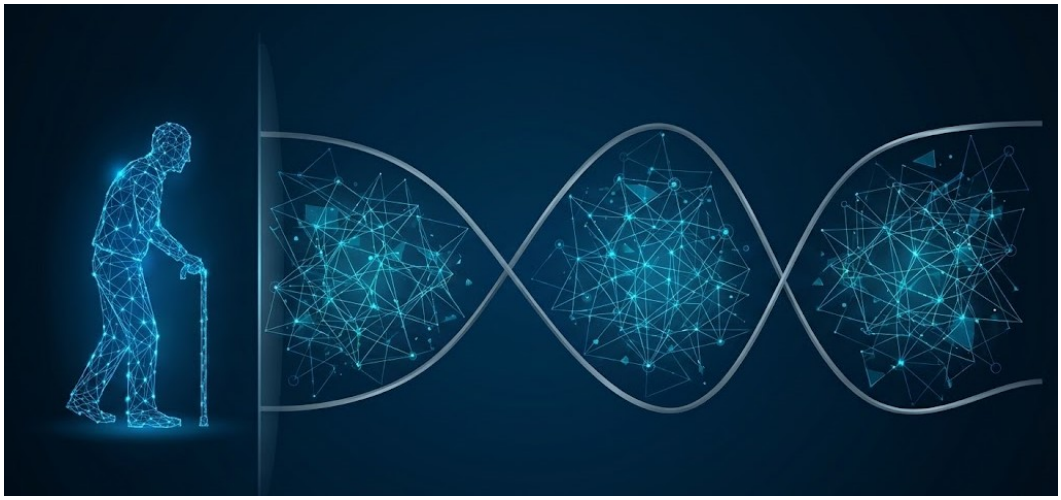
ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξατομικευμένες Παρεμβάσεις για Πρόβλεψη και Πρόληψη Πτώσεων μέσω Μηχανικής Μάθησης



Του φοιτητή
Νικόλαου Σαρακενίδη
Αρ. Μητρώου: 2020/151

Επιβλέπων
Παναγιώτης Αδαμίδης
Καθηγητής

Θεσσαλονίκη, Ιανουάριος, 2026

Τίτλος Δ.Ε.: Εξατομικευμένες Παρεμβάσεις για Πρόβλεψη και Πρόληψη Πτώσεων μέσω Machine Learning
Κωδικός Δ.Ε.: 25211
Ονοματεπώνυμο φοιτητή: Νικόλαος Σαρακενίδης
Ονοματεπώνυμο εισηγητή: Παναγιώτης Αδαμίδης
Ημερομηνία ανάληψης Δ.Ε.: 27-03-2025
Ημερομηνία περάτωσης Δ.Ε.: 23-01-2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

«Tell me and I forget, teach me and I may remember, involve me and I learn.»

Benjamin Franklin

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε με στόχο τη μελέτη και αξιοποίηση σύγχρονων τεχνικών μηχανικής μάθησης, με επίκεντρο το φαινόμενο των πτώσεων σε ηλικιωμένα άτομα. Το θέμα των πτώσεων αντιμετωπίσθηκε ως μία πρόκληση με έντονο κοινωνικό και οικονομικό αντίκτυπο. Η συμβολή μου σε πραγματικά προβλήματα υγείας και η επιθυμία μου να εφαρμόσω μεθόδους μηχανικής μάθησης, με ενέπνευσαν για την υλοποίηση της εργασίας.

Μέσα από τη διαδικασία αυτή, είχα την ευκαιρία να αποκομίσω θεωρητικές γνώσεις και να διευρύνω τις πρακτικές μου δεξιότητες στον τομέα της μηχανικής μάθησης και της ανάπτυξης σύγχρονων εφαρμογών λογισμικού. Παράλληλα, απέκτησα γνώση στη διεπιστημονική προσέγγιση μεταξύ της πληροφορικής και της ιατρικής, που ενισχύει τόσο την ακαδημαϊκή όσο και την επαγγελματική μου πορεία.

