ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Εξόρυξη γνώσης από διαχρονικές μελέτες πληθυσμών

Πτυχιακή εργασία

Μπουζέτος Εμμανουήλ

Αθήνα, 2022

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Τριμελής Εξεταστική Επιτροπή

Ηρακλής Βαρλάμης (Επιβλέπων)
Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Χρήστος Δίου
Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Δημήτριος Μιχαήλ
Αναπληρωτής Καθηγητή, Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Ο Εμμανουήλ Μπουζέτος

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Βαρλάμη Ηρακλή για όλη τη βοήθεια και συνεχή καθοδήγηση καθόλη τη διάρκεια προετοιμασίας της πτυχιακής μου εργασίας με θέμα «Εξόρυξη γνώσης από διαχρονικές μελέτες πληθυσμών».

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου που με στήριξαν σε όλο αυτό το διάστημα της φοίτησης μου για να μπορώ να επικεντρωθώ στους στόχους μου και στην σχολή μου.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη στα Ελληνικά

Η εξόρυξη γνώσης στις διαχρονικές μελέτες πληθυσμού είναι η διαδικασία εξερεύνησης συνόλων δεδομένων από μελέτες οι οποίες επαναλαμβάνονται ανά τακτά χρονικά διαστήματα σε ένα συγκεκριμένο κομμάτι του πληθυσμού με τη χρήση αλγορίθμων, στατιστικών μεθόδων και τεχνικών μηχανικής μάθησης. Σκοπός είναι η πληροφορία η οποία θα εξαχθεί να είναι κατανοητή από τον άνθρωπο ώστε να μπορεί να χρησιμοποιηθεί για τη λήψη αποφάσεων.

Ένας σημαντικός δείκτης για την υγεία των ηλικιωμένων είναι η "ευθραυστότητα" (frailty index) ο οποίος χρησιμοποιείται ως μέτρο της γήρανσης και της ευπάθειας αυτών των ατόμων. Η πρόβλεψή της μεταβολής του frailty index στους ηλικιωμένους ανθρώπους συμβάλει ευνοϊκά για την έγκαιρη παρέμβαση ιατρικής φροντίδας ώστε να βελτιωθεί η ποιότητά ζωής τους.

Για την πρόβλεψη επιλέχθηκαν 2820 άτομά από την English Lognitudinal Study of Ageing και με τη δημιουργία ενός μοντέλου πρόβλεψης γνωρίσματων με τον πυρήνα του να αποτελείται από ένα Long Short Memory Neural Network εκτιμήθηκαν τα μελλοντικά τους γνωρίσματα. Στην συνέχεια δημιουργήθηκαν επιπλέον 5 μοντέλα μηχανικής μάθησης όπου με βάση τα γνωρίσματα αυτά έγινε η πρόβλεψη του frailty index. Τα γνωρίσματα είναι παράγωγα ιατρικών εξετάσεων όπου επαναλαμβάνονται ανά τακτά χρονικά διαστήματα στους συμμετέχοντες της English Lognitudinal Study of Ageing

Για την αξιολόγησή των μοντέλων κατασκευαστήκαν και μοντέλα αναφοράς με μειωμένη προβλεπτική δύναμη ώστε να χρησιμοποιηθούν τα μέτρα R^2 ,adjusted R^2, Mean Squared Error, Mean Absolute Squared Error για την σύγκριση και την αξιολόγηση της απόδοσης τους

Λέξεις κλειδιά: Διαχρονικές Μελέτες Πληθυσμού, Εξόρυξη Γνώσης, Frailty Index, Μηχανική Μάθηση, Πρόβλεψη

Abstract

Data mining in Lognitudinal studies is the process of exploring datasets from studies that are repeated at regular intervals in a specific section of the population using algorithms, statistical methods and machine learning techniques. The purpose is for the information to be extracted and to be understood by humans so that it can be used for accurate decision making.

An important indicator for the health of the elderly is the frailty index which is used as a measure of aging and the vulnerability of these people. Predicting the change in the frailty index in the elderly contributes favorably to timely medical intervention and by result improving their quality of life.

2820 people were selected for the prediction by the English Lognitudinal Study of Aging and with the creation of a feature prediction model with its core consisting of a Long Short Memory Neural Network, their future features were assessed. Then an additional 5 models of machine learning were created where based on these features the frailty index was predicted. The traits are derivatives of medical examinations where they are repeated at regular intervals in the participants of the English Lognitudinal Study of Aging.

For the evaluation of the models, reference models with reduced predictive power were constructed to use the measures R^2 ,adjusted R^2, Mean Squared Error, Mean Absolute Squared Error for the comparison and evaluation of their performance.

Λέξεις κλειδιά: Lognitudinal Study, Data Mining, Frailty Index, Machine Learning, Prediction

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

ELSA	English Longitudinal Study of Ageing
FI	Frailty Index
AI	Artificial Intelligence
ML	Machine Learning
LSTM	Long-Short Term Memory
RF	Random Forest
mse	Mean Squared Error
mae	Mean Absolute Error

ΚΕΦ.1: ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή Προβλήματος

Η ανάγκη για την δημιουργία ενός μέτρου το οποίο θα περιγράφει την ευπάθεια παρουσιάσεις δυσμενών εκβάσεων στην υγειά ενός ηλικιωμένου ατόμου όπως η αναπηρία, η νοσηλεία ή και η θνησιμότητας οδήγησαν στην κατασκευή ενός δείκτη (frailty index)που θα προσδιορίζει ευστοχά αυτά τα ενδεχόμενα. Αν και η αξία του δείκτη αυτού είναι μεγάλη στην τρέχουσα ζωή των ηλικιωμένων τόσο για την ποιότητα ζωής όσο και για την ιατροφαρμακευτική περίθαλψη σημαντικό πλεονέκτημα θα ήταν η πρόβλεψη των μεταβολών του δείκτη στα επόμενα έτη ζωής ενός ατόμου. Αυτό θα έχει ως αποτέλεσμα να ληφθούν προληπτικά μέτρα για την αύξηση της ποιότητας και του προσδόκιμου ζωής του ατόμου.

1.2 Σκοπός Εργασίας

Σκοπός αυτής της πτυχιακής εργασίας είναι η ανάπτυξη μοντέλων πρόγνωσης του frailty index μέσω του dataset των ιατρικών εξετάσεων(Nurse Datasets) από την έρευνα English Lognitudinal Study of Ageing(ELSA). Αρχικά καθώς ο frailty index δεν παρέχετε από την ELSA με την χρήση των κυρίως δεδομένων(Core Datasets) από την ELSA θα υπολογιστεί ένας frailty index για κάθε άτομο και για χρόνια που συμμετείχαν στην έρευνα. Έπειτα θα γίνει πρόβλεψη των εξατομικευμένων frailty Index για τους συμμετέχοντες της έρευνας αυτής με την χρήση των διάφορων μοντέλων που αναπτύξαμε. Τέλος θα γίνει σύγκριση των αποτελεσμάτων των μοντέλων με τα πραγματικά δεδομένα που μας παρέχει η ELSA αλλά και σύγκριση των μοντέλων για την αξιολόγησή της χρήσης των τεχνικών αυτών σε πραγματικά δεδομένα.

Για τον στόχο της εργασίας χρησιμοποιήθηκαν regression μέθοδοί, δοκιμάζοντας διάφορες τεχνικές όπως Νευρωνικά δίκτυα, Γραμμική Παλινδρόμηση (Linear Regression), Δάση Τυχαίας Απόφασης (Random Forest), Μηχανές Διανυσματικής Υποστήριξης (SVM) και Gradient Boosting.

ΚΕΦ.2: ΥΠΟΒΑΘΡΟ

2.1 **Frailty**

Η ευθραυστότητα(frailty) είναι ένα κλινικό σύνδρομο το οποίο παρατηρείται σε ηλικιωμένα άτομα όπου το κυρίως γνώρισμα τους είναι η αυξημένη ευπάθεια σε στρεσογόνους παράγοντες εξαιτίας της επιδεινώσεις των συστημάτων οργάνων που είναι υπεύθυνα για την διατήρηση της ομοιόστασης. Αν και το σύνδρομο αυτό είναι συχνότερο στα ηλικιωμένα άτομα η ηλικία δεν είναι ο μονός παράγοντας για την παρουσίαση του καθώς οι αίτιες του είναι περίπλοκες και πιθανώς να περιλαμβάνουν όχι μόνο βιολογικούς παράγοντες αλλά κοινωνικούς και ψυχολογικούς.

2.1.1 **Frailty Index**

Για τον δείκτη ευθραυστότητας(FI) επιλέχθηκε το μοντέλο σωρευτικού ελλείματος (Cumulative deficit model) το οποίο αναπτύχθηκε αρχικά από τον Rockwood και τους συνάδελφους του από τα δεδομένα της Καναδικής Μελέτης Υγείας και Γήρανσης (Canadian Study of Health and Aging). Ο δείκτης για το κάθε άτομο ορίζετε από ον αριθμό προβλημάτων υγείας που παρουσίασε το άτομο από ένα προκαθορισμένο σύνολο προβλημάτων (χρονιές παθήσεις, αναπηρίες, ασθένειες) προς το σύνολο αυτό. Με αυτόν τον τρόπο δημιουργείτε ένα αριθμός από 0 να περιγράφει την καθόλου ευθραυστότητα έως το 1 να περιγράφει την μέγιστη (Searle , Mitnitski, Gahbauer, Gill, & Rockwood, 2008).

$$FI = \frac{Πρόβλημα\ Υγείας_1 + Πρόβλημα\ Υγείας_2 \ldots + Πρόβλημα\ Υγείας_n}{Σύνολο\ \ Προβλημάτων\ Υγείας}$$

2.2 **Εξόρυξη Γνώσης**

Στην σύγχρονη εποχή με την ραγδαία εξέλιξη της επεξεργαστικής ισχύος και παράλληλα με τον συνεχή διαμοιρασμό πληροφορίας έχουν δημιουργηθεί υπερμεγέθη ποσότητες δεδομένων.

Στην αναζήτηση χρήσιμης και κατανοητής από των άνθρωπο γνώσης μέσα από την ανάλυση αυτών των υπερμεγεθών συλλογών δεδομένων καλείτε η χρήση των εργαλείων της Εξόρυξης Γνώσης. Η εξόρυξη γνώσης είναι μια διαδικασία για την εύρεση μοτίβων και συσχετίσεων μέσα από μεγάλα σύνολα δεδομένων. Με τα εργαλεία που προσφέρει η Εξόρυξη Γνώσης όπως υπολογιστικά μοντέλα και αλγόριθμοι γίνεται ανάλυση των δεδομένων. Σημαντικό είναι επίσης ότι πέρα από την κατανόηση που προσφέρουν αυτές οι τεχνικές στο χρόνο που πάρθηκαν τα δεδομένα είναι δυνατή και η πρόβλεψη μελλοντικών τάσεων.

Για την εξαγωγή γνώσης από τα δεδομένα θα πρέπει θα εκτελεστούν κάποιες διεργασίες οι οποίες αποτελούνται από 4 βήματα με κάθε βήμα να παράγει ένα αποτέλεσμα το οποίο θα χρησιμοποιείτε από το αμέσως επόμενο. (Ιακωβίδου, 2015)

Στο πρώτο στάδιο καθορίζεται ο στόχος της εξόρυξης γνώσης και αφού εξεταστούν η πήγες δεδομένων που είναι διαθέσιμες γίνεται η επιλογή των συνόλων δεδομένων που θα αξιοποιηθούν στην διαδικασία της εξόρυξης γνώσης για την επίτευξη του στόχου.

Στο δεύτερο στάδιο πραγματοποιείται η προεργασία των δεδομένων. Τα δεδομένα που δίνονται από την ερευνά της ELSA δεν βρίσκονται σε μια μορφή που θα είναι εύκολη η κατανόηση τους από τους αλγορίθμους τις εξορύξεις δεδομένα. Για αυτόν τον λόγο με τεχνικές preprocessing τα κανονικοποιούνται και στην συνέχεια εξετάζονται οι ελλείψεις των δεδομένων και τα λάθη τα οποία περιέχουν. Οι ελλείψεις διορθώνονται οπού είναι δυνατόν με τεχνικές imputation ενώ ανάλογος το λάθος εφαρμόζονται τεχνικές για την διόρθωση τους. Το αποτέλεσμα είναι τα dataset που θα τροφοδοτήσουν τους αλγορίθμους να είναι ομοιοδιάστατα και με όσο το δυνατόν λιγότερες έλλειψης τιμών και λαθών.

Στο τρίτο στάδιο πραγματοποιείται η δημιουργία μοντέλων και εξόρυξη προτύπων: Ανάλογα με τον τύπο της ανάλυσης είναι δυνατή η διερευνήσει για τυχόν ενδιαφέρουσες σχέσεις δεδομένων, όπως μοτίβα ή συσχετίσεις. Οι αλγόριθμοι βαθιάς μάθησης (Deep Learning) μπορούν επίσης να εφαρμοστούν για την ταξινόμηση ή ομαδοποίηση ενός συνόλου δεδομένων ανάλογα με τα διαθέσιμα δεδομένα. Εάν τα δεδομένα εισόδου φέρουν ετικέτα (δηλαδή εποπτευόμενη μάθηση), μπορεί να χρησιμοποιηθεί ένα μοντέλο ταξινόμησης για την κατηγοριοποίηση των δεδομένων ή εναλλακτικά, μπορεί να εφαρμοστεί μια παλινδρόμηση για την πρόβλεψη της πιθανότητας μιας συγκεκριμένης ανάθεσης.

