ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Προβλεπτικά μοντέλα χρήσης ιατρικής εφαρμογής»

Βούζης Ελευθέριος

Επιβλέπων Καθηγητής:
Μαγκλογιάννης Ηλίας, Καθηγητής

ΠΕΙΡΑΙΑΣ

ΦΕΒΡΟΥΑΡΙΟΣ 2022

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Προβλεπτικά μοντέλα χρήσης ιατρικής εφαρμογής

Βούζης Ελευθέριος
Α.Μ.: ΜΕ2002

ΠΕΡΙΛΗΨΗ

Η ραγδαία ανάπτυξη της τεχνολογίας με την παράλληλη αύξηση στον όγκο των ιατρικών δεδομένων στα πληροφοριακά συστήματα των νοσοκομείων, σηματοδότησε την εποχή των Μεγάλων Δεδομένων στον τομέα της υγείας. Η ανάλυση ιατρικών δεδομένων αποτελεί σημαντική ευκαιρία παγκοσμίως για τα εθνικά συστήματα υγείας ώστε να μειωθεί το κόστος και ταυτόχρονα να βελτιωθεί η υγειονομική περίθαλψη. Η αξιοποίηση των τεχνολογιών αυτών γίνεται στο πλαίσιο παρακολούθησης θεμάτων υγείας, καταμέτρησης στόχων υγείας και φυσικής κατάστασης, καθώς και για καταγραφή ιατρικών δεδομένων. Σε ένα τέτοιο πλαίσιο, ο έγκαιρος εντοπισμός χρηστών με κίνδυνο χαμηλότερων ποσοστών συμμόρφωσης και μοτίβων χρήσης μιας εφαρμογής παρακολούθησης υγείας, που υποδηλώνουν κίνδυνο εγκατάλειψης είναι μια ανεκτίμητη ευκαιρία για την εφαρμογή προσαρμοσμένων στρατηγικών παρέμβασης που στοχεύουν στην ανάκαμψη και την αποφυγή αποδέσμευσης των χρηστών. Η παρούσα διπλωματική εργασία ασχολείται με την φετινή έκδοση του παγκόσμιου διαγωνισμού IFMBE Science Challenge 2022, στόχος του οποίου είναι ο εντοπισμός προτύπων πρόωρης εγκατάλειψης σε χρήστες μιας εφαρμογής για κινητές συσκευές παρεμβάσεων με τίτλο Active and Healthy Ageing (AHA). Οι συμμετέχοντες του διαγωνισμού έχουν πρόσβαση σε ένα σύνολο δεδομένων με περισσότερους από 150 χρήστες στη Μαδρίτη που έχουν δοκιμάσει τον αντίκτυπο μιας ψηφιακής εφαρμογής AHA για τη βελτίωση της ποιότητας ζωής τους για τουλάχιστον 6 μήνες στο δίκτυο Moving Active & Healthy Aging. Η πρόκληση του διαγωνισμού, είναι δεδομένου ενός παραθύρου με n=12 διαδοχικών προγραμματισμένων στιγμών απόκτησης δεδομένων, να προβλεφθεί η συμμόρφωση του χρήστη κατά τις προσεχείς 3 προγραμματισμένες στιγμές απόκτησης δεδομένων. Στο στάδιο της πειραματικής διαδικασίας υλοποιήθηκαν και παρουσιάζονται πολλές διαφορετικές προσεγγίσεις για την πρόβλεψη της εγκατάλειψης. Αρχικά στο σύνολο των δεδομένων και στην συνέχεια με διαφορετικό πλήθος χαρακτηριστικών, ώστε να επιλεγεί το καλύτερο. Συγκεκριμένα προτείνετε μια μεθοδολογία με την χρήση του κανόνα καθαρισμού δεδομένων Neighborhood Cleaning Rule (NCR) και ένας συγκεκριμένος αλγόριθμος ταξινόμησης με την μέθοδο μάθησης Stacked Generalization για την πρόβλεψη της πρόωρης εγκατάλειψης των χρηστών της εφαρμογής παρακολούθησης υγείας. Τα αποτελέσματα δείχνουν ότι ο προτεινόμενος αλγόριθμός ήταν ικανός να προβλέψει την πρόωρη εγκατάλειψη των χρηστών από την εφαρμογή με ποσοστό ακρίβειας 97,6%, ενώ παράλληλα, βάσει της μετρικής που επιλέχθηκε από τον διαγωνισμό σε ποσοστό 93,4%, κάτι και το οποίο φυσικά τον κάνει αρκετά αξιόπιστο, ώστε να χρησιμοποιείται ως ένα σύστημα έγκαιρης προειδοποίησης. Στο τέλος της παρούσας εργασίας παρουσιάζονται αναλυτικά τα συμπεράσματα που προέκυψαν από την έρευνα.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Κατηγοριοποίηση Δεδομένων Ιατρικής Εφαρμογής

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Μηχανική Μάθηση, Σύστημα Έγκαιρης Προειδοποίησης, Επιλογή Χαρακτηριστικών, Κατηγοριοποίηση, Stacked Generalization

ABSTRACT

The rapid development of technology with the parallel increase in the volume of medical data in the information systems of hospitals, marked the era of Big Data in the field of health. The analysis of medical data is a significant opportunity worldwide for national health systems to reduce costs and at the same time improve healthcare. The utilization of these technologies is done in the context of monitoring health issues, counting health goals and fitness, as well as for recording medical data. In such a context, early detection of users at risk of lower compliance rates and patterns of use of a health monitoring application suggesting a risk of abandonment is an invaluable opportunity to implement tailored intervention strategies aimed at recovering and avoiding abandonment thoughts. This dissertation deals with this year's edition of the global competition (IFMBE Science Challenge 2022), which aims to identify patterns of early dropout in users of an application for mobile intervention called Active and Healthy Aging (AHA). Contestants have access to a database of more than 150 users in Madrid who have experienced the impact of a digital AHA application to improve their quality of life for at least 6 months on the MAHA (Moving Active & Healthy Aging) network. The challenge of the competition is given a window of n = 12 consecutive scheduled moments of data acquisition, to predict the user's compliance during the next 3 scheduled data acquisition. At the experimental stage many different approaches to early dropout prediction were implemented and presented. First in the initial data set and then with a different set of features to choose the best one. Specifically, the current thesis proposes a methodology using the Neighborhood Cleaning Rule (NCR) and a specific classification algorithm with the Stacked Generalization learning method to predict the early abandonment of users of the health monitoring application. The results shown that the proposed algorithm was able to predict the early dropout of users from the application with an accuracy of 97.6%, while at the same time, based on the challenge metric at a rate of 93,4%, makes it reliable enough to be used as an early warning system. At the end of this paper are presented in detail the conclusions that emerged from the research.

SUBJECT AREA: Categorization of Medical Application Data

KEYWORDS: Machine Learning, Early Warning System, Feature Selection, Classification, Stacked Generalization

Στην μνήμη του αγαπημένου μου Παππού.

ΕΥΧΑΡΙΣΤΙΕΣ

Η ολοκλήρωση του παρόντος Προγράμματος Μεταπτυχιακών Σπουδών αποτελεί το αποτέλεσμα ομαδικής εργασίας. Οφείλω λοιπόν ένα θερμό ευχαριστώ στον αδερφικό μου φίλο Σαραντόπουλο Γιώργο για την αμέριστη βοήθειά του. Επιπλέον, αισθάνομαι την ανάγκη να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κύριο Μαγκλογιάννη Ηλία για τη στήριξη και τις συμβουλές του κατά την εκπόνηση της διπλωματικής εργασίας. Ιδιαιτέρως, όμως, ευχαριστώ από καρδιάς την οικογένεια μου για τη συνεχή υποστήριξη και κατανόηση τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε στο πλαίσιο του Προγράμματος Μεταπτυχιακών Σπουδών «Μεγάλα Δεδομένα & Αναλυτική» του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιά. Το θέμα της παρούσας διπλωματικής εργασίας είναι η εφαρμογή προβλεπτικών μοντέλων σε πραγματικά ανώνυμα δεδομένα προερχόμενα από ιατρική εφαρμογή παρακολούθησης ηλικιωμένων ατόμων.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1 Ανάλυση Δεδομένων

Κατά το διάστημα των τελευταίων ετών η τεχνολογική εξέλιξη έχει προκαλέσει αύξηση του παραγόμενου όγκου δεδομένων. Η πολυπλοκότητα των δεδομένων συνεχώς αυξάνεται και μαζί της επιτείνεται και η ανάγκη αποθήκευσης και διαχείρισής τους από σύγχρονες μεθόδους μοντελοποίησης και ανάλυσης. Ειδικότερα, στον τομέα της ιατρικής παράγεται τεράστιος όγκος ετερογενών και συνεχών μεταβαλλόμενων δεδομένων (Big Data), τα οποία αντλούνται από διαφορετικές πηγές, όπως βιοαισθητήρες, κάμερες, smartwatches, κοινωνικά δίκτυα και άλλα. Η ποικιλομορφία των δεδομένων είναι εντυπωσιακή, με τα δεδομένα να αναπαριστούν δομημένους πίνακες, βίντεο, χρονοσειρές, γραφήματα, εικόνες, κείμενα, ήχο και άλλα.

Η ανάλυση Δεδομένων (Data Analysis) ορίζεται ως η διαδικασία εξέτασης διάφορων συνόλων δεδομένων με σκοπό την εξαγωγή συμπερασμάτων αναφορικά με τις πληροφορίες που περιέχουν, με τη συνδρομή λογισμικού και εξειδικευμένων συστημάτων. Επιπλέον παρουσιάζεται ως η διαδικασία ανακάλυψης αξιοποιήσιμης πληροφορίας προερχόμενη από μεγάλες βάσεις δεδομένων με την χρήση αλγορίθμων για την ανακάλυψη κρυμμένων μοτίβων σε δεδομένα. Η ανάλυση δεδομένών μπορεί να τμηματοποιηθεί σε προγνωστική (predictive), διερευνητική (exploratory) και κανονιστική (prescriptive) ανάλυση δεδομένων. Οι προγνωστικές αναλύσεις υποδεικνύουν τι θα συμβεί στο μέλλον. Οι μέθοδοι που χρησιμοποιούνται κυρίως είναι η ανάλυση παλινδρόμησης (regression analysis), η μηχανική μάθηση (machine learning) και τα τεχνητά νευρωνικά δίκτυα (artificial neural networks). Οι αναλύσεις δεδομένων μπορούν επίσης να διαχωριστούν σε ανάλυση ποσοτικών δεδομένων και ανάλυση ποιοτικών δεδομένων. Στην πρώτη κατηγορία, περιλαμβάνεται η ανάλυση αριθμητικών δεδομένων με ποσοτικά προσδιορίσιμες μεταβλητές με δυνατότητα σύγκρισης και δημιουργίας στατιστικών. Στην δεύτερη κατηγορία, η ποιοτική ανάλυση είναι πιο ερμηνευτική, εστιάζει στην κατανόηση του περιεχομένου μη-αριθμητικών δεδομένων όπως είναι οι εικόνες, το κείμενο, το βίντεο και ο ήχος [1].

Οι τεχνικές ανάλυσης δεδομένων εφαρμόζονται ευρέως σε οργανισμούς για την λήψη ενημερωμένων και καλύτερων επιχειρηματικών αποφάσεων, από επιστήμονες και ερευνητές για την επαλήθευση και επικαιροποίηση επιστημονικών μοντέλων και υποθέσεων. Σε προηγμένες εφαρμογές ανάλυσης δεδομένων, μεγάλο μέρος της απαιτούμενης εργασίας λαμβάνει χώρα εκ των προτέρων, στο στάδιο συλλογής, ενσωμάτωσης και προετοιμασίας των επιθυμητών δεδομένων. Στη συνέχεια αναπτύσσονται και δοκιμάζονται τα ήδη δημιουργημένα αναλυτικά για να εξασφαλιστεί ότι παράγουν ακριβή αποτελέσματα. Αναλυτικότερα, η διαδικασία της ανάλυσης ξεκινά με το στάδιο συλλογής δεδομένων, στο οποίο γίνεται ο εντοπισμός των πληροφοριών που χρειάζονται για μια συγκεκριμένη εφαρμογή ανάλυσης δεδομένων. Με την ολοκλήρωση της συλλογής του επιθυμητού όγκου δεδομένων, ως επόμενο στάδιο ορίζεται ο εντοπισμός και η διόρθωση των προβλημάτων στην ποιότητα των δεδομένων που θα μπορούσαν να επηρεάσουν την ακρίβεια των εφαρμογών της ανάλυσης, για παράδειγμα πραγματοποιείται έλεγχος και εξάλειψη των σφαλμάτων και των διπλών εγγραφών.

Στην συνέχεια γίνεται εφαρμογή διαφόρων αλγορίθμων ανάλογα με την φύση του προβλήματος που καλείται η εφαρμογή να επιλύσει. Άλλωστε, ένα από τα κεντρικά προβλήματα της εποχής της πληροφορίας είναι η τεράστια ποσότητα δεδομένων που είναι διαθέσιμη. Ο όγκος των δεδομένων αυτών αυξάνεται συνεχώς και το χάσμα μεταξύ της παραγωγής και της ικανότητας κατανόησης διευρύνεται. Για την επιτυχή γεφύρωση αυτού του κενού γνώσης, δύναται να εφαρμοστούν μια ποικιλία τεχνικών γνωστών ως εξόρυξη δεδομένων [2].

Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί σε ένα μεγάλο πλήθος εφαρμογών. Σε αυτό ανήκουν λοιπόν και τα συστήματα υγειονομικής περίθαλψης, όπου οι αλγόριθμοι ταξινόμησης διανέμουν τα δεδομένα μεταξύ διαφορετικών κατηγοριών που ορίζονται σε ένα σύνολο δεδομένων. Κατά το στάδιο της εκπαίδευσης, οι αλγόριθμοι ταξινόμησης μαθαίνουν πώς να χρησιμοποιούν μοτίβο διανομής μέσω του σετ εκπαίδευσης (training set) που δίνεται από τον χρήστη. Στη συνέχεια, οι αλγόριθμοι αρχίζουν να ταξινομούν τα δεδομένα δοκιμής (test set) με σκοπό την σωστή τους ταξινόμηση.

Οι αποφάσεις που πρέπει να πάρει ένας γιατρός βασίζονται στην αξιολόγηση των εξετάσεων του ασθενή που περιγράφουν την τωρινή του κατάσταση, με την παραπομπή τους συνήθως σε αποφάσεις που είχε λάβει στο παρελθόν για ίδια περιστατικά. Γίνεται εύκολα αντιληπτό, λοιπόν, ότι το πλήθος των παραγόντων που πρέπει να αναλυθούν για τη διάγνωση των ασθενειών καθιστά δύσκολη την δουλειά της ιατρικής κοινότητας. Σε αυτό το κρίσιμο βήμα, μπορεί να χρειαστεί ένα ακριβές εργαλείο που να απαριθμεί τις προηγούμενες αποφάσεις του σχετικά με τον ασθενή που παρουσιάζει τα ίδια ή παρόμοια χαρακτηριστικά.

Η ανάγκη και η συχνότητα χρήσης συστημάτων ταξινόμησης στην ιατρική διάγνωση παρουσιάζει συνεχής αύξηση. Η εφαρμογή μοντέλων μηχανικής μάθησης σε ιατρικά σύνολα δεδομένων, άλλωστε, αποτελεί στην σημερινή εποχή ένα πολύ ενδιαφέρον θέμα για τους ερευνητές, καθώς υπάρχουν πολλά ζητήματα υγείας που χρειάζονται διερεύνηση. Με την χρήση των μοντέλων μηχανικής μάθησης οι γιατροί αναζητούν έναν σύμμαχο στον αγώνα τους για την έγκαιρη και ακριβέστερη ανάλυση κινδύνου της νόσου. Δεν υπάρχει αμφιβολία λοιπόν ότι η αξιολόγηση των δεδομένων που λαμβάνονται από τον ασθενή και το ιστορικό αποφάσεων των εμπειρογνωμόνων είναι οι πιο σημαντικοί παράγοντες στη διαδικασία της ιατρικής διάγνωσης. Όμως, αποτελεί κοινή παραδοχή της επιστημονικής κοινότητας ότι τα συστήματα κατηγοριοποίησης και οι διαφορετικές τεχνικές τεχνητής νοημοσύνης για ταξινόμηση βοηθούν επίσης τους ειδικούς σε έναν μεγάλο βαθμό στο δύσκολο έργο που συντελούν [3].

Εικόνα 1. Ανάλυση Δεδομένων

1.2 Αντικείμενο της Μεταπτυχιακής Διατριβής

Η έγκαιρη πρόβλεψη της εγκατάλειψης μιας ιατρικής εφαρμογής παρακολούθησης υγείας είναι ένα σοβαρό πρόβλημα στην σημερινή εποχή και παράλληλα δύσκολο στην επίλυσή του. Από την μία υπάρχουν πολλοί παράγοντες που μπορούν να επηρεάσουν τη διατήρηση των χρηστών μιας εφαρμογής, από την άλλη πλευρά η παραδοσιακή προσέγγιση ταξινόμησης με την δυνατότητα λύσης αυτού του προβλήματος κανονικά πρέπει να εφαρμοστεί στο τέλος μιας περιόδου παρακολούθησης των χρηστών για να συγκεντρωθούν οι μέγιστες πληροφορίες με σκοπό την υψηλότερη ακρίβεια πρόβλεψης. Υπό αυτό το πλαίσιο, για την ιατρική κοινότητα είναι υψίστης σημασίας να βρεθούν περισσότερο αποτελεσματικές πρακτικές και να μειωθούν οι μη ολοκληρωμένες ιατρικές παρακολουθήσεις μέσω συστημάτων καταγραφής και παρακολούθησης της υγείας ασθενών. Στην προσπάθεια αυτή, η συγκεκριμένη διπλωματική εργασία προτείνει μια μεθοδολογία και έναν συγκεκριμένο αλγόριθμο ταξινόμησης προς ανακάλυψη μοντέλων πρόβλεψης της εγκατάλειψης των χρηστών μιας εφαρμογής παρακολούθησης υγείας το συντομότερο δυνατό. Παρουσιάζει πλήθος πειραμάτων που υλοποιήθηκαν για την πρόβλεψη της εγκατάλειψης και με διαφορετικό πλήθος χαρακτηριστικών για τους χρήστες, ώστε να επιλεγεί το καλύτερο. Αναφορικά δε, τα αποτελέσματα δείχνουν ότι ο προτεινόμενος αλγόριθμός ήταν ικανός να προβλέψει την πρόωρη εγκατάλειψη των χρηστών από την εφαρμογή με εξαιρετική ακρίβεια και φυσικά μπορεί να χρησιμοποιείται ως ένα αξιόπιστο σύστημα έγκαιρης προειδοποίησης.

1.3 Δομή της Διατριβής

Στο πρώτο κεφάλαιο γίνεται συνοπτική εισαγωγή στην έννοια της ανάλυσης δεδομένων και παράλληλα, παρουσιάζεται το αντικείμενο και η δομή της ΜΔΕ. Στο δεύτερο κεφάλαιο γίνεται εισαγωγή στην έννοια και τη χρησιμότητα της μηχανικής μάθησης. Αναλύονται συνοπτικά οι όροι των μεγάλων δεδομένων και των μη σχεσιακών βάσεων. Ακολούθως εξετάζονται οι διάφοροι αλγόριθμοι εποπτευόμενης μάθησης. Στο τρίτο κεφάλαιο γίνεται παρουσίαση του συνόλου δεδομένων και του διαγωνισμού στον οποίον χρησιμοποιούνται. Ακόλουθα αναλύεται η μεθοδολογία που ακολουθήθηκε για να αντιμετωπιστεί το πρόβλημα της πρόβλεψης πρόωρης αποχώρησης του χρήστη από την εφαρμογή. Στο τέταρτο κεφάλαιο, παρουσιάζονται διαφορετικές προσεγγίσεις και αποτελέσματα αυτών, οι οποίες διενεργήθηκαν κατά το στάδιο της πειραματικής διαδικασίας. Στο πέμπτο κεφάλαιο, αναλύονται τα συμπεράσματα της πειραματικής διαδικασίας, όπως αυτή παρουσιάστηκε σε προηγούμενο κεφάλαιο. Τέλος, εντοπίζεται η καλύτερη μέθοδος ταξινόμησης των δεδομένων με βάση την ειδική μετρική που χρησιμοποιήθηκε στον διαγωνισμό, κατηγοριοποιώντας τα δεδομένα σε συνεπής ή μη συνεπής χρήστες της εφαρμογής. Όσο εκτεταμένο και αν είναι ένα κεφάλαιο, δεν μπορεί να αποτελεί συνολική επισκόπηση των τεχνικών επεξεργασίας δεδομένων ή αξιολόγησης αποτελεσμάτων εποπτευόμενων αλγορίθμων. Όμως, τα στοιχεία που αναφέρονται καλύπτουν τα σημαντικότερα θεωρητικά ζητήματα και καθοδηγούν τον ερευνητή σε ενδιαφέροντα συμπεράσματα.

ΚΕΦΑΛΑΙΟ 2. ΤΕΧΝΙΚΟ ΥΠΟΒΑΘΡΟ & ΕΠΙΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

2.1 Μεγάλα Δεδομένα & Μη Σχεσιακές Βάσεις Δεδομένων

Ο όρος Big Data αντιπροσωπεύει τις τεράστιες συλλογές δεδομένων, οι οποίες δεν μπορούν να επεξεργαστούν με παραδοσιακά μέσα διαχείρισης λογισμικού και ανάλυσης δεδομένων. Ο κύριος λόγος ραγδαίας ανάπτυξης των Big Data τα τελευταία χρόνια οφείλεται στην μακροπρόθεσμη επιχειρηματική αξία που παρέχουν. Όπως αναφέρεται και στο [4], για την διαχείριση, επεξεργασία και ανάλυση μεγάλων δεδομένων απαιτείται τεχνητή νοημοσύνη (Artificial intelligence - AI). Η καθημερινή χρήση του διαδικτύου έχει αυξήσει την παραγωγή δεδομένων, με αποτέλεσμα την δημιουργία τεράστιων όγκων δεδομένων μη διαχειρίσιμων από τις διαδεδομένες σχεσιακές βάσεις δεδομένων. Στην προσπάθεια αποθήκευσης και επεξεργασίας μεγάλης κλίμακας δεδομένων, δημιουργήθηκαν και οι μη σχεσιακές βάσεις.

Οι βάσεις δεδομένων NoSQL είναι ένα είδος μη σχεσιακών (non-relational) συστημάτων βάσεων δεδομένων, μια συλλογή ζεύγους κλειδιού-τιμής, εγγράφων, βάσεων δεδομένων γραφημάτων, που δεν έχουν τυπικούς ορισμούς σχήματος που πρέπει να τηρούνται. Συνήθως δεν απαιτείται ο καθορισμός του σχήματος δεδομένων (schema-less), αποτελούν δηλαδή ένα μη δομημένο τρόπο αποθήκευσης. Υποστηρίζουν την προσωρινή αποθήκευση συνδυάζοντας τη με το μικρότερο δυνατό κόστος για τη συντήρηση των διακομιστών τους. Επιπλέον, σημαντική προς αναφορά είναι η μη ύπαρξη περιορισμού ως προς το πλήθος και το scalability των δεδομένων. Άλλωστε, η αρχιτεκτονική κλιμάκωσης των συστημάτων NoSQL παρέχει μια σαφή πορεία προς την επεκτασιμότητα όταν αυξάνεται ο όγκος των δεδομένων, σε αντίθεση με τις βάσεις SQL, όπου η επίτευξη του ίδιου τύπου επεκτασιμότητας μπορεί να καταστεί δαπανηρή ή ανέφικτη.

Τα NoSQL συστήματα ακολουθούν τις ιδιότητες BASE. Αναλυτικότερα, διασφαλίζουν τη διαθεσιμότητα του συστήματος βάσης δεδομένων, ακόμα και σε περιπτώσεις που ένας ή περισσότεροι κόμβοι δεν είναι διαθέσιμοι (Basically Available). Η κατάσταση του συστήματος είναι πιθανόν να μεταβληθεί με την πάροδο του χρόνου, χωρίς να είναι απαραίτητη η αλληλεπίδρασή της με κάποιο χρήστη (Soft state) και επιπλέον τα δεδομένα που ενημερώνονται σε ένα κόμβο του συστήματος, μπορεί να μη διαδίδονται άμεσα στους υπόλοιπους κόμβους, με αποτέλεσμα την προσωρινή ασυνέπεια των δεδομένων του. Όμως, με το πέρασμα του χρόνου το σύστημα εγγυάται την επαναφορά της συνέπειας στο σύνολο των δεδομένων του (Eventual consistency).

Τα Data Models των NoSQL συστημάτων κατατάσσονται σε πέντε διαφορετικές κατηγορίες ανάλογα με τις ιδιότητες των δεδομένων που θέλουμε να χρησιμοποιήσουμε. Ονομαστικά οι κατηγορίες είναι Key-Value, Document oriented, Graph, Column oriented και Object oriented. Συγκεκριμένα, η MongoDB, που θα χρησιμοποιηθεί στην παρούσα εργασία, ανήκει στις Document Oriented Databases και εξυπηρετεί δεδομένα μεγάλου όγκου τα οποία μπορούν να αναπαρασταθούν με τη μορφή εγγράφων. Σε αυτού του είδους τις NoSQL βάσεις δεδομένων τα στοιχεία που αποθηκεύονται είναι σε μορφή είτε unstructured είτε semi-structured εγγράφων.

Οι βάσεις δεδομένων NoSQL δημιουργήθηκαν στο διαδίκτυο στις εποχές υπολογιστικού νέφους που κατέστησαν δυνατή την ευκολότερη εφαρμογή μιας αρχιτεκτονικής κλίμακας. Σε μια αρχιτεκτονική κλιμάκωσης, η επεκτασιμότητα επιτυγχάνεται με τη διάδοση της αποθήκευσης δεδομένων και την επεξεργασία των δεδομένων σε ένα μεγάλο σύμπλεγμα υπολογιστών. Συγκεκριμένα για την αύξηση της χωρητικότητας, περισσότεροι υπολογιστές προστίθενται στο σύμπλεγμα. Αυτή η αρχιτεκτονική κλιμάκωσης είναι ιδιαίτερα ανώδυνη για εφαρμογή σε περιβάλλοντα υπολογιστικού νέφους, όπου νέοι υπολογιστές μπορούν εύκολα να προστεθούν σε ένα σύμπλεγμα. Τα σχήματα των περισσότερων βάσεων δεδομένων NoSQL είναι ευέλικτα και υπό τον έλεγχο των προγραμματιστών, καθιστούν ευκολότερη την προσαρμογή της βάσης δεδομένων σε νέες μορφές δεδομένων.

2.1.1 MongoDB

Η MongoDB είναι μια βάση δεδομένων προσανατολισμένη στα έγγραφα που χρησιμοποιεί για την αποθήκευση των δεδομένων συλλογές τύπου JSON εγγράφων (BSON). Αυτά τα έγγραφα υποστηρίζουν ενσωματωμένα πεδία, παρέχοντας τη δυνατότητα αποθήκευσης σχετικών δεδομένων. Η MongoDB είναι επίσης μια βάση δεδομένων ακαθόριστου σχήματος (shema-less), οπότε δεν χρειάζεται να καθορίσουμε τον αριθμό ή τον τύπο στηλών πριν από την εισαγωγή των δεδομένων μας. Ενδεικτικά κάποιες από τις δυνατότητες που παρέχει η MongoDB:

- Ευέλικτα σχήματα εγγράφων
- Φιλικό προς το σχεδιασμό
- Εύκολη οριζόντια κλιμάκωση

Το μοντέλο εγγράφων της MongoDB επιτρέπει σχεδόν οποιοδήποτε είδος δομής δεδομένων να μοντελοποιείται και να χειρίζεται εύκολα. Η μορφή δεδομένων BSON της MongoDB, εμπνευσμένη από το JSON, μας επιτρέπει να έχουμε αντικείμενα σε μια συλλογή με διαφορετικά σύνολα πεδίων. Η απόφαση της MongoDB να αποθηκεύει και να αντιπροσωπεύει δεδομένα σε μορφή εγγράφου σημαίνει ότι μπορεί να επιτευχθεί η πρόσβαση σε αυτήν από οποιαδήποτε γλώσσα, σε δομές δεδομένων που είναι εγγενείς σε αυτήν τη γλώσσα (π.χ. λεξικά σε Python, συσχετισμένοι πίνακες σε JavaScript, Χάρτες σε Java κ.λπ.). Δεν απαιτείται χρόνος διακοπής λειτουργίας για την αλλαγή σχήματος των δεδομένων και είναι εφικτή η εγγραφή νέων δεδομένων στη MongoDB ανά πάσα στιγμή, χωρίς να διακοπούν οι λειτουργίες της. Κατά αυτόν τον τρόπο, παρέχεται η δυνατότητα για το χειρισμό μεγάλου όγκου δεδομένων σε υψηλή ταχύτητα με αρχιτεκτονική κλιμάκωσης.

Τα query της MongoDB υποστηρίζουν τις λειτουργίες CRUD (Create, Read, Update, Delete) [5]. Με την λειτουργία Create προστίθεται ένα νέο document σε ένα collection, στην περίπτωση δε που το collection δεν υπάρχει κατά την εισαγωγή του πρώτου document δημιουργείται αυτόματα. Μέσω της λειτουργίας Read γίνεται εφικτή η ανάγνωση documents ή πεδίων των documents προερχόμενα από ένα collection. Επιπλέον με τη χρήση διαφόρων φίλτρων δίνεται η δυνατότητα διαβάσματος εγγραφών που ικανοποιούν διάφορες προϋποθέσεις. Ακόμη η λειτουργία Update χρησιμοποιείται για την τροποποίηση ενός document που ήδη υπάρχει σε ένα collection. Τέλος, με τη delete ο χρήστης μπορεί να διαγράψει ένα document από ένα collection.