Περίληψη

Η παρούσα διπλωματική εργασία εστιάζει στην πρόβλεψη και πρόληψη πτώσεων σε ηλικιωμένα άτομα, μέσω της ανάπτυξης ενός ερμηνεύσιμου συστήματος βασισμένου σε αρχές μηχανικής μάθησης. Η μελέτη βασίστηκε σε ένα πραγματικό και ετερογενές σύνολο δεδομένων ηλικιωμένων, το οποίο δεν περιλάμβανε άμεση μεταβλητή-στόχο για την εκτίμηση του κινδύνου. Για τον σκοπό αυτό εφαρμόστηκε μια στρατηγική συσταδοποίησης και μεταγενέστερης ταξινόμησης, όπου αρχικά αξιοποιήθηκαν αλγόριθμοι συσταδοποίησης για τον εντοπισμό προφίλ κινδύνου, τα οποία στη συνέχεια χρησιμοποιήθηκαν ως ψευδο-ετικέτες για την εκπαίδευση επιβλεπόμενων μοντέλων. Μεταξύ των μοντέλων που αξιολογήθηκαν, ο ταξινομητής SVC παρουσίασε την καλύτερη επίδοση, επιτυγχάνοντας F1-score 95.32% σε πενταπλή διασταυρούμενη επικύρωση και 100% σε ανεξάρτητο σύνολο δοκιμής. Παράλληλα, αξιοποιήθηκαν τεχνικές ερμηνευσιμότητας (SHAP), επιτρέποντας την ανάλυση των χαρακτηριστικών που επηρεάζουν την κατάταξη κάθε ατόμου σε μία από τις ομάδες κινδύνου. Τα αποτελέσματα της μελέτης ενσωματώθηκαν σε μια διαδραστική εφαρμογή που υλοποιήθηκε με τη χρήση της πλατφόρμας Streamlit, προσφέροντας τη δυνατότητα άμεσης πρόβλεψης προφίλ κινδύνου για νέους χρήστες, και προτείνοντας εξατομικευμένες παρεμβάσεις πρόληψης.

Η εργασία συμβάλλει στην ανάπτυξη ενός ερμηνεύσιμου και πρακτικά εφαρμόσιμου συστήματος υποστήριξης αποφάσεων, σχεδιασμένου για χρήση από επαγγελματίες φυσικοθεραπευτές και ειδικούς στον χώρο της υγείας.

Personalized Interventions for Prediction and Prevention of Falls using Machine Learning

Nikolaos Sarakenidis

Abstract

This thesis focuses on the prediction and prevention of falls in older adults, through the development of an interpretable system based on machine learning principles. The study was conducted using a real-world and heterogeneous dataset of elder individuals falls, but lacks a direct target variable for fall risk assessment. To address this, a sequential clustering and classification strategy was implemented, in which unsupervised clustering algorithms were initially utilized to identify fall risk profiles, which were used as pseudo-labels for the supervised learning models. Among the models evaluated, the SVC classifier demonstrated the best performance, achieving F1-score 95.32% in 5-fold cross validation and 100% on an independent test set. In addition, interpretability techniques were employed (SHAP), to analyze the factors that most significantly influence the classification of each individual into one of the risk groups. The results of the study were incorporated into an interactive application developed using the Streamlit platform, offering an immediate risk profile prediction for new individuals and provisioning personalized prevention interventions.

The proposed work contributes to the development of an interpretable and practically applicable decision-support system, designed for use by professional physiotherapists and other healthcare experts.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω και συγχαρώ την οικογένεια και τους φίλους μου, καθώς και όλους εκείνους που συνέβαλαν και στάθηκαν στο πλευρό μου καθ' όλη τη διάρκεια εκπόνησης της εργασίας, ενθαρρύνοντάς με να ξεπερνάω κάθε δυσκολία και να προχωράω ένα βήμα τη φορά.