Στο τέταρτο στάδιο αφού συγκεντρωθούν τα δεδομένα και τα αποτελέσματα πρέπει να αξιολογηθούν και να ερμηνευθούν. Κατά την οριστικοποίηση των αποτελεσμάτων, θα πρέπει να είναι έγκυρα, πρωτότυπα, χρήσιμα και κατανοητά. Η αξιολόγηση μπορεί να γίνει με πόλους διαθετικούς τρόπους καθώς υπάρχει πληθώρα μετρικών για αυτόν τον σκοπό κυρίως όμως βασίζονται στην σύγκριση ενός πραγματικού αποτελέσματος με ένα προβλεπόμενο. Η διάφορες ανάμεσα σε αυτά τα δυο καθορίζουν την επιτυχία της εξόρυξης γνώσης. Τέλος εφόσον το αποτέλεσμα μας είναι ορθό γίνεται μια ερμηνεύσει αυτού στην οποία μπορούν να χρησιμοποιηθούν γραφήματα και αλλά σχήματα ώστε η αποικούμενη γνώση να είναι ευκόλως κατανοητής

2.3 English Lognitudinal Study of Ageing

Η English Lognitudinal Study of Ageing (ELSA) είναι μια ερευνά η οποία συλλεγεί δεδομένα από ανθρώπους άνω των 50 χρονών για να κατανοηθούν όλες οι πτυχές της γήρανσης στην Αγγλία. Πάνω από 18.000 άτομα έχουν συμμετάσχει από το 2002 που ξεκίνησε με τον ίδιο αριθμό άνθρωπόν να πραγματοποιούν εκ νέου συνεντεύξεις κάθε δύο χρονιά. Μέχρι στιγμής έχουν πραγματοποιηθεί 9 κύματα επαναλαμβανόμενο εξετάσεων (waves) με το τελευταίο κύμα να πραγματοποιήθηκε το 2018/2019. ELSA συλλεγεί πληροφορίες για την φυσική και ψυχική υγειά, την ευεξία, την οικονομική κατάσταση αλλά και γενικότερους παράγοντες της γήρανσης και πως εξελίσσονται ανά τον χρόνο.

2.4 **Σχετικές Εργασίες**

Σκοπός της μελέτης (Yang & Bath, 2020) ήταν η πρόβλεψη της άνοιας από τα δεδομένα της ELSA. Η άνοια θεωρήθηκε πρόβλημα classification δηλαδή εάν υπάρχει άνοια η όχι και αξιολογήθηκε με δυο τρόπους (το άτομο δήλωσε στο ερωτηματολόγιο της Elsa ότι έχει γίνει διαγνώσει από ειδικό ή/και αποτέλεσμα IQCODE test τα οποία έγιναν στα waves 7,8) Για την πρόβλεψη δημιουργήθηκαν 7 μοντέλα ενώ χρησιμοποιήθηκαν 400 γνωρίσματα από τα waves 7 και 8 από την ELSA. Κάποια από αυτά να εξάχθηκαν απευθείας από την ELSA ενώ κάποια αλλά ήταν παράγωγα των καταγεγραμμένων γνωρισμάτων. Για την εκπαίδευση τον μοντέλων χρησιμοποιήθηκαν τα δεδομένα από 7^{ο} wave της ELSA ενώ για την αξιολόγηση τους έγινε χρήση του 8^{ου} Wave

Τα γνωρίσματα κατηγοριοποιήθηκαν σε 6 βασικές ομάδες: Δημογραφικοί & Οικονομικοί Παράγοντες, Κοινωνικοί Παράγοντες, Φυσικής Υγείας & Παράγοντες Αναπηρίας, Ψυχολογικοί παράγοντες, Παράγοντες Τρόπου ζωής, Νοητική Παράγοντες.

Παράλληλα δημιουργήθηκαν 3 σύνολα γνωρισμάτων:

1. Όλο το σύνολο των μεταβλητών χωρίς τα παραγόμενα γνωρίσματα
2. Όλο το σύνολο των μεταβλητών χωρίς τα γνωρίσματα Νοητικής Λειτουργίας
3. Όλο το Σύνολο των γνωρισμάτων

Τα μοντέλα που χρησιμοποιήθηκαν για την πρόβλεψη της άνοιας είναι τα εξής:

- Gradient Boosting Machine (GBM: XGB &LGB & CatBoost)
- Keras Based CNN
- Random Forest
- Regularized Greedy Forests
- Logistic Regression

Καθώς η πρόβλεψη ήταν της μορφής classification για την αξιολόγηση των μοντέλων χρησιμοποιήθηκε το μετρικό Normalized Gini Coefficient (Gini) με όσο υψηλότερο το score τόσο μεγαλύτερη η ακρίβεια του μοντέλου. Στον πίνακα 1, οποίος καταγραφεί τις επιδόσεις του κάθε μοντέλου, φαίνεται το Random Forest να έχει την υψηλότερη ακρίβεια στην πρόγνωση και στα τρία σύνολα γνωρισμάτων.

Model	Feature set 1 (Validate/Test)	Feature set 2 (Validate/Test)	Feature set 3 (Validate/Test)
XGB	0.865/0.899	0.719/0.858	0.909/0.913
LGB	0.897/0.897	0.879/0.860	0.888/0.904
CatBoost	0.904/0.888	0.834/0.858	0.872/0.891
K-CNN	0.899/0.896	0.853/0.874	0.919/0.907
RF	0.937/0.912	0.908/0.872	0.946/0.918
RGF	0.861/0.904	0 .769/0.853	0.887/0.911
LR	0.926/0.862	0.863/0.831	0.917/0.868

Πίνακας 1 (Yang & Bath, 2020) Αξιολόγηση Πρόγνωσης

Στην μελέτη (Caballero, et al., 2017) στην οποία χρησιμοποιήθηκαν τα δεδομένα από τα πρώτα 6 waves της ELSA,στόχος ήταν η δημιουργία ενός μέτρου υγείας μέσω της Item Response

Theory(IRT) για το συνόλου των συμμετεχόντων ανά κάθε wave .Με το μέτρο αυτάο είναι δυνατή η σύγκριση της υγείας ανά Wave και παράλληλα να γίνει μια εξερεύνηση των κοινωνικοδημογραφικών που το επηρεάζει.

Για την δημιουργία του μέτρου υγείας πραγματοποιήθηκε factor analysis όπου χρησιμοποιήθηκαν 45 variables οι οποίες είτε ήταν αυτοαναφερόμενες στα ερωτηματολόγια της ELSA από τους συμμετέχοντες είτε αποτελέσματα από εξετάσεις. Για τις 45 αυτές μεταβλητές έγινε στο 70% Exploratory Factor Analysis(EFA) ώστε να αναγνωριστούν οι first order factors ενώ στο υπόλοιπο 30% των δειγμάτων πραγματοποιήθηκε η τεχνική Confirmatory Factor Analysis ώστε να δημιουργηθεί ένα μοντέλο με δομή Second Order. Τέλος εφαρμόστηκε ένα Minimum Average Partial Test ώστε να βρεθεί ο ελάχιστος αριθμός factors που θα χρησιμοποιούν από την EFA.

Στο άρθρο αναφέρεται ότι για την πρόβλεψη το μέτρου υγείας χωρίστηκε σε 4 classes
Μέτρο Υγείας<=20
20<S Μέτρο Υγείας <=40
40< Μέτρο Υγείας <=60
60< Μέτρο Υγείας
Και χρησιμοποιήθηκαν διάφοροι classifiers για να μοντελοποιηθούν οι διάφορες στα μοτίβα των factors. Ως τελικό αποτέλεσμα δίνεται ότι την μεγαλύτερη επίδοση σε αυτό το πρόβλημα το είχε ένας Random Forest Classifier με average accuracy 83%

Στην μελέτη (Su, et al.,2021) στόχος ήταν χρήση ενός LSTM νευρωνικό δικτύου για την πρόβλεψη των διαφορετικών παραγόντων για τον κίνδυνο της κατάθλιψης στην επόμενη διετία και παράλληλα η χρήση μοντέλων μηχανικής μάθησης για την πρόβλεψη της κατάθλιψης με βάση τους παράγοντες που προβλέφθηκαν από το νευρωνικό δίκτυο LSTM. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από την Chinese Lognitudinal Healthy Longevity Study (CLHLS) (παρόμοια δομή με την ELSA) και ποιοι συγκεκριμένα εξεταστήκαν τα δεδομένα των waves 3-7 (2002-2014).

Η προβλεπόμενη μεταβλητή είχε δυαδική μορφή (αν υπάρχει κατάθλιψη ή όχι) και για την εκπαίδευση των μοντέλων οποιαδήποτε θετική απάντηση στις παρακάτω ερωτήσεις θεωρήθηκε ως ύπαρξη κατάθλιψης.

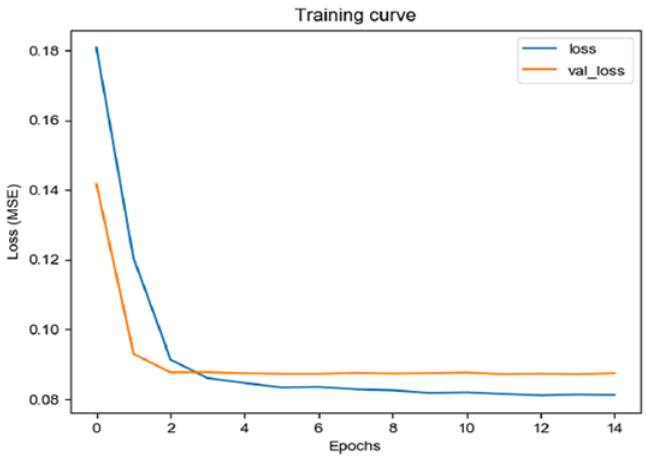
1) Υπήρξε στιγμή τους τελευτούν 12 μήνες που αισθανόσασταν στεναχωρημένος ή καταθλιπτικός για παραπάνω από δυο εβδομάδες
2) Υπήρξε στιγμή τους τελευταίους 12 μήνες που χάσατε το ενδιαφέρον σας σε αντικείμενα που σας αρέσαν (hobbies, εργασία ή άλλες δραστηριότητές)

Για την πρόβλεψη χρησιμοποιήθηκαν γνωρίσματα που υπάγοντες σε τρεις κατηγορίες:

- Δημογραφικά(ηλικία, οικογενειακή κατάσταση, κατοικία, αριθμός ατόμων συμβίωσης)
- Παράγοντες σχετικά με την υγειά (self-rated Health, άθληση, χοληστερίνη, νοητική λειτουργία και κάπνισμα)
- Χρονιές Παθήσεις (Διαβήτης καρδιακές παθήσεις, καταρράκτης, γλαύκωμα, καρκίνος, γαστρικό έλκος, αρθρίτιδα).

Από την πρόβλεψη των παραγόντων από το LSTM μας δίνεται το διάγραμμα της Loss Function στην εικόνα 1 που χρησιμοποιήθηκε από το LSTM μοντέλο και παρατηρείτε ότι το validation loss σταθεροποιήθηκε περίπου 0.09.

Εικόνα 1 MSE LOSS FUNCTION



Τα μοντέλα που χρησιμοποιήθηκαν για την πρόβλεψη της κατάθλιψης είναι: Logistic Regression (LR), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Support Vector Machine (SVM), Deep Neural Network (DNN) ενώ για την αξιολόγηση της αποδοτικότητάς

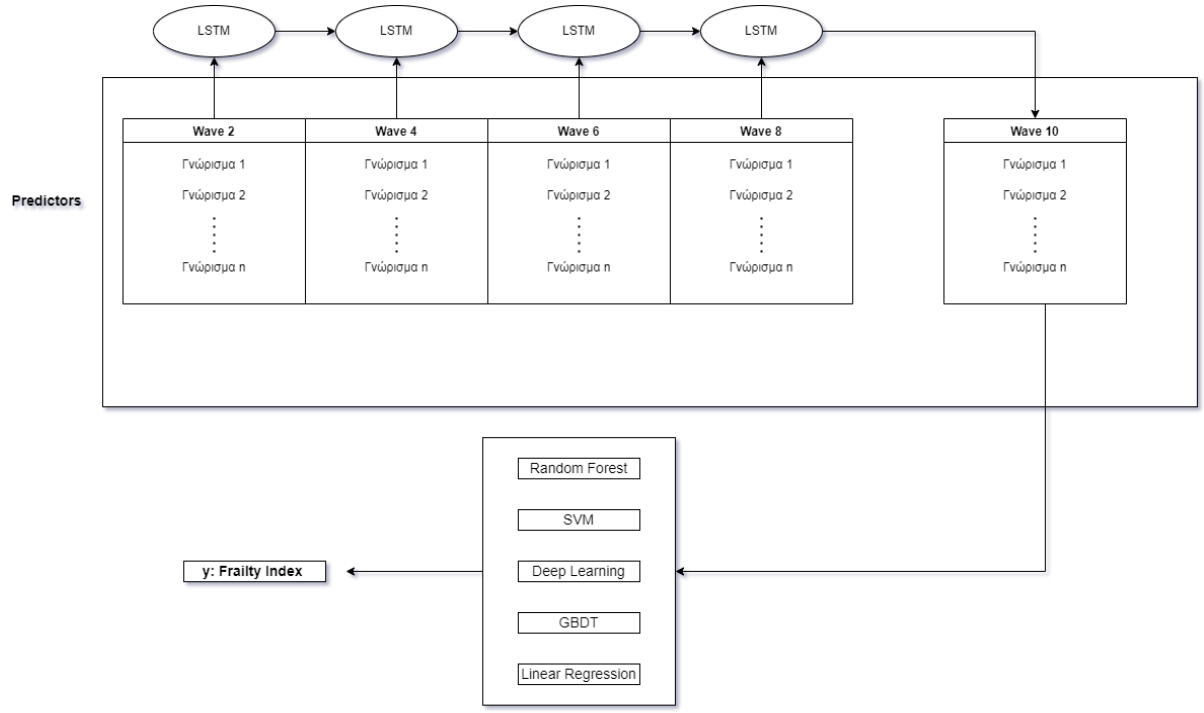
χρησιμοποιήθηκαν τα μετρικά sensitivity, specificity, Accuracy, Positive predictive value (PPV), Negative Predictive Value και Area Under Curve. Επίσης πραγματοποιήθηκε ένα Delong Test στις τιμές της AUC με σημείο αναφοράς το μοντέλο της Logistic Regression. Οι μετρήσεις αυτές περιγράφονται στον πίνακα 2.