2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση πρωτοεμφανίστηκε νωρίς την δεκαετία του 1980 και αποτελεί σήμερα έναν από τους βασικούς τομείς της Τεχνητής Νοημοσύνης. Η Μηχανική Μάθηση συνδυάζει στοιχεία από την επιστήμη των υπολογιστών, τα μαθηματικά, τη στατιστική, τη βιολογία και τη νευροεπιστήμη με σκοπό την ανίχνευση μοτίβων στα δεδομένα. Αναλυτικότερα ορίζεται ως το επιστημονικό πεδίο μελέτης και σχεδίασης υπολογιστικών προγραμμάτων, με την ικανότητα μάθησης αποσκοπώντας στην βελτίωση της απόδοσής τους μέσω της προηγούμενης εμπειρίας τους. Ουσιαστικά αναπτύχθηκε κατά την διάρκεια της μελέτης αναγνώρισης προτύπων στον τομέα της τεχνητής νοημοσύνης. Εννοιολογικά λοιπόν η μηχανική μάθηση ταυτίζεται με την απόκτηση γνώσης και εμπειρίας σε κάποιο τομέα.

Ένα υπολογιστικό σύστημα μπορεί αυτοματοποιημένα να μαθαίνει και να αποθηκεύει γνώση για διάφορους τομείς με στόχο τη βελτιστοποίησή του μέσω της επανάληψης σε μελλοντικές εφαρμογές, όπως ακριβώς αντίστοιχα ισχύει και για κάθε νοήμων οντότητα. Σημαντική, όμως, είναι η διαφοροποίηση στην διαδικασία απόκτηση αυτής της γνώσης και εμπειρίας, καθώς δεν προκύπτει από την αλληλεπίδραση του συστήματος με το περιβάλλον, αλλά από ένα υποσύνολο δεδομένων μιας βάσης δεδομένων που αποτελεί το σύνολο των δεδομένων εκπαίδευσης. Ο στόχος, άλλωστε, είναι να εκπαιδεύσει ένα μηχάνημα να μαθαίνει από δεδομένα και να βρίσκει κρυφές δομές και μοτίβα σε αυτά, να κάνει προβλέψεις για αόρατα δεδομένα και να παίρνει αποφάσεις υπό αβεβαιότητα. Η χρήση δε της μηχανικής μάθησης κρίνεται αναγκαία σε υπολογιστικές εφαρμογές, όπου ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος και συμβάλλει σημαντικά στο πεδίο της ανάλυσης δεδομένων, καθώς αποτελεί μία μέθοδο επινόησης πολύπλοκων μοντέλων και αλγορίθμων με σκοπό την πρόβλεψη μιας τιμής [6].

Η εφαρμογή του συστήματος αλληλοεπιδρώντας με την βάση δεδομένων για την ορθή εκτέλεση της επιθυμητής εργασίας διαιρείται σε δύο φάσεις. Η πρώτη ονομάζεται φάση της εκπαίδευσης, κατά την οποία, χρησιμοποιούνται δειγματοληπτικά ή ιστορικά δεδομένα για τη δημιουργία αντιπροσωπευτικού μοντέλου που να τα αναπαριστά. Ακολούθως κατά την φάση του ελέγχου, εφαρμόζεται το αντιπροσωπευτικό μοντέλο στα υπόλοιπα δεδομένα. Αναφορικά με την χρησιμότητα και την χρησιμοποίησή της στην καθημερινότητα, η μηχανική μάθηση δύναται να εφαρμοστεί σε πολλούς τομείς όπως η αναγνώριση ομιλίας, η υπολογιστική όραση και γενικά σε εργασίες ταξινόμησης και παλινδρόμησης, οι οποίες ασχολούνται με την πρόβλεψη της τιμής ενός πεδίου με βάση τις τιμές των άλλων πεδίων (χαρακτηριστικά). Η διαφορά μεταξύ αυτών των δύο είναι ότι η έξοδος της πρώτης εργασίας είναι κατηγορική, ενώ για τη δεύτερη είναι συνεχής. Οι αλγόριθμοι μηχανικής μάθησης έχουν την δυνατότητα να αναπαριστούν την γνώση ποικιλόμορφα με εξισώσεις, δέντρα αποφάσεων, κανόνες, αποστάσεις, χωρίσματα ακόμα και με πιθανοτικά ή γραφικά μοντέλα. Άλλωστε, στο χώρο της μηχανικής μάθησης για την επίλυση κάθε προβλήματος, υπάρχει ο κατάλληλος τρόπος μάθησης και για κάθε τρόπο μάθησης υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να χρησιμοποιηθεί για το συγκεκριμένο πρόβλημα. Ορισμένοι αλγόριθμοι δέχονται σαν είσοδο μόνο παρατηρήσεις ενώ άλλοι δίνουν βάση λίγο περισσότερο στην προϋπάρχουσα γνώση [7]. Σε αυτή τη διατριβή χρησιμοποιήθηκαν αλγόριθμοι μηχανικής μάθησης για προβλήματα ταξινόμησης.

2.3 Είδη Μηχανικής Μάθησης

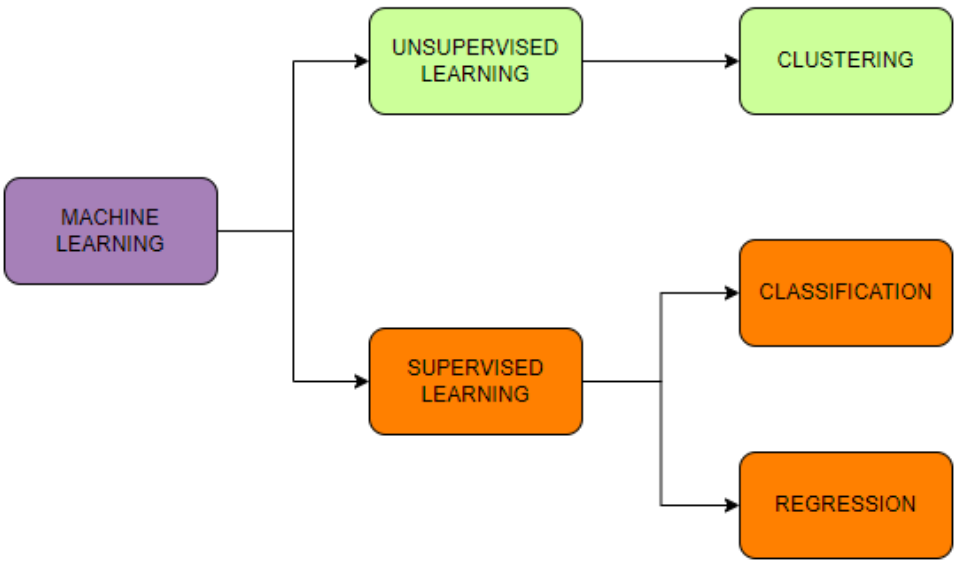
Με την πάροδο του χρόνου και την ραγδαία τεχνολογική εξέλιξη, αναπτύχθηκαν πολλές διαφορετικές τεχνικές μηχανικής μάθησης. Ανάλογα με την ικανότητα εφαρμογή τους σε προβλήματα διαφορετικής φύσεως κατατάσσονται στις παρακάτω δύο κατηγορίες [8].

2.3.1 Μάθηση χωρίς επίβλεψη (Unsupervised learning)

Μάθηση από παρατήρηση, όπου το σύστημα βασισμένο μονάχα στις δικές του ιδιότητες, ανακαλύπτει διαφορετικές κατηγορίες αντικειμένων. Στο συγκεκριμένο είδος μάθησης υπάρχουν τα δεδομένα εκπαίδευσης χωρίς όμως την γνώση της σωστής απάντησης, οπότε ο αλγόριθμος κατασκευάζει ένα μοντέλο για ένα σύνολο παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Ο στόχος της μη επιβλεπόμενης μάθησης είναι η ομαδοποίηση παρόμοιων στοιχείων σε ένα σύνολο δεδομένων.

2.3.2 Μάθηση με επίβλεψη (Supervised Learning)

Μάθηση με παραδείγματα, όπου το σύστημα τροφοδοτείται με διάφορα παραδείγματα αντικειμένων μιας κατηγορίας και στην συνέχεια καλείται να ανακαλύψει τις κοινές ιδιότητες αυτών των αντικειμένων. Στόχος της επιβλεπόμενης μάθησης δεν είναι μόνο η σωστή κατηγοριοποίηση παρόμοιων αντικειμένων, αλλά και η χαρτογράφηση κάθε αντικειμένου στην ομάδα που ανήκει. Στο συγκεκριμένο είδος μάθησης υπάρχουν τα δεδομένα εκπαίδευσης παράλληλα με την γνώση της σωστής απάντησης. Συγκεκριμένα το υπολογιστικό μοντέλο εφαρμόζει επιτυχώς κάθε καταχώρηση στο σύνολο εκπαίδευσης σε συνδυασμό με τις σωστές απαντήσεις.

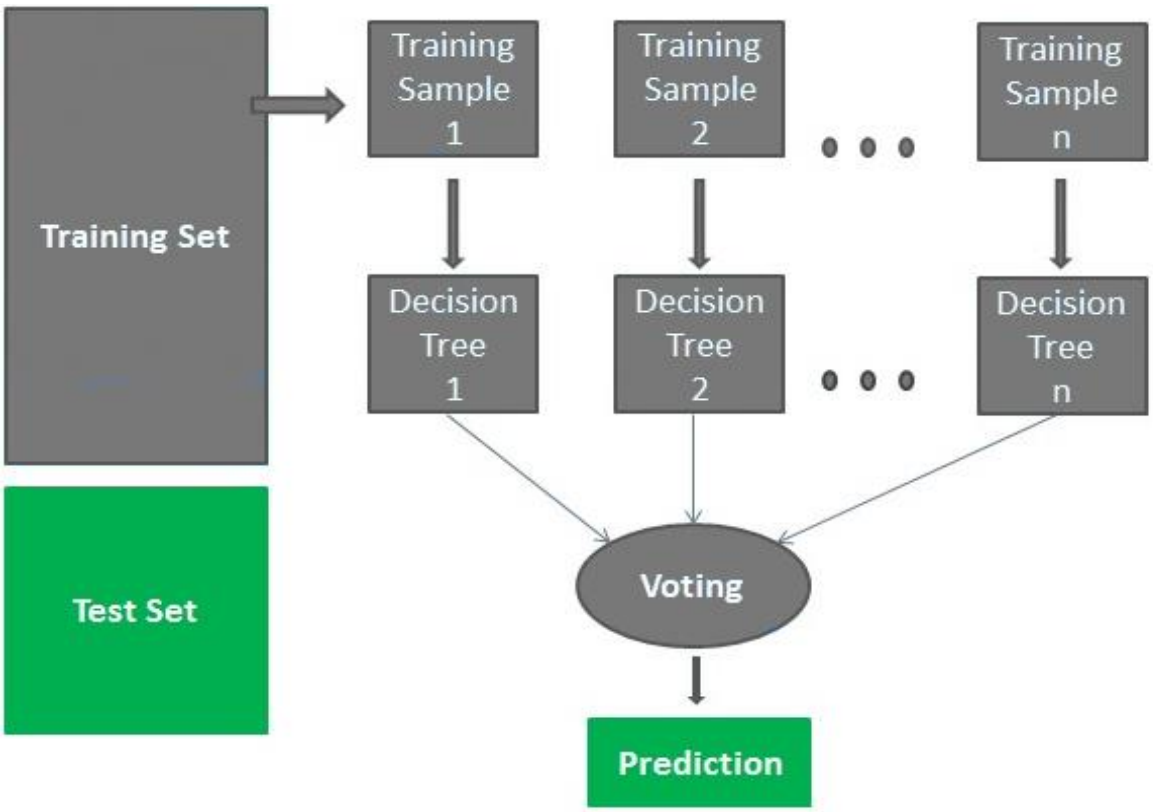


Εικόνα 2. Είδη Μηχανικής Μάθησης

2.4 Αλγόριθμοι Κατηγοριοποίησης

2.4.1 Τυχαία Δάση (Random Forest)

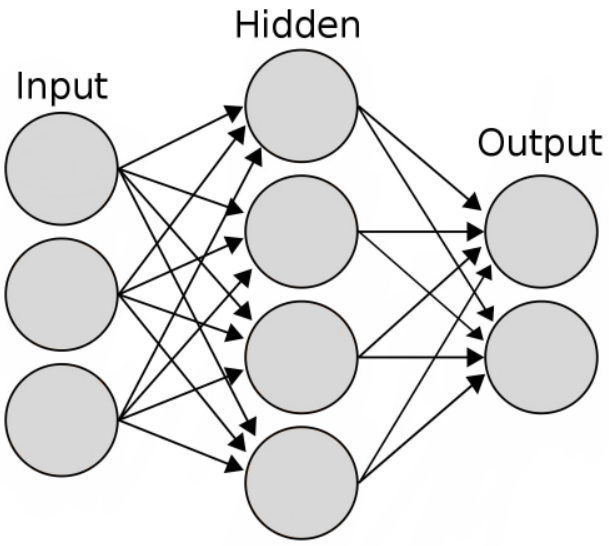
Ο τυχαίος ταξινομητής δασών ορίζεται ως συνδυασμός ταξινομητών δέντρων απόφασης. Ένα δένδρο απόφασης αποτελείται από κόμβους που αντιστοιχούν σε κάποιο χαρακτηριστικό του συνόλου εκπαίδευσης με εξερχόμενες ακμές που αντιστοιχούν σε μια συνθήκη διάσπασης των δεδομένων. Το βασικό ποσοτικό μέτρο που χρησιμοποιείται για επιλογή των διαχωριστών είναι το κέρδος πληροφορίας (information gain), το οποίο βασίζεται στην εντροπία πληροφορίας. Ο Random Forest αποτελεί, ίσως, τον δημοφιλέστερο αλγόριθμο στην κατηγορία του και αναπτύχθηκε από τους Leo Breiman και Adele Cutler το 2001 [9]. Γνωστός κυρίως για την ταχύτητα αλλά και την ακρίβειά του. Σύμφωνα με τους δημιουργούς του [10], προσφέρει καλύτερη ακρίβεια μεταξύ των υπαρχόντων αλγορίθμων με ταχύτητα ακόμα και σε μεγάλα σύνολα δεδομένων εκπαίδευσης. Επιπλέον, δεν χρειάζεται την χρήση διαφορετικού συνόλου δεδομένων για τον έλεγχο ακριβείας. Συγκεκριμένα, δεν κρίνεται απαραίτητη η διασταυρούμενη επικύρωση (cross validation), καθώς η εκτίμηση του λάθους γενίκευσης γίνεται από τον ίδιο τον αλγόριθμο κατά την εκτέλεσή του, ενώ δεν παρουσιάζει συχνά φαινόμενα υπερεκπαίδευσης (overfitting).



Εικόνα 3. Δομή Αλγόριθμου Τυχαίων Δασών

2.4.2 Δίκτυα Πολλαπλών Στρωμάτων (Multilayer Perceptron)

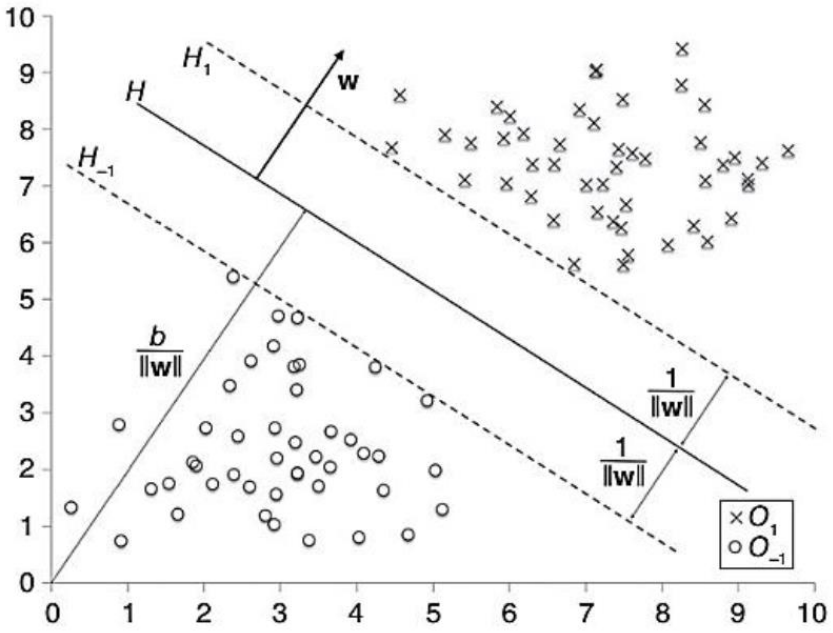
Τα τεχνητά νευρωνικά δίκτυα είναι εμπνευσμένα από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Ο ανθρώπινος εγκέφαλος μπορεί να θεωρηθεί ως ένας πολύ περίπλοκος και μη γραμμικός υπολογιστής που λαμβάνει, επεξεργάζεται πληροφορίες και εκτελεί πολλές περίπλοκες εργασίες. Ένα νευρωνικό δίκτυο προσπαθεί να μιμηθεί τον ανθρώπινο εγκέφαλο εντοπίζοντας μοτίβα μέσω της επεξεργασίας των δεδομένων και της διαδικασίας πρόβλεψης αόρατων τιμών. Η συγκεκριμένη κατηγορία τροφοδοτούμενων προς τα εμπρός δικτύων εκπαιδεύεται από έναν αλγόριθμο οπίσθιας διάδοσης και περιλαμβάνει περισσότερα από ένα κρυφά στρώματα υπολογιστικών νευρώνων τα οποία βρίσκονται μεταξύ του επιπέδου των κόμβων εισόδου και της εξόδου. Η μαθησιακή διαδικασία ενός νευρωνικού δικτύου είναι όπως αναφέρθηκε προηγουμένως παρόμοια με την μαθησιακή λειτουργία του ανθρώπινου εγκεφάλου [11]. Σημαντικό προς αναφορά είναι το γεγονός ότι ο αλγόριθμος δύναται να πραγματοποιήσει πιο σύνθετους υπολογισμούς με την προσθήκη επιπλέον στρωμάτων νευρώνων για την εξαγωγή αποτελεσμάτων υψηλότερης τάξης και πολυπλοκότητας. Η συγκεκριμένη ικανότητα είναι ιδιαίτερα χρήσιμη στην περίπτωση ύπαρξης αρκετών κόμβων στο αρχικό επίπεδο εισόδου. Όπως αναφέρει και ο Bishop [12], αυτά τα δίκτυα καλούνται Multilayer Perceptron με την λειτουργία τους να ενεργοποιείται με την τροφοδότηση του πρώτου κρυφού στρώματος με το διάνυσμα εισόδου. Στην συνέχεια πραγματοποιούνται οι υπολογισμοί στους νευρώνες του πρώτου κρυφού στρώματος με αποτέλεσμα την δημιουργία σημάτων εξόδου, τα οποία αποτελούν διανύσματα εισόδου του δεύτερου κρυφού στρώματος. Ουσιαστικά, κάθε στρώμα νευρώνων λειτουργεί ως είσοδος για το επόμενο στρώμα νευρώνων. Για να ενεργοποιηθεί κάθε νευρώνας έχει μια λειτουργία ενεργοποίησης. Αυτή η συνάρτηση δρα σε έναν σταθμισμένο συνδυασμό των εισόδων από κάθε νευρώνα στον οποίο είναι συνδεδεμένος ο εν λόγω νευρώνας. Η έξοδος γίνεται η είσοδος στους νευρώνες στα επόμενα στρώματα. Κάθε βάρος εκφράζει πληροφορίες που χρησιμοποιούνται από το δίκτυο για την επίλυση ενός προβλήματος, ενώ κάθε νευρώνας έχει μια συνάρτηση ενεργοποίησης που είναι το λαμβανόμενο σήμα από τον προηγούμενο νευρώνα. Η παραπάνω διαδικασία συνεχίζεται μέχρι την στιγμή που το σήμα φτάνει στο στρώμα εξόδου και επιτυγχάνεται η απόκριση του δικτύου.



Εικόνα 4. Δομή Νευρωνικού Δικτύου

2.4.3 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οικογένεια αλγορίθμων εποπτευόμενης μηχανικής μάθησης, οι οποίοι χρησιμοποιούνται κυρίως σε προβλήματα ταξινόμησης, αλλά και παλινδρόμησης. Κάθε στοιχείο δεδομένων σχεδιάζεται ως σημείο σε n-διάστατο χώρο ανάλογα με το πλήθος των χαρακτηριστικών, με την αξία κάθε χαρακτηριστικού να είναι η τιμή της συγκεκριμένης συντεταγμένης. Ακολούθως εκτελείται η ταξινόμηση με την εύρεση του υπερεπιπέδου που διαφοροποιεί πολύ καλά τα δεδομένα. Σημαντική επίσης ορίζεται η τεχνική kernel trick που χρησιμοποιούν οι αλγόριθμοι SVM, η οποία δέχεται σαν είσοδο έναν χώρο μικρής διάστασης και τον αναγάγει σε μεγαλύτερης. Αναλυτικότερα, υλοποιούν μερικούς περίπλοκους μετασχηματισμούς δεδομένων και στη συνέχεια με βάση τις ετικέτες ή τις εξόδους που έχουν οριστεί, ανακαλύπτουν τη διαδικασία διαχωρισμού των δεδομένων [13]. Τις περισσότερες φορές σε προβλήματα μεγάλων δεδομένων η εκπαίδευση των αλγορίθμων SVM γίνεται με αργό ρυθμό, αλλά χειρίζονται πολύ καλά μεγάλο πλήθος χαρακτηριστικών και παρουσιάζουν υψηλή απόδοση κατά την κατηγοριοποίηση αντικειμένων. Η έρευνα σε έναν αλγόριθμο SVM πραγματοποιείται για δύο λόγους, ένας είναι ο διαχωρισμός με το μεγαλύτερο ελάχιστο περιθώριο και ταυτόχρονα ο δεύτερος λόγος είναι μία γραμμή που θα διαχωρίζει σωστά όσο το δυνατόν περισσότερες περιπτώσεις. Η παράμετρος c καθορίζει το δεύτερο διαχωρισμό, η οποία εκφράζει την εξισορρόπηση σημαντικότητας μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης των λανθασμένων ταξινομήσεων. Η τιμή της c είναι επιλογή του χρήστη και γίνεται μετά την αξιολόγηση διάφορων δοκιμών και αφού το σφάλμα είναι ανάλογο με την c, χαμηλότερες τιμές του συντελεστή οδηγούν σε μικρότερη ποινή για τις λάθος ταξινομήσεις [14].



Εικόνα 5. Παράδειγμα κατηγοριοποίησης με την χρήση του ταξινομητή SVM [15]

2.4.4 Λογιστική Παλινδρόμηση (Logistic Regression)

Μοντέλο ταξινόμησης τιμών μιας μεταβλητής εξόδου με βάση τη θεωρία των πιθανοτήτων. Στο συγκεκριμένο μοντέλο, όπου η μεταβλητή εξόδου συνήθως έχει δυαδικό χαρακτήρα, στοχεύετε η πρόβλεψη του αποτελέσματός της από πλήθος μεταβλητών που μπορεί να είναι είτε ονομαστικές είτε ποσοτικές. Άλλωστε, στη λογιστική παλινδρόμηση η εκτίμηση παραμέτρων βασίζεται στη μέθοδο του λόγου πιθανοφάνειας. Ουσιαστικά γίνεται επιλογή των πιο πιθανοφανών τιμών των παραμέτρων, ώστε να οδηγήσουν σε κάποια αποτελέσματα. Η λογιστική παλινδρόμηση αποτελεί εναλλακτική της γραμμικής διακριτής ανάλυσης για την ταξινόμηση των στοιχείων της εξαρτημένης μεταβλητής και παρουσιάζει ευρεία απήχηση σε πληθώρα ερευνητικών κλάδων, όπως για παράδειγμα στην ιατρική [16].

2.4.5 Μπεϊζιανός Αλγόριθμος (Naïve Bayes)

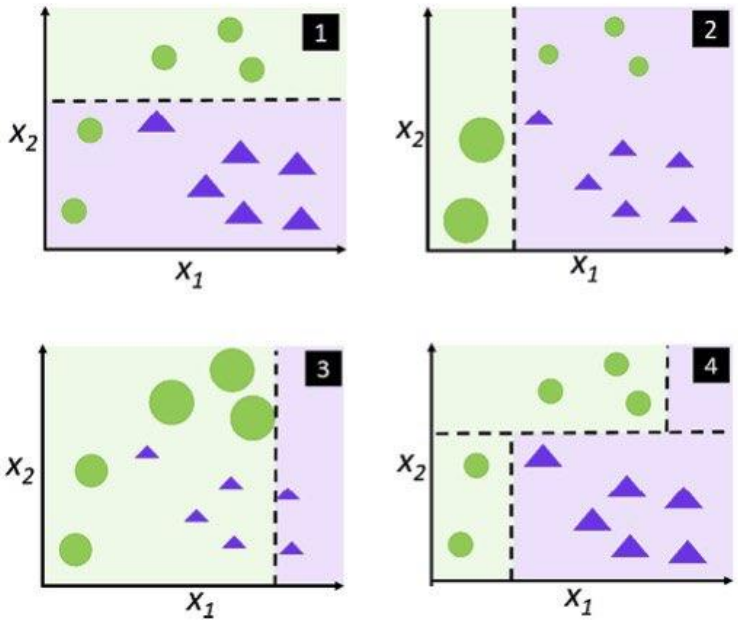
Οι ταξινομητές Naive Bayes έχουν αποδειχθεί ισχυρά εργαλεία για την επίλυση προβλημάτων ταξινόμησης με εφαρμογή σε διάφορα πεδία. Ουσιαστικά ο μπεϊζιανός αλγόριθμος είναι ένας πιθανοτικός ταξινομητής που χρησιμοποιεί το θεώρημα του Bayes, λειτουργώντας με υπό όρους πιθανότητες. Χρησιμοποιώντας την πιθανότητα υπό όρους, είναι σε θέση να υπολογίσει την πιθανότητα γεγονότος χρησιμοποιώντας τις προηγούμενες γνώσεις του. Ο τύπος για τον υπολογισμό της υπό όρους πιθανότητας δίνεται από την σχέση: $P(H/E) = P(E/H) * P(H)/P(E)$. Ως $P(H)$ ορίζεται η πιθανότητα της υπόθεσης να είναι αληθής (priori probability), ως $P(E)$ η πιθανότητα των στοιχείων (ανεξάρτητα από την υπόθεση) και ως $P(E|H)$ η δεσμευμένη πιθανότητα των αποδεικτικών στοιχείων, δεδομένου ότι η υπόθεση είναι αληθής. Προβλέπει επομένως πιθανότητες συμμετοχής για κάθε κλάση, όπως η πιθανότητα ότι μια δεδομένη εγγραφή ή ένα σημείο δεδομένων ανήκει σε μία συγκεκριμένη κλάση [17]. Ο ταξινομητής υποθέτει ότι όλα τα χαρακτηριστικά είναι ασυσχέτιστα, οπότε δεν μπορεί να μάθει τη μεταξύ τους σχέση, με αποτέλεσμα να θεωρείται αφελής και να είναι υπολογιστικά ελαφρύς. Επιπλέον η τεχνική κατηγοριοποίησης του μπεϊζιανού αλγορίθμου εμφανίζει ιδιαίτερα καλή εφαρμογή σε περιπτώσεις υψηλών εισροών. Παρά την απλή λειτουργία του, ο συγκεκριμένος αλγόριθμος έχει παρατηρηθεί ότι δύναται να υπερβαίνει σε απόδοση και αποτελεσματικότητα τις πιο εξελιγμένες μεθόδους ταξινόμησης. Χαρακτηρίζεται από ταχύτητα και παρέχει διαφορετικούς τύπους αλγορίθμων όπως GaussianNB, MultinomialNB και BernoulliNB, ενώ πολύ εύκολα μπορεί να ολοκληρώσει την διαδικασία εκπαίδευσής του ακόμα και με ένα μικρό σύνολο δεδομένων [18].

2.4.6 Κ Κοντινότεροι Γείτονες (KNN)

Ο αλγόριθμος Κ πλησιέστερων γειτόνων (KNN), αποτελεί μια χρήσιμη τεχνική εξόρυξης γνώσης και προσέγγισης του προβλήματος κατηγοριοποίησης [19]. Ανήκει στους αλγόριθμους εποπτευόμενης μάθησης και είναι ευρέως γνωστός για την χρήση μέτρων απόστασης, όπως η ομοιότητα συνημίτονου ή ο δείκτης συσχέτισης κατά Pearson. Αν και η περιγραφή του μοιάζει με την Παλινδρόμηση, εκείνη μπορεί να χρησιμοποιηθεί μόνο για αριθμητικές εξόδους. Επιπλέον, είναι εφικτός ο αριθμός των αποτελεσμάτων (γείτονες) που θέλουμε να βρούμε [20]. Η κεντρική ιδέα είναι ότι η τιμή της συνάρτησης-στόχου για μία νέα είσοδο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των k πιο «κοντινών» στιγμιότυπων εκπαίδευσης, τα οποία αποτελούν τους «γείτονές» του.

