ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΤΗΣ ΔΙΑΧΕΙΡΙΣΗΣ ΤΩΝ ΥΠΗΡΕΣΙΩΝ ΥΓΕΙΑΣ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Ελένη Φραντζεσκάκη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούλιος 2023

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΤΗΣ ΔΙΑΧΕΙΡΙΣΗΣ ΤΩΝ ΥΠΗΡΕΣΙΩΝ ΥΓΕΙΑΣ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Ελένη Φραντζεσκάκη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούλιος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

-Σωτήριος Μπερσίμης, Αναπληρωτής Καθηγητής (Επιβλέπων)

-Γεώργιος Τζαβελάς, Αναπληρωτής Καθηγητής

-Σωτήριος Τασουλής, Επίκουρος Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

OPTIMIZING HEALTHCARE SERVICES MANAGEMENT USING DATA ANALYTICS AND MACHINE LEARNING METHODS

By

Eleni Frantzeskaki

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
July 2023

Φτάσε όπου μπορείς παιδί μου.
Μη ντραπείς αν έπαιξες καλά κι έχασες.
Να ντραπείς αν έπαιξες κακά και κέρδισες.

Νίκος Καζαντζάκης

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου κύριο Σωτήριο Μπερσίμη για την καθοδήγησή του και για όλε στις γνώσεις που αποκόμισα.

Θα ήθελα επίσης να ευχαριστήσω την οικογένεια μου και το Δίκτυο για τη στήριξη, την κατανόηση και τη συμπαράστασή τους καθ' όλη τη διάρκεια των σπουδών μου.

Περίληψη

Σύμφωνα με διεθνείς μελέτες η σπατάλη στον κλάδο της υγείας είναι ευρύτατα διαδεδομένη. Παράλληλα, ο κλάδος της υγείας αναπτύσσεται με ταχείς ρυθμούς συλλέγοντας μεγάλο όγκο δεδομένων, η ανάλυση των οποίων, με μεθόδους αναλυτικής και στατιστικής μηχανικής μάθησης, μπορεί να συμβάλλει στον εξορθολογισμό των δαπανών αλλά και σε βελτίωση της εμπειρίας των ασθενών. Σε αυτή την εργασία γίνεται αναζήτηση στη βιβλιογραφία για εφαρμογές που βρίσκουν πεδίο εφαρμογής στον σχεδιασμό και στην οργάνωση μονάδων που παρέχουν υπηρεσίες υγείας. Στη συνέχεια, επιλεγμένες μέθοδοι αναλυτικής και αλγόριθμοι μηχανικής μάθησης εφαρμόζονται σε σύνολα δεδομένων που αντλήθηκαν από ιστότοπους ανοικτών δεδομένων, με σκοπό να διερευνηθεί η δυνατότητα των μεθόδων αυτών να συμβάλλουν στη μείωση δαπανών και στη βελτίωση εξυπηρέτησης των ασθενών.

Abstract

According to international studies, the economical waste in the health sector is widespread. At the same time, healthcare services are developing rapidly by collecting a large amount of data. The data analysis, by using analytics and statistical machine learning algorithms, can contribute to the rationalization of costs and to the improvement of patients' experience. In this work, applications from the literature on the optimization of healthcare services are presented. Finally, selected analytical methods and machine learning algorithms are applied to open datasets, so as to explore their value in reducing hospitals costs and improve patient care.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Υπηρεσίες Υγείας

Ο Παγκόσμιος Οργανισμός Υγείας (ΠΟΥ) δίνει τους ακόλουθους ορισμούς για την υγεία και την παροχή υπηρεσιών υγείας.

Ορισμός «Υγείας» κατά τον ΠΟΥ
Η υγεία είναι η «κατάσταση της πλήρους σωματικής, ψυχικής και κοινωνικής ευεξίας και όχι μόνο η απουσία ασθένειας ή αναπηρίας».

Ορισμός «Παροχή Υπηρεσιών Υγείας» κατά τον ΠΟΥ
Η παροχή υπηρεσιών υγείας περιλαμβάνει «την πρόληψη, τη θεραπεία καθώς και τη διαχείριση των ασθενειών, την προστασία της ψυχικής και σωματικής ευεξίας μέσω των υπηρεσιών που προσφέρονται από το ιατρικό, νοσηλευτικό προσωπικό και γενικότερα από τους επαγγελματίες υγείας».

Οι πρωταρχικό στόχοι των υπηρεσιών υγείας, σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, είναι η διασφάλιση δίκαιης πρόσβασης σε ποιοτική περίθαλψη, η προώθηση της ισότητας και η ενίσχυση των αποτελεσμάτων υγείας μέσω της παροχής βασικών υπηρεσιών όπως για παράδειγμα οι εμβολιασμοί, ο έλεγχος των μολυσματικών ασθενειών, η επείγουσα περίθαλψη.

Η παροχή υπηρεσιών υγείας ενδεχομένως διαφέρει μεταξύ των χωρών και των διαφορετικών συστημάτων και είναι ανάλογη παραγόντων όπως οι διαθέσιμοι οικονομικοί πόροι, οι υποδομές, πολιτιστικοί και κοινωνικοί παράγοντες.

Στα πλαίσια της δράσης του ΠΟΥ είναι η καθολική κάλυψη υγείας και η διασφάλιση ότι όλα τα άτομα θα έχουν πρόσβαση σε βασικές υπηρεσίες χωρίς να υποφέρουν.

1.2 Πάροχοι Υπηρεσιών Υγείας

Το Εθνικό Σύστημα Υγείας (Ε.Σ.Υ) της Ελλάδας συνυπάρχει με έναν εξαιρετικά ανεπτυγμένο ιδιωτικό τομέα που δραστηριοποιείται τόσο στην πρωτοβάθμια όσο και στην δευτεροβάθμια περίθαλψη (Μπιτσώρη και Μπαλάσκα, 2016).

Το Εθνικό Σύστημα Υγείας ιδρύθηκε το 1983 με το Νόμο 1397/1983 [1] στα πλαίσια μεταρρύθμισης και αναβάθμισης της δημόσιας υγείας αλλά και της αποτελεσματικής ενοποίησης των διαφόρων δημόσιων υποδομών με σκοπό την παροχή δωρεάν ιατροφαρμακευτικών και νοσηλευτικών υπηρεσιών στον πληθυσμό που διέμενε στην Ελλάδα.

Με τον ίδιο νόμο δημιουργήθηκαν και τα Κέντρα Υγείας (Κ.Υ.). Με το Νόμο ν.2889/2001 [2] συγκροτήθηκαν οι Υγειονομικές Περιφέρειες (Υ.Π.) με υπαγωγή στο Περιφερειακό Σύστημα Υγείας (ΠΕΣΥ).

Για τους ασφαλισμένους προβλέπεται η παροχή υπηρεσιών υγείας από τον Εθνικό Οργανισμό Παροχής Υπηρεσιών Υγείας (Ε.Ο.Π.Υ.Υ.) διαμέσου των εισφορών που καταβάλλουν στον Εθνικό Φορέα Κοινωνικής Ασφάλισης (Ε.Φ.Κ.Α) ενώ από το 2016 με το Νόμο ν.4368/2016 [3] διασφαλίζεται η δωρεάν πρόσβαση των ανασφάλιστων και των ευάλωτων ομάδων στο Ε.Σ.Υ..

Το Σύστημα αποτελείται από επτά Υ.Π., συγκεκριμένα Αττικής, Πειραιώς και Αιγαίου, Μακεδονίας και Θράκης, Θεσσαλίας και Στερεάς Ελλάδας, Πελοποννήσου, Ιονίων νήσων, Ηπείρου και Δυτικής Ελλάδας, Κρήτης οι οποίες διοικούν τρεις βαθμούς Φροντίδας Υγείας (Φ.Υ.) τον Α’, Β’ και Γ’.

Στην Πρωτοβάθμια Φροντίδα Υγείας ανήκουν τα Κέντρα Υγείας και οι Τοπικές Μονάδες Αυτοδιοίκησης (Το.Μ.Υ.) με σκοπό την πρόληψη, θεραπεία και αποκατάσταση των ασθενών. Στη Δευτεροβάθμια και Τριτοβάθμια Φροντίδα Υγείας ανήκουν τα Νοσοκομεία με σκοπό την ενδονοσοκομειακή περίθαλψη των ασθενών.

Τα διάσπαρτα Αγροτικά και Περιφερειακά Ιατρεία της Ελληνικής επικράτειας επίσης διοικούνται από τις Υ.Π.. Το Εθνικό Κέντρο Άμεσης Βοήθειας (Ε.Κ.Α.Β.) έχει σκοπό την αποστολή εξειδικευμένου προσωπικού στον τόπο έκτακτων συμβάντων για την παροχή άμεσης βοήθειας και διακομιδής των ασθενών προς τις πλησιέστερες μονάδες.

Τα τελευταία χρόνια η ανάπτυξη του ιδιωτικού τομέα της υγείας στην Ελλάδα είναι μεγάλη. Φαίνεται να έχει μειωθεί ο αριθμός ιδιωτικών νοσοκομείων και κλινών λόγω της μείωσης του αριθμού των μικρών ιδιωτικών νοσοκομείων, αλλά αυξάνονται σημαντικά οι ιδιώτες γιατροί και τα ιδιωτικά διαγνωστικά κέντρα.

Αυτό μπορεί να οφείλεται στη ποιότητα των δημόσιων υπηρεσιών υγείας που συχνά προκαλεί δυσαρέσκεια στους πολίτες αλλά και στη βελτίωση του επιπέδου διαβίωσης των πολιτών και την ανάπτυξη που έχει η ιδιωτική ασφάλιση υγείας (Μπιτσώρη και Μπαλάσκα, 2016).

1.3 Χρηματοδότηση Υπηρεσιών Υγείας

Οι βασικές πηγές χρηματοδότησης του ελληνικού συστήματος υγείας είναι δημόσιες και ιδιωτικές. Στις δημόσιες πηγές ανήκει ο κρατικός προϋπολογισμός με την άμεση και έμμεση φορολογία και η κοινωνική ασφάλιση από τις εισφορές εργαζόμενων και εργοδοτών. Στις ιδιωτικές πηγές ανήκει η ιδιωτική ασφάλιση, οι απευθείας πληρωμές από ιδιώτες και οι δωρεές. Το σύστημα υπηρεσιών υγείας στην Ελλάδα είναι μικτό σύστημα, όπου ο Έλληνας πολίτης έχει δωρεάν δημόσια περίθαλψη με μικρή συμμετοχή.

Οι υπηρεσίες υγείας αποτελούν κλάδο της οικονομίας που διεθνώς απορροφά σημαντικό μέρος του προϋπολογισμού της χώρας. Σύμφωνα με τον εθνικό προϋπολογισμό της χώρας που δημοσιεύθηκε πρόσφατα, προβλέφθηκαν για τη χρονιά 2023 να δοθούν στο Υπουργείο Υγείας 5.202.388.000 €.

1.4 Σημαντικότητα ποιότητας Υπηρεσιών Υγείας

Σύμφωνα με τον Donabedian (1980, Explorations in quality assessment and monitoring: the definition of quality and approaches to its assessment), η έννοια της ποιότητας στην υγεία γίνεται αντιληπτή ως το είδος της φροντίδας που αναμένεται να μεγιστοποιήσει το όφελος του ασθενή λαμβάνοντας υπόψη τις ωφέλειες, αλλά και τις απώλειες που περιέχει η διαδικασία της περίθαλψης.

Αναφέρει επίσης ότι η παροχή υπηρεσιών έχει τρεις διαστάσεις. Πρώτον, το διαπροσωπικό μέρος που περιλαμβάνει την αντιμετώπιση του ασθενούς από το ιατρικό προσωπικό αλλά και την αντιμετώπιση του προσωπικού που προσδιορίζεται από την επαγγελματική ιδεολογία. Δεύτερον, στο τεχνικό μέρος που αναφέρεται στην εφαρμογή της ιατρικής επιστήμης. Τρίτον, τις υποδομές που έχουν σχέση με το χώρο, το περιβάλλον και τις συνθήκες κάτω από τις οποίες προσφέρεται η φροντίδα.

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (1993) η ποιότητα των υπηρεσιών υγείας είναι η παροχή διαγνωστικών και θεραπευτικών πράξεων που είναι ικανές να διασφαλίσουν τα καλύτερα δυνατά αποτελέσματα στον τομέα της υγείας, στο πλαίσιο των δυνατοτήτων της σύγχρονης ιατρικής επιστήμης που στοχεύει στη μέγιστη δυνατή ικανοποίηση του ασθενή.

Η ποιότητα στις υπηρεσίες υγείας προσδιορίζεται βάσει κάποιων κριτηρίων, τα οποία σχετίζονται με άμεση παροχή της υπηρεσίας, την καταλληλόλητα, την αξιοπιστία, την ευκολία πρόσβασης, την υποστήριξη που παρέχεται κατόπιν. Μερικοί ακόμη βασικοί παράγοντες που σχετίζονται με την ποιότητα της υπηρεσίας είναι η συμπεριφορά του ιατρικού προσωπικού, η ταχύτητα επίλυσης προβλημάτων υγείας, η επάρκεια προσωπικού, η παροχή υγειονομικής περίθαλψης ανά πάσα στιγμή, οι γνώσεις, η εμπειρία, οι ικανότητες και η επαγγελματική

δεοντολογία των μονάδων, η συνεχής εκπαίδευση νοσηλευτικού, διοικητικού και τεχνικού προσωπικού, η σωστή επιλογή φαρμακευτικής αγωγής, ο σωστός σχεδιασμός των εγκαταστάσεων των νοσοκομείων ώστε να διευκολύνεται η μετακίνηση ασθενών και υλικών, η χρήση νέου τεχνολογικού εξοπλισμού, η αποφυγή νοσοκομειακών λοιμώξεων, η αξιοπιστία των διοικητικών υπηρεσιών και η καλή λειτουργία των μονάδων υγείας.

Η ποιότητα στις υπηρεσίες υγείας ενδιαφέρει τους ασθενείς που τις χρησιμοποιούν, ενδιαφέρει τους επαγγελματίες υγείας που τις προσφέρουν, τα διοικητικά και επιτελικά στελέχη της χώρας, καθώς και αυτούς που επιβαρύνονται με το κόστος των υπηρεσιών όπως το κράτος και τους ασφαλιστικούς φορείς και την κοινωνία στο σύνολο της.

Κατηγοριοποιώντας τα οφέλη για τον ασθενή, διακρίνουμε τρεις κατηγορίες. Τα υγειονομικά έχουν σχέση με τη γρήγορη διάγνωση και την επιλογή κατάλληλης θεραπείας. Η έγκαιρη διάγνωση συντελεί σε ταχύτερη ανάρρωση και κατά συνέπεια μικρότερο κόστος. Τα ψυχολογικά οφέλη έχουν σχέση με την ψυχολογία του ασθενούς, όπου όταν η ψυχολογία είναι θετική επηρεάζει θετικά τη θεραπεία του. Τέλος, συμβάλλουν οι καλές συνθήκες διαμονής στο νοσοκομείο καθώς και η ελαχιστοποίηση του χρόνου παραμονής εκεί.

Για τους επαγγελματίες υγείας τα οφέλη είναι ο σεβασμός και η εκτίμηση που λαμβάνουν από τους ασθενείς και η προσωπική ικανοποίηση που νιώθουν.

Για τις μονάδες υγείας τα οφέλη είναι οικονομικά, καθώς όσο πιο έγκαιρη είναι η διάγνωση και μετά η θεραπεία, τόσο μειώνονται τα λειτουργικά κόστη. Αντίστοιχα, λειτουργεί αυτό το όφελος και για τους ασφαλιστικούς φορείς καθώς μειώνεται το ύψος των δαπανών που θα χρειαστεί να καταβάλλουν. Τα οφέλη για το κράτος είναι η βελτίωση της εικόνας του και η αύξηση εμπιστοσύνης από τους πολίτες καθώς αυξάνεται η αποδοτικότητα και αποτελεσματικότητα των υπηρεσιών υγείας και αντίστοιχα μειώνεται ο χρόνος αναμονής των ασθενών.

1.5 Διοίκηση Υπηρεσιών Υγείας

Οι επαγγελματίες της υγείας έχουν στόχο την παροχή ποιοτικών υπηρεσιών και έρχονται αντιμέτωποι με την παραγωγικότητα και την αποδοτικότητα των υπηρεσιών που προσφέρουν. Δεν είναι λίγες οι φορές που στο χώρο της υγείας οι επαγγελματίες καλούνται να λάβουν ορθολογικές αποφάσεις οι οποίες απαιτούν ικανότητες διοίκησης/ management.

Τρεις ορισμοί για την έννοια της διοίκησης όπως παρατέθηκαν στη μελέτη του Τζαχρήστα (2005) είναι οι ακόλουθοι:

- Η διαδικασία του management μπορεί να λάβει χώρα σε οποιοδήποτε είδος Οργανισμού. Management είναι ο συντονισμός και η ενοποίηση/εναρμόνιση όλων των παραγωγικών πόρων (ανθρώπινων, υλικών, τεχνικών) για να επιτευχθούν συγκεκριμένα αποτελέσματα.

- Με τον όρο management εννοείται η μεθοδική προσπάθεια προγραμματισμού, οργάνωσης, διεύθυνσης και ελέγχου δραστηριοτήτων για την επιτυχία δεδομένων σκοπών.
- Διοίκηση είναι η διαδικασία του συντονισμού ανθρώπινων και άλλων πόρων με σκοπό την επίτευξη των στόχων ενός οργανισμού.

Οι λειτουργίες της διοίκησης σε έναν οργανισμό μπορούν να κατηγοριοποιηθούν στον Σχεδιασμό όπου θέτονται οι βασικές κατευθύνσεις και σκοποί του οργανισμού και διαμορφώνεται το αρχικό πλάνο προσέγγισης, η Οργάνωση όπου καθορίζονται οι απαραίτητες δραστηριότητες για την επίτευξη των σκοπών, ομαδοποιούνται, αναθέτονται σε συγκεκριμένες ομάδες και διαμορφώνονται οι βαθμίδες εξουσίας, η Διεύθυνση/Καθοδήγηση όπου είναι το στάδιο εποπτείας και καθοδήγησης των υφισταμένων για την επίτευξη των σκοπών και ο Έλεγχος όπου είναι η φάση αξιολόγησης ώστε να διατηρηθεί το σχέδιο και εάν το αποτέλεσμα δεν είναι το επιθυμητό αναθεωρείται ή τροποποιείται η διαδικασία ή μέρος της και ακολουθεί ανατροφοδότηση του συστήματος.

Ο τομέας της υγείας, όμως, έχει σημαντικές οργανωτικές και διοικητικές ιδιομορφίες. Αυτό οφείλεται στην αυξημένη κρατική παρέμβαση στους μηχανισμούς παραγωγής και διανομής των υπηρεσιών υγείας, στην αδυναμία του ασθενή να λάβει ο ίδιος αποφάσεις καθώς δεν έχει επιστημονική γνώση και υπάρχουν καταστάσεις που δεν μπορεί καν να διαμορφώσει άποψη, στην ένταση της εργασίας όπου οι υπηρεσίες προσφέρονται από ανθρώπινο δυναμικό, στην ένταση της οργάνωσης όπου πέρα από τα ιατρεία και τα εργαστήρια, οι υπόλοιπες μονάδες είναι σύνθετες και στον τρόπο επιμερισμού εξουσίας και ευθύνης τα οποία επιμερίζονται σε άτομα όχι μόνο βάσει της θέσης τους αλλά και με βάση το κύρος/ ειδική ισχύ του επαγγέλματος πχ οι γιατροί (Τζαχρήστας, 2005).

Λόγω οικονομικών, κοινωνικών και πολιτικών αιτιών, φαίνεται επιτακτική η ανάγκη για τη βέλτιστη δυνατή διοίκηση ενός νοσοκομείου και παράλληλα τη μέγιστη ποιότητα παροχής των υπηρεσιών που προσφέρει.

Από οικονομικής πλευράς, ένας λόγος που ωθεί σε αυτό είναι η ολοένα αυξανόμενη ζήτηση των υπηρεσιών αλλά και το γεγονός ότι οι οικονομικοί πόροι είναι περιορισμένοι. Από κοινωνικοπολιτική πλευρά, ένας λόγος που ωθεί σε αυτό είναι η προάσπιση του δικαιώματος για πρόσβαση στην υγεία.

Στρατηγικές όπως η διοίκηση ολικής ποιότητας, ο κλινικός και ιατρικός έλεγχος μπορούν να βοηθήσουν σε αυτή την ανάγκη των νοσοκομείων. Επιπλέον, η υιοθέτηση νέων τεχνολογιών είναι βασική συνιστώσα στην αποτελεσματική παροχή φροντίδας στους ασθενείς, στην άμεση διάγνωση, στην επιλογή κατάλληλης θεραπείας ασθενή, στον προγραμματισμό και στη διαχείριση μίας μονάδας νοσοκομείου.

Για τη διαχείριση των υπηρεσιών υγείας φαίνεται αναγκαίο να καθοριστεί ένα πλαίσιο όπου θα παρέχει στοιχεία για την ασφαλή εξαγωγή συμπερασμάτων και την αξιολόγηση αυτών

των υπηρεσιών στο κατά πόσο ανταποκρίνεται το τελικό αποτέλεσμα στον προσχεδιασμένο σκοπό. Κατά την αξιολόγηση θα λαμβάνονται υπόψη ο βαθμός επίτευξης των σκοπών και το χρονικό διάστημα με αντικειμενικά κριτήρια.

Η επιστήμη της διαχείρισης υπηρεσιών υγείας έχει παραθέσει πλήθος τυπολογίων για να κατηγοριοποιηθούν τα βασικά στοιχεία αξιολόγησης των υπηρεσιών υγείας τα οποία είναι η αποδοτικότητά, η αποτελεσματικότητα, η επιστημονική και τεχνική ποιότητα, η επάρκεια, η επίδραση, η επίπτωση και η οικονομική διάσταση (Φαρατζιάν, 2007).

Οι στόχοι που έχουν τεθεί θα πρέπει να είναι μετρήσιμοι, να έχουν αριθμητική μορφή και να ορίζεται το χρονικό διάστημα παρακολούθησης τους, ώστε να μπορέσουν να προκύψουν, συγκεντρωθούν και να αναλυθούν τα αποτελέσματα.

Οι δείκτες αξιολόγησης είναι τα εργαλεία που θα χρησιμοποιηθούν για τη σύγκριση των αποτελεσμάτων με τους στόχους που έχουν τεθεί και την εξαγωγή των συμπερασμάτων. Οι δείκτες οφείλουν να είναι έγκυροι, σύγχρονοι, αξιόπιστοι, να λαμβάνουν υπόψη την ευαισθησία και την εξειδίκευση.

Υπάρχει μεγάλο πλήθος δεικτών, οι οποίοι μπορούν να κατηγοριοποιηθούν σε πέντε βασικές κατηγορίες, τους δείκτες υγειονομικής πολιτικής, τους κοινωνικοοικονομικούς δείκτες, τους δείκτες επιπέδου υγείας πληθυσμού, τους δείκτες παροχής υπηρεσιών υγείας και τους δείκτες κάλυψης της πρωτοβάθμιας φροντίδας υγείας. Η μέτρηση της απόδοσης μπορεί να εφαρμοστεί σε μεμονωμένες μονάδες υγείας χρησιμοποιώντας δείκτες απόδοσης μονάδας οι οποίοι συγκρίνουν την πραγματική προσφορά της μονάδας με τη μέγιστη δυνατή που θα μπορούσε να έχει και λαμβάνουν υπόψη το ιατρικό και νοσηλευτικό προσωπικό, την εξειδίκευση των παρεχόμενων υπηρεσιών, τον αριθμό και τη δομή του προσωπικού, τη γενική διαχείριση που γίνεται.

Μερικοί βασικοί δείκτες απόδοσης της μονάδας είναι η μέση διάρκεια νοσηλείας, ο μέσος χρόνος αδράνειας κλίνης, η μέση κάλυψη κλινών, το ποσοστό μείωσης ενδονοσοκομειακών λοιμώξεων, η ικανοποίηση των ασθενών, το ποσοστό επανεισαγωγών, ο αριθμός επεμβατικών πράξεων που πραγματοποιούνται (Φαρατζιάν, 2007).

Για την ικανοποίηση των ασθενών, το ποσοστό επανεισαγωγών και τον αριθμό επεμβατικών πράξεων που πραγματοποιούνται θα ακολουθήσει βιβλιογραφική έρευνα στο επόμενο κεφάλαιο και στο κεφάλαιο 4 θα εφαρμοστεί ανάλυση σε σύνολα δεδομένων χρησιμοποιώντας τεχνικές στατιστικής μηχανικής μάθησης και αναλυτικής των δεδομένων για να δοθούν λύσεις σε μονάδες που αντιμετωπίζουν τα συγκεκριμένα προβλήματα.

Εν τέλει, ο βασικός σκοπός των επαγγελματιών της υγείας είναι η υψηλή ποιότητα υπηρεσιών. Οι διεθνείς οικονομικές καταστάσεις καταδεικνύουν τη σημαντικότητα της εκπαίδευσης και εξειδίκευσης του διοικητικού, του ιατρικού, του νοσηλευτικού προσωπικού αλλά και των λοιπών επαγγελματιών του κλάδου σε σύγχρονες μεθόδους και τεχνολογίες.

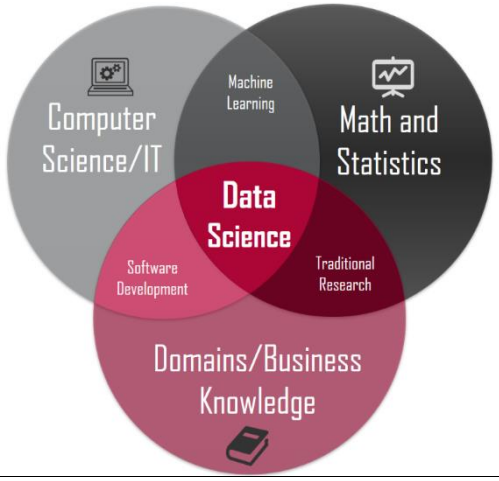
Έχει αποδειχθεί ότι ο μεγάλος όγκος δεδομένων των ημερών μας εάν αξιοποιηθεί και αναλυθεί μπορεί να οδηγήσει σε αύξηση της αποδοτικότητας αλλά και της αποτελεσματικότητας της παροχής των υπηρεσιών υγείας όπως και να συμβάλλει στην τροποποίηση των υπηρεσιών αυτών.

1.6 Στατιστική και Στατιστική Μηχανική Μάθηση

Στατιστική είναι η επιστήμη που ασχολείται με τη συλλογή, την επεξεργασία και την ανάλυση διαθέσιμων δεδομένων με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων για τους ευρύτερους πληθυσμούς.

Στατιστική Μάθηση ονομάζεται η δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων η οποία βασίζεται στη Στατιστική.

Στη Στατιστική Μηχανική Μάθηση έχουμε τη σύνθεση των στατιστικών τεχνικών μάθησης και του επιστημονικού προγραμματισμού και είναι συνδεδεμένη με τη μελέτη προτύπων και τη στατιστική μάθηση και αποτελεί τη βάση της τεχνητής νοημοσύνης. Αναλύονται στοιχεία από διάφορες πηγές και με τη χρήση των κατάλληλων τεχνικών και αλγορίθμων αυτά τα δεδομένα μετατρέπονται σε γνώση που μπορεί να αξιοποιηθεί όλους τους τομείς για αύξηση κερδοφορίας, μείωση κόστους, αποτελεσματικότερη διοίκηση και καλύτερη εξυπηρέτηση του πελάτη. Στο Σχήμα 1-1 που ακολουθεί φαίνεται η ένωση τριών τομέων που συνθέτει την Επιστήμη των Δεδομένων (Data Science).



Σχήμα 1-1: Data science concepts, published in Towards Data Science (2018)

1.7 Αναλυτική των Δεδομένων στον τομέα της Υγείας

Η προσφορά της τεχνολογίας στις πληροφορίες υγείας είναι μεγάλη. Η πρόσβαση σε δεδομένα φαίνεται να υποστηρίζεται από την τάση για υιοθέτηση ηλεκτρονικών φακέλων υγείας ασθενών, την αύξηση διαφόρων εφαρμογών, τη μείωση του κόστους τόσο απόκτησης όσο και αποθήκευσης αυτών των δεδομένων. Τα δεδομένα βρίσκονται επίσης σε εφαρμογές κοινωνικής δικτύωσης ή σε εφαρμογές κινητών συσκευών.

Η αναλυτική αυτών των πληροφοριών μπορεί να βοηθήσει το ιατρικό προσωπικό σε εγκυρότερες αποφάσεις περίθαλψης ή να δώσει χρόνο προετοιμασίας στο προσωπικό πριν την άφιξη ενός επείγοντος περιστατικού στο νοσοκομείο. Η χρήση αναλυτικών συστημάτων μπορεί να εντοπίσει τάσεις και μοτίβα στη φροντίδα και τα αποτελέσματα ασθενών με τον εντοπισμό συσχετίσεων ή ανισοτήτων. Και το βασικό είναι ότι τα συστήματα μπορούν να αξιοποιήσουν το μεγάλο όγκο δεδομένων, δηλαδή αυτές τις πληροφορίες που προηγουμένως αποτυπωνόντουσαν σε απλό χαρτί με ελεύθερο κείμενο.

Η ανάλυση της υγείας ξεκινά με την συλλογή, την οργάνωση και τη διαχείριση των δεδομένων υγείας και των ιατρικών δεδομένων και υποστηρίζεται από τέσσερα στάδια της Αναλυτικής.

Αρχικά η *περιγραφική ανάλυση (descriptive analytics),* η οποία περιγράφει τα δεδομένα ως έχουν και χρησιμοποιεί πολύ την οπτικοποίηση. Τέτοια μοντέλα συγκεντρώνουν, χαρακτηρίζουν και ταξινομούν τα δεδομένα μετατρέποντας τα σε χρήσιμες πληροφορίες. Έτσι οι ενδιαφερόμενοι μπορεί να λάβουν περιλήψεις δεδομένων, ουσιαστικά διαγράμματα και αναφορές που θα τους βοηθήσουν σε ταχύτερες και βέλτιστες αποφάσεις πχ «πόσους ασθενείς φρόντισε η κάθε εγκατάσταση» ή «ποια ήταν τα έσοδα και τα έξοδα μία συγκεκριμένη περίοδο» κλπ.

