνευστικό σύστημα και τα ενδοθωρακικά όργανα”. Το συμπέρασμα, το οποίο προκύπτει είναι ότι οι περισσότεροι ασθενείς με νεφρική ανεπάρκεια δεν πεθαίνουν την νόσο, αλλά η συγκεκριμένη ασθένεια δημιουργεί άλλα προβλήματα, στην καρδιά κυρίως, τα οποία είναι θανατηφόρα.



Εικόνα 4 Ιστόγραμμα συχνοτήτων με το πλήθος των ασθενών ανά αιτία θανάτου

Στην Εικόνα 5 φαίνεται ο αριθμός των ασθενών σε καθένα από τις τρεις καταστάσεις της νεφρικής ανεπάρκειας. Το πλήθος των δειγμάτων που χρησιμοποιήθηκε ήταν 500.000 το οποίο προήλθε από 500.000 ασθενείς. Εδώ πρέπει να αναφερθεί ότι έγινε αντικατάσταση των χαμένων τιμών με τυχαία σειρά. Η αντικατάσταση έγινε με βάση των ήδη υπάρχοντών εγγραφών οπότε είναι αντιπροσωπευτική.

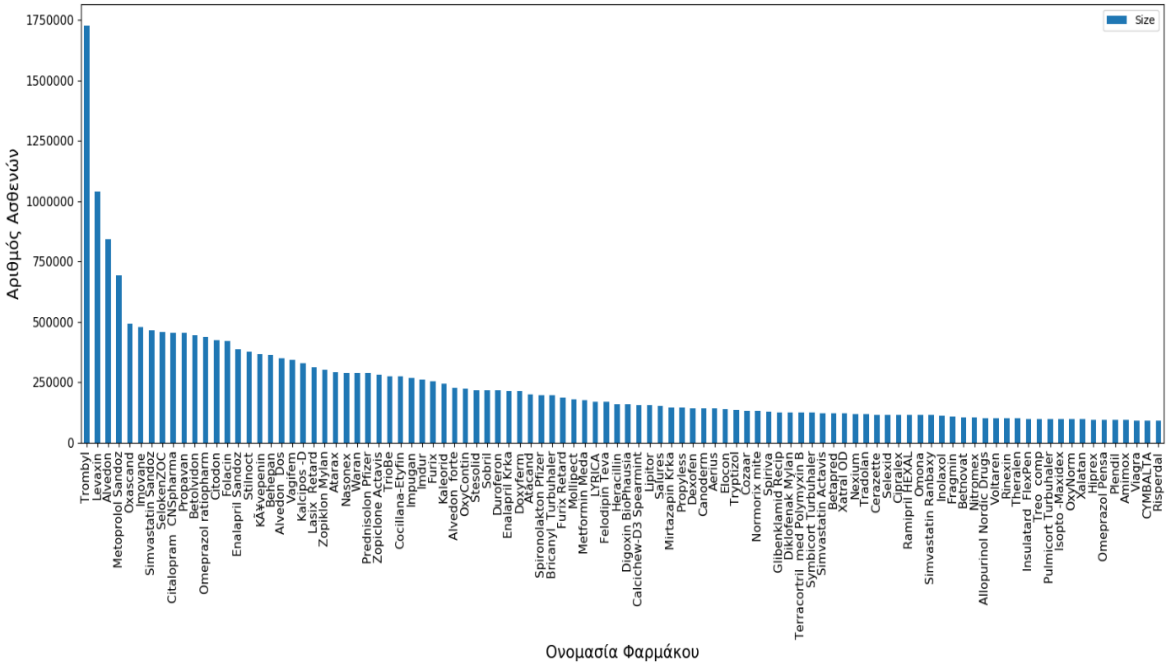
[2] ICD-10: Διεθνής Στατιστική Ταξινόμηση Νόσων και Συναφών Προβλημάτων Υγείας, 2008



Εικόνα 5 Ιστόγραμμα συχνοτήτων για το πλήθος των ασθενών που βρίσκεται σε καθεμία κατάσταση

Όπως φαίνεται στην Εικόνα 5 υπάρχουν τρεις καταστάσεις στις οποίες μπορεί να βρεθεί ένας ασθενής με νεφρική ανεπάρκεια. Η πρώτη κατάσταση είναι αυτή της αιμοκάθαρσης. Η δεύτερη κατάσταση είναι αυτή της μεταμόσχευσης. Η Τρίτη κατάσταση είναι αυτή της ανάρρωσης. Όπως φαίνεται και στην Εικόνα 5 οι περισσότεροι ασθενείς βρίσκονται στην κατάσταση της αιμοκάθαρσης. Αυτό είναι λογικό διότι αρχικά οι περισσότεροι ασθενείς ξεκινούν αιμοκάθαρση στα νεφρά. Αν υπάρχουν κατάλληλες προϋποθέσεις ο ασθενής μπορεί να κάνει μεταμόσχευση. Οι ασθενείς που είναι σε προχωρημένη ηλικία ή έχουν κάποιο καρδιακό πρόβλημα δεν ενδείκνυται να μπουν στην διαδικασία μεταμόσχευσης. Επιπρόσθετα, αν ο ασθενής έχει κάποια μορφής καρκίνο αυτόματα η μεταμόσχευση νεφρών γίνεται αρκετά ριψοκίνδυνη για την ζωή του. Τέλος αν ένας ασθενής είναι εθισμένος στο αλκοόλ ή στα ναρκωτικά τότε και σε αυτή την περίπτωση δεν ενδείκνυται η - μεταμόσχευση (BC Transplant. "Clinical guidelines for kidney transplantation." , 2018).

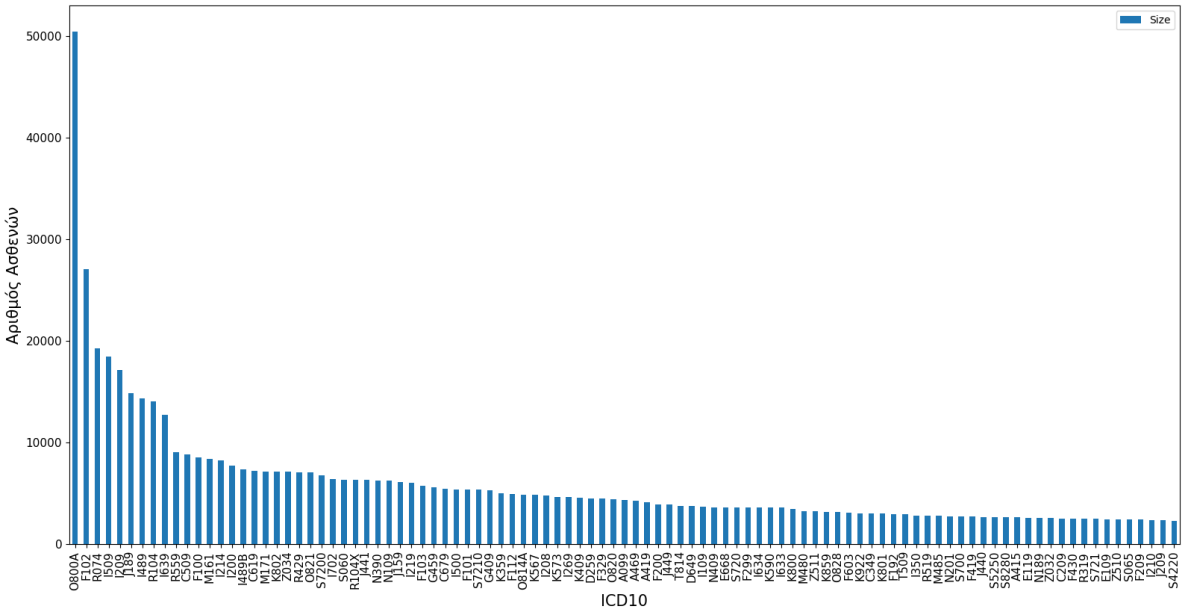
Από την επεξεργασία του συνόλου δεδομένων που περιέχει τα φάρμακα προέκυψαν τα παρακάτω αποτελέσματα. Το πλήθος των δειγμάτων που χρησιμοποιήθηκαν ήταν 46.898.746 το οποίο προέκυψε από 500.000 ασθενείς. Όπως φαίνεται στην Εικόνα 6 το πιο διαδεδομένο φάρμακο που χορηγήθηκε στους ασθενείς είναι το “Trombyl”. Το “Trombyl” είναι μια διεθνής επωνυμία χαμηλής δόσης ασπιρίνης, το οποίο είναι αποτελεσματικό στην διάλυση του αίματος και χρησιμοποιείται για την πρόληψη καρδιακών προβλημάτων και εγκεφαλικού επεισοδίου. Αυτό το αποτέλεσμα συνάδει με το συμπέρασμα που αναφέρθηκε νωρίτερα ότι η κύρια αιτία θανάτου είναι το “Οξύ έμφραγμα του μυοκαρδίου”. Το δεύτερο σε χρήση φάρμακο είναι το “Levaxin” το οποίο χρησιμοποιείται συνήθως σε μετεγχειριτικές καταστάσεις. Τέλος το τρίτο σε σειρά φάρμακο “Alvedon” χρησιμοποιείται σαν παυσίπονο για την μείωση του πόνου και την μείωση του πυρετού. Όλες οι χρήσεις των παραπάνω φαρμάκων συμπίπτουν με τα προβλήματα που προέρχονται από τη νεφρική ανεπάρκεια.



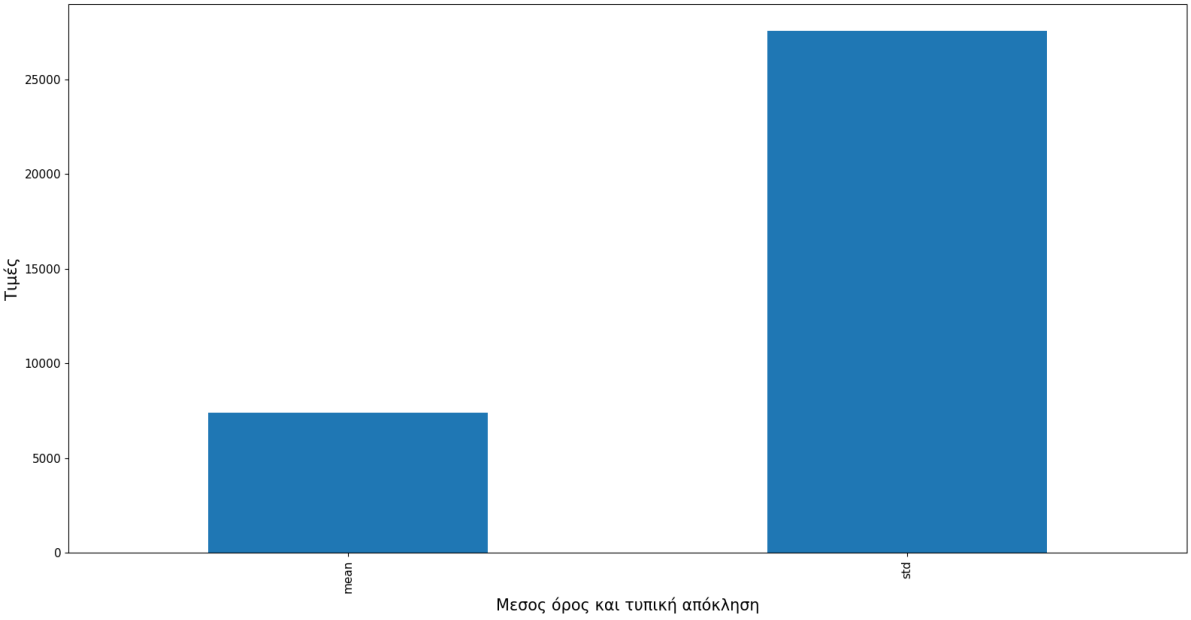
Εικόνα 6 Ιστόγραμμα συχνοτήτων με το πλήθος των ασθενών που τους χορηγήθηκε το εκάστοτε φάρμακο

Στην Εικόνα 7 φαίνεται ποια ασθένεια εμφανίζεται περισσότερο κατά την πρώτη διάγνωση. Το πλήθος των δειγμάτων που χρησιμοποιήθηκε ήταν 1.472.522 που δημιουργήθηκε από 500.000 ασθενείς. Οι χαμένες τιμές όσον αφορά την πρώτη διάγνωση ήταν 222.032 οι οποίες αφαιρέθηκαν. Έτσι το τελικό πλήθος των δειγμάτων ήταν 1.250.490. Όπως φαίνεται ο κωδικός που είναι ο επικρατέστερος είναι ο O800A, ο οποίος αντιστοιχεί σε εγκυμοσύνη (World Health Organization, 2018). Αυτό αρχικά φαίνεται περίεργο όμως αν γίνει μια πιο αναλυτική περιγραφή των συμπτωμάτων που δημιουργούν οι δύο ασθένειες θα φανεί η συσχέτιση. Οι δύο αυτές περιπτώσεις παρουσιάζουν σαν αρχικό σύμπτωμα την ναυτία. Επίσης και στις δυο περιπτώσεις υπάρχει η εύκολη κόπωση καθώς και πόνοι στο στήθος. Αυτές οι ενδείξεις σε συνδυασμό με το φύλο του ασθενή μπορούν να οδηγήσουν σε λάθος συμπεράσματα. Ο δεύτερος κωδικός που επικρατεί αντιστοιχεί στον αλκοολισμό κάτι που είναι φυσιολογικό για την συγκεκριμένη ασθένεια αν σκεφτεί κανείς ότι ο αλκοολισμός όπως και η νεφρική ανεπάρκεια επηρεάζουν κατά κύριο λόγο τα νεφρά.

Όσον αφορά το κόστος το οποίο κατά μέσο όρο χρειάζεται να πληρώσει εκάστοτε ασθενής για την αγορά φαρμάκων, αυτό κυμαίνεται γύρω 7.400 ευρώ όπως φαίνεται στην Εικόνα 8. Ο αριθμός των ασθενών που χρησιμοποιήθηκαν ήταν 500.000. Σύμφωνα με μια έρευνα το μέσο κόστος θεραπείας της αγχώδους διαταραχής, κυμαίνεται στα 5709 ευρώ για κάθε ασθενή (Martin D. Marciniak, 2005). Ακόμη, το ετήσιο κόστος των φαρμάκων για ασθενείς με διαβήτη κυμαίνεται μεταξύ 3000 και 4000 ευρώ (Chapel, 2017). Επιπλέον το ετήσιο κόστος φαρμάκων για ασθενείς με Χρόνια αποφρακτική πνευμονοπάθεια(COPD) κυμαίνεται περίπου στις 10.000 ευρώ (Chapel, 2017). Από τις παραπάνω αναφορές φαίνεται ότι η νεφρική ανεπάρκεια είναι μια δαπανηρή ασθένεια όσον αφορά το κόστος των φαρμάκων που χρειάζονται για να αντιμετωπιστεί.