Επίσης, θα ήθελα να ευχαριστήσω τον καθηγητή του τμήματος Φυσικοθεραπείας του ΔΙΠΑΕ, τον κ. Λύτρα για την προσφορά του, παρέχοντας το σύνολο δεδομένων πάνω στο οποίο βασίστηκε η διατριβή και τη συνεισφορά του στη διευκρίνιση κάθε δυσκολίας κατανόησης που αντιμετώπισα ως προς το σύνολο δεδομένων.

Τέλος, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Αδαμίδη, ο οποίος με καθοδήγησε από την πρώτη κιόλας στιγμή που ανέλαβα τη διπλωματική εργασία και με υποστήριξε καθ' όλη τη διάρκεια της εργασίας.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

ADL	Activities of Daily Living (Καθημερινές Δραστηριότητες)
API	Application Programming Interface
BBS	Berg Balance Scale
CHAIRSTANDTEST	30-Second Chair Stand Test
CONFbal	Confidence in Balance (Greek version)
CV	Cross Validation (Διασταυρούμενη Επικύρωση)
DBCV	Density-Based Clustering Validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DL	Deep Learning (Βαθιά Μάθηση)
EHR	Electronic Health Record
FAQ	Frequently Asked Questions
FOURSBT	Four-Stage Balance Test
IMU	Inertial Measurement Unit (Μονάδα Αδρανειακής Μέτρησης)
ML	Machine Learning (Μηχανική Μάθηση)
MMSE	Mini-Mental State Examination
NICE	National Institute for Health and Care Excellence
OHE	One-Hot Encoding
PCA	Principal Component Analysis (Ανάλυση Κυρίων Συνιστωσών)
RL	Reinforcement Learning (Ενισχυτική Μάθηση)
SHAP	Shapley Additive exPlanations
Short FES-I	Short Falls Efficacy Scale - International
SL	Supervised Learning (Επιβλεπόμενη Μάθηση)
SMOTE	Synthetic Minority Over-sampling Technique
SVC	Support Vector Classification
SVMs	Support Vector Machines (Μηχανές Διανυσμάτων Υποστήριξης)
TUG	Timed Up and Go
UL	Unsupervised Learning (Μη Επιβλεπόμενη Μάθηση)
WCSS	Within Cluster Sum of Squares
WHO	World Health Organization (Οργανισμός Παγκόσμιας Υγείας)
XGBoost	eXtreme Gradient Boosting
ΚΑΠΗ	Κέντρο Ανοικτής Προστασίας Ηλικιωμένων

ΤΝ	Τεχνητή Νοημοσύνη
ΤΝΔ	Τεχνητά Νευρωνικά Δίκτυα

Γλωσσάρι

EHR Ψηφιακά καταγεγραμμένες πληροφορίες που αφορούν το ιατρικό ιστορικό καθώς και τη συνολική φροντίδα υγείας ενός ασθενούς.

IMU Μία ηλεκτρονική συσκευή που μετράει και αναφέρει συγκεκριμένες τιμές δύναμης και βαρύτητας, γωνιακού ρυθμού, και μερικές φορές τον προσανατολισμό του σώματος, χρησιμοποιώντας έναν συνδυασμό από επιταχυνσιόμετρα και γυροσκόπια.

Διατομεακή μελέτη Ένας τύπος ερευνητικής μελέτης κατά τον οποίο δεδομένα συλλέγονται σε μία συγκεκριμένη χρονική στιγμή ή σε ένα σύντομο χρονικό διάστημα, με σκοπό την αποτύπωση της κατάστασης ενός πληθυσμού ή φαινομένου τη δεδομένη χρονική περίοδο.

Πολυτομεακές παρεμβάσεις Παρεμβάσεις που σχεδιάζονται και υλοποιούνται σε συνδυασμό πολλών διαφορετικών τομέων ή ειδικοτήτων, με στόχο την αντιμετώπιση σύνθετων προβλημάτων.