Η *προγνωστική ανάλυση (predictive analytics),* η οποία εξετάζει ιστορικά ή συνοπτικά δεδομένα υγείας αναζητώντας τα μοτίβα ώστε να προεκτείνει την απόδοση του παρελθόντος σε προσπάθεια πρόβλεψης του μέλλοντος. Εδώ οι ενδιαφερόμενοι μπορούν να πάρουν αποφάσεις όπως «ποια φάρμακα να χρησιμοποιήσουν σε συγκεκριμένες ομάδες ασθενών» ή «πως να κατανείμουν τους πόρους τους για να διασφαλίσουν αποτελεσματική παροχή υπηρεσιών υγείας» κλπ.

Η *κανονιστική ανάλυση (prescriptive analytics),* η οποία χρησιμοποιεί την υγειονομική και ιατρική γνώση εκτός από τα δεδομένα, όταν τα προβλήματα περιλαμβάνουν πάρα πολλές επιλογές ή εναλλακτικές, ώστε ένας πάροχος να εξετάσει αποτελεσματικά τις περιγραφικές και προγνωστικές αναλύσεις. Λειτουργεί «κανονιστικά» ως προς το τι θα έπρεπε να είναι η απάντηση. Εδώ οι ενδιαφερόμενοι μπορούν να εξετάσουν «εναλλακτικές επιλογές σταθμίζοντας πλεονεκτήματα και μειονεκτήματα» ή «να καθορίσουν μέγιστη δοσολογία φαρμάκου που είναι αποτελεσματική για τη μέγιστη θεραπεία» κλπ.

Η *αναλυτική της ανακάλυψης (cognitive analytics)*, στην οποία οι εφαρμογές μαθαίνουν συσχετισμούς για να αξιοποιήσουν αυτή τη γνώση για ανακάλυψη φαρμάκων, ασθενειών, εναλλακτικών θεραπειών. Οι ενδιαφερόμενοι μπορούν με χρήση προσομοιώσεων στον υπολογιστή «να ενισχύσουν τις κλινικές δοκιμές και να επιταχύνουν τις μελέτες αποτελεσματικότητας νέων φαρμάκων» ή «να μελετήσουν την πορείας μίας νέας επιδημίας» κλπ.

1.8 Τα Μεγάλα Δεδομένα στον τομέα της Υγείας

1.8.1 Τρόπος Συλλογής Δεδομένων

Σύμφωνα με έρευνα της Melanie Swan (2016), η προέλευση των Μεγάλων Δεδομένων στο χώρο της υγείας μπορεί να χωριστεί σε τρεις κατηγορίες (Σχήμα 1-2).

Σχήμα 1-2: Κατηγορίες προέλευσης Μεγάλων Δεδομένων, Melanie Swan 2016

- Τα «Omics» δεδομένα που προέρχονται από δεδομένα γονιδιομάτων, μικροβιομάτων, ασθενειών και περιβαλλοντικών παραγόντων και είναι σημαντικά για την κατανόηση της πορείας της ασθένειας με σκοπό την αντιμετώπισή της
- Τα Παραδοσιακά δεδομένα που προέρχονται από φακέλους υγείας των ασθενών, συνταγογραφήσεις, εργαστηριακά αποτελέσματα, δημογραφικά στοιχεία και είναι σημαντικά για την κατανόηση των αποτελεσμάτων της ασθένειας και της βελτιστοποίησης της παρεχόμενης περίθαλψης
- Τα Ποσοτικά δεδομένα που προέρχονται από καταγραφή μετρήσεων σε προσωπικές συσκευές ατόμων όπως κινητά ή έξυπνα ρολόγια για καθημερινές συνήθειες και δραστηριότητες που έχουν, από συσκευές που έχουν τοποθετηθεί σε ασθενείς για

καταγραφή αυτών των μετρήσεων ή από αισθητήρες και είναι σημαντικά στην κατανόηση του τρόπου ζωής του ασθενή

Σύμφωνα με την έρευνα του Κωνσταντάκη (2020) μία κατηγοριοποίηση της προέλευσης των δεδομένων υγείας είναι:

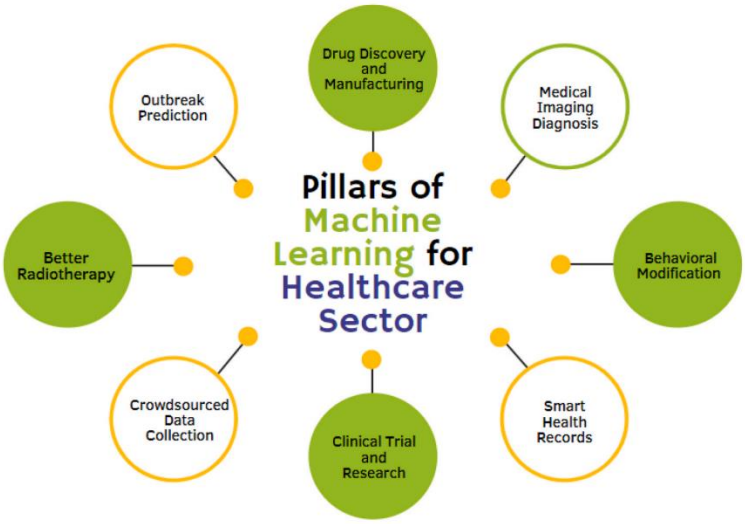
- Νοσοκομειακά συστήματα κλινικών πληροφοριών με κύρια πηγή τον ηλεκτρονικό φάκελο υγείας των ασθενών
- Πληροφοριακά συστήματα ηλεκτρονικής υγείας με πηγές την ηλεκτρονική συνταγογράφηση, το διαμοιρασμό πληροφοριών πρωτοβάθμιων μονάδων με τα νοσοκομειακά συστήματα και τη χρήση τεχνολογικών προϊόντων υγείας
- Νοσοκομειακά πληροφοριακά συστήματα οικονομικής διαχείρισης
- Βάσεις επιστημονικών δεδομένων και μητρώα μελετών που περιλαμβάνουν κλινικά και εργαστηριακά δεδομένα με ελεύθερες ή περιορισμένες προσβάσεις
- Δεδομένα βιοϊατρικών μετρήσεων και ιατρικών συσκευών που περιλαμβάνουν τις καταγραφές και μετρήσεις από φορητές συσκευές όπως κινητό, ρολόι, συσκευή holter, καταγραφής ύπνου κλπ. μέσω των οποίων αντλούνται πληροφορίες για τον τρόπο ζωής
- Βάσεις γενετικών δεδομένων για καταγραφή γονιδιακών εξετάσεων που συνδυαστικά με την κλινική εικόνα βοηθούν σε εξατομικευμένες θεραπείες
- Αναζητήσεις διαδικτύου, όπου το 2008 οι επιστήμονες της Google παρατήρησαν αύξηση αναζητήσεων για συμπτώματα γρίπης και λαμβάνοντας την τοποθεσία προέβλεψαν πιθανές επιδημίες σε κάποιες περιοχές
- Μέσα κοινωνικής Δικτύωσης που βοηθούν στην πρόβλεψη τάσεων στην υγεία
- Δεδομένα Ασφαλιστικών Οργανισμών
- Δημογραφικές και επιδημιολογικές εγγραφές που καταγράφονται από οργανισμούς με μελλοντικό στόχο τη διασύνδεσή τους με το σύστημα ηλεκτρονικής συνταγογράφησης και τον ηλεκτρονικό φάκελο
- Σημειώσεις και εξετάσεις έντυπης μορφής

1.8.2 Οφέλη από την αξιοποίηση Των Μεγάλων Δεδομένων στον τομέα της Υγείας

Τα Μεγάλα Δεδομένα στον τομέα της Υγείας σε συνδυασμό με μεθόδους Μηχανικής Μάθησης μπορούν να προσφέρουν μεγαλύτερη ταχύτητα και ακρίβεια στις ιατρικές υπηρεσίες, ειδικά σε συστήματα υγείας που ο αριθμός προσωπικού και οι οικονομικοί πόροι δεν είναι οι επιθυμητοί. Τεχνικές ανάλυσης των δεδομένων μπορούν να εντοπίσουν πρώιμα στάδια ή ρυθμούς εξάπλωσης πανδημιών, όπως είδαμε το 2020 να συμβαίνει με τον κορονοϊό. Αυτές οι τεχνικές λειτουργούν επικουρικά στα συστήματα υγείας καθώς δεν χρειάζεται να εξαντληθούν

οι διαθέσιμοι πόροι αλλά προσομοιώνουν καταστάσεις και δίνουν περιθώριο περιορισμού του φαινομένου ή προετοιμασίας για αυτό.

Από έρευνα των Javaid et al. (2022) εντοπίζονται σημαντικά οφέλη της ανάλυσης των Μεγάλων Δεδομένων στην υγειονομική περίθαλψη τα οποία βελτιώνουν την αποτελεσματικότητα των νοσοκομείων περιορίζοντας παράλληλα κόστη και τοποθετούνται σε οκτώ πυλώνες στο Σχήμα 1-3 που ακολουθεί.



Σχήμα 1-3: Πυλώνες μηχανικής μάθησης στην παροχή υπηρεσιών υγείας, M. Javaid et al 2022

Η Μηχανική Μάθηση αντλεί πληροφορίες και εκπαιδεύεται από Αρχεία Δεδομένων είτε αυτά προέρχονται από καταγραφή ηλεκτρονικού φακέλου ασθενούς είτε πλέον από σύγχρονες συσκευές όπως κινητά, ρολόγια αξιοποιώντας την κάθε πληροφορία προς όφελος του ασθενούς. Από αναλύσεις αυτών των πληροφοριών μέσω μοντέλων Μηχανικής Μάθησης διευκολύνεται η διάγνωση του γιατρού ή αυτές οι αναλύσεις λειτουργούν επικουρικά.

Πλέον τα δεδομένα που συλλέγονται εντοπίζουν μοτίβα στον τρόπο ζωής του ασθενούς ή μπορεί και να προλάβουν επιδημίες. Υπάρχουν μέθοδοι πρόβλεψης ασθένειας από έλεγχο εικόνων, εντοπισμός τύπων καρκίνου όπως καρκίνος του αίματος ή εντοπίζουν το διαβήτη. Συνεισφέρει στην Ακτινολογία καθώς με τους αλγορίθμους αναλύονται εικόνες και η εκπαίδευση από μία ευρεία γκάμα εικόνων βελτιώνει τη διαδικασία διάγνωσης. Οι Ρομποτικές επεμβάσεις είναι άλλη μία συνεισφορά της Μηχανικής Μάθησης στην υγειονομική περίθαλψη. Οι κλινικές μελέτες έχουν βοηθηθεί αφού δίδεται στους επιστήμονες η δυνατότητα να συλλέξουν ιδανικούς υποψήφιους και να αναλύουν τις πληροφορίες σε πραγματικό χρόνο, να εντοπίζουν τα προβλήματα, απροσδόκητες τάσεις και να καταλήγουν σε εξοικονόμηση κόστους και κυρίως τα φάρμακα να μελετώνται και να καταλήγουν στους ασθενείς ταχύτερα.

Κατατάσσοντας αυτά τα οφέλη σε κατηγορίες φαίνεται ότι ευνοούνται τέσσερα ενδιαφερόμενα μέρη, δηλαδή οι μονάδες υγείας, οι επαγγελματίες υγείας, η επιστήμη και οι ασθενείς.

- **Οφέλη για τη διοίκηση των μονάδων υγείας.**

Αποτελεσματικότερη διαχείριση των διαθέσιμων πόρων είτε οικονομικών είτε ανθρώπινων και αποτελεσματικότερη παροχή φροντίδας στους ασθενείς. Η αξιοποίηση των στατιστικών μοτίβων βοηθάει στη λήψη των αποφάσεων περίθαλψης και φαρμακευτικής αγωγής. Τα στοιχεία αντλούνται από ηλεκτρονικούς φακέλους ασθενή, από σύστημα ηλεκτρονικής συνταγογράφησης, από καταγραφή σε προϊόντα βιοτεχνολογίας. Δίνει τη δυνατότητα σύνδεσης των αποτελεσμάτων των Μονάδων υγείας και των οικονομικών τους δαπανών ώστε να υπάρχει αξιολόγηση. Με την ανάλυση των δεδομένων από διάφορες πηγές αξιολογούνται παράλληλα ιατρικά, διαχειριστικά, οικονομικά και επιδημιολογικά δεδομένα και μπορούν να προχωρήσουν σε προσαρμογές στο σύστημα υγείας. Από τα δεδομένα δίνεται η δυνατότητα εντοπισμού μοτίβων που μπορεί να οδηγήσουν σε επιδημίες και υπάρχει η ευκαιρία παρακολούθησης της πορείας ή μετριασμού της εξάπλωσης.

- **Οφέλη για τους επαγγελματίες υγείας.**

Οι ηλεκτρονικοί φάκελοι ασθενών και η ηλεκτρονική συνταγογράφηση βοηθούν στην ανταλλαγή δεδομένων και στην λήψη αποφάσεων. Έτσι, υπάρχει όλη η ιστορικότητα προς όφελος του ασθενή αλλά και του ιατρού ανά πάσα στιγμή. Τα δεδομένα αυτά ,όμως, μπορούν να αξιοποιηθούν και από κλινικές μελέτες και έρευνες.

- **Οφέλη για την Επιστήμη.**

Συμβάλει στην ιατρική επιστήμη και έρευνα βοηθώντας τόσο σε μελέτες όσο και σε αποφάσεις. Αναπτύσσονται μοντέλα πρόγνωσης της εξέλιξης ασθενειών. Μεγάλη συμβολή στην κλινική έρευνα για ανάπτυξη νέων φαρμάκων ταχύτερα με μείωση κόστους. Κατανόηση παραγόντων που ενεργοποιούν χρόνιες ασθένειες τα οποία το 2020 απασχολούσαν άνω του 40% του ενήλικου πληθυσμού και κατείχαν άνω του 70% των νοσημάτων.

- **Οφέλη για τον Ασθενή.**

Η έγκυρη πληροφόρηση, ισότητα στην πρόσβαση, αποτελεσματικότερη παροχή υπηρεσιών, διευκολύνσεις στις θεραπείες του ή μέχρι και αυτοματοποιημένες υπενθυμίσεις για επισκέψεις και εξετάσεις σε γιατρούς.

Κ Ε Φ Α Λ Α Ι Ο 2

Βιβλιογραφική Ανασκόπηση σε μελέτες βελτιστοποίησης των Υπηρεσιών Υγείας

2.1 Εισαγωγή

Εντοπίστηκαν ερευνητικές εργασίες οι οποίες εφάρμοσαν Αναλυτική των δεδομένων και Μηχανική Μάθηση με σκοπό να προτείνουν τρόπους βελτιστοποίησης της διοίκησης υπηρεσιών υγείας.

Παρακάτω παρουσιάζονται έρευνες για τρία θέματα που απασχολούν τη δημόσια διοίκηση υγείας και είναι η Εμφάνιση ή Απουσία ενός ασθενή από το προγραμματισμένο του ραντεβού σε εξωτερικό ιατρείο, η Επανεμφάνιση ενός ασθενή στο Νοσοκομείο εντός 30 ημερών από την ημέρα εξιτηρίου και ο προγραμματισμός κράτησης μίας χειρουργικής αίθουσας νοσοκομείου.

2.2 Εμφάνιση ή Απουσία ενός ασθενή στο προγραμματισμένο του ραντεβού σε εξωτερικά ιατρεία

Οι πάροχοι υγειονομικής περίθαλψης αντιμετωπίζουν συνεχώς την ανάγκη μείωσης του κόστους αλλά εξασφαλίζοντας υπηρεσίες υψηλής ποιότητας. Η σύγχρονη τεχνολογία ανάλυσης δεδομένων και η υιοθέτηση συστημάτων προγραμματισμού μπορούν να συμβάλλουν στην αποτελεσματική αξιοποίηση του προσωπικού και των λοιπών πόρων με παράλληλη μείωση του κόστους, αλλά κακή διαχείριση συστημάτων μπορεί να επιφέρει δυσαρέσκεια ασθενών.

Ανάλυση Οικονομικού Κόστους μη εμφάνισης ασθενών σε εξωτερικά ιατρεία Ισπανίας (2017)

Η μελέτη των Mesa et al. (2017) είχε σκοπό να εκτιμηθεί το οικονομικό κόστος των απουσιών από τέσσερα εξωτερικά ιατρεία της Ισπανίας. Από τα αποτελέσματα της μελέτης φαίνεται η δαπάνη των απουσιών που αντιστοιχεί στο 13.8% των ραντεβού να ανέρχεται σε

περίπου 3.300.000 €. Τα δεδομένα συλλέχθηκαν από το νοσοκομείο Costa del Sol, από το High Resolution Hospital of Benalmadena και από το High Resolution Center of Mijas.

Ακολούθησε μελέτη περίπτωσης όπου οι ομάδες ήταν αυτοί που παρευρέθηκαν στα ραντεβού τους και αυτοί που δεν παρευρέθηκαν. Συγκεκριμένα, στο πρώτο νοσοκομείο η απουσία από το ραντεβού έφτασε στο 14,2% που αντιστοιχεί σε 256.377 ραντεβού, από το δεύτερο νοσοκομείο έφτασε 12,2% που αντιστοιχεί σε 44.848 ραντεβού και από το τρίτο νοσοκομείο έφτασε 13,5% που αντιστοιχεί σε 99.536 ραντεβού.

Για την ανάλυση κόστους που πραγματοποιήθηκε κρίθηκε σκόπιμο να συμπεριλάβουν δαπάνες ιατρικού προσωπικού, αναλωσίμων, εξοπλισμού και φαρμάκων, αλλά και δαπάνες για ασφάλεια, συντήρηση, διοίκηση. Κατά την ανάλυση των δεδομένων, ελέγχθηκε η κανονικότητα των ομάδων με τεστ Kolmogorov-Smirnov (παρουσία/απουσία ασθενούς σε ραντεβού), ενώ ελέγχθηκε και η διαφορά των μέσων όρων των παραπάνω ομάδων μέσω t-test στην κανονική κατανομή και Mann-Whitney U/ Wilcoxon W στην μη κανονική κατανομή. Επίσης, ελέγχθηκε η συσχέτιση Pearson και Spearman μεταξύ ηλικίας και κόστους.

Το υψηλότερο συνολικό μέσο μοναδιαίο κόστος παρατηρήθηκε σε ραντεβού πεπτικής ειδικότητας (134,22 € με τ.α.=66.54). Επίσης, το 57,8% των ασθενών που απουσίασαν από το ραντεβού τους ζήτησαν να το προγραμματίσουν σε νέα ημέρα/ ώρα, κάτι το οποίο συνεπάγεται πρόσθετη αύξηση κόστους. Όσον αφορά τον έλεγχο σχέσης μεταξύ ηλικίας ασθενούς και κόστους δεν βρέθηκε συσχέτιση. Στους άνδρες ο μέσος όρος έφτασε το 108,8€ με τ.α.=77,1 , ενώ στις γυναίκες ήταν στο 103,1€ με τ.α.=73,4 και παρουσίασαν και οι δύο ομάδες στατιστική σημαντικότητα ($\rho < 0,001$).

<u>Προγνωστικές Αναλύσεις AI στη διαχείριση μη εμφάνισης σε ραντεβού μαγνητικής τομογραφίας εξωτερικού ιατρείου (2020)</u>

Στην έρευνα τους οι Chong et al. (2020) χρησιμοποίησαν μεθόδους μηχανικής μάθησης για την πρόβλεψη μη εμφάνισης ασθενούς σε ραντεβού και τη μείωση αυτών των περιπτώσεων.

Χρησιμοποίησαν δεδομένα για πάνω από 30.000 ραντεβού μεταξύ 2016 και 2018. Το ποσοστό μη εμφάνισης ήταν 17,4%. Το μοντέλο πρόβλεψης που επιλέχτηκε να αναπτυχθεί μεταξύ άλλων ήταν το XGBoost, ένας αλγόριθμος βασισμένος σε δέντρα αποφάσεων που χρησιμοποιεί πλαίσιο ενίσχυσης κλίσης.

Ως μέτρο παρέμβασης στην απουσία από το ραντεβού έθεσαν για διάστημα 6 μηνών την τηλεφωνική υπενθύμιση στο 25% των ασθενών που εμφάνιζαν την υψηλότερη πιθανότητα. Από την ημέρα εφαρμογής του προβλεπτικού μοντέλου και σε συνδυασμό με τις τηλεφωνικές υπενθυμίσεις παρουσιάστηκε βελτίωση κατά 17,2% (με στατιστική σημαντικότητα $\rho<0,001$). Εκείνο το διάστημα συλλέχθηκαν επιπλέον 1.080 δεδομένα που χρησιμοποιήθηκαν ως test set.

Πιο συγκεκριμένα, οι επιδόσεις του μοντέλου πρόβλεψης στα δεδομένα εκπαίδευσης ήταν ROC AUC=0.746 , F1 score=0.708 , Precision=0.606 και Recall=0.852 . Για το Test Set προέκυψε AUC=0.738 , F1 score=0.721 , Precision=0.605 και Recall=0.893. Από τους 10 παράγοντες που φάνηκε να επηρεάζουν περισσότερο στην απουσία από το ραντεβού πιο σημαντικοί εμφανίζονται η ηλικία και ο χρόνος αναμονής από το ραντεβού (σε ημέρες) και ακολουθούν οι φορές επαναπρογραμματισμού, το φύλο (άνδρες), κάποιες γεωγραφικές περιοχές και κάποιες μέρες εβδομάδας.

2.3 Επανεμφάνιση ενός ασθενή στο Νοσοκομείο εντός 30 ημερών από την ημέρα εξιτηρίου

Η ικανότητα πρόβλεψης επανεισαγωγών ασθενών δίνει στα Νοσοκομεία τη δυνατότητα έγκαιρης παρέμβασης ώστε να αποφευχθούν μελλοντικά απειλητικά περιστατικά υγείας μειώνοντας έτσι τον κίνδυνο για τους ασθενείς, ενισχύοντας την ποιότητα υπηρεσιών υγείας και μειώνοντας τα κόστη για το σύστημα υγείας. Η εφαρμογή μεθόδων Μηχανικής Μάθησης με δεδομένα από τα συστήματα των μονάδων θα μπορούσε να εντοπίσει πιθανούς ασθενείς με υψηλές πιθανότητες να υποτροπιάσουν. Σε αρκετές χώρες το ποσοστό επανεισδοχής ασθενών σε νοσοκομείο εντός 30 ημερών από την ημέρα εξιτηρίου λειτουργεί ως δείκτης αξιολόγησης των μονάδων και επιβάλλονται οικονομικά πρόστιμα σε περίπτωση υπέρβασης της τιμής του δείκτη.

Πρόβλεψη Νοσοκομειακών Επανεισαγωγών με χρήση Τεχνητής Νοημοσύνης (2022)

Στην έρευνα των Michailidis et al. (2022) χρησιμοποιήθηκαν τέσσερα μοντέλα τεχνητής νοημοσύνης για να προβλέψουν επανεισαγωγές σε νοσοκομείο και συγκεκριμένα SVM με πυρήνα RBF και γραμμικό πυρήνα, Balances RF και Weighted RF. Τα ανωνυμοποιημένα δεδομένα συλλέχθηκαν από το σύστημα διαχείρισης ασθενών, το σύστημα επιχειρηματικών πληροφοριών και το σύστημα εργαστηριακών πληροφοριών του Σισμανόγλειου Νοσοκομείου Κομοτηνής για τα έτη 2018 και 2019 και περιλάμβαναν μεταβλητές διοικητικές/δημογραφικές, ιατρικές/κλινικές και δεδομένα για τη λειτουργική κατάσταση του νοσοκομείου. Στο τελικό σύνολο δεδομένων που κατέληξαν μετά την προεπεξεργασία, το 15,6% των ασθενών επανεισήχθησαν εντός 30 ημερών από την έξοδο τους. Τα μοντέλα τους έφτασαν ακρίβεια πρόβλεψης 88% και ακρίβεια επανεισαγωγής 70% με το Balanced Random Forest να πετυχαίνει την καλύτερη απόδοση εντοπισμού επανεισδοχών. Σύμφωνα με τα αποτελέσματά τους, οι δύο πιο σημαντικές μεταβλητές είναι η διάγνωση εισαγωγής και εξιτηρίου και η τρίτη σημαντικότερη μεταβλητή είναι το ποσοστό πληρότητας της κλινικής. Φαίνεται, λοιπόν, ότι υπάρχουν ασθένειες που δίνουν μεγαλύτερες πιθανότητες

επανεισδοχής των ασθενών όπως επίσης υπάρχουν και λειτουργικοί παράγοντες του νοσοκομείου που το ενισχύουν. Η πληρότητα αναδείχθηκε ο σημαντικότερος λειτουργικός παράγοντας σύμφωνα με τα μοντέλα τους, αλλά υψηλή θέση στην κατάταξη έχει και ο αριθμός γιατρών που εργάζονται την ημέρα εξιτηρίου όπως και ο αριθμός νοσηλευτικού προσωπικού.

Πρόβλεψη Επανεισαγωγής ασθενών με οδοντιατρικά προβλήματα εντός 90 ημερών με χρήση Μηχανικής Μάθησης (2021)

Στην έρευνα τους οι Li et al. (2021) εξετάζουν την επανεισδοχή ασθενών με οδοντιατρικά θέματα, χρησιμοποιώντας σύνολο δεδομένων από το Nationwide Readmissions Database (NRD) και εφαρμόζοντας 5 αλγορίθμους μηχανικής μάθησης που περιλαμβάνουν Decision Trees, Logistic Regression, Support Vector Machine, K-Nearest Neighbors και Artificial Neural Networks.

Οι επανεισαγωγές έφταναν περίπου το 19% των περιπτώσεων του συνόλου δεδομένων που εξετάστηκαν και οι παράγοντες που επηρεάζουν περισσότερο φαίνεται να είναι οι συνολικές χρεώσεις, ο αριθμός των διαγνώσεων, η ηλικία, οι χρόνιες παθήσεις, η διάρκεια παραμονής στο νοσοκομείο, ο αριθμός των διαδικασιών, ο κύριος πληρωτής και η σοβαρότητα της ασθένειας ανάμεσα σε παράγοντες δημογραφικούς, κοινωνικοοικονομικούς και ιατρικούς. Η απόδοση των μοντέλων ήταν παρόμοια με το ANN να ξεπερνά ελαφρώς τα υπόλοιπα (AUC=0,743), η Logistic Regression ακολούθησε (AUC=0,738), τα Decisions Trees (AUC=0,721), SVM (AUC=0,679) και K-NN (AUC=0,623). Τη μεγαλύτερη Ευαισθησία σημείωσε ο Decision Trees (0,734) και ακολούθησε ο ANN (0,719) ενώ η Ακρίβεια ήταν παρόμοια ανάμεσα σε SVM (0,667) και σε ANN (0,665).

Καταλήγουν στην έρευνα τους ότι θα μπορούσαν να αποταμιευθούν έως και 500 εκατομμύρια δολάρια εάν όλες οι περιπτώσεις επανεισαγωγής μπορούσαν να αποφευχθούν σε όλες τις πολιτείες των ΗΠΑ ενώ με μία άλλη πιο συντηρητική εκτίμηση εάν 1 στις 4 επανεισδοχές μπορούσε να αποφευχθεί θα εξοικονομούνταν 100 δολάρια. Για το σύνολο δεδομένων τους εάν είχαν αποφευχθεί οι επανεισαγωγές που αντιστοιχούν σε 1.746 ασθενείς θα είχαν εξοικονομηθεί από το νοσοκομείο 103.272.408 εκατομμύρια δολάρια.

2.4 Προγραμματισμός κράτησης μίας χειρουργικής αίθουσας νοσοκομείου

Οι μεγάλες Λίστες Αναμονής για τα χειρουργεία φαίνεται να προβληματίζουν τις χώρες, ανάμεσα τους και η Ελλάδα η οποία κρίνει ότι θα χρειαστούν μέτρα για τη μείωση τους. Σύμφωνα με δημοσίευση του 2022 (Ευθυμιάδου Δ., ΕΘΝΟΣ), το Υπουργείο Υγείας της Ελλάδας έκρινε ότι θα χρειαστεί να προχωρήσεις σε έκτακτο σχέδιο ώστε να μειώσει τις λίστες

χειρουργείων στα δημόσια νοσοκομεία. Η πανδημία έφερε ακόμη μεγαλύτερη αναμονή στη λίστα, καθώς υπήρχαν επεμβάσεις που αναβλήθηκαν για πάνω από δύο χρόνια. Τα τρία μέτρα που πιστεύει ότι θα μειώσουν το μέγεθος του προβλήματος είναι η επανεξέταση των περιστατικών που αναγράφονται στις λίστες για να διαπιστωθεί εάν έχουν ακόμη ανάγκη την επέμβαση ή έχουν αναζητήσει λύση στον ιδιωτικό τομέα, η συνεργασία με το Υπουργείο Ψηφιακής Διακυβέρνησης για να δημιουργηθεί μία ενιαία λίστα αναμονής χειρουργικής αίθουσας ώστε να αποφευχθούν οι διπλοεγγραφές και τέλος η μεταφορά ασθενών σε ιδιωτικές κλινικές με τις οποίες το κράτος θα διαπραγματευτεί την οικονομική συμφωνία. Πέρα από αυτά τα μέτρα, μελετώνται τρόποι για καλύτερη κατανομή του χρόνου προγραμματισμού των αιθουσών και παρουσιάζονται παρακάτω.

Η χρήση Μηχανικής Μάθησης βοηθάει σε αποδοτικότερο προγραμματισμό χειρουργικών αιθουσών την εποχή του Covid-19 (2020)

Στην έρευνα της οι Rozario et al. (2020) παρουσιάζουν μοντέλο βελτιστοποίησης αποτελεσματικότερου χρόνου κράτησης αίθουσας χειρουργείου, το οποίο μπορεί να οδηγήσει σε μείωση των υπερωριών του νοσηλευτικού προσωπικού κατά 21% με το θεωρητικό οικονομικό όφελος να αντιστοιχεί σε 469.000€ σε βάθος τριετίας.

Αυτή η έρευνα στηρίχθηκε σε ανωνυμοποιημένα δεδομένα του νοσοκομείου Oakville Trafalgar Memorial του Οντάριο που διαθέτει 10 αίθουσες χειρουργείων. Τα δεδομένα αφορούσαν 36 μήνες από το 2017 έως το 2019 και περιλάμβαναν 10.553 περιπτώσεις από 15 χειρουργούς διαφορετικών τμημάτων.