Πίνακας 2 (Su, et al.,2021) Αξιολόγηση Πρόγνωσης

Model	AUC	Delong Test	Accuracy	Sensitivity	Specificity	PPV	NPV
LR	0.605	Reference	0.659	0.571	0.677	0.262	0.887
LRLR	0.629	0.020	0.670	0.623	0.680	0.281	0.900
RF	0.589	0.624	0.482	0.753	0.427	0.209	0.896
GBDT	0.630	0.454	0.759	0.429	0.826	0.330	0.878
SVM	0.578	0.545	0.564	0.571	0.563	0.208	0.867
DNN	0.644	0.427	0.571	0.688	0.547	0.234	0.797

ΚΕΦ.3: Προτεινομένη Προσέγγιση και Προ επεξεργασία

Για την επίτευξη της πρόβλεψης του μελλοντικού frailty index θα χρησιμοποιηθούν διαφορετικοί αλγόριθμοι εξόρυξης γνώσεις ώστε να υπάρχει μεταξύ τους σύγκριση ενώ τα γνωρίσματα που θα χρησιμοποιηθούν από αυτά τα μοντέλα θα προβλεφθούν από ένα νευρωνικό δίκτυο LSTM. Το πρώτο βήμα το οποίο απαιτείται να γίνει είναι ο προσδιορισμός των γνωρισμάτων για την πρόβλεψη του frailty index. Στην συνέχεια θα χρειαστεί να εφαρμοστούν τεχνικές preprocessing ώστε με την χρήση του LSTM να προβλεφθούν οι μελλοντικές τους τιμές. Έπειτα θα χρειαστεί να κατασκευαστούν οι frailty indexes για κάθε άτομο του οπίου τα δεδομένα θα χρησιμοποιούν για την εκπαίδευση των μοντέλων πρόβλεψης του frailty index. Αφού έχει κατασκευαστεί και frailty index με βάση τα αποτελέσματα από το LSTM θα γίνει η πρόβλεψη των μελλοντικών τους τιμών. Η αρχιτεκτονική της προτεινομένης προσέγγισης φαίνεται στην εικόνα 7.



Εικόνα 11 Αρχιτεκτονική Πρόβλεψης FI

3.1 Δεδομένα

Το σύνολο των δεδομένων που χρησιμοποιήθηκαν προέρχεται από την μελέτη ELSA (English Longitudinal Study of Ageing) στην οποία οι συμμετέχοντες καλούνται να συμπληρώσου ερωτηματολόγια αλλά και να υποβληθούν σε εξετάσεις. Τα δεδομένα ομαδοποιούνται σε ξεχωριστά dataset ανάλογος με την κατηγορία που ανήκουν. Καθώς η ανάγκες της έρευνας αλλάζουν με τον χρόνο προστίθενται και αφαιρούνται δευτερεύοντα datasets από τα waves της ερευνάς με αποτέλεσμα μερικά να dataset να είναι μοναδικά και αλλά να έχουν δημιουργηθεί σποραδικά αναμεσά στα waves. Τα δεδομένα τα οποία χρησιμοποιούνται σε αυτήν την πτυχιακή εργασία προέρχονται από τα core_data datasets (μετά το wave 2 ονομάζονται elsa_data) τα οποία επαναλαμβάνονται σε κάθε wave της ερευνάς καθώς περιγράφουν το κυρίως ζητούμενο της. Η δεύτερη η κατηγορία δεδομένων που χρησιμοποιήθηκε είναι τα nurse_data τα οποία περιέχουν αποτελέσματα ιατρικών εξετάσεων και ξεκίνησαν να πραγματοποιούνται από 2^{ο} wave και από τότε επαναλαμβάνονται ανά 4 χρονιά. Μια εξαίρεση στον ρυθμό των επαναλήψεων στην κατηγορία nurse έγινε στο 8ο wave καθώς δεν εξεταστικέ ολόκληρο το σύνολο των συμμετεχόντων στο χρονικό περιθώριο του 8^{ου} wave έτσι οι εξετάσεις συνεχίστηκαν στο 9^{ο} wave οπού και ολοκληρωθήκαν. Για αυτόν τον λόγο γίνεται μια παραδοχή ότι τα nurse_data του 9^{ου} wave, τα οποία εμπεριέχουν και τα αποτελέσματα του 8^{ου}, ανήκουν στο 8ο wave. Με βάση τους παραπάνω λόγους η επιλογή των datasets ανά wave για τον στόχο της εργασίας περιγράφεται στον πίνακα 3.

Όνομα Dataset	**Wave of Dataset**	**Όνομα Dataset**	**Wave of Dataset**
Core_data	Wave 2	Elsa_data	Wave 6
Nurse_data	Wave 2	Nurse_data	Wave 6
Elsa_data	Wave 4	Elsa_data	Wave 8
Nurse_data	Wave 4	Nurse_data	Wave 9

Πίνακας 3 Datasets ανά wave

3.1.1 Δεδομένα Frailty Index

Για την δημιουργία του frailty index χρησιμοποιήθηκαν γνωρίσματα τα οποία περιγράφουν προβλήματα υγείας όπως αναπηρίες, ασθενείς, χρόνιες παθήσεις και ελλείματα στην καθημερινή ζωή τα οποία βρίσκονται στα Core/Elsa_Data. Τα προβλήματα αυτά πρέπει να

προστεθούν ώστε να υπάρχουν τουλάχιστον 30-40 σε σύνολο καθώς όσο περισσότερο το πλήθος τους τόσο ποιο ακριβής θα είναι και ο frailty index. Τα γνωρίσματα για να συμπεριληφθούν στην δημιουργία του Frailty Index θα πρέπει να υπακούνε σε 5 κανόνες. (Searle , Mitnitski, Gahbauer, Gill, & Rockwood, 2008)

1) Τα γνωρίσματα θα πρέπει σχετίζονται με την κατάσταση υγείας.
2) Η επικράτηση του γνωρίσματος θα πρέπει να αυξάνεται γενικά με την ηλικία
3) Τα επιλεγμένα προβλήματα δεν θα πρέπει να κορέζονται πολύ νωρίς. Για παράδειγμα η πρεσβυωπία αν και είναι πρόβλημα υγείας παρουσιάζεται σχεδόν καθολικά στα άτομα ηλικίας 55 και άνω.
4) Τα γνωρίσματα θα πρέπει να καλύπτουν ένα ευρύς φάσμα των συστημάτων του ανθρωπίνου οργανισμού. Αν για παράδειγμα όλα τα γνωρίσματα αφορούσαν την νοητική λειτουργία τότε θα είχαμε έναν δείκτη νοητικής λειτουργίας και όχι έναν δείκτη ευθραυστότητας.
5) Αν γίνει σύγκριση δεικτών μεταξύ ατόμων τότε τα γνωρίσματα που καθορίζουν τους δείκτες θα πρέπει να είναι ιδιά. (Searle , Mitnitski, Gahbauer, Gill, & Rockwood, 2008)

Ακολουθώντας του 5 κανόνες αυτούς επιλέχτηκαν 31 γνωρίσματα τα οποία κατηγοριοποιούνται σε 3 ομάδες για τους συμμετέχοντες της ερευνάς.

Η πρώτη ομάδα περιλαμβάνει 14 γνωρίσματα για ελλείματα τα οποία μπορεί να αντιμετωπίζει ο συμμετέχοντας την καθημερινή του ζωή και περιγράφονται από τον πίνακα 4

Όνομα	**Περιγραφή**	**Όνομα**	**Περιγραφή**
Help_Medication	Χρειάζεται βοήθεια για την λήψη φαρμακευτικής αγωγής	Help_mealprep	Χρειάζεται βοήθεια για την προετοιμασία γεύματος
Help_Dressing	Χρειάζεται βοήθεια για να ντυθεί	Help_Shopping	Χρειάζεται βοήθεια για να ψωνίσει

Help_Chair	Χρειάζεται βοήθεια για σηκωθεί από την καρεκλά	Help_10lbs	Χρειάζεται βοήθεια για να σηκώσει 10 λίβρες (4.53 kg)
Help_Walking	Χρειάζεται βοήθεια για να περπατήσει	Help_Houswork	Χρειάζεται βοήθεια στις οικιακές δουλείες
Help_Toilet	Χρειάζεται βοήθεια για χρησιμοποιήσει την τουαλέτα	Help_Finances	Χρειάζεται βοήθεια για την διαχείριση των οικονομικών
Help_Stairs	Χρειάζεται βοήθεια για ανέβει τις σκάλες	Walking 1/4 mile	Αν μπορεί να περπατήσει ¼ του μιλίου (402 m)
Help_Eating	Χρειάζεται βοήθεια για να φάει	Help_Bathing	Χρειάζεται βοήθεια για να κάνει μπάνιο

Πίνακας 4 Κατασκευή FI Γνωρίσματα ομάδα 1

Η δεύτερη ομάδα περιλαμβάνει 7 γνωρίσματα που περιγράφουν την ψυχική και συναισθηματική κατάσταση του συμμετέχοντα και την προσωπική του εκτίμηση της υγείας του και περιγράφονται στον πίνακα 5.

Όνομα	**Περιγραφή**	**Όνομα**	**Περιγραφή**
Feel Depressed	Αν αισθάνεται καταθλιπτικός	Self-Rating of Health	Αυτοαναφερόμενο Score Υγείας
Feel Happy	Αν Αισθάνεται Χαρούμενος	Feel Lonely	Αν αισθάνεται μόνος
ENP_problems	Ύπαρξη Συναισθηματικής, Ψυχικής, Νευρικής Διαταραχής	Have Trouble getting going	Αν έχει κίνητρο να ξεκινήσει την ημέρα του
		Effort	Αν αισθάνεται ότι για τα πάντα χρειάζεται προσπάθεια

Πίνακας 5 Κατασκευή FI Γνωρίσματα ομάδα 2

Η τρίτη ομάδα περιλαμβάνει 10 γνωρίσματα που περιγράφουν χρονιές παθήσεις και προβλήματα με τη φυσική υγειά του συμμετέχοντα και περιγράφονται στον πίνακα 6.

Όνομα	Περιγραφή	Όνομα	Περιγραφή
CHF	*Ύπαρξη Ιστορικού Καρδιακής Ανεπάρκειας*	CLD	*Ύπαρξη/Υπήρξε* Χρόνιας Πνευμονικής Νόσου
Stroke	*Ύπαρξη Ιστορικού Εγκεφαλικού*	Alzheimer	Εκδήλωση Νόσου Alzheimer
Arthritis	*Ύπαρξη/Υπήρξε Ιστορικού Αρθρίτιδας*	Heart_Attack	Ιστορικό Εμφράγματος
Cancer	*Υπάρξει Ιστορικού Καρκίνου*	Diabetes	*Υπάρξει Ιστορικού Διαβήτη*
High_BP	*Ύπαρξη/Υπήρξε Ιστορικού* υψηλής πίεσης	Dementia	*Ύπαρξη/Υπήρξε Ιστορικού* Ιστορικό Άνοιας

Πίνακας 6 Κατασκευή FI Γνωρίσματα ομάδα 3

3.1.2 Κατασκευή Frailty Index Wave 2

Άφωτου ορίστηκε μια κοινή βάση για τα γνωρίσματα στο σύνολο των waves πραγματοποιήθηκε μια διαδικασία preprocessing στα δεδομένα για την μορφοποίηση τους ώστε να μπορεί να υπολογιστεί ο frailty index ανά άτομο.

Στην δημιουργία των dataset η ELSA δεν αποσκοπούσε στην χρήση τους από αλγορίθμους εξόρυξης δεδομένων. Αυτό έχει ως αποτέλεσμα η προ επεξεργασία των δεδομένων να είναι ιδιαίτερα δύσκολη και μέθοδοι που εφαρμόζονται στο ένα wave να μην απαραίτητα εφαρμόσιμοι στο άλλο. Κυρίες διάφορες παρουσιάζουν το wave 2 με τα waves 4,6,8.