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Οι πτώσεις αποτελούν μείζον ζήτημα δημόσιας υγείας, καθώς συνδέονται με αυξημένη θνησιμότητα, νοσηρότητα και απώλεια λειτουργικής κατάστασης [1]. Η σύγχρονη κλινική πρακτική μετατοπίζεται από την απλή αντιμετώπιση των συνεπειών προς την πρόληψη των πτώσεων, εστιάζοντας στη πολυπαραγοντική στρατηγική, που περιλαμβάνει ετερογενή προγράμματα πρόληψης. Ωστόσο, στις περισσότερες μελέτες, η υπάρχουσα βιβλιογραφία επικεντρώνεται στην ανίχνευση ή πρόβλεψη της πτώσης προκειμένου να ενημερωθούν οι συγγενείς ή οι υγειονομικοί, ή στη γενική περιγραφή παραγόντων κινδύνου. Στην εργασία αυτή παρουσιάζεται ένα υβριδικό σύστημα, το οποίο συμβάλει στην αντιμετώπιση των παραπάνω ζητημάτων.

1.2 Σκοπός και Στόχοι Διπλωματικής

Η διατριβή αποσκοπεί να αποτελέσει ένα συμπληρωματικό εργαλείο αξιολόγησης πτώσεων, προς τους ανθρώπους υγείας που ασχολούνται με τον τομέα των πτώσεων, για τη διευκόλυνση λήψης αποφάσεων σε περιπτώσεις ατόμων που έχουν ήδη υποστεί πτώση.

Αξιοποιώντας σύγχρονες τεχνικές Μηχανικής Μάθησης (Machine Learning – ML) και εστιάζοντας στην εύρεση των πιο σημαντικών παραγόντων κινδύνου πτώσης, η παρούσα εργασία επιδιώκει να αναδείξει διακριτά προφίλ ηλικιωμένων, τα οποία μπορούν να υποστηρίξουν την αναγνώριση συγκεκριμένου επιπέδου κινδύνου, για τη διαμόρφωση στοχευμένων παρεμβάσεων πρόληψης στο πλαίσιο της κλινικής λήψης αποφάσεων. Πιο συγκεκριμένα η εργασία στοχεύει στην:

- εύρεση των πιο σημαντικών (πολυτροπικών) παραγόντων που οδηγούν σε πτώση
- ομαδοποίηση των στιγμιοτύπων δεδομένων σε προφίλ κινδύνου ως προς τα κοινά τους χαρακτηριστικά
- ερμηνεία των διακριτών προφίλ κινδύνου
- πρόβλεψη των προφίλ που ανήκουν νέα άτομα
- παροχή προσωποποιημένων πολυτομεακών παρεμβάσεων
- οπτικοποίηση και αλληλεπίδραση όλων των παραπάνω μέσω μίας διαδραστικής εφαρμογής

Οι παραπάνω στόχοι, συγκροτούν μία ενιαία μεθοδολογία ανάπτυξης ενός συστήματος πρόβλεψης κινδύνου και παροχής εξατομικευμένων συστάσεων πρόληψης πτώσεων, αξιοποιώντας τεχνικές ML, το οποίο μπορεί να συντελέσει συμπληρωματικό εργαλείο υποστήριξης στη διαδικασία πρόληψης πτώσεων.

1.3 Δομή Εργασίας

Η εργασία είναι δομημένη με αρθρωτό τρόπο, εξασφαλίζοντας τη σταδιακή ανάπτυξη του ερευνητικού προβλήματος, ξεκινώντας με τα θεωρητικά στοιχεία και συνεχίζοντας με τα πρακτικά, μέχρι την εφαρμογή και την ερμηνεία των αποτελεσμάτων.