Η μέθοδος που ακολουθούν για πρόβλεψη χρόνου επεμβάσεων είναι ο μέσος όρος των προηγούμενων 10 φορών της συγκεκριμένης επέμβασης, το οποίο οδηγεί σε υπέρβαση χρόνου κράτησης στο 50% των περιστατικών. Για την κράτηση των αιθουσών χρησιμοποιείται το σύστημα κρατήσεων «PICIS OR» του οποίου τα δεδομένα αξιοποιήθηκαν μέσω Python και συνδυαστικά με τη σουίτα βελτιστοποίησης «OR-Tools» εισήγαγαν τους επιθυμητούς χρόνους βελτιστοποίησης για να επιτύχουν ιδανικούς χρόνους επεμβάσεων και κατόπιν αξιολόγησαν το μοντέλο τους στα δεδομένα για να συγκρίνουν εάν είχαν αλλάξει τα αποτελέσματα.

Όρισαν το κόστος της Υπερωρίας και το κόστος της Ελλείπουσας ώρας με διαφορετικά βάρη κατάλληλα για το συγκεκριμένο νοσοκομείο. Υπερωρία χειρουργείου θεωρούν εάν το άθροισμα του πραγματικού χρόνου συν τις αλλαγές υπερβαίνουν τον προγραμματισμένο χρόνο περιστατικού για μία δεδομένη ημέρα, ενώ Ελλιπή θεωρούν το χρόνο χειρουργείου εάν ο πραγματικός χρόνος λήξης είναι περισσότερος από 15 λεπτά νωρίτερα του προγραμματισμένου για μία δεδομένη ημέρα. Στόχος της βελτιστοποίησης τους είναι η ελαχιστοποίηση τόσο των περιπτώσεων Υπερωριών όσο και των περιπτώσεων Ελλείπουσας ώρας και να φτάσουν σε ποσοστό 80% των περιπτώσεων τους προγραμματισμένους χρόνους.

Τα αποτελέσματα τους έδειξαν ότι εάν είχαν χρησιμοποιηθεί οι χρόνοι προγραμματισμού του μοντέλου μηχανικής μάθησης θα είχαν Υπερωρίες στο 27% του χρόνου και Υποαπασχόληση στο 18% του χρόνου έναντι του 48% του χρόνου και 37% του χρόνου αντίστοιχα όπως συνέβη στην πράξη χωρίς το μοντέλο. Έφτασαν σε ποσοστό 55% των περιπτώσεων να ολοκληρωθεί το χειρουργείο εντός του προγραμματισμένου χρόνου σε αντίθεση με την πράξη όπου το νοσοκομείο αυτό το χρόνο μόνο στο 15% των περιστατικών. Υπολογίστηκε ότι το 97% του όγκου των περιστατικών που είχαν αναβληθεί και συσσωρευτεί το προηγούμενο διάστημα θα είχε ολοκληρωθεί στον ίδιο συνολικό χρόνο λεπτών χειρουργείου, οδηγώντας σε εξοικονόμηση κόστους στο βάθος τριετίας.

Βελτιστοποίηση προγραμματισμού χειρουργικών αιθουσών με Προβλετικά μοντέλα (2022)

Στη μελέτη τους οι Abbou et al. (2022), εκπαιδεύουν ένα προγνωστικό μοντέλο για τη διάρκεια των χειρουργικών επεμβάσεων με σκοπό τη βέλτιστη αξιοποίηση της διαθεσιμότητας των χειρουργικών αιθουσών, τη βελτίωση της παραγωγικότητας και κατ' επέκταση μείωση κόστους και βελτίωση εξυπηρέτησης ασθενών.

Τα δεδομένα που συλλέχθηκαν ήταν κλινικά και διοικητικά από δύο μεγάλα γενικά και δημόσια νοσοκομεία του Ισραήλ. Αφορούσαν την περίοδο από Δεκέμβρη του 2009 μέχρι Μάιο του 2020. Τα δεδομένα μέχρι και το 2018 χρησιμοποιήθηκαν ως δείγμα εκπαίδευσης ενώ τα υπόλοιπα ως δείγμα επικύρωσης του μοντέλου που ανέπτυξαν.

Σύγκριναν την εμπειρική μέθοδο που εφάρμοζαν τα νοσοκομεία για τη διαχείριση χειρουργικών αιθουσών με ένα αλγοριθμικό μοντέλο. Ο προγραμματισμός των χειρουργείων στηρίζεται στην εκτίμηση των γιατρών για τον απαιτούμενο χρόνο, αλλά αποδεικνύεται ότι αυτό δεν είναι επαρκές. Το μοντέλο των Abbou et al., που χρησιμοποίησε τεχνικές Μηχανικής Μάθησης με τον αλγόριθμο XGBoost και 5-fold-cross validation, πέτυχε καλύτερες επιδόσεις και εξήγησε το 70% της διασποράς στη διάρκεια κράτησης των αιθουσών.

Το μοντέλο τους επικεντρώθηκε στις κλινικές παραμέτρους των ασθενών όπως προηγούμενες διαγνώσεις, συνταγογραφήσεις, καπνιστική συνήθεια, εργαστηριακές εξετάσεις καθώς και στην συσσωρευμένη εμπειρία των χειρουργών σε διάστημα επταετίας. Από αξιολόγηση του F-score προέκυψε ότι οι σημαντικότεροι παράγοντες που συμβάλλουν στο χρόνο κράτησης των χειρουργικών αιθουσών είναι η εμπειρία του χειρουργού, η ηλικία του ασθενούς, ο αριθμός των χειρουργών που έχουν αναλάβει την επέμβαση, ο αριθμός των διαγνώσεων, ο αριθμός των φαρμάκων και ο αριθμός προγραμματισμένων επεμβάσεων. Ανάμεσα στις μετρικές που αξιολόγησαν ήταν το Μέσο Απόλυτο Σφάλμα (MAE) όπου είναι ο μέσος όρος των απόλυτων σφαλμάτων και προέκυψε ότι το XGBoost μοντέλο είχε καλύτερες επιδόσεις (21,5 και 25,3 στο κάθε ένα από τα νοσοκομεία έναντι 25,4 και 28,7 του μοντέλου που δείχνει τον τρόπο που δουλεύουν μέχρι τώρα).

Καταλήγουν ότι η μεγιστοποίηση χρόνου χρήσης των διαθέσιμων αιθουσών σε κάθε μονάδα μπορεί να φέρει όφελος στο σύστημα υγειονομικής περίθαλψης αφού βελτιώνεται η περίθαλψη ασθενών αλλά αποφορτίζει και το υπερωριακό σύστημα προσωπικού.

ΚΕΦΑΛΑΙΟ 3

Στατιστική και Μηχανική Μάθηση

3.1 Εισαγωγή

Τόσο η Στατιστική όσο και η Μηχανική Μάθηση έχουν κοινό στόχο, τη μάθηση από τα δεδομένα και επικεντρώνονται στην άντληση γνώσεων. Όμως έχουν διαφορετικό σκοπό και η κύρια διαφορά τους είναι ότι τα στατιστικά μοντέλα έχουν σχεδιαστεί για εξαγωγή συμπερασμάτων σχετικά με τις σχέσεις των μεταβλητών, ενώ τα μοντέλα μηχανικής μάθησης έχουν σχεδιαστεί για να κάνουν όσο το δυνατόν πιο ακριβείς προβλέψεις. Τα στατιστικά στοιχεία για τη μηχανική μάθηση αποτελούν σημαντικό εργαλείο στη μελέτη των ακατέργαστων δεδομένων για να αναγνωρίσουν μοτίβα ορατά ή αόρατα παρέχοντας σωστή κατεύθυνση κατά την ανάλυση και χρήση των δεδομένων.

3.2 Μηχανική Μάθηση

Δύο διαδεδομένα είδη μηχανικής μάθησης, τα οποία θα παρουσιαστούν και στη συνέχεια του κεφαλαίου, είναι η επιβλεπόμενη (supervised) και η μη επιβλεπόμενη (unsupervised) μηχανική μάθηση, όπου στην επιβλεπόμενη στόχος είναι να προβλεφθεί το αποτέλεσμα με βάση τις τιμές εισόδου, ενώ στη μη επιβλεπόμενη στόχος είναι η περιγραφή των συσχετίσεων και τα μοτίβα που διακρίνονται από τις τιμές εισόδου. Επιπλέον αυτών, διακρίνεται η ημί-εποπτευόμενη (semi-supervised) μάθηση και η ενισχυμένη (reinforcement) μάθηση, όπου στην ημί-εποπτευόμενη δημιουργείται ένα μικτό μοντέλο με στόχο την πρόβλεψη, αλλά χρησιμοποιώντας ένα μικρό μέρος των τιμών εισόδου για βελτίωση, ενώ στην ενισχυμένη εμφανίζεται έντονα η αλληλεπίδραση με το περιβάλλον και η ανάγκη για ανατροφοδότηση. Στη συνέχεια του κεφαλαίου αυτού η αναφορά που θα γίνει στις έννοιες και στις μεθόδους θα είναι σε επίπεδο βασικής κατανόησης και όχι εκτεταμένης ανάλυσης.
Η εποπτευόμενη μηχανική μάθηση χρησιμοποιείται σε προβλήματα:

- Παλινδρόμησης (Regression) που έχουν στόχο τη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών.
- Ταξινόμησης (Classification) που έχουν στόχο τη δημιουργία μοντέλων πρόβλεψης διακριτών τιμών.

Η μη εποπτευόμενη μάθηση χρησιμοποιείται σε προβλήματα:

- Ομαδοποίησης (Clustering) που έχουν στόχο τη δημιουργία του αριθμού των κλάσεων που θα ζητηθεί με βάση τα κοινά χαρακτηριστικά του συνόλου των δεδομένων.
- Κανόνων Συσχέτισης (Association Rules) που έχουν στόχο να ανακαλύψουν τις συσχετίσεις από το σύνολο των δεδομένων.
- Μείωσης των Διαστάσεων (Dimensionality Reduction) που έχουν στόχο να μειωθούν οι διαστάσεις του συνόλου των δεδομένων δημιουργώντας παράγοντες από σύνολα χαρακτηριστικών με υψηλές συσχετίσεις μεταξύ τους.

Μερικά παραδείγματα για τα παραπάνω είναι:

- Πρόβλεψη μεταβολής της αξίας αγοράς στο λιανικό εμπόριο ή στα κτηματομεσιτικά (Εποπτευόμενη Μάθηση)
- Πρόβλεψη εμφάνισης ή απουσίας ασθενούς σε προγραμματισμένο ραντεβού (Εποπτευόμενη Μάθηση)
- Πρόταση Προϊόντος ή Υπηρεσίας σε καταναλωτή βάσει τις αγοραστικές του συνήθειες (μη Εποπτευόμενη Μάθηση)
- Στοχευμένες διαφημίσεις ή προτάσεις σε χρήστες μέσων κοινωνικής δικτύωσης με βάση τα ενδιαφέροντα τους και τη συμπεριφορά τους (μη Εποπτευόμενη Μάθηση)

3.3 Αλγόριθμοι Μηχανικής Μάθησης

Υπάρχουν πολλοί αλγόριθμοι μηχανικής μάθησης και ο κάθε ένας αναπτύσσεται με σκοπό να εξυπηρετήσει καλύτερα σε ένα συγκεκριμένο πρόβλημα. Με αποτέλεσμα η επιλογή και η απόδοση του κάθε αλγορίθμου να επηρεάζεται από το σύνολο των δεδομένων, τη φύση των δεδομένων αλλά και από την κατηγορία του προβλήματος.

3.3.1 Παλινδρόμηση

Η παλινδρόμηση είναι η στατιστική τεχνική μοντελοποίησης για την έρευνα συσχέτισης μίας εξαρτώμενης μεταβλητής (Y) και μίας ή περισσότερων ανεξάρτητων μεταβλητών $(X_i, για\ i = 1, \dots)$, όπου χρησιμοποιείται ένα μοντέλο πρόβλεψης των τιμών της κατηγορίας για τα νέα δεδομένα και περιλαμβάνει τις άγνωστες παραμέτρους συσχέτισης (β). Παράδειγμα παλινδρόμησης είναι η πρόβλεψη της ζήτησης μίας νέας υπηρεσίας συναρτήσει διαφόρων δαπανών.

Η παλινδρόμηση γράφεται ως εξίσωση με τη μορφή

$$y = f(X, \beta) + \varepsilon\ ,$$

με το ε να είναι το τυχαίο σφάλμα πρόβλεψης λόγω μη ελεγχόμενων παραγόντων. Οι υπολογισμοί για την ελαχιστοποίηση σφαλμάτων και οι διαφορετικές συναρτήσεις παλινδρόμησης ανάλογα τη φύση του προβλήματος, βοηθιούνται από τη χρήση υπολογιστών και προγραμμάτων ανοικτού ή κλειστού κώδικα. Μερικοί συνήθεις αλγόριθμοι παλινδρόμησης παρουσιάζονται παρακάτω.

3.3.1.1 Γραμμική Παλινδρόμηση (Linear Regression)

Αποτελεί την πιο διαδεδομένη τεχνική παλινδρόμησης για τη μοντελοποίηση και ανάλυση των αριθμητικών δεδομένων μίας εξαρτημένης μεταβλητής και κάποιων ανεξάρτητων μεταβλητών. Το τελικό μοντέλο εκφράζεται μέσω μίας γραμμικής συνάρτησης. Οι υποθέσεις της γραμμικής παλινδρόμησης είναι η γραμμικότητα της σχέσης μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών, η σταθερότητα της διασποράς (ομοσκεδαστικότητα) των σφαλμάτων, η ανεξαρτησία των σφαλμάτων και η κανονικότητα των σφαλμάτων.

Αναφερόμαστε σε απλή γραμμική παλινδρόμηση όταν υπάρχει μόνο μία ανεξάρτητη μεταβλητή, ενώ όταν οι ανεξάρτητες μεταβλητές είναι περισσότερες αναφερόμαστε σε πολλαπλή γραμμική παλινδρόμηση.

3.3.1.2 Παλινδρόμηση Ridge

Είναι μία τεχνική για να αντιμετωπίσει την ισχυρή συσχέτιση των ανεξάρτητων μεταβλητών όπου όταν συμβαίνει αυτό οι διακυμάνσεις των αμερόληπτων εκτιμητών είναι υψηλές. Στην παλινδρόμηση Ridge οι εκτιμητές δεν είναι αμερόληπτοι, αλλά έχουν μικρότερη διακύμανση από τους εκτιμητές ελαχίστων τετραγώνων με το μειονέκτημα τους να είναι ότι τείνουν να υποεκτιμούν τις πραγματικές τιμές.

3.3.1.3 Παλινδρόμηση Διανυσμάτων Υποστήριξης (Support Vector Regression)

Στόχος είναι η εύρεση συνάρτησης που πλησιάζει τα παραδείγματα εκπαίδευσης μέσω ελαχιστοποίησης του σφάλματος πρόβλεψης. Τα σημεία με απόσταση μεγαλύτερη από την καθορισμένη τιμή ε απορρίπτονται και έτσι τα διανύσματα υποστήριξης είναι όλα εκείνα τα σημεία που βρίσκονται κοντά στην επιφάνεια της συνάρτησης.

3.3.1.4 Παλινδρόμηση Gradient Boosting

Είναι μία τεχνική ενδυνάμωσης της κλίσης η οποία χρησιμοποιεί τον αλγόριθμο Gradient Descent. Με αυτή την τεχνική η διαδικασία εκμάθησης προσαρμόζεται σταδιακά στα μοντέλα ώστε να παρέχει ακριβέστερη εκτίμηση της τιμής της μεταβλητής απόκρισης. Με τον αλγόριθμο αυτό επιλέγεται η συνάρτηση κόστους βάσει του τύπου του προβλήματος (βελτιστοποίηση συνάρτησης κόστους). Δημιουργείται μοντέλο από υποσύνολο του συνόλου,

οι προβλέψεις διενεργούνται στα δεδομένα εκπαίδευσης τα οποία χρησιμοποιούν παραμετρικά βάρη και μαθαίνουν από τα λάθη (weak learner), το επόμενο μοντέλο που δημιουργείται έχει λάβει υπόψη τα σφάλματα που παρουσιάστηκαν ώστε να τα διορθώσει και η διαδικασία επαναλαμβάνεται μέχρις ότου δε μπορούν να διορθωθούν επιπλέον σφάλματα ή έχει οριστεί τερματικό κριτήριο.

3.3.1.5 Παλινδρόμηση Random Forest

Σε αυτήν την τεχνική παράγονται πολλές παραλλαγές του πρωτότυπου συνόλου δεδομένων με τη μέθοδο Bootstrap, εκπαιδεύοντας δέντρα όπου το τελικό μοντέλο συνδυάζει όλα τα δέντρα που έχουν δημιουργηθεί (τεχνική Bagging). Όσο πιο ασυσχέτιστες είναι οι προβλέψεις των δέντρων που δημιουργούνται τόσο ικανοποιητικότερα είναι τα αποτελέσματα του μοντέλου. Η διαδικασία επαναλαμβάνεται όσες φορές ορίζει ο χρήστης και το τελικό μοντέλο παίρνει το μέσο όρο των προβλέψεων συνδυάζοντας όλα τα παραγόμενα δέντρα.

3.3.2 Ταξινόμηση

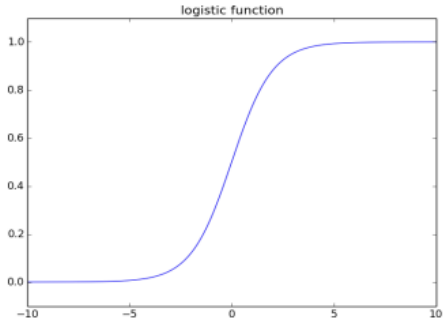
Αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών κλάσεων και ο αλγόριθμος που υλοποιεί την ταξινόμηση αποκαλείται ταξινομητής. Στην ταξινόμηση το σύνολο δεδομένων χωρίζεται σε σετ εκπαίδευσης το οποίο είναι διαθέσιμο για την εκπαίδευση του αλγορίθμου και περιέχει παρατηρήσεις των οποίων η κλάση είναι γνωστή και στη συνέχεια δημιουργείται το προγνωστικό μοντέλο για να ταξινομήσει τα νέα δεδομένα στις υπάρχουσες κλάσεις. Κατά την εκπαίδευση του αλγορίθμου διερευνώνται πρότυπα και συσχετίσεις στα δεδομένα τα οποία θα αυξήσουν την προβλεπτική ικανότητα του μοντέλου. Μερικοί συνήθεις αλγόριθμοι παρουσιάζονται παρακάτω.

3.3.2.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Είναι μία τεχνική για ανάλυση δεδομένων που αφορά τη μελέτη και την πρόβλεψη τιμών μίας κατηγορικής εξαρτημένης μεταβλητής ενώ οι ανεξάρτητες μεταβλητές που χρησιμοποιούνται είναι είτε ποσοτικές είτε ποιοτικές. Το αποτέλεσμα πρόβλεψης δεν είναι πλέον γραμμικό (γραμμική παλινδρόμηση) αλλά είναι δίτιμο ή κατηγορικό και η εκτίμηση των παραμέτρων στηρίζεται στη μέθοδο μέγιστης Πιθανοφάνειας. Για την ανάλυση υπολογίζεται αρχικά ο λόγος πιθανοτήτων που ονομάζεται odds και είναι $\frac{p}{1-p}$ όπου p είναι η πιθανότητα επιτυχίας του γεγονότος ενώ $1-p$ η πιθανότητα αποτυχίας. Μετά, καθορίζεται η logit, ο φυσικός λογάριθμος του λόγου πιθανότητας ώστε να μπορεί να ενσωματωθεί στο μοντέλο παλινδρόμησης και ισχύει ότι:

$$logit(p) = \ln\left(\frac{p}{1-\rho}\right) = b + b_1x_1 + \cdots + b_nx_n \quad = \quad \frac{e^{b_0+b_1x_1+\cdots+b_nx_n}}{1+e^{b_0+b_1x_1+\cdots+b_nx_n}}$$

Στο παρακάτω Σχήμα 3-1 παρουσιάζεται η σιγμοειδής συνάρτηση.

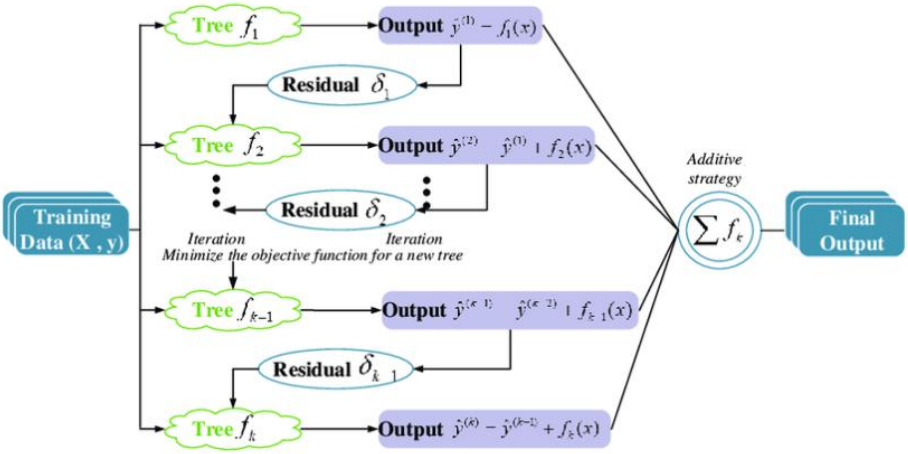


Σχήμα 3-1: Σιγμοειδής συνάρτηση (Τηλλύρος, 2019)

3.3.2.2 Extreme Gradient Boosting (XGBoost)

Ο αλγόριθμος χρησιμοποιεί δέντρα αποφάσεων και την τεχνική ενίσχυσης κλίσης (Gradient Boosting) όπως αυτή εξηγήθηκε σε παραπάνω ενότητα. Εστιάζει στην υπολογιστική ταχύτητα και στην απόδοση του μοντέλου και σχεδιάστηκε με στόχο να βελτιστοποιήσει τον υπολογιστικό χρόνο και τη χρήση των πόρων μνήμης κατά την εκπαίδευση των μοντέλων. Χαρακτηριστικά του είναι η υλοποίηση Sparse Aware όπου διαχειρίζεται αυτόματα ελλείπουσες τιμές, η δομή Block όπου υποστηρίζει παράλληλη κατασκευή δέντρων και η συνεχής εκπαίδευση ώστε να ενισχύεται η απόδοση ενός ήδη εκπαιδευμένου μοντέλου σε νέα δεδομένα.

Όπως φαίνεται και στο παρακάτω Σχήμα 3-2, τα δέντρα λειτουργούν σαν «αδύναμοι μαθητές» και εκπαιδεύονται λαμβάνοντας υπόψη σφάλματα του προηγούμενου δενδροειδές μοντέλου προσπαθώντας να τα διορθώσουν και αυτή η διαδικασία επαναλαμβάνεται έως ότου τα σφάλματα δεν μπορούν να διορθωθούν ή έως να επιτευχθεί ο μέγιστος αριθμός μοντέλων με τελικό σκοπό να βελτιστοποιηθεί η συνάρτηση απώλειας.



Σχήμα 3-2: XGBoost Classification (Cheng et al., 2020)

3.3.2.3 Κ Κοντινότεροι Γείτονες (K Nearest Neighbors)

Ο αλγόριθμος αυτός στηρίζει την ταξινόμηση των δεδομένων στη χρήση μέτρων απόστασης. Υποθέτει ότι σε ένα σύνολο δεδομένων με μία εξαρτημένη μεταβλητή δύο κλάσεων (X) και n ανεξάρτητες μεταβλητές (Y_n) , κάθε σημείο μπορεί να θεωρηθεί ως σημείο στο χώρο των n διαστάσεων και επομένως η απόσταση ανάμεσα σε 2 σημεία X και Y του χώρου είναι ίση με $d(X,Y)$. Ορισμένες συναρτήσεις αποστάσεων που χρησιμοποιεί είναι η Ευκλείδεια απόσταση, η Σταθμισμένη Ευκλείδεια απόσταση, η απόσταση Mahalanobis, η απόσταση Manhattan, η απόσταση Chebychev, η απόσταση Minkowsky. Στον Πίνακα 3-1 παρουσιάζονται οι ορισμένες συναρτήσεις αποστάσεων.

Ευκλείδεια απόσταση: $d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	Σταθμισμένη Ευκλείδεια απόσταση: $d(X,Y) = \sqrt{\sum_{i=1}^{n} w_i(x_i - y_i)^2}$
Απόσταση Mahalanobis: $d(x,y)\sqrt{(x-y)^T s^{-1}(x-y)}$	Απόσταση Manhattan: $d(x,y) = \sum_{i=1}^{m}\lvert x_i - y_i\rvert$
Απόσταση Chebychev: $d(x,y) = \max_{i=1,2,\dots,m}\lvert x_i - y_i\rvert$	Απόσταση Minkowsky: $d(x,y) = (\sum_{i=1}^{m}\lvert x_i - y_i\rvert^r)^{1/r}$

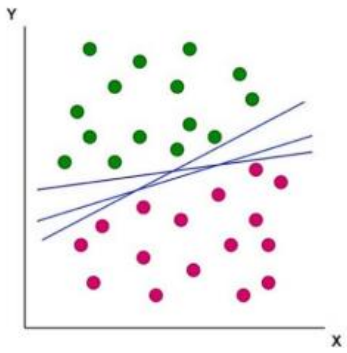
Πίνακας 3-1: Συναρτήσεις αποστάσεων

3.3.2.4 Naïve Bayes

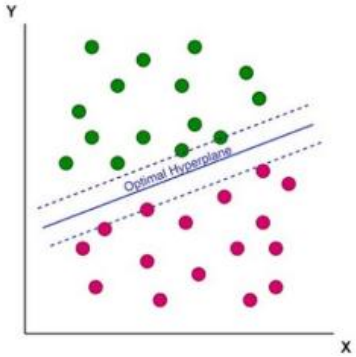
Είναι μία τεχνική ταξινόμησης όπου υποθέτει ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού δεδομένης της μεταβλητής κλάσης. Χρησιμοποιείται η μέθοδος μεγίστης Πιθανοφάνειας και ένα από τα πλεονεκτήματα της τεχνικής αυτής είναι απαιτεί μικρό αριθμό δεδομένων εκπαίδευσης για την εκτίμηση παραμέτρων που χρειάζονται στην ταξινόμηση.

3.3.2.5 Support Vector Machines

Σκοπός είναι η εύρεση εξίσωσης που περιγράφει το υπερεπίπεδο μεγίστου περιθωρίου διαχωρίζοντας τα θετικά από τα αρνητικά παραδείγματα των 2 κατηγοριών. Σε περιπτώσεις προβλημάτων 2 διαστάσεων ο διαχωρισμός θα επιτευχθεί με τη διαχωριστική ευθεία (Σχήμα 3-3), ενώ σε περιπτώσεις περισσότερων διαστάσεων αναζητείται το βέλτιστο υπερεπίπεδο (Σχήμα 3-4). Η μεγιστοποίηση του περιθωρίου παρέχει κάποια ενίσχυση ώστε τα μελλοντικά παραδείγματα να ταξινομηθούν ακριβέστερα σε μία από τις δύο κατηγορίες. Τα διανύσματα υποστήριξης είναι αυτά με τη μικρότερη απόσταση από το υπερεπίπεδο μεγίστου περιθωρίου. Στα Σχήματα που ακολουθούν παρουσιάζεται ο διαχωρισμός.



Σχήμα 3-3: Απεικόνιση με διαχωριστική ευθεία



Σχήμα 3-4: Απεικόνιση με βέλτιστο υπερεπίπεδο

Ένα μοντέλο SVM δίτιμων δεδομένων αποτελείται από τη μεταβλητή απόκρισης Y όπου παίρνει τιμές {-1,1} και από τις ερμηνευτικές μεταβλητές X_i. Η συνάρτηση απώλειας είναι αυτή που πρέπει να ελαχιστοποιηθεί και έχει ως εξής:

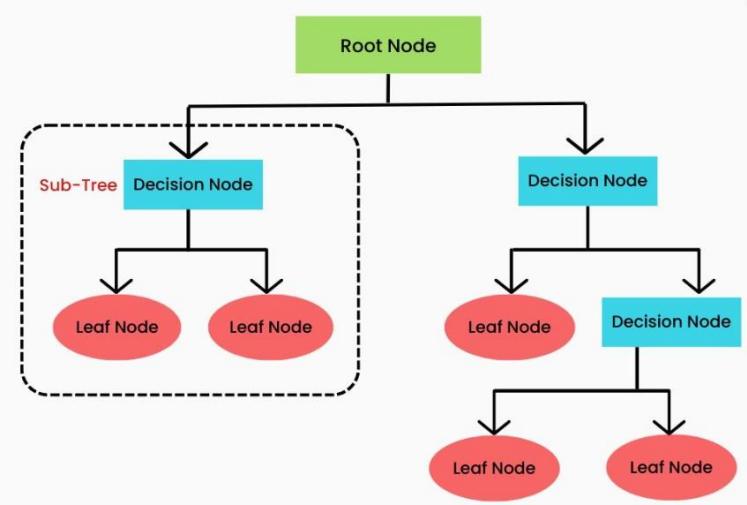
$$\min_{w} \ \frac{1}{2}\sum_{i=1}^{n} w_i^2 + c\sum_{j\cdot=1}^{m} \max\left(0,1 - k(x_i, x_j)y_j\right),$$

όπου το πρώτο άθροισμα αναφέρεται στο πλήθος των χαρακτηριστικών n, ενώ το δεύτερο στον αριθμό των δειγμάτων των δεδομένων m. Το w εκφράζει τα βάρη των ερμηνευτικών μεταβλητών X_i , το C τη σταθερά κανονικοποίησης, ενώ το $k(\chi_i, \chi_j)$ εκφράζει τη συνάρτηση πυρήνα της οποίας ο τύπος διαφέρει ανάλογα τη φύση του προβλήματος δηλαδή εάν είναι

γραμμικά πλήρως διαχωρίσιμες ή μη διαχωρίσιμες κλάσεις, μη απόλυτα γραμμικά διαχωρίσιμες κλάσεις.

3.3.2.6 Δέντρα Απόφασης (Decision Trees)

Ο αλγόριθμος δημιουργεί μία δομή που μοιάζει με δέντρο και μπορεί να χρησιμοποιηθεί για επίλυση προβλημάτων. Το δέντρο αρχίζει με ένα ριζικό κόμβο (κόμβος απόφασης, root node) και διακλαδίζεται σε υποκόμβους που αντιπροσωπεύουν πιθανά αποτελέσματα. Κάθε αποτέλεσμα μπορεί να δημιουργήσει θυγατρικούς κόμβους οι οποίοι μπορούν να οδηγήσουν σε νέες δυνατότητες. Ο χρόνος που απαιτείται στα δέντρα απόφασης είναι συνάρτηση του αριθμού εγγραφών και των χαρακτηριστικών του συνόλου. Τα δέντρα απόφασης είναι μέθοδος που δεν εξαρτάται από υποθέσεις κατανομής και μπορούν να διαχειριστούν δεδομένα μεγάλων διαστάσεων με υψηλή ακρίβεια. Υπάρχουν αρκετοί αλγόριθμοι για την κατασκευή ενός δέντρου απόφασης όπως ο ID3, ο C4.5, ο CART. Στο Σχήμα 3-5 παρουσιάζονται τα χαρακτηριστικά ενός δέντρου απόφασης.