Στο wave 2 τα γνωρίσματά τα οποία περιγράφουν ελλείματα ζωής (Πίνακας 4), εκτός από το Walking 1/4 mile, αντιστοιχούν στις επαναλαμβανόμενες μεταβλητές headb01-headb13, head01-10 από το core_data dataset. Για παράδειγμα η ερώτηση με την οποία συμπληρωθήκαν στο dataset οι μεταβλητές headb01- headb014είναι η εξής:

''Στην καθημερινή σας ζωή συναντάτε κάποια από τα παρακάτω προβλήματα εξαετίας κάποιου προβλήματος υγείας/φυσικής κατάστασης''

Με τις παρακάτω αριθμημένες απαντήσεις

Value = 96.0 Label = None of these
Value = 1.0 Label = Walking 100 yards
Value = 2.0 Label = Sitting for about two hours
Value = 3.0 Label = Getting up from a chair after sitting for long periods
Value = 4.0 Label = Climbing several flights of stairs without resting
Value = 5.0 Label = Climbing one flight of stairs without resting
Value = 6.0 Label = Stooping, kneeling, or crouching
Value = 7.0 Label = Reaching or extending arms above shoulder level
Value = 8.0 Label = Pulling/pushing large objects like a living room chair
Value = 9.0 Label = Lifting/carrying over 10 lbs., like a heavy bag of groceries
Value = 10.0 Label = Picking up a 5p coin from a table
Value = -9.0 Label = Refusal
Value = -8.0 Label = Don't know
Value = -1.0 Label = Not applicable

Η ερώτηση επαναλαμβάνεται επιπλέον 12 φορές ή έως ότου εξεταζόμενος συμπληρώσει την τιμή -1= Not applicable που υποδεικνύει ότι δεν υπάρχει κάποιο άλλο έλλειμα να προστεθεί. Στην περίπτωση που απαντηθεί η πρώτη ερώτηση με την τιμή 96 τότε η ερώτηση δεν επαναλαμβάνετε. Αυτό συμβαίνει ώστε να καλυφθούν όλα τα εν δύναμη προβλήματα τα οποία μπορεί να έχουν παρουσιαστεί στην ζωή του συμμετέχοντα.

Για την μετατροπή των δεδομένών στα γνωρίσματα που τέθηκαν στον πίνακα 5. Αντιστοιχήθηκε το κάθε γνώρισμα με την ανάλογη τιμή και μεταβλητή του core_data dataset και περιγράφονται στον πίνακα 7.

Όνομα	Μεταβλητή	Τιμή	Όνομα	Μεταβλητή	Τιμή
Help_Bathing	headb01-headb13	3	Help_10lbs	heada01-head10	9
Help_Dressing	headb01-headb13	1	Help_Shopping	headb01-headb13	9
Help_Chair	heada01-head10	3	Help_Houswork	headb01-headb13	12

Help_Walking	headb01-headb13	2	Help_mealprep	headb01-headb13	8
Help_Eating	headb01-headb13	4	Help_Medication	headb01-headb13	11
Help_Toilet	headb01-headb13	6	Help_Medication	headb01-headb13	13
Help_Stairs	heada01-head10	5			

Πίνακας 7 Μεταβλητές Γνωρίσματων

Από τον ορισμό του Frailty index κάθε γνώρισμα θα πρέπει να παίρνει την τιμή 1 εφόσον εμφανίζεται στο άτομο και την 0 στην αντίθετη περίπτωση, προσθέτοντας τις τιμές μεταξύ τους και διαιρώντας τες με τον αριθμό του συνόλου τους αποδίδεται ο frailty index. Όταν δεν υπάρχει καταχωρημένη τιμή στα core Datasets τότε εισάγετε η τιμή -1 υποδεικνύοντας την έλλειψη της. Έτσι δημιουργήθηκε ένας κανόνας για την μετατροπή των μεταβλητών στην νέα τους μορφή. Ο κανόνας αυτός δέχεται σαν arguments την ενδιαφερομένη τιμή (var_val), το όνομα του γνωρίσματος(col_name), το όνομα της πρώτης επαναλαμβανομένης μεταβλητής(start_col) και το όνομα της τελευταίας επαναλαμβανομένης μεταβλητής(end_col). Έχοντας εισάγει το dataset core_data ως df και δημιουργώντας ένα καινούριο dataframe με όνομα ind_df. Η παρακάτω function μετατρέπει τα δεδομένα στην νέα τους μορφή διατηρώντας τυχών ελλιπείς τιμές.

```
def  headb_classifier(var_val, col_name, start_col, end_col):
   start_point=df.columns.get_loc(start_col)
   end_point=df.columns.get_loc(end_col)

   for i in reversed(range(start_point+1,end_point+1)):
      ind_df.loc[df.iloc[:,i] == -1, col_name] = 0
      ind_df.loc[df.iloc[:,i] == var_val, col_name] = 1

   ind_df.loc[df.iloc[:,start_point] == -1, col_name] = -1
   ind_df.loc[df.iloc[:,start_point] == var_val, col_name] = 1
   ind_df.loc[df.iloc[:,start_point] == 96, col_name] = 0
```

Καλώντας την function με τα ανάλογα arguments για το κάθε γνώρισμα τοποθετεί στο νέο dataframe ind_df το γνώρισμα αυτό στην νέα του μορφή. Πχ.

```
# 1)Help Bathing
headb_classifier(3, 'Help_Bathing', 'headb01', 'headb13')
```

Τα γνωρίσματα τα οποία περιγράφουν χρονιές παθήσεις και προβλήματα με τη φυσική υγειά (Πίνακας 6) συμπεριλαμβανομένου και του γνωρίσματος ENP_Problems (Πίνακας 5)ορίζονται μέσα από τρεις διαφορετικές μεταβλητές στα Core_data περιγράφοντας διαφορετικές πτυχές της εξέλιξης της κάθε νόσου. Κατηγοριοποιούνται σε τρεις ομάδες. Η πρώτα ομάδα εξετάζει εάν η πάθηση στο καταγράφηκε στο προηγούμενο wave. Η δεύτερη ομάδα εξετάζει εάν η πάθηση συνεχίζει στην ζωή του ασθενή στον παρών wave και η τρίτη ομάδα εξετάζει εάν έχει διαγνωσθεί κάποια νέα πάθηση στο παρών wave. Η τρίτη ομάδα ακολουθεί την δομή ερωτήσεων των ελλειμάτων ζωής και για αυτό τον λόγο χρησιμοποιήθηκε η ιδιά function ώστε να απομονωθεί το κάθε γνώρισμα. Έπειτα για να εξεταστούν για γνώρισμα και οι υπόλοιπες ομάδες δημιουργήθηκε ο εξής λογικός κανόνας:

Εάν η ομάδα 1 επιβιώνει την ασθένεια
Ή η ομάδα 2 επιβιώνει την ασθένεια
Ή η ομάδα 3 επιβιώνει την ασθένεια
Τότε το γνώρισμα =1
Διαφορετικά =0
Εάν η ομάδα 1 έχει ελλιπής τιμή
και η ομάδα 2 έχει ελλιπής τιμή
και η ομάδα 3 έχει ελλιπής τιμή
Τότε το γνώρισμα = ελλιπής τιμή

Κάποια γνωρίσματα όπως το εγκεφαλικό επεισόδιο δεν μπορούν να συνεχίζουν να ισχύουν στην παρών στιγμή που λαμβάνεται το ερωτηματολόγιο και αρά δεν υφίσταστε η ομάδα 2. Σε αυτές τις περιπτώσεις ισχύει ο ίδιος κανόνας με την παράλειψη της ομάδας 2.

Στο ερωτηματολόγιο τα γνωρίσματα Self-Rating of Health (Πίνακας 5) και Walking 1/4 mile (πίνακας 4) δεν απαντώνται με δυαδικές τιμές (Ναι, Όχι) αλλά δίνεται κάποια αύξουσα βαθμολόγηση. Αυτό έχει ως αποτέλεσμα το γνώρισμα με την σειρά του να μην αποδίδεται με τις δυαδικοί τιμές (0,1) αλλά και με ενδιάμεσα κλάσματα.Το γνώρισμα Self-Rating of Health αντιστοιχεί στην μεταβλητή Self-reported general health του core_dataset με τις έξεις εν δυνάμει απαντήσεις:

Value = 1.0 Label = excellent,

Value = 2.0 Label = very good,

Value = 3.0 Label = good,

Value = 4.0 Label = fair,

Value = 5.0 Label = or, poor?

Value = -9.0 Label = Refusal

Value = -8.0 Label = Don't know

Value = -1.0 Label = Not applicable

Έτσι δημιουργείτε μια αντιστοιχία τιμών τις μεταβλητής και τιμών του γνωρίσματος που περιγράφεται στον πίνακα 8.

Τιμή Μεταβλητής	Τιμή Γνωρίσματος	Τιμή Μεταβλητής	Τιμή Γνωρίσματος
1	0	4	0.75
2	0.25	5	1
3	0.50	<0	Ελλιπής τιμή(-1)

Πίνακας 8 Αντιστοιχία Τιμών 1 wave 2

Παρομοίως η ερώτηση Walking 1/4 mile έχει την δομή:

Value = 1.0 Label = ...no difficulty,
Value = 2.0 Label = some difficulty,
Value = 3.0 Label = much difficulty,
Value = 4.0 Label = Unable to do this
Value = -9.0 Label = Refusal
Value = -8.0 Label = Don't know
Value = -1.0 Label = Not applicable

Και δημιουργήθηκε η αντιστοιχία που περιγράφεται στον πίνακα 9.

Τιμή Μεταβλητής	Τιμή Γνωρίσματος
1	0
2	0.33
3	0.66
4	1
<0	Ελλιπής τιμή(-1)

Πίνακας 9 Αντιστοιχία Τιμών 2 wave 2

Οι μεταβλητές Effort, Feel Lonely, Have Trouble getting going απαιτούνται με δυαδική μορφή Ναι=1, Όχι=2 με αποτέλεσμα η απαντήσεις να μετατρέπονται στην τιμή 1 και τιμές όχι στην τιμή 0 αντίστοιχος. Παρομοίως και η μεταβλητή Feel Happy απαντάτε δυαδικά αλλά αντιθέτως η θετική απάντηση μετατρέπεται στην τιμή 0 ενώ η αρνητική στην τιμή 1.

Εφόσον έχουν υπολογιστεί όλες οι τιμές των γνωρισμάτων γίνεται μια ανάλυση για το ποσοστό ελλείπων τιμών. Οι έλλειπες τιμές εκφράζονται ως ποσοστό τις 100 για κάθε γνώρισμα προς το σύνολο των συμμετεχόντων και περιγράφεται στον πίνακα 10

Όνομα Μεταβλητής	%Ελλιπής Τιμή	Όνομα Μεταβλητής	%Ελλιπής Τιμή
Help_Bathing	0.00	CHF	0.00
Help_Dressing	0.00	Stroke	0.00
Help_Chair	0.00	Cancer	0.00
Help_Walking	0.00	Diabetes	0.00
Help_Eating	0.00	Arthritis	0.00
Help_Toilet	0.00	CLD	0.00
Help_Stairs	0.00	Alzheimer	0.00
Help_10lbs	0.00	Dementia	0.00
Help_Shopping	0.00	ENP_problems	0.00
Help_Houswork	0.00	Self Rating of Health	1.46
Help_mealprep	0.00	Effort	2.27
Help_Medication	0.00	Feel Depressed	2.26
Help_Finances	0.25	Feel Happy	2.49
High_BP	0.00	Feel Lonely	2.29
Heart_Attack	0.00	Have Trouble getting going	2.36
		Walking 1/4 mile	0.10

Πίνακας 10 Ελλείπεις Τιμές Μεταβλητών wave 2

Παρατηρείτε ότι τα ποσοστά ελλείπων τιμών είναι πολύ μικρά και ένα μεγάλος μέρος από αυτές συναντάται στα ιδιά άτομα. Για αυτό τον λόγο έγινε η επιλογή να διαγραφθούν τα άτομα με ελλιπείς τιμές χωρίς να εφαρμοστεί κάποια τεχνική statistical's imputation. Από το αρχικό σύνολο 9432 ατόμων διαγράφθηκαν 265 άτομα δηλαδή το 2.81% με το τελικό dataset να απαρτίζετε από 9167 άτομα.

Προσθέτοντας τις τιμές κάθε γνωρίσματος για κάθε συμμετέχοντα και διαιρώντας το σύνολο αυτό με το συνολικό αριθμό των γνωρισμάτων υπολογίζεται ο frailty index του κάθε ατόμου και προστίθεται ως νέα στήλη με τον όνομα FI στο dataframe ind_df.