2.4.7 Αλγόριθμος Προσαρμοσμένης Ενίσχυσης (AdaBoost)

Ο στατιστικός αλγόριθμος Προσαρμοσμένης Ενίσχυσης (AdaBoost) χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης και δύναται να συνδυαστεί με διάφορους αλγόριθμους για να βελτιωθεί η απόδοση του μοντέλου. Αναλυτικά στον αλγόριθμο προσαρμοσμένης ενίσχυσης, τα αποτελέσματα του συνόλου των αδύναμων αλγόριθμων μάθησης (weak learners) συνδυάζονται ως ένα σταθμισμένο άθροισμα το οποίο αντιπροσωπεύει την τελική έξοδο του ταξινομητή. Κατά την παραπάνω διαδικασία, κάθε νέος weak learner που διαδέχεται έναν άλλον, τροποποιείται υπέρ των δειγμάτων που ταξινομήθηκαν λανθασμένα από τους learners που έχουν προηγηθεί. Η απόδοση αυτών των learners μπορεί να μην είναι η βέλτιστη ή να είναι πολύ κακή, είναι σημαντικό, όμως, να είναι καλύτερη από την τυχαία ταξινόμηση για να μετατραπεί το μοντέλο σε έναν δυνατό learner. Με αυτόν τον τρόπο, ο AdaBoost χαρακτηρίζεται ως προσαρμοστικός. Η δημοφιλέστερη υλοποίηση του AdaBoost είναι η χρήση ως weak learners δέντρων απόφασης (Decision Trees) και θεωρείται από τους κορυφαίους «έξυπνους» αλγόριθμους ταξινόμησης. Σε αυτή την υλοποίηση έχουμε την χρήση Decision Tree Stumps. Το κάθε Stump αποτελείται από τη ρίζα και δύο φύλλα και αξιολογεί μόνο ένα χαρακτηριστικό των δεδομένων εισόδου, κάτι το οποίο συνεπάγεται ότι λαμβάνοντας υπόψιν μόνο ένα χαρακτηριστικό, κάθε stump είναι ένας πολύ αδύναμος learner. Ο συνδυασμός, όμως, μεγάλου αριθμού αυτών μπορεί να οδηγήσει στην κατασκευή ενός αξιόπιστου και μεγάλης ακρίβειας συνδυαστικού μοντέλου ταξινόμησης (ensemble model) [21]. Όπως παρουσιάζεται και στην εικόνα 6, στους κακώς ταξινομημένους βαθμούς δίνονται περισσότερα βάρη και τα βάρη των σωστά ταξινομημένων πόντων μειώνονται. Ένας νέος ταξινομητής εκπαιδεύεται με ένα νέο σύνολο δεδομένων εκπαίδευσης που έχει περισσότερα βάρη στα λανθασμένα ταξινομημένα σημεία και μικρότερα βάρη σε σωστά ταξινομημένα σημεία. Η διαδικασία επαναλαμβάνεται στο σύνολο των αδύναμων ταξινομητών που θα χρησιμοποιηθούν, με τον προσαρμοστικό ταξινομητή ενίσχυσης συνόλου να κατασκευάζεται από τον συνδυασμό των τριών ταξινομητών που εκπαιδεύτηκαν χρησιμοποιώντας διαφορετικά σύνολα δεδομένων εκπαίδευσης αποτέλεσμα της προσαρμοστικής επαναδειγματοληψίας.

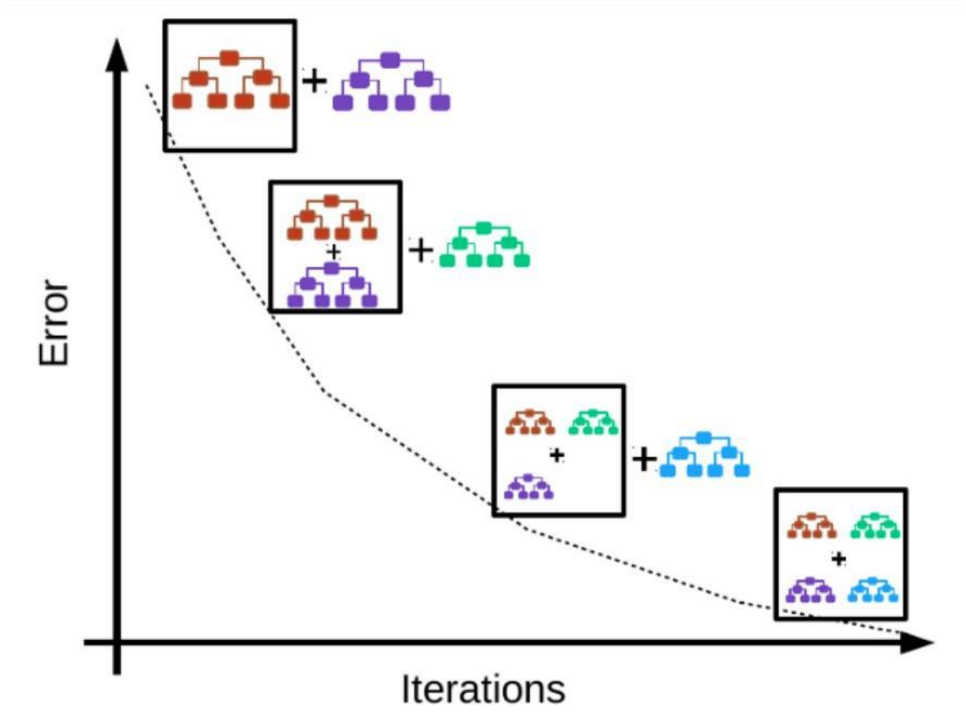


Εικόνα 6. Αλγόριθμος Προσαρμοσμένης Ενίσχυσης

2.4.8 Αλγόριθμος Ακραίας Ενίσχυσης Κλίσης (XGBoost)

Ο XGBoost ταξινομητής ανήκει στην οικογένεια αλγόριθμων δέντρων αποφάσεων και χρησιμοποιεί αρκετά μικρότερα δέντρα αποφάσεων με βαθμολογημένα αποτελέσματα, όπου κάθε αποτέλεσμα παίρνει διαφορετική βαθμολογία όταν διατρέχει το δέντρο. Ο αλγόριθμος XGBoost παρέχει μια μέθοδο εξαιρετικά ενισχυμένης κλίσης (extreme regularizing gradient boosting framework) για την βελτίωση της επίδοσης του μοντέλου και επιλέγεται μεταξύ άλλων υλοποιήσεων ενισχυμένων δέντρων λόγω της ανωτερότητάς του στην ταχύτητα. Ο πιο σημαντικός παράγοντας πίσω από την επιτυχία του XGBoost είναι η επεκτασιμότητα του σε όλα τα σενάρια. Το σύστημα τρέχει τουλάχιστον δέκα φορές πιο γρήγορα από τις υπάρχουσες λύσεις σε ένα μόνο μηχάνημα και κλιμακώνεται σε δισεκατομμύρια παραδείγματα σε κατανεμημένα συστήματα, με την επεκτασιμότητα του να οφείλεται σε αλγοριθμικές και έξυπνες βελτιστοποιήσεις όπως το κλάδεμα δέντρων και η βελτιστοποίηση υπολογιστικών πόρων [22].

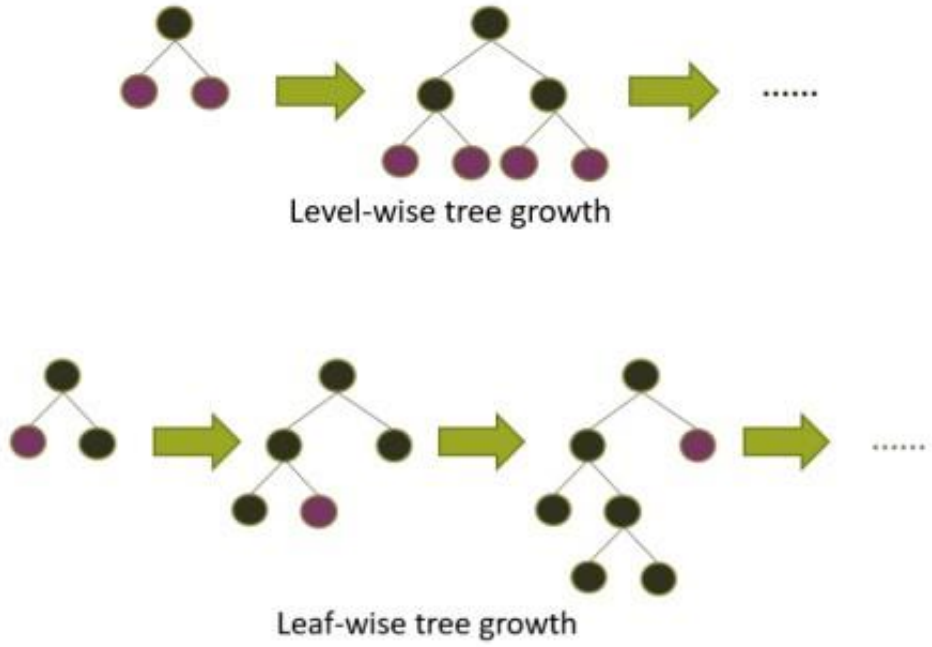
Στο σημείο αυτό, αξιοσημείωτο είναι ότι αντί να αλλάξει η βαρύτητα των στιγμιότυπων όπου ταξινομήθηκαν λάθος, γίνεται εκπαίδευση κάθε νέου μοντέλου αξιοποιώντας τα υπολειπόμενα σφάλματα του προηγούμενου, αποτελώντας με αυτόν τον τρόπο την διαφορά μεταξύ της ενίσχυσης κλίσης και της προσαρμοσμένης ενίσχυσης. Η επαναληπτική αυτή διαδικασία σταματά στην περίπτωση που το σφάλμα του μοντέλου παραμείνει αμετάβλητο ή επιτευχθεί το μέγιστο όριο του αριθμού των μοντέλων.



Εικόνα 7. Αλγόριθμος Ακραίας Ενίσχυσης Κλίσης

2.4.9 Αλγόριθμος Light Gradient Boosting Machine (LightGBM)

Ο Light Gradient Boosting Machine ανήκει στην οικογένεια αλγορίθμων που βασίζονται σε δέντρα αποφάσεων και ενίσχυση κλίσης. Ο αλγόριθμος αυτός δημιουργήθηκε από τον Guolin Ke στην Microsoft [23]. Όπως και άλλοι αλγόριθμοι GBDT (όπως ο XGBoost), ο LightGBM εξακολουθεί να χρησιμοποιεί τη μέθοδο προσαρμογής αρνητικής διαβάθμισης κατά την εκπαίδευση των δέντρων αποφάσεων. Με στόχο το πρόβλημα της χαμηλής απόδοσης και του μεγάλου υπολογιστικού κόστος των προηγούμενων αλγόριθμων GBDT, ο LightGBM εισάγει βελτιωμένους μηχανισμούς, όπωςGradient-based One-Side Sampling (GOSS) και Exclusive Feature Bundling (EFB), για την επίλυση των ελαττωμάτων του προηγούμενου αλγόριθμου GBDT, ο οποίος είναι χρονοβόρος και αναποτελεσματικός ενόψει των υψηλών διαστάσεων χαρακτηριστικών ή του μεγάλου όγκου δεδομένων. Ο μηχανισμός GOSS διατηρεί μια ισορροπία μεταξύ της μείωσης των δεδομένων εκπαίδευσης και της διασφάλισης της ακρίβειας. Διατηρεί όλα τα δείγματα με μεγάλη κλίση και τυχαία δείγματα με μικρή διαβάθμιση για να αποτρέψει την αλλαγή της διανομής δεδομένων και να επηρεάσει την ακρίβεια της εκπαίδευσης, κάτι που είναι παρόμοιο με την αρχή του αλγορίθμου Adaboost. Ο LightGBM διαφέρει ως προς το γεγονός ότι αναπτύσσει οριζόντια δέντρα, δηλαδή επιλέγει να αναπτύξει το φύλλο που πιστεύει ότι θα οδηγήσει στη μεγαλύτερη μείωση της απώλειας (Leaf-wise tree growth), ενώ οι υπόλοιποι αλγόριθμοι αναπτύσσουν δέντρα ανά επίπεδο (Level-wise tree growth), βασιζόμενοι στο βάθος του δέντρου. Επιπλέον, δεν χρησιμοποιεί τον σύνηθες αλγόριθμο δέντρων αποφάσεων, αλλά έναν εξαιρετικά βελτιστοποιημένο αλγόριθμο που βασίζεται σε ιστογράμματα. Όπως παρουσιάζεται και στην εικόνα 8, είναι ευδιάκριτη η διαφορετική προσέγγιση που επιλέγει ο LightGBM σε αντιπαράθεση με τον XGBoost. Τέλος η μικρή χρήση μνήμης, η παράλληλη μάθηση και η ευκολία του να διαχειρίζεται μεγάλους όγκους δεδομένων αποτελούν τα βασικά πλεονεκτήματα του αλγόριθμου [24].



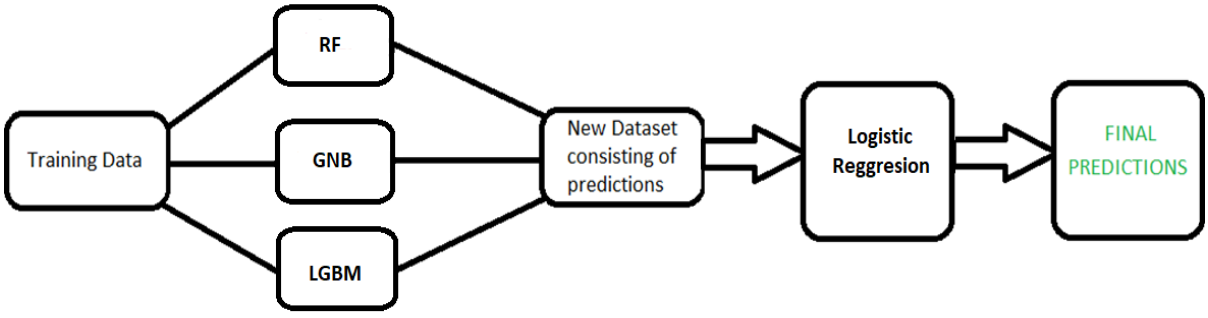
Εικόνα 8. Αλγόριθμος Light Gradient Boosting Machine

2.4.10 Stacked Ensemble learning

Η μέθοδος μάθησης Stacking προτάθηκε αρχικά από τον David H. Wolpert [25] και έγινε ευρέως γνωστή με την ονομασία Stacked Generalization. Μία από τις πιο αποτελεσματικές προσεγγίσεις στα προβλήματα ταξινόμησης που χρησιμοποιεί πολλούς ταξινομητές για να επιτύχει υψηλή γενίκευση. Η λογική είναι ότι συνδυάζοντας ταξινομητές με διαφορετικές επαγωγικές προκαταλήψεις, ο χώρος χαρακτηριστικών θα διερευνηθεί με διαφορετικό τρόπο, με αποτέλεσμα διαφορετικούς ταξινομητές των οποίων τα λάθη δεν συσχετίζονται. Ως εκ τούτου, όταν συνδυάζονται σε ένα μετα-ταξινομητή, αυτοί οι ταξινομητές θα μάθουν ο ένας από τον άλλον τα λάθη τους με αποτέλεσμα την δημιουργία ενός αποδοτικότερου μοντέλου [26].

Η κύρια ιδέα λοιπόν είναι η χρησιμοποίηση προβλέψεων μοντέλων μηχανικής μάθησης από το προηγούμενο επίπεδο ως μεταβλητές εισόδου για μοντέλα στο επόμενο επίπεδο. Η χρήση μοντέλων πολλαπλών επιπέδων με τη χρήση της μεθόδου stacking είναι πολύ δημοφιλής. Ουσιαστικά η στοίβαξη (stacking) είναι η διαδικασία διδασκαλίας ενός αλγορίθμου μάθησης για τη συγχώνευση των προβλέψεων πολλών αλγορίθμων μάθησης. Αρχικά κάθε ένας από τους αλγόριθμους που ανήκουν στο πρώτο επίπεδο, εκπαιδεύεται πρώτα με το αρχικό σύνολο δεδομένων και στην συνέχεια διδάσκεται ένας αλγόριθμος σύντηξης χρησιμοποιώντας όλες τις προβλέψεις των άλλων αλγορίθμων. Τα αποτελέσματα στις περισσότερες περιπτώσεις με την χρήση της μεθόδου εκπαίδευσης stacking υπερέχουν όλων των μοντέλων ξεχωριστά.

Επίσης, σε αντίθεση με την ενίσχυση που περιγράφηκε προηγουμένως, η στοίβαξη χρησιμοποιεί ένα συγκεκριμένο πρότυπο για να μάθει την καλύτερη ενσωμάτωση των παρατηρήσεων από τα συμμετέχοντα μοντέλα, αντί της χρησιμοποίησης μιας σειράς αλγορίθμων που διορθώνουν τις προβλέψεις προηγούμενων μοντέλων. Τα μοντέλα στο stacking είναι συνήθως διαφορετικά και λειτουργούν στο ίδιο σύνολο δεδομένων. Ο τελικός αλγόριθμος γνωρίζει πότε να χρησιμοποιήσει κάθε μοντέλο στο σύνολο δεδομένων και χρησιμοποιείται για να αντιμετωπίσει το πρόβλημα ως τελικός εκτιμητής. Η δομή ενός μοντέλου στοίβαξης, όπως παρουσιάζεται στην εικόνα 9, αποτελείται από πολλαπλά βασικά μοντέλα, γνωστά και ως μοντέλα επιπέδου-0, τα οποία ταιριάζουν στο σύνολο δεδομένων εκπαίδευσης και ένα μετα-μοντέλο που ενσωματώνει τις προβολές των βασικών μοντέλων, γνωστό ως μοντέλο επιπέδου-1, που μαθαίνει πώς να συνδυάζει καλύτερα τις προηγούμενες προβλέψεις [27]. Όπως κάθε τεχνική έχει ποικίλες δυνατότητες για την μοντελοποίηση των δεδομένων και παρουσιάζει πλεονεκτήματα και μειονεκτήματα, ανάλογα με την εφαρμογή ή το πρόβλημα που πρέπει να αντιμετωπιστεί. Όμως, η μέθοδος εκπαίδευσης stacking απαντά πολλές φορές στο ζήτημα επιλογής μεταξύ διαφόρων μοντέλων μηχανικής μάθησης, που είναι ειδικευμένα στην επίλυση ενός προβλήματος με διαφορετικούς τρόπους.



Εικόνα 9. Stacked Ensemble learning

2.5 Μη ισορροπημένα δεδομένα

Οι συνεχείς εξελίξεις στην επιστήμη και την τεχνολογία έχουν οδηγήσει στην συνεχώς αυξανόμενη διαθεσιμότητα ακατέργαστων δεδομένων με εκρηκτικό ρυθμό. Αυτό έχει δημιουργήσει μία τεράστια ευκαιρία για ανακάλυψη γνώσης, με την μηχανική μάθηση να διαδραματίζει ουσιαστικό ρόλο σε ένα ευρύ φάσμα εφαρμογών από την καθημερινή ζωή των πολιτών έως την παρακολούθηση της υγείας τους. Εύκολα λοιπόν διακρίνεται η αναγκαιότητα της γνώσης από ακατέργαστα δεδομένα για την υποστήριξη των διαδικασιών λήψης αποφάσεων. Αν και οι υπάρχουσες τεχνικές εξόρυξης γνώσης έχουν δείξει μεγάλη επιτυχία σε πολλές εφαρμογές του πραγματικού κόσμου, τα συστήματα υποστήριξης λήψης αποφάσεων, έρχονται αντιμέτωπα με το πρόβλημα της ανισορροπίας των συνόλων δεδομένων. Ένα πρόβλημα που έχει συγκεντρώσει σημαντικό ενδιαφέρον στην επιστημονική κοινότητα, τη βιομηχανία και τους κρατικούς φορείς χρηματοδότησης. Το θεμελιώδες ζήτημα είναι η ικανότητα του μη ισορροπημένου συνόλου δεδομένων να επηρεάζει σημαντικά την απόδοση των περισσότερων τυπικών αλγορίθμων μηχανικής μάθησης.

Ουσιαστικά, η μη ισορροπημένη μάθηση, προσδιορίζει την αδυναμία στην απόδοση των αλγορίθμων μάθησης παρουσία υποεκπροσωπούμενων δεδομένων. Οι αλγόριθμοι αναμένουν ισορροπημένες κατανομές κλάσεων ή ίσο κόστος λανθασμένης ταξινόμησης, οπότε όταν εφαρμόζονται σε πολύπλοκα μη ισορροπημένα σύνολα δεδομένων, αποτυγχάνουν να αντιπροσωπεύουν σωστά τα χαρακτηριστικά των δεδομένων. Λόγω της πολυπλοκότητας των χαρακτηριστικών, η εκμάθηση από τέτοια δεδομένα απαιτεί νέες αρχές και αλγόριθμους για τη μετατροπή τεράστιων ποσοτήτων ακατέργαστων δεδομένων αποτελεσματικά σε πληροφορία και γνώση [28].

Η διαδικασία εξισορρόπησης ενός συνόλου δεδομένων, βασικά χωρίζεται σε τρεις μεθόδους: υποδειγματοληψία, υπερδειγματοληψία και υβριδική υπερδειγματοληψία και υποδειγματοληψία. Στην παρούσα έρευνα, για την εξισορρόπηση των δεδομένων πραγματοποιήθηκε υπερδειγματοληψία συνθετικής μειονότητας και υποδειγματοληψία με την χρήση του κανόνα καθαρισμού γειτονιάς (NCR).

2.5.1 Smote: Τεχνική Υπερδειγματοληψίας Συνθετικής Μειονότητας

Η υπερδειγματοληψία πραγματοποιήθηκε μόνο στο σετ εκπαίδευσης, προκειμένου να αντιμετωπιστεί το πρόβλημα της «ανισορροπίας τάξης» που αναφέρεται σε πολύ περισσότερες περιπτώσεις με κλάση 0 («ασυνέπεια») από ό,τι στην κατηγορία 1 («συνέπεια») που δυσκολεύει το μοντέλο στην ακριβή διάκριση μεταξύ τους. Η υπερδειγματοληψία υλοποιήθηκε με την τεχνική Smote που δημιουργεί παραδείγματα από την τάξη μειοψηφίας με βάση τους K πλησιέστερους γείτονες των παραδειγμάτων αυτής, πολλαπλασιάζοντας τη διαφορά διανυσμάτων με ένα τυχαίο αριθμό στο διάστημα [0,1] [29]. Μέθοδοι που εξετάστηκαν:

- Simple SMOTE: κατασκευή νέων παραδειγμάτων επιλέγοντας τυχαίες παρουσίες από την κατηγορία μειοψηφίας.
- Borderline SMOTE: κατασκευή νέων παραδειγμάτων επιλέγοντας εκείνες τις παρουσίες της κατηγορίας μειοψηφίας που δεν έχουν ταξινομηθεί σωστά, δηλαδή τις «δύσκολες περιπτώσεις».

Με την εφαρμογή της τεχνικής Borderline SMOTE παράχθηκαν καλύτερα αποτελέσματα ως προς την ακρίβεια των αλγορίθμων, όπως αυτά παρουσιάζονται σε ενότητα των πειραματικών διεργασιών.

2.5.2 Neighborhood Cleaning Rule: Κανόνας καθαρισμού θορύβου

Ο κανόνας καθαρισμού Neighborhood Cleaning Rule (NCR) με σκοπό τον καθαρισμό των δεδομένων, είναι μια τεχνική υποδειγματοληψίας που συνδυάζει τόσο τον κανόνα Condensed Nearest Neighbor (CNN) για την κατάργηση περιττών παραδειγμάτων όσο και τον κανόνα Edited Nearest Neighbors (ENN) για την αφαίρεση θορυβωδών ή διφορούμενων παραδειγμάτων. Αρχικά η μέθοδος NCR αφαιρεί τα αρνητικά παραδείγματα που ταξινομούνται εσφαλμένα από τους 3 πλησιέστερους γείτονές τους. Ακολούθως, εντοπίζει τους γείτονες κάθε θετικού παραδείγματος και αφαιρεί όσους ανήκουν στην πλειοψηφική τάξη. Ουσιαστικά, χρησιμοποιεί την αρχή της μονόπλευρης επιλογής (one-sided selection principle), αλλά εξετάζει πιο προσεκτικά την ποιότητα των δεδομένων που πρέπει να αφαιρεθούν. Ο λόγος δεν είναι άλλος από το γεγονός ότι τα θορυβώδη παραδείγματα πιθανότατα θα ταξινομηθούν εσφαλμένα και πολλά από αυτά θα χρησιμοποιηθούν στην διαδικασία εκπαίδευσης, θα οδηγήσουν και στην εσφαλμένη ταξινόμηση αρκετών επόμενων δοκιμαστικών παραδειγμάτων [28]. Εστιάζει λοιπόν λιγότερο στη βελτίωση της ισορροπίας της κατανομής της τάξης και περισσότερο στην ποιότητα των παραδειγμάτων που διατηρούνται στην πλειοψηφική τάξη.

Η συγκεκριμένη προσέγγιση αντιμετώπισης μη ισορροπημένων συνόλων δεδομένων, προτάθηκε από την Jorma Laurikkala το 2001, υπογραμμίζοντας ότι η ποιότητα των αποτελεσμάτων ταξινόμησης δεν εξαρτάται απαραίτητα από το μέγεθος της τάξης. Επομένως, θα πρέπει να ληφθούν υπόψη, εκτός από την κατανομή κλάσης, και άλλα χαρακτηριστικά των δεδομένων, όπως ο θόρυβος, που μπορεί να εμποδίσουν την ταξινόμηση [30].

Data: T (the original training set)
Result: S (reduced data)
begin
  Split data T into the class of interest C and the remaining data 0;
  Identify noisy data A_1 in 0 with edited nearest neighbor rule;
  for *each class* C_i *in* 0 **do**
    if ($x \in C_i$ *in 3-nearest neighbors of misclassified* $y \in C$) *and* ($|C_i| > 0.5|C|$) **then**
      $A_2 = \{x\} \cup A_2$;
    end
  end
  $S = T - (A_1 \cup A_2)$;
end

Εικόνα 10. Neighborhood Cleaning Rule [30]

2.6 Cross Validation

Για την υλοποίηση εκπαίδευσης και επικύρωσης δεν χρησιμοποιήθηκε κάποιο train-test split, αλλά έγινε χρήση του stratified cross validation για την αξιολόγηση της ταξινόμησης με τον δυνατότερο αμερόληπτο τρόπο. Σύμφωνα με το stratified cross validation τα δεδομένα χωρίζονται τυχαία σε κομμάτια, ανάλογα με τις δίπλες (folds) που ορίζει ο χρήστης, διατηρώντας την μη ισορροπημένη κατανομή κλάσεων σε κάθε fold. Όπως αναφέρεται και στο [31], είναι σύνηθες ιδίως στην περίπτωση μη ισορροπημένων κλάσεων, να χρησιμοποιείται η στρωματοποιημένη δεκαπλάσια διασταυρούμενη επικύρωση (10-fold stratified cross-validation), η οποία διασφαλίζει ότι η αναλογία των θετικών προς των αρνητικών παραδειγμάτων που βρέθηκαν στην αρχική κατανομή τηρείται σε όλες τις πτυχές. Αναλυτικότερα το σύνολο των παραδειγμάτων χωρίζεται σε Ν υποσύνολα με το κάθε ένα από αυτά να χρησιμοποιείται διαδοχικά ως σύνολο ελέγχου και τα υπόλοιπα Ν-1 υποσύνολα να ενώνονται και να δημιουργούν το σύνολο εκπαίδευσης για τους αλγορίθμους. Στο τέλος των Ν εκπαιδεύσεων, χρησιμοποιούνται τα αποτελέσματα για να βγει ένας μέσος όρος ακρίβειας για το κάθε μοντέλο ξεχωριστά [32].

2.7 Μετρικές Αξιολόγησης

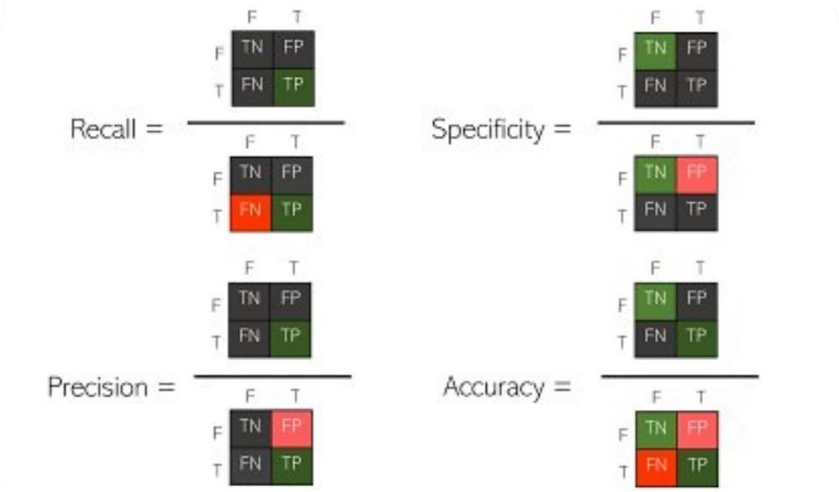
Στην προσπάθεια λοιπόν να συγκρίνουμε τα αποτελέσματα των αλγορίθμων ταξινόμησης προκειμένου να βρούμε την καλύτερη δυνατή μέθοδο για την πρόβλεψη πρόωρου τερματισμού χρήσης της εφαρμογής από τον χρήστη, παρουσιάζονται παρακάτω οι μετρικές με τις οποίες γίνεται εφικτή η διαδικασία της σύγκρισης. Στους μαθηματικούς υπολογισμούς που χρησιμοποιούνται ως μέτρα αξιολόγησης για τις τεχνικές ταξινόμησης και παρουσιάζονται παρακάτω, ορίζονται σύμφωνα με την [33] ως:

- True Positives (TP): Ο αριθμός των περιπτώσεων που προβλέπονται ως «συνεπής» (π.χ.1) από έναν δυαδικό ταξινομητή και είναι επίσης θετικές στην πραγματικότητα.