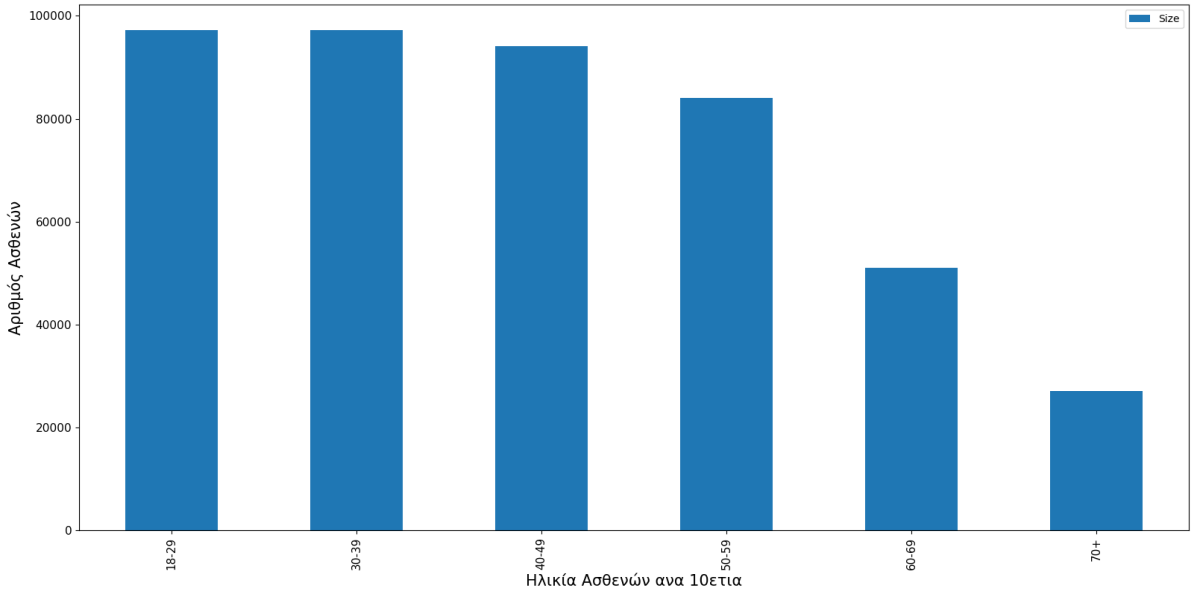


Εικόνα 7 Ιστόγραμμα συχνοτήτων με το πλήθος εμφάνισης της εκάστοτε ασθένειας κατά την πρώτη διάγνωση



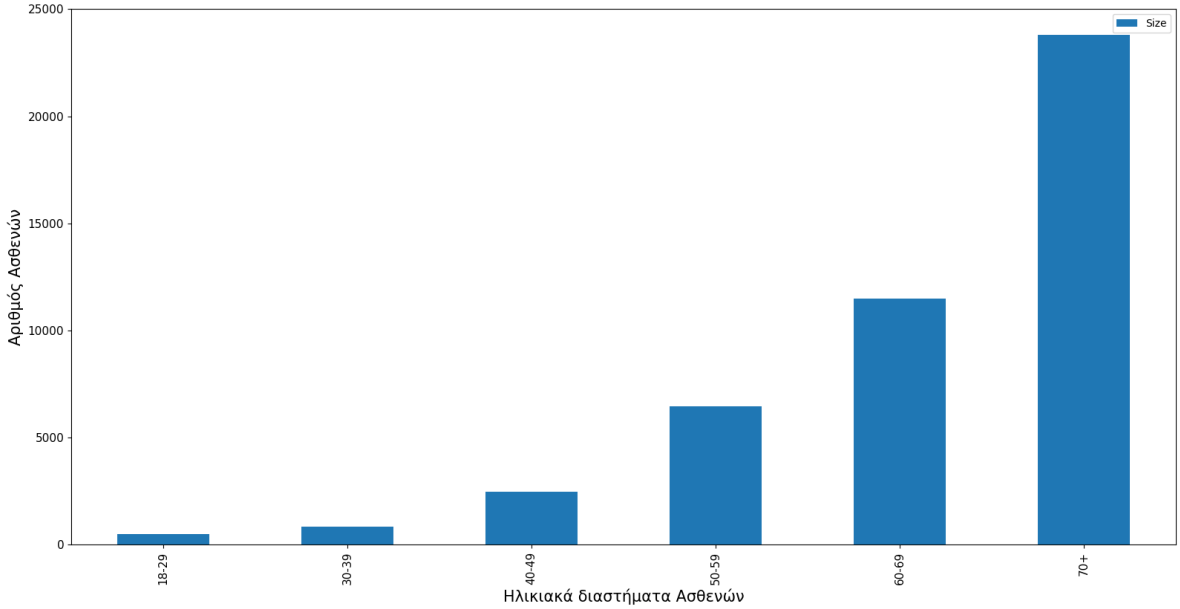
Εικόνα 8 Ιστόγραμμα συχνοτήτων με το μέσο κόστος ανά ασθενή

Κοιτώντας την ημερομηνία γέννησης των ασθενών καθώς και την ημερομηνία κατά την οποία υπολογίστηκε η κρεατινίνη για πρώτη φορά, η οποία είναι η ένδειξη ότι ο ασθενής πάσχει από νεφρική ανεπάρκεια, παρατηρήθηκε, όπως φαίνεται και στην Εικόνα 9, ότι οι περισσότεροι ασθενείς που διαπίστωσαν την ασθένεια βρίσκονταν μεταξύ των 18-29. Το πλήθος των δειγμάτων που χρησιμοποιήθηκε ήταν 500.000 το οποίο προέκυψε από 500.000 ασθενείς. Οι χαμένες τιμές όσον αφορά την ημερομηνία πρώτης μέτρησης της κρεατίνης ήταν 35.583. Οπότε το τελικό πλήθος των δειγμάτων ήταν 464.417. Όπως φαίνεται και από την Εικόνα 10 η ασθένεια διαπιστώνεται σε μικρές ηλικίες.



Εικόνα 9 Ιστόγραμμα συχνοτήτων με το πλήθος των ασθενών ανά ηλικιακό διάστημα κατά την πρώτη μέτρηση της κρεατινίνης

Παράλληλα παρατηρείται, όπως φαίνεται στην Εικόνα 10, ότι οι περισσότεροι ασθενείς πεθαίνουν πάνω από τα 70. Το πλήθος των δειγμάτων που χρησιμοποιήθηκαν ήταν 45.659 που προέκυψε από 500.000 ασθενείς. Όπως φαίνεται και από την Εικόνα 10 πολλοί λίγοι ασθενείς πεθαίνουν στις ηλικίες μεταξύ 18-29.



Εικόνα 10 Ιστόγραμμα συχνοτήτων με το μέσο όρο αριθμού θανάτων των ασθενών ηλικιακά

4.3. Αποτελέσματα Μηχανικής Μάθησης

Στην ενότητα αυτή γίνεται χρήση αλγορίθμων μηχανικής μάθησης με σκοπό την εύρεση μοντέλων με την καλύτερη απόδοση όσον αφορά την ακρίβεια για την πρόβλεψη των παρακάτω ερωτημάτων.

- Αρχικά την κατηγοριοποίηση των ασθενών που έκαναν μεταμόσχευση και πέθαναν ως προς τη χρονολογία θανάτου.
- Έπειτα την κατηγοριοποίηση των ασθενών με νεφρική ανεπάρκεια που πέθαναν ως προς την αιτία θανάτου.
- Τέλος γίνεται πρόβλεψη για το κόστος των φαρμάκων με το οποίο θα επιβαρυνθεί ο ασθενής.

Παράλληλα έγινε και Feature Ranking με τη χρήση αλγορίθμων Feature Selection σε κάθε περίπτωση λόγω των πολλών μεταβλητών που υπάρχουν στο σύνολο των δεδομένων και δυσκολεύουν τη γρήγορη απόκριση των αλγορίθμων. Επιπλέον η χρήση τεχνικών Feature Selection βοηθάει στην εύρεση των μεταβλητών οι οποίες παίζουν σημαντικότερο ρόλο στην πρόβλεψη της εκάστοτε μεταβλητής στόχου. Έτσι χρησιμοποιήθηκαν αυτές που είναι πιο αποτελεσματικές για την εκάστοτε περίπτωση σύμφωνα με τα αποτελέσματα του Feature Ranking. Το Feature Selection έγινε με την χρήση των αλγορίθμων extra tree classifier και Lasso. Πρέπει να σημειωθεί εδώ ότι σε όλες τις εφαρμογές των αλγορίθμων μηχανικής μάθησης οι κενές τιμές των εκάστοτε συνόλων δεδομένων αφαιρέθηκαν.

4.3.1. Αριθμητικά Αποτελέσματα

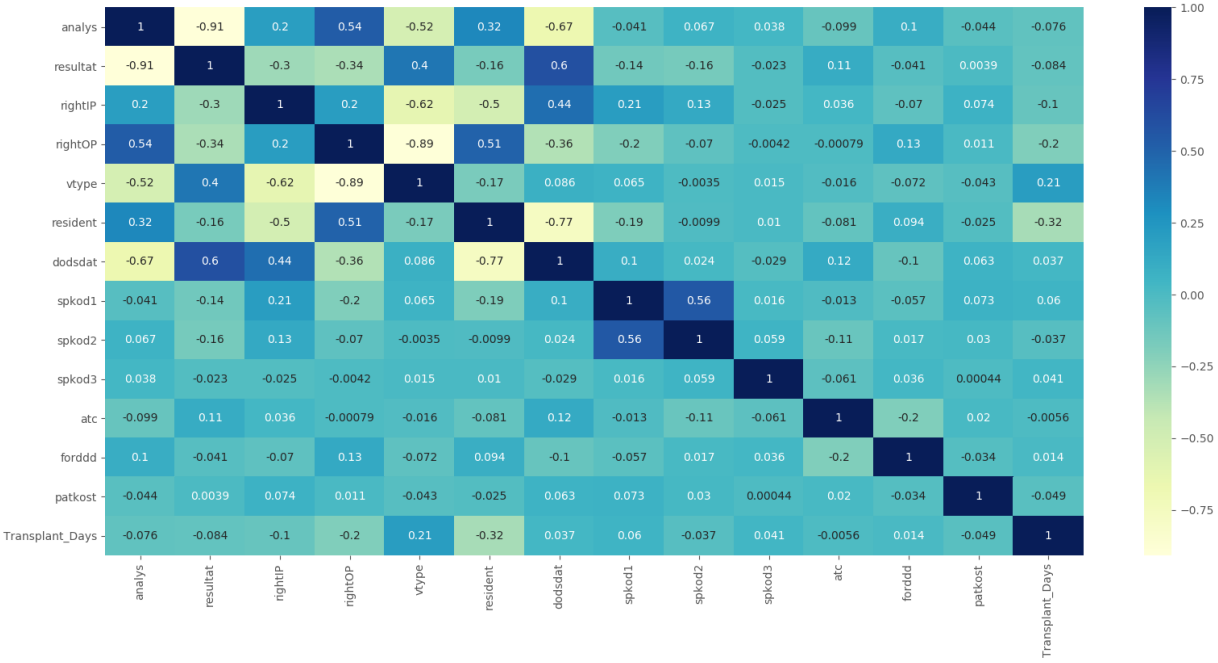
Στην ενότητα αυτή θα παρουσιαστούν αριθμητικά αποτελέσματα από τη χρήση αλγορίθμων μηχανικής μάθησης για τα προβλήματα που διατυπώθηκαν στην αρχή του κεφαλαίου. Αρχικά χρησιμοποιώντας αλγορίθμους Μηχανικής Μάθησης στα δεδομένα, έγινε κατηγοριοποίηση για τους ασθενείς που έκαναν μεταμόσχευση και πέθαναν.

4.3.1.1. Χρήση Αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη χρόνου ζωής των ασθενών από την μέρα εισαγωγής τους σε κατάσταση Μεταμόσχευσης.

Γίνεται προσπάθεια να προβλεφθεί η διάρκεια ζωής από την στιγμή που έγινε η μεταμόσχευση μέχρι το θάνατό τους. Δημιουργήθηκε μια νέα μεταβλητή “Transplant_Days”, η οποία περιέχει τα έτη ζωής των ασθενών από τη μέρα της μεταμόσχευσης μέχρι το θάνατό τους. Τα έτη ζωής που προκύπτουν είναι διακριτές τιμές και κυμαίνονται από 5 ως 40. Έτσι δημιουργήθηκαν 5 κλάσεις οι οποίες χωρίστηκαν με βάση το πλήθος των δειγμάτων. Η κάθε κλάση περιέχει περίπου ίδιο αριθμό δειγμάτων. Οι κλάσεις είναι [4-9] έτη, (9-16] έτη, (16,]. Χρησιμοποιήθηκε η μέθοδος 10-Fold Cross Validation για να αποφευχθεί το overfitting. Μετά από αφαίρεση των κενών τιμών από τις μεταβλητές που περιέχουν την ημερομηνία του θανάτου και την ημερομηνία που έγινε η μεταμόσχευση απέμειναν 2921 παρατηρήσεις οι οποίες προέκυψαν μετά από την συγχώνευση χρησιμοποιώντας την μέθοδο join, η οποία περιλαμβάνεται στη βιβλιοθήκη Pandas της Python. Πιο συγκεκριμένα χρησιμοποιήθηκε Inner Join χρησιμοποιώντας το μοναδικό αριθμό lopnr του κάθε ασθενή. Η συγκεκριμένη μέθοδος επιλέγει εγγραφές σύμφωνα με μια μεταβλητή ΄κλειδί΄ που περιέχεται κα στα δύο σύνολα δεδομένων. Οι μεταβλητές που χρησιμοποιήθηκαν στο Feature Selection και τα αποτελέσματα των οποίων φαίνονται στον πίνακα 12 με τον αλγόριθμο Extra Tree Classifier ήταν: ‘vtype’(The aggregation of rightIP and rightOP), ‘analys’(The kind of the performed laboratory test), ’atc’(The ATC code of the medication), Spkod (1,2) (The prescriber’s speciality code), Resident (The binary indicator of residency in the County of Stockholm for each measurement), ‘resultalt’(The value of the performed test), ’rightOP’(Δυαδική μεταβλητή με την απάντηση αν ήταν εξωτερικός ασθενείς η όχι), ‘rightIP’(The answer whether it was an in-hospital measurement or not),

'forddd'(The recommended daily drug dosage), 'patcost'(The total cost of the medication, covered by the citizen), και 'Transplant_Days'(Days in Transplant).
Οι ανεξάρτητες μεταβλητές που επιλέχθηκαν με βάση την τεχνική είναι: vtype (The aggregation of rightIP and rightOP), analys (The kind of the performed laboratory test), Atc (The ATC code of the medication), Spkod (1,2,3) (The prescriber's speciality code), Resident (The binary indicator of residency in the County of Stockholm for each measurement), το lan(The county of residency relative to each measurement), resultat(The value of the performed test).

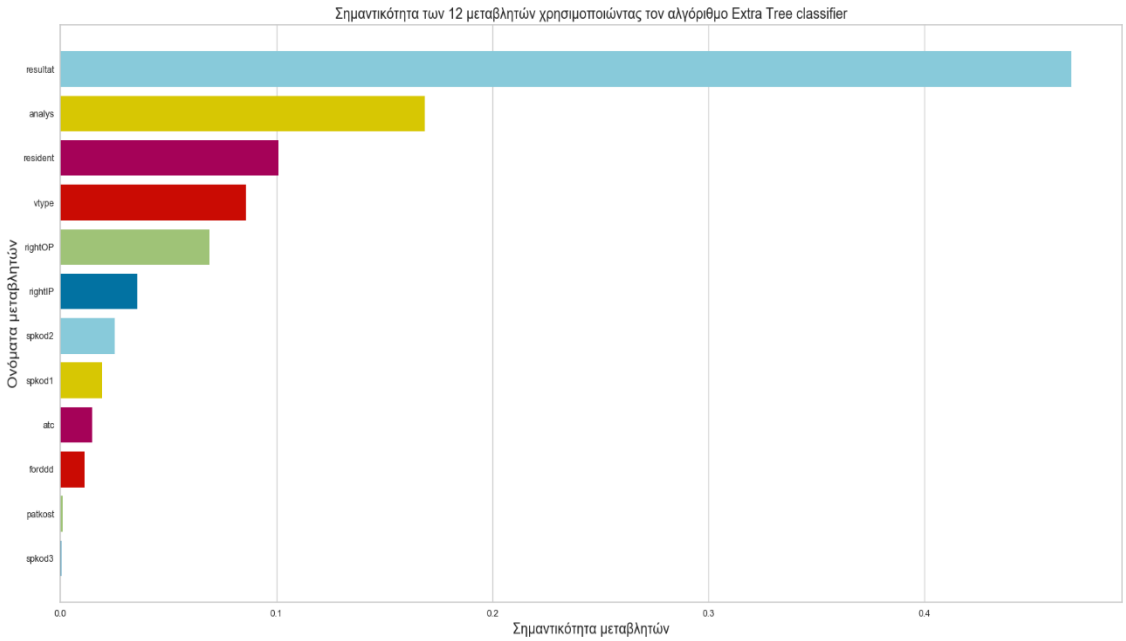
Στην Εικόνα 11 φαίνεται ένα Heat map με τη συσχέτιση των μεταβλητών χρησιμοποιώντας Spearman test. Με την χρήση του συγκεκριμένου τεστ γίνεται η εύρεση των πιο συσχετισμένων μεταβλητών σε σχέση με τη μεταβλητή που πρέπει να προβλεφθεί. Όπως φαίνεται και στην Εικόνα 11 οι πιο συσχετισμένες μεταβλητές με την μεταβλητή "Transplant_Days" που πρέπει να προβλεφθεί είναι: 'analys'(το είδος του εκτελέσιμου τεστ), 'vtype'(η ένωση των μεταβλητών rightOP kai rightIP), 'lan'(Κωδικός της περιοχής που έγινε το εκάστοτε τεστ), 'atc'(Ο κωδικός ATC κάθε φαρμάκου),'rightOP'(Δυαδική μεταβλητή με την απάντηση αν ήταν εξωτερικός ασθενείς η όχι),'resident'(Δυαδική μεταβλητή που υποδεικνύει αν ο ασθενής ήταν κάτοικος Στοκχόλμης ή όχι). Οι αριθμητικές μεταβλητές που χρησιμοποιήθηκαν σαν ανεξάρτητες μεταβλητές, επιλέχθηκαν χρησιμοποιώντας feature ranking με την χρήση του αλγορίθμου extra tree classifier. Εδώ πρέπει να σημειωθεί ότι όπως προαναφέρθηκε σε προηγούμενη ενότητα, ο αριθμός k που χρησιμοποιεί ο αλγόριθμος extra tree classifier είναι $k = \sqrt{8}$. Η εξαρτημένη μεταβλητή είναι Transplant_days (). Οι μεταβλητές που προέκυψαν από το Spearman test συμπίπτουν με τις μεταβλητές που προέκυψαν και από την χρήση του αλγορίθμου extra tree classifier για την εύρεση της σημαντικότητας των μεταβλητών που φαίνεται στον Πίνακα 12.