Κεφάλαιο 2

Στο πρώτο κεφάλαιο γίνεται μία εισαγωγή και αναπτύσσονται οι στόχοι της εργασίας. Στο δεύτερο κεφάλαιο παρουσιάζεται το γενικό πλαίσιο του προβλήματος των πτώσεων και των μεθόδων αντιμετώπισής τους. Στο ίδιο, αναλύεται το θεωρητικό υπόβαθρο, περιγράφοντας βασικές έννοιες της ML, τεχνικές προεπεξεργασίας δεδομένων, τα μοντέλα που χρησιμοποιήθηκαν, μέθοδοι ανάπτυξης και αξιολόγησης των μοντέλων καθώς και σχετικές εργασίες. Το τρίτο κεφάλαιο επικεντρώνεται στα δεδομένα και στη μεθοδολογία, αναφέροντας τα υποψήφια σύνολα δεδομένων και τεκμηριώνοντας την επιλογή αυτού που χρησιμοποιήθηκε στην εργασία. Παράλληλα, τα δεδομένα αναλύονται διεξοδικά, ενώ περιγράφεται η μεθοδολογία που ακολουθήθηκε στο πρακτικό τμήμα. Στο τέταρτο κεφάλαιο παρουσιάζονται οι τεχνικές προεπεξεργασίας και ο τρόπος που εφαρμόστηκαν, διαμορφώνοντας το τελικό σύνολο δεδομένων της εργασίας. Συγχρόνως, εμφανίζονται τα πειραματικά αποτελέσματα των τεχνικών ML. Στο πέμπτο κεφάλαιο λαμβάνει χώρα η ερμηνεία των προφίλ κινδύνου που προέκυψαν από το προηγούμενο, αποδίδονται υπολογιστικές αναπαραστάσεις ερμηνείας του κάθε προφίλ με χρήση τεχνικών ερμηνευσιμότητας, ενώ ακόμη γίνεται αναφορά για τις εξατομικευμένες παρεμβάσεις πρόληψης πτώσεων. Στο έκτο κεφάλαιο, παρουσιάζεται η τελική εφαρμογή της εργασίας. Τέλος, στο έβδομο κεφάλαιο συνοψίζονται τα κύρια ευρήματα της διατριβής, συζητούνται οι περιορισμοί και οι προκλήσεις, καθώς ακόμη προτείνονται μελλοντικές επεκτάσεις.

1.4 Επίλογος

Στο παρόν κεφάλαιο παρουσιάστηκε το πλαίσιο της εργασίας, τονίζοντας τα σημεία τα οποία καλείται να καλύψει. Επισημάνθηκαν, ο σκοπός και οι στόχοι της εργασίας καθώς και οι χρήστες στους οποίους απευθύνεται. Επιπλέον, έγινε μία συνοπτική παρουσίαση της δομής των κεφαλαίων που ακολουθούν.

Κεφάλαιο 2ο: Θεωρητικό Υπόβαθρο

Το δεύτερο κεφάλαιο της παρούσας εργασίας παρουσιάζει το θεωρητικό υπόβαθρο που σχετίζεται με τις πτώσεις, εστιάζοντας όμως στη θεωρητική θεμελίωση εννοιών, μεθόδων και τεχνικών που αξιοποιούνται στη μελέτη της πρόληψης πτώσεων μέσω Μηχανικής Μάθησης.