- True Negatives (TN): Ο αριθμός των περιπτώσεων που προβλέπονται ως «ασυνεπής» (π.χ. 0) από έναν δυαδικό ταξινομητή και είναι επίσης αρνητικές στην πραγματικότητα.

- False Positives (FP): Ο αριθμός των περιπτώσεων που προβλέφθηκαν λανθασμένα ως «θετικές» (είναι αρνητικές στην πραγματικότητα).

- False Negatives (FN): Ο αριθμός των περιπτώσεων που προβλέπονται λανθασμένα ως «αρνητικές» (είναι θετικές στην πραγματικότητα).



Εικόνα 11. Μετρικές Αξιολόγησης Αλγορίθμων

Συγκεκριμένα οι μετρικές αξιολόγησης των προβλέψεων που παρατηρήθηκαν με την χρήση των παραπάνω αλγορίθμων ταξινόμησης αναφέρονται παρακάτω:

Sensitivity (Recall) είναι ένας δείκτης ο οποίος μετράει τις θετικές προβλέψεις και υπολογίζεται από την μαθηματική σχέση:

$$\text{Sensitivity (Recall)} = TP/(TP+FN)$$

Specificity είναι ένας δείκτης ο οποίος μετράει τις αρνητικές προβλέψεις και υπολογίζεται ως εξής:

$$\text{Specificity} = TN/(TN+FP)$$

Precision είναι ο λόγος των μονάδων που κατετάγησαν ορθά στην κατηγορία TP προς το σύνολο των μονάδων που κατετάγησαν σε αυτήν είτε ορθά είτε λανθασμένα:

$$\text{Precision} = TP/(TP+FP)$$

Ακρίβεια (Accuracy) είναι ο δείκτης επιτυχίας του μοντέλου πρόβλεψής, σταθμίζει τα θετικά και τα αρνητικά εξίσου, αντί να δίνει προτεραιότητα το ένα έναντι του άλλου και υπολογίζεται με τον μαθηματικό τύπο:

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

Βαθμολογία F1 (F1 score) είναι ο δείκτης ο οποίος δίνει ίση βαρύτητα στην ακρίβεια και την ανάκληση (αρμονικός μέσος) και υπολογίζεται μαθηματικά:

$$F1\ \text{score} = 2(\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$$

Σε αυτό το σημείο, πρέπει να υπογραμμιστεί ότι η βαθμολογία F1 θα πρέπει να χρησιμοποιείται όταν η αποφυγή λαθών είναι πιο σημαντική, καθώς τα ψευδώς θετικά και τα ψευδώς αρνητικά τιμωρούνται περισσότερο, ενώ η ακρίβεια θα πρέπει να χρησιμοποιείται όταν ο στόχος του μοντέλου είναι να βελτιστοποιήσει την απόδοση. Και οι δύο μετρήσεις χρησιμοποιούνται με βάση το περιβάλλον και έχουν διαφορετική απόδοση ανάλογα με τα δεδομένα. Γενικά, ωστόσο, η βαθμολογία F1 είναι καλύτερη για μη ισορροπημένες κατηγορίες (για παράδειγμα στα ιατρικά δεδομένα, όταν υπάρχουν πολύ περισσότερα αρνητικά από θετικά αποτελέσματα), ενώ η ακρίβεια είναι καλύτερη για πιο ισορροπημένες κατηγορίες.

2.8 Challenge Metric

Η βαθμολογία κάθε πρόβλεψης θα καθορίζεται από τον γεωμετρικό μέσο όρο της επιτευχθείσας ευαισθησίας (Sensitivity) και ειδικότητας (Specificity):

$$\text{score} = \text{GE} = \sqrt{\text{SE} \times \text{SP}}$$

Ο κώδικας python που αναπτύχθηκε για τον υπολογισμό του συγκεκριμένου δείκτη αναφέρεται και στο Παράρτημα ΙΙ.

2.9 Γλώσσα & Βιβλιοθήκες Προγραμματισμού

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για την επίλυση του συγκεκριμένου προβλήματος είναι η Python 3.9.4, μία πανίσχυρη γλώσσα προγραμματισμού με ευρύ φάσμα εφαρμογών και στον τομέα της Ιατρικής [34]. Η επιλογή έγινε λόγω της ύπαρξη μιας σειράς από εργαλεία που εξυπηρετούν τον τομέα αυτόν, αλλά και την δυνατότητα που παρέχει η συγκεκριμένη γλώσσα με την δημιουργία βιβλιοθηκών, οι οποίες εσωτερικά τρέχουν σε C++ πετυχαίνοντας ταχύτατες επιδόσεις. Το σύνολο των python βιβλιοθηκών που χρειάζονται για την εύρυθμη λειτουργία του συστήματος έγκαιρης προειδοποίησης, αναφέρονται στο Παράρτημα ΙΙΙ.

ΚΕΦΑΛΑΙΟ 3. ΠΑΡΟΥΣΙΑΣΗ ΠΡΟΒΛΗΜΑΤΟΣ & ΔΕΔΟΜΕΝΩΝ

3.1 Ορισμός του Προβλήματος

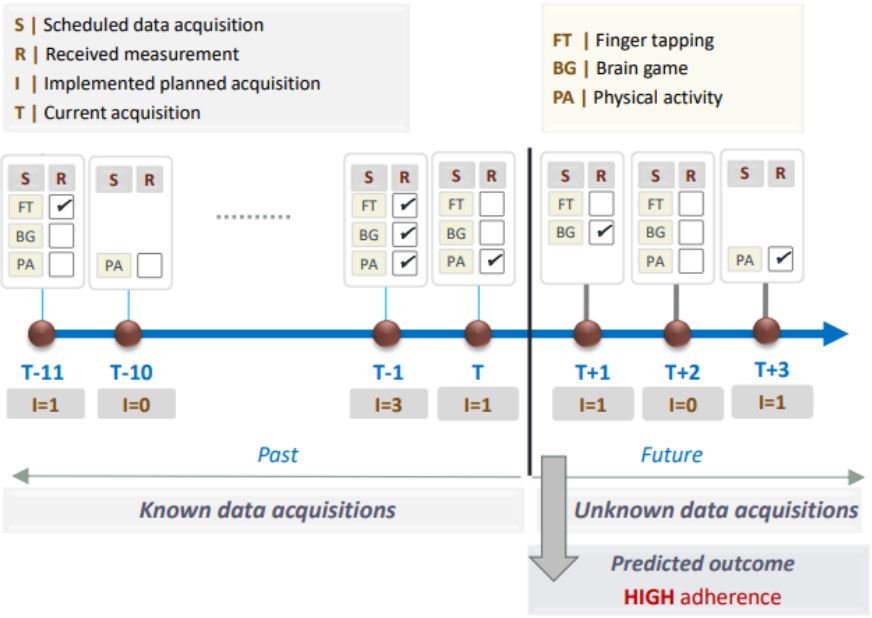
Οι εξελίξεις στον τομέα των τηλεπικοινωνιακών τεχνολογιών, με τη στιβαρότητα και την αξιοπιστία που παρέχουν, έχουν συμβάλει σημαντικά στην πρόοδο και ανάπτυξη στον τομέα της ιατρικής, προβάλλοντας παράλληλα την ανάγκη αξιοποίησής τους στον τομέα της υγείας, έναν κλάδο που έχει χαρακτηριστεί από πολλούς ως τομέας εντατικής πληροφορίας και απαιτητικής γνώσης. Έτσι, είναι προφανές ότι οι λύσεις e-Health είναι κρίσιμης σημασίας [35]. Ο όρος e-Health ορίστηκε από τον Παγκόσμιο Οργανισμό Υγείας (Π.Ο.Υ.) ως «η αποδοτική και ασφαλής χρήση των τεχνολογιών πληροφορίας και επικοινωνιών για την υποστήριξη της υγείας αλλά και πεδίων που σχετίζονται με την υγεία, συμπεριλαμβανομένης της υγειονομικής περίθαλψης, της παρακολούθησης και της αγωγής υγείας, της γνώσης και της έρευνας». Με μια ευρύτερη έννοια, ο όρος χαρακτηρίζει όχι μόνο μια τεχνική εξέλιξη, αλλά και έναν τρόπο σκέψης, για τη βελτίωση της υγειονομικής περίθαλψης σε παγκόσμιο επίπεδο χρησιμοποιώντας πληροφορίες και τεχνολογία επικοινωνίας.

Οι νέες τεχνολογίες μπορεί να έχουν την δυνατότητα να συμβάλλουν ριζικά στον μετασχηματισμό του κλάδου υγείας, ωστόσο έρχονται αντιμέτωπες με σημαντικές προκλήσεις. Τα ρυθμιστικά εμπόδια, οι δυσκολίες στην ψηφιοποίηση των δεδομένων των ασθενών και οι οικονομικοί περιορισμοί, καθιστούν επιτακτική ανάγκη τον άμεσο και έγκυρο μετασχηματισμό του κλάδου της υγείας. Στο πλαίσιο αυτό αν συνυπολογιστεί και η γήρανση του πληθυσμού σε παγκόσμιο επίπεδο, εφόσον η βελτίωση των συνθηκών διαβίωσης, η καλύτερη εκπαίδευση και η ευκολότερη πρόσβαση σε υπηρεσίες υγείας οδηγούν στην αύξηση του προσδόκιμου ζωής, αναμένεται να αυξηθεί σημαντικά και η ζήτηση για τις υπηρεσίες υγείας. Στην αναγκαία αυτή αλλαγή λοιπόν, οι ψηφιακές τεχνολογίες διαδραματίζουν κεντρικό ρόλο. Από την πλευρά τους, οι ασθενείς επενδύουν στην πρόληψη με την αυξημένη χρήση ψηφιακών εφαρμογών υγείας. Η αξιοποίηση των τεχνολογιών αυτών γίνεται στο πλαίσιο παρακολούθησης θεμάτων υγείας, καταμέτρησης στόχων υγείας και φυσικής κατάστασης, καθώς και για καταγραφή ιατρικών δεδομένων. Με αυτό τον τρόπο, άλλωστε, οι ασθενείς αποκτούν μεγαλύτερο έλεγχο στη διαχείριση της υγείας τους και για τους ασθενείς με χρόνιες ασθένειες, η συνεχής παρακολούθηση των κρίσιμων δεδομένων μπορεί να είναι σωτήρια [36].

Η ψηφιοποίηση των παρεμβάσεων στον τομέα της υγειονομικής περίθαλψης έχει τη δυνατότητα να ανακουφίσει το βάρος που προκαλείται στα συστήματα υγείας από την αυξανόμενη γήρανση του πληθυσμού και τις χρόνιες ασθένειες. Η απομακρυσμένη παρακολούθηση ασθενών, γνωστή και ως τηλε-παρακολούθηση, είναι μία μέθοδος παροχής υπηρεσιών υγείας με την χρήση συγκεκριμένων τεχνολογιών για την ηλεκτρονική μετάδοση πληροφορίας μεταξύ ασθενών και ιατρών, αποσκοπώντας στην παρακολούθηση ασθενών εκτός νοσοκομειακών εγκαταστάσεων. Ωστόσο, αυτοί οι τύποι τεχνολογικών λύσεων απαιτούν υψηλά ποσοστά συμμόρφωσης για να είναι αποτελεσματικοί και όπως και σε οποιαδήποτε άλλη εφαρμογή, οι χρήστες συχνά εγκαταλείπουν τη χρήση της για διάφορους λόγους.

Σε ένα τέτοιο πλαίσιο, ο έγκαιρος εντοπισμός χρηστών με κίνδυνο χαμηλότερων ποσοστών συμμόρφωσης και μοτίβων χρήσης που υποδηλώνουν κίνδυνο εγκατάλειψης είναι μια ανεκτίμητη ευκαιρία για την εφαρμογή προσαρμοσμένων στρατηγικών παρέμβασης που στοχεύουν στην ανάκαμψη και την αποφυγή αποδέσμευσης των χρηστών. Για αυτό το λόγο, στην φετινή έκδοση του παγκόσμιου διαγωνισμού (IFMBE Science Challenge 2022) [37], στόχος είναι να εντοπιστούν πρότυπα πρόωρης εγκατάλειψης σε χρήστες μιας εφαρμογής για κινητές συσκευές παρεμβάσεων με τίτλο Active and Healthy Ageing (AHA). Παραδοσιακά, η αποδοχή των χρηστών αξιολογήθηκε χρησιμοποιώντας στατικές μεθοδολογίες, επομένως η μεσοπρόθεσμη και μακροπρόθεσμη αποδοχή και δέσμευση δεν έχουν αναλυθεί συστηματικά. Για να επιτευχθεί λοιπόν η δημιουργία ενός συστήματος προειδοποίησης για την εφαρμογή, θα πρέπει να εντοπιστούν προφίλ χρηστών με αυξημένο κίνδυνο πρόωρης εγκατάλειψης.

Η πρόκληση του διαγωνισμού, είναι δεδομένου ενός παραθύρου n=12 διαδοχικών προγραμματισμένων στιγμών απόκτησης δεδομένων, να προβλεφθεί η συμμόρφωση του χρήστη κατά τις προσεχείς 3 προγραμματισμένες στιγμές απόκτησης δεδομένων. Η απόκτηση δεδομένων ακολουθεί ένα καλά καθορισμένο πρωτόκολλο που αποτελείται από τακτικά προγραμματισμένες στιγμές, όπου οι συμμετέχοντες πρέπει να τροφοδοτήσουν με δεδομένα τη μελέτη. Αναλυτικά, η έγκυρη συχνότητα απόκτησης δεδομένων περιοδικών δραστηριοτήτων ορίστηκε δύο φορές ανά εβδομάδα για κάθε δραστηριότητα. Επιπλέον, κάθε προγραμματισμένη λήψη δεδομένων μπορεί να περιλαμβάνει την πραγματική μέτρηση μιας ή περισσότερων περιοδικών δραστηριοτήτων (brain games, physical activity, finger tapping, mindfulness) ανεξάρτητα από την επιτυχία ή αποτυχία του συμμετέχοντα. Ακόμη παρέχονται ορισμένες πληροφορίες περιβάλλοντος, όπως η ημερομηνία. Στο πλαίσιο αυτό, θα πρέπει να προγραμματιστούν διαδοχικές λήψεις δεδομένων με βάση το πρωτόκολλο απομακρυσμένης παρακολούθησης για χρονικές περιπτώσεις i=1,…,n για το διάστημα χρήσης της εφαρμογής από τον χρήστη.



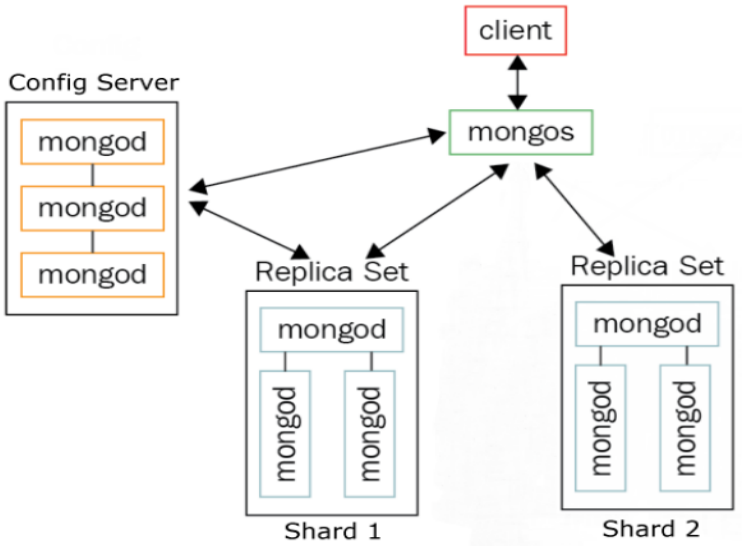
Εικόνα 12. Παράδειγμα υψηλού επίπεδου συμμόρφωσης [38]

3.2 Παρουσίαση Συνόλου Δεδομένων

Στο πλαίσιο της πρόβλεψης πρόωρων εγκαταλείψεων από την εφαρμογή παρακολούθησης ηλικιωμένων, οι συμμετέχοντες του διαγωνισμού έχουν πρόσβαση σε ένα σύνολο δεδομένων με περισσότερους από 150 χρήστες στη Μαδρίτη που έχουν δοκιμάσει τον αντίκτυπο μιας ψηφιακής εφαρμογής AHA για τη βελτίωση της ποιότητας ζωής τους για τουλάχιστον 6 μήνες στο δίκτυο MAHA (Moving Active & Healthy Aging). Το σύνολο δεδομένων αντιπροσωπεύει τα ερωτηματολόγια έναρξης (πριν από την έναρξη χρήσης της εφαρμογής) και τελικής αξιολόγησης (μετά από 6 μήνες χρήσης) από τους συμμετέχοντες. Ακόμα περιέχει εσωτερικές μετρήσεις των λύσεων που προέκυψαν από την αλληλεπίδραση κάθε χρήστη με καθεμία από τις λειτουργίες και τις παρεμβάσεις στην εφαρμογή. Αυτά τα αποτελέσματα αξιολογούν την επιτυχία και την τήρηση της παρέμβασης καθώς και τα στοιχεία καταγραφής κάθε συμμετέχοντα κατά τη χρήση της εφαρμογής (κουμπιά, μενού, κ.λπ.). Κατά τη διάρκεια αυτής της εξάμηνης περιόδου, οι συμμετέχοντες χρησιμοποιούσαν την εφαρμογή για την αντιμετώπιση συγκεκριμένων προβλημάτων του ηλικιωμένου προφίλ τους είτε είχαν να αντιμετωπίσουν προβλήματα γνωστικά, σωματικά ή ακόμα και κοινωνικοποίησης ή κατάθλιψης. Σημαντική λεπτομέρεια κατά το στάδιο της συλλογής των δεδομένων, υπήρξε η οδηγία προς τους χρήστες, όπου ζητήθηκε η συχνή χρήση της εφαρμογής για τουλάχιστον δύο φορές την εβδομάδα, ανάλογα φυσικά και με τις ανάγκες τους.

Αναλυτικότερα, το σύνολο δεδομένων περιέχει ένα σύνολο πινάκων που αποτελείται από ασύγχρονα συμβάντα και δεδομένα που ποικίλλουν μεταξύ κατηγοριών και χρηστών. Επιπλέον αναφορικά με τα διαθέσιμα σύνολα & τύπους δεδομένων, τα δεδομένα στο σύνολο τους είναι αποθηκευμένα σε μία μη σχεσιακή βάση δεδομένων MongoDB και έχουν κατανεμηθεί σε 10 σύνολα δεδομένων, όπως αυτά παρουσιάζονται στην συνέχεια.

- EQ5D3L
- SPQ
- Sociodemo
- UCLA
- brain_games
- fingertapping
- physical_activity
- mindfulness
- UTAUT
- digital_phenotyping



Εικόνα 13. Παράδειγμα της Βάσης Δεδομένων της Εφαρμογής

Οι τύποι δεδομένων που εντοπίστηκαν στις παραπάνω συλλογές ποικίλλουν και συμβάλλουν στην πληρέστερη καταγραφή δεδομένων για έναν χρήστη της εφαρμογής.

- Κοινωνικοδημογραφικές πληροφορίες (Sociodemographic)

  Οι κοινωνικοδημογραφικές πληροφορίες ασθενών, περιλαμβάνουν επιπλέον τον τύπο της συσκευής που χρησιμοποιούν για να έχουν πρόσβαση στις εφαρμογές (κινητό ή tablet), τις λύσεις στον ΜΑΗΑ, βασικές πληροφορίες για την παρέμβαση που δέχονται, καθώς και τις πληροφορίες σχετικά με την εμπλοκή τους κατά το διάστημα παρακολούθησης (αποχώρηση ή όχι).

- Ποιότητα ζωής (Quality of Life - QoL)

  Δεδομένα QoL που αντιπροσωπεύουν την αυτοαντιλαμβανόμενη ποιότητα ζωής και την κατάσταση ευημερίας του χρήστη κατά την έναρξη της μελέτης και στο τέλος (έξι μήνες αργότερα).

- Δεδομένα αποδοχής (Acceptability)

  Δεδομένα χρηστικότητας και αξιολόγησης αποδοχής των λύσεων.

- Μετρήσεις εφαρμογής (Application Measurements)

  Αναπαριστούν εσωτερικές μετρήσεις των λύσεων ενός χρήστη, ανάλογα με τον τελικό σκοπό τους, δηλαδή την διάρκεια και την δυσκολία της συνεδρίας, την τελική βαθμολογία και την οριστικοποίηση της συνεδρίας.

- Ερωτηματολόγιο αυτοαντίληψης (Self-Perception Questionnaire - SPQ)

  Περιλαμβάνει πεδία που αντιπροσωπεύουν την αυτοαντιληπτή επιρροή των δοκιμασμένων λύσεων της εφαρμογής στην ποιότητα ζωής των χρηστών.

- Ψηφιακός φαινότυπος χρήστη (User Phenotype)

  Περιέχει όλες τις καταγεγραμμένες πληροφορίες των ενεργειών που πραγματοποιεί ο χρήστης της εφαρμογής του δικτύου ΜΑΗΑ. Συγκεκριμένα περιλαμβάνει όλα τα πατημένα κουμπιά και τις ενέργειες που πραγματοποιήθηκαν κατά τη διάρκεια της αλληλεπίδρασης με τα μενού εφαρμογών που στοχεύει στην παροχή πληροφοριών σχετικά με τη χρήση των εφαρμογών για κινητά από τους χρήστες και στην λήψη στατιστικών στοιχείων για τη συνολική χρήση της εφαρμογής.

- Αποδοχή και χρήση τεχνολογίας (Unified Theory of Acceptance and Use of Technology - UTAUT)

  Ερωτηματολόγιο που ακολουθεί τα αποτελέσματα της προσέγγισης του μοντέλου UTAUT, για την αξιολόγηση της αποδοχής της λύσης στο τέλος της μελέτης.

USER ID: MDS000

ACCEPTABILITY

SOCIODEMOGRAPHIC	
record_id	**MDS000**
Questionnaire	sociodemo
gender	1
year_of_birth	1947
educational_level	2
technology_level	1
living_environment	1
living_conditions	1
living_status	1
date_of_termination	11/7/18
Date_of_finalization	15/1/19
Device	Own mobile
Status	Dropout

ACCEPTABILITY SPQ*		
ID	**MDS000**	
Instance	1	3
Q1	6	9
Q2		5
Q3	5	8
Q4		5
Q5	7	9
Q6		5

ACCEPTABILITY UTAUT	
ID	**MDS000**
ee1_my_interaction_with_th	4
ee2_it_is_easy_for_me_to_b	6
ee3_i_find_the_iot_device	6
ee4_learning_to_operate_th	6
pe1_i_find_the_iot_device	3
pe2_using_the_iot_device_e	4
pe3_using_the_iot_device_i	6
pe4_if_i_use_the_iot_devic	3
at1_using_the_iot_device_i	6
at2_the_iot_device_makes_s	6
at3_working_with_the_iot_d	3
at4_i_like_working_using_w	4
si1_people_who_influence_m	2
si2_people_who_are_importa	5
si3_professors_in_my_class	6
si4_in_general_the_doctor	7
fc1_i_have_the_resources_n	3
fc2_i_have_the_knowledge_n	4
fc3_the_iot_device_is_not	4
fc4_a_specific_person_or_g	7
se1_i_can_complete_a_job_o	7
se2_i_can_complete_a_job_o	3
se3_i_can_complete_a_job_o	4
se4_i_can_complete_a_job_o	5
ax1_i_feel_apprehensive_ab	5
ax2_it_scares_me_to_think	6
ax3_i_hesitate_to_use_the	3
ax4_the_iot_device_is_some	4
bi1_i_intend_to_use_the_io	5
bi2_i_predict_i_would_use	5
bi3_i_plan_to_use_the_iot	5

USER PHENOTYPE			
uid	**MDS000**		
start	19/12/19 17:16	19/12/19 17:16	19/12/19 17:22
log	Brain_Games/ Go_to_Cognitiv e_training	Brain_Games/Cogniti ve_training/Go_to_ Word_Search_game	Brain_Games/ Cognitive_training/ Word_search_completed

QUALITY OF LIFE

QUALITY OF LIFE	
ID	**MDS000**
Questionnaire	EQ5D3L
Instance	1
Mobility	1
Self-care	1
UsualActivities	1
Pain	1
Anxiety	2

UCLA*2	
ID	**MDS000**
Questionnaire	EQ5D3L
Instance	3
Q1	4
Q2	2
...	...
Q19	1
!20	2

APPLICATION MEASUREMENTS

PHYSICAL ACTIVITY		
uid	**MDS000**	
start	21/3/19 0:02	12/12/18 21:02
solved	0	0
type	upperlimbs	upperlimbs

MINDFULNESS	
uid	**MDS000**
start	6/7/18 18:29
status_practice	INCOMPLETE
duration	27

FINGER TAPPING	
uid	**MDS000**
start	6/2/19 19:53
taps	3
errors	4
mean_rt	1067,14
std_rt	633,49
max_rt	2000
min_rt	491
solved	False
bilateral	False
type	fta_tapOnTheDrumOrNot

BRAIN GAMES	
uid	**MDS000**
start	5/7/18 10:58
difficulty	2
duration	21
solved	0
type	puzzle

Εικόνα 14. Παράδειγμα διαθέσιμων δεδομένων χρήστη [38]

ΚΕΦΑΛΑΙΟ 4. Σύστημα Έγκαιρης Προειδοποίησης

Στόχος της εργασίας, όπως παρουσιάστηκε αναλυτικά και σε προηγούμενο κεφάλαιο, είναι η δημιουργία των καλύτερων δυνατών μοντέλων για τον εντοπισμό πρόωρων εγκαταλείψεων. Μια πιθανή εφαρμογή ενός τέτοιου μοντέλου θα μπορούσε να είναι η ελαχιστοποίηση των προβλημάτων συμμόρφωσης που μπορούν να θέσουν σε κίνδυνο τον αντίκτυπο αυτών των τύπων εφαρμογών υγειονομικής περίθαλψης, παρατηρώντας τις μετρήσεις της εφαρμογής, τα πρότυπα χρήσης και τα αποτελέσματα αξιολόγησης. Η έγκαιρη αναγνώριση της κλινικής επιδείνωσης είναι ένα από τα κύρια βήματα για τη μείωση της νοσηρότητας και της θνησιμότητας των ασθενών [39]. Στο πλαίσιο αυτό, ο σκοπός ενός συστήματος έγκαιρης προειδοποίησης στην λειτουργία μιας εφαρμογής παρακολούθησης υγείας ηλικιωμένων ανθρώπων, είναι να εκδίδει προειδοποιητικά σήματα πριν από γεγονότα πιθανής εγκατάλειψης από τον χρήστη.

4.1 Προεπεξεργασία Δεδομένων & Δημιουργία Χαρακτηριστικών

Το στάδιο της προεπεξεργασίας δεδομένων είναι απαραίτητο, καθώς η ύπαρξη προβλημάτων είναι ο κανόνας στα δεδομένα του πραγματικού κόσμου. Στα διάφορα είδη προβλημάτων συγκαταλέγονται οι ασυνέπειες ως προς την κωδικοποίηση, την ονομασία και τις μονάδες μέτρησης πεδίων, η ύπαρξη χαμένων τιμών και δεδομένων χωρίς ουσιαστικό περιεχόμενο (θόρυβος). Στη διεθνή βιβλιογραφία της εξόρυξης δεδομένων, ένας από τους όρους που έχει επικρατήσει και περιγράφει δεδομένα με χαμένες τιμές, θόρυβο και άλλα προβλήματα, είναι ο όρος «ακάθαρτα δεδομένα» (dirty data) [40]. Τα ακάθαρτα δεδομένα μπορούν να προκαλέσουν σύγχυση στους αλγορίθμους εξόρυξης και για τον λόγο αυτό, άλλωστε, κρίνεται αναγκαίο να προηγηθεί χρονικά η αντιμετώπιση των προβλημάτων από την έναρξη της ανάλυσης.

Εικόνα 15. Στάδια Ανάλυσης Δεδομένων

Η διαδικασία αντιμετώπισης των χαμένων τιμών, του θορύβου, των ασυνεπειών και άλλων προβλημάτων των δεδομένων ονομάζεται «καθαρισμός δεδομένων» (data cleaning) και αποτελεί μέρος των εργασιών της προεπεξεργασίας τους. Αναλυτικά, ως καθαρισμό δεδομένων ορίζεται η διαδικασία εντοπισμού και διόρθωσης ή αφαίρεσης κατεστραμμένων ή ανακριβών εγγραφών από ένα σύνολο δεδομένων αποθηκευμένο σε πίνακες ή μία βάση δεδομένων. Σημαντικό προς αναφορά είναι δε το γεγονός, ότι με γνώμονα την βελτίωση της ποιότητας των δεδομένων τους, πολλοί οργανισμοί προχώρησαν στην καθιέρωση κανόνων καταγραφής δεδομένων. Αν και τα κατάφεραν σε σημαντικό βαθμό να περιορίσουν το πρόβλημα, δεν είναι σε θέση να το εξαλείψουν τελείως και ο λόγος δεν είναι άλλος, από τα ιστορικά δεδομένα, που έχουν συλλεχτεί σε προγενέστερο χρόνο.