Εικόνα 11 Heat map με τις συσχετίσεις των μεταβλητών

Μεταβλητές	**Feature Importance(ExtraTreesClassifier)**
Vtype	0.08580675
Analys	0.16866229
Atc	0.01466746
Spkod1	0.01927921
Spkod2	0.02504293
Spkod3	0.00047595
Resident	0.10109589
Resultat	0.46798468
RightIP	0.03577102
RightOP	0.06907239
Forddd	0.01112436
Patkost	0.00101708

Πίνακας 12 Σημαντικότητα των μεταβλητών με την χρήση του αλγορίθμου extra tree classifier



Εικόνα 12 Σημαντικότητα των 12 μεταβλητών του συνόλου των δεδομένων

Στον Πίνακα 13 φαίνεται η ακρίβεια με την οποία ο κάθε αλγόριθμος κατάφερε να κάνει κατηγοριοποίηση. Στην Εικόνα 13 φαίνονται τα θηκογράμματα των αλγορίθμων.

Αλγόριθμος	**Ακρίβεια**
Decision Tree	0.793793
Random Forest	0.796185

Πίνακας 13 Ακρίβεια Αλγορίθμων

Στον Πίνακα 14 φαίνονται οι μετρικές Precision, recall, F1-score για τον αλγόριθμο Random Forest. Στον Πίνακα 15 φαίνονται οι μετρικές Precision, recall, F1-score για τον αλγόριθμο Decision Tree.

Class	**Precision**	**Recall**	**F1-Score**	**Αριθμός δειγμάτων**
[4-9]	1	0.71	0.83	992
(9-16]	0.88	0.75	0.81	1088
16+	0.63	0.97	0.76	841

Πίνακας 14 Μετρικές ακρίβειας του αλγορίθμου Random Forest

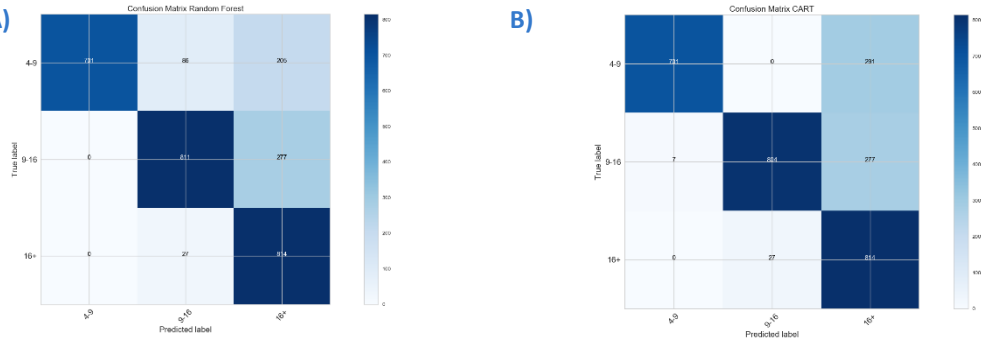
Class	**Precision**	**Recall**	**F1-Score**	**Αριθμός δειγμάτων**
[4-9]	0.99	0.71	0.82	992
(9-16]	0.97	0.74	0.84	1088
16+	0.59	0.97	0.73	841

Πίνακας 15 Μετρικές ακρίβειας του αλγορίθμου Decision Tree

A) B)

Εικόνα 13 Θηκογράμματα των αλγορίθμων: A) Random Forest, B) Decision Tree

A) B)



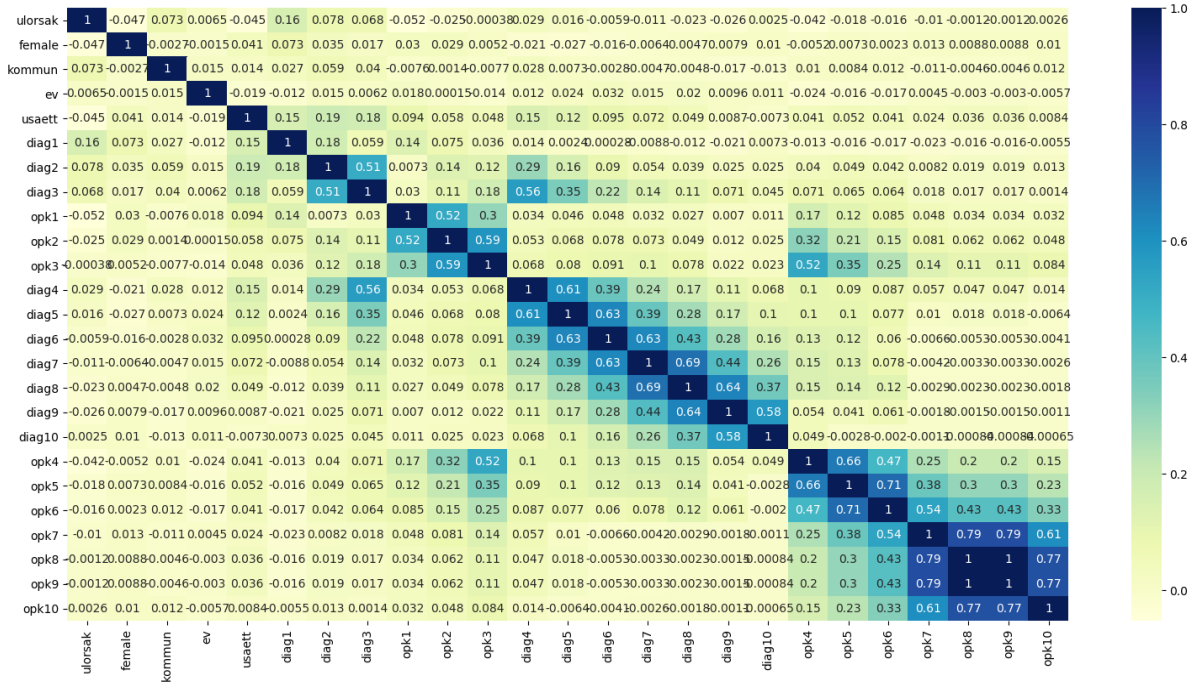
Εικόνα 13 Πίνακες σύγχυσης των αλγορίθμων κατηγοριοποίησης A) Random Forest, B) Decision Tree

Για την κατηγοριοποίηση χρησιμοποιήθηκαν τα σύνολα δεδομένων Living Status Registry (dors_large.csv), Swedish Renal Registry (SNR_large.csv), Laboratory values(lab _values.csv) καθώς και Prescribed Drugs National Registry (Lmed_large.csv). Στην Εικόνα 14 φαίνονται τα Confusion matrixes των αλγορίθμων τα οποία δείχνουν πόσες παρατηρήσεις ταξινομήθηκαν σωστά και πόσες λάθος.

4.3.1.2. Χρήση Αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη της αιτίας θανάτου των ασθενών.

Έπειτα έγινε κατηγοριοποίηση για την αιτία θανάτου όλων των ασθενών ανεξάρτητα από τη κατάσταση στην οποία βρίσκονταν. Για το συγκεκριμένο σκοπό χρησιμοποιήθηκαν μόνο δείγματα που η αιτία ανήκει σε μια από τις παρακάτω 2 κατηγορίες ICD10: “I219 η οποία αντιστοιχεί σε “οξύ έμφραγμα του μυοκαρδίου”, “C349” η οποία αντιστοιχεί σε κακοήθη νεοπλάσματα των βρόγχων και του πνεύμονα. Οι συγκεκριμένες αιτίες θανάτου είναι οι συνηθέστερες. Εδώ χρησιμοποιήθηκαν τα σύνολα δεδομένων Living Status Registry (dors_large.csv), τα δημογραφικά στοιχεία των ασθενών, καθώς και Swedish Renal Registry (SNR_large.csv). Χρησιμοποιήθηκε η μέθοδος 10 Fold-Cross Validation για να αποφευχθεί το overfitting. Σε αυτήν την περίπτωση χρησιμοποιήθηκαν 4.623 παρατηρήσεις, οι οποίες προέκυψαν μετά από την συγχώνευση χρησιμοποιώντας την μέθοδο join, η οποία περιλαμβάνεται στη βιβλιοθήκη Pandas της Python. Έγινε ένα Spearman correlation test ώστε να βρεθούν οι μεταβλητές που είναι περισσότερο συσχετισμένες μεταξύ τους όσο και με την εξαρτημένη μεταβλητή για την οποία έγινε η πρόβλεψη. Στην Εικόνα 15 φαίνεται ένας Heat map με το κατά πόσο συσχετισμένες είναι οι μεταβλητές. Στην συγκεκριμένη περίπτωση έγινε Feature Selection χρησιμοποιώντας τον αλγόριθμο Lasso. Οι μεταβλητές που χρησιμοποιήθηκαν για να γίνει Feature Selection ήταν: ‘kommun’(περιοχή πρώτης μέτρησης της κρεατίνης), ‘diag1-7’(κωδικός

διάγνωσης), 'opk1-6'(Κωδικός Nomesco χειρουργικής επέμβασης) καθώς και Ulorsak (η αιτία θανάτου). Στην Εικόνα 16 και στον Πίνακα 21 φαίνονται τα αποτελέσματα του αλγορίθμου Lasso για την σημαντικότητα των μεταβλητών. Οπότε οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν μετά το Feature Selection είναι: Opk5 (The NOMESCO surgical procedure codes), Kommun (The municipality of residency at date of first creatinine measurement), diag1(The ICD-10 diagnostic codes), diag7(The ICD-10 diagnostic codes), Opk4 (The NOMESCO surgical procedure codes), Opk2 (The NOMESCO surgical procedure codes). Η εξαρτημένη μεταβλητή είναι Ulorsak (The cause that led to the specific citizen's death). Στη συγκεκριμένη περίπτωση χρησιμοποιήθηκε μια άλλη τεχνική για feature selection, η τεχνική Lasso regularization (Jain, 2016).

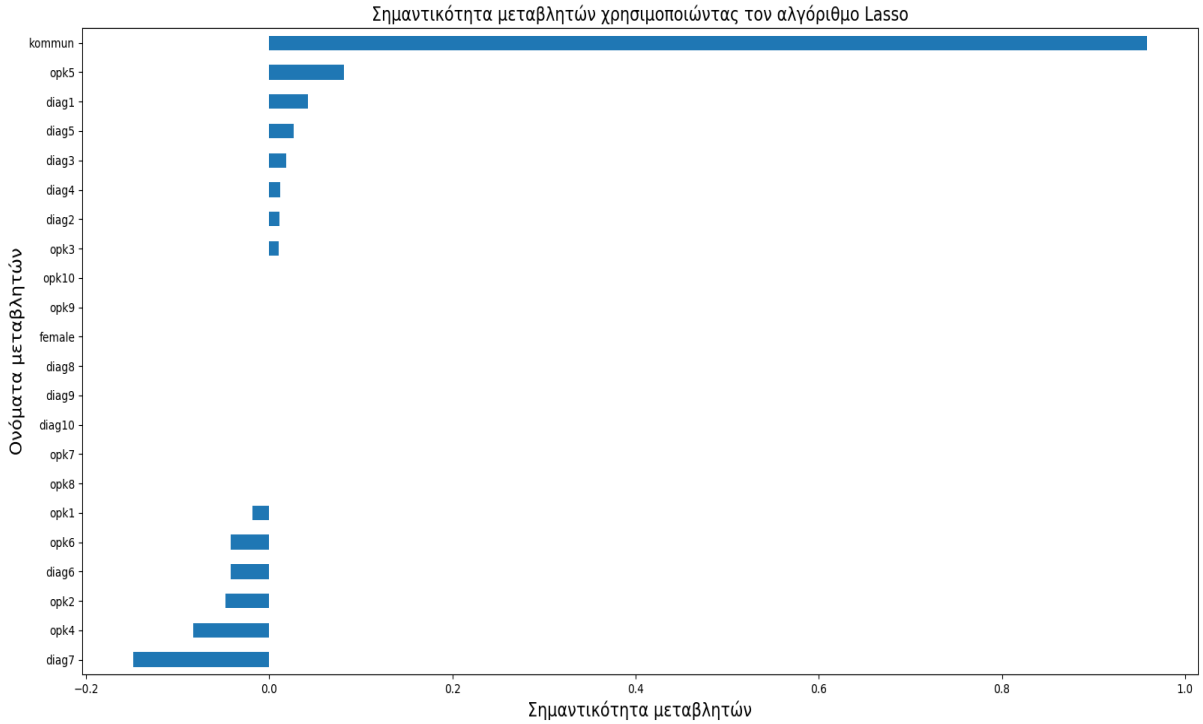


Εικόνα 14 Heat map με τις συσχετίσεις των μεταβλητών

Μεταβλητές	**Feature Importance(Lasso)**
Kommun	0.958612
Diag1	0.042272
Diag2	0.010979
Diag3	0.018467
Diag4	0.011932
Diag5	0.026696
Diag6	-0.042045

Diag7	-0.148814
Opk1	-0.018308
Opk2	-0.047932
Opk3	0.010144
Opk4	-0.083404
Opk5	0.081480
Opk6	-0.041983

Πίνακας 16 Σημαντικότητα των μεταβλητών με την χρήση του αλγορίθμου Lasso



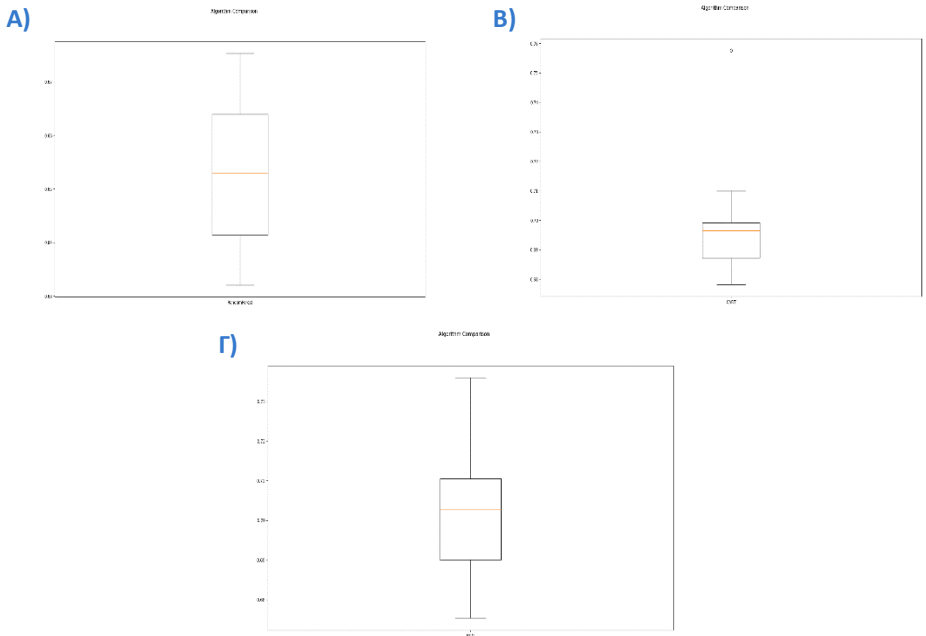
Εικόνα 15 Η σημαντικότητα των μεταβλητών με την χρήση του αλγορίθμου Lasso