Αρχικά, παρουσιάζονται οι πτώσεις, τα αίτια και οι συνέπειες γύρω από αυτές σε θεωρητικό επίπεδο, καθώς και τα παραδοσιακά εργαλεία αξιολόγησης του κινδύνου πτώσεων. Στη συνέχεια, το κεφάλαιο μεταβαίνει στη νέα γενιά τεχνολογιών και μεθόδων αξιολόγησης κινδύνου πτώσεων. Παρουσιάζεται ο ρόλος των τεχνολογιών αυτών και η συμβολή τους στη δημιουργία πλουσιότερων αναπαραστάσεων δεδομένων. Ακολούθως, εισάγεται το ML ως βασικό αναλυτικό εργαλείο, περιγράφοντας τις κατηγορίες αλγορίθμων που χρησιμοποιήθηκαν και αποτελούν το επίκεντρο της εργασίας. Ιδιαίτερη έμφαση δίνεται στις τεχνικές προεπεξεργασίας δεδομένων, οι οποίες αποτελούν κρίσιμο στάδιο για την επιτυχή εφαρμογή των αλγορίθμων. Το κεφάλαιο συνεχίζει με την παρουσίαση των μετρικών αξιολόγησης που χρησιμοποιήθηκαν για την αξιολόγηση των αλγορίθμων, ενώ επίσης αναδεικνύονται κρίσιμα θεωρητικά τμήματα που αξιοποιήθηκαν στο πρακτικό μέρος της εργασίας, όπως οι υπερπαράμετροι των αλγορίθμων και η ερμηνευσιμότητα των αποτελεσμάτων των μοντέλων. Το θεωρητικό υπόβαθρο ολοκληρώνεται με την αναφορά σχετικών εργασιών και τον επίλογο του κεφαλαίου, συνοψίζοντας όσα ειπώθηκαν.

2.1 Πτώσεις

2.1.1 Πτώσεις στην τρίτη ηλικία

Πτώση χαρακτηρίζεται μία απροσδόκητη κατάσταση στην οποία ένα άτομο βρίσκεται ξαπλωμένο στο έδαφος, στο πάτωμα ή στο χαμηλότερο επίπεδο [2]. Η πτώση μπορεί να συμβεί σε όλους τους ανθρώπους ανεξαρτήτως ηλικίας, όμως για τους ανθρώπους μεγαλύτερης ηλικίας, ακόμη και χωρίς σχετικές παθήσεις (όπως Πάρκινσον, Άνοια κλπ), ο κίνδυνος είναι πολύ μεγαλύτερος. Καθώς η ηλικία του ανθρώπου αυξάνεται, συγχρόνως αποκτά δυσκολίες κίνησης και μείωση λειτουργικής ικανότητας. Οι πτώσεις μπορεί να είναι επαναλαμβανόμενες, δηλαδή περισσότερες από μία σε χρονικό διάστημα ενός έτους, και ως αποτέλεσμα προκαλούν τραυματισμούς, όπως συχνά καταλήγουν σε κατάγματα ισχίου [3]. Οι πτώσεις είναι η πιο συχνή αιτία εισαγωγής σε νοσοκομείο από μη-θανατηφόρο τραύμα μεταξύ ηλικιωμένων ανθρώπων [3], ενώ είναι η δεύτερη σε σειρά αιτία θανάτου λόγω ακούσιου τραυματισμού παγκοσμίως [1]. Περίπου το 30% των ανθρώπων άνω των 65 ετών του παγκόσμιου πληθυσμού έρχεται αντιμέτωπο με τουλάχιστον μία πτώση ανά έτος, ενώ το ποσοστό αυξάνεται στα 42% για τα άτομα άνω των 70 ετών, και στο 50% για άνω των 80 [1], [2].