3.1.3 Κατασκευή Frailty Index Wave 4,6,8

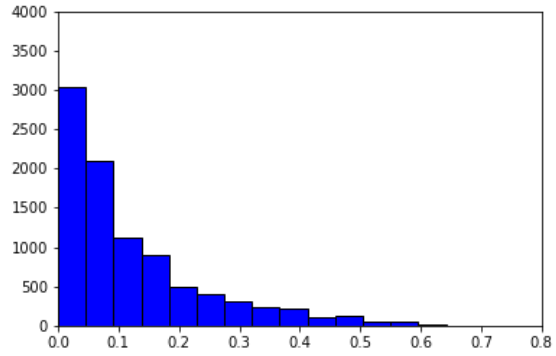
Η κατασκευή του frailty index διαφοροποιείτε ελαφρά και καταστείτε πιο εύκολη στα waves 4,6,8 καθώς αλλάζουνε οι δομές των ερωτήσεων με την κύρια διαφορά να παρατηρείτε στην δομή των ερωτήσεων για τις ομάδες 1 (Πίνακας 4) και 3(Πίνακας 6) των μεταβλητών. Στα waves 4,6,8 η κάθε ερώτηση όσον αφορά αυτές τις ομάδες περιγράφει ένα συγκεκριμένο γνώρισμα και απαντάνε δυαδικά με ναι η όχι. Έτσι δημιουργήθηκε μια αρκετά απλή function η οποία δεχόμενη ως arguments το όνομα της μεταβλητής και το όνομα του γνωρίσματος στο οποίο θα μετατραπεί μεταμορφώνει τις τιμές στην επιθυμητή μορφή.

```
def yesno(var,colname):
    ind_df.loc[df[var] == 1, colname] = 1
    ind_df.loc[df[var] == 0, colname] = 0
    ind_df.loc[df[var] < 0, colname] = -1
```

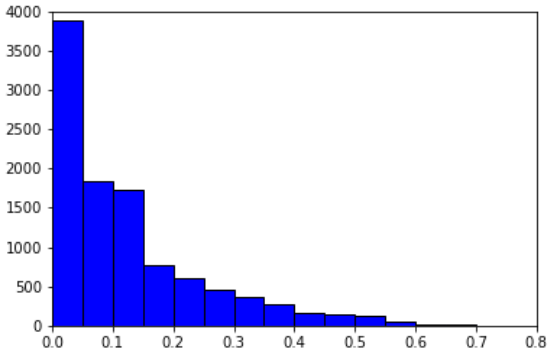
Έπειτα ακολουθώντας τα ιδιά βήματα που περιγράφονται για την κατασκευή του frailty index για το wave 2 δημιουργήθηκαν και οι υπόλοιποι frailty indexes για τα waves 4,6,8 Οι συμμετέχοντες με ελλείπεις τιμές διαγράφθηκαν και από τα υπόλοιπα waves καθώς δεν ξεπερνούσαν το 10%. Τα ιστογράμματα που περιγράφουν τις τιμές των frailty indexes ανά τον πληθυσμό για τα waves 2, 4, 6 και 8 παρουσιάζονται στις εικόνα, 2 εικόνα 3 , εικόνα 4, εικόνα 5 αντίστοιχα ενώ στον πίνακα 11 παρουσιάζονται ο συνολικός αριθμός των ατόμων για τα οποία υπολογιστικέ ο Frailty Index, η κατανομή και η διάμεσος του frailty index σε κάθε Wave.

Αριθμός Wave	Αριθμός Ατόμων	Διάμεσος	Κατανομή
Wave 2	9167	0.119771	0.122890
Wave 4	10478	0.124834	0.128924
Wave 6	9868	0.128841	0.132453
Wave 8	7906	0.133261	0.135000

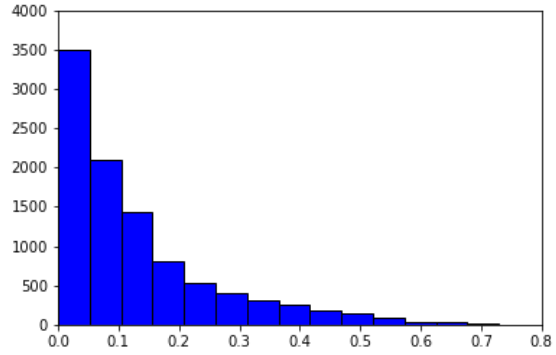
Πίνακας 11 Στοιχεία FI



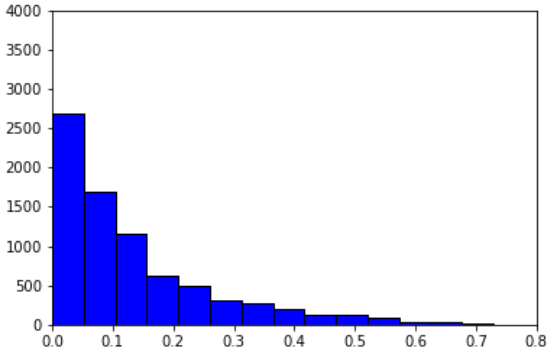
Εικόνα 2 Wave 2 FI



Εικόνα 3 Wave 4 FI



Εικόνα 4 Wave 6 FI



Εικόνα 5 Wave 8 FI

3.2 Προ επεξεργασία Δεδομένων για πρόβλεψη γνωρισμάτων

Για να προβλεφθεί ο frailty index αρχικά θα κατασκευαστεί ένα Long-Short Term Memory (LSTM) Neural Network το οποίο λαμβάνοντας υπόψιν έναν αριθμό προεπιλεγμένων γνωρισμάτων από το παρελθόν θα πραγματοποίηση μια πρόβλεψη των τιμών τους στο μέλλον. Καθώς τα δεδομένα μπορεί να θεωρηθούν ότι είναι χρονομέτρες δηλαδή μεταβάλλονται από wave σε wave ένα νευρωνικό δίκτυο τύπου LSTM θα ήταν ιδανικό για την πρόβλεψη τους. Αυτό συμβαίνει διότι το LSTM έχει μια «κατάσταση» που αποθηκεύει τις πληροφορίες που αφορούν ό,τι έχει παρατηρήσει/επεξεργαστεί μέχρι τώρα, και επεξεργάζεται διαδοχικά δεδομένα μέσω ενός αριθμού επαναλήψεων λαμβάνοντας υπόψιν την προηγούμενος κατάσταση. (Rahman & Adjeroh, 2019)Τα δεδομένα που θα χρησιμοποιηθούν από το LSTM πηγάζουν από τις Ιατρικές Εξετάσεις(nurse_data) που έγιναν στα waves 2,4,6,8 και επιλέχθηκαν με την προϋπόθεση της

ύπαρξη τους και στα 4 αυτά waves. Για αυτό το λόγο εξετάστηκαν τα 4 αυτά datasets και επιλέχθηκαν 24 γνωρίσματα τα οποία περιγράφονται στον πίνακα 12.

Όνομα Γνωρίσματος	Περιγραφή Γνωρίσματος	Όνομα Γνωρίσματος	Περιγραφή Γνωρίσματος
sys1	Μέτρηση 1 Συστολικής Πίεσης (mmHg)	trig	Επίπεδο Τριγλυκεριδίων Αίματος (mmol/L)
sys2	Μέτρηση 2 Συστολικής Πίεσης (mmHg)	ldl	Επίπεδο LDL Χοληστερίνης Αίματος (mmol/L)
sys3	Μέτρηση 3 Συστολικής Πίεσης (mmHg)	rtin	Επίπεδο Φεριτίνης Αίματος (ng/mL)
dias1	Μέτρηση 1 Διαστολικής Πίεσης (mmHg)	hscrp	Επίπεδο C-αντιδρώσα πρωτεΐνης (mg/L)
dias2	Μέτρηση 2 Διαστολικής Πίεσης (mmHg)	hgb	Επίπεδο αιμοσφαιρίνης αίματος(g/dL)
dias3	Μέτρηση 3 Διαστολικής Πίεσης (mmHg)	hba1c	Επίπεδο γλυκοζιωμένης αιμοσφαιρίνη (%)
pulse1	Μέτρηση 1 Παλμών (bpm)	mmgsd1	Μέτρηση 1 Δύναμης Λαβής Μη κυρίαρχο Χέρι (kg)
pulse2	Μέτρηση 2 Παλμών (bpm)	mmgsd2	Μέτρηση 2 Δύναμης Λαβής Μη κυρίαρχο Χέρι (kg)
pulse3	Μέτρηση 3 Παλμών (bpm)	mmgsd3	Μέτρηση 3 Δύναμης Λαβής Μη κυρίαρχο Χέρι (kg)

cfib	Επίπεδο Ινωδογόνου Αίματος (g/L)	mmgsn1	Μέτρηση 1 Δύναμης Λαβής κυρίαρχο Χέρι (kg)
chol	Επίπεδο Συνολικής Χοληστερίνης Αίματος (mmol/L)	mmgsn2	Μέτρηση 2 Δύναμης Λαβής κυρίαρχο Χέρι (kg)
hdl	Επίπεδο HDL Χοληστερίνης Αίματος (mmol/L)	mmgsn3	Μέτρηση 3 Δύναμης Λαβής κυρίαρχο Χέρι (kg)

Πίνακας 12 Περιγραφή Γνωρίσματων

Αφού καθοριστήκαν τα γνωρίσματα για κάθε wave βρέθηκαν τα κοινά άτομα τα οποία έχουν συμμετάσχει και στα 4 waves με τον συνολικό τους αριθμό να είναι 2820. Η επιλογή των ατόμων έγινε με βάση το γνώρισμα 'idauniq' το οποίο είναι ένας μοναδικός αριθμός που ταυτοποιεί κάθε άτομο σε όλα waves. Έτσι δημιουργήθηκαν 4 dataset με διαστάσεις (2820,25). Στην συνέχεια χρειάστηκε να αναλυθούν οι έλλειπες τιμές και να αποφασιστεί εάν θα χρησιμοποιηθεί κάποια τεχνική statistical imputation. Τα ποσοστά ελλείπων τιμών ανά κάθε γνώρισμα παρουσιάζονται στον πίνακα 13.

Όνομα	% Ελλιπής	Όνομα	% Ελλιπής	Όνομα	% Ελλιπής	Όνομα	% Ελλιπής
sys1	0.257693	pulse1	0.257693	trig	23.722904	mmgsd1	3.350008
sys2	0.257693	pulse2	0.257693	ldl	24.602092	mmgsd2	3.350008
sys3	0.257693	pulse3	0.257693	rtin	23.768380	mmgsd3	3.350008
dias1	0.257693	cfib	29.574049	hscrp	23.707746	mmgsn1	3.456116
dias2	0.257693	chol	23.738063	hgb	24.799151	mmgsn2	3.471275
dias3	0.257693	hdl	23.738063	hba1c	25.162953	mmgsn3	3.516750

Πίνακας 13 Ελλείπεις Τιμές Γνωρισμάτων 1

Καθώς τα ποσοστά των ελλείπων τιμών είναι ιδιαίτερα υψηλά στα γνωρίσματα cfi, chol, hdl, trig, ldl, rtin, hscrp, hgb, hba1c επιλέχθηκε να χρησιμοποιηθεί η τεχνική k-nearest neighbor (k-NN) η οποία αντικαθιστά κάθε ελλιπή τιμή με μια τιμή η οποία αποκτάτε από παρόμοιες περιπτώσεις σε όλο το μήκος του dataset.

Τα πιο αξιοσημείωτα χαρακτηριστικά του αλγορίθμου k-NN είναι τα εξής,

1) η καινούργια τιμή είναι η τιμή που έχει πραγματικά εμφανίστηκε. Δεν υπάρχει δευτερεύουσα επεξεργασία.
2) Η κατανομή των αρχικών δεδομένων διατηρείται σύμφωνα με την διακύμανση των δεδομένων (Su, Zhang, He, & Chen, 2021).

Τα δεδομένα χρειάστηκε να κανονικοποιηθούν καθώς το νευρωνικό δίκτυο LSTM είναι επιρρεπής σε απότομες αυξήσεις και μειώσεις της καμπύλης μάθησης του μοντέλου κάτι το οποίο επηρεάζει αρνητικά την σωστή υιοθέτηση βαρών(weights) και πόλωσης (bias). (Beeksma, et al., 2019). Για τον λόγο αυτό χρησιμοποιήθηκε από την βιβλιοθήκη της Python sklrean η function MinMaxScaler. Η function MinMaxScaler κανονικοποιεί τα δεδομένα σε κάθε γνώρισμα με βάση την μεγαλύτερη και μικρότερη τιμή σε αυτό το γνώρισμα αναθέτοντας τους τιμές από το 0 έως το 1, εφόσον δεν παραμετροποιηθεί διαφορετικά. Καθώς υπάρχουν 4 διαφορετικά dataset, για να δημιουργηθεί ένας κοινός Scaler, τα dataset συνενωθήκαν (concatenated) σε ένα νέο και ο Scaler έγινε fit σε αυτό. Αφού πραγματοποιήθηκε το fit στα δεδομένα τότε σε κάθε ένα από τα 4 dataset πραγματοποιήθηκε ο μετασχηματισμός τους (transform) με τον Scaler. Ο Scaler αποθηκευτικέ σε ένα αρχείο .plk για μελλοντική χρήση και για να από-μετασχηματιστούν τα δεδομένα που θα παράγει το LSTM.