Στον Πίνακα 17 φαίνεται η ακρίβεια πρόβλεψης 4 αλγορίθμων:

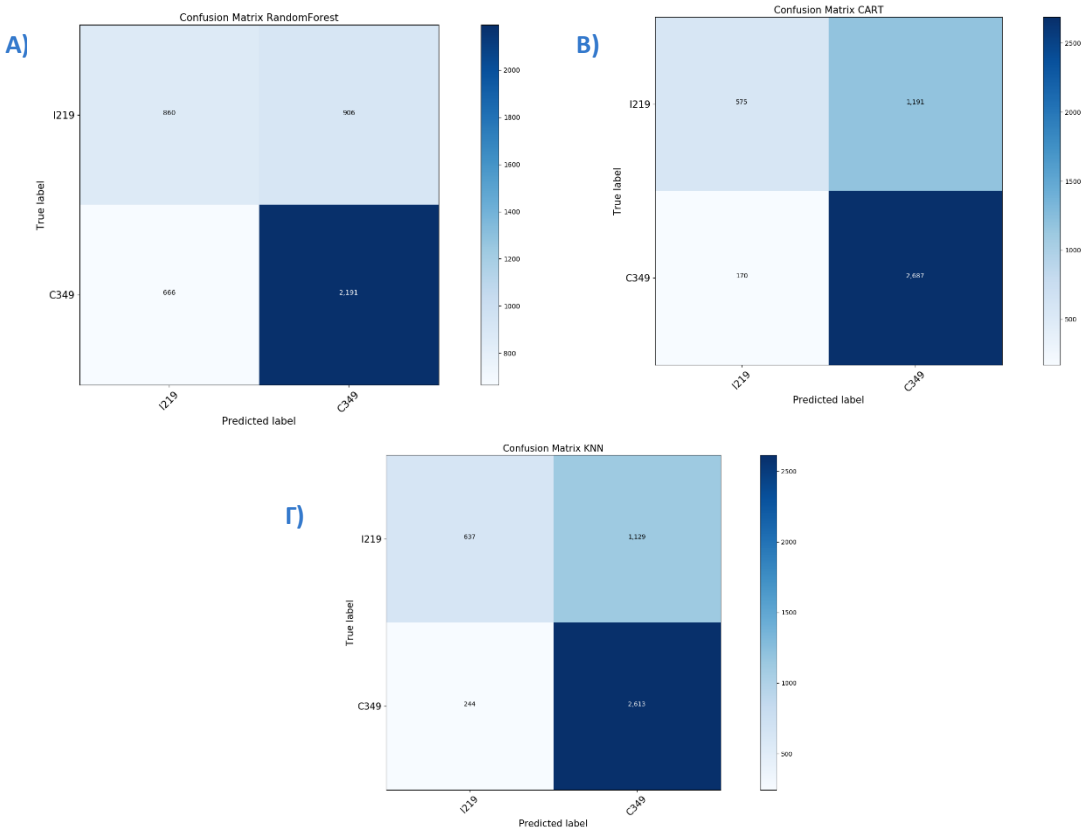
Αλγόριθμος	**Ακρίβεια**
KNN: K Nearest Neighbours	0.703013
Decision Tree	0.699552
Random Forest	0.652828

Πίνακας 17 Ακρίβεια Αλγορίθμων

Όπως φαίνεται και στον Πίνακα 17 την καλύτερη ακρίβεια την πετυχαίνει ο αλγόριθμος KNN.



Εικόνα 16 Θηκογράμματα των αλγορίθμων: A) Random Forest, B) Decision Tree, C) KNN



Εικόνα 17 Πίνακες Σύγχυσης: A) Random Forest, B) Decision Tree, Γ) KNN

Στον Πίνακα 18, στον Πίνακα 19 και στον Πίνακα 20 φαίνονται κάποιες μετρικές για την ακρίβεια των αλγορίθμων. Όπως φαίνεται και από τους Πίνακες οι αλγόριθμοι έχουν καλύτερα αποτελέσματα όσον αφορά την πρόβλεψη της κατηγορίας ‘I219’.

Class	Precision	Recall	F1-Score	Αριθμός δειγμάτων
I219	0.71	0.77	0.74	2857
C349	0.56	0.49	0.52	1766

Πίνακας 18 Μετρικές ακρίβειας του αλγορίθμου Random Forest

Class	Precision	Recall	F1-Score	Αριθμός δειγμάτων
I219	0.69	0.94	0.80	2857
C349	0.77	0.33	0.46	1766

Πίνακας 19 Μετρικές ακρίβειας του αλγορίθμου Decision Tree

Class	Precision	Recall	F1-Score	Αριθμός δειγμάτων
I219	0.70	0.91	0.79	2857
C349	0.72	0.36	0.48	1766

Πίνακας 20 Μετρικές ακρίβειας του αλγορίθμου KNN

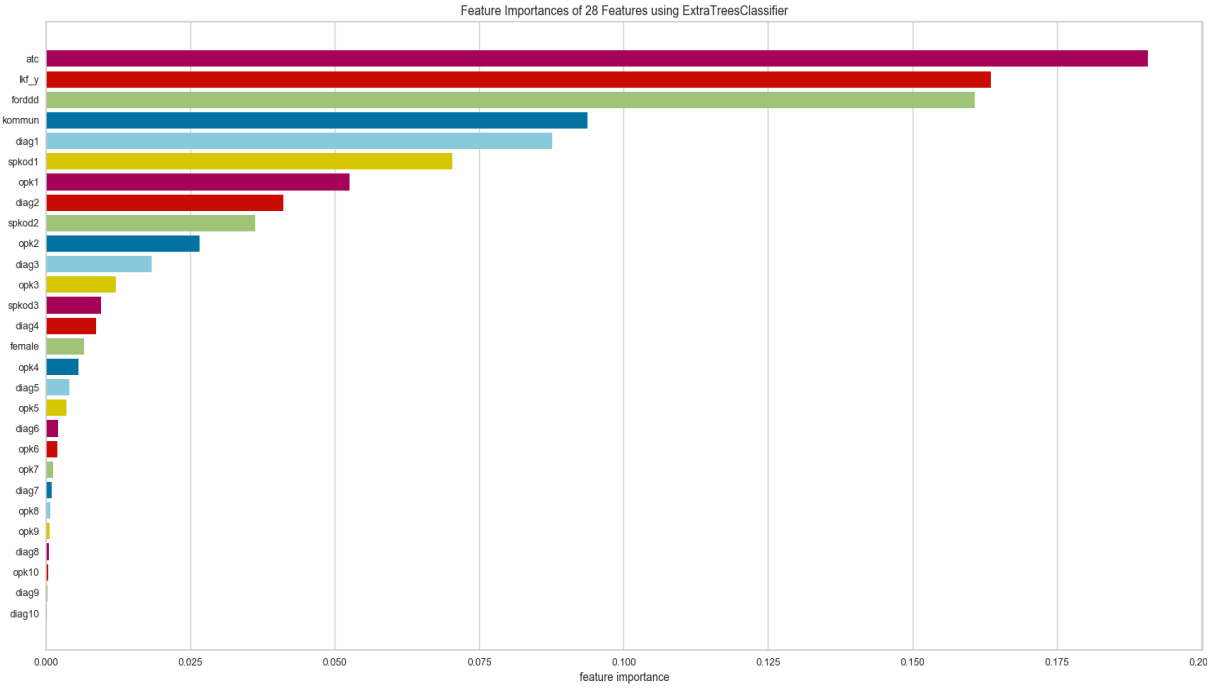
4.3.1.3. Χρήση Αλγορίθμων Μηχανικής Μάθησης για την πρόβλεψη του κόστους της θεραπείας των ασθενών

Παράλληλα, έγινε πρόβλεψη για το κόστος των φαρμάκων, με το οποίο θα επιβαρυνθεί ο ασθενής. Ο αριθμός των παρατηρήσεων που χρησιμοποιήθηκαν ήταν 365.409 παρατηρήσεις οι οποίες προέκυψαν μετά από την συγχώνευση χρησιμοποιώντας την μέθοδο join, η οποία περιλαμβάνεται στη βιβλιοθήκη Pandas της Python. Πιο συγκεκριμένα χρησιμοποιήθηκε Inner Join χρησιμοποιώντας το μοναδικό αριθμό lopnr του κάθε ασθενή. Η συγκεκριμένη μέθοδος επιλέγει εγγραφές σύμφωνα με μια μεταβλητή ΄κλειδί΄ που περιέχεται κα στα δύο σύνολα δεδομένων. Έπειτα έγινε διαγραφή των διπλότυπων τιμών. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν ήταν το demo_large το οποίο περιέχει δημογραφικά στοιχεία των ασθενών, το Slv_all.csv που περιέχει στοιχεία για την εισαγωγή στο νοσοκομείο καθώς και το lmed_large που περιέχει τα φάρμακα που χορηγήθηκαν σε κάθε ασθενή. Έγινε επιλογή κατάλληλων μεταβλητών χρησιμοποιώντας Feature Selection και συγκεκριμένα τον χρησιμοποιώντας Extra tree Classifier. Οι μεταβλητές που χρησιμοποιήθηκαν για να γίνει Feature Selection είναι: ‘atc’ (The ATC code of the medication), ‘lkf’(The county/municipality/assembly that the Creatinine was first calculated), ‘forddd’(The recommended daily drug dosage), kommun(The municipality of residency at date of first creatinine measurement), ‘diag1-10’(The ICD-10 diagnostic codes), ‘spkod1-3’(The prescriber’s speciality code), ‘opk1-10’(The NOMESCO surgical procedure codes), ‘female’(The gender of the citizen). Τα αποτελέσματα φαίνονται στον Πίνακα 21.

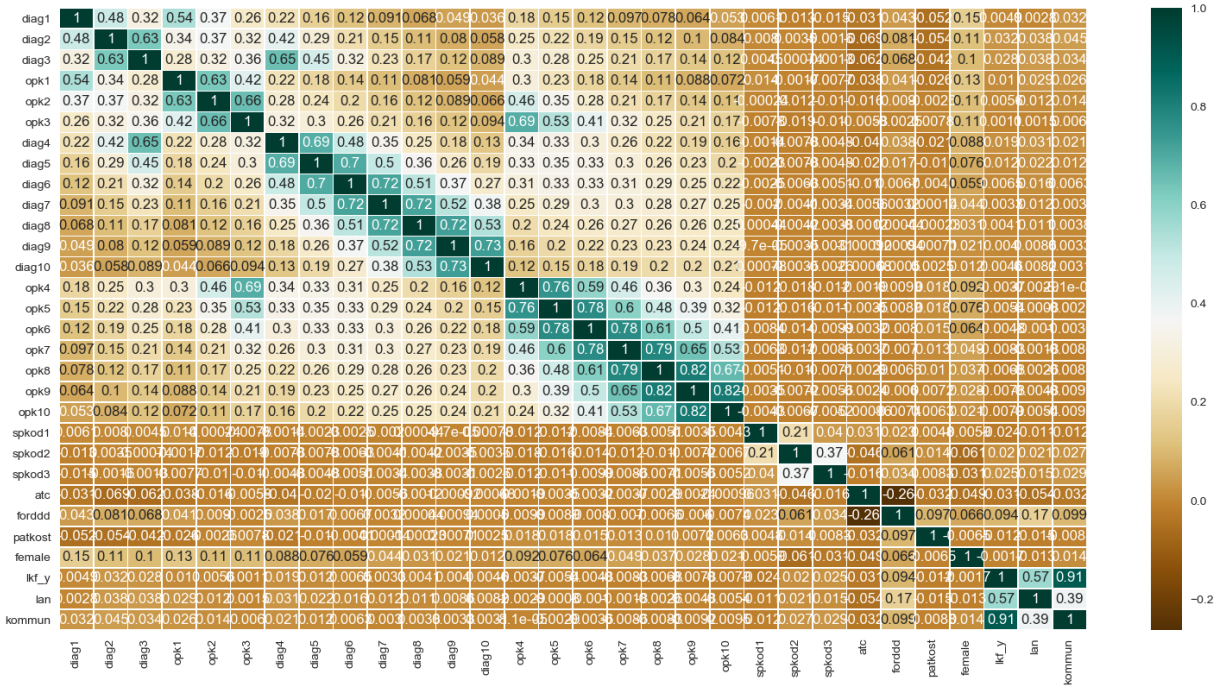
Μεταβλητές	Feature Importance(Lasso)
atc	0.190715654
lkf	0.163559384
forddd	0.160779037
Kommun	0.0937637970
Diag1	0.0876272385
Spkod1	0.0703743075
Opk1	0.0525809129
Diag2	0.0410221012

Spkod2	0.0362216060
Opk2	0.0265556874
Diag3	0.0183159992
Opk3	0.0120158422
Spkod3	0.00946342463
Diag4	0.00863765292
Female	0.00661275070
Opk4	0.00558723487
Diag5	0.00400876136
Opk5	0.00352435286
Diag6	0.00201070203
Opk6	0.00194497506
Opk7	0.00120352820
Diag7	0.000923408415
Opk8	0.000761661011
Opk9	0.000555455909
Diag8	0.000474238624
Opk10	0.000409065343
Diag9	0.000237594676
Diag10	0.000113626224

Πίνακας 21 Σημαντικότητα μεταβλητών με την χρήση του αλγορίθμου Extra Tree Classifier



Εικόνα 18 Σημαντικότητα μεταβλητών με την χρήση του αλγορίθμου Extra Tree Classifier



Εικόνα 19 Heat map με τις συσχετίσεις των μεταβλητών

Οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν μετά την χρήση Feature Selection ήταν: ο κωδικός του εκάστοτε φαρμάκου (atc: The ATC code of the medication), η ημερήσια δόση του εκάστοτε φαρμάκου (forddd: The recommended daily drug dosage),η περιοχή στην οποία μετρήθηκε πρώτη φορά η κρεατινίνη (lkf: The county/municipality/assembly that the Creatinine was first calculated), ο δήμος κατοικίας κατά την ημερομηνία της πρώτης μέτρησης κρεατινίνης (kommun: The

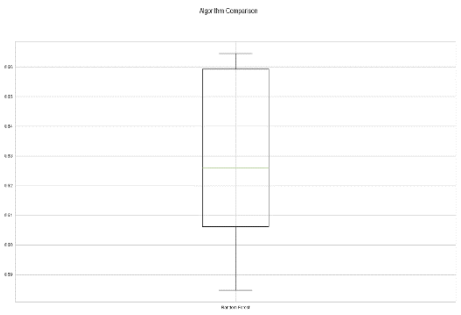
municipality of residency at date of first creatinine measurement, Spkod1 (The prescriber's speciality code), Spkod2(The prescriber's speciality code), Diag1(The ICD-10 diagnostic codes), Diag2(The ICD-10 diagnostic codes), Diag3(The ICD-10 diagnostic codes),Opk1(The NOMESCO surgical procedure codes), Opk2(The NOMESCO surgical procedure codes), Χρησιμοποιήθηκε ο αλγόριθμος Extra Tree Classifier για την εύρεση των καταλληλότερων μεταβλητών.

Χρησιμοποιήθηκαν αλγόριθμοι παλινδρόμησης, η ακρίβεια των οποίων φαίνεται στον Πίνακα 22. Όπως φαίνεται και στον Πίνακα 22 την καλύτερη ακρίβεια την πέτυχε ο αλγόριθμος Decision Tree.