Για την δημιουργία προβλεπτικών μοντέλων, δημιουργήθηκε ένα σύνολο δεδομένων με την εξαγωγή δεδομένων από την μη σχεσιακή βάση δεδομένων της εφαρμογής. Πριν ξεκινήσουμε, όμως, τον καθαρισμό δεδομένων για ένα πρόβλημα μηχανικής μάθησης, είναι ζωτικής σημασίας η κατανόηση των δεδομένων και ο στόχος που θέλουμε να επιτύχουμε. Χωρίς αυτήν την κατανόηση, δεν έχουμε καμία βάση για να λάβουμε αποφάσεις σχετικά με τα δεδομένα στο στάδιο καθαρισμού και προετοιμασίας τους. Όπως έχει ήδη υπογραμμιστεί, η βάση δεδομένων περιέχει ένα σύνολο πινάκων που αποτελείται από διάφορες μετρήσεις της εφαρμογής και δημογραφικά δεδομένα που ποικίλλουν μεταξύ των χρηστών. Στο στάδιο της προεπεξεργασίας των δεδομένων, τα διάφορα σύνολα δεδομένων (EQ5D3L, SPQ, Sociodemo, UCLA, brain_games, fingertapping, physical_activity, UTAUT, digital_phenotyping, mindfulness) που χρησιμοποιήθηκαν στην παρούσα εργασία, επεξεργάστηκαν επιτυχώς αναφορικά με τα διάφορα προβλήματα που εμφάνισαν.

Κατά το πρώτο στάδιο της προεπεξεργασίας, εφόσον οι αλγόριθμοι μηχανικής μάθησης δέχονται μόνο αριθμητικές εισόδους, κρίνεται απαραίτητη η κωδικοποίηση αυτών των κατηγορικών μεταβλητών σε αριθμητικές τιμές με την χρήση τεχνικών κωδικοποίησης. Στο πλαίσιο αυτό, οι κατηγορικές τιμές που εντοπίστηκαν στους διάφορους πίνακες δεδομένων, κωδικοποιήθηκαν σε αριθμούς για την καλύτερη κατανόησή τους από τους αλγορίθμους. Αναλυτικότερα, ακολουθήθηκε η τεχνική της τακτικής κωδικοποίησης (Ordinal Coding), όπως αναφέρεται και στο [41], κατά την οποία εκχωρείται ένας ακέραιος αριθμός σε κάθε κατηγορία, με την προϋπόθεση γνώσης του αριθμού των υπαρχουσών κατηγοριών, όπως για παράδειγμα η πληροφορία σχετικά με την ηλεκτρονική συσκευή του χρήστη (Tablet-> 0, Mobile-> 1).

Σημαντική λεπτομέρεια στην διαδικασία κωδικοποίησης, αποτέλεσε και η επεξεργασία της πληροφορίας σχετικά με την ομάδα στην οποία ανήκουν οι χρήστες. Συγκεκριμένα, οι χρήστες που συμμετείχαν στην μελέτη, ανήκαν σε μία περίπτωση χρήσης, ανάλογα με τη σωματική και γνωστική τους κατάσταση κατά το στάδιο της εγγραφής τους. Στις κατηγορίες UC6 και UC3 ανήκουν οι ηλικιωμένοι χωρίς σωματικά και γνωστικά προβλήματα, η κατηγορία UC5 περιλαμβάνει χρήστες με σωματικά προβλήματα, ενώ στο UC7 εμπλέκονται άτομα που είναι σε κίνδυνο απομόνωσης. Επομένως κατά την διαδικασία της κωδικοποίησης, οι κατηγορίες UC6 και UC3 αναπαραστάθηκαν με τον ίδιο ακέραιο αριθμό, χωρίς να χάνεται η αξία της πληροφορίας.

Η προεπεξεργασία των δεδομένων συνεχίστηκε με την μετονομασία χαρακτηριστικών σε διάφορους πίνακες για την διατήρηση της συνοχής των δεδομένων. Αναλυτικότερα, στον πίνακα που περιέχει την πληροφορία από το παιχνίδι fingertapping της εφαρμογής, η στήλη του μοναδικού αναγνωριστικού των χρηστών μετονομάστηκε για να έχει σε όλους τους πίνακες την ίδια ονομασία. Ακόμη για την δημιουργία του συνόλου δεδομένων και την καλύτερη κατανόηση των χαρακτηριστικών του, οι τρεις ερωτήσεις που αναφέρονται στα ερωτηματολόγια αυτοαντίληψης, μετονομάστηκαν σε SPQ1, SPQ3, SPQ5 αντίστοιχα.

Στη συνέχεια, όπως δηλώνεται και στην περιγραφή των διαθέσιμων δεδομένων [38], επιλέχθηκε στους διάφορους πίνακες δεδομένων που αντιστοιχούν σε ερωτηματολόγια που συμπλήρωσαν οι χρήστες, η μη διαθέσιμη πληροφορία να αντιπροσωπεύεται με μηδέν. Για αυτό το λόγο, αντικαταστάθηκε στον πίνακα κοινωνικοδημογραφικών πληροφοριών (Sociodemographic) η αναπαράσταση της μη διαθέσιμης πληροφορίας για το τεχνολογικό επίπεδο των χρηστών με μηδέν.

Η κατανόηση των διαθέσιμων δεδομένων, εν συνεχεία οδήγησε στην χρησιμοποίηση της πληροφορίας που παρείχαν οι χρήστες, από την συμπλήρωση των ερωτηματολογίων EQ5DL3, SPQ και UCLA, κατά το στάδιο εγγραφής τους στην εφαρμογή μόνο και όχι κατά την διάρκεια ή το τέλος της μελέτης. Ο λόγος δεν ήταν άλλος από την έλλειψη της πληροφορίας για την πλειοψηφία των χρηστών της εφαρμογής.

Επιπλέον, κατά την παρατήρηση των δεδομένων, εντοπίστηκαν και αντιμετωπίστηκαν προβλήματα στους τύπους των δεδομένων και συγκεκριμένα στα πεδία με τιμές ημερομηνίας. Στην ίδια κατεύθυνση, παρατηρήθηκε η ύπαρξη προβληματικών εγγραφών σχετικά με τις μετρήσεις της εφαρμογής είτε λόγω διάρκειας αλληλεπίδρασης του χρήστη είτε λόγω υπερκάλυψης στο ίδιο χρονικό διάστημα δύο ή τριών εγγραφών για τον ίδιο χρήστη. Για την επεξεργασία και την αντιμετώπιση του προβλήματος, δημιουργήθηκε ένας έλεγχος διάρκειας των εγγραφών, με αποτέλεσμα την απόρριψη όσων είχαν διάρκεια μικρότερη των πέντε δευτερολέπτων και ακόλουθα ένας κανόνας φιλτραρίσματος των εγγραφών σχετικά με την υπερκάλυψη τους, κατά τον οποίο διατηρούταν η εγγραφή με την μεγαλύτερη διάρκεια.

Ένα άλλο σύνολο εργασιών, που εκτελούνται στα πλαίσια της προεπεξεργασίας, αφορούν τη μείωση των διαστάσεων των δεδομένων. Υπό το πρίσμα αυτό, μπορεί να καταγράφονται διαφορετικές εκδοχές της ίδιας πληροφορίας σε διάφορα χαρακτηριστικά, με τον αναλυτή να κατασκευάζει νέα πεδία για να μπορέσει να αποδώσει καλύτερα το πραγματικό περιεχόμενο των δεδομένων. Τα νέα χαρακτηριστικά υπολογίζονται με κατάλληλες πράξεις από τα δεδομένα άλλων πεδίων. Με τον τρόπο αυτό, μπορεί να αναδειχθεί η σημαντικότητα της πληροφορίας, με τον κίνδυνο, όμως, της μη χρήσης των κατάλληλων δεδομένων που μπορεί να οδηγήσει σε τελείως εσφαλμένα συμπεράσματα.

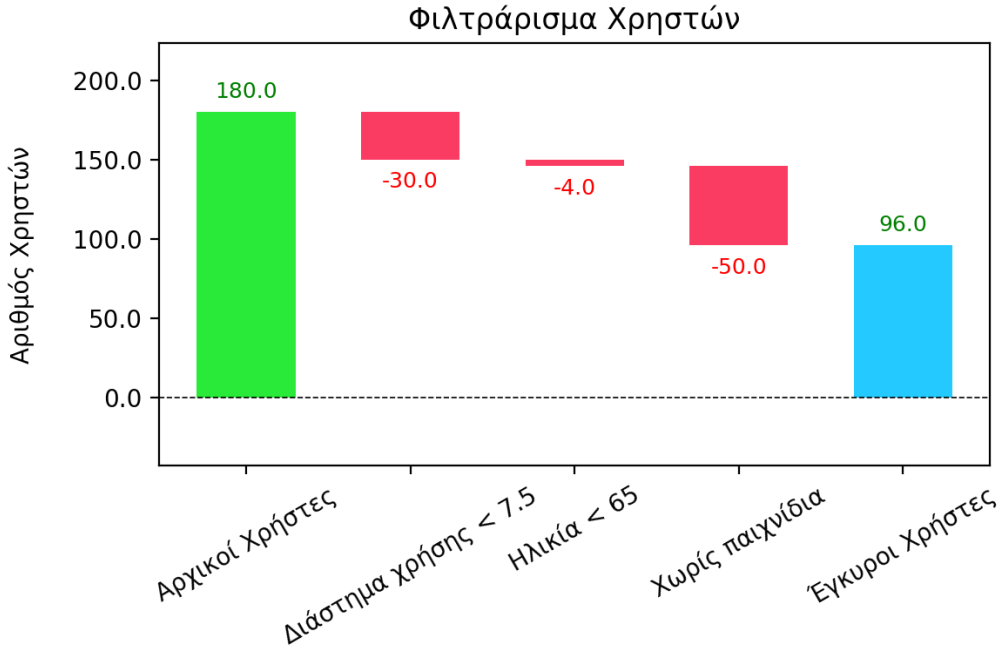
Κατά την μελέτη των διαθέσιμων δεδομένων, στο σύνολο δεδομένων του προαιρετικού ερωτηματολογίου UCLA, το οποίο αναφέρεται στην κλίμακα μοναξιάς του UCLA, παρατηρήθηκε καταγραφή παραπλήσιων εκδοχών της ίδιας πληροφορίας. Ουσιαστικά είναι ένα μέτρο 20 στοιχείων που αξιολογεί πόσο συχνά ένα άτομο αισθάνεται αποκομμένο από τους άλλους, μέσω της καταγραφής απαντήσεων σε ερωτήσεις που εκφράζουν την μοναξιά που νιώθει ο χρήστης. Για τον λόγο αυτό και με σκοπό την μείωση διαστάσεων, επιλέχθηκε η δημιουργία ενός νέου χαρακτηριστικού (level of loneliness), το οποίο χρησιμοποιεί το άθροισμα των απαντήσεων του χρήστη στο σύνολο των 20 ερωτήσεων για να κατηγοριοποιήσει τον χρήστη σε μία από τις πέντε κλάσεις που θα δημιουργηθούν για την περιγραφή του αισθήματος της μοναξιάς, με τους χρήστες που δεν απάντησαν στο ερωτηματολόγια να κατηγοριοποιούνται στην κλάση 0. Κατά αυτό τον τρόπο, όπως αποδείχτηκε και κατα την διάρκεια της πειραματικής διαδικασίας, η μείωση των διαστάσεων για την περιγραφή του συναισθήματος της μοναξιάς από είκοσι σε μία, οδήγησε στην αποφυγή υπερεκπαίδευσης των μοντέλων και την καλύτερη προβλεπτική λειτουργία τους. Ομοίως παρατηρήθηκε και με την μείωση των διαστάσεων στις μετρήσεις της εφαρμογής ανά χρήστη στις στιγμές απόκτησης δεδομένων, όπως θα περιγραφεί στην συνέχεια. Στο σημείο αυτό, για την ολοκλήρωση του πρώτου σταδίου της προεπεξεργασίας δεδομένων, απορρίφθηκαν τα χαρακτηριστικά που δεν παρέχουν πληροφορία για την διαδικασία της πρόβλεψης των πρόωρων αποχωρήσεων των χρηστών και συνεχίστηκε η επεξεργασία των δεδομένων με το φιλτράρισμα των χρηστών της εφαρμογής βάσει των προτύπων της μελέτης.

4.2 Επικύρωση Χρηστών

Το σύνολο δεδομένων που χρησιμοποιήθηκε περιείχε περισσότερους από 150 χρήστες που έχουν δοκιμάσει τον αντίκτυπο μιας ψηφιακής εφαρμογής ΑΗΑ για τη βελτίωση της ποιότητας ζωής τους για τουλάχιστον 6 μήνες στο δίκτυο ΜΑΗΑ στη Μαδρίτη. Όπως διαπιστώθηκε, όμως, η βάση δεδομένων περιείχε πολλούς χρήστες που δεν ήταν έγκυροι για την συμμετοχή τους στην έρευνα που διεξήχθη. Για την επιβεβαίωση της εγκυρότητας των χρηστών, πραγματοποιήθηκε μια σειρά ελέγχων για την ποιότητα και την επικύρωση των δεδομένων.

Η επικύρωση δεδομένων διαφέρει από τον καθαρισμό δεδομένων που περιγράφηκε προηγουμένως, καθώς η διαδικασία επικύρωσης σχεδόν πάντα συνεπάγεται με την απόρριψη δεδομένων κατά την εισαγωγή τους σε ένα σύστημα εξόρυξης γνώσης. Κατά το πρώτο στάδιο παρατήρησης των δεδομένων λοιπόν, χρησιμοποιήθηκαν οι αποθηκευμένες κοινωνικοδημογραφικές πληροφορίες που παρείχαν οι χρήστες. Συγκεκριμένα, απορρίφθηκαν αρχικά οι χρήστες με περίοδο χρήσης της εφαρμογής μικρότερης των 7.5 εβδομάδων, από την στιγμή που το διάστημα πρόβλεψης της συνέπειας των χρηστών αφορά την διενέργεια 15 στιγμών απόκτησης δεδομένων. Ακολούθως σύμφωνα με τις οδηγίες των υπεύθυνων της μελέτης, πραγματοποιήθηκε φιλτράρισμα των χρηστών ως προς την ηλικία τους. Για την έγκυρη συμμετοχή τους στην μελέτη, κάθε συμμετέχων έπρεπε να είναι άνω των 65 ετών.

Στο δεύτερο στάδιο ελέγχου, χρησιμοποιήθηκαν οι επεξεργασμένες καταγραφές αλληλεπίδρασης των χρηστών με την εφαρμογή, όπως αυτές προέκυψαν με την ολοκλήρωση της προεπεξεργασίας των αρχικών εγγραφών. Σκοπός ήταν η απόρριψη των αδρανών χρηστών, οι οποίοι δεν χρησιμοποίησαν ούτε μια φορά την εφαρμογή. Ως αποτέλεσμα του παραπάνω ελέγχου, όπως παρουσιάζεται και στην εικόνα 16, υπήρξε η σημαντική μείωση στο σύνολο των έγκυρων χρηστών της εφαρμογής βάσει των κανονισμών της μελέτης.



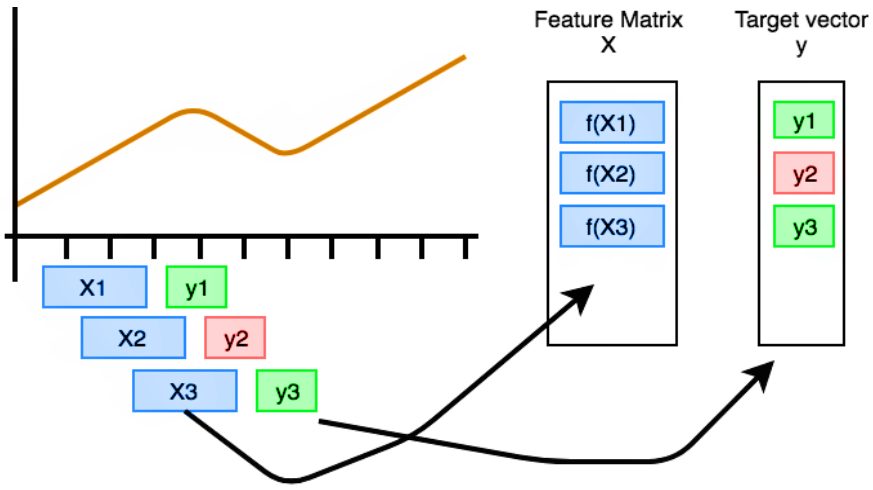
Εικόνα 16. Έλεγχος έγκυρων χρηστών

4.3 Συλλογή Δεδομένων ανά Χρήστη

Το στάδιο της προεπεξεργασίας των δεδομένων συνεχίστηκε με την συλλογή δεδομένων ανά έγκυρο χρήστη. Η απόκτηση δεδομένων ακολουθεί βάσει των κανονισμών της μελέτης ένα καλά καθορισμένο πρωτόκολλο που αποτελείται από τακτικά προγραμματισμένες στιγμές, όπου οι συμμετέχοντες πρέπει να τροφοδοτήσουν με δεδομένα τη μελέτη, με την έγκυρη συχνότητα απόκτησης δεδομένων περιοδικών δραστηριοτήτων να ορίζεται στις δύο φορές ανά εβδομάδα χρήσης της εφαρμογής για κάθε περιοδική δραστηριότητα. Επιπλέον, κάθε προγραμματισμένη λήψη δεδομένων μπορεί να περιλαμβάνει την πραγματική μέτρηση μιας ή περισσότερων περιοδικών δραστηριοτήτων (brain games, physical activity, finger tapping, mindfulness) ανεξάρτητα από την επιτυχία ή αποτυχία του συμμετέχοντα.

Η κυλιόμενη ανάλυση χρησιμοποιείται συνήθως για τον έλεγχο ενός στατιστικού μοντέλου σε ιστορικά δεδομένα, με σκοπό την αξιολόγηση της σταθερότητας και της προγνωστικής του ακρίβειας. Στο πλαίσιο λοιπόν δημιουργίας προγραμματισμένων διαδοχικών λήψεων δεδομένων με βάση το πρωτόκολλο απομακρυσμένης παρακολούθησης για χρονικές περιπτώσεις i=1,…,n για το διάστημα χρήσης της εφαρμογής από τον χρήστη, χρησιμοποιήθηκε η μέθοδος rolling window time series. Ουσιαστικά η συγκεκριμένη μέθοδος θα μπορούσε να περιγραφεί ως ένα παράθυρο που ολισθαίνει πάνω από τα δεδομένα «χρονοσειρών» και εξάγει όλα τα δεδομένα που υπάρχουν στο συγκεκριμένο παράθυρο [42].

Αυτό το συρόμενο παράθυρο είναι η βάση για το πώς μπορούμε να μετατρέψουμε οποιοδήποτε σύνολο δεδομένων χρονοσειρών σε εποπτευόμενο μαθησιακό πρόβλημα. Στην συγκεκριμένη ιδιότητα της μεθόδου, βασίστηκε και η κλάση προγραμματισμού που δημιουργήθηκε, όπως παρουσιάζεται λεπτομερώς στο Παράρτημα ΙΙ, για την δημιουργία των προφίλ των χρηστών που χρησιμοποιούν την εφαρμογή. Ουσιαστικά, με την συνεχή επαναχρησιμοποίηση των χρήσιμων συναρτήσεων που περιέχονται στην κλάση προγραμματισμού, ανά βδομάδα και ανά χρήστη, δύναται να περιγραφεί η συμπεριφορά κάθε χρήστη στο περιβάλλον της εφαρμογής και να καταστεί δυνατή η δημιουργία ενός συνόλου δεδομένων που περιέχει όλη την δραστηριότητα και πληροφορία για το σύνολο των χρηστών.



Εικόνα 17. Rolling Window Method for Time Series Analysis

Η διαδικασία συλλογής δεδομένων ξεκινάει με την συλλογή όλων των διαθέσιμων δεδομένων από την βάση δεδομένων ανά χρήστη και διαγράφεται το μοναδικό αναγνωριστικό του χρήστη, διότι δεν είναι σημαντικό για την διαδικασία της πρόβλεψης. Στην συνέχεια πραγματοποιείται έλεγχος για την διαθεσιμότητα των απαντήσεών του χρήστη στα ερωτηματολόγια EQ5DL3, SPQ και UCLA, κατά το στάδιο εγγραφής του χρήστη στην εφαρμογή. Στις περιπτώσεις που οι πληροφορίες αυτές δεν είναι διαθέσιμες, συμπληρώνονται αυτόματα τα ερωτηματολόγια για τον χρήστη με τον ακέραιο αριθμό 0 που αντιπροσωπεύει την μη διαθέσιμη πληροφορία βάσει των κανόνων της μελέτης.

Με την ολοκλήρωση της συλλογής της «στατιτικής» πληροφορίας, εφαρμόζεται η προσαρμοσμένη μέθοδος που δημιουργήθηκε με βάση την μέθοδο rolling window time series. Αρχικά συλλέγεται το καταγεγραμμένο διάστημα αλληλεπίδρασης και παρουσίας του χρήστη στην εφαρμογή. Ακολούθως το διαθέσιμο χρονικό διάστημα, διασπάται σε χρονικά «παράθυρα» που περιέχουν 15 στιγμές λήψης δεδομένων. Σε αυτό το σημείο, είναι σημαντική προς αναφορά η λεπτομέρεια και η ακρίβεια που χρησιμοποιεί η προσαρμοσμένη μέθοδος, καθώς για την δημιουργία χρονικών πλαισίων, μετατρέπεται το χρονικό διάστημα από εβδομάδες σε ώρες. Η απόκτηση δεδομένων ακολουθεί ένα καλά καθορισμένο πρωτόκολλο με την έγκυρη συχνότητα απόκτησης δεδομένων περιοδικών δραστηριοτήτων να ορίζεται για τους συμμετέχοντες στις δύο φορές ανά εβδομάδα για κάθε δραστηριότητα. Σύμφωνα λοιπόν με το πρωτόκολλο, η χρονική περίοδος μίας εβδομάδας αναγάγεται σε 168 ώρες και το χρονικό πλαίσιο των στιγμών απόκτησης δεδομένων ορίζεται στις 84 ώρες. Ακολουθώντας λοιπόν την ίδια στρατηγική, δημιουργούνται τα χρονικά «παράθυρα» 15 στιγμών απόκτησης δεδομένων, με την διαδικασία να προχωράει κάθε φορά κατά μία εβδομάδα μπροστά στον χρόνο και να ολοκληρώνεται ύστερα από τον έλεγχο μη διαθέσιμων παρατηρήσεων για το σύνολο του προσεχές διαστήματος των 7.5 εβδομάδων (1260 ώρες).

Παράλληλα με την δημιουργία των χρονικών πλαισίων, συλλέγονται οι μετρήσεις των περιοδικών δραστηριοτήτων (brain games, physical activity, finger tapping, mindfulness) του χρήστη ανεξάρτητα από την επιτυχία ή αποτυχία του. Με αυτόν τον τρόπο, γίνεται εφικτή η καταμέτρηση της δραστηριότητας του χρήστη ανά χρονική στιγμή λήψης δεδομένων. Συγκεκριμένα, η μέθοδος αναζητάει τις εγγραφές του χρήστη στο καθορισμένο χρονικό διάστημα και ελέγχει αν ακολουθείται το πρωτόκολλο έγκυρης συχνότητας. Αποτέλεσμα της μεθόδου είναι η δημιουργία παραθύρων 12 στιγμών λήψης δεδομένων, όπως ορίζεται και από το πρόβλημα που παρουσιάζει η μελέτη. Στο σημείο αυτό, για την δημιουργία προβλέψεων και αξιολόγησης των μοντέλων μηχανικής μάθησης, η προσαρμοσμένη μέθοδος εφαρμόζεται στο διάστημα των τριών τελευταίων στιγμών λήψης δεδομένων για την δημιουργία της κλάσης που θα εφαρμοστεί η πρόβλεψη και περιγράφει την συνέπεια του χρήστη στο διάστημα αυτό. Στο πλαίσιο του [37], θεωρείται ότι μια προγραμματισμένη απόκτηση έχει υλοποιηθεί αποτελεσματικά από έναν συμμετέχοντα εάν έχει ληφθεί τουλάχιστον μία από τις προγραμματισμένες μεταβλητές για μέτρηση. Η συνέπεια λοιπόν κατά τις επόμενες 3 προγραμματισμένες στιγμές απόκτησης δεδομένων, θεωρείται χαμηλή εάν ο αριθμός των ληφθέντων μετρήσεων είναι 0 ή 1 και υψηλός εάν είναι μεγαλύτερος των 2 έγκυρων μετρήσεων κατά τη διάρκεια της χρονικής περιόδου.

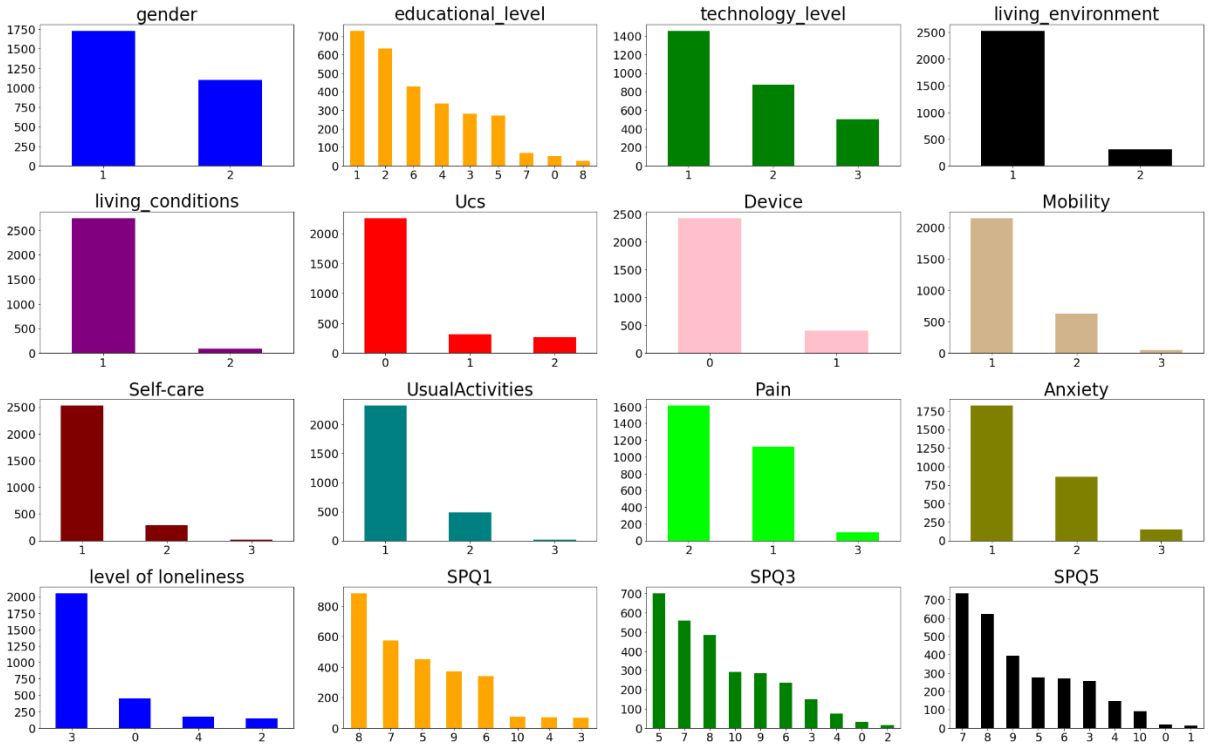
4.4 Διερευνητική Ανάλυση Δεδομένων

Με το πέρας της προεπεξεργασίας των διάφορων συνόλων δεδομένων που χρησιμοποιήθηκαν (EQ5D3L, SPQ, Sociodemo, UCLA, brain_games, fingertapping, physical_activity, mindfulness), καθώς και της επικύρωσης των έγκυρων χρηστών βάσει των προδιαγραφών της μελέτης, δημιουργήθηκε με την συλλογή δεδομένων ανά χρήστη, το σύνολο δεδομένων που θα χρησιμοποιηθεί για την εφαρμογή προβλεπτικών μοντέλων. Στο πλαίσιο αυτό, πραγματοποιήθηκε και η διερευνητική ανάλυση (Exploratory Data Analysis - EDA) του συνόλου δεδομένων, όπως αυτό παρουσιάζεται στον Πίνακα 1. Η διερευνητική ανάλυση παρουσιάζει χρήσιμες πληροφορίες για το σύνολο των δεδομένων με στόχο την καλύτερη δυνατή αξιοποίηση αυτών, για την επίλυση του προβλήματος των πρόωρων εγκαταλείψεων χρήσης μιας εφαρμογής [43].

Πίνακας 1. Περιγραφή Συνόλου Δεδομένων

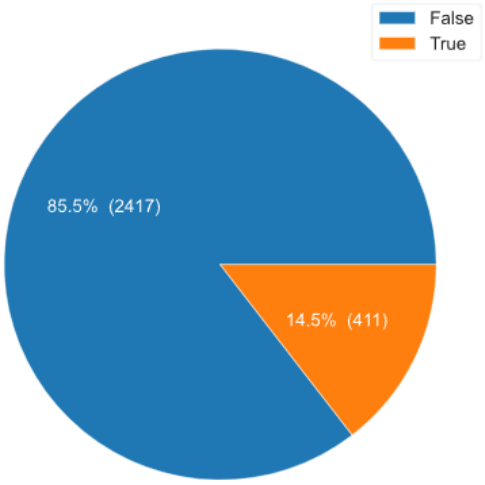
Αριθμός Γνωρίσματος	Όνομα Γνωρίσματος	Τιμές Γνωρίσματος
1	gender	1, 2
2	educational_level	0, 1, 2, 3, 4, 5, 6, 7, 8
3	technology_level	1, 2, 3
4	living_environment	1, 2
5	living_conditions	1, 2
6	Ucs	0, 1, 2
7	Device	0, 1
8	Mobility	1, 2, 3
9	Self-care	1, 2, 3
10	UsualActivities	1, 2, 3
11	Pain	1, 2, 3
12	Anxiety	1, 2, 3
13	level of loneliness	0, 2, 3, 4
14	SPQ1	3, 4, 5, 6, 7, 8, 9, 10
15	SPQ3	0, 2, 3, 4, 5, 6, 7, 8, 9, 10
16	SPQ5	0, 1, 3, 4, 5, 6, 7, 8, 9, 10
17	T-11	0, 1, 2, 3, 4
18	T-10	0, 1, 2, 3, 4
19	T-9	0, 1, 2, 3, 4
20	T-8	0, 1, 2, 3, 4
21	T-7	0, 1, 2, 3, 4
22	T-6	0, 1, 2, 3, 4
23	T-5	0, 1, 2, 3, 4
24	T-4	0, 1, 2, 3, 4
25	T-3	0, 1, 2, 3, 4
26	T-2	0, 1, 2, 3, 4
27	T-1	0, 1, 2, 3, 4
28	T-0	0, 1, 2, 3, 4
29	class	False, True

Αρχικά μελετήθηκε η κατανομή των χαρακτηριστικών που προκύπτουν από τα ερωτηματολόγια που έχουν συμπληρώσει οι χρήστες της εφαρμογής και το πέρας του σταδίου της προεπεξεργασίας των δεδομένων.