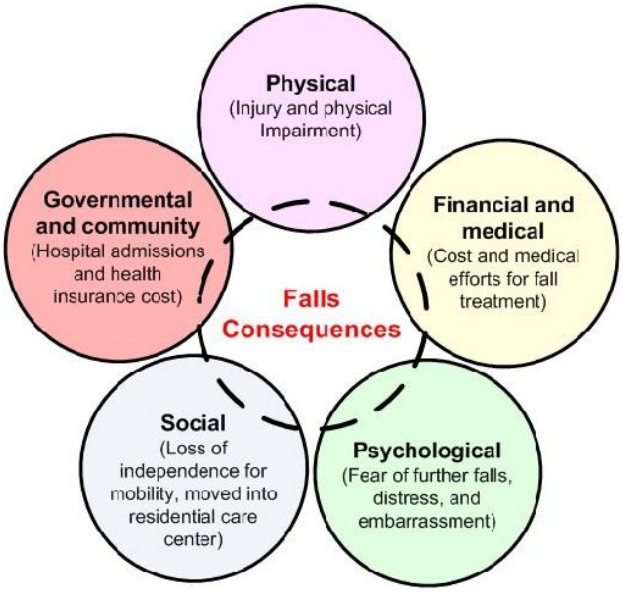
Η ραγδαία αύξηση του γηραιού παγκόσμιου πληθυσμού (>65) προκαλεί ανησυχητική τάση στον τομέα της υγειονομικής περίθαλψης. Σύμφωνα με το World Health Organization (WHO) περίπου 37.3 εκατομμύρια πτώσεις που απαιτούν εισαγωγή σε νοσοκομείο ή απλώς ιατρική περίθαλψη συμβαίνουν κάθε έτος [1]. Εκτιμάται, ότι σχεδόν 1 εκατομμύριο εισαγωγές σε νοσοκομεία λόγω πτώσης πραγματοποιούνται ετησίως στις ΗΠΑ με το αντίστοιχο κόστος επιβάρυνσης να φθάνει τα $50 δισεκατομμύρια [4]. Για το έτος 2020, το κόστος σε μη-θανάσιμες πτώσεις υπολογίζεται στα $80 δισεκατομμύρια για τις υγειονομικές εγκαταστάσεις στις ΗΠΑ [5]. Στην Αυστραλία, το 2015-2016 καταγράφηκαν 34 000 εισαγωγές ατόμων με κύρια αιτία την πτώση [6]. Στο Ηνωμένο Βασίλειο 75 000 κατάγματα ισχίου συμβαίνουν ετησίως, όπου το κόστος στην υγεία και την κοινωνία να ξεπερνάει τα €2 δισεκατομμύρια [7]. Όλες αυτές οι καταγραφές δεν έχουν μόνο

προσωπικό αντίκτυπο στους ίδιους τους ηλικιωμένους που έχουν υποστεί πτώση αλλά και στα άτομα και τα νοσοκομεία που τους φροντίζουν ή γενικώς τους χώρους περίθαλψης. Η σταδιακή αύξηση του μέσου όρου ηλικίας και των διαφόρων νοσηροτήτων που συνοδεύουν του ηλικιωμένους, συνεισφέρουν στην αύξηση της ευαισθησίας και την μη διαθεσιμότητα κλινών των υγειονομικών υποδομών. Εκτιμάται ότι μέχρι το 2050 τουλάχιστον ένας στους πέντε ανθρώπους θα έχει ηλικία άνω των 65 ετών [8]. Για αυτό κρίνεται αναγκαία η πρόληψη των ατόμων αυτών από τις πτώσεις.

2.1.2 Αίτια και συνέπειες των πτώσεων

Οι συχνότητα εμφάνισης πτώσεων μπορεί να διαφέρει μεταξύ διαφορετικών περιοχών και πληθυσμών. Μία αιτία μπορεί να οφείλεται στην κουλτούρα της περιοχής και στον μεγαλύτερο αριθμό ηλικιωμένων που μπορεί να διαμένουν στη εκεί. Για παράδειγμα, σε Κινέζικους πληθυσμούς στην νοτιοανατολική Ασία ο ρυθμός πτώσεων βρίσκεται σε κλίμακα 15%-34%, ενώ σε περιοχές όπως η Λατινική Αμερική και η Καραϊβική ο ρυθμός ανέρχεται από 22% και φθάνει έως τα 34% στην Χιλή [9].