ΚΕΦ.4: Πειράματα

4.1 Υλοποίηση Μοντέλου Πρόβλεψης γνωρισμάτων

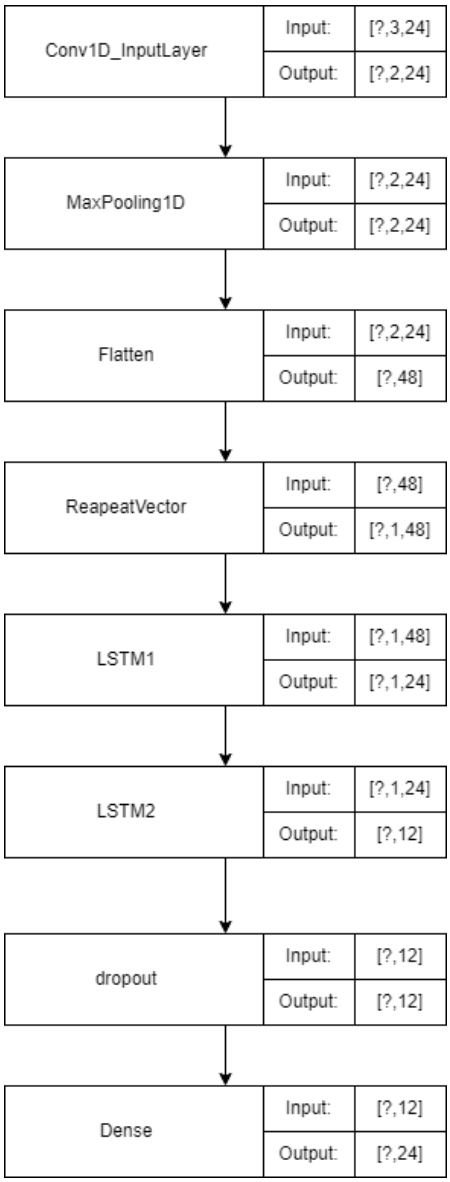
Τα δεδομένα που παρέχονται από τις διαχρονικές μελέτες μπορούν να θεωρηθούν ως χρονοσειρές διότι μεταβάλλονται στην διάσταση του χρόνου. Στην περίπτωση της ELSA το κάθε wave συμβολίζει μια διαφορετικοί χρονική στιγμή. Για αυτόν τον λόγο το μοντέλο πρόβλεψης που θα χρησιμοποιηθεί θα πρέπει να είναι ικανό να διαχειριστεί την επιπλέον πολυπλοκότητα των χρονοσειρών αυτόν. Μια κατηγορία νευρωνικών δικτυών που μπορούν να διαχειριστούν της ακολουθίες αυτές είναι τα Reccurant Neural Networks (RNNs) από τα οποία το LSTM ξεχωρίζει για τους σκοπούς της παρούσας πτυχιακής είναι καθώς καθιστά εφικτή και αποδοτική την εκπαίδευση μεγάλων αρχιτεκτονικών. (Brownlee, 2020)

Για την εισαγωγή των δεδομένων στο Νευρωνικό δίκτυο απαιτητέ η μετατροπή των 4 dataset που περιγράφονται στην παράγραφο 3.2 σε ένα νέο τρισδιάστατο dataset ώστε τα γνωρίσματα του κάθε συμμετέχοντα να εμφανίζονται ως χρονοσειρές. Ως αποτέλεσμα τα 4 datasets με διαστάσεις 2820,24 μετατρέπονται σε ένα νέο array με διαστάσεις 2820,4,25. Έπειτα το dataset διαχωρίζετε σε δυο arrays στην είσοδο x και στην έξοδο y και με την σειρά τους αυτά τα δυο arrays χωρίζονται, με την function της βιβλιοθήκης sklearn train_test_split(), σε 4 νέα arrays X_train, X_test, y_train, y_test, με τα δυο train να είναι το 80% του αρχικού τρισδιάστατου dataset, και τα τεστ το 20% ώστε να αξιολογηθεί μετέπειτα το μοντέλο.

Για την δημιουργία του μοντέλου χρησιμοποιήθηκε η αρχιτεκτονική κωδικοποιητή - αποκωδικοποιητή με την χρήση ενός Convolution Layer ως την είσοδο ώστε να εξαχθούν τα γνωρίσματα. Η αρχιτεκτονική αυτή είναι μια καλή λύση στο πρόβλημα της μάθησης με ακολουθίες καθώς το κάθε γνώρισμα για τον κάθε συμμετέχοντα αναπαραστήστε σαν μια ακολουθία στον χρόνο (Halil, 2021). Το σύνολο των Layers ο αριθμός των νευρώνων αλλά και οι activation functions επιλέχθηκαν μετρά από αρκετό πειραματισμό συγκρίνοντας κάθε φορά τις απόδοσης μοντέλου. Έτσι τελικά δημιουργήθηκε ένα μοντέλο που απαρτίζεται από 1 Conv1D layers, 1MaxPooling1D layer, 1 Flatten layer, 1 RepeatVector Layer, 2 LSTM layers, 1 Dropout Layer και ένα Dense Layer .Το μοντέλο αυτό θα εκπαιδευτεί προσπαθώντας να προβλέψει τα γνωρίσματα του wave 8. Η ακριβής δομή εμφανίζεται στην εικόνα 6 ενώ διάγραμμα για την ακριβής περιγραφή των εισόδων εξόδων του κάθε layer εμφανίζεται στην εικόνα 7

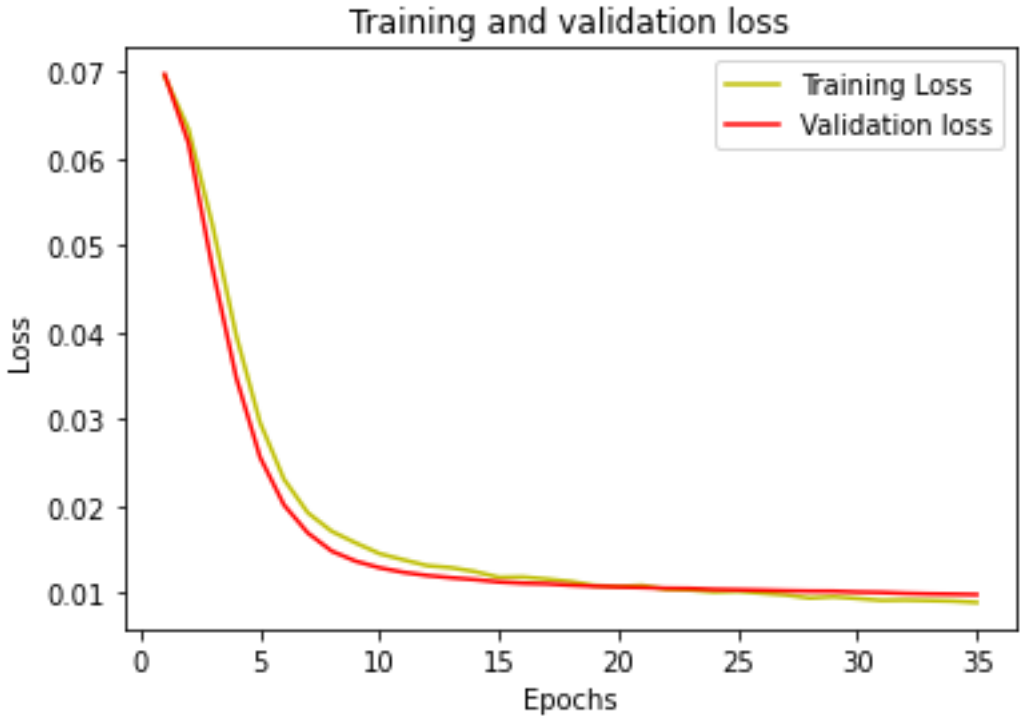
```
model= Sequential()
model.add(Conv1D(filters=24, kernel_size=2, activation='relu', input_shape=(X_train.shape[1],X_train.shape
model.add(MaxPooling1D(pool_size=1))
model.add(Flatten())
model.add(RepeatVector(1))
model.add(LSTM(24,return_sequences=True,activation='relu'))
model.add(LSTM (12,activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(X_train.shape[2], activation='linear'))
model.compile(loss = "mse", optimizer = opt , metrics=['mae'])
```

Εικόνα 6 Δημιουργία Μοντέλου Προβλεψης Γνωρισμάτων



Εικόνα 7 Διάγραμμά Εισόδων Εξόδων Layers

Στο μοντέλο φάνηκε να αποδίδει καλυτέρα ο adam optimizer με ένα μικρό learning rate ίσο με 0.00005 ενώ για Loss Function οριστικέ το Mean Squared Error. Στην συνέχεια το μοντέλο έγινε fit στα arrays X_train και y_train με 32 epochs και το batch_size οριστικέ ίσο με 8. Από το fit του μοντέλου εξάχθηκαν οι σχεδιαστήκαν σε ένα διάγραμμα οι καμπύλες της Loss Function (εικόνα 8) το οποίο υποδεικνύει ένα αρκετά καλό fit του μοντέλου στα δεδομένα. Το training loss μειώνετε έως ότου φτάσει σε ένα σημείο σταθερότητας με παρόμοια συμπεριφορά να ακολουθείται και από το Validation Loss ενώ η διαφορά στο τελευταίο epoch ισούται με 0.0001 (Brownlee, 2019).



Εικόνα 8 Διάγραμμα Loss Function Μοντέλου Προβλεψης Γνωρισμάτων

4.2 Υλοποίηση Μοντέλων Πρόβλεψης Frailty index

Για την τελική πρόβλεψη του frailty index θα δοκιμαστούν 4 διαφορετικά regression μοντέλα ώστε να αξιολογηθούν οι αποδόσεις τους. Τα μοντέλα που θα δοκιμαστούν είναι:

- Random Forest
- SVM
- Deep Neural Network

- GBDT

Η αξιολόγηση τους θα γίνει με την σύγκριση τους με ένα απλό Linear Regression Model. Για την εκπαίδευση των μοντέλων τα δεδομένα που θα χρησιμοποιηθούνε περιλαμβάνουν τα γνωρίσματα για όλο το σύνολο των ατόμων που υπολογίστηκε ο frailty index και συμμετείχαν στις ιατρικές εξετάσεις και όχι μόνο τα κοινά άτομα αναμεσά στα Waves 2,4,6,8 που επιλέχτηκαν για την εκπαίδευση του μοντέλου πρόβλεψης γνωρισμάτων. Επίσης ο κάθε συμμετέχοντας της ερευνάς και τα δεδομένα του εμφανίζονται στο dataset αυτό όσες φορές αυτός συμμετείχε στο κάθε wave της ELSA καταλήγοντας έτσι σε ένα dataset με διαστάσεις 29912, 25 (αριθμός ατόμων ,γνωρίσματα + frailty index). Με αυτό τον τρόπο αυξάνονται κατά μεγάλο βαθμό τα διαθέσιμα δεδομένα με αποτέλεσμα την αύξηση της απόδοσης των μοντέλων καθώς καλούνται να βρούνε μοτίβα της σχέσεως των γνωρισμάτων με τον frailty index.

Καθώς τα δεδομένα αυτά αποτελούν ένα μεγαλύτερο σύνολο των δεδομένων που χρησιμοποιήθηκαν για το μοντέλο πρόβλεψης γνωρισμάτων παρουσιάζουν παρόμοια ποσοστά ελλείπων τιμών με τις ακριβείς τιμές τους να παρουσιάζονται στον πίνακα 14. Η στήλη με τον frailty index εξαιρείται καθώς άτομα για τα οποία δεν υπολογιστικέ δεν συμμετάσχουν στο dataset (παράγραφος 3.1.1) και επόμενος είναι 0.

Όνομα	% Ελλιπής	Όνομα	% Ελλιπής	Όνομα	% Ελλιπής	Όνομα	% Ελλιπής
sys1	0.618481	pulse1	0.628510	trig	24.117411	mmgsd1	2.657796
sys2	0.725461	pulse2	0.715432	ldl	25.528216	mmgsd2	2.664482
sys3	0.842471	pulse3	0.845814	rtin	24.117411	mmgsd3	2.674512
dias1	0.621824	cfib	26.427521	hscrp	24.127440	mmgsn1	2.878443
dias2	0.722118	chol	24.114068	hgb	25.140412	mmgsn2	2.881787
dias3	0.842471	hdl	24.160872	hba1c	25.200588	mmgsn3	2.901845

Πίνακας 14 Ελλείπεις Τιμές Γνωρισμάτων 2

Έτσι εφαρμόστηκε στο σύνολο των δεδομένων και πάλι η τεχνική της KNN imputation ώστε να εξαλειφθούν οι ελλείπεις τιμές.

4.2.1 Random Forest

Τα random forest μοντέλα χρησιμοποιούνται για προβλήματα classification kai regression. Ρέουν οπτικά ως δέντρα και ξεκινούν από την ριζά του δέντρου ακολουθώντας τα μεταβλητά

αποτελέσματα μέχρι να επιτευχθεί ένας κόμβος φύλου και να δοθεί το αποτέλεσμα. (Beheshti, 2019) Για την πρόβλεψη με το random Forest model δεν εφαρμόστηκε κάποια τεχνική scaling καθώς δεν φάνηκε να επηρεάζει την απόδοση του. Έπειτα χρησιμοποιήθηκε η function train_test_split() για την κατανομή των δεδομένων σε 4 arrays: X_train, y_train, X_test, y_test.

Το μοντέλο εισάχθηκε από την βιβλιοθήκη sklearn και εκπαιδευτικέ με τα δεδομένα X_train και y_train.

```
rf= RandomForestRegressor()
rf.fit(X_train, y_train)
```

Στην συνέχεια έγινε μια πρόβλεψη με το X_test και το αποτέλεσμα συγκρίθηκε με το y_test για την δημιουργία μετρικών της πρόβλεψης .

4.2.2 Gradient Boosted Regression

Το Gradient Boosting είναι μια μέθοδος πρόβλεψής που αποτελείται από σύνολά άλλο πιο αδυνάμων μοντέλων κύριος δέντρα αποφάσεων (Masui, 2019). Για την πρόβλεψη με Gradient Boosted Regression χρησιμοποιήθηκε η ιδιά τεχνική με την train_test_split function και δεν χρησιμοποιήθηκε κάποιος scaler.

```
params = {"n_estimators": 500, "max_depth": 4, "min_samples_split": 5, "learning_rate": 0.01,
   "loss": "squared_error",}
reg = ensemble.GradientBoostingRegressor(**params)
reg.fit(X_train, y_train)
```

Στην συνέχεια έγινε μια πρόβλεψη με το X_test και το αποτέλεσμα συγκρίθηκε με το y_test για την δημιουργία μετρικών της πρόβλεψης .