Σχήμα 3-5: Δέντρο αποφάσεων (Why Do We Use Decision Trees in Machine Learning, Turing, 2023)

3.3.3 Ομαδοποίηση κατά Συστάδες (Clustering)

Είναι η διαδικασία διαχωρισμού του συνόλου δεδομένων σε ομάδες βάσει των ομοιοτήτων τους χωρίς να υπάρχει πρότερη γνώση για κριτήριο διαχωρισμού. Μερικοί συνήθεις αλγόριθμοι ομαδοποίησης παρουσιάζονται παρακάτω.

3.3.3.1 Ιεραρχικοί Αλγόριθμοι (Hierarchical algorithms)

Οι Συστάδες δημιουργούνται σε επίπεδα και κάθε επίπεδο αντιπροσωπεύει ένα σύνολο από Συστάδες. Ανάλογα του τρόπου προσέγγισης της κατηγοριοποίησης, οι αλγόριθμοι χωρίζονται σε Συσσωρευτικούς και Διαιρετικούς. Αυτό σημαίνει ότι σε κάθε περίπτωση θα προκύπτουν ομάδες από τα δεδομένα οι οποίες θα πρέπει να αξιολογούνται από τον αναλυτή για τη συνεισφορά τους.

1.Οι Συσσωρευτικοί αλγόριθμοι (Agglomerative algorithms) θεωρούν κάθε στοιχείο ότι αποτελεί μία συστάδα και η διαδικασία συγχώνευσης επαναλαμβάνεται μέχρι να δημιουργηθούν οι τελικές συστάδες βάσει των κριτηρίων του χρήστη.

2.Οι Διαιρετικοί αλγόριθμοι (Divisive algorithms) θεωρούν ότι όλα τα στοιχεία αποτελούν μία συστάδα και η διαδικασία διαίρεσης επαναλαμβάνεται μέχρι να δημιουργηθούν οι τελικές συστάδες βάσει των κριτηρίων του χρήστη.

3.3.3.2 Αλγόριθμος K-means

Με αυτήν την τεχνική ομαδοποίησης ο χρήστης δηλώνει τον αριθμό των συστάδων που επιθυμεί να διαμερίσει το σύνολο δεδομένων και στον αλγόριθμο ξεκινάει μία επαναληπτική διαδικασία ταξινόμησης των στοιχείων στις συστάδες που λαμβάνει υπόψη κάποιο μέτρο απόστασης μέχρις ότου να μην εμφανίζονται πλέον μεταβολές στις θέσεις των στοιχείων. Συγκεκριμένα, δέχεται ως είσοδο ένα σύνολο στοιχείων χ_n και τον αριθμό k συστάδων που ορίζει ο χρήστης και ξεκινάει η επαναληπτική διαδικασία με την οποία γίνεται η ανάθεση των k τυχαίων σημείων (K_n) που ονομάζονται κεντροειδή και δηλώνουν το κέντρο βάρους της συστάδας. Για κάθε χ υπολογίζεται το κοντινότερο $K_j(argmin_j D(\chi_i, K_j)$ και με τη χρήση μέτρων απόστασης αντιστοιχίζεται σε μία συστάδα K_j. Σε κάθε μία διαδικασία υπολογίζονται ξανά τα γεωμετρικά κέντρα των συστάδων K_n βάσει του μέσου όρου κάθε συστάδας από όλα τα χ που προέκυψαν από το προηγούμενο βήμα και η διαδικασία ολοκληρώνεται όταν δεν εμφανίζονται πλέον μεταβολές στα στοιχεία των συστάδων, δηλαδή όταν τα κεντροειδή των συστάδων μετατοπίζονται ελάχιστα.

$$K_j = \frac{1}{n_j} \sum_{x_i \to k_j} x_i$$

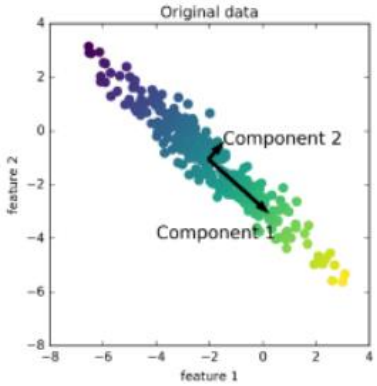
3.3.4 Κανόνες Συσχέτισης (Association Rules)

Βασίζονται σε τεχνικές που βοηθάνε στον περιορισμό του εκθετικά μεγάλου χώρου αναζήτησης. Μία διαδεδομένη τεχνική είναι ο αλγόριθμος Apriori, ο οποίος έχει ως αρχή ότι όλα τα υποσύνολα ενός συχνού συνόλου, πρέπει να είναι επίσης συχνά. Εάν το υποσύνολο $\{A, B\}$ είναι συχνό, τότε τα αντικείμενα $\{A\}$ και $\{B\}$ θα πρέπει να είναι συχνά. Η υποστήριξη (support) ως μέτρο δείχνει πόσο συχνά ένα αντικείμενο εμφανίζεται στα δεδομένα. Εφόσον ξέρουμε ότι το αντικείμενο $\{A\}$ δεν ικανοποιεί ένα επιθυμητό όριο υποστήριξης δεν υπάρχει λόγος να ληφθεί υπόψη κανένα υποσύνολο το οποίο περιλαμβάνει το $\{A\}$. Ο αλγόριθμος χρησιμοποιεί τη λογική για να απορρίψει πιθανούς κανόνες συσχέτισης με χαμηλή υποστήριξη που δεν παρουσιάζουν κάποιο ενδιαφέρον.

3.3.5 Μείωση Διαστασιμότητας (Dimensionality Reduction)

Η τεχνική αυτή στοχεύει στην ανεύρεση νέων και συνήθως λιγότερων στο πλήθος μεταβλητών από τις p μεταβλητές ενός συνόλου δεδομένων, με τρόπο που θα είναι αντιπροσωπευτικές, ισχυρά επεξηγηματικές των αρχικών μεταβλητών, θα είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα ασυσχέτιστες μεταξύ τους. Μία διαδεδομένη τεχνική μείωσης διαστάσεων είναι η Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis – PCA).

Συγκεκριμένα, ο PCA μπορεί και εξηγεί μεγάλο μέρος της συνολικής μεταβλητότητας μεταξύ των αρχικών p μεταβλητών μέσω των νέων μειωμένων μεταβλητών. Αρχικά γίνεται τυποποίηση των αρχικών X_i μεταβλητών και στη συνέχεια δημιουργούνται οι p συνδυασμοί με τέτοιο τρόπο ώστε να μην υπάρχει συσχέτιση μεταξύ τους και κάθε διάσταση να μετράει διαφορετικές διαστάσεις των στοιχείων. Στο παρακάτω Σχήμα 3-6 παρουσιάζονται διάφορα σημεία με διαφορετικό χρώμα με σκοπό να διαχωριστούν και αρχικά ο αλγόριθμος εντοπίζει την κατεύθυνση της μέγιστης διακύμανσης με την ένδειξη component 1 και στη συνέχεια εντοπίζει εκείνη την κατεύθυνση που είναι ορθογώνια ως προς την πρώτη κατεύθυνση και περιέχει περισσότερες πληροφορίες με την ένδειξη component 2.



Σχήμα 3-6: Μείωση διαστάσεων με τη μέθοδο PCA (Τηλλύρος, 2019)

3.4 Μετρικές Αξιολόγησης (Evaluation Metrics)

Η επίδοση των μοντέλων Μηχανικής Μάθησης αξιολογείται για να επιλεχθεί το καλύτερο μοντέλο για την εκάστοτε περίπτωση. Σε προβλήματα παλινδρόμησης ο έλεγχος γίνεται με εκτίμηση του σφάλματος πρόβλεψης μεταξύ της πραγματικής και της προβλεπόμενης τιμής εξόδου, ενώ σε προβλήματα ταξινόμησης υπολογίζεται το πλήθος των παρατηρήσεων που ο ταξινομητής ταξινόμησε ορθά ή λανθασμένα. Παρακάτω παρουσιάζονται τέτοιες τεχνικές αξιολόγησης.

Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE)

Είναι η μέση τιμή του τετραγώνου του σφάλματος πρόβλεψης σε όλα τα παραδείγματα του συνόλου δεδομένων. Το καλύτερο αποτέλεσμα είναι η τιμή μηδέν ενώ οι μεγαλύτερες τιμές δηλώνουν μεγάλη απόκλιση από την πραγματικότητα κατά συνέπεια είναι κακό αποτέλεσμα.

Μέσο Τετραγωνικό Σφάλμα (Root Mean Squared Error - RMSE)

Είναι η ρίζα του αντίστοιχου μέσου τετραγωνικού σφάλματος και το αποτέλεσμα ερμηνεύεται αντίστοιχα. Το πλεονέκτημα του είναι ότι βρίσκεται στην ίδια διάσταση με την προβλεπόμενη τιμή.

Μέσο Απόλυτο Σφάλμα (Mean Absolute Error – MAE)

Είναι η μέση τιμή της απόλυτης τιμής του σφάλματος πρόβλεψης σε όλα τα παραδείγματα του συνόλου δεδομένων και το αποτέλεσμα ερμηνεύεται αντίστοιχα. Είναι λιγότερο ευαίσθητο σε ακραίες τιμές από τα MSE, RMSE.

Συντελεστής Προσδιορισμού (R^2)

Προσδιορίζει τι ποσοστό της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής ερμηνεύουν οι ανεξάρτητες μεταβλητές και οι τιμές που παίρνει είναι από μηδέν (δεν ερμηνεύεται καθόλου από τις επεξηγηματικές μεταβλητές) έως 1 (οι ανεξάρτητες μεταβλητές ερμηνεύουν πλήρως τη μεταβλητότητα της εξαρτημένης μεταβλητής).

Συντελεστής Συσχέτισης (Correlation Coefficient – CC)

Υπολογίζει τη συσχέτιση που υπάρχει μεταξύ των πραγματικών τιμών εξόδου και των προβλεπόμενων. Οι τιμές που μπορεί να πάρει είναι από -1 έως 1, με την τιμή 0 να δηλώνει ότι δεν υπάρχει καμία συσχέτιση μεταξύ των μεταβλητών ενώ οι τιμές -1 και 1 δηλώνουν πλήρη συσχέτιση των μεταβλητών αρνητική και θετική αντίστοιχα.

Στις παρακάτω μετρικές θα γίνει αναφορά στην Ορθή αποδοχή (True Positive – TP) δηλαδή το πλήθος των παρατηρήσεων που ταξινομήθηκαν ορθά ως θετικά, Εσφαλμένη αποδοχή (False Positive - FP) δηλαδή το πλήθος των παρατηρήσεων που ταξινομήθηκαν εσφαλμένα ως θετικά, Ορθή απόρριψη (True Negative - TN) δηλαδή το πλήθος των παρατηρήσεων που ταξινομήθηκαν ορθά ως αρνητικά και στην Εσφαλμένη απόρριψη (False Negative - FN) δηλαδή το πλήθος των παρατηρήσεων που ταξινομήθηκαν εσφαλμένα ως αρνητικά. Αυτές οι μετρικές παρέχουν πληροφορίες σχετικά με την απόδοση ενός μοντέλου, βοηθούν στην κατανόηση των αποτελεσμάτων με ακρίβεια και στην εξαγωγή συμπερασμάτων σχετικά με την ικανότητα του μοντέλου να διαχωρίζει με επιτυχία τις θετικές και αρνητικές παρατηρήσεις.

Ορθότητα (Accuracy)

Είναι το ποσοστό των ορθά ταξινομημένων περιπτώσεων, αλλά σε περιπτώσεις που τα μεγέθη των κλάσεων δεν είναι ισοπληθή το μέτρο αυτό δεν είναι αντιπροσωπευτικό.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Ακρίβεια (Precision)

Είναι το ποσοστό των περιπτώσεων που ταξινομήθηκαν ως θετικά και στην πραγματικότητα είναι θετικά.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Ανάκληση (Recall)

Είναι το ποσοστό των περιπτώσεων που είναι θετικά και ταξινομήθηκαν ορθώς ως θετικά.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Ειδικότητα (Specificity)

Είναι το ποσοστό των περιπτώσεων που είναι αρνητικά και ταξινομήθηκαν ως αρνητικά.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Ευαισθησία (Sensitivity)

Είναι το ποσοστό των περιπτώσεων που είναι θετικά και ταξινομήθηκαν ως θετικά.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

F-score

Είναι η μετρική του αρμονικού μέσου όρου της Ανάκλησης και της Ακρίβειας.

$$\text{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.5 Προεπεξεργασία Δεδομένων (Preprocessing)

Η προεπεξεργασία των δεδομένων είναι μία διαδικασία προετοιμασίας του συνόλου δεδομένων πριν εφαρμοστούν οι αλγόριθμοι Μηχανικής Μάθησης. Είναι σημαντικό μέρος της ανάλυσης καθώς τα ακατέργαστα δεδομένα που συλλέγονται έχουν «θόρυβο» και χρειάζονται «καθάρισμα» για να μπορέσουν να εφαρμοστούν οι αλγόριθμοι και να δώσουν πληροφορία. Τα κριτήρια αυτής της διαδικασίας είναι υποκειμενικά και οι αναλυτές εφαρμόζουν αυτά που θεωρούν κατάλληλα, με τις συνθήκες που οι ίδιοι ορίζουν ως κατάλληλες και ανάλογα τη φύση του προβλήματος που καλούνται να διαχειριστούν. Τα πιο συνήθη βήματα που εφαρμόζονται είναι η κωδικοποίηση των κατηγορικών χαρακτηριστικών, η διαχείριση ακραίων τιμών, η

διαχείριση ελλειπουσών τιμών, ο μετασχηματισμός των δεδομένων για να βρίσκονται σε κοινή κλίμακα, τεχνικές για εξισορρόπηση του πλήθους δείγματος όταν αυτό δεν είναι ισορροπημένο. Παρακάτω περιγράφονται κάποιες από αυτές τις τεχνικές.

Ελλείπουσες Τιμές

Σε περιπτώσεις που το ποσοστό των ελλειπουσών τιμών είναι πολύ μικρό μπορούν να αφαιρεθούν οι εγγραφές από το σύνολο δεδομένων. Σε περιπτώσεις που το ποσοστό των ελλειπουσών τιμών μίας μεταβλητής είναι πολύ μεγάλο μπορεί να αφαιρεθεί η μεταβλητή από το σύνολο δεδομένων καθώς δεν θα προκύψει κάποια αξιόλογη πληροφορία. Άλλη τεχνική είναι η αντικατάσταση των ελλειπουσών τιμών με το μέσο όρο ή τη διάμεση τιμή της κατανομής τους σε περιπτώσεις ποσοτικών μεταβλητών ή επικρατούσα τιμή σε περιπτώσεις κατηγορικών μεταβλητών.

Κωδικοποίηση κατηγορικών μεταβλητών

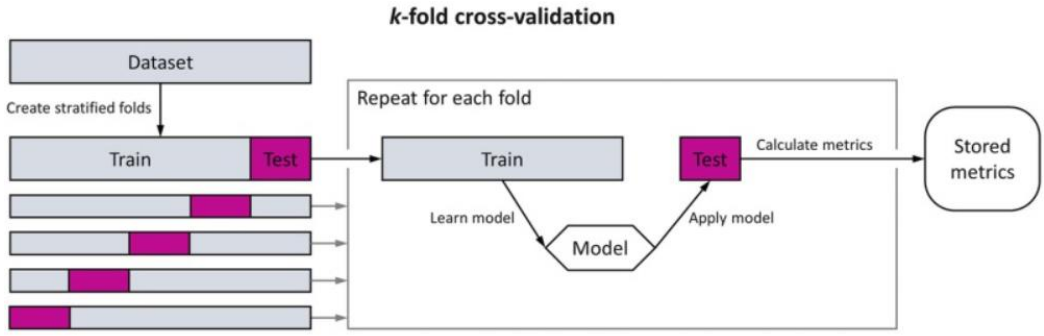
Οι ετικέτες των κατηγορικών μεταβλητών θα πρέπει να κωδικοποιηθούν για να μπορέσουν να αξιοποιηθούν από τους αλγορίθμους που στην πλειοψηφία τους αναγνωρίζουν αριθμητικές τιμές. Η δημιουργία ψευδομεταβλητών είναι απαραίτητη σε ορισμένες περιπτώσεις που οι κατηγορικές μεταβλητές είναι άνω των 2 τάξεων.

Διαχωρισμός του συνόλου δεδομένων σε σετ Εκπαίδευσης και σετ Δοκιμής

Σκοπός είναι το μοντέλο που θα δημιουργηθεί να έχει υψηλό ποσοστό πρόβλεψης σε νέα δεδομένα. Για το λόγο αυτό γίνεται ο διαχωρισμός των υπαρχόντων δεδομένων σε σύνολο εκπαίδευσης (training set) από το οποίο μαθαίνει το μοντέλο τις τάσεις των δεδομένων και έπειτα δοκιμάζεται η προβλεπτική του ικανότητα στο σύνολο δοκιμής (test set) το οποίο μας ενημερώνει για την τελική ακρίβεια του μοντέλου. Σε πολλές περιπτώσεις ένα μέρος του συνόλου εκπαίδευσης χρησιμοποιείται ως σετ επικύρωσης (validation set) το οποίο συμβάλλει στη βελτίωση της απόδοσης του μοντέλου προσαρμόζοντας το μετά από κάθε epoch. Το ποσοστό διαχωρισμού του συνόλου σε σετ εκπαίδευσης, επικύρωσης και δοκιμής αποφασίζεται από το χρήστη. Στις παρακάτω μελέτες έγινε διαχωρισμός 75% σετ εκπαίδευσης και 25% σετ δοκιμής, δε χρησιμοποιήθηκε σετ επικύρωσης.

Η Διασταυρούμενη Επικύρωση k Τμημάτων (k-Fold Cross Validation) είναι μία μέθοδος επαναδειγματοληψίας που χρησιμοποιεί διαφορετικά τμήματα των δεδομένων για να δοκιμάσει και να εκπαιδεύσει ένα μοντέλο σε διαφορετικές επαναλήψεις. Η παράμετρος *k* που

καθορίζει τον αριθμό των τμημάτων που θα χωριστεί το σετ ορίζεται από τον αναλυτή και ένα χαρακτηριστικό είναι ότι τα τμήματα έχουν ίσο μέγεθος. Στο παρακάτω Σχήμα 3-7 απεικονίζεται η ροή της διασταυρούμενης επικύρωσης *k* τμημάτων. Σε μελέτη που ακολουθεί εφαρμόστηκε 8-folds cross validation.



Σχήμα 3-7: k- fold cross-validation (Prediction Modeling Methodology, Frank et al., 2018)

Μετασχηματισμός Δεδομένων σε κοινή κλίμακα

Δύο συνήθεις μέθοδοι κανονικοποίησης είναι ο StandarScaler και ο MinMaxScaler. Στην πρώτη περίπτωση τυποποιείται ένα χαρακτηριστικό αφαιρώντας το μέσο όρο και στη συνέχεια διαιρούνται όλες οι τιμές με την τυπική απόκλιση. Επηρεάζεται από ακραίες τιμές καθώς εμπλέκει τον εμπειρικό μέσο όρο και την τυπική απόκλιση κάθε στοιχείου του συνόλου. Στη δεύτερη περίπτωση για κάθε τιμή σε ένα χαρακτηριστικό αφαιρείται η ελάχιστη τιμή στο χαρακτηριστικό και διαιρείται με το εύρος. Το εύρος είναι η διαφορά του αρχικού μεγίστου και του αρχικού ελαχίστου.

Τυχαία Υπερδειγματοληψία

Η τεχνική SMOTE (Synthetic Minority Oversampling Technique) είναι μία τέτοια μέθοδος όπου για να αντιμετωπίσει το πρόβλημα μη ισορροπημένου συνόλου δεδομένων ως προς κάποιο χαρακτηριστικό, δημιουργεί νέα συνθετικά δείγματα για την τάξη της μειοψηφίας με μία επαναλαμβανόμενη διαδικασία μέχρι να επιτύχει ισορροπημένη κατανομή τάξεων. Ένα μειονέκτημα της τεχνικής SMOTE είναι ότι μπορεί να αυξήσει την πιθανότητα υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης και η προβλεπτική ικανότητα του μοντέλου να στηρίζεται στα τεχνητά δεδομένα αποδίδοντας χειρότερα σε πραγματικά δεδομένα. Για να διατηρηθεί η αξιοπιστία και η ικανότητα γενίκευσης των μοντέλων σε πραγματικές συνθήκες, στις παρακάτω αναλύσεις παρόλο που εφαρμόστηκε η τεχνική δεν προτάθηκαν τα συγκεκριμένα μοντέλα

ΚΕΦΑΛΑΙΟ 4

Εφαρμογές

4.1 Εισαγωγή

Παρακάτω παρουσιάζονται τρεις εφαρμογές στις οποίες έχουν χρησιμοποιηθεί τεχνικές Στατιστικής Μηχανικής Μάθησης και Αναλυτικής των δεδομένων για να δώσουν λύσεις στα προβλήματα της διοίκησης υπηρεσιών υγείας που μελετήθηκαν και παραπάνω.

Αυτά είναι η Εμφάνιση ή Απουσία ενός ασθενή από το προγραμματισμένο του ραντεβού σε εξωτερικό ιατρείο, η Επανεμφάνιση ενός ασθενή στο Νοσοκομείο εντός 30 ημερών από την ημέρα εξιτηρίου και ο Προγραμματισμός κράτησης μίας χειρουργικής αίθουσας νοσοκομείου.

4.2 1η Εφαρμογή – Πρόβλεψη Εμφάνισης ασθενή σε Προγραμματισμένο Ραντεβού

Πρόβλημα

Ο προγραμματισμός των ραντεβού ασθενών σε περιβάλλοντα εξωτερικών ιατρείων είναι μεγάλης σημασίας καθώς αντιστοιχίζει τη ζήτηση με τη χωρητικότητα. Μία κακή διαχείριση μπορεί να φέρει συνωστισμό ασθενών, υπερφόρτωση ή έλλειψη ιατρικού προσωπικού. Ένα από τα μεγαλύτερα προβλήματα στον προγραμματισμό είναι η μη εμφάνιση προγραμματισμένων ραντεβού ασθενών, όπου μία απουσία κοστίζει στην εγκατάσταση μία ευκαιρία εσόδων αφήνοντας παράλληλα πολύτιμους πόρους ανεκμετάλλευτους.

Σκοπός

Σκοπός της παρακάτω ανάλυσης είναι αρχικά η διερεύνηση των δεδομένων και στη συνέχεια η δημιουργία προβλεπτικού μοντέλου μη εμφάνισης ασθενών με βάση τα διαθέσιμα δεδομένα. Το μοντέλο θα συμβάλλει στον προγραμματισμό των ραντεβού ώστε να βοηθήσει

τη διοίκηση της εγκατάστασης να κατανείμει τις βάρδιες προσωπικού με τέτοιο τρόπο ώστε να αποφευχθούν οι ελλείψεις ή υπεράριθμα άτομα σε βάρδια.

Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων αντλήθηκε από τον ιστότοπο προέρχεται της Kaggle. Αφορά μία μονάδα Νοσοκομείου στην Αμερική και περιέχει ανωνυμοποιημένα δεδομένα για 110.527 ασθενείς και πληροφορίες για το φύλο, την ηλικία, τη γειτονιά τους, το είδος ασφάλισης που έχουν, μετρήσεις δεικτών υγείας όπως η υπέρταση ή ο διαβήτης, εάν υπάρχει αναπηρία, πληροφορίες για την ημερομηνία του ραντεβού τους, το διάστημα που μεσολάβησε από τον προγραμματισμό μέχρι την ημερομηνία ραντεβού και εάν πραγματοποίησαν την επίσκεψη ή όχι.

Συγκεκριμένα οι δεκατέσσερεις μεταβλητές που περιέχονται στο σύνολο καθώς και ο τύπος τους περιγράφονται στον παρακάτω Πίνακα 4-1. Τα δεδομένα αφορούν διάστημα 19 μηνών (Νοέμβριος 2015 έως Ιούνιος 2016).

Πίνακας 4-1: Περιγραφή μεταβλητών του συνόλου δεδομένων

ΑΑ	Μεταβλητή	Περιγραφή	Ετικέτες	Τύπος
1	Patient Id	Αναγνωριστικό ασθενούς		Ποιοτική
2	Appointment ID	Αναγνωριστικό ραντεβού		Ποιοτική
3	Gender	Φύλο	• Αρσενικό • Θηλυκό	Ποιοτική
4	Appointment Day	Ημερομηνία ραντεβού		Ποιοτική
5	Scheduled Day	Ημερομηνία Προγραμματισμού		Ποιοτική
6	Age	Ηλικία		Ποσοτική
7	Neighbourhood	Γειτονιά		Ποιοτική
8	Scholarship	Ύπαρξη ασφάλισης	• Ναι • Όχι	Ποιοτική
9	Hipertension	Υπέρταση	• Ναι • Όχι	Ποιοτική
10	Diabetes	Διαβήτης	• Ναι • Όχι	Ποιοτική

11	Alcoholism	Αλκοολισμός	• Ναι • Όχι	Ποιοτική
12	Handcap	Αναπηρία (προσδιορίζει το πλήθος)	• Καμία • 1 • 2 • 3 • 4 • 5 και άνω	Ποιοτική
13	SMS_received	Μήνυμα υπενθύμισης ραντεβού	• Ελήφθη • Δεν ελήφθη	Ποιοτική
14	No-show	Εμφάνιση στο ραντεβού (Μεταβλητή Στόχος)	• Ναι • Όχι	Ποιοτική

Διερευνητική Ανάλυση

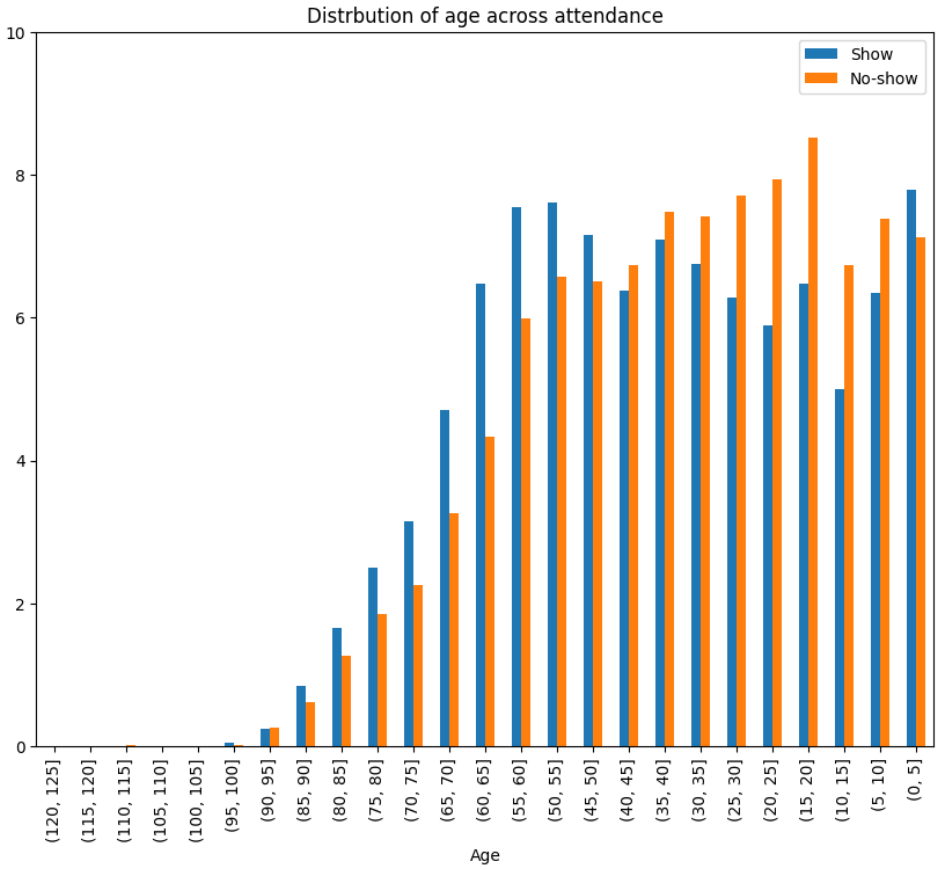
Η παρακάτω διερευνητική ανάλυση θα βοηθήσει να κατανοηθούν τα χαρακτηριστικά του συνόλου δεδομένων και οι μεταβλητές ώστε να γίνει καλύτερη διαχείριση τους. Από τις 110.527 εγγραφές η αναλογία Εμφάνισης – Μη εμφάνισης στο ραντεβού είναι περίπου 80-20 %, ενώ στο σύνολο δεδομένων η μέση τιμή ηλικίας είναι τα 37 έτη.

Από τη διερεύνηση συσχέτισης μεταξύ των μεταβλητών προέκυψε μέτρια θετική συσχέτιση μεταξύ ηλικίας, υπέρτασης και διαβήτη η οποία παρουσιάζεται στο παρακάτω Σχήμα 4-1.