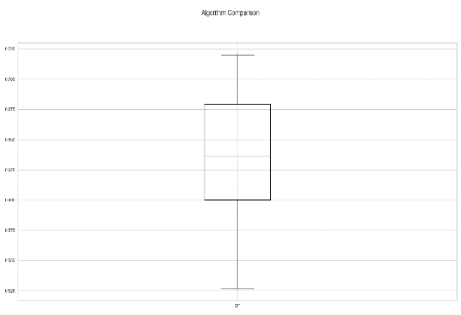
Αλγόριθμος	R^2
Decision Tree Regressor	0.638145
XGB Regressor	0.634015
Random Forest Regressor	0.629102

Πίνακας 22 Ακρίβεια Αλγορίθμων

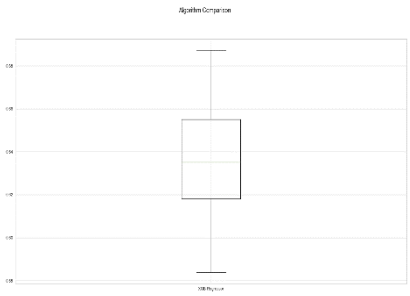
Α)



Β)



Γ)



Εικόνα 20 Θηκογράμματα των αλγορίθμων: Α) Random Forest, Β) Decision Tree, Γ) XGBoost

Στην συγκεκριμένη περίπτωση το μέτρο ακρίβειας που χρησιμοποιήθηκε είναι το R^2. Το R^2 είναι ένα στατιστικό μέτρο που αντιπροσωπεύει το ποσοστό της διακύμανσης για μια εξαρτημένη μεταβλητή που εξηγείται από μια ή περισσότερες ανεξάρτητες μεταβλητές σε ένα μοντέλο παλινδρόμησης. Παρέχει ένα μέτρο για το πόσο καλά αποτελέσματα αναπαράγονται από το μοντέλο, με βάση το ποσοστό της συνολικής διακύμανσης των αποτελεσμάτων που εξηγείται από το μοντέλο. Δίνεται από τον τύπο $R^2 = \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$, όπου το $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, y_i είναι η i τιμή της εξαρτημένης μεταβλητής και f_i

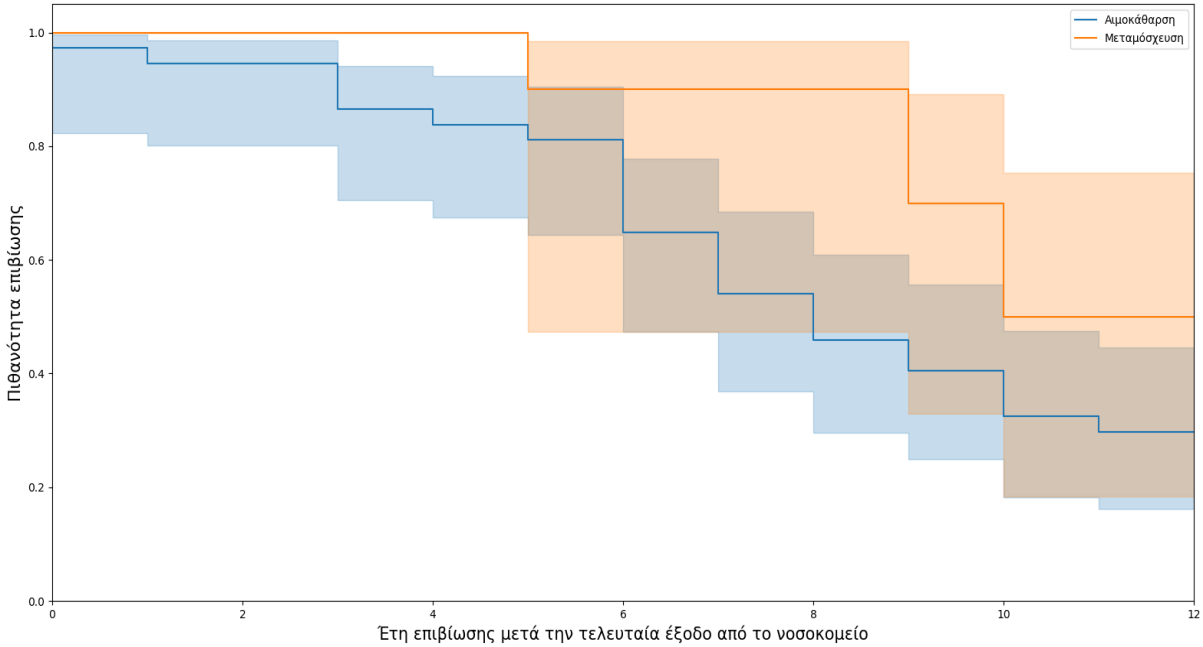
είναι η προβλεπόμενη i τιμή του μοντέλου. Οι τιμές του R^2 κυμαίνονται στο διάστημα [0,1]. Όσο μεγαλύτερη είναι η τιμή του R^2 τόσο καλύτερο είναι το μοντέλο.

4.4. Αποτελέσματα Causal Analysis και Ανάλυσης Επιβίωσης

Στο κεφάλαιο αυτό μελετώνται τεχνικές για την εκτίμηση του χρόνου ζωής των ασθενών και συγκεκριμένα ο εκτιμητής Kaplan-Meier (Andersen, 2014). Η χρήση του εκτιμητή Kaplan-Meier έγινε διότι μπορεί να υπολογίσει το ποσοστό επιβίωσης ακόμα και αν ο ασθενής φύγει ή πεθάνει κατά την διάρκεια του πειράματος, κάτι που συμβαίνει στο σύνολο των δεδομένων που χρησιμοποιείται. Αυτό δίνει την δυνατότητα ακριβέστερων αποτελεσμάτων ακόμη και με censored Data. Θα παρουσιαστούν συμπεράσματα της χρήσης του συγκεκριμένου εκτιμητή στο σύνολο δεδομένων. Ακόμη θα μελετηθούν και εκτιμητές που χρησιμοποιούνται για να υπολογιστεί πόσο επηρεάζεται το αποτέλεσμα μιας θεραπείας από τη δοσολογία ενός φαρμάκου με τη χρήση της Αιτιολογικής Ανάλυσης. Οι εκτιμητές που χρησιμοποιήθηκαν στην παρούσα εργασία είναι ο Ordinary Least Squares (OLS) (Rahman, 2018), ο εκτιμητής Matching (Becker, 2002), και ο εκτιμητής Weighting (Matschinger, 2020). Θα παρουσιαστούν συμπεράσματα της χρήσης των εκτιμητών αυτών στο σύνολο δεδομένων.

4.4.1. Αριθμητικά Αποτελέσματα της Ανάλυσης Επιβίωσης

Οι ασθενείς που χρησιμοποιήθηκαν ήταν αυτοί που απεβίωσαν σε διάστημα 12 χρόνων από την τελευταία έξοδο από το νοσοκομείο. Οι ασθενείς αυτοί χρησιμοποιήθηκαν διότι σύμφωνα με (Neild, 2017) ο μέσος όρος ζωή των ασθενών με χρόνια νεφρική ανεπάρκεια είναι περίπου 12 χρόνια. Το εύρος των ηλικιών που χρησιμοποιήθηκαν είναι μεταξύ 50 και 80 χρονών. Στην Εικόνα 22 φαίνεται η πιθανότητα επιβίωσης των ασθενών που είναι σε κατάσταση μεταμόσχευσης και σε κατάσταση αιμοκάθαρσης μετά την έξοδο από το νοσοκομείο και σε διάστημα 12 χρόνων. Εδώ πρέπει να σημειωθεί ότι κρατήθηκε μόνο η τελευταία φορά που βγήκε από το νοσοκομείο ο ασθενής.



Εικόνα 21 Απεικόνιση της συνάρτησης επιβίωσης (survival function) για χρονική περίοδο 12 χρόνων αναφορικά με τους ασθενείς σε αιμοκάθαρση (μπλε χρώμα) και τους ασθενείς με μεταμόσχευση (πορτοκαλί χρώμα).

Ο Αριθμός των ασθενών που χρησιμοποιήθηκαν ήταν 47. Από αυτούς οι 37 ήταν σε κατάσταση αιμοκάθαρσης και οι 10 σε κατάσταση μεταμόσχευσης. Όπως φαίνεται από την Εικόνα 22 το ποσοστό επιβίωσης των ασθενών σε μεταμόσχευση είναι μεγαλύτερο από αυτό των ατόμων σε αιμοκάθαρση κατά την διάρκεια των 12 χρόνων.

Στον Πίνακα 23 και στον Πίνακα 24 φαίνεται το ποσοστό επιβίωσης των ασθενών σε μεταμόσχευση και αιμοκάθαρση αντίστοιχα.

Χρονοδιάγραμμα(Έτη)	Ποσοστό Επιβίωσης
0	1
5	0.9
9	0.7
10	0.5
12	0.5

Πίνακας 23 Ποσοστό επιζώντων σε μεταμόσχευση ανά έτος μετά την τελευταία έξοδο από το νοσοκομείο

Χρονοδιάγραμμα(Έτη)	Ποσοστό Επιβίωσης
0	0.972973
1	0.945946
3	0.864865
4	0.837838
5	0.810811
6	0.648649
7	0.540541
8	0.459459
9	0.405405
10	0.324324
11	0.297297
12	0.297297

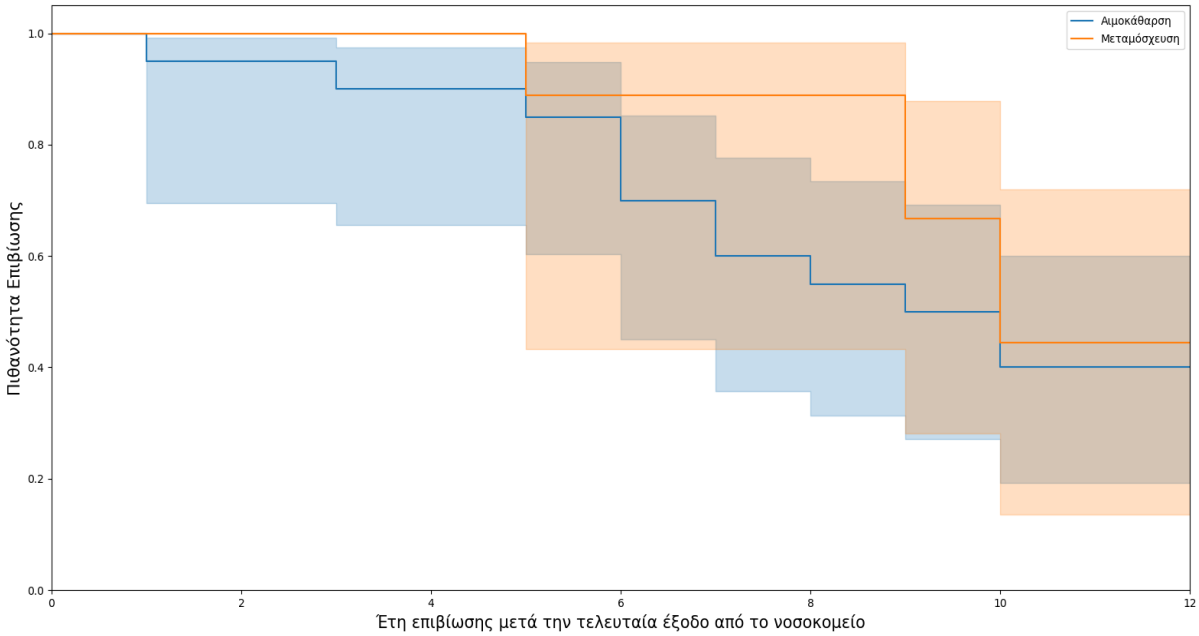
Πίνακας 24 Ποσοστό επιζώντων σε αιμοκάθαρση ανά έτος μετά την τελευταία έξοδο από το νοσοκομείο

Χρονοδιάγραμμα(Έτη)	Ποσοστό Επιβίωσης
0	1
5	0.888889
9	0.666667
10	0.444444
12	0.444444

Πίνακας 25 Ποσοστό επιζώντων σε μεταμόσχευση ανά έτος για ασθενείς με ηλικία μικρότερη των 65 χρόνων μετά την έξοδο από το νοσοκομείο σε διάστημα 12 χρόνων

Χρονοδιάγραμμα(Έτη)	Ποσοστό Επιβίωσης
0	1
1	0.95
3	0.90
5	0.85
6	0.70
7	0.60
8	0.55
9	0.50
10	0.40
12	0.40

Πίνακας 26 Ποσοστό επιζώντων σε αιμοκάθαρση ανά έτος για ασθενείς με ηλικία μικρότερη των 65 χρόνων μετά την έξοδο από το νοσοκομείο σε διάστημα 12 χρόνων



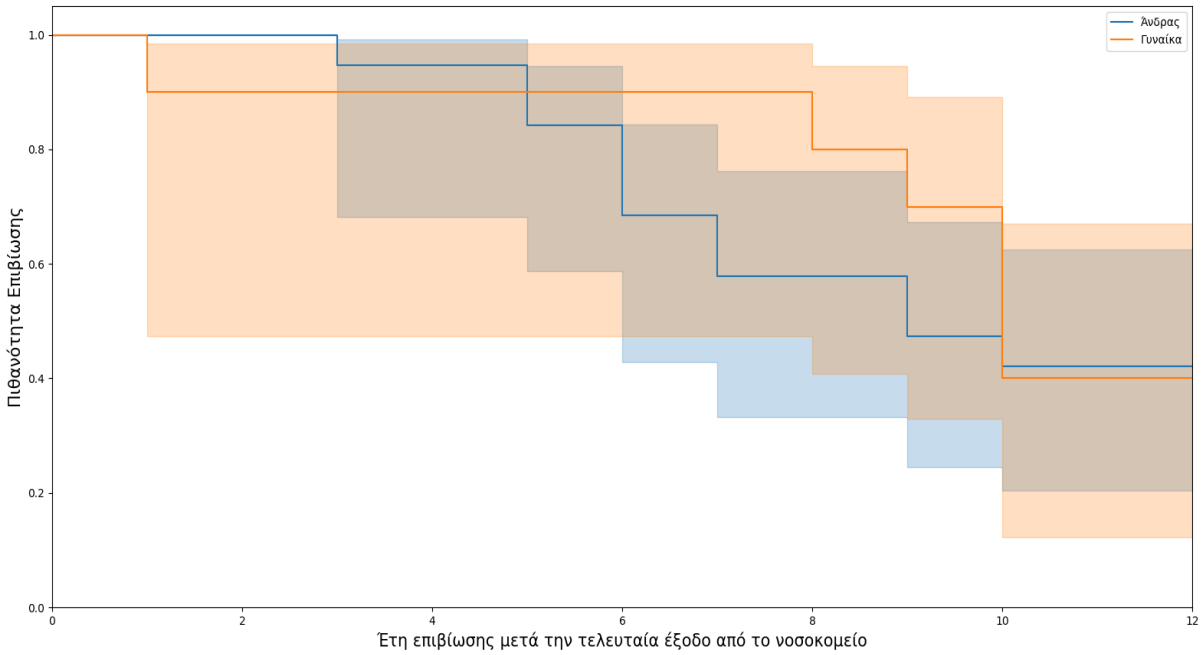
Εικόνα 22 Απεικόνιση της συνάρτησης επιβίωσης (survival function) για χρονική περίοδο 12 χρόνων αναφορικά με τους ασθενείς με ηλικία μικρότερη των 65 ετών σε αιμοκάθαρση (μπλε χρώμα) και τους ασθενείς με μεταμόσχευση (πορτοκαλί χρώμα).

Στην Εικόνα 23 φαίνεται πως κυμαίνεται ο χρόνος ζωής για τους ασθενείς με ηλικία μικρότερη των 65 μετά από την τελευταία έξοδο από το νοσοκομείο και σε διάστημα 12 ετών. Σε αυτήν τη περίπτωση διατηρήθηκε μόνο η τελευταία φορά που ο ασθενής βγήκε από το νοσοκομείο. Εδώ πρέπει να σημειωθεί ότι ο αριθμός των ασθενών που χρησιμοποιήθηκαν ήταν 29, από τους οποίους οι 6 ήταν σε μεταμόσχευση και οι 20 σε αιμοκάθαρση. Πρέπει να σημειωθεί εδώ ότι οι κενές τιμές από τις μεταβλητές που περιέχουν την ημερομηνία εξόδου από το νοσοκομείο και την ημερομηνία θανάτου διαγράφηκαν.