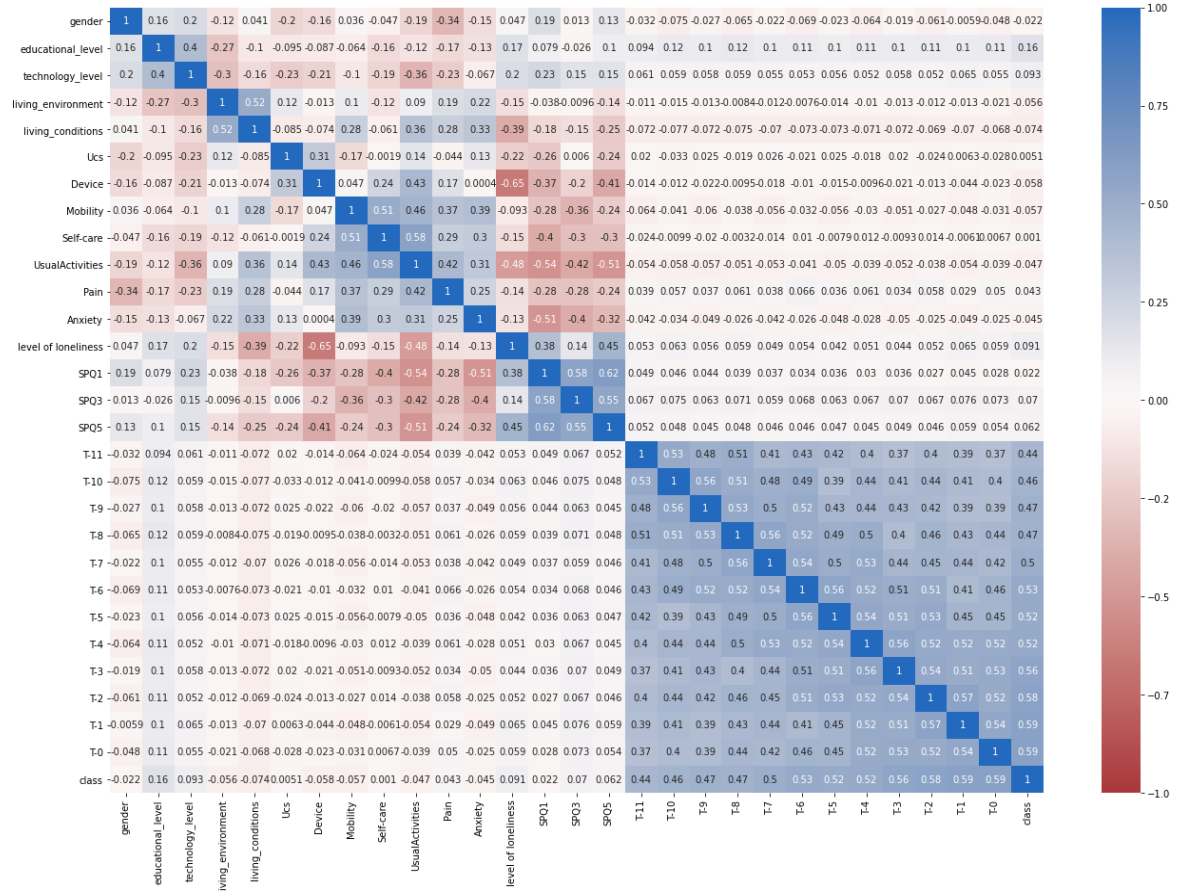
Εικόνα 18. Κατανομή χαρακτηριστικών συνόλου δεδομένων

Ένα από τα χαρακτηριστικά που παρουσιάζονται στο σύνολο δεδομένων, είναι η τάξη που αναφέρεται στην μελλοντική συνέπεια ή ασυνέπεια των χρηστών της εφαρμογής παρακολούθησης υγείας. Σε αυτό το σημείο της ανάλυσης παρουσιάστηκε και το πρόβλημα μη ισοκατανεμημένων δεδομένων, το οποίο αντιμετωπίστηκε κατά την πειραματική διαδικασία με διάφορες προσεγγίσεις, με σκοπό την εύρυθμη λειτουργία και καλύτερη απόδοση των αλγορίθμων ταξινόμησης.



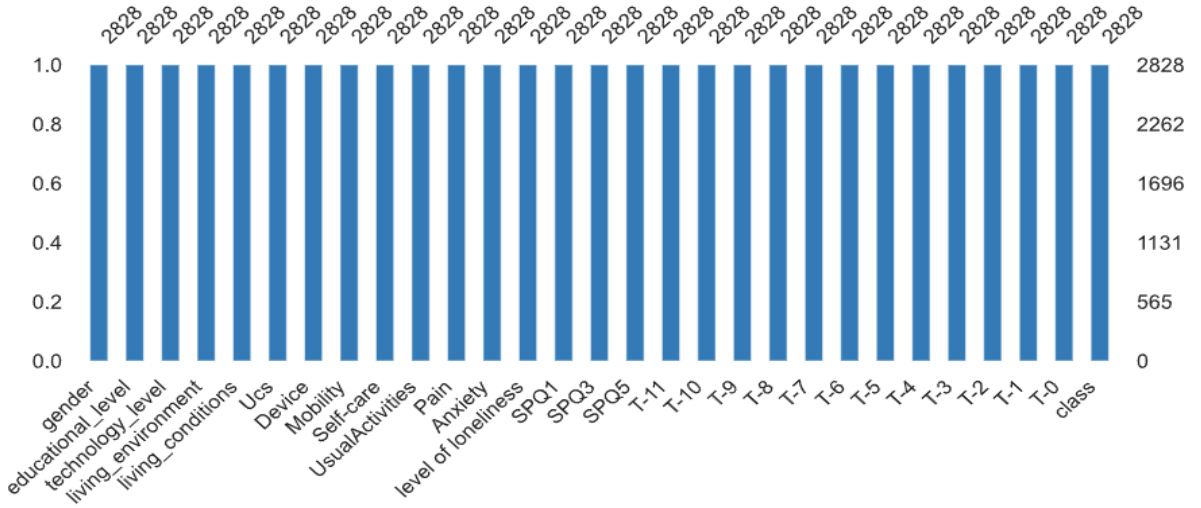
Εικόνα 19. Κλάση Μελλοντικής Συνέπειας

Ακολούθως, δημιουργήθηκε και μελετήθηκε ο πίνακας συσχετίσεων για όλα τα χαρακτηριστικά, αποσκοπώντας στην καλύτερη κατανόηση του συνόλου δεδομένων. Επιπλέον, βάσει των συσχετίσεων των χαρακτηριστικών με την κλάση πρόβλεψης, στο στάδιο της πειραματικής διαδικασίας πραγματοποιήθηκε πληθώρα διαφορετικών προσεγγίσεων για την πρόβλεψη της συνέπειας των χρηστών.



Εικόνα 20. Πίνακας Συσχετίσεων

Σαν τελευταίο στάδιο της διερευνητικής ανάλυσης του συνόλου δεδομένων, έγινε έλεγχος για την ύπαρξη κενών τιμών, με την εικόνα 21 να επιβεβαιώνει την πληρότητα του συνόλου δεδομένων που δημιουργήθηκε.



Εικόνα 21. Έλεγχος κενών τιμών

4.5 Μείωση Διαστάσεων & Επιλογή Χαρακτηριστικών

Στην παγκόσμια βιβλιογραφία, οι στήλες ορίζονται τις περισσότερες φορές ως χαρακτηριστικά (features), διαστάσεις (dimensions) ή γνωρίσματα (attributes) και σύμφωνα με την επιστήμη των δεδομένων, το πλήθος των στηλών αποτελεί μια από τις σημαντικότερες ιδιότητες των δεδομένων. Όπως έχει υπογραμμιστεί και σε προηγούμενη ενότητα, η ύπαρξη πολλών χαρακτηριστικών ισοδυναμεί με σημαντικό πρόβλημα για τη διαδικασία εξόρυξης προτύπων. Αναλυτικότερα πολλές φορές μπορεί είτε να υπάρχουν γνωρίσματα που περιέχουν άσχετη πληροφορία, είτε γνωρίσματα με υψηλό βαθμό συσχέτισης, των οποίων η ταυτόχρονη παρουσία στο σύνολο δεδομένων είναι ανούσια. Εύλογα, όμως, σε αυτό το σημείο μπορεί να προκύψει η απορία γιατί όσα αναφέρθηκαν αποτελούν πρόβλημα. Τα διαθέσιμα δεδομένα περιλαμβάνουν πολλά χαρακτηριστικά, ωστόσο για μια συγκεκριμένη εργασία εξόρυξης γνώσης, δεν είναι χρήσιμα όλα αυτά τα χαρακτηριστικά, καθώς δεν περιέχουν πληροφορίες σχετικές με το αντικείμενο της ανάλυσης. Ακόμη ορισμένες στατικές μέθοδοι ανάλυσης, υποθέτουν ότι στα μοντέλα περιλαμβάνονται μόνο οι σημαντικές διαστάσεις με το περιεχόμενο των συγκεκριμένων στηλών να είναι ασυσχέτιστο. Υπό αυτό το πρίσμα, η ύπαρξη πολλών χαρακτηριστικών οδηγεί στην αύξηση της πολυπλοκότητας του προβλήματος και παράλληλα στην καθυστέρηση της εκπαίδευσης των μοντέλων.

Μια βάση δεδομένων τυπικά περιέχει εκατοντάδες γνωρίσματα και συχνά τίθεται ζήτημα για την επιλογή υποσυνόλου των συγκεκριμένων χαρακτηριστικών. Η απαλοιφή κάποιων γνωρισμάτων σε ορισμένες περιπτώσεις μπορεί να είναι εύκολη και προφανής, όπως για παράδειγμα στην παρούσα εργασία η μη χρησιμοποίηση του συνόλου δεδομένων «digital_phenotyping», διότι η πληροφορία του είναι διαμοιρασμένη στα υπόλοιπα σύνολα δεδομένων που αντιστοιχούν στις μετρήσεις της εφαρμογής. Ωστόσο υπάρχουν προβλήματα, στα οποία η επιλογή γνωρισμάτων δεν είναι καθόλου προφανής, καθώς η συμπεριφορά των δεδομένων δεν είναι εκ των προτέρων γνωστή και απουσιάζει η επαρκής προγενέστερη γνώση. Επομένως, κατά αυτό τον τρόπο τίθεται ζήτημα χρήσης μεθόδων για την επιλογή χαρακτηριστικών και τον περιορισμό των διαστάσεων. Τα δεδομένα μεγάλου όγκου και η αύξηση της πολυπλοκότητας προκαλούν αύξηση του υπολογιστικού κόστους και μεγάλη καθυστέρηση στη διεξαγωγή των αναλύσεων, η οποία όμως δεν επιφέρει τις περισσότερες φορές βελτίωση των αποτελεσμάτων. Στην παγκόσμια επιστημονική κοινότητα, το πρόβλημα αυτό είναι γνωστό ως «κατάρα της διαστατικότητας» (curse of dimensionality) [44].

Η μείωση διαστάσεων αποτελεί σημαντικό στάδιο στην αντιμετώπιση προβλημάτων απόδοσης των αλγορίθμων μηχανικής μάθησης, εφόσον ορίζεται ως ο μετασχηματισμός δεδομένων υψηλών διαστάσεων σε μια ουσιαστική αναπαράσταση μειωμένων διαστάσεων. Η μείωση των διαστάσεων μπορεί να επέλθει με την προβολή των δεδομένων σε ένα διαφορετικό χώρο, αλλά και με την επιλογή χαρακτηριστικών λιγότερων διαστάσεων. Η έννοια, όμως, της μείωσης των διαστάσεων (dimensionality reduction) διαφέρει με την έννοια της επιλογής σημαντικών χαρακτηριστικών (feature selection). Συγκεκριμένα, η επιλογή χαρακτηριστικών περιλαμβάνεται στην ευρύτερη έννοια της μείωσης των διαστάσεων.

Τα αποτελέσματα της ανάλυσης των μειωμένων δεδομένων πρέπει να είναι τα ίδια ή κοντινά με τα αποτελέσματα της ανάλυσης του συνόλου των δεδομένων, επομένως η μείωση του όγκου δεδομένων δεν είναι σε καμία περίπτωση μια τετριμμένη εργασία. Σε αυτό το σημείο λοιπόν, η επιλογή χαρακτηριστικών συνίσταται στην επιλογή ενός υποσυνόλου Κ χαρακτηριστικών από ένα αρχικό σύνολο Λ χαρακτηριστικών, με το επιλεγμένο υποσύνολο να είναι το πλέον κατάλληλο για την εξόρυξη προτύπων και την παράλληλη διατήρηση ουσιαστικής πληροφορίας σχετικά με τη διασπορά και τη συμπεριφορά των δεδομένων. Ο χώρος αυτός είναι διαφορετικών διαστάσεων από τον αρχικό, με τις νέες διαστάσεις του, όμως, να έχουν καθοριστεί με σκοπό την διατήρηση της ουσιαστική πληροφορίας για τη συμπεριφορά των δεδομένων. Επίκεντρο της επιλογής χαρακτηριστικών αποτελεί η επιλογή αντιπροσωπευτικών λειτουργιών του συνόλου δεδομένων, με τον αποκλεισμό περιττών και άσχετων δεδομένων και κρίνεται αναγκαία η συνεισφορά της σε περιπτώσεις εξαγωγής κανόνων με νόημα. Στόχος της είναι η βελτίωση της απόδοσης της κατηγοριοποίησης, ο ευκολότερος υπολογισμός της και η διασφάλιση ως προς το μοντέλο μηχανικής μάθησης ότι χρησιμοποιεί το πιο ουσιαστικό σύνολο δεδομένων. Επιπλέον παίζει καθοριστικό ρόλο στην διαδικασία απλοποίησης των μοντέλων μάθησης, βοηθώντας τον χρήστη στην καλύτερη και ευκολότερη ερμηνεία του μοντέλου και των αποτελεσμάτων του. Παράλληλα με την χρήση μόνο του σχετικού υποσυνόλου δεδομένων μειώνεται ο χρόνος επεξεργασίας των δεδομένων κάτι το οποίο συνεπάγεται με μικρότερο χρόνο εκπαίδευσης για το μοντέλο μηχανικής μάθησης [45].

Η επιλογή χαρακτηριστικών, δύναται να πραγματοποιηθεί με την εφαρμογή πολλών και διαφορετικών μεθόδων. Για τις μεθόδους τύπου διήθησης (filter), βάση αποτελούν τα χαρακτηριστικά των δεδομένων και χρησιμοποιούν μεθόδους διαφορετικές από τους αλγορίθμους που θα εφαρμοστούν στην διαδικασία κατηγοριοποίησης των δεδομένων. Χαρακτηριστικό τους είναι η ταχύτητα, η εφαρμογή τους με ποικίλους αλγορίθμους και ότι δεν λαμβάνουν υπόψιν τις σχέσεις μεταξύ των μεταβλητών. Από την άλλη πλευρά, οι μέθοδοι τύπου ενσωμάτωσης (wrapper) χρησιμοποιούν για την αξιολόγηση των διαφορετικών υποσυνόλων χαρακτηριστικών τον ίδιο αλγόριθμο που θα εφαρμοστεί στην διαδικασία κατηγοριοποίησης. Επομένως με την συγκεκριμένη υλοποίηση επιτυγχάνονται καλύτερα αποτελέσματα, διότι τα συγκεκριμένα υποσύνολα χαρακτηριστικών είναι προσαρμοσμένα στις μεθόδους που θα χρησιμοποιηθούν για την τελική ανάλυση. Σημαντικό μειονέκτημα τους είναι η ταχύτητα, με τις μεθόδους τύπου filter να αποδεικνύονται ιδανικές για μεγάλα σύνολα δεδομένων, ενώ οι μέθοδοι τύπου wrapper για την εφαρμογή τους σε μικρά σύνολα δεδομένων [46].

Στην παρούσα έρευνα, κατόπιν πειραματικού ελέγχου που διενεργήθηκε, πραγματοποιήθηκε επιλογή χαρακτηριστικών με την χρήση διαφορετικών μεθόδων τύπου διήθησης. Συγκεκριμένα εφαρμόστηκαν οι στατιστικοί έλεγχοι που παρουσιάζονται παρακάτω και βάσει των στατιστικών δεικτών, πραγματοποιήθηκε και η επιλογή του υποσυνόλου χαρακτηριστικών. Το επιλεγμένο υποσύνολο δεδομένων, σε σύγκριση με την χρήση ολόκληρου του συνόλου δεδομένων οδήγησε σε αξιοσημείωτες αποκλίσεις στην ακρίβεια των αλγορίθμων. Παράλληλα επιτεύχθηκε μείωση του θορύβου στο σύνολο δεδομένων. Άλλωστε, λιγότερα περιττά δεδομένα σημαίνει λιγότερες πιθανότητες για λήψη αποφάσεων με βάση το θόρυβο, μείωση του χρόνου εκπαίδευσης και αποφυγή φαινομένων υπερεκπαίδευσης των μοντέλων μηχανικής μάθησης.

4.5.1 Στατιστικός Έλεγχος Chi-squared

Στην συγκεκριμένη υλοποίηση πραγματοποιήθηκε επιλογή χαρακτηριστικών με την χρήση του στατιστικού δείκτη Chi-squared. Αναλυτικότερα, μεταξύ των μέτρων εξάρτησης, υπάρχει μια δοκιμή Chi-squared, η οποία αξιολογεί την αξία ενός χαρακτηριστικού υπολογίζοντας την τιμή του στατιστικού δείκτη σε σχέση με την εξαρτημένη μεταβλητή. Ουσιαστικά, μετριέται η ικανότητα πρόβλεψης της τιμής ενός χαρακτηριστικού από την τιμή της εξαρτημένης μεταβλητής, μέσω του ελέγχου ανεξαρτησίας των χαρακτηριστικών [47]. Ο στατιστικός δείκτης Chi-squared προκύπτει από την σχέση:

$$X_C^2 = \sum \frac{(O-E)^2}{E}$$

Στην παραπάνω σχέση, ως X_C^2 ορίζεται ο στατιστικός έλεγχος, ως E οι αναμενόμενες μετρήσεις και ως O οι παρατηρούμενες μετρήσεις [48]. Με τον υπολογισμό του στατιστικού δείκτη για όλα τα χαρακτηριστικά, πραγματοποιήθηκε η επιλογή των δεκατεσσάρων πρώτων χαρακτηριστικών βάσει αυτού, όπως αυτά παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 2. Στατιστικός Έλεγχος Chi-squared

Κατάταξη Γνωρίσματος	Όνομα Γνωρίσματος	P-Value
1	T-0	4.463518e-277
2	T-1	1.270621e-273
3	T-2	5.714030e-268
4	T-3	2.986645e-245
5	T-4	9.432530e-218
6	T-6	3.871255e-217
7	T-5	1.962693e-212
8	T-7	8.384336e-196
9	T-9	5.969369e-173
10	T-8	8.384336e-196
11	T-10	4.463518e-277
12	T-11	1.270621e-273
13	educational_level	4.098910e-21
14	level of loneliness	4.808311e-04
15	SPQ3	3.103145e-03
16	technology_level	3.356497e-03
17	Device	4.132446e-03
18	SPQ5	1.117858e-02
19	Mobility	1.953092e-01
20	Anxiety	2.322798e-01
21	Pain	3.321059e-01
22	UsualActivities	3.551005e-01
23	living_environment	3.837095e-01
24	SPQ1	4.985621e-01
25	living_conditions	5.015815e-01
26	gender	6.227800e-01
27	Ucs	7.529325e-01
28	Self-care	9.859490e-01

4.5.2 Συντελεστής συσχέτισης Spearman

Ο συντελεστής συσχέτισης Spearman είναι ένα μη παραμετρικό μέτρο της μονοτονικότητας της σχέσης μεταξύ δύο συνόλων δεδομένων. Σε αντίθεση με τη συσχέτιση Pearson, η συσχέτιση Spearman δεν προϋποθέτει ότι και τα δύο σύνολα δεδομένων είναι κανονικά κατανεμημένα. Όπως και άλλοι συντελεστές συσχέτισης, αυτός κυμαίνεται μεταξύ -1 και +1 με 0 όπου δεν συνεπάγεται συσχέτιση. Οι συσχετίσεις -1 ή +1 υποδηλώνουν μια ακριβή μονοτονική σχέση με τους θετικούς συσχετισμούς να υπονοούν ότι καθώς το x αυξάνεται, το ίδιο ισχύει και για το y. Οι αρνητικοί συσχετισμοί υπονοούν ότι καθώς το x αυξάνεται, το y μειώνεται. Ο συντελεστής Spearman, όπως κάθε συντελεστής συσχέτισης, είναι κατάλληλος και για συνεχείς και για διακριτές μεταβλητές, συμπεριλαμβανομένων των τακτικών διακριτών μεταβλητών [49]. Με τον υπολογισμό του συντελεστή συσχέτισης Spearman για όλα τα χαρακτηριστικά, πραγματοποιήθηκε μια πειραματική διαδικασία με την απαλοιφή των χαρακτηριστικών με απόλυτη συσχέτιση μικρότερη του 0.06 με την κλάση πρόβλεψης, όπως αυτά παρουσιάζονται στον παρακάτω πίνακα.

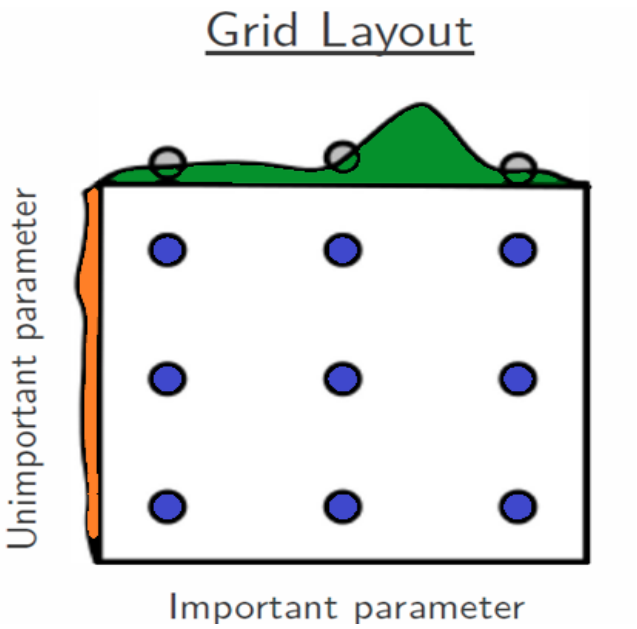
Πίνακας 3. Συντελεστής συσχέτισης Spearman

Κατάταξη Γνωρίσματος	Όνομα Γνωρίσματος	Correlation
1	T-0	0.591417
2	T-1	0.590678
3	T-2	0.577596
4	T-3	0.556381
5	T-6	0.527788
6	T-4	0.524424
7	T-5	0.519706
8	T-7	0.498493
9	T-9	0.468905
10	T-8	0.468252
11	T-10	0.455565
12	T-11	0.436491
13	educational_level	0.158441
14	technology_level	0.093481
15	level of loneliness	0.090558
16	SPQ5	0.069838
17	SPQ3	0.062411
18	Pain	0.042661
19	SPQ1	0.021753
20	gender	0.005135
21	Ucs	0.001041
22	Self-care	-0.022362
23	Anxiety	-0.045170
24	UsualActivities	-0.046945
25	living_environment	-0.055653
26	Mobility	-0.056890
27	Device	-0.058310
28	living_conditions	-0.073901

4.6 Υπερπαράμετροι Προγνωστικών Μοντέλων

Υπερπαράμετροι ορίζονται οι παράμετροι ανώτατου επιπέδου των οποίων οι τιμές ελέγχουν τη διαδικασία εκμάθησης και καθορίζουν τις τιμές των παραμέτρων του μοντέλου που καταλήγει να μαθαίνει ένας αλγόριθμος μηχανικής μάθησης. Τα μοντέλα μηχανικής μάθησης δεν είναι αρκετά έξυπνα για να γνωρίζουν ποιες υπερπαράμετροι θα οδηγούσαν στην υψηλότερη δυνατή ακρίβεια στο δεδομένο σύνολο δεδομένων. Κατά τον σχεδιασμό ενός μοντέλου, πριν την έναρξη της εκπαίδευσης του, πραγματοποιείται η επιλογή και η τιμή των υπερπαραμέτρων που θα χρησιμοποιηθούν. Για αυτό το λόγο αναφέρονται στην βιβλιογραφία ως εξωτερικοί παράμετροι, εφόσον το μοντέλο δεν μπορεί να αλλάξει τις τιμές τους κατά τη διάρκεια της εκπαίδευσης. Με την τιμή τους να μην μπορεί να εκτιμηθεί από τα δεδομένα, δεν είναι εφικτή η γνώση της καλύτερης δυνατής τιμής για μια υπερπαράμετρο μοντέλου σε ένα δεδομένο πρόβλημα. Σίγουρα, όμως, αυτό το σύνολο τιμών επηρεάζει την απόδοση, τη σταθερότητα και την ερμηνεία ενός μοντέλου. Σε αυτό το σημείο λοιπόν, η χρήση εμπειρικών κανόνων, η αντιγραφή τιμών που χρησιμοποιούνται σε παρόμοια προβλήματα ή η αναζήτηση της καλύτερης τιμής με την διαδικασία δοκιμής και σφάλματος, μπορούν να βοηθήσουν στον καθορισμό των υπερπαραμέτρων ενός μοντέλου [50].

Κατά την πειραματική διαδικασία στην παρούσα εργασία, επιλέχθηκε η επιλογή αναζήτησης των τιμών των υπερπαραμέτρων με την διαδικασία αναζήτησης πλέγματος και την χρήση ενός υποσυνόλου των δεδομένων για ταχύτερη απόκριση του αλγορίθμου. Η αναζήτηση πλέγματος συνδυάζει μια επιλογή υπερπαραμέτρων που καθορίζονται από τον χρήστη. Ο αλγόριθμος ξεκινά την εκμάθηση για καθεμία από τις διαμορφώσεις υπερπαραμέτρων για να αξιολογήσει την απόδοση του μοντέλου. Πλεονέκτημα του είναι η απλότητα, καθώς θα περάσει από όλους τους προγραμματισμένους συνδυασμούς, αλλά το μεγαλύτερο μειονέκτημα του είναι η μεγάλη υπολογιστική δύναμη που χρειάζεται για να ολοκληρωθεί η παράλληλη εκπαίδευση πολλών μοντέλων, εφόσον ο αριθμός των διαμορφώσεων που πρέπει να δοκιμάσει είναι εκθετικός σε σχέση με τον αριθμό των υπερπαραμέτρων που πρέπει να επιλέξει.



Εικόνα 22. Επιλογή Υπερπαραμέτρων: Αναζήτηση Πλέγματος

Κάθε αλγόριθμος απαιτεί ένα συγκεκριμένο πλέγμα υπερπαραμέτρων που μπορεί να προσαρμοστεί ανάλογα με το πρόβλημα που καλείται να αντιμετωπίσει. Μερικά παραδείγματα υπερπαραμέτρων μοντέλου περιλαμβάνουν:

- Το επιθυμητό βάθος κάθε δέντρου στον αλγόριθμο τυχαίων δασών.
- Την συνάρτηση ενεργοποίησης σε ένα στρώμα νευρικού δικτύου.
- Το ποσοστό εκμάθησης για την εκπαίδευση ενός νευρωνικού δικτύου.
- Τον αριθμό κρυφών επιπέδων σε ένα νευρωνικό δίκτυο.
- Τις υπερπαραμέτρους C και sigma για τις μηχανές διανυσμάτων υποστήριξης.
- Τον αριθμό k στον αλγόριθμο k-πλησιέστερων γειτόνων.

Τα αποτελέσματα της επιλογής υπερπαραμέτρων, όπως παρουσιάζονται αναλυτικά στον Πίνακα 2, βελτίωσαν την επίδοση των μοντέλων που διερευνήθηκαν για τον εντοπισμό χαμηλών μελλοντικών επιπέδων συμμόρφωσης των χρηστών της εφαρμογής.