	PatientId	AppointmentID	Age	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received
PatientId	1.000000	0.004039	-0.004139	-0.002880	-0.006441	0.001605	0.011011	-0.007916	-0.009749
AppointmentID	0.004039	1.000000	-0.019126	0.022615	0.012752	0.022628	0.032944	0.014106	-0.256618
Age	-0.004139	-0.019126	1.000000	-0.092457	0.504586	0.292391	0.095811	0.078033	0.012643
Scholarship	-0.002880	0.022615	-0.092457	1.000000	-0.019729	-0.024894	0.035022	-0.008586	0.001194
Hipertension	-0.006441	0.012752	0.504586	-0.019729	1.000000	0.433086	0.087971	0.080083	-0.006267
Diabetes	0.001605	0.022628	0.292391	-0.024894	0.433086	1.000000	0.018474	0.057530	-0.014550
Alcoholism	0.011011	0.032944	0.095811	0.035022	0.087971	0.018474	1.000000	0.004648	-0.026147
Handcap	-0.007916	0.014106	0.078033	-0.008586	0.080083	0.057530	0.004648	1.000000	-0.024161
SMS_received	-0.009749	-0.256618	0.012643	0.001194	-0.006267	-0.014550	-0.026147	-0.024161	1.000000

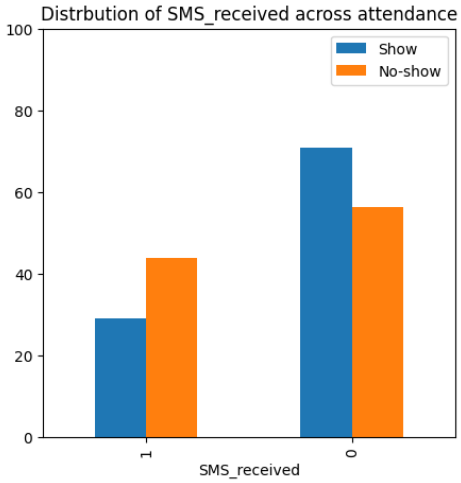
Σχήμα 4-1: Συσχέτιση μεταξύ των μεταβλητών

Κατά τη διερεύνηση της ηλικίας φάνηκε η τάση άτομα μεταξύ 7 έως 32 ετών να χάνουν τα ραντεβού τους πιο συχνά και από 50 έως 70 ετών να παρευρίσκονται στα ραντεβού τους. Αλλά πέρα από αυτή την τάση δεν φαίνεται η ηλικία να επηρεάζει περισσότερο στην εμφάνιση του ασθενούς στο ραντεβού και αυτά παρουσιάζονται στο Σχήμα 4-2.

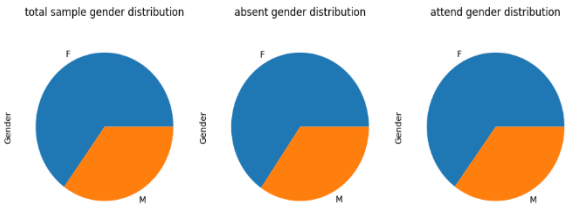


Σχήμα 4-2: Ηλικία

Από τη διερεύνηση της υπενθύμισης μέσω μηνύματος φάνηκε ότι μεταξύ των ατόμων που έχασαν το ραντεβού τους το 55% δεν είχε λάβει μήνυμα ενώ το 45% είχε λάβει μήνυμα, ενώ μεταξύ των ατόμων που παρευρέθηκαν στο ραντεβού τους το 70% δεν είχε λάβει μήνυμα και μόλις το 30% είχε λάβει μήνυμα. Δεν φαίνεται η συγκεκριμένη μεταβλητή να επηρεάζει στην εμφάνιση του ασθενούς. Αντίστοιχα, το φύλο φαίνεται να μην επηρεάζει τη μεταβλητή στόχο. Τα παραπάνω παρουσιάζονται στο Σχήμα 4-3 και Σχήμα 4-4 αντίστοιχα.

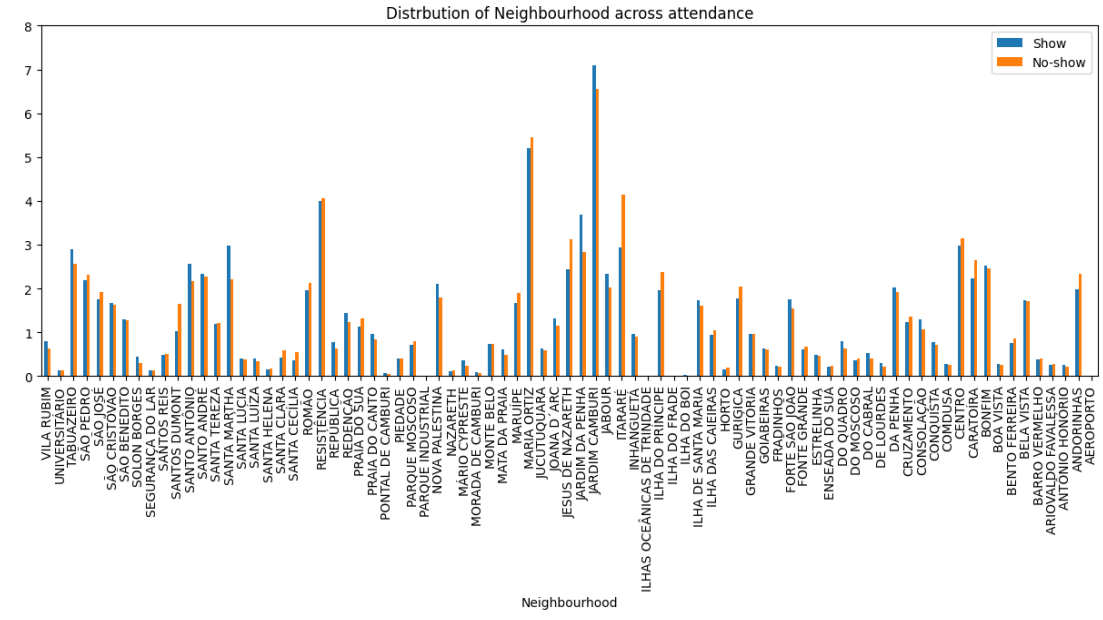


Σχήμα 4-3: Μήνυμα Υπενθύμισης



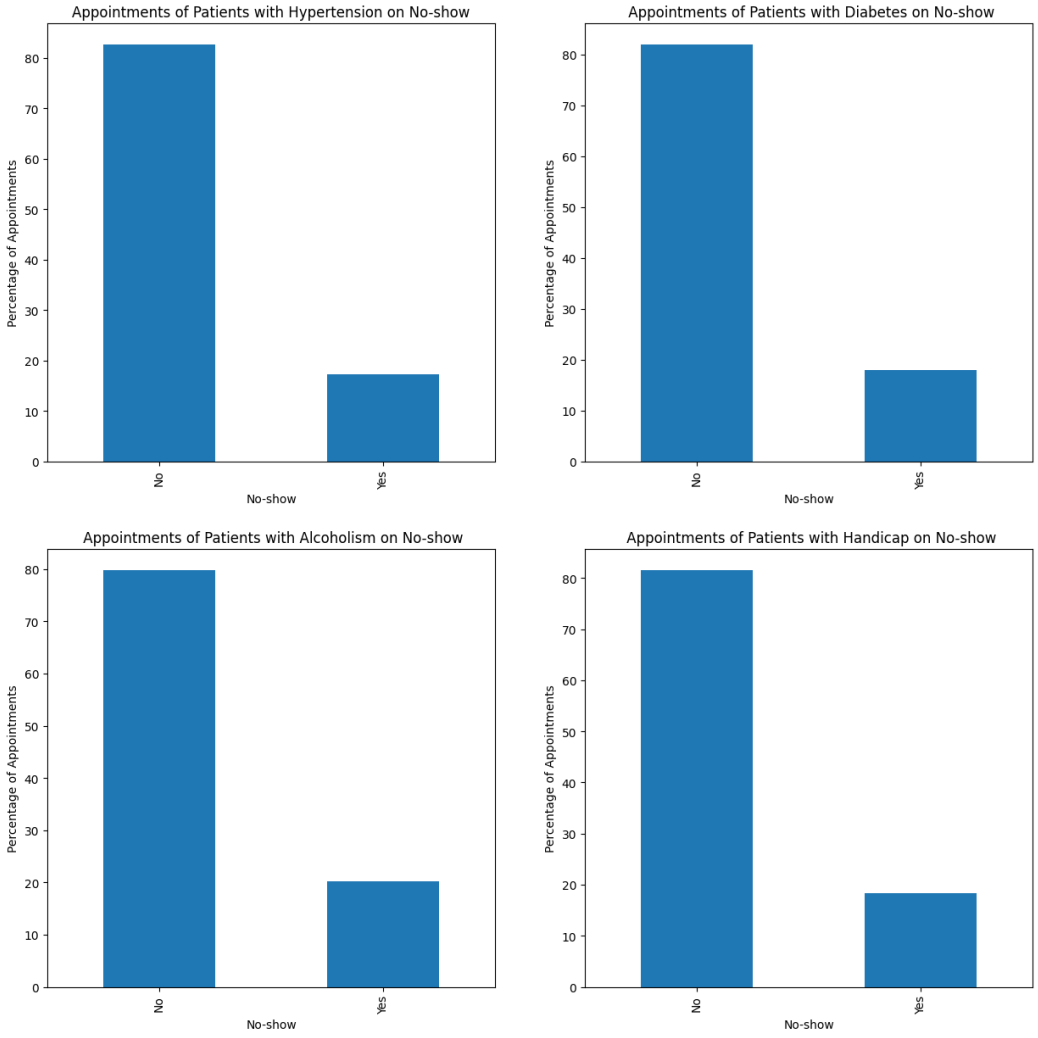
Σχήμα 4-4: Φύλο

Από τη διερεύνηση της γειτονιάς φάνηκε η μεταβλητή να μην επηρεάζει την εμφάνιση ή όχι στο ραντεβού του ασθενή και παρουσιάζεται στο παρακάτω Σχήμα 4-5.



Σχήμα 4-5: Γειτονιά

Από τη διερεύνηση σχέσης μεταξύ της εμφάνισης ή όχι στο ραντεβού και των ασθενειών του συνόλου δηλαδή της υπέρτασης, του διαβήτη αλλά και του αλκοολισμού και των δηλωμένων αναπηριών δεν φάνηκε να υπάρχει κάποια σύνδεση και στο παρακάτω Σχήμα 4-6 παρουσιάζονται τα σχετικά γραφήματα.



Σχήμα 4- 6: Ασθένειες

Από τη διερεύνηση των ημερομηνιών μεταξύ προγραμματισμού του ραντεβού και της ημερομηνίας του ραντεβού φάνηκε όταν οι δύο ημερομηνίες είναι αυθημερόν το ποσοστό εμφάνισης να αγγίζει το 95%, ενώ όταν μεσολαβεί διάστημα άνω της 1 ημέρας το ποσοστό εμφάνισης στο ραντεβού να ανέρχεται σε 70%. Αυτό δείχνει τάση οι ασθενείς να εμφανίζονται στα ραντεβού τους όταν πραγματοποιούνται την ίδια ημέρα με αυτήν που το προγραμματίζουν.

Προεπεξεργασία Δεδομένων

Κατά την προεπεξεργασία των δεδομένων, αφαιρέθηκε μία εγγραφή με αρνητική τιμή στην ηλικία. Μετατράπηκαν σε 0,1 οι τιμές των ποιοτικών μεταβλητών ώστε να μπορούν να διαχειριστούν από τους αλγορίθμους. Σε αυτό το βήμα οι Αναπηρίες μετατράπηκαν σε 0 εάν δεν υπάρχει καμία ή 1 εάν εμφανίζεται τουλάχιστον μία αναπηρία. Δημιουργήθηκε η νέα μεταβλητή «Day_diff» που δείχνει τις ημέρες που μεσολάβησαν μεταξύ του προγραμματισμού

και της ημερομηνίας του ραντεβού. Επιπλέον, αφαιρέθηκαν οι εγγραφές που αφορούσαν τον ίδιο ασθενή κρατώντας μόνο τη μία φορά που προγραμμάτισε επίσκεψη επειδή θεωρήθηκε ότι εάν ένα άτομο έχει συγκεκριμένη συμπεριφορά απέναντι στην εμφάνιση ή όχι σε ένα ραντεβού αυτό μπορεί να προκαλέσει μία τάση στα δεδομένα η οποία δεν θα είναι η πραγματική. Ακολούθησε η αφαίρεση των μεταβλητών που αφορούσαν την ημερομηνία προγραμματισμού και ραντεβού, ο κωδικός του ασθενή και του ραντεβού, η γειτονιά. Τέλος, διενεργήθηκε η τυποποίηση των μεταβλητών.

Μοντέλα Μηχανικής Μάθησης

Εφαρμόστηκαν αλγόριθμοι Logistic Regression, Decision Tree, Random Forest και Gradient Boost και η σύγκριση μεταξύ τους έγινε αξιολογώντας τα μέτρα Accuracy, Precision, Recall, F-score.

Ο διαχωρισμός του συνόλου δεδομένων έγινε με αναλογία 75% για εκπαίδευση και 25% για επικύρωση.

Ακολούθησε επανάληψη της διαδικασίας με την τεχνική SMOTE (Synthetic minority oversampling technique) καθώς το σύνολο δεδομένων δεν ήταν ισορροπημένο ως προς το πλήθος εμφάνισης ή μη εμφάνισης ασθενούς στο ραντεβού του. Παρόλα αυτά δεν προτάθηκαν τα συγκεκριμένα μοντέλα με κύριο λόγο την υπερπροσαρμογή τους στα δεδομένα εκπαίδευσης μειώνοντας έτσι την ικανότητα των μοντέλων να γενικεύουν και να προβλέπουν σωστά νέες παρατηρήσεις.

Στον παρακάτω Πίνακα 4-2 παρουσιάζεται η απόδοση του κάθε αλγορίθμου πριν την εφαρμογή της τεχνικής SMOTE και μετά την εφαρμογή.

Πίνακας 4-2: Απόδοση μοντέλων ταξινόμησης

Αλγόριθμος	Τεχνική	Accuracy	Precision	Recall	F-score
Logistic Regression	original	76%	36%	3%	5%
	SMOTE	65%	36%	58%	44%
Decision Tree	original	72%	36%	25%	30%
	SMOTE	69%	35%	33%	34%
Random Forest	original	71%	36%	27%	31%
	SMOTE	68%	35%	40%	37%
Gradient Boost	original	76%	54%	2%	4%
	SMOTE	60%	35%	81%	49%

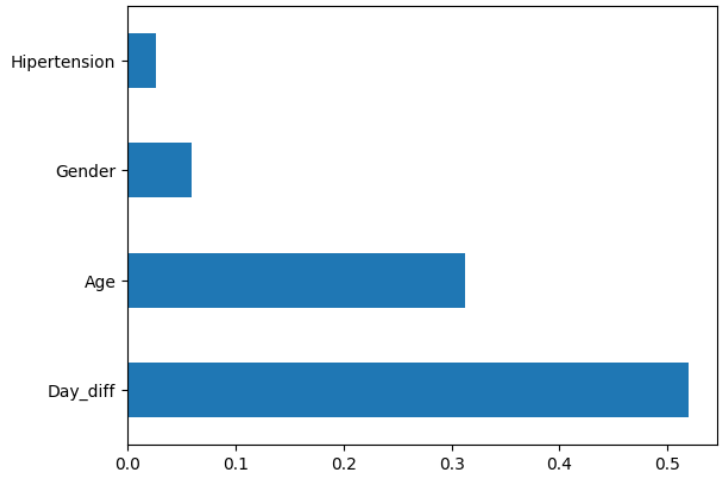
Με βάση τα αποτελέσματα φαίνεται να λειτουργούν καλύτερα ο Decision Tree και ακολουθεί ο Random Forest πριν την εφαρμογή SMOTE.

Ακολούθησε διασταυρούμενη επικύρωση 8 τμημάτων στους αλγορίθμους με την καλύτερη επίδοση και η ακρίβεια παρέμεινε το ίδιο.

Τα μοντέλα Decision Tree και Random Forest έχουν μέτρια ορθότητα πρόβλεψης 72% και 71% αντίστοιχα. Ο λόγος δημιουργίας του μοντέλου είναι η πρόβλεψη των ασθενών που δεν θα εμφανιστούν στο προγραμματισμένο τους ραντεβού τους ώστε η διοίκηση των μονάδων να κατανείμει αποτελεσματικότερα τις βάρδιες του προσωπικού τόσο για καλύτερη εξυπηρέτηση των ασθενών όσο και για μείωση κόστους αποφεύγοντας υπερωρίες σε πολυάσχολες ημέρες ή πολύ προσωπικό σε αδρανείς ημέρες.

Παρόλα αυτά τα μοντέλα μας παρουσιάζουν αδυναμία στη συγκεκριμένη πρόβλεψη κάτι που μπορεί να οφείλεται και στην ανισορροπία του συνόλου μας που υπερτερεί σε πλήθος ατόμων που παρευρέθηκαν στην επίσκεψή τους και αυτό το γεγονός μπορεί να επηρεάζει τις μετρικές υπέρ της πρόβλεψης ασθενών που παρευρίσκονται στα ραντεβού τους. Το αποτέλεσμα εφαρμόζοντας τον αλγόριθμο Decision Tree δείχνει ότι ταξινομήθηκε σωστά το 72% των περιπτώσεων. Πιο συγκεκριμένα, από το σύνολο των ασθενών που ταξινομήθηκαν ως πιθανό να εμφανιστούν, όντως το 79% παρευρέθηκε ενώ από αυτούς που είχε κριθεί πιθανό να μην εμφανιστούν μόνο το 36% προβλέφθηκε σωστά. Το μοντέλο έφτασε στο 25% στην αναγνώριση των ασθενών που δεν εμφανίστηκαν στο ραντεβού τους κάτι το οποίο δεν είναι πολύ ικανοποιητικό.

Από την εξέταση της σημαντικότητας των μεταβλητών του μοντέλου φαίνεται δύο να είναι οι μεταβλητές που επηρεάζουν περισσότερο στην εμφάνιση του ασθενούς στο ραντεβού, το χρονικό διάστημα που μεσολαβεί μεταξύ προγραμματισμού και ραντεβού να είναι η σημαντικότερη και ακολουθεί η ηλικία που είχε παρατηρηθεί στη διερευνητική ανάλυση ότι επηρεάζει. Οι επόμενες μεταβλητές που κρίθηκαν σημαντικές για το μοντέλο έχουν μικρότερη τιμή σε σχέση με τις παραπάνω δύο και είναι το φύλο και η υπέρταση παρόλο που δεν είχε διακριθεί αυτό από τη διερευνητική ανάλυση και ακολουθεί το Σχήμα 4-7 που τις απεικονίζει.



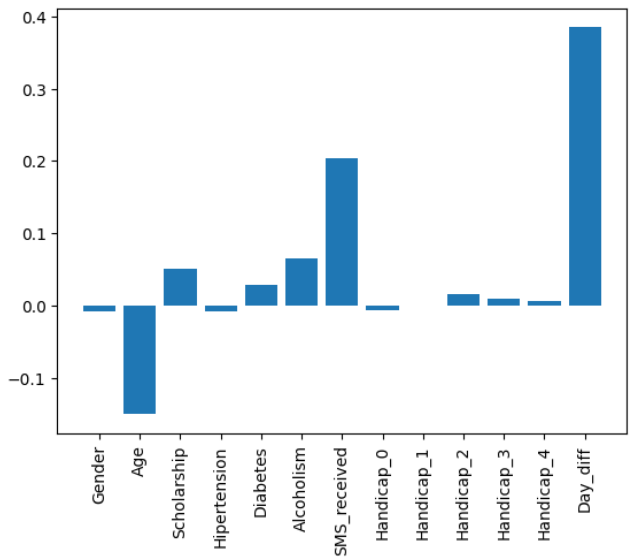
Σχήμα 4-7: Απεικόνιση των τεσσάρων πιο σημαντικών μεταβλητών του μοντέλου Decision Tree

Παρακάτω παρουσιάζεται ο σχετικός Πίνακας 4-3 ταξινόμησης του μοντέλου Decision Tree.

Πίνακας 4-3: Πίνακας ταξινόμησης του μοντέλου Decision Tree

#	Πρόβλεψη		Σύνολο
Πραγματικό	**No show**	**Show**	
No show	1078	3208	4286
Show	1897	11771	13668
Σύνολο	2975	14979	17954

Μετά την εφαρμογή της τεχνικής υπερδειγματοληψίας SMOTE ο Logistic Regression και Gradient Boost φαίνεται να αποδίδουν καλύτερα με ακρίβεια 65% και 60% αντίστοιχα. Το αποτέλεσμα εφαρμόζοντας τον αλγόριθμο Logistic Regression με SMOTE δείχνει ότι ταξινομήθηκε σωστά το 60% των περιπτώσεων. Πιο συγκεκριμένα, από το σύνολο των ασθενών που ταξινομήθηκαν ως πιθανό να εμφανιστούν, όντως το 84% παρευρέθηκε. Επίσης, το μοντέλο αναγνωρίζει το 58% του συνόλου των περιπτώσεων που είναι πιθανό να απουσιάσουν, αλλά αναγνωρίζει σωστά το 36% του συνόλου αυτών των περιπτώσεων. Σε αυτό το μοντέλο οι τρεις σημαντικότερες μεταβλητές βρέθηκαν να είναι η διαφορά ημερών μεταξύ προγραμματισμού και ραντεβού, εάν οι ασθενείς έλαβαν υπενθύμιση με μήνυμα και η ηλικία οι οποίες και απεικονίζονται στο παρακάτω Σχήμα 4-8.



Σχήμα 4-8: Απεικόνιση σημαντικότητας μεταβλητών του μοντέλου Logistic Regression

Στον παρακάτω Πίνακα 4-4 παρουσιάζεται η ταξινόμηση του μοντέλου Logistic Regression με SMOTE.

Πίνακας 4-4: Πίνακας ταξινόμησης του μοντέλου Logistic Regression με SMOTE

#	Πρόβλεψη		Σύνολο
Πραγματικό	**No show**	**Show**	
No show	3487	799	4286
Show	6353	7315	13668
Σύνολο	9840	8114	17954

Συμπεράσματα

Με εφαρμογή του αλγορίθμου Decision Tree το μοντέλο ταξινομεί σωστά το 72% των περιπτώσεων. Το ποσοστό αναγνώρισης των μη εμφανίσεων φτάνει το 36%, αλλά από την οπτική της πρόγνωσης των ασθενών που είναι πιθανό να εμφανιστούν στο ραντεβού το μοντέλο αγγίζει το 79%. Επιπλέον βρέθηκε ότι σημαντικός παράγοντας εμφάνισης είναι η ημέρα προγραμματισμού του ραντεβού να είναι κοντά χρονικά με την ημέρα του ραντεβού καθώς και η ηλικία του ασθενή. Σε σύντομο χρονικό διάστημα θα μπορούσε να γίνει σύγκριση της υφιστάμενης κατάστασης με το αποτέλεσμα του προγραμματισμού βάσει του μοντέλου και επιπλέον θα μπορούσε να αξιολογηθεί εκ νέου το μοντέλο έχοντας συλλέξει και περισσότερα δεδομένα.

4.3 2η Εφαρμογή – Πρόβλεψη Επανεμφάνισης ασθενή στο Νοσοκομείο εντός 30 ημερών

Πρόβλημα

Η ικανότητα πρόβλεψης επανεισαγωγών ασθενών δίνει στα Νοσοκομεία τη δυνατότητα έγκαιρης παρέμβασης ώστε να αποφευχθούν μελλοντικά απειλητικά περιστατικά υγείας μειώνοντας έτσι τον κίνδυνο για τους ασθενείς, ενισχύοντας την ποιότητα υπηρεσιών υγείας και μειώνοντας τα κόστη για το σύστημα υγείας

Σκοπός

Σκοπός της παρακάτω ανάλυσης είναι αρχικά η διερεύνηση των δεδομένων και στη συνέχεια η δημιουργία προβλεπτικού μοντέλου επανεμφάνισης ασθενών εντός 30 ημερών με βάση τα διαθέσιμα δεδομένα. Το μοντέλο θα συμβάλλει υπέρ του νοσηλευτικού προσωπικού σε βελτίωση διαδικασίας εξιτηρίου δίνοντας στοχευμένη φαρμακευτική αγωγή και συστηματική παρακολούθηση ασθενών με αυξημένες πιθανότητες επανεισαγωγής. Κατ' επέκταση αυτό μπορεί να οδηγήσει σε βελτίωση των παρεχόμενων υπηρεσιών υγείας και σε αποφυγή κόστους επανεισδοχής.

Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων αντλήθηκε από τον ιστότοπο προέρχεται της Kaggle. Αφορά μονάδες Νοσοκομείων στις Ηνωμένες Πολιτείες και περιέχει ανωνυμοποιημένα δεδομένα για 101.766 ασθενείς και πληροφορίες για το φύλο, την ηλικία, το βάρος, το είδος ασφάλισης που έχουν, μετρήσεις εργαστηριακών αιματολογικών εξετάσεων, αριθμό φαρμάκων που χρησιμοποιούν ή επεμβάσεων, το διάστημα νοσηλείας, εάν επανεισήχθησαν ή όχι. Συγκεκριμένα οι πενήντα μεταβλητές που περιέχονται στο σύνολο καθώς και ο τύπος τους περιγράφονται στον παρακάτω Πίνακα 4-5. Τα δεδομένα αφορούν διάστημα 10 ετών (1999-2008).

Πίνακας 4-5: Περιγραφή μεταβλητών του συνόλου δεδομένων

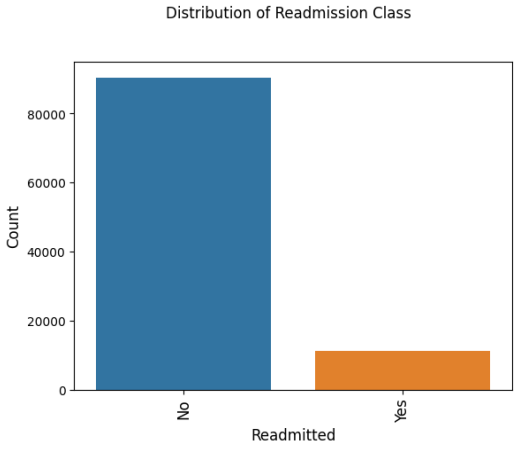
ΑΑ	Μεταβλητή	Περιγραφή	Τύπος
1	Encounter ID	Αναγνωριστικό εισαγωγής	Ποιοτική
2	Patient number	Αναγνωριστικό ασθενούς	Ποιοτική
3	Race	Φυλή	Ποιοτική
4	Gender	Φύλο	Ποιοτική
5	Age	Ηλικία	Ποσοτική
6	Weight	Ύψος	Ποσοτική
7	Admission type	Αντιστοιχεί σε 9 διακριτές τιμές, πχ έκτακτη ανάγκη,προαιρετικό, νεογένητο	Ποιοτική
8	Discharge disposition	Αντιστοιχεί σε 29 διακριτές τιμές-θέσεις όπου ο ασθενής μεταφέρεται κατά την έξοδό του πχ σπίτι, μη διαθέσιμη πληροφορία	Ποιοτική

9	Admission source	Αντιστοιχεί σε 21 διακριτές τιμές, πχ παραπομπή γιατρού, έκτακτη ανάγκη, μεταφορά από νοσοκομείο	Ποιοτική
10	Time in hospital	Αριθμός ημερών μεταξύ εισαγωγής και εξόδου	Ποσοτική
11	Payer code	Αντιστοιχεί σε 21 διακριτές τιμές, πχ ασφαλιστική, αυτοπληρωμή	Ποιοτική
12	Medical specialty	Αντιστοιχεί σε 84 διακριτές τιμές πχ καρδιολογία, γενική ιατρική	Ποιοτική
13	Number of lab procedures	Αριθμός εργαστηριακών εξετάσεων κατά την εισαγωγή	Ποσοτική
14	Number of procedures	Αριθμός εξετάσεων (εκτός των εργαστηριακών αναλύσεων) κατά την εισαγωγή	Ποσοτική
15	Number of medications	Αριθμός διαφορετικών φαρμάκων που δόθηκαν κατά την παραμονή	Ποσοτική
16	Number of outpatient visits	Αριθμός επισκέψεων του ασθενούς σε εξωτερικούς ασθενείς την προηγούμενη χρονιά της εισαγωγής	Ποσοτική
17	Number of emergency visits	Αριθμός επειγουσών επισκέψεων του ασθενούς την προηγούμενη χρονιά της εισαγωγής	Ποσοτική
18	Number of inpatient visits	Αριθμός επισκέψεων του ασθενούς σε ενδοσοκομειακούς ασθενείς την προηγούμενη χρονιά της εισαγωγής	Ποσοτική
19	Diagnosis 1	Αντιστοιχεί σε 848 διακριτές τιμές κωδικοποιημένης ονομασίας της κύριας διάγνωσης	Ποιοτική
20	Diagnosis 2	Αντιστοιχεί σε 923 διακριτές τιμές κωδικοποιημένης ονομασίας της δευτερεύουσας διάγνωσης	Ποιοτική
21	Diagnosis 3	Αντιστοιχεί σε 954 διακριτές τιμές κωδικοποιημένης ονομασίας της επιπλέον της δευτερευούσης διάγνωσης	Ποιοτική
22	Number of diagnoses	Αριθμός διαγνώσεων	Ποσοτική
23	Glucose serum test result	Υποδεικνύει το εύρος του αποτελέσματος εάν έγινε η μέτρηση με τιμές «>200», «>300», «κανονικό» ή «κανένα»	Ποιοτική
24	A1c test result	Υποδεικνύει το εύρος του αποτελέσματος εάν έγινε η μέτρηση με	Ποιοτική

		τιμές «>7», «>8», «κανονικό» ή «κανένα»	
25	23 features for medications	Υποδεικνύει για 23 ουσίες εάν συνταγογραφήθηκαν και εάν υπήρξε αλλαγή στη δοσολογία με τιμές «πάνω», «κάτω», «σταθερή», «όχι»	Ποιοτική
26	Change of medications	Υποδεικνύει εάν υπήρξε αλλαγή στα διαβητικά φάρμακα είτε στη δοσολογία είτε στην ονομασία του φαρμάκου με τιμές «αλλαγή» και «καμία αλλαγή»	Ποιοτική
27	Diabetes medications	Υποδεικνύει εάν συνταγογραφήθηκε διαβητικό φάρμακο με τιμές «ναι», «όχι»	Ποιοτική
28	Readmitted	Υποδεικνύει εάν υπήρξε επανεισδοχή του ασθενούς ή όχι σύμφωνα με το αρχείο με τιμές «<30», «>30», «όχι»	Ποιοτική

Διερευνητική Ανάλυση

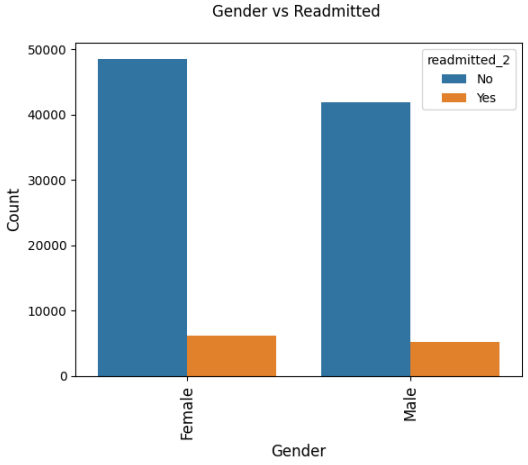
Από τη διερευνητική ανάλυση βρέθηκε ότι στο σύνολο των 101.766 εγγραφών, οι 71.518 εγγραφές (70%) αφορούν μοναδικό ασθενή και οι 54.745 ασθενείς (53%) έχουν μόνο μία εισαγωγή στο νοσοκομείο. Από τις 101.766 εγγραφές η αναλογία Επανεισαγωγής ή όχι στο Νοσοκομείο (σε διάστημα < 30 ημερών) είναι περίπου 13- 87 % το οποίο παρουσιάζεται στο παρακάτω Σχήμα 4-9.