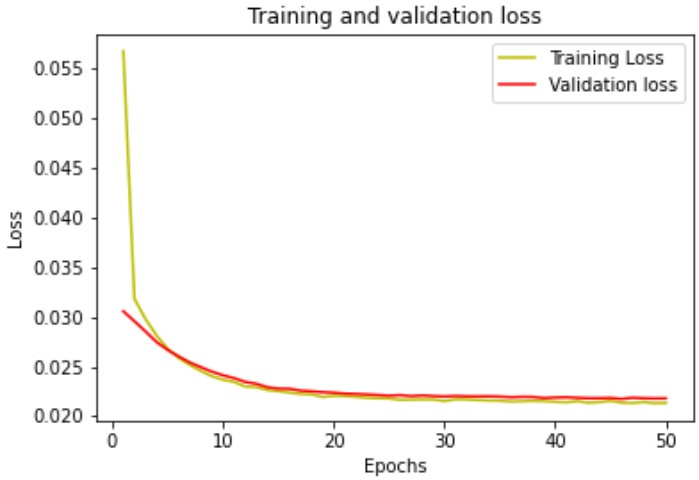
4.2.3 Deep Neural Network

Για την δημιουργία του Deep Neural Network model δημιουργήθηκε ένα απλό μοντέλο το οποίο αποτελείται από 3 Dense Layers και ένα dropout. Τα δεδομένα πριν εισαχθούν κανονικοποιήθηκαν με την χρήση του MinMaxScaler() και διαχωριστήκαν σε 4 datasets με την

function test_train_split() με τα train datasets να αποτελούν το 95% του σύνολού των δεδομένων. Η ακριβής δομή του μοντέλου περιγράφεται στον κώδικα

```
model = Sequential()
model.add(Dense(24,  activation='relu',input_dim=24))
model.add(Dense(5, activation='relu'))
model.add(Dropout(0.2))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer=opt)
```

Για την καλύτερη απόδοση του χρησιμοποιήθηκε ένα χαμηλό Learning Rate = 0.00005 Η εκπαίδευση του μοντέλο παρήγαγε το γράφημά της Loss Function που εμφανίζεται στην εικόνα 9



Εικόνα 9 Διάγραμμα Loss Function Μοντέλου DNN προβλεψης FI

Μετά την εκπαίδευση του το μοντέλο χρησιμοποιήθηκε για να κάνει μια πρόβλεψη με το dataset X_test οπού η εξόδους συγκρίθηκε με το y_test για την δημιουργία μετρικών αξιολόγησης

4.2.4 Support Vector Machine

Ο στόχος του μοντέλου Support Vector Machine είναι να βρει ένα υπερεπίπεδο σε αν χώρο n-διαστάσεων (n=αριθμός γνωρισμάτων) και να τα ταξινομήσει σε ευδιάκριτά σημεία για την εξέταση της απόδοσης ενός Support vector machine επιλέχθηκε. Για την πρόβλεψη με το

support Vector Machine επιλέχθηκε ο αλγόριθμος SVR (Support Vector Regression) και δεν εφαρμόστηκε κάποια τεχνική scaling καθώς δεν φάνηκε να επηρεάζει την απόδοση του. Έπειτα χρησιμοποιήθηκε η function train_test_split() για την κατανομή των δεδομένων σε 4 arrays: X_train, y_train, X_test, y_test. Το μοντέλο εκπαιδεύτηκε με τα δεδομένα X_train και y_train.

```
from sklearn.svm import SVR
regressor = SVR(kernel = 'rbf')
regressor.fit(X_train, y_train)
```

Μετά την εκπαίδευση του το μοντέλο χρησιμοποιήθηκε για να κάνει μια πρόβλεψη με το dataset X_test οπού η εξόδους συγκρίθηκε με το y_test για την δημιουργία μετρικών αξιολόγησης.

4.2.5 Linear Regression

Ως σημείο αναφοράς για την αξιολόγησης των παραπάνω μοντέλων δημιουργήθηκε και ένα μοντέλο το οποίο χρησιμοποιεί την γραμμική παλινδρόμηση για την πρόβλεψη. Χρησιμοποιήθηκαν τα ιδιά δεδομένα με τα υπόλοιπα μοντέλα και έγινε διαίρεση τον δεδομένων σε μικρότατα σύνολα με την χρήση της function train_test_split(). Στο τέλος και πάλι έγινε μια πρόβλεψη με το σύνολο δεδομένων X_test για την δημιουργία μετρικών αξιολόγησης

ΚΕΦ.5: Αποτελέσματα και Συγκρίσεις Μοντέλων

5.1 Μετρικά Αξιολογήσεις

Για την αξιολόγηση των μοντέλων καθώς είναι όλα regression χρησιμοποιήθηκαν 5 μετρικά τα οποία συγκρίνουν ένα αποτέλεσμα πρόβλεψης με ένα πραγματικό αποτέλεσμα. Αυτό όπως περιγράφεται και παραπάνω με την χρήση της function train_test_split επιτυγχάνετε με την διαίρεση των δεδομένων κρατώντας ένα μικρό σύνολο το οποίο δεν εισάγεται στο μοντέλο. Έτσι οι τιμές αυτές παραμένουν άγνωστες στο μοντέλο και η πρόβλεψη που θα πραγματοποίηση αναπαριστά τις πραγματικές του δυνατότητες. Τα μετρικά που χρησιμοποιήθηκαν είναι το Mean Squared Error ,Mean Absolute Error, το R^2και το AdjustedR^2

- Το Mean Squared Error (mse) είναι ένα μέτρο το οποίο το οποίο υπολογίζει την τετραγωνισμένη αποστάτη μιας προβλεπόμενης τιμής με την πραγματική
  Mean Squared Error = $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2$
  n= αριθμός προβλέψεων, x_i= πραγματική τιμή, y_i=προβλεπόμενη τιμή
  Υπολογίζεται με την χρήση της βιβλιοθήκης sklearn.metrics με την function
  mse = mean_squared_error(y_true, y_pred)

- Το Mean Absolute Error (mae) είναι ένα πολύ απλό μετρό και το οποίο υπολογίζει την απολυτή διαφορά ανάμεσα στο προβλεπόμενο και πραγματικό dataset.
  Mean Absolute Error= $\frac{1}{n}\sum_{i=}^{n}|x_i - y_i|$
  n= αριθμός προβλέψεων, x_i= πραγματική τιμή, y_i=προβλεπόμενη τιμή
  Υπολογίζεται με την χρήση της βιβλιοθήκης sklearn.metrics με την function
  mae = mean_absolute_error(y_true, y_pred)

- Το R^2 είναι ένα πολύ χρήσιμο μέτρο καθώς υπολογίζει την αναλογία της διακύμανσης μεταξύ της πραγματικής και προβλεπόμενης τιμής.
  $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{x_i})^2}$

n= αριθμός προβλέψεων, x_i= πραγματική τιμή, y_i=προβλεπόμενη τιμή, $\bar{x}_i$=μέσος ορός πραγματικών τιμών.

Υπολογίζεται με την χρήση της βιβλιοθήκης sklearn.metrics με την function

Rsquared= r2_score (y_true, y_pred)

- AdjustedR^2 είναι μια παραλλαγή του R^2 καθώς λαμβάνει υπόψιν του και τον αριθμό των γνωρισμάτων καθιστώντας το πιο εύστοχο.

$$adjusted\mathrm{R}^2 = 1 - \frac{(1 - \mathrm{R}^2)(\mathrm{n} - 1)}{\mathrm{n} - \mathrm{p} - 1}$$

n= αριθμός προβλέψεων, p= αριθμός γνωρισμάτων

Καθώς δεν υπάρχει βιβλιοθήκη στην python υπολογιστικέ με τον τρόπο:

Adjusted2=1-(((1- Rsquared)*(564-1))/((564-24-1)))

5.2 Αξιολόγηση μοντέλου πρόβλεψης γνωρισμάτων

Για την αξιολόγηση του μοντέλου πρόβλεψης γνωρισμάτων δημιουργήθηκε άλλο ένα deep neural Network (benchmark model). Το νευρωνικό δυτικό αυτό αποτελείται από 1 Convolution1D layer, 1 Flatten layer και ένα Dense layer.

```
model= Sequential()
model.add(Conv1D(filters=24, kernel_size=2, activation='relu',
input_shape=(X_train.shape[1],X_train.shape[2])))
model.add(Flatten())
model.add(Dropout(0.1))
model.add(Dense(X_train.shape[2], activation='linear'))
model.compile(loss = "mse", optimizer = opt , metrics=['mae'])
```

Καθώς σκοπός του είναι η δημιουργία ενός μέτρου σύγκρισης δεν χρειάζεται να έχει κάποια δύναμη πρόβλεψης και εξαιτίας αυτού δικαιολογείται και η απλοϊκότητα του**.** Τα δεδομένα τα οποία χρησιμοποιήθηκαν είναι ακριβώς της ιδιά μορφής που χρησιμοποιήθηκαν και στο μοντέλο πρόβλεψης γνωρισμάτων. Με την χρήση της function test_train_split() όπως περιγράφεται και στο κεφάλαιο 4.1 εκπαιδευτικέ και έγινε μια πρόβλεψη κι έτσι παράχθηκαν τα μετρικά αξιόλογης η σύγκριση των δυο μοντέλων γίνεται στον πίνακα 15

Μοντέλο	MSE	MAE	R^2	$adjusted\mathrm{R}^2$
Μοντέλο πρόβλεψης γνωρισμάτων	0.006	0.042	0.014	0.013
Benchmark	0.008	0.047	-180	-182

Πίνακας 14 Σύγκριση Μοντέλων Πρόβλεψης Γνωρισμάτων

Σύμφωνα με τα αποτελέσματα των μετρήσεων από τον πίνακα 14 η δύναμη πρόβλεψης του μοντέλου γνωρίσματων είναι αρκετά μειωμένη και σύμφωνα με το R^2 και το $adjusted\mathrm{R}^2$ η πρόβλεψη που έγινε σε έναν πάρα πολύ μικρό βαθμό να εξηγήσει την διακύμανση του πραγματικού συνόλου δεδομένων. Σε σύγκριση όμως με το μοντέλο benchmark φαίνεται να αποδίδει ελαφρώς καλυτέρα στα μετρικά MSE και MAE ενώ είναι σαφώς καλύτερο R^2 και το $adjusted\mathrm{R}^2$ οπού το μοντέλο αξιολογήσεις υστερεί κατά μεγάλο βαθμό. Συνήθως οι αρνητικές τιμές υποδεικνύουνε ότι το μοντέλο δεν έχει καταφέρει να περιγράψει το αποτέλεσμα με τα δεδομένα της εκπαίδευσης του.

5.3 Αξιολόγηση μοντέλων πρόβλεψης Frailty Index

Και τα 5 μοντέλα που δημιουργήθηκαν εκπαιδευτήκαν με την χρήση όλων τον πραγματικών διαθέσιμων δεδομένων εκτός από ένα 5% το οποίο θα χρησιμοποιηθεί για την πρώτη αξιολόγηση τους. Ο πίνακας 16 παρουσιάζει τα μετρικά τα οποία αποκτήθηκαν με μια πρόβλεψη με την χρήση 5% συνόλου δεδομένών με bold να είναι τα καλυτέρα αποτελέσματα.

Μοντέλο	**MSE**	**MAE**	R^2	$adjusted\mathrm{R}^2$
LR	0.014	0.086	-3.016	-3.019
SVM	0.013	0.087	0.198	0.197
RF	**0.012**	**0.082**	0.247	0.247
GBDT	**0.012**	0.081	**0.250**	**0.250**
DNN	**0.012**	**0.082**	0.189	0.188

Πίνακας 15 Σύγκριση Μοντέλων Πρόβλεψης FI

Από τον πίνακα 16 παρατηρείται ότι σε σχέση με το μοντέλο της Γραμμικής Παλινδρόμησης (LR) τα υπόλοιπα απέδωσαν αρκετά καλύτερα ειδικότερα στα μετρικά R^2 και το $adjusted\mathrm{R}^2$ δείχνοντας έτσι ότι οι τεχνικές μηχανικής μάθησης μπορούν να συνεισφέρουν σημαντικά στον κλάδο της υγείας.