Πίνακας 4. Υπερπαράμετροι Μοντέλων

Μοντέλο Μηχανικής Μάθησης	Υπερπαράμετροι
KNN	algorithm='ball_tree', leaf_size=9 metric='minkowski', n_neighbors=5
SVC	C=2, gamma='scale', kernel = 'rbf'
GNB	var_smoothing = 1e-11
XGB	verbosity =0, booster ="dart"
AdaBoost	n_estimators=32
LogisticRegression	C=0.1, penalty="l2" solver="newton-cg"
MLP	hidden_layer_sizes = (50,100,20) early_stopping = True, alpha=0.097
LightGBM	objective="binary" n_estimators=32, max_depth=8 num_leaves=64, learning_rate= 0.2
Random Forest	bootstrap=False, criterion='entropy' class_weight = {0:03, 1:07} max_depth=15, max_features='sqrt' min_samples_leaf=2, oob_score=False min_samples_split=3, n_estimators=100

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ & ΠΡΟΤΕΙΝΟΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ

5.1 Πειραματική Διαδικασία

Στόχος της παρούσας μεταπτυχιακής διπλωματικής εργασίας είναι να δημιουργηθεί ένα προβλεπτικό μοντέλο για την ιατρική εφαρμογή παρακολούθησης υγείας ηλικιωμένων ατόμων με σκοπό τον εντοπισμό προφίλ χρηστών που αποκλίνουν από το πλαίσιο δραστηριότητας και είναι πιθανό να εγκαταλείψουν την χρήση της. Στο κεφάλαιο αυτό, αρχικά πραγματοποιείται σύγκριση των αλγορίθμων μηχανικής μάθησης που μελετήθηκαν. Στην συνέχεια πραγματοποιείται έλεγχος για πιθανή υπερεκπαίδευσης των μοντέλων με βάση το μέγεθος του συνόλου δεδομένων για την καλύτερη πειραματική διαδικασία. Τέλος παρουσιάζεται ο αλγόριθμος που δημιουργήθηκε στα πλαίσια της έρευνας με τις επιδόσεις του ανά πειραματικό σενάριο για τον εντοπισμό πρόωρων εγκαταλείψεων χρηστών της ιατρικής εφαρμογής.

Στο σημείο αυτό, όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, λόγω της μη ισορροπημένης κατανομής δεδομένων μεταξύ συνεπών και ασυνεπών εγγραφών που εντοπίστηκε στο στάδιο διερευνητικής ανάλυσης του συνόλου δεδομένων, επιλέχθηκε για την αξιολόγηση των αλγορίθμων η μέθοδος stratified cross validation. Ο λόγος δεν είναι άλλος, από την ιδιότητα της μεθόδου να χωρίζει τα δεδομένα τυχαία σε κομμάτια, ανάλογα με τις δίπλες (folds) που ορίζει ο χρήστης, διατηρώντας παράλληλα την μη ισορροπημένη κατανομή κλάσεων σε κάθε fold. Με αυτό τον τρόπο θα πραγματοποιηθεί και με τον δυνατότερο αμερόληπτο τρόπο η αξιολόγηση των αλγορίθμων ταξινόμησης. Ουσιαστικά όπως παρουσιάζεται και στην [51], η στρωματοποιημένη δεκαπλάσια διασταυρούμενη επικύρωση (10-fold stratified cross-validation) διασφαλίζει ότι η αναλογία των «θετικών» προς των «αρνητικών» παραδειγμάτων που βρέθηκαν στην αρχική κατανομή τηρείται σε όλες τις πτυχές. Με το πέρας των 10 εκπαιδεύσεων, χρησιμοποιούνται τα αποτελέσματα για τον υπολογισμό ενός μέσου όρου κάθε μετρικής ταξινόμησης σε κάθε μοντέλο ξεχωριστά.

Οι αλγόριθμοι οι οποίοι μελετήθηκαν, παρουσιάστηκαν αναλυτικά στο 2ο κεφάλαιο της παρούσας εργασίας και επιγραμματικά είναι οι Random Forest, Multilayer Perceptron, SVM, Logistic Regression, Naïve Bayes, KNN, AdaBoost, XGBoost, LightGBM καθώς και ο Stacked Ensemble learning algorithm που δημιουργήθηκε στα πλαίσια της εργασίας. Στα παραπάνω μοντέλα με την χρήση του μηχανισμού αναζήτησης πλέγματος (Grid Search), επιλέχθηκαν οι υπερπαράμετροι των μοντέλων για την μελέτη τους στην πειραματική διαδικασία. Στόχος ήταν να μειωθεί η πιθανότητα εμφάνισης φαινομένων υπερεκπαίδευσης των μοντέλων. Για τον ίδιο λόγο, άλλωστε, επιλέχθηκε και η διενέργεια πειραματικών σεναρίων με την επιλογή χαρακτηριστικών του συνόλου δεδομένων βάση στατιστικών ελέγχων συσχέτισης μεταβλητών. Τα αποτελέσματα έδειξαν ότι εκτός από την αποφυγή υπερεκπαίδευσης των αλγορίθμων επιτεύχθηκαν και καλύτερες επιδόσεις των μοντέλων μηχανικής μάθησης στο σύνολο των μετρικών αξιολόγησης.

Τέλος, αναφορικά με την αντιμετώπιση της μη ισορροπημένης κατανομής δεδομένων στην κλάση πρόβλεψης που εφαρμόστηκαν τα μοντέλα, επιλέχθηκαν και παρουσιάζονται οι τεχνικές Smote και Neighborhood Cleaning Rule με τα αποτελέσματα τους να κρίνονται σημαντικά στην βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης.

5.1.1 1ο Σενάριο: Χρήση ολόκληρου του συνόλου δεδομένων

Κατά την πρώτη πειραματική διαδικασία, χρησιμοποιήθηκε το σύνολο των χαρακτηριστικών του συνόλου δεδομένων με την μέση τιμή των μετρήσεων των αλγορίθμων ταξινόμησης ανά μετρική ταξινόμησης, να παρουσιάζονται στον παρακάτω πίνακα. Ακόλουθα παρουσιάζεται από το γράφημα της εικόνας 23 η διακύμανση των μετρήσεων για την καλύτερη απεικόνιση και ερμηνεία των παραγόμενων αποτελεσμάτων.

Πίνακας 5. Αποτελέσματα Αλγορίθμων στο σύνολο των χαρακτηριστικών

ALL FEATURES OF THE DATASET					
MODEL	Precision	Recall	F1 Score	Accuracy	Challenge Metric
Stacking Model	92.02%	92.05%	91.16%	92.05%	73.35%
Random Forest	92.83%	92.42%	92.28%	92.42%	78.92%
KNN	91.14%	91.49%	91.24%	91.49%	63.62%
SVC	92.57%	92.62%	92.58%	92.62%	79.31%
GNB	91.47%	83.45%	85.49%	83.45%	87.26%
Logistic Regression	91.94%	92.28%	92.04%	92.28%	78.45%
XGB	91.97%	91.89%	91.91%	91.89%	70.19%
Adaboost	91.46%	91.74%	91.53%	91.74%	75.66%
LGBM	91.84%	91.94%	91.86%	91.94%	72.81%
MLP	92.12%	92.21%	92.06%	92.21%	81.15%

Εικόνα 23. Σύγκριση Αλγορίθμων ταξινόμησης στο σύνολο των χαρακτηριστικών

5.1.2 2ο Σενάριο: Επιλογή Χαρακτηριστικών βάσει συσχέτισης Spearman

Κατά την δεύτερη πειραματική διαδικασία, πραγματοποιήθηκε επιλογή χαρακτηριστικών του συνόλου δεδομένων βάσει της συσχέτισης κατά Spearman. Συγκεκριμένα επιλέχθηκαν 18 από τα 28 χαρακτηριστικά, τα οποία εμφάνιζαν είτε θετική είτε αρνητική συσχέτιση με την κλάση πρόβλεψης μεγαλύτερη του 0.06 με τον μέσο όρο των μετρήσεων των αλγορίθμων ταξινόμησης ανά μετρική ταξινόμησης, να παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 6. Αποτελέσματα Αλγορίθμων - Spearman Correlation

18 SELECTED FEATURES BASED ON SPEARMAN CORRELATION					
MODEL	Recall	Precision	F1 Score	Accuracy	Challenge Metric
Stacking Model	92.25%	92.30%	92.26%	92.25%	75.39%
Random Forest	92.92%	93.46%	93.12%	92.92%	82.75%
KNN	92.11%	91.83%	91.91%	92.11%	68.38%
SVC	92.68%	92.63%	92.63%	92.68%	80.41%
GNB	83.68%	91.53%	85.67%	83.68%	88.07%
Logistic Regression	92.19%	91.91%	92.00%	92.19%	78.42%
XGB	92.21%	92.23%	92.20%	92.21%	72.56%
Adaboost	91.66%	91.35%	91.44%	91.66%	77.72%
LGBM	92.19%	92.20%	92.17%	92.19%	75.08%
MLP	92.29%	92.57%	92.38%	92.29%	81.10%

5.1.3 3ο Σενάριο: Επιλογή Χαρακτηριστικών βάσει Chi square

Κατά την τρίτη πειραματική διαδικασία, πραγματοποιήθηκε επιλογή χαρακτηριστικών του συνόλου δεδομένων βάσει του στατιστικού ελέγχου Chi square. Συγκεκριμένα επιλέχθηκαν τα 14 πρώτα σε κατάταξη χαρακτηριστικά με τον μέσο όρο των μετρήσεων των αλγορίθμων ταξινόμησης ανά μετρική ταξινόμησης, να παρουσιάζονται στον παρακάτω πίνακα.

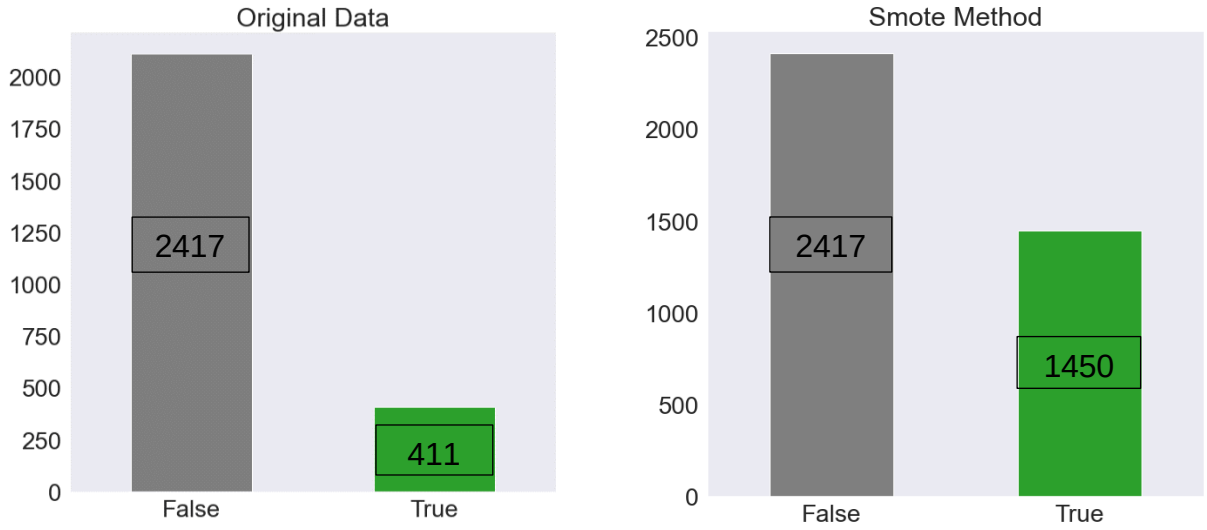
Πίνακας 7. Αποτελέσματα Αλγορίθμων - Chi Square

14 SELECTED FEATURES BASED ON CHI SQUARE TEST					
MODEL	Recall	Precision	F1 Score	Accuracy	Challenge Metric
Stacking Model	92.31%	92.34%	92.31%	92.31%	80.51%
Random Forest	92.34%	92.92%	92.56%	92.34%	85.51%
KNN	91.59%	91.21%	91.33%	91.59%	74.50%
SVC	91.76%	91.76%	91.82%	91.97%	77.62%
GNB	89.22%	92.44%	90.12%	89.22%	89.86%
Logistic Regression	92.33%	91.99%	92.08%	92.33%	79.56%
XGB	91.68%	91.62%	91.64%	91.68%	79.26%
Adaboost	91.54%	91.21%	91.31%	91.54%	78.62%
LGBM	92.25%	92.29%	92.25%	92.25%	80.06%
MLP	92.77%	92.73%	92.70%	92.77%	81.56%

Με το πέρας της τρίτης πειραματικής διαδικασίας, η πειραματική διερεύνηση συνεχίστηκε με την επιλογή των 14 χαρακτηριστικών βάσει του στατιστικού ελέγχου Chi square, διότι η συγκεκριμένη μέθοδος παρουσίασε τα βέλτιστα αποτελέσματα.

5.1.4 4ο Σενάριο: Επιλογή Χαρακτηριστικών βάσει Chi square & Smote

Κατά το τέταρτο πειραματικό σενάριο, αρχικά χρησιμοποιήθηκε η επιλογή των 14 πρώτων σε κατάταξη χαρακτηριστικών του συνόλου δεδομένων βάσει του στατιστικού ελέγχου Chi square. Στην συνέχεια εφαρμόστηκε η τεχνική υπερδειγματοληψίας συνθετικής μειονότητας για την αντιμετώπιση της μη ισορροπημένης κατανομής εγγραφών στην κλάση πρόβλεψης, που δυσκολεύει τα μοντέλα στην ακριβή διάκριση μεταξύ τους. Συγκεκριμένα εφαρμόστηκε η τεχνική Borderline SMOTE, η οποία κατασκευάζει νέα συνθετικά παραδείγματα, τα οποία προέρχονται από τις εγγραφές της κατηγορίας μειοψηφίας που δεν έχουν ταξινομηθεί σωστά, δηλαδή τις «δύσκολες περιπτώσεις». Η διαδικασία αυτή πραγματοποιείται βάσει του αλγόριθμου Κ πλησιέστερων γειτόνων των παραδειγμάτων και τον πολλαπλασιασμό της διαφορά των διανυσμάτων με ένα τυχαίο αριθμό στο διάστημα [0,1]. Ακόμη επιλέχθηκε η στρατηγική, ο αριθμός των συνθετικών δεδομένων της τάξης μειονότητας να είναι ίσος με το 60% του αριθμού των παρατηρήσεων της πλειοψηφικής τάξης, ώστε να βοηθήσει τους αλγόριθμους στην καλύτερη απόδοση τους στο στάδιο της πρόβλεψης.



Εικόνα 24. Κατανομή παρατηρήσεων - Smote Method

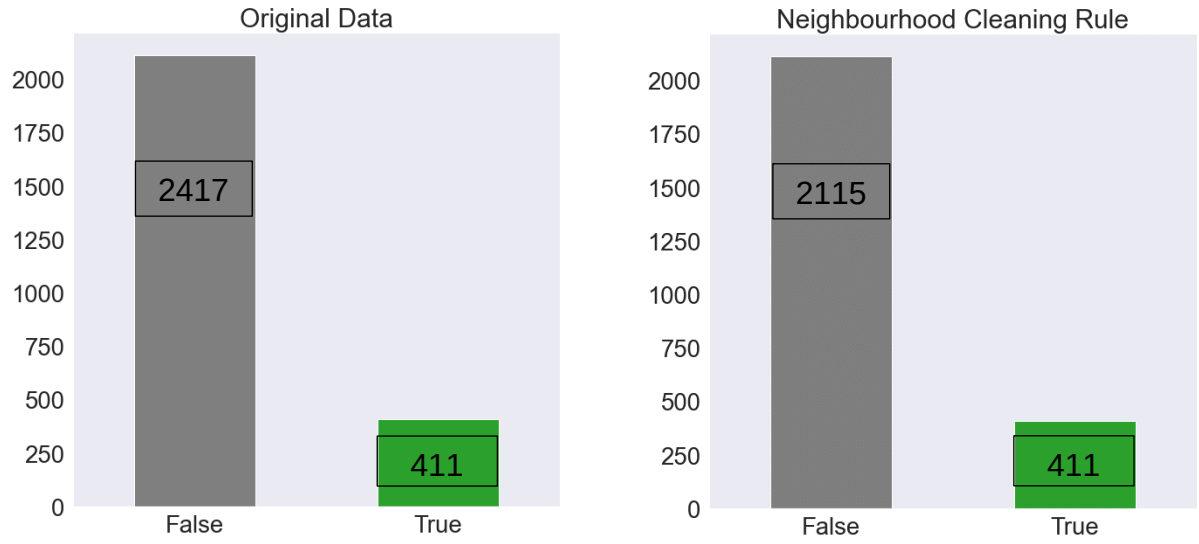
Ο μέσος όρος των μετρήσεων των αλγορίθμων ταξινόμησης ανά μετρική ταξινόμησης της συγκεκριμένης προσέγγισης παρουσιάζεται στον παρακάτω πίνακα.

Πίνακας 8. Αποτελέσματα Αλγορίθμων - Chi Square & Smote Method

14 SELECTED FEATURES BASED ON CHI SQUARE TEST & Smote Method					
MODEL	Recall	Precision	F1 Score	Accuracy	Challenge Metric
Stacking Model	90.09%	90.25%	90.14%	90.09%	88.19%
Random Forest	90.03%	90.52%	90.12%	90.03%	89.01%
KNN	89.68%	89.95%	89.74%	89.68%	89.22%
SVC	88.31%	88.67%	88.40%	88.31%	87.20%
GNB	84.04%	84.17%	84.09%	84.04%	83.29%
Logistic Regression	84.05%	83.99%	83.79%	84.05%	80.46%
XGB	90.81%	90.90%	90.83%	90.81%	87.91%
Adaboost	83.70%	83.74%	83.35%	83.70%	79.85%
LGBM	89.66%	89.84%	89.71%	89.66%	87.56%
MLP	87.32%	87.40%	87.33%	87.32%	85.99%

5.1.5 5ο Σενάριο: Επιλογή Χαρακτηριστικών βάσει Chi square & NCR

Κατά το πέμπτο πειραματικό σενάριο, αρχικά χρησιμοποιήθηκε η επιλογή των 14 πρώτων σε κατάταξη χαρακτηριστικών του συνόλου δεδομένων βάσει του στατιστικού ελέγχου Chi square. Στην συνέχεια εφαρμόστηκε ο κανόνας καθαρισμού δεδομένων Neighborhood Cleaning Rule. Σε αντίθεση με την τεχνική Smote, ο κανόνας NCR δεν δημιουργεί νέα συνθετικά δεδομένα. Η εφαρμογή του στην πλειοψηφική τάξη, βασιζόμενη στις ιδιότητες του κανόνα Condensed Nearest Neighbor (CNN) για την κατάργηση περιττών παραδειγμάτων όσο και του κανόνα Edited Nearest Neighbors (ENN) για την αφαίρεση θορυβωδών παραδειγμάτων, απέρριψε περίπου 300 εγγραφές από το σύνολο δεδομένων (Εικόνα 25). Ο κανόνας NCR λοιπόν, εστιάζει λιγότερο στη βελτίωση της ισορροπίας των δύο τάξεων και περισσότερο στην ποιότητα των παραδειγμάτων που διατηρούνται στην πλειοψηφική τάξη.



Εικόνα 25. Κατανομή παρατηρήσεων - Neighborhood Cleaning Rule

Με την ολοκλήρωση του φιλτραρίσματος του συνόλου δεδομένων και την εφαρμογή της στρωματοποιημένης δεκαπλάσιας διασταυρούμενης επικύρωσης, πραγματοποιήθηκε ο έλεγχος για την ακρίβεια των αλγορίθμων κατά το στάδιο της εκπαίδευσης. Ο παρακάτω πίνακας παρουσιάζει τα αποτελέσματα του ελέγχου και συγκρίνει τον μέσο όρο ακρίβειας στο στάδιο της εκπαίδευσης με αυτόν στο στάδιο της πρόβλεψης νέων δεδομένων για το σύνολο των αλγορίθμων.

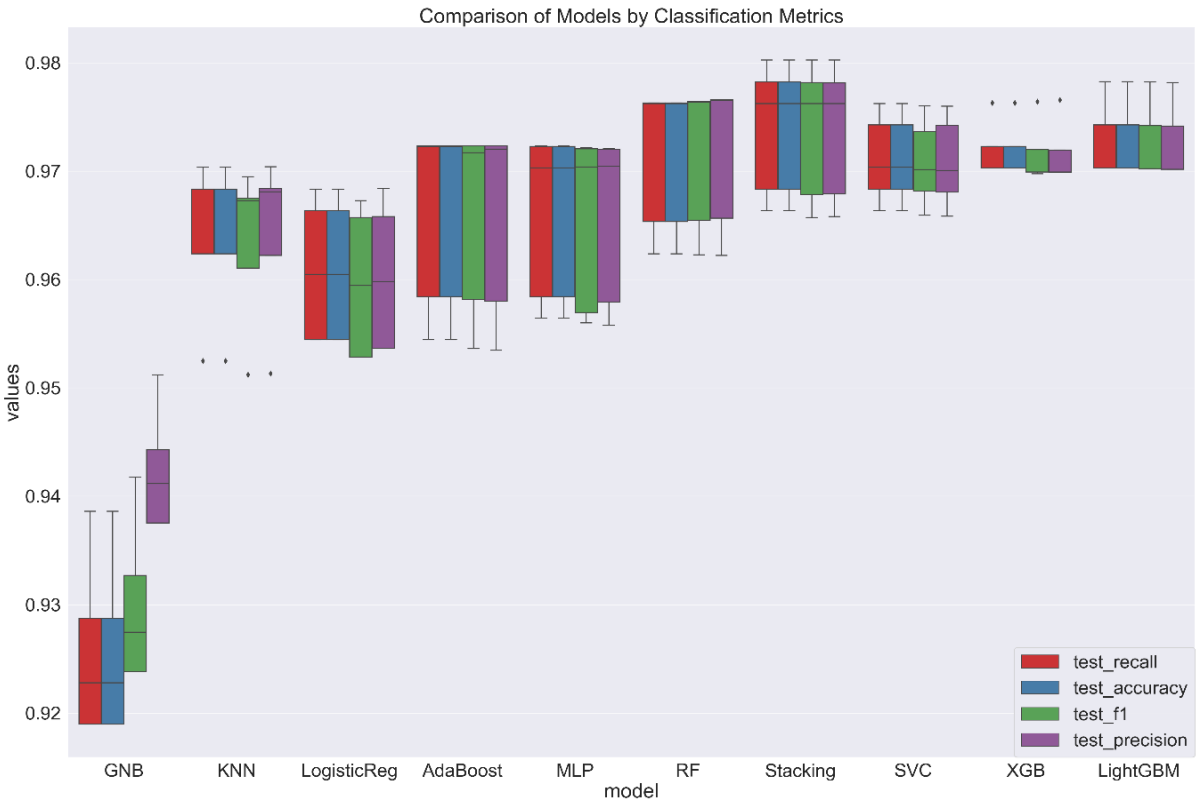
Πίνακας 9. Σύγκριση Αλγορίθμων: Training/Testing Score

MODEL	Training Score	Testing Score
Stacking Model	98.97%	97.62%
Random Forest	98.97%	97.25%
KNN	97.62%	96.43%
SVC	98.17%	97.06%
GNB	92.91%	92.76%
Logistic Regression	97.26%	96.47%
XGB	99.32%	97.17%
Adaboost	97.03%	96.62%
LGBM	98.97%	97.38%
MLP	97.38%	96.35%

Επιπλέον, όσο αφορά τον μέσο όρο μετρήσεων των αλγορίθμων ταξινόμησης ανά μετρική ταξινόμησης της συγκεκριμένης προσέγγισης, παρουσιάζονται αναλυτικά τα αποτελέσματα της πειραματικής διαδικασίας στον παρακάτω πίνακα (Πίνακας 9). Ακόλουθα, παρουσιάζεται από το γράφημα της εικόνας 26 η διακύμανση των μετρήσεων για την καλύτερη απεικόνιση και ερμηνεία των παραγόμενων αποτελεσμάτων.

Πίνακας 10. Βέλτιστα Αποτελέσματα Αλγορίθμων

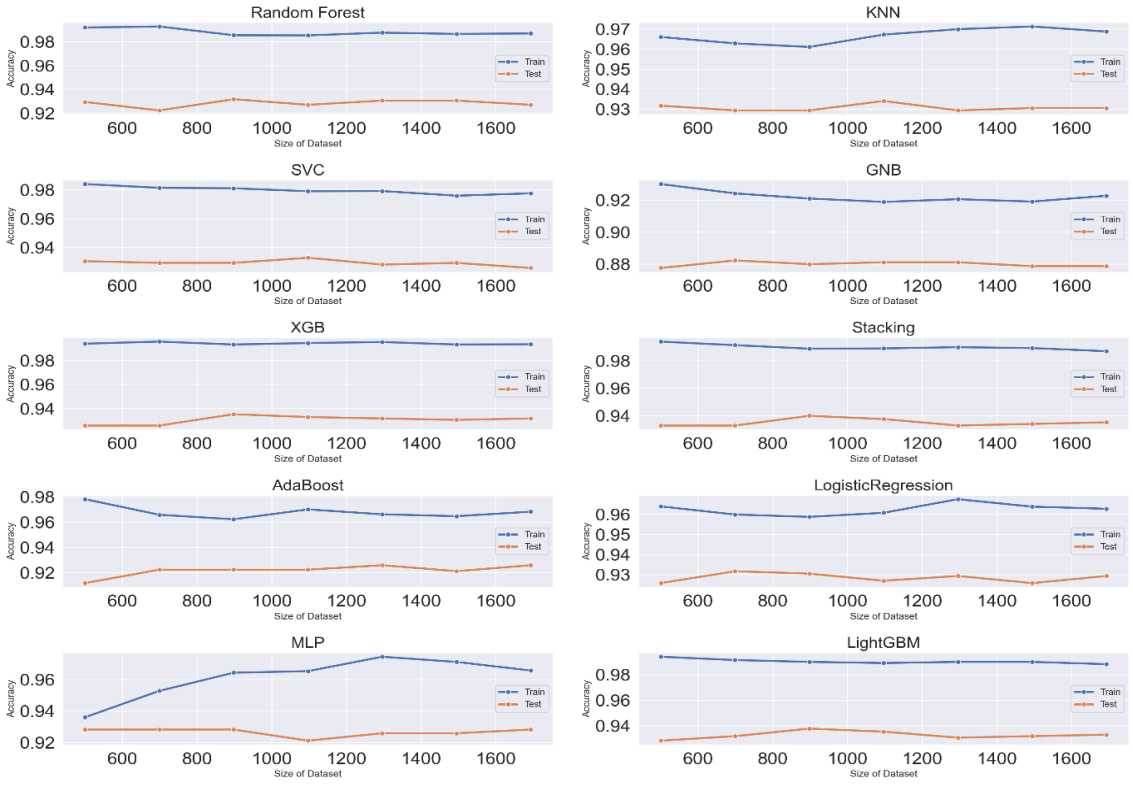
14 SELECTED FEATURES BASED ON CHI SQUARE TEST & NCR RULE					
MODEL	Recall	Precision	F1 Score	Accuracy	Challenge Metric
Stacking Model	97.62%	97.62%	97.63%	97.62%	93.45%
Random Forest	97.25%	97.25%	97.26%	97.25%	92.55%
KNN	96.43%	96.40%	96.33%	96.43%	87.31%
SVC	97.06%	97.03%	97.03%	97.06%	91.62%
GNB	92.76%	94.38%	93.18%	92.76%	92.70%
Logistic Regression	96.47%	96.43%	96.38%	96.47%	89.85%
XGB	97.17%	97.14%	97.14%	97.17%	92.12%
Adaboost	96.62%	96.58%	96.59%	96.62%	91.25%
LGBM	97.49%	97.52%	97.51%	97.38%	92.16%
MLP	96.35%	96.31%	96.32%	96.35%	92.35%



Εικόνα 26. Σύγκριση Βέλτιστων Αλγορίθμων Ταξινόμησης

5.2 Μέγεθος συνόλου Δεδομένων & Ανάλυση Υπερεκπαίδευσης

Το μέγεθος του συνόλου δεδομένων αποτελεί σημαντική ανησυχία στον ιατρικό τομέα, όπου η έλλειψη δεδομένων είναι σύνηθες φαινόμενο και η ταξινόμηση είναι από μόνη της μια πρόκληση. Γίνεται λοιπόν πιο δύσκολη πρόκληση όταν εφαρμόζεται σε μικρά σύνολα δεδομένων εκπαίδευσης, τα οποία οδηγούν σε ένα αναξιόπιστο και μεροληπτικό μοντέλο ταξινόμησης. Με στόχο λοιπόν την διερεύνηση της επίδρασης του μεγέθους των δεδομένων εκπαίδευσης στην συνολική απόδοση των εποπτευόμενων μοντέλων ταξινόμησης, μελετήθηκε η απόδοση των μοντέλων όταν εκπαιδεύτηκαν σε διαφορετικού μεγέθους υποσύνολα των συνολικών διαθέσιμων εγγραφών. Για τις ανάγκες της ανάλυσης, αρχικά έγινε διαχωρισμός του συνόλου δεδομένων σε δεδομένα εκπαίδευσης και δοκιμής και ακολούθως δημιουργήθηκαν 7 διαφορετικά υποσύνολα δεδομένων από το σύνολο εκπαίδευσης, που χρησιμοποιήθηκαν για την εκπαίδευση των αλγορίθμων. Η αύξηση της ποσότητας δεδομένων στο σύνολο εκπαίδευσης αυξάνει την αναλογία δεδομένων προς θόρυβο. Εάν τα δεδομένα εκπαίδευσης και δοκιμής προέρχονται λοιπόν από διαφορετικές κατανομές, η αύξηση του όγκου των δεδομένων δεν θα μειώσει αυτήν την πηγή θορύβου. Επομένως, η αύξηση του όγκου των δεδομένων μπορεί να επιδεινώσει την υπερπροσαρμογή μόνο με την παράλληλη αύξηση της πολυπλοκότητας του μοντέλου. Διαφορετικά, η απόδοση στο σύνολο δεδομένων δοκιμής θα πρέπει να βελτιωθεί ή να παραμείνει η ίδια, αλλά σε καμία περίπτωση να μην χειροτερέψει σημαντικά. Τα αποτελέσματά της διερεύνησης, όπως παρουσιάζονται στην εικόνα 27, έδειξαν ότι η συνολική απόδοση των ταξινομητών εξαρτάται από το πόσο ένα σύνολο δεδομένων αντιπροσωπεύει την αρχική κατανομή και όχι από το μέγεθός του, επαληθεύοντας την μελέτη [52]. Σημαντικά ευρήματα της ανάλυσης αποτέλεσαν οι μικρές διακυμάνσεις στην απόδοση των αλγορίθμων, καθώς και ο έγκυρος έλεγχος απουσίας υπερεκπαίδευσης των αλγορίθμων.