Σχήμα 4-9: Αναλογία επανεισαγωγής εντός 30 ημερών

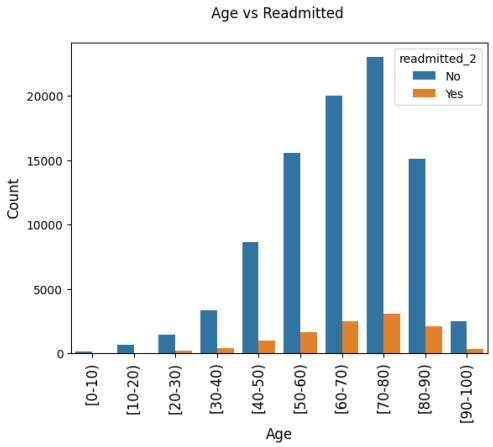
Από διερεύνηση της φυλής των ασθενών προκύπτει ότι η πλειοψηφία είναι Καυκάσιοι (άνω των 70.000) και ακολουθούν οι Αφροαμερικανοί (περίπου 20.000). Σχετικά με την αναλογία φύλου, είναι περίπου ίση με τις γυναίκες να είναι ελαφρώς περισσότερες (53% -

47%). Δεν παρουσιάζεται σημαντική επιρροή του φύλου στην επανεισδοχή στο νοσοκομείο και αυτό παρουσιάζεται στο επόμενο Σχήμα 4-10.



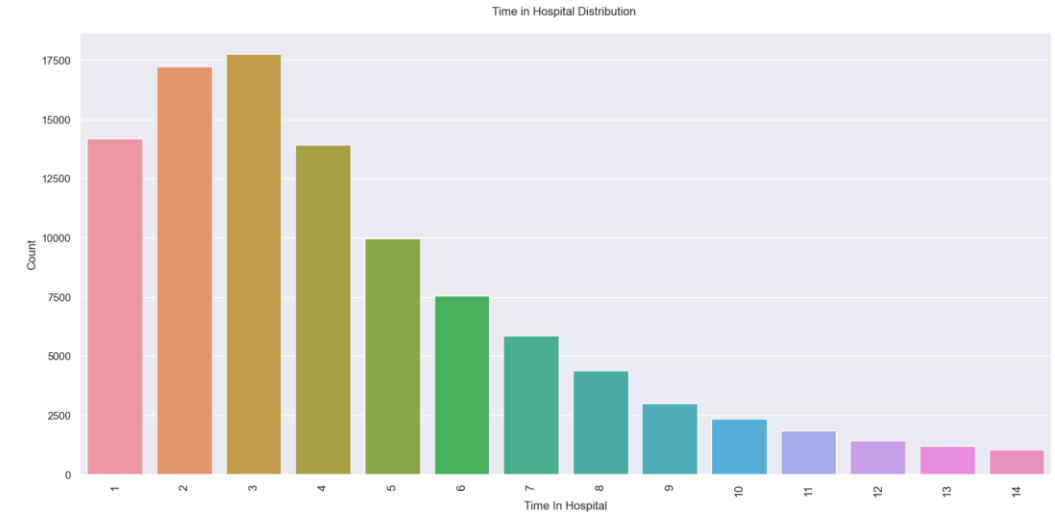
Σχήμα 4-10: Επανεισδοχή σε σχέση με το φύλο

Κατά τη διερεύνηση παρατηρήθηκε ότι οι επανεισαγωγές ασθενών παρατηρούνται κυρίως μεταξύ των ηλικιών από 50 έως 90 έτη, αλλά το 80% των ασθενών του συνόλου δεδομένων ανήκει σε αυτές τις ηλικίες και αυτό απεικονίζεται στο παρακάτω Σχήμα 4-11.

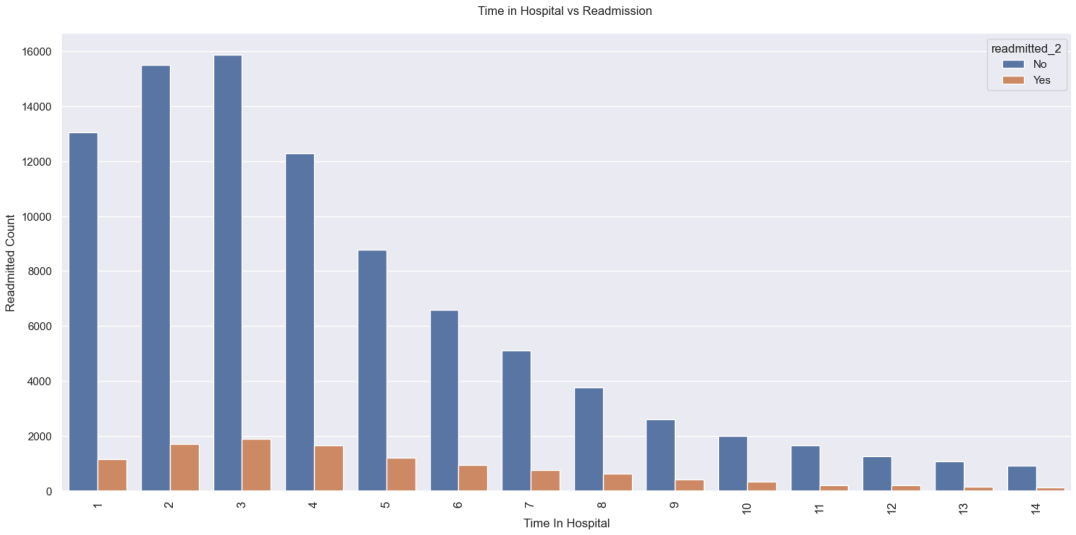


Σχήμα 4-11: Επανεισδοχή σε σχέση με την ηλικία

Σχετικά με τις ημέρες διαμονής στο νοσοκομείο, βρέθηκε ότι η μέση τιμή είναι 4,4 ημέρες, ενώ η πλειοψηφία επανεισαγωγών εντοπίζεται σε ασθενείς που η διαμονή τους ήταν από 1 έως 5 ημέρες. Παρακάτω ακολουθούν τα Σχήματα 4-12 και 4-13 όπου απεικονίζουν τα ευρήματα αντίστοιχα.

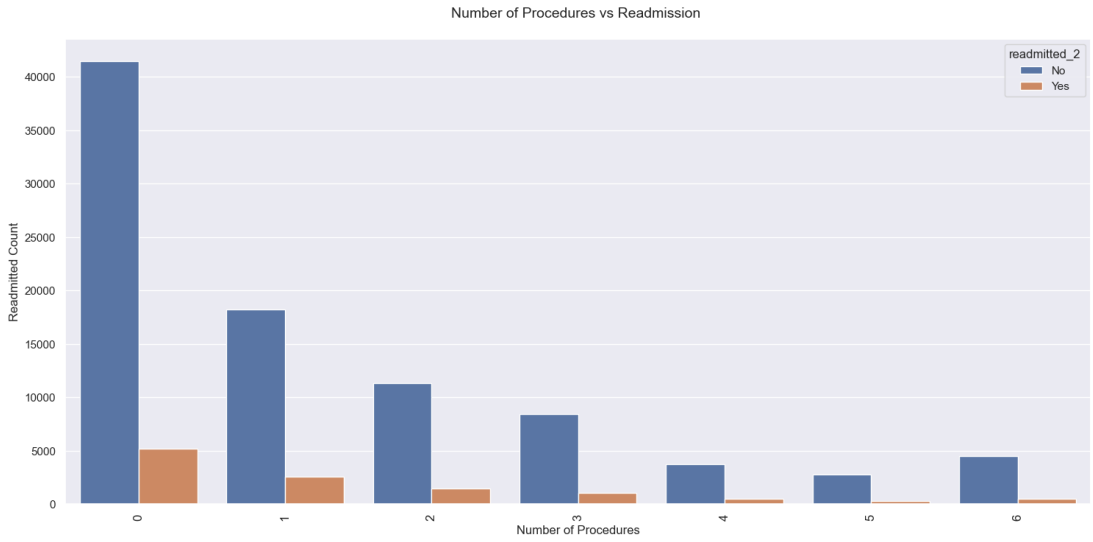


Σχήμα 4-12: Ημέρες διαμονής στο νοσοκομείο



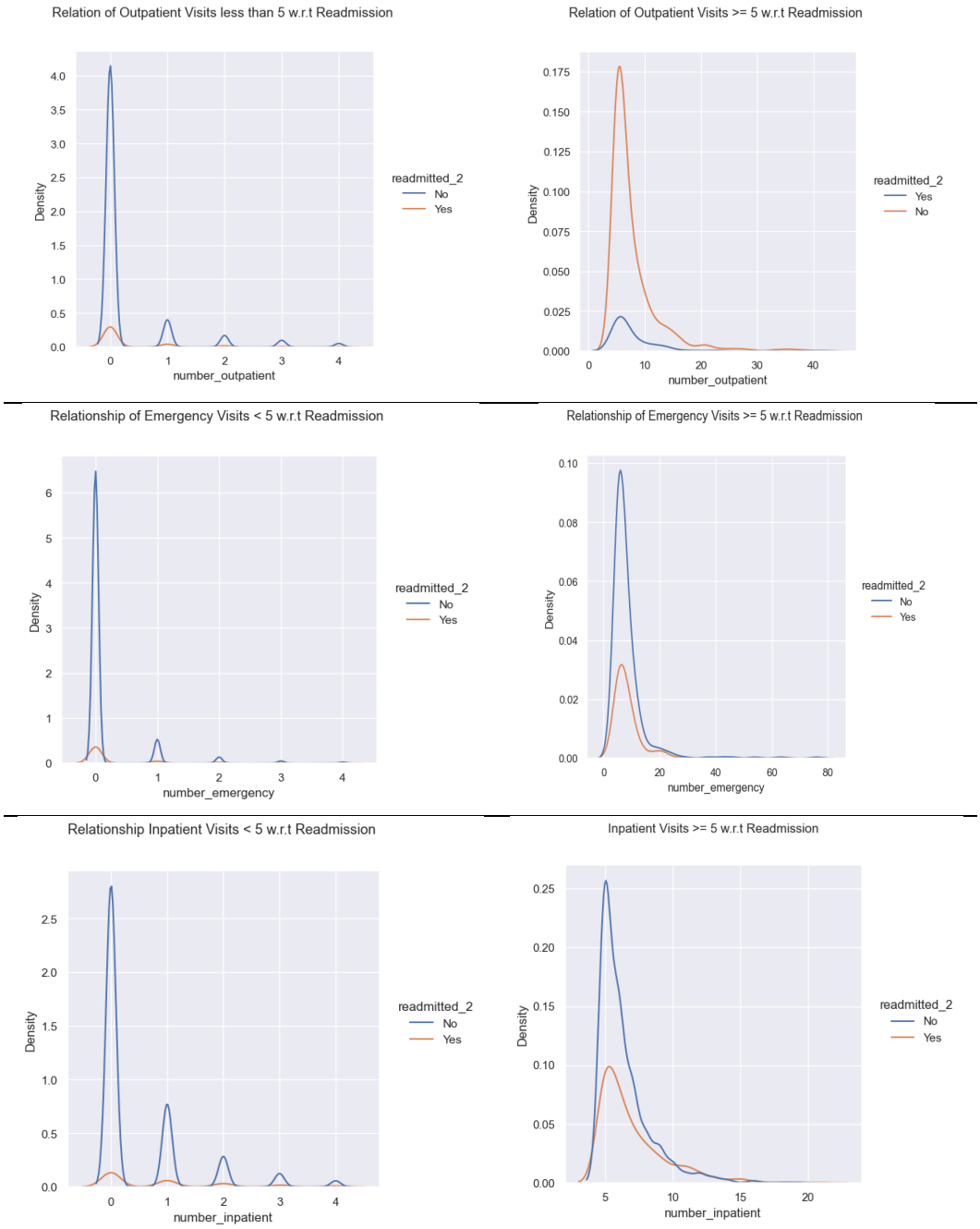
Σχήμα 4-13: Απεικόνιση ημερών διαμονής και Επανεισαγωγή

Επίσης, μελετήθηκε διαγραμματικά η ύπαρξη σχέσης μεταξύ επανεισαγωγών και αριθμού εργαστηριακών εξετάσεων κατά τη διαμονή του ασθενή στο νοσοκομείο, αριθμού άλλων εξετάσεων (εκτός των εργαστηριακών) και παρουσιάζονται στο Σχήμα 4-14. Παρατηρούνται περισσότερες επανεισαγωγές σε ασθενείς που τους έγινε μία ή καμία εξέταση.



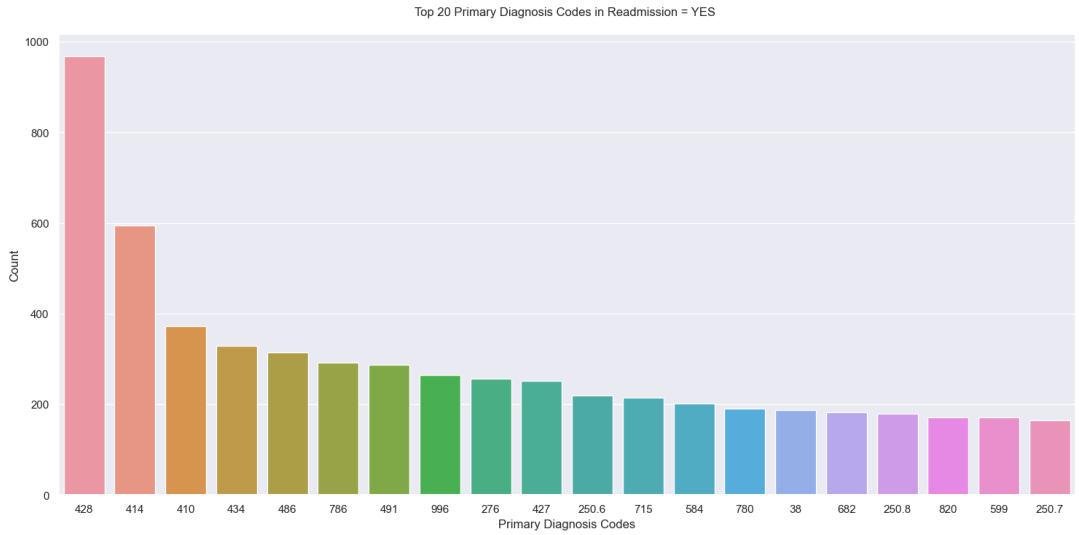
Σχήμα 4-14: Εξετάσεις και Επανεισαγωγή

Παρακάτω μελετήθηκε μέσω διαγραμμάτων ο αριθμός επισκέψεων των ασθενών στο νοσοκομείο σε εξωτερικά ιατρεία, σε επείγονται και εντός του νοσοκομείου την προηγούμενη χρονιά από αυτήν της νοσηλείας τους και παρουσιάζονται στο Σχήμα 4-15. Παρατηρείται συχνότερη εμφάνιση για επισκέψεις από 5 φορές και άνω σε σχέση με έως τέσσερεις επισκέψεις στους ασθενείς που επανεισήχθησαν.



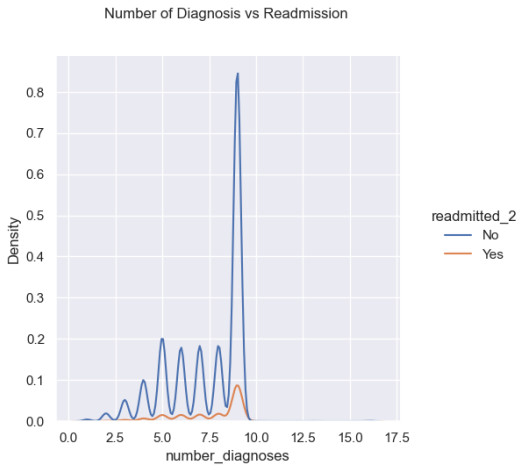
Σχήμα 4-15: Σχέση επισκέψεων προηγούμενης χρονιάς και επανεισαγωγής

Από τη διερεύνηση της Διάγνωσης σε σχέση με την επανεισδοχή στο νοσοκομείο (Σχήμα 4-16) παρατηρήθηκε ότι ασθενείς με καρδιακά προβλήματα είναι πιθανότερο να επιστρέψουν στο νοσοκομείο εντός των 30 ημερών (προηγείται η Συμφορητική καρδιακή ανεπάρκεια και έπεται η Ισχαιμική καρδιοπάθεια).



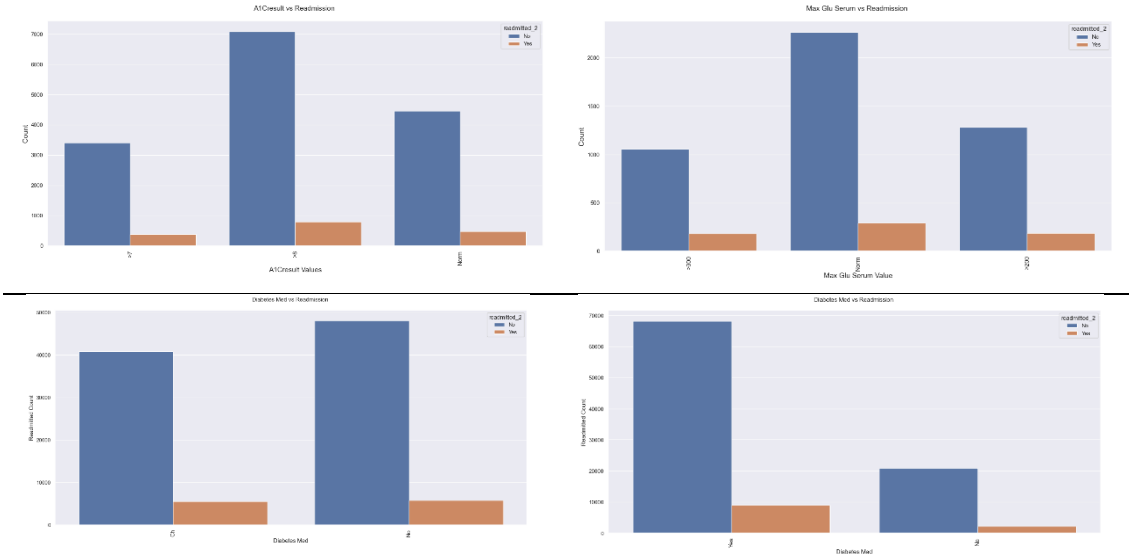
Σχήμα 4-16: Ασθένειες με μεγαλύτερη πιθανότητα επανεισδοχής

Από τη διερεύνηση του αριθμού διαγνώσεων σε σχέση με την επανεισδοχή ασθενή (Σχήμα 4-17) εντοπίστηκε μεγαλύτερη τάση για αυξημένο αριθμό διαγνώσεων από πέντε έως δέκα σε σχέση με λιγότερες παθήσεις.



Σχήμα 4-17: Αριθμός διαγνώσεων και επανεισαγωγή ασθενή

Από τη διερεύνηση τιμών για A1c, επίπεδα γλυκόζης, πιθανότητα για διαβητική αγωγή και ύπαρξη συνταγογραφημένων διαβητικών φαρμάκων σε σχέση με την επανεισαγωγή ασθενών εντοπίζεται ότι ασθενείς με διαβήτη έχουν μεγαλύτερα ποσοστά επανεισδοχής και παρουσιάζεται στο παρακάτω Σχήμα 4-18.



Σχήμα 4-18: Μετρήσεις επιπέδων A1c, γλυκόζης και διαβήτη για επανεισαγωγή ασθενών

Κατά τη διερεύνηση συσχέτισης μεταξύ των ανεξάρτητων αριθμητικών μεταβλητών δεν προέκυψε ισχυρή σχέση.

Προεπεξεργασία Δεδομένων

Στο σύνολο δεδομένων υπήρχαν τρεις ετικέτες για την επανεισδοχή ασθενών, εάν ένας ασθενής δεν εισήχθη εντός 30 ημερών από την ημέρα εξιτηρίου, εάν επέστρεψε εντός 30 ημερών και εάν επέστρεψε σε διάστημα άνω των 30 ημερών. Σε αυτή την ανάλυση θεωρήσαμε ως επανεισδοχή ασθενή μόνο τις εγγραφές που αφορούσαν διάστημα επιστροφής τις έως 30 ημέρες. Στη μεταβλητή ηλικία αφαιρέσαμε τις εγγραφές που δεν είχαν έγκυρη τιμή. Το ίδιο έγινε με τον αριθμό διαγνώσεων των ασθενών. Επιπλέον, αφαιρέθηκαν μεταβλητές που είχαν μεγάλο αριθμό ελλειπουσών τιμών, όπως το βάρος, το είδος ασφαλιστικής κάλυψης, το ιατρικό τμήμα. Αντίστοιχα αφαιρέθηκαν μεταβλητές που δε θα χρησιμοποιηθούν σε αυτή την ανάλυση όπως ο κωδικός ασθενή, ο κωδικός εισαγωγής του ασθενή και η αρχική μεταβλητή επανεισαγωγής με τις τρεις ετικέτες.

Ακολούθησε η μετατροπή σε κατηγορικές μεταβλητές των απαραίτητων μεταβλητών όπως η φυλή, το φύλο, η ηλικία.

Τέλος, ακολούθησε η τυποποίηση των μεταβλητών.

Μοντέλα Μηχανικής Μάθησης

Εφαρμόστηκαν αλγόριθμοι Logistic Regression, Extreme Gradient Boost, Random Forest και Support Vector Machines και η σύγκριση μεταξύ τους έγινε αξιολογώντας τα μέτρα

Accuracy, Precision, Recall, F-score. Ο διαχωρισμός του συνόλου δεδομένων έγινε με αναλογία 75% για εκπαίδευση και 25% για επικύρωση.

Ακολούθησε επανάληψη της διαδικασίας με την τεχνική SMOTE καθώς το σύνολο δεδομένων δεν ήταν ισορροπημένο ως προς το πλήθος ασθενών που επανεισήχθησαν και όσων δεν πραγματοποίησαν νέα εισαγωγή εντός 30 ημερών. Θα παρουσιαστούν τα αποτελέσματα για να μπορέσουν να συγκριθούν με αυτά πριν εφαρμοστεί η τεχνική, παρόλο που η τεχνική έχει επιδράσει στην αξιοπιστία των αποτελεσμάτων.

Στον παρακάνω Πίνακα 4-6 παρουσιάζεται η απόδοση του κάθε αλγορίθμου πριν την εφαρμογή της τεχνικής SMOTE και μετά την εφαρμογή.

Πίνακας 4-6: Απόδοση μοντέλων ταξινόμησης

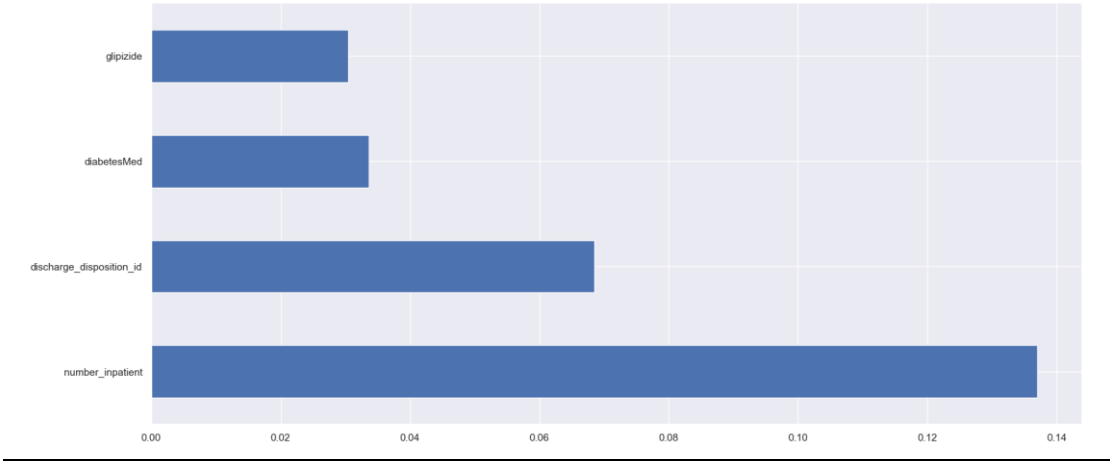
Αλγόριθμος	Τεχνική	Accuracy	Precision	Recall	F-score
Logistic Regression	original	89%	51%	2%	3%
	SMOTE	64%	16%	54%	25%
XGradient Boost	original	89%	48%	4%	7%
	SMOTE	89%	49%	3%	6%
Random Forest	original	89%	50%	0%	0%
	SMOTE	81%	21%	25%	23%

Με βάση τα αποτελέσματα φαίνεται να λειτουργούν καλύτερα ο Logistic Regression και ακολουθεί ο XGBoost πριν την εφαρμογή SMOTE. Τα μοντέλα Logistic Regression και XGBoost έχουν υψηλή ακρίβεια πρόβλεψης 89% και τα δύο.

Ο λόγος δημιουργίας του μοντέλου είναι η πρόβλεψη Επανεισαγωγής ενός ασθενή μετά το εξιτήριο του ώστε να βοηθηθεί το νοσηλευτικό προσωπικό με παρέχοντας στον ασθενή στοχευμένη φαρμακευτική αγωγή και συστηματική παρακολούθηση με σκοπό την αποφυγή του συμβάντος. Παρόλα αυτά τα μοντέλα μας παρουσιάζουν αδυναμία στη συγκεκριμένη πρόβλεψη κάτι που μπορεί να οφείλεται και στην ανισορροπία του συνόλου μας που υπερτερεί σε πλήθος ατόμων τα οποία δεν χρειάστηκαν επανεισαγωγή στο νοσοκομείο. Το αποτέλεσμα εφαρμόζοντας τον αλγόριθμο XGBoost δείχνει ότι ταξινομήθηκε σωστά το 89% των περιπτώσεων. Πιο συγκεκριμένα, από το σύνολο των ασθενών που ταξινομήθηκαν ως πιθανό να εμφανιστούν ξανά, όντως το 48% επέστρεψε ενώ από αυτούς που είχε κριθεί πιθανό να μην εμφανιστούν ξανά το 89% προβλέφθηκε σωστά. Το μοντέλο έφτασε στο 4% στην αναγνώριση των επανεισαγωγών κάτι το οποίο δεν είναι πολύ ικανοποιητικό.

Από την εξέταση της σημαντικότητας των μεταβλητών του μοντέλου φαίνεται δύο να είναι οι μεταβλητές που επηρεάζουν περισσότερο στην επανεισαγωγή του ασθενούς στο νοσοκομείο, ο αριθμός επισκέψεων του ασθενούς στο νοσοκομείο την προηγούμενη χρονιά και το μέρος διαμονής του ασθενούς μετά το εξιτήριο του. Οι επόμενες μεταβλητές που

κρίθηκαν σημαντικές για το μοντέλο έχουν μικρότερη τιμή σε σχέση με τις παραπάνω δύο και είναι η φαρμακευτική αγωγή για διαβήτη (είχε διακριθεί και κατά τη διερευνητική ανάλυση) και η παρουσία της φαρμακευτικής ουσίας glipizide τα οποία παρουσιάζονται στο παρακάτω Σχήμα 4-19.



Σχήμα 4-19: Απεικόνιση των τεσσάρων πιο σημαντικών μεταβλητών του μοντέλου XGBoost

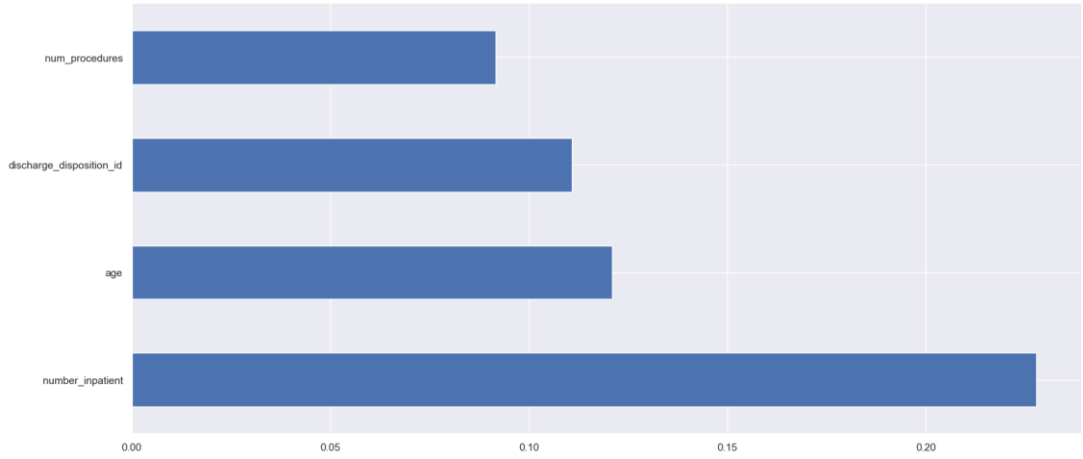
Παρακάτω παρουσιάζεται ο σχετικός Πίνακας 4-7 ταξινόμησης του μοντέλου XGBoost.

Πίνακας 4-7: Πίνακας ταξινόμησης του μοντέλου XGBoost

#	Πρόβλεψη		Σύνολο
Πραγματικό	**Readmitted**	**Not Readmitted**	
Readmitted	100	2675	2775
Not Readmitted	110	22176	22286
Σύνολο	210	24851	25061

Μετά την εφαρμογή της τεχνικής υπερδειγματοληψίας SMOTE ο XGBoost και ο Random Forest φαίνεται να αποδίδουν καλύτερα με ακρίβεια 89% και 81% αντίστοιχα. Το αποτέλεσμα εφαρμόζοντας τον αλγόριθμο XGBoost με SMOTE δείχνει ότι ταξινομήθηκε σωστά το 89% των περιπτώσεων, ενώ η προβλεπτική ικανότητα του μοντέλου φτάνει το 100% σε αναγνώριση των μη επανεισδοχών επί του συνόλου. Πιο συγκεκριμένα, από το σύνολο των ασθενών που ταξινομήθηκαν ως πιθανό να μην εμφανιστούν ξανά, όντως το 89% επαληθεύτηκε. Επίσης, το μοντέλο αναγνωρίζει το 49% του συνόλου των περιπτώσεων που θεωρήθηκε πιθανό να επανεισαχθούν, αλλά αναγνωρίζει σωστά μόνο το 3% του συνόλου αυτών των περιπτώσεων. Με την τεχνική SMOTE φαίνεται να αλλάζει η σημαντικότητα των μεταβλητών για το μοντέλο καθώς εμφανίζεται η ηλικία στη δεύτερη θέση, αλλά παραμένει

σημαντικότερη μεταβλητή ο αριθμός επισκέψεων του ασθενούς στο νοσοκομείο την προηγούμενη χρονιά. Ακολουθούν το μέρος διαμονής του ασθενούς μετά το εξιτήριο του και ο αριθμός των επεμβάσεων που έχει υποβληθεί ο ασθενής. Ακολουθεί το Σχήμα 4-20 με την απεικόνιση των μεταβλητών.