Όπως φαίνεται και στην Εικόνα 23, το ποσοστό των επιζώντων ασθενών σε μεταμόσχευση είναι μεγαλύτερο από το ποσοστό επιβίωσης σε αιμοκάθαρση.

Στον Πίνακα 25 και στον Πίνακα 26 φαίνεται το ποσοστό επιζώντων ανά έτος για ασθενείς μικρότερους των 65 ετών σε μεταμόσχευση και αιμοκάθαρση σε διάστημα 12 χρόνων. Το ποσοστό των επιζώντων για τα άτομα ηλικίας μικρότερη των 65 ετών σε αιμοκάθαρση είναι μεγαλύτερο από το ποσοστό επιβίωσης του συνολικού πλήθους των ασθενών.

Ενδιαφέρον έχει και πως συμπεριφέρονται οι ασθενείς ανάλογα με το φύλο τους. Στην Εικόνα 24 φαίνεται αν οι ασθενείς ανάλογα με το φύλο και ηλικία μικρότερη των 65 ετών ζουν περισσότερο από 12 χρόνια από την έξοδο από το νοσοκομείο ανεξαρτήτως κατάστασης. Πρέπει να σημειωθεί ότι και σε αυτήν την περίπτωση κρατήθηκε μόνο η τελευταία φορά που βγήκε ο ασθενής από το νοσοκομείο. Εδώ πρέπει να σημειωθεί ότι από τις 29 παρατηρήσεις, οι 10 είναι γυναίκες και 19 είναι άντρες. Πρέπει να σημειωθεί εδώ ότι οι κενές τιμές από τις μεταβλητές που περιέχουν την ημερομηνία εξόδου από το νοσοκομείο και την ημερομηνία θανάτου διαγράφηκαν. Όπως φαίνεται και από την Εικόνα 24 το ποσοστό επιβίωσης των γυναικών και των ανδρών είναι σχεδόν ίδιο. Στον Πίνακα 27 και στον Πίνακα 28 φαίνεται το ποσοστό επιβίωσης γυναικών και αντρών αντίστοιχα. Από τους 2 πίνακες παρατηρείται ότι γυναίκες και άντρες έχουν το ίδιο ποσοστό επιβίωσης μετά τα 12 χρόνια.



Εικόνα 23 Απεικόνιση της συνάρτησης επιβίωσης (survival function) για χρονική περίοδο 12 χρόνων αναφορικά με τους άντρες ασθενείς (μπλε χρώμα) και τις γυναίκες ασθενείς (πορτοκαλί χρώμα) για ηλικία μικρότερη των 65 χρονών

Χρονοδιάγραμμα(Έτη)	Ποσοστό Επιβίωσης
0	1.000000
3	0.947368
5	0.842105
6	0.684211
7	0.578947
9	0.473684
10	0.421053
12	0.421053

Πίνακας 27 Ποσοστό επιζώντων Ανδρών ανά έτος για ασθενείς με ηλικία μικρότερη των 65 χρόνων μετά την έξοδο από το νοσοκομείο σε διάστημα 12 ετών

Χρονοδιάγραμμα(Έτη)	Ποσοστό Επιβίωσης
0	1.0
1	0.9
8	0.8
9	0.7
10	0.4
12	0.4

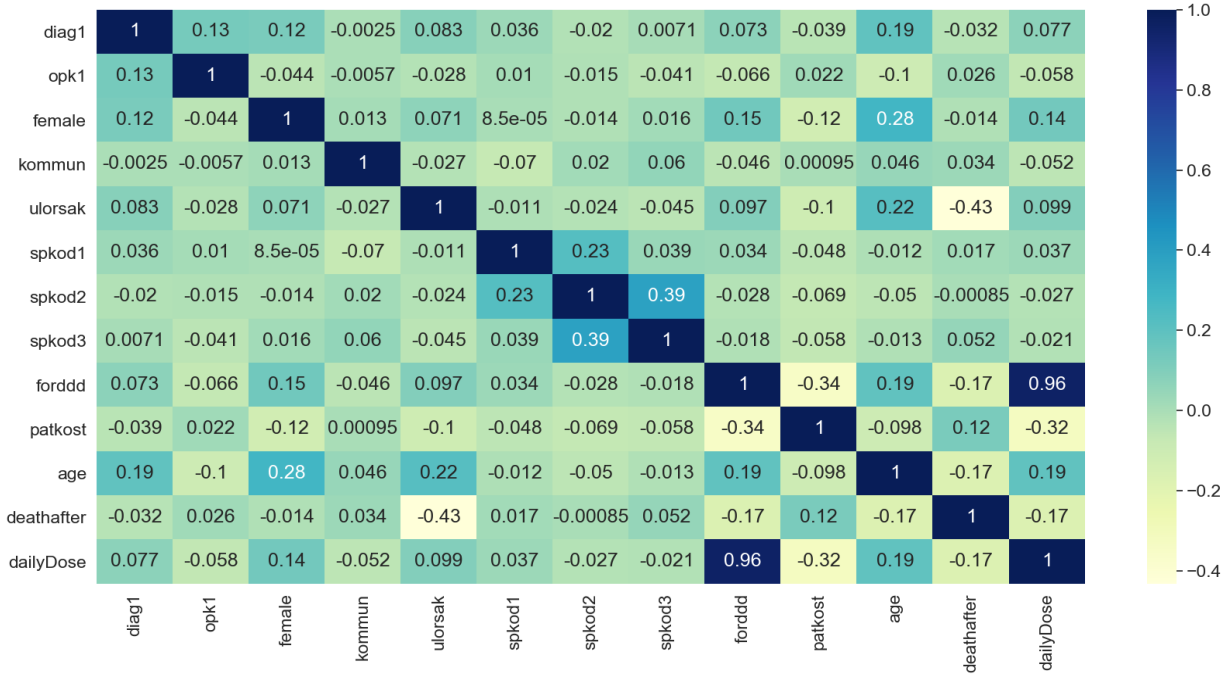
Πίνακας 28 Ποσοστό επιζώντων γυναικών ανά έτος για ασθενείς με ηλικία μικρότερη των 65 χρόνων μετά την έξοδο από το νοσοκομείο σε διάστημα 12 ετών

4.4.2. Αριθμητικά Αποτελέσματα Αιτιολογικής Ανάλυσης

Στο κεφάλαιο αυτό θα παρουσιαστούν κάποια συμπεράσματα όσον αφορά την επίδραση της χρήσης των φαρμάκων και συγκεκριμένα της ημερήσιας δόσης στην επιμήκυνση της ζωής των νεφροπαθών ασθενών. Το δείγμα που επιλέχθηκε δεν είναι αντιπροσωπευτικό του πληθυσμού καθώς περιέχει ασθενείς με ηλικίες από 30 ως 90 με τους περισσότερους να κυμαίνονται μεταξύ 80-85 χρονών. Χρησιμοποιήθηκαν μοντέλα Αιτιολογικής Ανάλυσης για να βρεθεί η επιρροή που έχει η δοσολογία ενός φαρμάκου στην επιμήκυνση της ζωής. Πρέπει να σημειωθεί εδώ ότι έγινε χρήση της βιβλιοθήκης *Causalinference της Python.*

Αρχικά πρέπει να σημειωθεί ότι επιλέχθηκε το φάρμακο που είχε τις περισσότερες συνταγογραφήσεις. Αυτό είναι το φάρμακο Trombyl. Ο αριθμός των δειγμάτων που χρησιμοποιήθηκαν για το πείραμα ήταν 2039 τα οποία προέκυψαν ύστερα από συνένωση των συνόλων δεδομένων “demo_large.csv”, ”dors_large.csv ”, “Lmed_large.csv ”, “Slv_all.csv ”. Πιο συγκεκριμένα χρησιμοποιήθηκε Inner Join χρησιμοποιώντας μοναδικό αριθμό lopnr του κάθε ασθενή. Η συγκεκριμένη μέθοδος επιλέγει εγγραφές σύμφωνα με μια μεταβλητή ΄κλειδί΄ που περιέχεται και στα δύο σύνολα δεδομένων. Έπειτα έγινε διαγραφή των διπλότυπων τιμών. Τα παραπάνω σύνολα δεδομένων αντιστοιχούν σε δεδομένα δημογραφικών στοιχείων των ασθενών, σε στοιχεία θνησιμότητας, στοιχεία συνταγογραφούμενων φαρμάκων και σε στοιχεία νοσηλείας σε νοσοκομείο. Οι 1063 ήταν άνδρες και οι 976 ήταν γυναίκες. Παράλληλα, βρέθηκε η ηλικία των ασθενών την μέρα χορήγησης του φαρμάκου “Trombyl” ώστε να χρησιμοποιηθεί στην Αιτιολογική Ανάλυση σαν παράμετρος επιρροής. Το σύνολο τιμών της ημερήσιας δόσης του συγκεκριμένου φαρμάκου είναι 100 ή 1000. Από το σύνολο των ατόμων οι 1846 παίρνουν ημερήσια δόση 100 και οι 193 ημερήσια δόση 1000. Ο συγκεκριμένος αριθμός αποτελεί καθαρό αριθμό απαλλαγμένο από μονάδες μέτρησης. Με τη χρήση των τριών εκτιμητών έγινε προσπάθεια να εκτιμηθεί αν οι ασθενείς

που παίρνουν δοσολογία 100 του φαρμάκου Trombyl παρουσιάζουν καλύτερα αποτελέσματα από τους υπόλοιπους ασθενείς όσον αφορά την επιμήκυνση της ζωής. Τα χαρακτηριστικά που χρησιμοποιήθηκαν σαν επιπλέον μεταβλητές επιρροής ήταν αρχικά η ηλικία κατά την οποία ξεκίνησε η φαρμακευτική αγωγή καθώς και αν ο ασθενής ήταν γυναίκα ή άντρας, όπου αν ο ασθενής είναι γυναίκα συμβολίζεται με 1 και αν είναι άντρας με 0. Η επιλογή των συγκεκριμένων χαρακτηριστικών έγινε με βάση την συσχέτιση των μεταβλητών αυτών με την ημερήσια δόση του φαρμάκου. Επιλέχθηκαν οι μεταβλητές με την μεγαλύτερη συσχέτιση διότι επηρεάζουν περισσότερο το αποτέλεσμα. Όπως φαίνεται στον Heat map Εικόνα 25.



Εικόνα 24 Heat map με τις συσχετίσεις των μεταβλητών

Στον Πίνακα 29 παρουσιάζεται μια στατιστική επισκόπηση των δεδομένων του μοντέλου.

Summary Statistics

Variable	Controls (N_c=193) Mean	S.d.	Treated (N_t=1846) Mean	S.d.	Raw-diff
Y	2.497	1.971	3.822	2.231	1.324

Variable	Controls (N_c=193) Mean	S.d.	Treated (N_t=1846) Mean	S.d.	Nor-diff
X0	84.912	8.779	79.131	9.716	-0.624

Πίνακας 29 Απεικόνιση στατιστικής επισκόπησης των δεδομένων.

Παρατηρώντας τον Πίνακα 29, η στήλη Nor-diff υποδεικνύει την κανονικοποιημένη διαφορά των παραγόντων που επηρεάζουν το αποτέλεσμα και δίνεται από τον τύπο:

$$Nor\_diff = \frac{\bar{X}_{k,t} - \bar{X}_{k,c}}{\sqrt{\frac{s^2_{k,t} - s^2_{k,c}}{2}}} \quad (18)$$

Όπου τα $\bar{X}_{k,t}$ και $s_{k,t}$ είναι η μέση τιμή και η τυπική απόκλιση του κ-οστού παράγοντα του συνόλου των ατόμων που έλαβαν την θεραπεία, και τα $\bar{X}_{k,c}$ και $s_{k,c}$ είναι η μέση τιμή και η τυπική απόκλιση του κ-οστού παράγοντα του συνόλου των ατόμων που δεν έλαβαν την θεραπεία.

Όπως φαίνεται και από τον πίνακα 28 η μεταβλητή X0 που αντιπροσωπεύει την ηλικία των ασθενών έχει Nor-diff=-0.624 μεγαλύτερη του 0.5. Αυτό σημαίνει ότι υπάρχει μια ανισορροπία στις ηλικίες στα άτομα που λαμβάνουν ημερήσια δόση 100 του φαρμάκου και σε αυτά που λαμβάνουν 1000. Η ανισορροπία αυτή σημαίνει ότι υπάρχουν περισσότερες ηλικίες ασθενών που παίρνουν ημερήσια δόση του φαρμάκου 100 από ότι την ημερήσια δόση 1000. Στον Πίνακα 30 φαίνονται οι συντελεστές των αριθμητικών μεταβλητών που χρησιμοποιήθηκαν μέσω λογιστικής παλινδρόμησης για τον υπολογισμό του Propensity score.

Estimated Parameters of Propensity Score

	Coef.	S.e.	z	P>|z|	[95% Conf.	int.]
Intercept	-6.388	4.011	-1.592	0.111	-14.250	1.475
X3	0.090	0.042	2.145	0.032	0.008	0.173
X1	0.273	0.101	2.709	0.007	0.075	0.470
X0	0.601	0.630	0.954	0.340	-0.634	1.836
X2	-0.095	0.060	-1.591	0.112	-0.212	0.022
X3*X3	-0.001	0.000	-7.936	0.000	-0.002	-0.001
X1*X1	-0.002	0.001	-3.410	0.001	-0.003	-0.001
X2*X2	0.003	0.001	2.383	0.017	0.001	0.006
X3*X1	0.001	0.000	2.283	0.022	0.000	0.002
X0*X2	-0.045	0.024	-1.849	0.064	-0.092	0.003

Πίνακας 30 Απεικόνιση των παραμέτρων του Propensity score όπου X0=Age και X1=Female or not, X2=kommun(The municipality of residency at date of first creatinine measurement), X3=Patkost(The total cost of the medication, covered by the citizen.)

Το Propensity score είναι χρήσιμο για την βελτίωση της ισορροπίας των παραγόντων που επηρεάζουν την πρόβλεψη, όπως θα εξηγηθεί στην επόμενη παράγραφο. Η μεταβλητές που χρησιμοποιήθηκαν στις μεθόδους ήταν η μεταβλητή ‘age’ και η μεταβλητή ‘female’.

Στον Πίνακα 31 φαίνονται η στατιστική επισκόπηση των δεδομένων από τη χρήση της μεθόδου Trimming. Η μέθοδος αυτή χρησιμοποιεί το Propensity score για να μειώσει την ανισορροπία των παραγόντων που επηρεάζουν την πρόβλεψη. Αυτό το πετυχαίνει διαγράφοντας τα στοιχεία με ακραίες τιμές Propensity score. Ένας καλός κανόνας είναι να διαγραφούν οι τιμές των οποίων η το Propensity Score είναι μικρότερο από 0.1 ή μεγαλύτερη από 0.9. (Imbens, 2015) Αυτό έγινε χρησιμοποιώντας την μέθοδο Trimming της βιβλιοθήκης causalinference της Python. Όπως φαίνεται και στον Πίνακα 31 η κανονικοποιημένη τιμή της μεταβλητής X0 που αντιστοιχεί στην ηλικία μειώθηκε.

Summary Statistics

Variable	Controls (N_c=175) Mean	S.d.	Treated (N_t=1452) Mean	S.d.	Raw-diff
Y	2.440	1.967	3.784	2.227	1.344

Variable	Controls (N_c=175) Mean	S.d.	Treated (N_t=1452) Mean	S.d.	Nor-diff
X0	86.714	6.588	82.409	7.580	-0.606

Πίνακας 31 Απεικόνιση στατιστικής επισκόπησης των δεδομένων μετά το Trimming.

Στον Πίνακα 32 παρουσιάζονται τα αποτελέσματα της χρήσης του OLS εκτιμητή. Ο OLS εκτιμητής χρησιμοποιεί τη εξίσωση παλινδρόμησης στην προκειμένη περίπτωση:

$$Y_i = a + \beta D_i + \gamma'(X_i - \bar{X}_i) + \varepsilon_i\,, \tag{19}$$

όπου το D_i αντιπροσωπεύει αν πήρε ο ασθενής η όχι 100 του φαρμάκου, το X_i αντιπροσωπεύει τα covariates, το $\bar{X}_i$ αντιπροσωπεύει το μέσο όρο των covariates.