Εικόνα 27. Dataset size - Overfitting Analysis

5.3 Προτεινόμενος Αλγόριθμος

Στα πλαίσια της παρούσας εργασίας και κατά το στάδιο των πειραματικών διεργασιών επιλέχθηκε η δημιουργία και η δοκιμή ενός stacked generalized αλγόριθμου, εφόσον αποτελεί μία από τις πιο αποτελεσματικές προσεγγίσεις στα προβλήματα ταξινόμησης. Η επιλογή των μοντέλων για την τελική δομή του stacked generalized αλγόριθμου, που προτείνει η παρούσα εργασία, έγινε με την δοκιμή όλων των πιθανών συνδυασμών αλγορίθμων μηχανικής μάθησης που μελετήθηκαν ξεχωριστά. Συγκεκριμένα ο stacked generalized αλγόριθμος που επιλέχθηκε, χρησιμοποιεί ως μοντέλα επιπέδου-0, τους αλγορίθμους Random Forest, Naïve Bayes και LightGBM με τις υπερπαραμέτρους που παρουσιάζονται στον πίνακα 4 του προηγούμενου κεφαλαίου, οι οποίοι εφαρμόζονται στο σύνολο δεδομένων εκπαίδευσης. Με την ολοκλήρωση των προβλέψεών τους, χρησιμοποιείται στην συνέχεια ο αλγόριθμος Logistic Regression ως ένα μετα-μοντέλο (μοντέλο επιπέδου-1), που ενσωματώνει τις προβολές των αρχικών μοντέλων, μαθαίνοντας πώς να συνδυάζει καλύτερα τις προηγούμενες προβλέψεις. Ο τελικός αλγόριθμος γνωρίζει πότε να χρησιμοποιήσει κάθε μοντέλο στο σύνολο δεδομένων και χρησιμοποιείται για να αντιμετωπίσει το πρόβλημα ως τελικός εκτιμητής.

Η κύρια ιδέα για την δημιουργία του stacked generalized αλγόριθμου δημιουργήθηκε κατά την παρατήρηση των πειραματικών διεργασιών με την απόδοση του αλγορίθμου Naïve Bayes να υστερεί των υπολοίπων ταξινομητών. Η λογική λοιπόν ήταν ότι συνδυάζοντας ταξινομητές με διαφορετικές επαγωγικές προκαταλήψεις, ο χώρος χαρακτηριστικών θα διερευνηθεί με διαφορετικό τρόπο, με αποτέλεσμα διαφορετικούς ταξινομητές των οποίων τα λάθη δεν συσχετίζονται. Ως εκ τούτου, χρησιμοποιήθηκαν μαζί του στο επίπεδο-0 και οι αλγόριθμοι Random Forest και LightGBM, οι οποίοι παρουσίαζαν τα καλύτερα αποτελέσματα.

Ο συνδυασμός των αποτελεσμάτων αυτών των μοντέλων σε ένα μετα-ταξινομητή, οδήγησε τους ταξινομητές να μάθουν ο ένας από τα λάθη του άλλου με αποτέλεσμα την δημιουργία ενός αποδοτικότερου μοντέλου. Αναφορικά δε με την απόδοση του προτεινόμενου ταξινομητή, όπως παρουσιάζεται και στον παρακάτω πίνακα, τα καλύτερα αποτελέσματα προήλθαν από την προσέγγιση του προβλήματος με την επιλογή χαρακτηριστικών βάσει του στατιστικού ελέγχου Chi Square και την χρήση του κανόνα καθαρισμού δεδομένων NCR.

Πίνακας 11. Αποτελέσματα Stacking Model σε κάθε πειραματικό σενάριο

Stacking Model						
Process	Feature Number	Recall	Precision	F1 Score	Accuracy	Challenge Metric
Chi Square & NCR Rule	14	97.62%	97.62%	97.63%	97.62%	93.45%
Chi Square & Smote Method	14	90.09%	90.25%	90.14%	90.09%	88.19%
Chi Square	14	92.31%	92.34%	92.31%	92.31%	80.51%
Spearman Correlation	18	92.25%	92.30%	92.26%	92.25%	75.39%
All Features of Dataset	28	92.02%	92.05%	91.16%	92.05%	73.35%

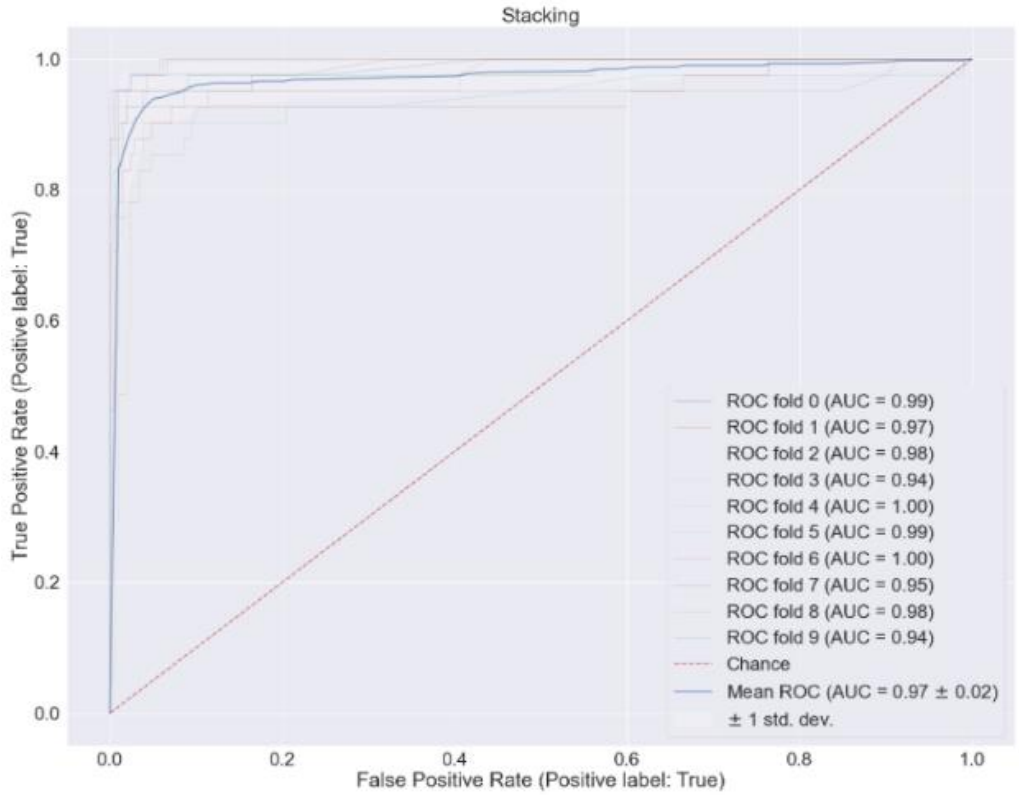
Επιπλέον στο πλαίσιο της καλύτερης προσέγγισης, όσο αφορά τις μετρήσεις του προτεινόμενου αλγορίθμου ταξινόμησης ανά μετρική, παρουσιάζονται αναλυτικά τα αποτελέσματα της πειραματικής διαδικασίας στους πίνακες 12 και 13, με την παράλληλη σύγκριση απόδοσης του μοντέλου με την εφαρμογή της μεθόδου train test split και 10-Fold Stratified Cross Validation. Ακόλουθα, παρουσιάζεται από το γράφημα στην εικόνα 28, η διακύμανση της καμπύλης Roc ως μετρική ταξινόμησης ανά fold δεδομένων δοκιμής για την καλύτερη απεικόνιση και ερμηνεία των παραγόμενων αποτελεσμάτων.

Πίνακας 12. Train Test Split Method Results

Stacking Model	
70% - 30% Train Test Split Method	
Classification Metrics	
Precision False	97%
Precision True	91%
Recall False	98%
Recall True	87%
F1 False	98%
F1 True	89%
Accuracy	96%
Challenge Metric	93.7%

Πίνακας 13. 10-Fold Cross Validation Results

Stacking Model	
10-Fold Cross Validation Method	
Classification Metrics	
Accuracy	97.6%
Precision	97.6%
Roc Curve	97.4%
F1 Score	97.6%
Specificity	98.3%
Recall	97.6%
Challenge Metric	93.4%



Εικόνα 28.10-Fold Cross Validation Roc Curve

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εφαρμογή μοντέλων μηχανικής μάθησης σε σύνολα δεδομένων εφαρμογών ιατρικής παρακολούθησης αποτελεί στην σημερινή εποχή ένα πολύ ενδιαφέρον θέμα για τους ερευνητές, καθώς υπάρχουν πολλά ζητήματα υγείας που χρειάζονται διερεύνηση. Στην παρούσα διπλωματική εργασία, μελετήθηκαν διάφορες τεχνικές ταξινόμησης, όπως οι αλγόριθμοι Random Forest, Multilayer Perceptron, SVM, Logistic Regression, Naïve Bayes, KNN, AdaBoost, XGBoost, LightGBM καθώς και ο Stacked Ensemble learning algorithm που δημιουργήθηκε στα πλαίσια της εργασίας. Τα πειράματα διενεργήθηκαν στο σύνολο δεδομένων που δημιουργήθηκε στα πλαίσια του προβλήματος, μέσω της διαδικασίας δημιουργίας χαρακτηριστικών από μία μη σχεσιακή βάση δεδομένων. Με την ολοκλήρωση της επικύρωσης των δεδομένων, αναζητήθηκε η καλύτερη τεχνική ταξινόμησης για την πρόβλεψης πιθανής πρόωρης εγκατάλειψης των χρηστών από την χρήση της εφαρμογής. Με την χρήση της τεχνικής 10 fold stratified cross validation στα πειραματικά σενάρια που πραγματοποιήθηκαν, αποδείχθηκε ότι όλες οι τεχνικές ταξινόμησης λειτουργούν με πάνω από 90% ακρίβεια. Αναφορικά με την χρησιμοποίηση του στατιστικού δείκτη Chi-squared, ως γνώμονα για την επιλογή των κατάλληλων χαρακτηριστικών γνωρισμάτων, με την μείωση των διαστάσεων, μειώθηκε ο κίνδυνος overfitting των αλγορίθμων και ο χρόνος εκπαίδευσης του μοντέλου, ενώ παράλληλα αυξήθηκε σημαντικά η ακρίβεια των αποτελεσμάτων. Εν συνεχεία, η προσέγγιση του προβλήματος, με την χρήση του αλγόριθμου καθαρισμού δεδομένων Neighborhood Cleaning Rule οδήγησε στην βέλτιστη απόδοση όλων των αλγορίθμων ταξινόμησης, υπογραμμίζοντας ότι η ποιότητα των αποτελεσμάτων ταξινόμησης δεν εξαρτάται απαραίτητα από το μέγεθος της τάξης. Επομένως, θα πρέπει να λαμβάνονται υπόψη, εκτός από την κατανομή κλάσης, και άλλα χαρακτηριστικά των δεδομένων, όπως ο θόρυβος, που μπορεί να εμποδίσουν την βέλτιστη ταξινόμηση. Η προτεινόμενη προσέγγιση θα μπορούσε να αποτελέσει μελλοντικά μέρος μιας τεχνολογίας τεχνητής νοημοσύνης (AI) για την απομακρυσμένη παρακολούθηση υγείας ηλικιωμένων ατόμων, με σκοπό την μέγιστη δυνατή συμμόρφωση των χρηστών της. Συγκεκριμένα προτείνεται η χρήση του κανόνα καθαρισμού δεδομένων Neighborhood Cleaning Rule (NCR) και ο αλγόριθμος ταξινόμησης με την μέθοδο μάθησης Stacked Generalization που δημιουργήθηκε για την βέλτιστη πρόβλεψη της πρόωρης εγκατάλειψης των χρηστών της εφαρμογής παρακολούθησης υγείας. Τα αποτελέσματα δείχνουν ότι ο προτεινόμενος αλγόριθμός ήταν ικανός να προβλέψει την πρόωρη εγκατάλειψη των χρηστών από την εφαρμογή με ποσοστό ακρίβειας 97,6%, ενώ παράλληλα, βάσει της μετρικής που επιλέχθηκε από τον διαγωνισμό σε ποσοστό 93.4%.

Εικόνα 29. Σύστημα Έγκαιρης Προειδοποίησης

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός Όρος
Classifier	Ταξινομητής
Data Analysis	Ανάλυση Δεδομένων
Big Data	Μεγάλος όγκος δεδομένων
Predictive	Προγνωστική
Exploratory	Διερευνητική
Prescriptive	Κανονιστική
Regression Analysis	Ανάλυση Παλινδρόμησης
Machine Learning	Μηχανική Μάθηση
Artificial Neural Networks	Τεχνητά Νευρωνικά Δίκτυα
Training set	Δεδομένα Εκπαίδευσης
Test set	Δεδομένα Δοκιμής
NoSQL Databases	Μη Σχεσιακές Βάσεις Δεδομένων
Artificial intelligence	Τεχνητή Νοημοσύνη
Shema-less	Ακαθόριστου Σχήματος
Supervised Learning	Επιβλεπόμενη Μάθηση
Unsupervised Learning	Μάθηση Χωρίς Επίβλεψη
Random Forest	Τυχαία Δάση
Information Gain	Κέρδος Πληροφορίας
Cross Validation	Διασταυρούμενη Επικύρωση
Overfitting	Υπερεκπαίδευση
Multilayer Perceptron	Δίκτυα Πολλαπλών Στρωμάτων
Logistic Regression	Λογιστική Παλινδρόμηση
Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
K-Nearest Neighbors	Κ Κοντινότεροι Γείτονες
Naïve Bayes	Μπεϊζιανός Αλγόριθμος
AdaBoost	Αλγόριθμος Προσαρμοσμένης Ενίσχυσης
Weak Learners	Αδύναμοι Αλγόριθμοι μάθησης
Decision Trees	Δέντρα Απόφασης
Ensemble Model	Συνδυαστικού Μοντέλου Ταξινόμησης
XGBoost	Αλγόριθμος Ακραίας Ενίσχυσης Κλίσης
Accuracy	Ακρίβεια
Sensitivity	Ευαισθησία
Attributes	Γνωρίσματα
Features	Χαρακτηριστικά
Specificity	Ειδικότητα
Curse of Dimensionality	Κατάρα των Διαστάσεων
Dimensionality Reduction	Μείωση Διαστάσεων
Feature Selection	Επιλογή Χαρακτηριστικών

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

Σύντμηση	Έννοια Σύντμησης
ΜΔΕ	Μεταπτυχιακή Διπλωματική Εργασία
AI	Artificial intelligence
CRUD	Create, Read, Update, Delete
RF	Random Forest
MLP	Multilayer Perceptron
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
XGBoost	Extreme Regularizing Gradient Boosting
LightGBM	Light Gradient Boosting Machine
GOSS	One-Side Sampling
EFB	Exclusive Feature Bundling
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
QoL	Quality of Life
SPQ	Self-Perception Questionnaire
UTAUT	Unified Theory of Acceptance and Use of Technology
EDA	Exploratory Data Analysis

ΠΑΡΑΡΤΗΜΑ Ι. Python Code: Class User - Rollout Method

```
class User:
    """User object describing a user of the experiment"""

    def __init__(self, uid, df_fingertapping, df_brain_games,
    df_physical_activity, df_mindfulness, df_SOCIODEMOGRAPHIC=None,
    df_EQ5D3L=None, df_UCLA=None, df_SPQ=None):

       # Set variables
         self.uid = uid
         self.fingertapping =
         df_fingertapping[df_fingertapping['record_id'] ==
         self.uid].drop(['record_id'], axis=1)
         self.brain_games =
         df_brain_games[df_brain_games['record_id'] ==
         self.uid].drop(['record_id'], axis=1)
         self.physical_activity =
         df_physical_activity[df_physical_activity['record_id'] ==
         self.uid].drop(['record_id'], axis=1)
         self.SOCIODEMOGRAPHIC =
         df_SOCIODEMOGRAPHIC[df_SOCIODEMOGRAPHIC['record_id'] ==
         self.uid].drop(['record_id'], axis=1)
         self.EQ5D3L = df_EQ5D3L[df_EQ5D3L['record_id'] ==
         self.uid].drop(['record_id'], axis=1)
         self.UCLA = df_UCLA[df_UCLA['record_id'] ==
         self.uid].drop(['record_id'], axis=1)
         self.SPQ = df_SPQ[df_SPQ['record_id'] ==
         self.uid].drop(['record_id'], axis=1)

         # Fill with zeros if non-existant
         if len(self.EQ5D3L.values) == 0:
             self.EQ5D3L.loc[1] = [0] * len(self.EQ5D3L.columns)
         if len(self.UCLA.values) == 0:
             self.UCLA.loc[1] = [0] * len(self.UCLA.columns)
         if len(self.SPQ.values) == 0:
             self.SPQ.loc[1] = [0] * len(self.SPQ.columns)

         # Create dynamic fields
         self.set_start_time()
         self.set_end_time()
         self.set_time_periods()
         self.set_acquisition()
         self.set_adherence()


    def revert(self):

         self.set_start_time()
         self.set_time_periods()
         self.set_acquisition()
         self.set_adherence()
```

```
def forward_week(self, weeks=1):

    if self.start_time + timedelta(hours=weeks*7*24 + 7*24*7.5) >
    self.end_time:
        return False

    self.set_start_time(weeks)
    self.set_time_periods()
    self.set_acquisition()
    self.set_adherence()
    return True

def set_start_time(self, weeks=None):

    if weeks is None:
        self.start_time = datetime.strptime(
    self.SOCIODEMOGRAPHIC['date_of_entering'].values[0],
    '%Y-%m-%d')
    else:
        self.start_time = self.start_time +
        timedelta(hours=weeks*7*24)

def set_end_time(self):

    self.end_time = datetime.strptime(
  self.SOCIODEMOGRAPHIC['date_of_finalization'].values[0],
  '%Y-%m-%d')

def set_time_periods(self):

    time_periods = []
    for week in range(16):
        time_periods.append(self.start_time +
    timedelta(hours=week*3.5*24))
    self.time_periods = time_periods

def get_partial_acquisition(self, df):

    acquisition = []
    for time_period in range(len(self.time_periods) - 1):
        times = 0
        for start_time in df['start']:
            if self.time_periods[time_period] <= start_time <
            self.time_periods[time_period + 1]:
                times += 1
                acquisition.append(1)
                break
        if times < 1:
            acquisition.append(0)
    return acquisition

def set_adherence(self):

    self.adherence =
    sum([any(aq) for aq in self.acquisitions[-3:]]) >= 2
```

```
    def set_acquisition(self):

        fingertapping =
        self.get_partial_acquisition(self.fingertapping)
        brain_games = self.get_partial_acquisition(self.brain_games)
        mindfulness = self.get_partial_acquisition(self.mindfulness)
        physical_activity =
        self.get_partial_acquisition(self.physical_activity)
        self.acquisitions = list(zip(fingertapping, mindfulness,
        brain_games,physical_activity))

def get_values_from_user_object(user):
    values = []
    tmp_df = user.SOCIODEMOGRAPHIC.drop(['date_of_entering',
    'date_of_finalization',"living_status"], axis=1)
    values.extend([int (val) for val in tmp_df.values[0]])
    values.extend([int (val) for val in user.EQ5D3L.values[0]])
    values.extend([int (val) for val in user.UCLA.values[0]])
    values.extend([int (val) for val in user.SPQ.values[0]])
    del tmp_df
    return values

users = []

for uid in uids:
    users.append(User(uid, df_fingertapping, df_brain_games,
    df_physical_activity, df_mindfulness, df_SOCIODEMOGRAPHIC,
    df_EQ5D3L, df_UCLA, df_SPQ))

columns =
    list(df_SOCIODEMOGRAPHIC.drop(['record_id','living_status',
    'date_of_entering','date_of_finalization'], axis=1))+
    list(df_EQ5D3L.drop(['record_id'], axis=1))+
    list(df_UCLA.drop(['record_id'], axis=1))+
    list(df_SPQ.drop(['record_id'], axis=1))+
    [f'T-{i}_{j}' for i in range(11, -1, -1)
    for j in range(1,4)]+ ['class']

ml_list = []
i = 1
for user in users:
    period_tmp=0
    user.revert()
    static_values = get_values_from_user_obj(user)
    week = [item for sublist in user.acquisitions[:-3]
            for item in sublist]
    tmp = []
    tmp.extend(static_values)
    tmp.extend(week)
    tmp.append(user.adherence)
    ml_list.append(tmp)
    period_tmp+=1
```

```
    while user.forward_week():

        week = [item for sublist in user.acquisitions[:-3]
                for item in sublist]
        tmp = []
        tmp.extend(static_values)
        tmp.extend(week)
        tmp.append(user.adherence)
        ml_list.append(tmp)
        period_tmp+=1

    print(f'Done user {i}/{len(users)}')
    i += 1

# Final_Data_Ready_for_Analysis

ml_df = pd.DataFrame(ml_list, columns=columns)
```

ΠΑΡΑΡΤΗΜΑ ΙΙ. Python Code: Challenge Metric

```
def challengemetric(y_true, y_pred):

    TP = sum(y_true & y_pred)
    TN = sum(~y_true & ~y_pred)
    FP = sum(~y_true & y_pred)
    FN = sum(y_true & ~y_pred)

    Specificity = TN / (TN+FP)
    Sensitivity = TP / (TP+FN)

    return math.sqrt(Sensitivity*Specificity)
```

ΠΑΡΑΡΤΗΜΑ ΙΙΙ. Βιβλιοθήκες προγραμματισμού Python

Οι βιβλιοθήκες προγραμματισμού Python που χρησιμοποιήθηκαν για την αντιμετώπιση του προβλήματος είναι οι παρακάτω:

numpy, pandas, sklearn, scikit-learn, xgboost, lightgbm, warnings, pickle, matplotlib.pyplot, seaborn, math, datetime, pyarrow, pickle, requests, json, waterfallcharts, pytools, imblearn, collections, .

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] Grus, Joel. “Data science from scratch.”, 2015.

[2] Jiawei Han, Micheline Kamber, Jian Pei, «Data Mining (Third Edition)», 3 – “Data Preprocessing”, Morgan Kaufmann, pages 83-124, 2012.

[3] K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 910-914, 2018.

[4] Dey, N., Hassanien, A. E., Bhatt, C., Ashour, “Internet of Things and Big Data Analytics Toward Next-Generation Intelligence. Studies in Big Data”, 2018.

[5] Aghi, Rajat, Sumeet Mehta, Rahul Chauhan, Siddhant Chaudhary and N. Bohra. “A comprehensive comparison of SQL and MongoDB databases.”, 2015.

[6] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[7] Tom Mitchell, McGraw Hill, Machine Learning, 1997.

[8] Murphy, Kevin P., 1970-, *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press, 2012.

[9] L. Breiman, “Random Forests” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001/10/01, 2001.

[10] P. Sittidech, and N. Nai-arun, “Random Forest Analysis on Diabetes Complication Data”, Proceedings of the IASTED International Conference on Biomedical Engineering, BioMed 2014, pp. 315-320, 01/01, 2014.

[11] A. Agresti, “An introduction to categorical data analysis Wiley”, New York, 1996.

[12] C. M. Bishop, *Neural networks for pattern recognition*: Oxford university press, 1995.

[13] M. Hofmann, "Support Vector Machines Kernels and the Kernel Trick An elaboration for the Hauptseminar Reading Club : Support Vector Machines ", 2006.

[14] A. Pradhan, “Support vector machine-A survey,” IJETAE, vol. 2, 09/01, 2012.

[15] W. Homenda and W. Pedrycz, “Pattern Recognition: A Quality of Data Perspective”, John Wiley & Sons, Inc. USA, 2018.

[16] Boateng, Ernest Yeboah & Abaye, Daniel. A Review of the Logistic Regression Model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing, 2019.

[17] B. A. Thakkar, M. I. Hasan, and M. A. Desai, "Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier." pp. 101-105, 2014.

[18] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala, and S. Saxena, "Sentimental analysis using fuzzy and naive bayes." pp. 945-950, 2017.

[19] J. Laaksonen, and E. Oja, "Classification with learning k-nearest neighbors." pp. 1480-1483 vol.3, 1996.

[20] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of Improving K-Nearest-Neighbor for Classification." pp. 679-683, 2014.

[21] Schapire, R. E. The Boosting Approach to Machine Learning: An Overview. Lecture Notes in Statistics, pp.149–171, 2003.

[22] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, pp. 785–794, 2016.

[23] Ke Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” *NIPS*, 2017.

[24] X. Shi, Y. Cheng, and D. Xue, “Classification Algorithm of Urban Point Cloud Data based on LightGBM”, IOP Conference Series: Materials Science and Engineering, vol. 631, no. 5, 2019.

[25] Wolpert, David, Stacked Generalization, Neural Networks 5,241-259, 1992.

[26] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models", IEEE Second International Conference on Data Stream Mining & Processing (DSMP), pp. 255-258. 2018.

[27] Nakano Felipe Kenji, Mastelini Saulo, Barbon Junior Sylvio, Cerri Ricardo, Stacking Methods for Hierarchical Classification. 289-296. 2017.

[28] Haibo He, Yunqian Ma., Imbalanced Learning: Foundations, Algorithms and Applications, Wiley-IEEE Press, 2013.

[29] Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357, 2002.

[30] Agustianto Khafidurrohman, Destarianto Prawidya,Imbalance, Data Handling using Neighborhood Cleaning Rule (NCR) Sampling Method for Precision Student Modeling, 2019.

[31] Laurikkala, Jorma, Improving Identification of Difficult Small Classes by Balancing Class Distribution, Proc 8th Conf AI Med Eur Artif Intell Med. 63-66. 10.1007/3-540-48229-6_9, 2001.

[32] M. Stone, “Cross-validation: A review *Statistics”: A Journal of Theoretical and Applied Statistics,* vol. 9, no. 1, pp. 127-139, 1978.

[33] D. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” *Mach. Learn. Technol.,* vol. 2, 01/01, 2008.

[34] Jake VanderPlas, “Python Data Science Handbook: Essential Tools for Working with Data”, O'Reilly Media, Inc, 2016.

[35] Olsson S, Lymberis A, Whitehouse D. European Commission activities in eHealth. Int J Circumpolar Health, 63(4):310-6, 2004.

[36] Kordatzakis Antonios & Perakis Kostas & Haritou Maria & Maglogiannis Ilias & Koutsouris, Dimitris, A Novel Telematics Platform for Remote Monitoring of Patients, The Journal on Information Technology in Healthcare. 5. 248-254, 2007.

[37] IFMBE Science Challenge 2022, https://wchallenge2022.lst.tfo.upm.es/

[38] 2022 IFMBE Science Challenge-World Congress 2022, Boosting adherence in AHA - Dataset Overview, https://wchallenge2022.lst.tfo.upm.es/ wp-content/ uploads/ 2021/ 10/ DATASET-SCORE-EXAMPLE-WC-CHALLENGE-2021.pdf

[39] J. Kobylarz Ribeiro et al., "A Machine Learning Early Warning System: Multicenter Validation in Brazilian Hospitals," 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 321-326, 2020.

[40] Wickham Hadley, Tidy data, Journal of Statistical Software 14(10), 2014.

[41] Potdar Kedar & Pardawala Taher & Pai Chinmay, A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers, International Journal of Computer Applications. 175. 7-9. 10.5120/ijca2017915495, 2017.

[42] L. B. Amor, I. Lahyani and M. Jmaiel, "Recursive and Rolling Windows for Medical Time Series Forecasting: A Comparative Study," IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016, pp. 106-113, 2016.

[43] Camizuli Estelle & Carranza Emmanuel John, Exploratory Data Analysis (EDA). 1-7, 2018.

[44] Venkat Naveen, The Curse of Dimensionality: Inside Out, 2018.

[45] Esra Mahsereci Karabulut, Selma Ayşe Özel, Turgay İbrikçi, A comparative study on the effect of feature selection on classification accuracy, Procedia Technology, Pages 323-327, 2012.

[46] Nnamoko Nonso, Arshad Farath, England David, Vora Jiten , Norman James, Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning, 2014.

[47] M. Trabelsi, N. Meddouri, and M. Maddouri, “A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis,” Procedia Computer Science, vol. 112, pp. 186-194, 2017.

[48] M. L. McHugh, “The chi-square test of independence,” Biochemia medica, vol. 23, no. 2, pp. 143-149, 2013.

[49] Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, NY, 2008.

[50] Probst, Philipp, Anne-Laure Boulesteix and B. Bischl, “Tunability: Importance of Hyperparameters of Machine Learning Algorithms.”, J. Mach. Learn. Res. 20: 53:1-53:32, 2019.

[51] Bey R, Goussault R, Grolleau F, Benchoufi M, Porcher R., Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. *J Am Med Inform Assoc*, 2020.

[52] Althnian Alhanoof, AlSaeed Duaa, Al-Baity Heyam, Samha Amani, Dris Alanoud, Alzakari Najla, Abou Elwafa Afnan, Kurdi Heba, Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. Applied Sciences, 2021.