Σχήμα 4-20: Απεικόνιση σημαντικότητας μεταβλητών του μοντέλου Logistic Regression

Παρακάτω παρουσιάζεται ο σχετικός Πίνακας 4-8 ταξινόμησης του μοντέλου XGBoost με SMOTE.

Πίνακας 4-8: Πίνακας ταξινόμησης του μοντέλου XGBoost με SMOTE

#	Πρόβλεψη		Σύνολο
Πραγματικό	**Readmitted**	**Not Readmitted**	
Readmitted	84	2691	2775
Not Readmitted	87	22199	22286
Σύνολο	171	24890	25061

Συμπεράσματα

Το μοντέλο XGBoost θα μπορούσε να βοηθήσει τη νοσοκομειακή μονάδα να εντοπίσει ασθενείς με πιθανότητα επανεισδοχής τους και θα τους έδινε τη δυνατότητα να λάβουν μέτρα ώστε να το αποφύγουν. Από την οπτική της προβλεπτικής ικανότητα σε επίπεδο 89% μεταξύ των πιθανών μη εμφανίσεων και το 100% που φτάνει το μοντέλο σε αναγνώριση των μη επανεισδοχών επί του συνόλου, το ιατρικό προσωπικό θα μπορούσε να κρίνει ότι δε θα χρειαστούν επιπλέον συστηματική παρακολούθηση ή εξειδικευμένη φαρμακευτική αγωγή. Από την οπτική της προβλεπτικής ικανότητας του μοντέλου για τους ασθενείς που είναι πιθανό

να επανεισαχθούν , το ιατρικό προσωπικό θα μπορούσε να λάβει μέτρα για συστηματική παρακολούθηση ή εξειδικευμένη φαρμακευτική αγωγή και σε βάθος χρόνου να αξιολογήσει εκ νέου το μοντέλο έχοντας περισσότερα δεδομένα και παράλληλα να αξιολογήσει εάν αυτά τα προληπτικά μέτρα είχαν κέρδος ή ζημία σε σχέση με το εάν δεν είχαν ληφθεί.

4.4 3η Εφαρμογή – Διερευνητική μελέτη χρόνου κράτησης Αιθουσών Χειρουργείου

Πρόβλημα

Οι μεγάλες Λίστες Αναμονής για τα χειρουργεία φαίνεται να προβληματίζουν τις χώρες, ανάμεσα τους και η Ελλάδα η οποία κρίνει ότι θα χρειαστούν μέτρα για τη μείωση τους. Η Λίστα αναμονής υπήρχε και πριν την πανδημία, αλλά εκείνο το διάστημα οι επεμβάσεις μειώθηκαν επιπλέον 80% λόγω αποφυγής μετάδοσης του ιού.

Σκοπός

Σκοπός της παρακάτω ανάλυσης είναι η διερεύνηση των δεδομένων με σκοπό να ανακαλυφθούν τάσεις λαμβάνοντας υπόψη πολύπλοκες σχέσεις μεταξύ τους. Έτσι μπορεί να επιτευχθεί βέλτιστη αξιοποίηση της διαθεσιμότητας των χειρουργικών αιθουσών. Αυτό μπορεί να οδηγήσει σε μείωση υπερωριών του προσωπικού και κατά συνέπεια οικονομικό όφελος.

Περιγραφή Συνόλου Δεδομένων

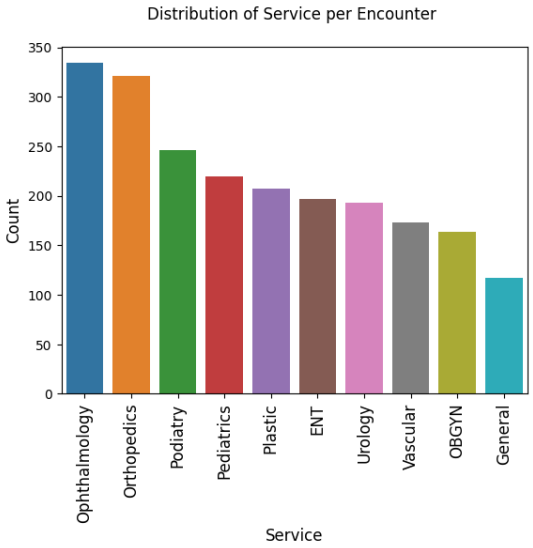
Το σύνολο δεδομένων αντλήθηκε από τον ιστότοπο της Kaggle. Περιέχει ανωνυμοποιημένα δεδομένα για 2172 επεμβάσεις και πληροφορίες για τον κωδικό επέμβασης, ημερομηνία επέμβασης, νούμερο χειρουργικής αίθουσας, ιατρική ειδικότητα και κωδικό ιατρικής ειδικότητας, κατηγορία επέμβασης, προτεινόμενο χρόνο επέμβασης με βάση την κατηγορία, και ώρα για είσοδο ασθενή στη χειρουργική αίθουσα, έναρξη επέμβασης, λήξη επέμβασης, έξοδο ασθενή από τη χειρουργική αίθουσα. Συγκεκριμένα οι δώδεκα μεταβλητές που περιέχονται στο σύνολο καθώς και ο τύπος τους περιγράφονται στον παρακάτω Πίνακα 4-9. Τα δεδομένα αφορούν διάστημα 3 μηνών (Ιανουάριος έως Μάρτιος 2022).

Πίνακας 4-9: Περιγραφή μεταβλητών του συνόλου δεδομένων

ΑΑ	Μεταβλητή	Περιγραφή	Τύπος
1	Encounter ID	Αναγνωριστικό επέμβασης	Ποιοτική
2	Date	Ημερομηνία επέμβασης	Ποσοτική
3	OR Suite	Αριθμός χειρουργείου	Ποιοτική
4	Service	Ιατρική ειδικότητα	Ποιοτική
5	CPT Code	Κωδικός ιατρικής ειδικότητας	Ποιοτική
6	CPT Description	Κατηγορία επέμβασης	Ποιοτική
7	Booked Time (min)	Καθορισμένος χρόνος σε λεπτά	Ποσοτική
8	OR Schedule	Προγραμματισμένη ημερομηνία και ώρα έναρξης επέμβασης	Ποσοτική
9	Wheels In	Ώρα εισόδου ασθενή στην αίθουσα	Ποσοτική
10	Start Time	Ώρα έναρξης επέμβασης	Ποσοτική
11	End Time	Ώρα λήξης επέμβασης	Ποσοτική
	Wheels Out	Ώρα εξόδου ασθενή από την αίθουσα	Ποσοτική

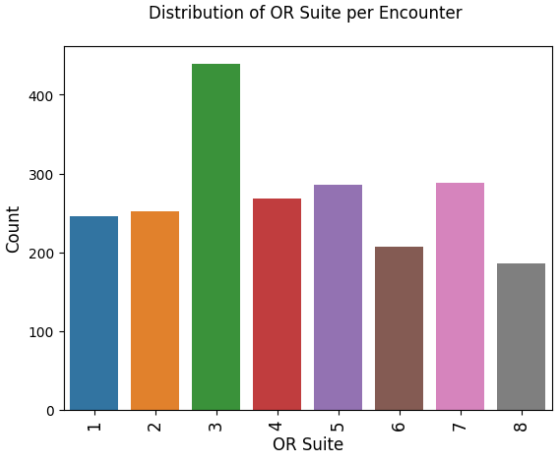
Διερευνητική Ανάλυση

Έγινε διερεύνηση του κωδικού επέμβασης εάν υπάρχει μόνο μία φορά ο κάθε κωδικός στο σύνολο και προέκυψε ότι δεν υπάρχουν πολλαπλές εγγραφές. Διερευνήθηκαν οι ιατρικές ειδικότητες του συνόλου και το πλήθος των επεμβάσεων για κάθε μία από αυτές. Οι ειδικότητες οι οποίες παρουσιάζονται στο παρακάτω Σχήμα 4-21 είναι δέκα και συγκεκριμένα η οφθαλμολογική για την οποία πραγματοποιήθηκαν 334 επεμβάσεις, η ορθοπεδική όπου καταγράφηκαν 321 επεμβάσεις, η ποδιατρική με 246 επεμβάσεις, η παιδιατρική με 220 επεμβάσεις, η πλαστική με 207 επεμβάσεις, η ω.ρι.λα με 197 επεμβάσεις, η ουρολογική με 193 επεμβάσεις, η αγγειοχειρουργική με 173 επεμβάσεις, η μαιευτική – γυναικολογική με 164 επεμβάσεις και η γενική με 117 επεμβάσεις.



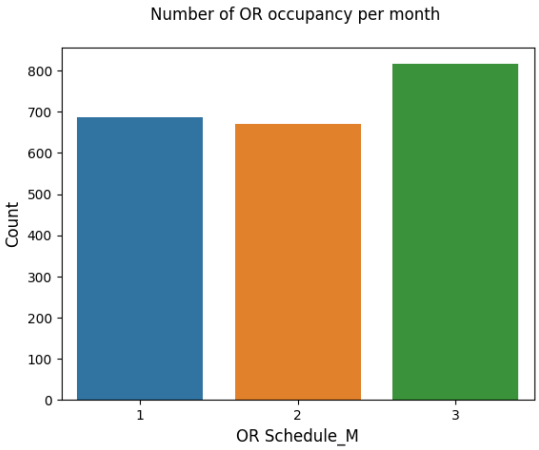
Σχήμα 4-21: Πλήθος επεμβάσεων ανά ειδικότητα

Οι χειρουργικές αίθουσες είναι οκτώ στο πλήθος και μετά από διερεύνηση εντοπίστηκε ότι οι περισσότερες επεμβάσεις πραγματοποιήθηκαν στην αίθουσα 3 (439) ενώ οι λιγότερες επεμβάσεις έγιναν στην αίθουσα 8 (186). Ακολουθεί η απεικόνιση στο Σχήμα 4-22.



Σχήμα 4-22: Πλήθος επεμβάσεων ανά χειρουργική αίθουσα

Διερευνήθηκε το πλήθος των επεμβάσεων ανά μήνα (Σχήμα 4-23) και προέκυψε ότι ο μήνας με τις περισσότερες επεμβάσεις ήταν ο Μάρτιος (815) ενώ ο μήνας με τις λιγότερες επεμβάσεις ήταν ο Φεβρουάριος (671).



Σχήμα 4-23: Πλήθος επεμβάσεων ανά μήνα

Η χρονική διάρκεια με βάση την οποία γίνεται η κράτηση της αίθουσας είναι τυποποιημένη ανάλογα την κατηγορία της κάθε ιατρικής ειδικότητας. Για τη διερεύνηση αποκλίσεων μεταξύ του τυποποιημένου προτεινόμενου χρόνου και του πραγματικού χρόνου διάρκειας των χειρουργικών επεμβάσεων χρειάστηκε να δημιουργηθεί μία νέα μεταβλητή «Actual_Time» η οποία είναι η διαφορά της ώρας εισόδου του ασθενή στη χειρουργική αίθουσα (Wheels In) και της ώρας εξόδου του ασθενή από τη χειρουργική αίθουσα (Wheels Out). Η μέση διάρκεια επέμβασης του συνόλου με βάση τους τυποποιημένους χρόνους κράτησης ισούται με 77 περίπου λεπτά, ενώ βάσει του πραγματικού χρόνου ισούται με 80 περίπου λεπτά. Στον παρακάτω Πίνακα 4-10 παρουσιάζονται οι μέσοι χρόνοι των επεμβάσεων για τον τυποποιημένο χρόνο κράτησης αίθουσας και για την πραγματική διάρκεια χρήσης της αίθουσας ανά ιατρική ειδικότητα.

Πίνακας 4-10: Μέσος χρόνος κράτησης χειρουργικής αίθουσας (εκτιμώμενος και πραγματικός)

	Booked Time (min)	Actual Time (min)
Οφθαλμολογική	44,64	35,87
Ορθοπεδική	87,38	100,96
Ποδιατρική	89,51	94,33
Παιδιατρική	60	66
Πλαστική	110,43	103,42
ΩΡΛ	67	69
Ουρολογική	66,06	70,76
Αγγεία	68,24	81,18
Μαιευτ.-Γυναικολ.	97,5	91,75
Γενική	110	113
Γενικός Μέσος Χρόνος	77,18	79,69

Συμπεράσματα

Από τον παραπάνω Πίνακα 4-10 παρατηρείται ότι εάν ο εκτιμώμενος χρόνος που χρησιμοποιείται για την κράτηση μίας αίθουσας χειρουργείου είναι δυναμικός και υπολογίζεται ανά τακτά χρονικά διαστήματα από το μέσο όρο των πραγματικών χρόνων χρήσης μίας αίθουσας τότε ενδέχεται οι αποκλίσεις να είναι μικρότερες και να βελτιωθεί ο προγραμματισμός και κατά συνέπεια να υπάρξει μείωση υπερωριών του προσωπικού αλλά και οικονομικό όφελος. Γενικά με την ανάλυση των δεδομένων από αντλούνται από διάφορες πηγές δίνεται η δυνατότητα για παράλληλη αξιολόγηση σε ιατρικά, διαχειριστικά, οικονομικά και επιδημιολογικά δεδομένα και έτσι οι ενδιαφερόμενοι μπορούν να προχωρήσουν σε προσαρμογές στο σύστημα υγείας.

ΚΕΦΑΛΑΙΟ 5

Συμπεράσματα

Σε αυτή την εργασία έγινε αναφορά σε μεθόδους αναλυτικής και μηχανικής μάθησης στη διοίκηση υπηρεσιών υγείας. Έγινε εκτενής αναφορά στους αλγορίθμους της μηχανικής μάθησης, στον τρόπο λειτουργίας τους και επιλεγμένοι αλγόριθμοι χρησιμοποιήθηκαν σε τρεις εφαρμογές-προβλήματα που αντιμετωπίζει η διοίκηση του χώρου υγείας και παράλληλα έρχονται αντιμέτωποι και οι ασθενείς με αυτά τα προβλήματα. Σκοπός της ανάλυσης των δεδομένων που αντλήθηκαν ήταν η βελτίωση της υφιστάμενης κατάστασης με προτάσεις, με τη χρήση αλγορίθμων κατηγοριοποίησης αλλά και με διερευνητική ανάλυση.

Οι αναλύσεις που έγιναν και οι μέθοδοι που χρησιμοποιήθηκαν στηρίχθηκαν στην βιβλιογραφική ανασκόπηση που παρουσιάστηκε στο κεφάλαιο 2 από την οποία φαίνεται ότι οι σύγχρονες τεχνικές μηχανικής μάθησης έχουν καλύτερες επιδόσεις και είναι πιο εύχρηστες από τα παραδοσιακά μοντέλα στατιστικής.

Για κάθε μία από τις τρεις εφαρμογές έγινε ανάλυση στο εκάστοτε σύνολο δεδομένων. Αρχικά με τη διερευνητική ανάλυση έγινε η κατανόηση των χαρακτηριστικών και των μεταβλητών του κάθε συνόλου. Επιπλέον, στην εφαρμογή για εμφάνιση ή απουσία ασθενούς από το προγραμματισμένο ραντεβού και στην εφαρμογή για επανεισδοχή στο νοσοκομείο εντός 30 ημερών, παρουσιάστηκαν συσχετίσεις και η σχέση της μεταβλητής στόχου με τις υπόλοιπες μεταβλητές.

Για δύο από τις παραπάνω εφαρμογές έγινε προεπεξεργασία των συνόλων δεδομένων, αφαιρέθηκαν μεταβλητές οι οποίες κρίθηκε ότι δεν συνεισφέρουν στην ανάλυση, άλλες μεταβλητές μετασχηματίστηκαν για να μπορέσουν να επεξεργαστούν κατάλληλα και τελικά να χρησιμοποιηθούν στους αλγορίθμους μηχανικής μάθησης. Αξιολογήθηκαν ο Logistic Regression, XGradient Boost, Random Forest, Decision Tree, Gradient Boost και σε κάθε εφαρμογή το προγνωστικό μοντέλο με την καλύτερη επίδοση ήταν διαφορετικό.

Πιο συγκεκριμένα, στην εφαρμογή για εμφάνιση ή απουσία ασθενούς από το προγραμματισμένο ραντεβού, το ποσοστό αναγνώρισης των μη εμφανίσεων έφτασε το 36% με Decision Tree, αλλά ακόμη και έτσι υπάρχει η πρόταση για βελτίωση του προγραμματισμού βαρδιών του προσωπικού. Επίσης κατά την ανάλυση αναδείχθηκαν σημαντικές μεταβλητές η ημέρα προγραμματισμού του ραντεβού καθώς και η ηλικία του ασθενή οπότε και αυτές οι πληροφορίες θα μπορούσαν να αξιοποιηθούν από τη διοίκηση.

Στην εφαρμογή για επανεισδοχή στο νοσοκομείο εντός 30 ημερών, το μοντέλο αναγνωρίζει σωστά μόνο το 4% των περιπτώσεων επανεισαγωγής εφαρμόζοντας XGBoost κάτι το οποίο δεν είναι καλό ποσοστό, αλλά φτάνει το 89% σε προβλεπτική ικανότητα μεταξύ των πιθανών μη εμφανίσεων, δίνοντας τη δυνατότητα στο ιατρικό προσωπικό να αναγνωρίσει περιπτώσεις όπου δε θα χρειαστεί επιπλέον συστηματική παρακολούθηση ή εξειδικευμένη φαρμακευτική αγωγή. Επιπλέον πληροφορία που αναδείχθηκε από το μοντέλο είναι ότι η σημαντικότερη μεταβλητή επανεμφάνισης του ασθενή είναι το πλήθος των επισκέψεων του ασθενούς στο νοσοκομείο την προηγούμενη χρονιά και έπεται το μέρος διαμονής του ασθενούς μετά το εξιτήριό του. Παρόλα αυτά, λόγω της απόδοσης του, το μοντέλο θα πρέπει να αξιολογηθεί εκ νέου.

Στην εφαρμογή της μελέτης του χρόνου κράτησης αιθουσών χειρουργείου πραγματοποιήθηκε διερευνητική ανάλυση με στόχο να παρατηρηθούν τάσεις μεταξύ των δεδομένων με σκοπό τη συμβολή στη βέλτιστη αξιοποίηση της διαθεσιμότητας μίας αίθουσας. Παρατηρήθηκε ότι οι εκτιμώμενοι χρόνοι που χρησιμοποιούνται για να γίνει η κράτηση της αίθουσας διαφέρουν από τον πραγματικό χρόνο είτε με μικρές είτε με μεγαλύτερες αποκλίσεις, κάτι το οποίο μπορεί να οφείλεται στην εξέλιξη της τεχνολογίας και στη χρήση πιο σύγχρονων μηχανημάτων και εργαλείων που απαιτούν λιγότερο χρόνο ολοκλήρωσης μίας επέμβασης. Εάν, λοιπόν, εκτιμώμενος χρόνος κράτησης της αίθουσας επιλέγεται από μία δυναμική λίστα η οποία τον υπολογίζει κάθε φορά βάσει των προηγούμενων πραγματικών δεδομένων, τότε αυτές οι αποκλίσεις θα είναι μικρότερες τουλάχιστον ως προς τις σύγχρονες τεχνολογίες. Ένα εμπόδιο που υπήρξε στη μελέτη αυτού του συνόλου ήταν το μικρό εύρος ημερομηνιών ΄καθώς αφορούσε διάστημα μόνο 3 μηνών, ενώ εάν το σύνολο ήταν χρονικά ευρύτερο ενδέχεται να είχαν παρατηρηθεί και άλλες τάσεις.

ΠΑΡΑΡΤΗΜΑΤΑ

Π1 ΠΗΓΑΙΟΣ ΚΩΔΙΚΑΣ ΣΕ ΡΥΤΗΟΝ ΓΙΑ ΤΗΝ 1η ΕΦΑΡΜΟΓΗ

```
#LIBRARIES
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from importlib import reload
%matplotlib inline

#READ FILE
csv_url = 'KaggleV2-May-2016.csv'
df = pd.read_csv(csv_url)
#ΠΕΡΙΓΡΑΦΙΚΑ
df.info()
df.describe()
for i in df.columns:
    print(i+":",len(df[i].unique()))
df.shape
corr = df.corr()
corr.style.background_gradient(cmap='coolwarm')
#Αφαίρεση αρνητικής τιμής ηλικίας
df=df.loc[( df.Age != -1) ]
df.describe()
#'ScheduledDay'and'AppointmentDay' data types are 'object', I'll change them into dates
df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'], format='%Y-%m-%d
%H:%M:%S')
df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'], format='%Y-%m-%d
%H:%M:%S')
print(df.ScheduledDay.max())
print(df.ScheduledDay.min())
print(df.AppointmentDay.max())
print(df.AppointmentDay.min())
#compare dates for checking a relation between dates and patients attendance to their
appointments
def ratio (x) :
    for i in x.columns :
        print(i , '\n\n' ,x[i].value_counts(normalize=True).mul(100).round(1).astype(str) + '%')
        print('\n')
# Unfiltered datasets - we have 80% attendance ratio to 20% absence
ratio(df)
showed = df['No-show'] == 'No'
```

```
not_showed = df['No-show'] == 'Yes'
df['showed'] = showed
df['not_showed'] = not_showed
df
def pie_plot(s):
    allP = df[s].value_counts()
    pieChart = allP.plot.pie(figsize=(10,10), autopct='%1.1f%%', fontsize = 18);
    pieChart.set_title("Status per appointment \n", fontsize = 15);
    plt.legend();
pie_plot('showed')
#match 'ScheduledDay' and 'AppointmentDay' to look if there is a relation when tha date
matches
Matching_Dates = df.loc[(df.ScheduledDay.dt.date == df.AppointmentDay.dt.date)]
ratio(Matching_Dates)
#match 'ScheduledDay' and 'AppointmentDay' to look if there is a relation when tha date
NOT matches
Not_Matching_Dates = df.loc[(df.ScheduledDay.dt.date != df.AppointmentDay.dt.date)]
ratio(Not_Matching_Dates)
#τα ορίσματα της στήλης No-show που είναι yes/no
attend = df.loc[(df["No-show"] == 'No')]
absent = df.loc[(df["No-show"] == 'Yes')]
#Age
attend_grouped_age = pd.DataFrame (pd.cut(attend["Age"], np.arange(0, 130, 5)))
attend_age = pd.DataFrame(attend_grouped_age.value_counts(normalize=True).mul(100))
absent_grouped_age = pd.DataFrame (pd.cut(absent["Age"], np.arange(0, 130, 5)))
absent_age = pd.DataFrame(absent_grouped_age.value_counts(normalize=True).mul(100))
age_merged = attend_age.merge(absent_age,how='outer',on='Age',suffixes=("_No",'_Yes'))
age_merged.sort_values('Age',axis=0,ascending=False).plot.bar(figsize=(10,8),ylim=(0, 10),
use_index=True,legend=False,stacked=False);
plt.legend(['Show', "No-show"]);
plt.ylabel=('Percentage');
plt.title('Distrbution of age across attendance');
#SMS Received
attend_grouped_SMS = pd.DataFrame (attend["SMS_received"])
attend_SMS = pd.DataFrame(attend_grouped_SMS.value_counts(normalize=True).mul(100))
absent_grouped_SMS = pd.DataFrame (absent["SMS_received"])
absent_SMS =
pd.DataFrame(absent_grouped_SMS.value_counts(normalize=True).mul(100))
SMS_merged =
attend_SMS.merge(absent_SMS,how='outer',on='SMS_received',suffixes=("_attend",'_absent
'))
SMS_merged.sort_values('SMS_received',axis=0,ascending=False).plot.bar(figsize=(5,5),yli
m=(0, 100), use_index=True,legend=False,stacked=False)
plt.legend(['Show', "No-show"]);
plt.title('Distrbution of SMS_received across attendance');
#Neighbourhood
attend_grouped_Neighbourhood = pd.DataFrame (attend["Neighbourhood"])
```

```
attend_Neighbourhood =
pd.DataFrame(attend_grouped_Neighbourhood.value_counts(normalize=True).mul(100))
absent_grouped_Neighbourhood = pd.DataFrame (absent["Neighbourhood"])
absent_Neighbourhood =
pd.DataFrame(absent_grouped_Neighbourhood.value_counts(normalize=True).mul(100))
Neighbourhood_merged =
attend_Neighbourhood.merge(absent_Neighbourhood,how='outer',on='Neighbourhood',suffix
es=("_attend",'_absent'))
Neighbourhood_merged.sort_values('Neighbourhood',axis=0,ascending=False).plot.bar(figsiz
e=(15,5),ylim=(0, 8), use_index=True,legend=False,stacked=False)
plt.legend(['Show', "No-show"]);
plt.title('Distrbution of Neighbourhood across attendance');
#Gender
plt.figure(figsize=(10,4))
ax1 = plt.subplot(1,3,1)
df.Gender.value_counts(normalize=True).mul(100).plot.pie(y = df["No-show"]
,subplots=True,figsize=(5, 5), ax=ax1)
plt.title('\ntotal sample gender distribution\n')
ax2 = plt.subplot(1,3,2)
absent.Gender.value_counts(normalize=True).mul(100).plot.pie(y = Matching_Dates["No-
show"] ,subplots=True,figsize=(5, 5), ax=ax2)
plt.title('\nabsent gender distribution\n')
ax3 = plt.subplot(1,3,3)
attend.Gender.value_counts(normalize=True).mul(100).plot.pie(y =
Not_Matching_Dates["No-show"] ,subplots=True,figsize=(5, 5), ax=ax3)
plt.title('\nattend gender distribution\n')
plt.tight_layout()
##conveting to boolean format
for i in df[['Scholarship', 'Hipertension', 'Alcoholism', 'Diabetes']]:
    df[i]=df[i].astype('bool')
# relationship between the diseases and the probability of patients showing up for their
scheduled appointments or not?
no_disease_showed_up = df[(df['No-show'] == False) & (df['Hipertension'] == False) &
(df['Diabetes'] == False) & (df['Alcoholism'] == False) & (df['Handcap'] == False)]
no_disease_no_show = df[(df['No-show'] == True) & (df['Hipertension'] == False) &
(df['Diabetes'] == False) & (df['Alcoholism'] == False) & (df['Handcap'] == False)]
amount_no_disease = no_disease_showed_up['AppointmentDay'].count() +
no_disease_no_show['AppointmentDay'].count()
amount_Hipertension = sum(df.groupby(['No-show']).sum()['Hipertension'])
amount_Diabetes = sum(df.groupby(['No-show']).sum()['Diabetes'])
amount_Alcoholism = sum(df.groupby(['No-show']).sum()['Alcoholism'])
amount_Handcap_combined = sum(df.groupby(['No-show']).sum()['Handcap'])
##Calculation the proportions of No-show and Show-up per disease and for no-disease
hipertension = df.groupby(['No-show']).sum()['Hipertension']/amount_Hipertension *100
diabetes = df.groupby(['No-show']).sum()['Diabetes']/amount_Diabetes * 100
alcoholism = df.groupby(['No-show']).sum()['Alcoholism']/amount_Alcoholism * 100
handcap_combined = df.groupby(['No-show']).sum()['Handcap']/amount_Handcap_combined
* 100
```

```
##Plotting results of calculations above.
# Styling the graphs
plt=reload(plt)
fig = plt.figure(figsize=(15,55))
## Defining the graphs
plt.subplot(7,2,1);
hipertension.plot.bar();
plt.title('Appointments of Patients with Hypertension on No-show')
plt.xlabel('No-show')
plt.ylabel('Percentage of Appointments')
## precentage of showed patient that has Hypertension is high compared to no-showed
patients
plt.subplot(7,2,2);
diabetes.plot.bar();
plt.title(' Appointments of Patients with Diabetes on No-show')
plt.xlabel('No-show')
plt.ylabel('Percentage of Appointments')
## precentage of showed patient that has Diabetes is high compared to no-showed patients
plt.subplot(7,2,3);
alcoholism.plot.bar();
plt.title(' Appointments of Patients with Alcoholism on No-show')
plt.xlabel('No-show')
plt.ylabel('Percentage of Appointments')
## precentage of showed patient that has Alcoholism is high compared to no-showed patients
plt.subplot(7,2,4);
handcap_combined.plot.bar();
plt.title(' Appointments of Patients with Handicap on No-show')
plt.xlabel('No-show')
plt.ylabel('Percentage of Appointments')
## precentage of showed patient that has Handicap is high compared to no-showed patients

#Preprocessing
df['No-show'] = df['No-show'].map({'No':0, 'Yes':1})
df['Gender'] = df['Gender'].map({'F':0, 'M':1})
##Convert to Categorical
df['Handcap'] = pd.Categorical(df['Handcap'])
##Convert to Dummy Variables
Handicap = pd.get_dummies(df['Handcap'], prefix = 'Handicap')
df = pd.concat([df, Handicap], axis=1)
df.drop(['Handcap'], axis=1, inplace = True)
## New Feature: Day Difference between AppointmentDay & ScheduledDay
df['Day_diff'] = (df['AppointmentDay'] - df['ScheduledDay']).dt.days
df['Day_diff'].unique()
#Duplicated values #number of duplicated raws
df.duplicated().sum()
#number of duplicated values in patient id
df['PatientId'].duplicated().sum() #out of 110527
#Number of duplicated patient Id AND No show TOGETHER
```

```
df[['PatientId','No-show']].duplicated().sum()
#we can drop them as we don`t need them because a patient habitat affects the pattern
df.drop_duplicates(['PatientId','No-show'],inplace=True)
df.shape
#now it`s not found
df[['PatientId','No-show']].duplicated().sum()
# Drop Columns
df.drop('AppointmentID', axis=1,inplace = True)
df.drop(['ScheduledDay'], axis=1, inplace=True)
df.drop('PatientId', axis=1,inplace = True)
df.drop('Neighbourhood', axis=1,inplace = True)
df.drop('AppointmentDay', axis=1,inplace = True)
df.drop('showed', axis=1,inplace = True)
df.drop('not_showed', axis=1,inplace = True)
df=df.copy() #too much errors, so that to reindex dataframe
df.info()

# Standardization / Split dataset into train & test set
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
scaler = StandardScaler()
X = df.drop("No-show", axis=1)
y = df["No-show"]
X = scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)
sm = SMOTE(random_state=42)
X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)

#Classification
#Logistic Regression
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(solver='newton-cg',)
lr.fit(X_train, y_train)
print(lr.score(X_train,y_train))
y_pred_lr = lr.predict(X_test)
clf_report = classification_report(y_test, y_pred_lr)
print(f"Classification Report : \n{clf_report}")

clf_report = confusion_matrix(y_test, y_pred_lr)
print(f"confusion_matrix : \n{clf_report}")
# plot feature importance
importance = lr.coef_[0]
plt.bar([x for x in range(len(importance))], importance)
columns_len = df.drop("No-show", axis=1).columns
tickvalues = range(0,len(columns_len))
plt.xticks(ticks = tickvalues, labels = columns_len, rotation = 'vertical')
```

```
plt.show()

#SMOTE - Logistic Regression
lr = LogisticRegression(solver='newton-cg',)
lr.fit(X_train_sm, y_train_sm)
print(lr.score(X_train_sm,y_train_sm))
y_pred_lr = lr.predict(X_test)
clf_report = classification_report(y_test, y_pred_lr)
print(f"Classification Report (SMOTE) : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_lr)
print(f"confusion_matrix (SMOTE) : \n{clf_report}")
# plot feature importance
importance = lr.coef_[0]
plt.bar([x for x in range(len(importance))], importance)
columns_len = df.drop("No-show", axis=1).columns
tickvalues = range(0,len(columns_len))
plt.xticks(ticks = tickvalues, labels = columns_len, rotation = 'vertical')
plt.show()


#Decision Tree
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(X_train, y_train)
y_pred_dtc = dtc.predict(X_test)
clf_report = classification_report(y_test, y_pred_dtc)
print(f"Classification Report : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_dtc)
print(f"confusion_matrix : \n{clf_report}")
#Plot Feature Importance
(pd.Series(dtc.feature_importances_, index=df.drop("No-show", axis=1).columns)
   .nlargest(4)
   .plot(kind='barh'))

#SMOTE - Decision Tree
dtc = DecisionTreeClassifier()
dtc.fit(X_train_sm, y_train_sm)
y_pred_dtc = dtc.predict(X_test)
clf_report = classification_report(y_test, y_pred_dtc)
print(f"Classification Report (SMOTE) : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_dtc)
print(f"confusion_matrix (SMOTE) : \n{clf_report}")
#Plot Feature Importance
(pd.Series(dtc.feature_importances_, index=df.drop("No-show", axis=1).columns)
   .nlargest(4)
   .plot(kind='barh'))

#Random Forest
```

```
from sklearn.ensemble import RandomForestClassifier
rd_clf = RandomForestClassifier()
rd_clf.fit(X_train, y_train)
y_pred_rd_clf = rd_clf.predict(X_test)
clf_report = classification_report(y_test, y_pred_rd_clf)
print(f"Classification Report : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_rd_clf)
print(f"confusion_matrix : \n{clf_report}")
#Plot Feature Importance
(pd.Series(rd_clf.feature_importances_, index=df.drop("No-show", axis=1).columns)
   .nlargest(4)
   .plot(kind='barh'))

#SMOTE - Random Forest
rd_clf = RandomForestClassifier()
rd_clf.fit(X_train_sm, y_train_sm)
y_pred_rd_clf = rd_clf.predict(X_test)
clf_report = classification_report(y_test, y_pred_rd_clf)
print(f"Classification Report (SMOTE) : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_rd_clf)
print(f"confusion_matrix (SMOTE) : \n{clf_report}")
#Plot Feature Importance
(pd.Series(rd_clf.feature_importances_, index=df.drop("No-show", axis=1).columns)
   .nlargest(4)
   .plot(kind='barh'))
#GBOOST
from sklearn.ensemble import GradientBoostingClassifier
gb = GradientBoostingClassifier()
gb.fit(X_train, y_train)
y_pred_gb = gb.predict(X_test)
clf_report = classification_report(y_test, y_pred_gb)
print(f"Classification Report : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_gb)
print(f"confusion_matrix : \n{clf_report}")
#Plot Feature Importance
(pd.Series(gb.feature_importances_, index=df.drop("No-show", axis=1).columns)
   .nlargest(4)
   .plot(kind='barh'))

#SMOTE - GBOOST
gb = GradientBoostingClassifier()
gb.fit(X_train_sm, y_train_sm)
y_pred_gb = gb.predict(X_test)
clf_report = classification_report(y_test, y_pred_gb)
print(f"Classification Report (SMOTE) : \n{clf_report}")
clf_report = confusion_matrix(y_test, y_pred_gb)
print(f"confusion_matrix (SMOTE) : \n{clf_report}")
#Plot Feature Importance
```

```
(pd.Series(gb.feature_importances_, index=df.drop("No-show", axis=1).columns)
   .nlargest(4)
   .plot(kind='barh'))

#Cross Validation
from sklearn.model_selection import cross_val_score
accuracy = cross_val_score(estimator = dtc, X = X, y =y, cv = 8)
print("avg acc: ",np.mean(accuracy))
print("acg std: ",np.std(accuracy))
accuracy = cross_val_score(estimator = rd_clf, X = X, y =y, cv = 8)
print("avg acc: ",np.mean(accuracy))
print("acg std: ",np.std(accuracy))
```