Treatment Effect Estimates: OLS

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	1.171	0.163	7.202	0.000	0.852	1.490

Πίνακας 32 Πίνακας απεικόνισης του ATE score του OLS εκτιμητή

Εδώ το ATE=Average Treatment Effect υποδεικνύει την μέση θεραπευτική αγωγή (Average Treatment Effect), η οποία δίνεται από τον τύπο

$$ATE = \frac{1}{n}\sum_i \big(Y_i(1) - Y_i(0)\big), \tag{23}$$

οπού n είναι ο αριθμός των ασθενών, $Y_i(1), Y_i(0)$υποδεικνύουν το αποτέλεσμα του ασθενή παίρνοντας ή όχι τη θεραπεία . Όπως φαίνεται από τον Πίνακα 32 η εκτίμηση ATE=1.171, κάτι που σημαίνει ότι οι δόσεις των 100 του φαρμάκου βοηθούν στην αύξηση του χρόνου ζωής των ασθενών. Ακόμη χρησιμοποιήθηκαν δυο επιπλέον εκτιμητές. Ο πρώτος είναι nearest neighborhood Matching εκτιμητής και ο δεύτερος είναι ο weighting estimator. Όσον αφορά τον εκτιμητή nearest neighborhood Matching σε όλες τις περιπτώσεις υπολογίζεται το bias μέσω μιας βοηθητικής παλινδρόμησης. Χρησιμοποιώντας παλινδρόμηση διατηρείται η εκτίμηση του εκτιμητή nearest neighborhood Matching unbiased. Ο πρώτος εκτιμητής αντί για το propensity score χρησιμοποιεί ζεύγη από τα σύνολα τον ασθενών που έλαβαν την θεραπεία και αυτών που δεν την έλαβαν προσαρμόζοντας απευθείας τις μεταβλητές που τους επηρεάζουν. Πιο συγκεκριμένα χρησιμοποιεί

την εξίσωση (14) όπου το $\left|\left|X_j - X_i\right|\right|$ είναι κάποια απόσταση όπως η ευκλείδεια. Τα X_j, X_i είναι τα διανύσματα που περιέχουν τα χαρακτηριστικά των ασθενών.

Ο δεύτερος εκτιμητής χρησιμοποιεί τον παρακάτω τύπο με βάρη:

$$Y_i = a + \beta D_i + \gamma'(X_i) + \varepsilon, \tag{20}$$

Όπου το βάρος για την μονάδα i είναι $\frac{1}{\hat{P}(X)}$, αν το i είναι στο σύνολο των ασθενών που έλαβαν την θεραπεία(treatment group) και $\frac{1}{1-\hat{P}(X)}$, αν το i είναι στο σύνολο των ασθενών που δεν έλαβαν την θεραπεία(control group). Στον Πίνακα 33 και στον Πίνακα 34 φαίνονται τα αποτελέσματα της εφαρμογής των παραπάνω εκτιμητών στο πρόβλημα.

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.895	0.239	3.750	0.000	0.427	1.363

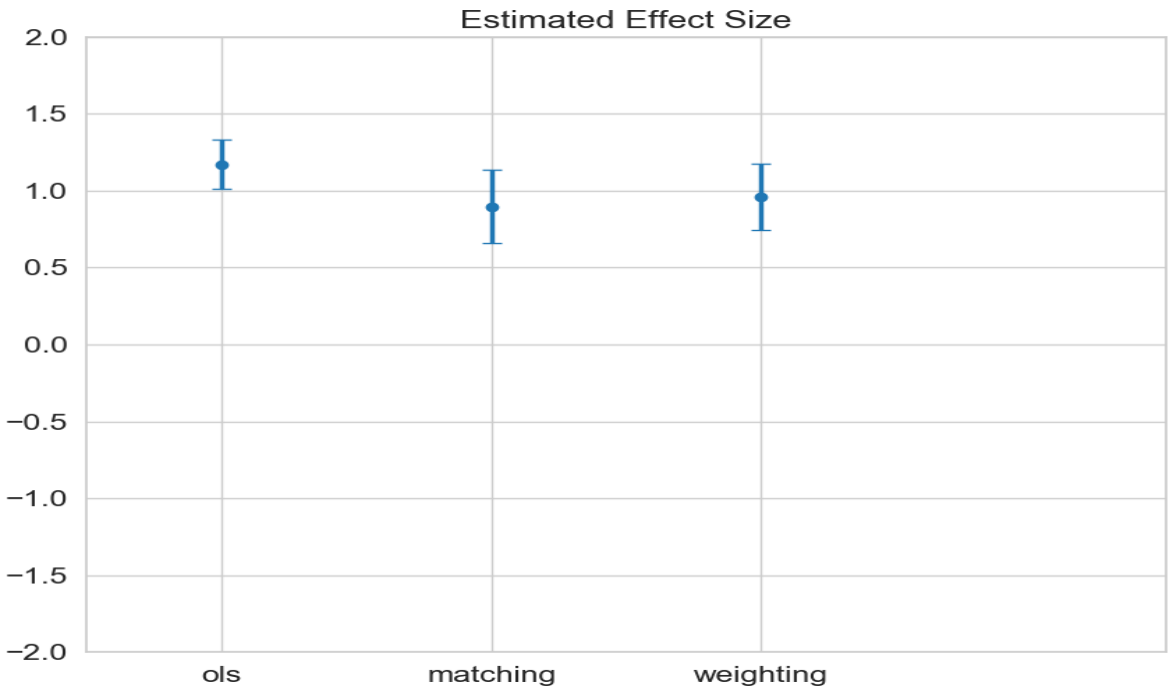
Πίνακας 33 Πίνακας απεικόνισης του ATE score nearest neighborhood Matching εκτιμητή

Treatment Effect Estimates: Weighting

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.961	0.217	4.437	0.000	0.536	1.385

Πίνακας 34 Πίνακας απεικόνισης του ATE score του Weighting εκτιμητή

Όπως φαίνεται στον Πίνακα 34, ο εκτιμητής Weighting, δίνει τιμή $ATE = 0{,}961$. Η θετική τιμή του εκτιμητή υποδεικνύει ότι η δοσολογία 100 του φαρμάκου αυξάνει τον χρόνο ζωής των ασθενών. Πρέπει να σημειωθεί εδώ ότι ο εκτιμητής Weighting εξαρτάται από το Propensity Score. Επιπλέον, επειδή τα εκτιμώμενα propensity score εισέρχονται ως παρονομαστής, οποιοσδήποτε θόρυβος στα εκτιμώμενα Propensity score μπορεί να προκαλέσει ουσιαστική μεροληψία.



Εικόνα 25 Αποτελέσματα των τριών εκτιμητών

Στην Εικόνα 26 φαίνεται πως λειτούργησαν οι τρεις εκτιμητές όσον αφορά τα δεδομένα. Από τις εκτιμήσεις των τριών εκτιμητών προέκυψε ότι η ημερήσια δόση 100 του φαρμάκου 'Trombyl' αυξάνει την διάρκεια ζωής των ασθενών 1-2 χρόνια κατά μέσο όρο.

Ακόμη χρησιμοποιήθηκαν οι τρεις εκτιμητές ώστε να φανεί αν υπάρχει επιρροή στην διάρκεια ζωής των ασθενών ανάλογα με το αν ο ασθενής είναι γυναίκα ή άντρας. Εδώ πρέπει να σημειωθεί ότι το σύνολο δεδομένων όπως και πριν είναι οι ασθενείς που τους χορηγήθηκε το φάρμακο 'Trombyl'. Έτσι οι παρατηρήσεις που έγιναν προηγουμένως ισχύουν και σε αυτή την περίπτωση. Σαν μεταβλητές επιρροής χρησιμοποιήθηκαν η ηλικία, η ημερήσια δόση. Όπως και πριν στον Πίνακα 35 παρουσιάζονται κάποια στατιστικά στοιχεία. Ο Μέσος όρος που εμφανίζεται στον Πίνακα 35 είναι το άθροισμα όλων των τιμών της μεταβλητής διά το πλήθος των εγγραφών. Η μέση τιμή της μεταβλητής 'age' διαφέρει από τις προηγούμενες μετρήσεις και αυτό γίνεται διότι τα δείγματα έχουν μοιραστεί διαφορετικά βασισμένα στο φύλο και όχι στην ημερήσια δόση όπως πριν.

Summary Statistics

Variable	Controls (N_c=1063) Mean	S.d.	Treated (N_t=976) Mean	S.d.	Raw-diff
Y	3.725	2.273	3.665	2.206	-0.060

Variable	Controls (N_c=1063) Mean	S.d.	Treated (N_t=976) Mean	S.d.	Nor-diff
X0	77.082	10.099	82.505	8.564	0.579
X1	147.269	201.203	222.591	308.951	0.289

Πίνακας 35 Πίνακας απεικόνισης στατιστικής επισκόπησης των δεδομένων.

Παρατηρώντας τον Πίνακα 35 , η στήλη Nor-diff υποδεικνύει την κανονικοποιημένη διαφορά των παραγόντων που επηρεάζουν το αποτέλεσμα. Όπως φαίνεται και από τον Πίνακα 35 η μεταβλητή X0 που αντιπροσωπεύει την ηλικία των ασθενών έχει Nor-diff=0.579 μεγαλύτερη του 0.5. Αυτό σημαίνει ότι υπάρχει μια ανισορροπία στις ηλικίες στα άτομα που λαμβάνουν ημερήσια δόση 100 του φαρμάκου και σε αυτά που λαμβάνουν 1000. Η ανισορροπία αυτή σημαίνει ότι υπάρχουν περισσότερες ηλικίες ασθενών που παίρνουν ημερήσια δόση του φαρμάκου 100 από ότι την ημερήσια δόση 1000.

Στον Πίνακα 36 φαίνονται οι συντελεστές των μεταβλητών που χρησιμοποιήθηκαν μέσω λογιστικής παλινδρόμησης για τον υπολογισμό του Propensity score. Η μεταβλητές που χρησιμοποιήθηκαν στις μεθόδους ήταν η μεταβλητή ‘age’ και η μεταβλητή ‘daily dose’.

Estimated Parameters of Propensity Score

	Coef.	S.e.	z	P>|z|	[95% Conf.	int.]
Intercept	-0.104	2.166	-0.048	0.962	-4.350	4.142
X1	-0.081	0.057	-1.424	0.154	-0.193	0.031
X0	0.001	0.000	2.963	0.003	0.000	0.001
X3	-0.005	0.002	-2.925	0.003	-0.009	-0.002
X2	0.055	0.028	1.966	0.049	0.000	0.109
X1*X1	0.001	0.000	2.446	0.014	0.000	0.002
X2*X2	-0.001	0.001	-1.726	0.084	-0.002	0.000

Πίνακας 36 Πίνακας απεικόνισης των παραμέτρων του Propensity score όπου X0=Age και X1=daily dose, X2=kommun(The municipality of residency at date of first creatinine measurement), X3=Patkost(The total cost of the medication, covered by the citizen

Στον Πίνακα 37 φαίνονται η στατιστική επισκόπηση των δεδομένων μετά τη χρήση της μεθόδου Trimming. Η μέθοδος αυτή χρησιμοποιεί το Propensity score για να μειώσει την ανισορροπία των παραγόντων που επηρεάζουν την πρόβλεψη. Αυτό το πετυχαίνει διαγράφοντας τα στοιχεία με ακραίες τιμές Propensity score. Ένας καλός κανόνας είναι να διαγραφούν οι τιμές των οποίων η το Propensity Score είναι μικρότερο από 0.1 ή μεγαλύτερη από 0.9. (Imbens, 2015) Αυτό έγινε χρησιμοποιώντας την μέθοδο Timm της βιβλιοθήκης causalinference της Python. Όπως φαίνεται και στον Πίνακα 37 η κανονικοποιημένες τιμές των μεταβλητών μειώθηκαν.

Summary Statistics

Variable	Controls (N_c=1061) Mean	S.d.	Treated (N_t=972) Mean	S.d.	Raw-diff
Y	3.732	2.269	3.676	2.204	-0.056

Variable	Controls (N_c=1061) Mean	S.d.	Treated (N_t=972) Mean	S.d.	Nor-diff
X0	77.039	10.059	82.439	8.520	0.579
X1	145.662	197.951	219.392	305.523	0.286

Πίνακας 37 Απεικόνιση στατιστικής επισκόπησης των δεδομένων μετά το Trimming.

Στον Πίνακα 38 παρουσιάζονται τα αποτελέσματα του OLS εκτιμητή. Έχει χρησιμοποιηθεί ξανά η εξίσωση 23. Εδώ το D_i αντιπροσωπεύει το φύλο των ασθενών, το X_i αντιπροσωπεύει τα covariates, το $\bar{X}_i$ αντιπροσωπεύει το μέσο όρο των covariates. Όπως φαίνεται και από τον Πίνακα 38 ο OLS εκτιμητής δείχνει ότι οι γυναίκες ζουν λίγο περισσότερο από τους άντρες κατά 0.19 χρόνια. Το φύλο έχει επιρροή στην διάρκεια ζωής των ασθενών. Υποθέτοντας ότι το μοντέλο περιγράφει με ακρίβεια τα covariates x, το CausalModel παρέχει ένα διάστημα εμπιστοσύνης 95%. Αυτό σημαίνει ότι, εάν επρόκειτο να επαναληφθεί το πείραμα, στο 95% των περιπτώσεων το μέσο αποτέλεσμα της θεραπείας θα ήταν εντός αυτού του διαστήματος. Συνολικά, φαίνεται ότι η αν ο ασθενής είναι γυναίκα έχει θετική επίδραση στην διάρκεια ζωής.

Treatment Effect Estimates: OLS

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.186	0.102	1.821	0.069	-0.014	0.385

Πίνακας 38 Πίνακας απεικόνισης του ATE score του OLS εκτιμητή

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.257	0.111	2.303	0.021	0.038	0.475

Πίνακας 39 Πίνακας απεικόνισης του ATE score του nearest neighborhood Matching εκτιμητή

Όπως φαίνεται και από τον Πίνακα 39 ο Matching εκτιμητής εκτιμά ξανά ότι οι γυναίκες ζουν παραπάνω από τους άντρες.

Στον Πίνακα 40 παρουσιάζονται τα αποτελέσματα του weighting εκτιμητή.

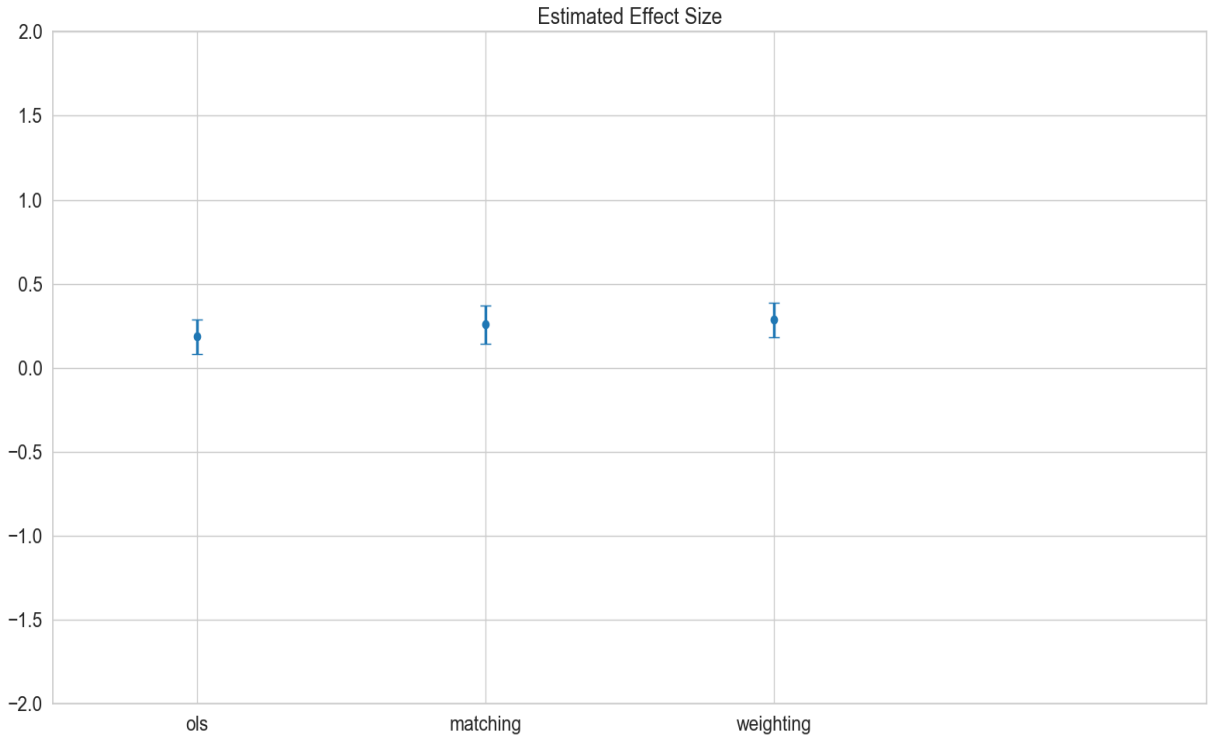
Treatment Effect Estimates: Weighting

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.284	0.103	2.758	0.006	0.082	0.486

Πίνακας 40 Πίνακας απεικόνισης του ATE score του Weighting εκτιμητή

Ο Weighting εκτιμητής δείχνει όπως και οι προηγούμενοι ότι οι γυναίκες ζουν παραπάνω από τους άντρες. Συνολικά, φαίνεται ότι η αν ο ασθενής είναι γυναίκα είχε θετική επίδραση στην διάρκεια ζωής. Πράγματι, η τιμή P>|z|= 0,004 απορρίπτει την μηδενική υπόθεση η οποία λέει ότι υπάρχει αρνητική επίδραση στην διάρκεια ζωής αν ο ασθενής είναι γυναίκα με επίπεδο εμπιστοσύνης 97,5%.