5.4 Αξιολόγηση συνολικού αποτελέσματος

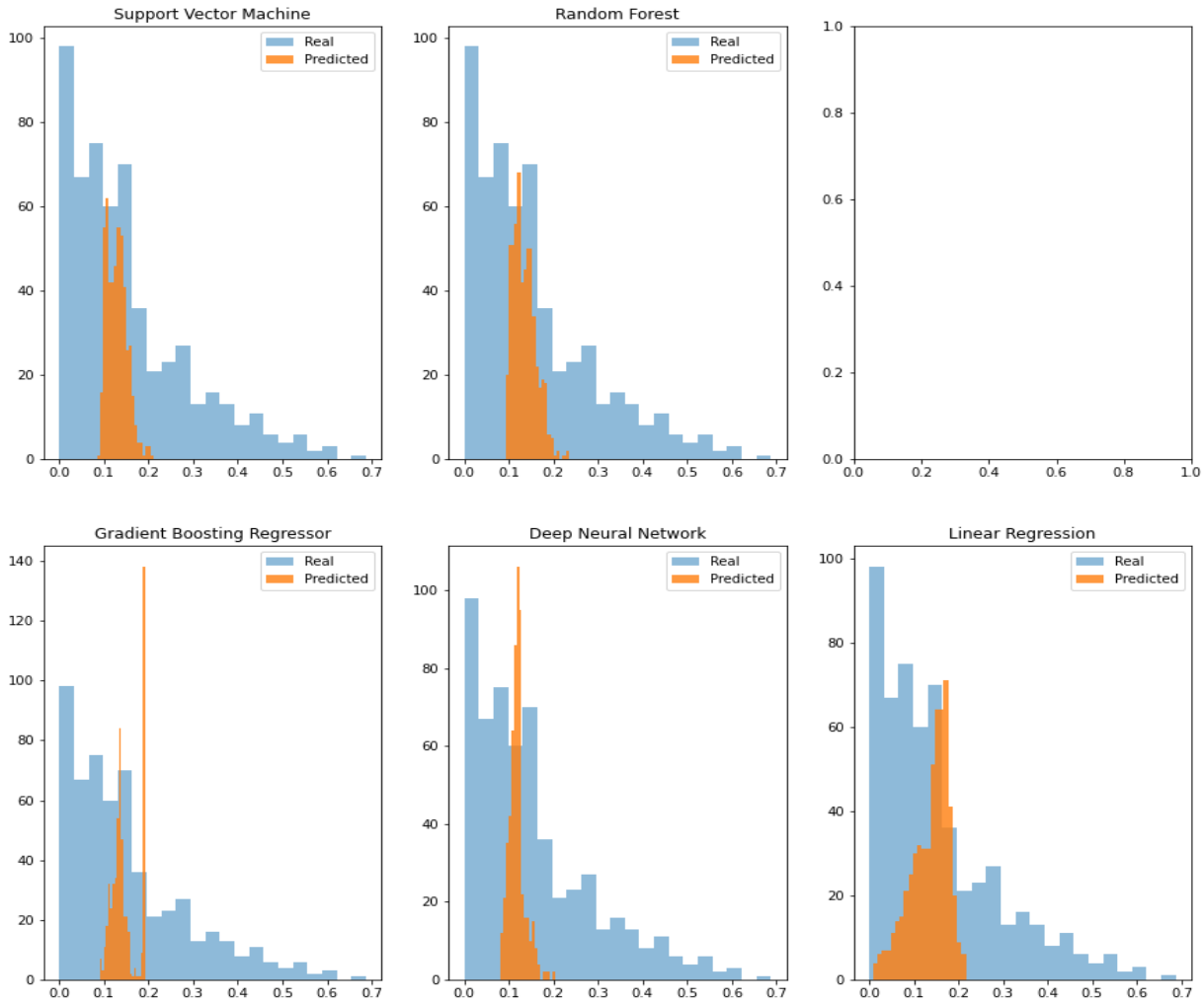
Για την αξιολόγηση του συνολικού αποτελέσματος της εξόρυξης γνώσης από επαναλαμβανόμενες μελέτες παρακρατήθηκε από την εκπαίδευση του μοντέλου πρόβλεψης γνωρισμάτων ένα συνόλου δεδομένων το οποίο ισούται με το 20% δηλαδή 564 δείγματα του γενικού dataset το οποίο προσδιορίζεται στο κεφάλαιο 4.1.

Με την χρήση αυτού του dataset γίνεται αρχικά μια πρόβλεψη της εξέλιξης των γνωρισμάτων για το wave 8 και στην συνέχεια τα αποτελέσματα αυτά αποκανωνικοποιούνται. Έπειτα στο κάθε μοντέλο πρόβλεψης του Frailty Index, αφού εκπαιδευτήκαν βάση του κεφαλαίου 4.2, εισάγονται τα αποτελέσματα του πρώτου μοντέλου και το καθένα υπολογίζει τους frailty indexes στο wave 8 για τα 564 αυτά δείγματα. Για την παραγωγή μέτρων αξιολόγησης εντοπίζετε ο πραγματικός frailty index στο wave 8 για τα συγκεκριμένα 564 αυτά δείγματα. Έχοντας τις πραγματικές τιμές και αυτές που προβλέφθηκαν παράγονται και πάλι τα μετρικά mse,mae, R^2 και $adjusted\mathrm{R}^2$ των οποίων οι τιμές φαίνονται στον πίνακα 17

Μοντέλο	MSE	MAE	R^2	$adjusted\mathrm{R}^2$
LR	0.021	0.109	-0.140	-0.191
SVM	0.019	0.103	-0.065	-0.113
RF	0.019	0.103	-0.058	-0.105
GBDT	0.019	0.104	-0.043	-0.090
DNN	0.020	0.102	-0.104	-0.153

Πίνακας 16 Συνολικά μετρικά τεχνικής πρόβλεψης

Η κατανομή των frailty index ανά πληθυσμό για κάθε μοντέλο σε σύγκριση με τους πραγματικούς περιγράφεται από τα διαγράμματα στην εικόνα 10.



Εικόνα 10 Κατανομή Πληθυσμού ανά FI σύγκριση με wave 8

Η συνολική εικόνα του μοντέλου δεν θα μπορούσε να θεωρηθεί ικανοποιητική καθώς μπορεί να παρατηρηθεί από τα R^2 και $adjusted\mathrm{R}^2$ ότι τα αποτελέσματα του μοντέλου πρόβλεψης γνωρισμάτων δεν είναι ικανά να περιγράψουνε των frailty index για κάθε άτομο. Ειδικότερα οι κατανομές των δεδομένων περιορίζονται στον μέγιστο frailty index =0.2.

ΚΕΦ.6: Συμπεράσματα

Σε αυτήν την πτυχιακή εργασία αποπειράθηκε μια μέθοδος για την εξόρυξη γνώσης από διαχρονικές μελέτες υγείας. Τα δεδομένα προήλθαν από την ερευνά της English Longitudinal Study of Ageing οι οποία συλλεγεί δεδομένα σχετικά με την υγιή γήρανση με πάνω από 18000 συμμετέχοντες

Πριν γίνει η εξόρυξη απαιτήθηκε να υποστούν κάποια προ εξεργασία το οποίο κόστισε και τον περισσότερο χρόνο. Σημαντικό είναι ότι εξαιτίας των αλλαγών στα ερωτηματολόγια και στις εξετάσεις ανά wave αλλά και την έλλειψη επανα-συμμετοχής των ατόμων σε κάποια waves τα χρήσιμα δεδομένα μειωθήκαν σε μεγάλο βαθμό και ειδικότερα αυτά που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου πρόβλεψης γνωρισμάτων.

Το μοντέλο πρόβλεψης γνωρίσματων το οποίο χρησιμοποιείτε στον πυρήνα του 2 Layers LSTM ώστε να συνυπολογιστεί και η χρονική σειρά των γνωρίσματων φάνηκε να αποδίδει σχετικά καλύτερα από ότι ένα απλό νευρωνικό δίκτυο. Σαν απόλυτη μέτρηση οι εκτιμήσεις των γνωρίσματων φαίνεται να απέχουν αρκετά από τις πραγματικές κάτι το οποίο θα μπορούσε να αποδοθεί και στην έλλειψη δεδομένων και κυρίως στην έλλειψη περισσότερων Wave. Στην έρευνα αυτή ήταν διαθέσιμα 4 waves από τα οποία τα 3 χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου δίνοντας έτσι μικρές χρονοσειρές μήκους 3.

Τα μοντέλα πρόβλεψης του frailty index απέδωσαν σε σημαντικά καλύτερο βαθμό από ότι το μοντέλο αναφοράς LR. Κατάφεραν να εξηγήσουν σε πολύ μεγαλύτερο βαθμό το frailty index βάση των ιατρικών δεδομένων που τους δοθήκαν. Πρέπει να αναφερθεί όμως ότι ο frailty index δεν στηρίζεται μονό σε ιατρικά δεδομένα και ίσως μια πιο ενδιαφέρουσα και αποτελεσματική μέθοδος να είναι η προσθήκη κοινωνικών και οικονομικών γνωρίσματων στην πρόβλεψη του frailty index καθώς βάση της καλύτερης μέτρησης που μας έδωσε ο GBDT στο r2 στον πίνακα 15 τα βιολογικά γνωρίσματα εξηγούν το ¼ (0.250) της συνολικής διακύμανσης του δείκτη.

Συνολικά η τεχνική αυτή δεν κατάφερε να προβλέψει τα γνωρίσματα και να υπολογίσει έναν ρεαλιστικό νέο δείκτη βάση αυτών. Συνυπολογίζοντας της απώλειες από και από το μοντέλο πρόβλεψης γνωρίσματων αλλά και από τα μοντέλα πρόβλεψης του frailty index το αποτέλεσμα αυτό είναι αναμενόμενο. Παρόλα αυτά ίσως αξίζει να επαναληφθεί η τεχνική αυτή σε έρευνες όπου θα υπάρχουν περισσότερα διαθέσιμα δεδομένα ανά wave καθώς φαίνεται ότι είναι δυνατή η πρόβλεψη του frailty index μέσω από διαχρονικές μελέτες και τεχνικές εξόρυξης δεδομένων κάτι το οποίο θα προσφέρει σημαντική αξία στον χώρο της ιατρικής.

Επίλογος

Είναι εύλογο επομένως ότι ο frailty index διαδραματίζει ένα σημαντικό ρολό για την αξιολόγηση της υγείας μας και ειδικότερα έχει καταστεί ως ένας πολύ χρήσιμος δείκτης στον κλάδο της γηριατρικής. Προβλέποντας τον frailty index θα είναι δυνατή η καλύτερη εκτίμηση της υγείας των ηλικιωμένων κάνοντας έτσι πιο εύκολη την προσφορά έγκυρής ιατρικής περιθάλψεις βελτιώνοντας έτσι όχι μονό το προσδόκιμο αλλά και την ποιότητα ζωής.

Όπως είδαμε η δημιουργία μοντέλων μηχανικής μάθησης είναι μια αρκετά εύκολη ενέργεια ενώ η κύρια δυσκολία συναντιέται στην συλλογή των δεδομένων στην επιλογή τους και στην μορφοποίηση τους. Όμως καθώς η τεχνικές εξορύξεις δεδομένων γίνονται όλο και πιο δημοφιλής και ταυτόχρονα με την άνθηση της συλλογής της πληροφορίας που βιώνουμε στην εποχή μας θα είναι όλο και πιο εύκολη η δημιουργία dataset για αυτές τις χρήσεις. Μοντέλα τα οποία θα μπορούν επεξεργάζονται δεδομένα στην διαστατή τύπου χρόνου όπως το LSTM θα φανούν ιδιαίτερα χρήσιμα για την πρόβλεψη μέσα από αυτά τα datasets.

Έτσι είτε χρειαζόμαστε να κατανοήσουμε καλύτερα τους τομείς της υγείας είτε επιθυμούμε να αυξήσουμε την ποιότητας της οι τεχνικές εξορύξεις γνώσεις θα αποτελέσουν ένα ιδανικό εργαλείο για αυτόν τον σκοπό. Εν κατακλείδι θεωρώ ότι είναι σημαντικό να ασχοληθούμε και βελτιώσουμε αυτόν τον κλάδο της πληροφορικής καθώς έχει να συνεισφέρει πολλά τόσο στην υγειά όσο στην γνώση του ανθρώπου.

Βιβλιογραφία

Beeksma, M., Verberne, S., van den Bosch, A., Das, E., Hendrickx, I., & Groenewoud, S. (2019). Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Medical Informatics and Decision Making*.

Caballero, F. F., Soulis, G., Engchuan, W., ánchez-Niubó, A., Arndt, H., Ayuso-Mateos, J. L., . . . Panagiotakos, D. B. (2017). dvanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the ATHLOS project. *Scientific reports*.

Gale, C. R., Mõttus, R., Deary, I. J., Cooper, C., & Sayer, A. A. (2016). Personality and risk of frailty: The English longitudinal study of Ageing. *Annals of Behavioral Medicine*.

Rahman, S. A., & Adjeroh, D. A. (2019). Deep learning using convolutional LSTM estimates biological age from physical activity. *Scientific Reports*.

Searle , S. D., Mitnitski, A., Gahbauer, E. A., Gill, T. M., & Rockwood, K. (2008). A standard procedure for creating a frailty index. *BMC Geriatrics*.

Su, D., Zhang, X., He, K., & Chen, Y. (2021). Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. *Journal of Affective Disorders*.

Yang, H., & Bath, P. A. (2020). he Use of Data Mining Methods for the Prediction of Dementia: Evidence From the English Longitudinal Study of Aging. *IEEE journal of biomedical and health informatics*.

Ιακωβίδου, Ε. (2015). *Εξαγωγή γνώσης από διατροφικά δεδομένα και παράγοντες που σχετίζονται με καρδιαγγειακά νοσήματα.* Αθηνά: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Δικτυογραφία

Beheshti, N. (2019, 3 2). *Random Forest Regression*. Ανάκτηση από Towards Data Science: https://towardsdatascience.com/random-forest-regression-5f605132d19d

Brownlee, J. (2019, 8 6). *How to use Learning Curves to Diagnose Machine Learning Model Performance*. Ανάκτηση από Machine Learning Mastery: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Brownlee, J. (2020, 8 28). *Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras*. Ανάκτηση από Machine Learning Mastery: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

Halil, E. (2021, 11 17). *CNN-LSTM-Based Models for Multiple Parallel Input and Multi-Step Forecast*. Ανάκτηση από Towards Data Science: https://towardsdatascience.com/cnn-lstm-based-models-for-multiple-parallel-input-and-multi-step-forecast-6fe2172f7668

Masui, T. (2019, 1 20). *All You Need to KNow about Gradient Boosting Algorith- Part1. Regression*. Ανάκτηση από Towards Data Science: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502