Π2 ΠΗΓΑΙΟΣ ΚΩΔΙΚΑΣ ΣΕ PYTHON ΓΙΑ ΤΗΝ 2η ΕΦΑΡΜΟΓΗ

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

#Read File
df = pd.read_csv('diabetic_data.csv')

#Perigrafika
df.shape
df.describe(include = 'all').T
# look for '?' in each column for null values
for i in df.columns:
    print(i, df[df[i] == '?'].shape[0])
# unique patients in the data
len(df['patient_nbr'].unique())
# only 1 encounter
df_encounters_check = df.groupby(['patient_nbr']).agg(encounters = ('encounter_id',
'count')).reset_index().sort_values(['encounters'], ascending = False)
df_encounters_check[df_encounters_check['encounters']==1]

#Visualization/Cleaning the Dataset
df['readmitted'].value_counts()
ax = sns.barplot(x=df['readmitted'].value_counts().index,   y=df['readmitted'].value_counts())
plt.xlabel('labels', size = 12)
plt.ylabel('# of Readmitted', size = 12)
plt.title('Class Distribution \n', size = 12)
plt.show()
# Created another column, Label the <30 and >30 as YES and Other "N0" as No
```

```
def check_label(text):
    if text == '<30':
        return 'Yes'
    else:
        return 'No'
df['readmitted_2'] =df['readmitted'].apply(check_label)
ax = sns.countplot(x='readmitted_2',   data= df)
plt.xlabel('Readmitted', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Distribution of Readmission Class  \n\n', size = 12)
plt.show()
#Race
df['race'].value_counts()
#Replace " ? " with Other
df.loc[df['race'] == '?', 'race'] = 'Other'
ax = sns.barplot(x=df['race'].value_counts().index,   y=df['race'].value_counts())
plt.xlabel('Race', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Distribution of Race of Patients \n', size = 12)
plt.show()
#Gender
df['gender'].value_counts()
ax = sns.countplot(x='gender',   data= df)
plt.xlabel('Gender', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Gender Distribution \n', size = 12)
plt.show()
# Drop the "Unknown/Invalid" gender of the data
df.drop(df[df['gender'] == 'Unknown/Invalid'].index, inplace = True)
df.reset_index(inplace = True, drop = True)
# Relationship of Gender and Readmitted Overall
ax = sns.countplot(x="gender", hue="readmitted_2", data=df)
plt.xlabel('Gender', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Gender vs Readmitted \n', size = 12)
plt.show()
#Age
df['age'].value_counts()
ax = sns.countplot(x='age',   data= df)
plt.xlabel('Age', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Age Distribution \n', size = 12)
plt.show()
```

```
# RelationShip Between and Age and Readmission
ax = sns.countplot(x="age", hue="readmitted_2", data=df)
plt.xlabel('Age', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Age vs Readmitted \n', size = 12)
plt.show()
#Weight
df['weight'].value_counts()
# Drop Weight column
df.drop(columns = ['weight'], inplace = True)
#admission_type_id
df['admission_type_id'].value_counts()
ax = sns.countplot(x='admission_type_id',   data= df)
plt.xlabel('Admission Type ID', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Admission Type Id Distribution \n', size = 12)
plt.show()
#discharge_disposition_id, #admission_source_id
len(df['discharge_disposition_id'].unique())
df['admission_source_id'].unique()
#Time in hospital
sns.set(rc={'figure.figsize':(18,8.2)})
ax = sns.countplot(x='time_in_hospital',   data= df)
plt.xlabel('Time In Hospital', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Time in Hospital Distribution \n', size = 12)
plt.show()
df['time_in_hospital'].mean()
# Relation of Stay in Hospital and Readmission
sns.set(rc={'figure.figsize':(18,8.2)})
ax = sns.countplot(x='time_in_hospital',  hue= 'readmitted_2',  data= df)
plt.xlabel('Time In Hospital', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Readmitted Count', size = 12)
plt.title('Time in Hospital vs Readmission \n', size = 12)
plt.show()
#Payer_code
df['payer_code'].value_counts()
# Drop the column -40256 Empty values
df.drop(columns = ['payer_code'], inplace = True)
#Medical Speciality
sns.set(rc={'figure.figsize':(18,8.2)})
ax = sns.countplot(x='medical_specialty',   data= df)
plt.xlabel('Medical Speciality', size = 12)
plt.xticks(rotation=90, size = 12)
```

```
plt.ylabel('Count', size = 12)
plt.title('Medical Speciality Distribution \n', size = 12)
plt.show()
# Drop the column -many missing values
df.drop(columns =['medical_specialty'], inplace = True)
# Number of Procedures/ Lab Procedures
# relation of Number of Procedures and Readmission
sns.set(rc={'figure.figsize':(18,8.2)})
ax = sns.countplot(x='num_procedures',  hue= 'readmitted_2',  data= df)
plt.xlabel('Number of Procedures', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Readmitted Count', size = 12)
plt.title('Number of Procedures vs Readmission \n', size = 14)
plt.show()
# relation of Number of Lab Procedures and Readmission
sns.displot(df, x="num_lab_procedures", hue= 'readmitted_2', kind="kde")
plt.title('Realtionship of Lab Procedures with Readmission \n\n', size = 13)
plt.show()
#Num_medications
sns.displot(df, x="num_medications", hue= 'readmitted_2', kind="kde")
plt.title('Number of Medications VS Readmission \n\n')
plt.show()
# Outpatient Visits
# Outpatient Visits less than 5
sns.displot(df.loc[df['number_outpatient']<5], x="number_outpatient", hue= 'readmitted_2',
kind ='kde')
plt.title('Relation of Outpatient Visits less than 5 w.r.t Readmission \n\n', size = 13)
plt.show()
# Outpatient Visits greater than 5
sns.displot(df.loc[df['number_outpatient']>=5], x="number_outpatient", hue= 'readmitted_2',
kind ='kde')
plt.title('Relation of Outpatient Visits >= 5 w.r.t Readmission \n\n', size = 13)
plt.show()
#Emergency visits
#Visits less than 5
sns.displot(df.loc[df['number_emergency']<5], x="number_emergency", hue= 'readmitted_2',
kind='kde')
plt.title('Relationship of Emergency Visits < 5 w.r.t Readmission \n\n', size = 13)
plt.show()
#Visits greater than 5
sns.displot(df.loc[df['number_emergency']>=5], x="number_emergency", hue=
'readmitted_2', kind='kde')
plt.title('Relationship of Emergency Visits >= 5 w.r.t Readmission \n\n', size = 13)
plt.show()
#Inpatient visits
#Visits less than 5
sns.displot(df.loc[df['number_inpatient']<5], x="number_inpatient", hue= 'readmitted_2',
kind='kde')
```

```
plt.title('Relationship Inpatient Visits < 5 w.r.t Readmission \n\n', size = 13)
plt.show()
#Visits greater than 5
sns.displot(df.loc[df['number_inpatient']>=5], x="number_inpatient", hue= 'readmitted_2',
kind='kde')
plt.title(' Inpatient Visits >= 5 w.r.t Readmission \n\n')
plt.show()
# Top 20 Diagnosis in the Readmitted = YES
ax = sns.barplot(x=df[df['readmitted_2'] == 'Yes']['diag_1'].value_counts().index[:20],
            y=df[df['readmitted_2'] == 'Yes']['diag_1'].value_counts()[:20])
plt.xlabel('Primary Diagnosis Codes', size = 12)
plt.ylabel('Count', size = 12)
plt.title('Top 20 Primary Diagnosis Codes in Readmission = YES \n', size = 12)
plt.show()
#Number of Diagnosis
sns.displot(df, x="number_diagnoses", hue= 'readmitted_2', kind='kde')
plt.title('Number of Diagnosis vs Readmission \n\n')
plt.show()
#Delete empty rows
df = df[~((df['diag_1'] == "?") | (df['diag_2'] == "?") | (df['diag_3'] == "?"))]
# max_glu_serum
sns.set(rc={'figure.figsize':(18,8.2)})
ax = sns.countplot(x='max_glu_serum',  hue= 'readmitted_2', data=
df[df['max_glu_serum']!='None'])
plt.xlabel('Max Glu Serum Value', size = 14)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 14)
plt.title('Max Glu Serum vs Readmission \n', size = 14)
plt.show() ## max_glu_serum >300 there is high chance of Readmission
# A1Cresult
sns.set(rc={'figure.figsize':(18,8.2)})
ax = sns.countplot(x='A1Cresult', hue = 'readmitted_2', data=df[df['A1Cresult']!='None'])
plt.xlabel('A1Cresult Values', size = 14)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 14)
plt.title('A1Cresult vs Readmission \n', size = 14)
plt.show()
#Diabetic med change
ax = sns.countplot(x='change',  hue= 'readmitted_2',  data= df)
plt.xlabel('Diabetes Med', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Readmitted Count', size = 12)
plt.title('Diabetes Med vs Readmission \n', size = 12)
plt.show()
#Diabet med prescribed
ax = sns.countplot(x='diabetesMed',  hue= 'readmitted_2',  data= df)
plt.xlabel('Diabetes Med', size = 12)
plt.xticks(rotation=90, size = 12)
```

```
plt.ylabel('Readmitted Count', size = 12)
plt.title('Diabetes Med vs Readmission \n', size = 12)
plt.show() ##Generally diabets has high percentage of Readmission

# Transform Categorical Features
# Make copy of data
df_ = df.copy()
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
le = LabelEncoder()
categorical_features =['race', 'gender', 'age',
      'admission_type_id', 'discharge_disposition_id', 'admission_source_id', 'diag_1', 'diag_2',
'diag_3', 'number_diagnoses',
      'max_glu_serum', 'A1Cresult', 'metformin', 'repaglinide', 'nateglinide',
      'chlorpropamide', 'glimepiride', 'glipizide', 'glyburide',
      'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'insulin',
      'glyburide-metformin', 'change', 'diabetesMed', 'acetohexamide', 'tolbutamide',
'troglitazone', 'tolazamide', 'examide', 'citoglipton',
      'glipizide-metformin', 'glimepiride-pioglitazone', 'metformin-rosiglitazone', 'metformin-
pioglitazone']
for i in categorical_features:
   df_[i] = le.fit_transform(df_[i])
label = le.fit(df_['readmitted_2'])
df_['readmitted_2_encoded'] = label.transform(df_['readmitted_2'])
df_.head()
## Drop Columns
df_ = df_.drop(columns= ['encounter_id', 'patient_nbr', 'readmitted','readmitted_2'])
#Correlation
f_[['time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications',
  'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses']].corr()

#Split
X = df_.drop(columns= ['readmitted_2_encoded'])
Y = df_['readmitted_2_encoded']

#Standardization
from sklearn import preprocessing
scaled_X = preprocessing.StandardScaler().fit_transform(X)

#Train/ Test
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
X_train, X_test, y_train, y_test = train_test_split(scaled_X, Y, test_size=0.25,
random_state=42)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
sm = SMOTE(random_state=42)
X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)
from sklearn.metrics import classification_report, confusion_matrix
```

```
#Classification
#Logistic Regression
from sklearn.linear_model import LogisticRegression
# Define Model
lr = LogisticRegression()
# Training
lr.fit(X_train, y_train)
# Prediction
lr_prediction = lr.predict(X_test)
print(classification_report(y_test, lr_prediction, target_names= ['Not Readmitted',
'Readmitted']))
clf_report = confusion_matrix(y_test, lr_prediction)
print(f"confusion_matrix : \n{clf_report}")
# plot feature importance
importance = lr.coef_[0]
plt.bar([x for x in range(len(importance))], importance)
columns_len = X.columns
tickvalues = range(0,len(columns_len))
plt.xticks(ticks = tickvalues, labels = columns_len, rotation = 'vertical')
plt.show()

#SMOTE - Logistic Regression
lr = LogisticRegression()
lr.fit(X_train_sm, y_train_sm)
lr_prediction = lr.predict(X_test)
print(classification_report(y_test, lr_prediction, target_names= ['Not Readmitted',
'Readmitted']))
clf_report = confusion_matrix(y_test, lr_prediction)
print(f"confusion_matrix (SMOTE): \n{clf_report}")
# plot feature importance
importance = lr.coef_[0]
plt.bar([x for x in range(len(importance))], importance)
columns_len = X.columns
tickvalues = range(0,len(columns_len))
plt.xticks(ticks = tickvalues, labels = columns_len, rotation = 'vertical')
plt.show()

#XGBoost
import xgboost
xgb =  xgboost.XGBClassifier()
xgb.fit(X_train, y_train)
xgb_prediction = xgb.predict(X_test)
print(classification_report(y_test, xgb_prediction, target_names= ['Not Readmitted',
'Readmitted']))
clf_report = confusion_matrix(y_test, xgb_prediction)
print(f"confusion_matrix : \n{clf_report}")
#Plot Feature Importance
(pd.Series(xgb.feature_importances_, index=X.columns)
```

```
  .nlargest(4)
  .plot(kind='barh'))

#SMOTE - XGBOOST
xgb =  xgboost.XGBClassifier()
xgb.fit(X_train_sm, y_train_sm)
xgb_prediction = xgb.predict(X_test)
print(classification_report(y_test, xgb_prediction, target_names= ['Not Readmitted',
'Readmitted']))
clf_report = confusion_matrix(y_test, xgb_prediction)
print(f"confusion_matrix (SMOTE): \n{clf_report}")
#Plot Feature Importance
(pd.Series(xgb.feature_importances_, index=X.columns)
  .nlargest(4)
  .plot(kind='barh'))

#Random Forest
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators = 450, max_depth=9, random_state=43)
rf.fit(X_train, y_train)
rf_prediction =  rf.predict(X_test)
print(classification_report(y_test, rf_prediction, target_names= ['Not Readmitted',
'Readmitted']))
clf_report = confusion_matrix(y_test, rf_prediction)
print(f"confusion_matrix : \n{clf_report}")
#Plot Feature Importance
(pd.Series(rf.feature_importances_, index=X.columns)
  .nlargest(4)
  .plot(kind='barh'))

#SMOTE – Random Forest
rf = RandomForestClassifier(n_estimators = 450, max_depth=9, random_state=43)
rf.fit(X_train_sm, y_train_sm)
rf_prediction =  rf.predict(X_test)
print(classification_report(y_test, rf_prediction, target_names= ['Not Readmitted',
'Readmitted']))
clf_report = confusion_matrix(y_test, rf_prediction)
print(f"confusion_matrix (SMOTE): \n{clf_report}")
#Plot Feature Importance
(pd.Series(rf.feature_importances_, index=X.columns)
  .nlargest(4)
  .plot(kind='barh'))
```

ΠΣ ΠΗΓΑΙΟΣ ΚΩΔΙΚΑΣ ΣΕ PYTHON ΓΙΑ ΤΗΝ 3η ΕΦΑΡΜΟΓΗ

```
#LIBRARIES
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from importlib import reload
%matplotlib inline

#READ FILE
csv_url = '2022_Q1_OR_Utilization.csv'
df = pd.read_csv(csv_url)

#ΠΕΡΙΓΡΑΦΙΚΑ
df.shape
df.info()
# Check for Duplicates in Encounter Id
len(df['Encounter ID'].unique()) #No Duplicates
#Service
#Check for unique categories in Service and number of data for each one
df['Service'].value_counts()
ax = sns.barplot(x=df['Service'].value_counts().index,   y=df['Service'].value_counts())
plt.xlabel('Service', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Distribution of Service per Encounter \n', size = 12)
plt.show()
#OR Suite
print(df['OR Suite'].value_counts())
ax = sns.barplot(x=df['OR Suite'].value_counts().index,   y=df['OR Suite'].value_counts())
plt.xlabel('OR Suite', size = 12)
plt.xticks(rotation=90, size = 12)
plt.ylabel('Count', size = 12)
plt.title('Distribution of OR Suite per Encounter \n', size = 12)
plt.show()

#Date/Time Format
df['OR Schedule'] = pd.to_datetime(df['OR Schedule']).dt.strftime('%Y-%m-%d-
%H:%M:%S')
df['OR Schedule'] = pd.to_datetime(df['OR Schedule'])
df['OR Schedule']

df['Wheels In'] = pd.to_datetime(df['Wheels In']).dt.strftime('%Y-%m-%d-%H:%M:%S')
df['Wheels In'] = pd.to_datetime(df['Wheels In'])
```

```
df['Wheels In']

df['Start Time'] = pd.to_datetime(df['Start Time']).dt.strftime('%Y-%m-%d-%H:%M:%S')
df['Start Time'] = pd.to_datetime(df['Start Time'])
df['Start Time']

df['End Time'] = pd.to_datetime(df['End Time']).dt.strftime('%Y-%m-%d-%H:%M:%S')
df['End Time'] = pd.to_datetime(df['End Time'])
df['End Time']

df['Wheels Out'] = pd.to_datetime(df['Wheels Out']).dt.strftime('%Y-%m-%d-%H:%M:%S')
df['Wheels Out'] = pd.to_datetime(df['Wheels Out'])
df['Wheels Out']

# OR Schedule per Year/ Month / Day
df['OR Schedule_M'] = df['OR Schedule'].dt.month
df['OR Schedule_D'] = df['OR Schedule'].dt.day
df["OR Schedule_Time"] = df['OR Schedule'].dt.hour + df['OR Schedule'].dt.minute/60
# print(df['OR Schedule_Y'].value_counts())
print(df['OR Schedule_M'].value_counts())
# print(df['OR Schedule_D'].value_counts())

ax = sns.barplot(x=df['OR Schedule_M'].value_counts().index,   y=df['OR
Schedule_M'].value_counts())
plt.xlabel('OR Schedule_M', size = 12)
plt.ylabel('Count', size = 12)
plt.title('Number of OR occupancy per month \n', size = 12)
plt.show()

# Create Actual_Time: by calculating the Difference between WheelsIn-WheelsOut Time
df['Actual_Time'] = ((df['Wheels Out'] - df['Wheels In']).dt.seconds/60).astype("int")

#Mean of BookedTime in Min
print(df['Booked Time (min)'].mean())
print(df.groupby('Service')['Booked Time (min)'].mean())

#Mean of Actual Time
print(df['Actual_Time'].mean())
print(df.groupby('Service')['Actual_Time'].mean())

#υπολογίζω το μ.ο. ανά CPT για να μπορώ να συγκρίνω Actual Time / Mean Time κάθε
επέμβασης με το Booked Time
df.groupby("CPT Code")["Actual_Time"].mean()
df.merge(df.groupby("CPT Code")["Actual_Time"].mean(), on="CPT Code", how="left")
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Κωνσταντάκης, Ε. (2020). Συμβολή των Big Data στη Διοίκηση Μονάδων Υγείας.

Φαρατζιάν, Α. (2007). Στρατηγική και Ποιότητα Μονάδων και Υπηρεσιών Υγείας. Διπλωματική εργασία, Τμήμα οργάνωσης και διοίκησης επιχειρήσεων, Πανεπιστήμιο Πειραιά.

Τζαχρήστας, Ν. (2015). Νέα μοντέλα διοίκησης στον τομέα της δημόσιας υγείας: προβλήματα και προοπτικές.

Δεβελέγκα, Α., & Παπαδοπούλου, Μ. (2015). Μεταρρυθμίσεις των συστημάτων υγείας στην Ελλάδα: ανασκόπηση και κριτική προσέγγιση.

Τήλλυρος, Χ. (2019). Συγκριτική αξιολόγηση αλγορίθμων μηχανικής μάθησης σε δεδομένα ασθενών με διαβήτη (Doctoral dissertation, University of Piraeus (Greece)).

Κόκκινος, Ι. (2011). Παράλληλοι αλγόριθμοι εξόρυξης γνώσης από βάσεις δεδομένων με τεχνητά νευρωνικά δίκτυα και μηχανές διανυσμάτων υποστήριξης.

Ντάλλα, Μ. (2009). Εφαρμογή αλγορίθμων επαγωγικού λογικού προγραμματισμού στη σχεσιακή εξόρυξη δεδομένων (Doctoral dissertation).

Ταρακτσίδης, Γ. (2008). Εξόρυξη γνώσης από βάση δεδομένων ηλεκτρονικών δημοπρασιών απο τον δικτυακό τόπο eBay.

Κεχαγιά-Παρδάλη, Ε (2006), Αλγόριθμοι Εξόρυξης Χωρικών Δεδομένων Εφαρμογή σε Αλγορίθμους Συσταδοποίησης

Μπαλάσκα, Δ., & Μπιτσώρη, Ζ. (2015). Ποιότητα των παρεχόμενων υπηρεσιών υγείας και ο βαθμός ικανοποίησης των ασθενών. Περιεγχειρητική Νοσηλευτική, 4(3), 106-120.

ΜΠΙΤΣΩΡΗ, Ζ., & ΜΠΑΛΑΣΚΑ, Δ. (2016). Υπηρεσίες υγείας και η χρηματοδότησή τους. Περιεγχειρητική Νοσηλευτική, 5, 113-124.

Χαλκίδη, Μ., & Βαζιργιάννης, Μ. (2005). Εξόρυξη γνώσης από βάσεις δεδομένων και τον Παγκόσμιο Ιστό. Εκδόσεις Τυπωθήτω-Γιώργος Δάρδανος.

Μπερσίμης Σ., Μπάρτζης Γ., Παπαδάκης Γ., Σαχλάς Α. (2021). Εφαρμοσμένη Στατιστική και Στατιστική Μηχανική Μάθηση, Με χρήση των IBM SPSS Statistics, R, Python. . Εκδόσεις Τζιόλα

Κρατικός Προϋπολογισμός 2023 (Νοέμβριος 2022), Άρθρο 2

Νόμος 4368/2016, Μέτρα για την επιτάχυνση του κυβερνητικού έργου και άλλες διατάξεις. Εφημερίδα της Κυβέρνησης (ΦΕΚ 21/Α/21-2-2016)

Νόμος 2889/2001, Βελτίωση και εκσυγχρονισμός του Εθνικού Συστήματος υγείας Ε.Σ.Υ. και άλλες διατάξεις. Εφημερίδα της Κυβέρνησης (ΦΕΚ 37/Α/2-3-2001)

Νόμος 1397/1983, Εθνικό Σύστημα Υγείας. Εφημερίδα της Κυβέρνησης (ΦΕΚ 143/Α/7-10-1983)

Νόμος 965/1937, Περί οργανώσεως των Δημοσίων νοσηλευτικών και Υγειονομικών Ιδρυμάτων. Εφημερίδα της Κυβέρνησης (ΦΕΚ 476/Α/24-11-1937)

Ευθυμιάδου, Δ. (2022, October 7). Υπουργείο Υγείας: Τρία έκτακτα μέτρα για να αδειάσουν οι λίστες χειρουργείων στα νοσοκομεία - Όλο το σχέδιο. ΕΘΝΟΣ. https://www.ethnos.gr/health/article/227323/ypoyrgeioygeiastriaektaktametragianaadeiasoynoilistesxeiroyrgeionstanosokomeiaolotosxedio

Ξένη

Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks, 3, 58-73.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Raghupathi, W., & Raghupathi, V. (2013). An overview of health analytics. J Health Med Informat, 4(132), 2.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. Big data, 1(2), 85-99.

Glowacka, K. J., Henry, R. M., & May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. Journal of the Operational Research Society, 60, 1056-1068.

Chong, L. R., Tsai, K. T., Lee, L. L., Foo, S. G., & Chang, P. C. (2020). Artificial intelligence predictive analytics in the management of outpatient MRI appointment no-shows. American Journal of Roentgenology, 215(5), 1155-1162.

Michailidis, P., Dimitriadou, A., Papadimitriou, T., & Gogas, P. (2022, May). Forecasting Hospital Readmissions with Machine Learning. In Healthcare (Vol. 10, No. 6, p. 981). MDPI.

Li, W., Lipsky, M. S., Hon, E. S., Su, W., Su, S., He, Y., ... & Hung, M. (2021). Predicting all-cause 90-day hospital readmission for dental patients using machine learning methods. BDJ open, 7(1), 1.

Rozario, D. (2020). Can machine learning optimize the efficiency of the operating room in the era of COVID-19?. Canadian Journal of Surgery, 63(6), E527.

Abbou, B., Tal, O., Frenkel, G., Rubin, R., & Rappoport, N. (2022). Optimizing Operation Room Utilization—A Prediction Model. Big Data and Cognitive Computing, 6(3), 76.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Cheng, F., Yang, C., Zhou, C., Lan, L., Zhu, H., & Li, Y. (2020). Simultaneous determination of metal ions in zinc sulfate solution using UV–Vis spectrometry and SPSE-XGBoost method. Sensors, 20(17), 4936.

Donabedian, A. (1996). Quality improvement through monitoring health care.

Donabedian, A. (1980). Explorations in quality assessment and monitoring: the definition of quality and approaches to its assessment.

Mesa, M. J., Asencio, J. M., Ruiz, F. R., & González, M. P. (2017). Análisis del coste económico del absentismo de pacientes en consultas externas. Revista de Calidad Asistencial, 32(4), 194-199.

WHO Working Group. (1989). The principles of quality assurance. International Journal for Quality in Health Care, 1(2-3), 79-95.

World Health Organization. (1993). Continuous quality development: A proposed national policy. In Continuous quality development: a proposed national policy.

World Health Organization. (2000). The world health report 2000: health systems: improving performance. World Health Organization.

Seldon. (2023). Supervised vs Unsupervised Learning Explained. Seldon. https://www.seldon.io/supervised-vs-unsupervised-learning-explained

R, A. (2022). Why Do We Use Decision Trees in Machine Learning? www.turing.com. https://www.turing.com/kb/importance-of-decision-trees-in-machine-learning

Devopedia. 2022. "Data Science." Version 7, February 15. Accessed 2023-05-02. https://devopedia.org/data-science