Στην Εικόνα 27 φαίνονται οι εκτιμήσεις των τριών μεταβλητών. Όπως φαίνεται και στην Εικόνα 27 οι γυναίκες ασθενείς ζουν περισσότερο από τους άντρες ασθενείς.



Εικόνα 26 Αποτελέσματα των τριών εκτιμητών

5.Συμπεράσματα

Στην παρούσα εργασία μελετήθηκαν δεδομένα ασθενών με νεφρική ανεπάρκεια και δόθηκε λύση στα προβλήματα που είχαν τεθεί στην αρχή αυτής της εργασίας. Για να καταστεί αυτό εφικτό χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python. Αρχικά χρησιμοποιήθηκε Περιγραφική Στατιστική για την εξαγωγή συμπερασμάτων από τα δεδομένα. Επιπλέον χρησιμοποιήθηκαν αλγόριθμοι Μηχανικής Μάθησης ώστε να γίνει η πρόβλεψη για την διάρκεια ζωής των ασθενών και έγινε χρήση τεχνικών feature selection για να βρεθούν ποιες μεταβλητές επηρεάζουν περισσότερο την διάρκεια ζωής. Ακόμη με την χρήση αλγορίθμων Μηχανικής Μάθησης έγινε πρόβλεψη για την αιτία θανάτου των ασθενών καθώς και χρήση τεχνικών feature selection για την εύρεση των μεταβλητών που επηρεάζουν την πρόβλεψη.

Χρησιμοποιώντας τεχνικές Feature Selection συνδυαστικά με τους αλγορίθμους Μηχανικής Μάθησης εξήχθησαν συμπεράσματα για τις μεταβλητές που επηρεάζουν την διάρκεια ζωής των ασθενών όπως η περιοχή που κατοικούν και η αρχική διάγνωση. Εξάγοντας παραμέτρους που επηρεάζουν την διάρκεια ζωής των ασθενών σε μεταμόσχευση εκτός από την πρόβλεψη του χρόνου ζωής δίνεται η δυνατότητα στους κατάλληλους ανθρώπους όπως οι γιατροί να λάβουν υπόψη παραμέτρους που δεν είχαν σκεφτεί ότι επηρεάζουν ώστε να έχουν μια πιο καλή εικόνα για τον τρόπο αντιμετώπισης αυτών των ασθενών. Ακόμη χρησιμοποιώντας τεχνικές Feature Selection επιλέχθηκαν οι σημαντικότερες μεταβλητές που επηρεάζουν την αιτία θανάτου των ασθενών με νεφρική ανεπάρκεια. Έτσι προκύπτουν συμπεράσματα ώστε να γίνει σωστότερη συνταγογράφηση φαρμάκων αλλά και διαφόρων άλλων ενεργειών ανάλογα με την προβλεπόμενη αιτία θανάτου. Επιπρόσθετα, όσον αφορά το κόστος της θεραπείας χρησιμοποιώντας Feature Selection διαπιστώθηκε ποιες μεταβλητές το επηρεάζουν. Έτσι δίνεται η δυνατότητα τόσο στους ασθενείς όσο και στους γιατρούς να έχουν μια ξεκάθαρη εικόνα των παραμέτρων που επηρεάζουν το κόστος της θεραπείας και κάνοντας κατάλληλες ενέργειες να διαπιστώσουν αν το κόστος μειώνεται. Επιπρόσθετα χρησιμοποιήθηκε Ανάλυση Επιβίωσης ώστε να βρεθεί η διάρκεια ζωής των ασθενών μετά από την τελευταία έξοδο τους από το νοσοκομείο και σε σχέση με το φύλο και την κατάσταση του ασθενή. Όπως φάνηκε από τα πειράματα οι ασθενείς οι οποίοι βρίσκονται σε κατάσταση μεταμόσχευσης επιζούν περισσότερο. Σύμφωνα με αυτά τα αποτελέσματα οι επιστήμονες έχουν μια πιο έγκυρη εικόνα για την διάρκεια ζωής των ασθενών τόσο σε αιμοκάθαρση όσο και σε μεταμόσχευση και κατά πόσο αυτή επηρεάζεται από παράγοντες όπως η ηλικία και το φύλο. Τέλος χρησιμοποιήθηκε Αιτιολογική Ανάλυση για την εξαγωγή αποτελεσμάτων για την επιρροή του φαρμάκου 'Trombyl' στην διάρκειας ζωής των ασθενών. Από τα συμπεράσματα που προέκυψαν φάνηκε κατά πόσο η ηλικία και το φύλο επηρεάζουν την απόδοση του φαρμάκου.

6. Μελλοντικές επεκτάσεις

Στην παρούσα εργασία τέθηκαν κάποια προβλήματα σε σχέση με το σύνολο των δεδομένων ασθενών με νεφρική ανεπάρκεια. Χρησιμοποιώντας τεχνικές Ανάλυσης δεδομένων απαντήθηκαν όσο το δυνατόν καλύτερα. Μελλοντικά θα μπορούσε να γίνει ανάλυση των δεδομένων και σε συνδυασμό με αλλά δεδομένα να εξεταστούν προβλήματα όπως η εύρεση παραμέτρων που επηρεάζουν την διάρκεια ζωής των ασθενών σε αιμοκάθαρση ή την διάρκεια των ασθενών μετά την μεταμόσχευση. Ακόμα χρησιμοποιώντας επιπλέον δεδομένα φαρμάκων να εξεταστεί ποιο φάρμακο είναι πιο

αποτελεσματικό στην επιμήκυνση της ζωής των ασθενών. Τέλος έχοντας περισσότερα δεδομένα τα ερωτήματα που τέθηκαν θα μπορέσουν να απαντηθούν με μεγαλύτερη ακρίβεια καθώς και να δημιουργηθούν νέα ερωτήματα που θα χρειαστούν λύση χρησιμοποιώντας τεχνικές που αναφέρθηκαν στην παρούσα εργασία.

Αναφορές

Akbilgic, O. O. (2019). *Machine learning to identify dialysis patients at high death risk.* Kidney international reports 4.9 (2019): 1219-1229.

Al-Aidaroos, K. M. (2012). Medical data classification with Naive Bayes approach. *Information Technology Journal, 11(9), 1166*.

Alam, M. Z. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked 15 (2019): 100180*.

Andersen, P. K. (2014). *Survival analysis, overview.*

Baranidharan, B. P. (2019, September). "Cardio-Vascular Disease Prediction based on Ensemble technique enhanced using Extra Tree Classifier for Feature Selection. *International Journal of Recent Technology and Engineering (IJRTE)*.

BC Transplant. "Clinical guidelines for kidney transplantation." . (2018).

Becker, S. O. (2002). Estimation of average treatment effects based on propensity scores. . *The stata journal, 2(4), 358-377.*

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. . Στο A. P. Bradley, *Pattern recognition* (σσ. 1145-1159).

Cai, R. Q. (2018). Causal discovery from discrete data using hidden compact representation. *In Advances in neural information processing systems, pp. 2666-2674. 2018.*

Carroll, K. J. (2003). *On the use and utility of the Weibull model in the analysis of survival data.*

Chandrahasan, R. K. (2011, November). An empirical comparison of boosting and bagging algorithms. *International Journal of Computer Science and Information Security, 9(11), 147.*

Chapel, J. M. (2017). Prevalence and medical costs of chronic diseases among adult Medicaid beneficiaries. American journal of preventive medicine, 53(6), S143-S154. *American journal of preventive medicine, 53(6), S143-S154.*, 143-154.

Chatton, A. L. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports, 10(1), 1-13.*

Chen, C. L. (2004). *Predicting disease risks from highly imbalanced data using random forest.* University of California, Berkeley, 110(1-12), 24.

Chipman, H. A. (2007). *Bayesian ensemble learning. In Advances in neural information processing systems.*

Chipman, H. A. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*.

Cichosz, P. (2015). Data Mining Algorithms: Explained Using R. Στο P. Cichosz, *Data Mining Algorithms: Explained Using R* (σ. 716). Wiley-Blackwell .

Cole, S. R. (2010). *Survival analysis in infectious disease research: describing events in time.*

Dey, A. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies, 7(3), 1174-1179.*

El-Houssainy, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*.

Fonti, V. &. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics, 30, 1-25.*

Geurts, P. E. (2006). *Extremely randomized trees. Machine learning.*

Goel, M. K. (2010, December). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*.

Guo, G. W. (2003, November). KNN model-based approach in classification. *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.*

Guyon, I. W. (2002). *Gene selection for cancer classification using support vector machines. Machine learning.*

Harrison, O. (2018, September 10). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Ανάκτηση από Towards Data Science: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Haury, A. C. (2011). *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures.*

Healthcare Cost and Utilization Project. "Overview of the nationwide inpatient sample (NIS). (2020, October 22). Ανάκτηση από Healthcare Cost and Utilization Project. "Overview of the nationwide inpatient sample (NIS): www. hcup-us. ahrq. gov/databases. jsp

Hemphill, E. L. (2014, December). Feature selection and classifier performance on diverse bio-logical datasets. *BMC bioinformatics (Vol. 15, No. S13, p. S4). BioMed Central.*

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20(1), 217-240.*

Hu, R. (2011, July). Medical data mining based on decision tree algorithm. *Computer and Information Science*.

Ichikawa, D. S. (2016). How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. *Journal of biomedical informatics, 64, 20-24.*

İlkerEtikan*, O. B. (2018, February). Survival Analysis: A Major Decision Technique in Healthcare Practices. *International Journal of Science and Research Methodology*.

Imbens, G. &. (2015). Causal inference in statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

Indraja, B. &. (2018). Classification of medicines using naive bayes classifier. *Research Journal of Pharmacy and Technology*.

Jain, A. (2016, January 28). A Complete Tutorial on Ridge and Lasso Regression in Python.

Jović, A. B. (2015, May). A review of feature selection methods with applications. *In 2015 38th international convention on information and communication technology, electronics and microelectronics*.

Julian, D. (2016). Designing Machine Learning Systems with Python.

Kamkar, I. G. (2015, February). Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. *Journal of biomedical informatics, 53, 277-290.*

Kartsonaki, C. (2016). *Survival analysis.*

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(2), 119-127.*

Kathleen F. Weaver, V. M. (2017). An Introduction to Statistical Analysis in Research.

Kaushik, S. C. (2019). Comparative Analysis of Features Selection Techniques for Classification in Healthcare. Στο S. C. Kaushik, *In MLDM (2)* (σσ. (pp. 458-472).).

Khalilia, M. C. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*.

Khan, S. H. (2010). *Predictive models for chronic renal disease using decision trees, naïve bayes and case-based methods.*

Kingsford, C. &. (2008). What are decision trees?. Στο C. &. Kingsford, *Nature biotechnology* (σσ. 26(9), 1011-1013.).

Lee, J. W. (2005). An extensive comparison of recent classification tools applied to microarray data. Στο *Computational Statistics & Data Analysis,* (σσ. 869-885).

Li, C. Z. (2012, October). Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. *Computational and mathematical methods in medicine*.

Liu, Y. W. (2012). New machine learning algorithm: Random forest. *In International Conference on Information Computing and Applications*, (σσ. (pp. 246-252).).

Luo, L. L. (2019, November 11). Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China. *Health informatics journal, 26(3), 1577-1598.*

Magidson, J. (1982). Some common pitfalls in causal analysis of categorical data. *Journal of Marketing Research, 19(4), 461-471.*

Martin D. Marciniak, M. J. (2005). *The cost of treating anxiety: The medical and demographic correlates that impact total medical costs.* Indianapolis: WILEY-LISS, INC.

Matschinger, H. H. (2020). *A Comparison of Matching and Weighting Methods for Causal Inference Based on Routine Health Insurance Data, or: What to do If an RCT is Impossible.*

Mo, X. C. (2019, October). Early and accurate prediction of clinical response to methotrexate treatment in juvenile idiopathic arthritis using machine learning. *Frontiers in pharmacology, 10, 1155.*

NATIONAL KIDNEY FOUNDATION. (2020). Ανάκτηση από National Kidney Foundation Inc: https://www.kidney.org/

Oo, A. N. (2019). *Decision Tree Models for Medical Diagnosis.*

Panda, D. R. (2019, November). Predictive Systems: Role of Feature Selection in Prediction of Heart Disease. *Journal of Physics: Conference Series (Vol. 1372, No. 1, p. 012074). IOP Publishing.*

Quan, H. S. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care, 1130-1139.*

Rahman, M. N. (2018). Macroeconomic variables of India and finite sample properties of OLS under classical assumptions. *Pacific Business Review International, 10(8), 7-14.*

Rajkomar, A. O. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*.

Ramteke, R. J. (2012). Automatic medical image classification and abnormality detection using k-nearest neighbour. *International Journal of Advanced Computer Research, 2(4), 190*.

Razzak, M. I. (2020, October). Big data analytics for preventive medicine. *Neural Computing & Applications* .

Robert Thomas, M. A. (2008, June). Chronic Kidney Disease and Its Complications.

Rogers, J. &. (2005). *Identifying feature relevance using a random forest.*

Rosa, R. G. (2017). *Mortality of adult critically ill subjects with cancer. Respiratory Care.*

Sahoo, A. K. (2019). Intelligence-based health recommendation system using big data analytics. Στο N. D. Behera, *Big data analytics for intelligent healthcare management* (σσ. 227-246).

Sanders, R. (1987). The Pareto principle: its use and abuse. *Journal of Services Marketing.*

Sehra, C. (2018, January 19). *Decision Trees Explained Easily*. Ανάκτηση από https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248

Spathis, D. &. (2019). Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health informatics journal, 25(3), 811-827.*

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1), 1.*

Synced. (2017, October 25). *How Random Forest Algorithm Works in Machine Learning*. Ανάκτηση από https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. . *Journal of the American Statistical Association, 101(476), 1607-1618.*

Tjortjoglou, S. (2016, August 5). Surviving the NFL - Survival Analysis using Python.

Urrutia, J. D. (2015, June). Survival Analysis of Patients with End Stage Renal Disease. *Journal of Physics: Conference Series (Vol. 622, No. 1, p. 012014).*

Wong, L. (χ.χ.). Causal Inference in Python Blog .

World Health Organization. (2018, March 22). Ανάκτηση από https://www.who.int/en/news-room/fact-sheets/detail/depression

Zhang, X. Y. (2020). Predicting Missing Values in Medical Data Via XGBoost Regression. *ournal of Healthcare Informatics Research, 1-12.*

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine, 4(11).*

Zhao, J. H. (2020). Causal inference for the effect of environmental chemicals on chronic kidney disease. *Computational and Structural Biotechnology Journal, 18, 93-99.*