Οι αιτίες των πτώσεων είναι πολυπαραγοντικές, δηλαδή δεν προέρχονται από έναν μόνο παράγοντα, αλλά από την αλληλεπίδραση πολλαπλών παραγόντων κινδύνου, όπως φαίνεται στο σχήμα 2.1. Σύμφωνα με οργανισμούς που παρέχουν οδηγίες αποτροπής πτώσεων [1], [4], [9], [10], αναφέρουν ότι οι παράγοντες διακρίνονται σε προσωπικούς και περιβαλλοντικούς. Οι προσωπικοί ή ατομικοί λόγοι σχετίζονται με τη χρήση φαρμάκων, με μειωμένες ή πολύπλοκες καθημερινές δραστηριότητες, με ψυχολογικές διαταραχές, με μυικά και σκελετικά προβλήματα καθώς και με προβλήματα στάσης και ισορροπίας τα οποία εκδηλώνουν υψηλό κίνδυνο πτώσης. Από την άλλη, οι περιβαλλοντικοί λόγοι σχετίζονται με το περιβάλλον που ζει ή αλληλεπιδρά ο ηλικιωμένος και φέρουν ιδιαίτερη ευθύνη για την πρόκληση πτώσης. Συχνά, ανισόπεδο ή ολισθηρό δάπεδο, αντικείμενα στο πάτωμα, κακή εργονομία χώρου, σκαλοπάτια, μη καλή ορατότητα λόγω κακού φωτισμού ή νύχτας, αποτελούν μερικές από τις πιο συχνές αιτίες που οι ηλικιωμένοι υπόκεινται σε πτώση. Ο συνδυασμός των προσωπικών και περιβαλλοντικών παραγόντων, οι οποίοι μπορεί να είναι περισσότεροι από ένας από την κάθε κατηγορία, διαδραματίζουν σημαντικό ρόλο στην αύξηση των πτώσεων που συμβαίνουν καθημερινά παγκοσμίως.



Σχήμα 2.1: Οι κύριες επιπτώσεις των πτώσεων [8]

2.1.3 Προσεγγίσεις Αντιμετώπισης

Στο παρελθόν, χρησιμοποιούνταν ιδιαίτερα παραδοσιακά εργαλεία αξιολόγησης του κινδύνου πτώσης των ατόμων. Μερικά από αυτά περιγράφονται συντόμως παρακάτω.

Το Morse Fall Scale (MFS), δημοσιεύτηκε το 1989 ως ένα απλό και γρήγορο εργαλείο εκτίμησης του κινδύνου πτώσης σε διάφορες υγειονομικές υποδομές (π.χ. νοσοκομεία) [6], [11]. Σχεδιάστηκε για να διαχειρίζεται από νοσοκόμους/ες. Αναπτύχθηκε με τη χρήση 6 στοιχείων-παραγόντων, σύμφωνα με τα οποία αποδίδεται μία τιμή που μέσω αυτής συμπεραίνεται το ύψος του κινδύνου πτώσης των ασθενών. Ένα άλλο εργαλείο αποτελεί το εργαλείο αξιολόγησης κινδύνου πτώσης St. Thomas (STRATIFY) σε ηλικιωμένους ασθενείς που αναπτύχθηκε το 1997 [6], [12]. Το συγκεκριμένο εμφανίζεται σε πολλές παραλλαγές, όπου η κάθε μία αναπτύχθηκε για την βελτιστοποίηση της τεχνικής ή για μερική αλλαγή των μετρήσεων των δειγμάτων. Σε όλες τις παραλλαγές του εργαλείου, τουλάχιστον 5 στοιχεία χρησιμοποιούνται για την πρόβλεψη πτώσεων σε κλινικές υποδομές. Ακόμη ένα άλλο εργαλείο, το μοντέλο κινδύνου πτώσης Hendrick II, πρωτοεμφανίστηκε το 2003, περιέχοντας 3 στοιχεία, περιλαμβάνοντας παράγοντες όπως η πρόσληψη πολλών φαρμάκων, κατάσταση υγείας και συμπεριφοράς [13]. Και τα τρία εργαλεία λειτουργούν με την ανάθεση αριθμητικής βαθμολογίας στους παράγοντες που εξετάζουν πάνω στα παρατηρούμενα άτομα. Το κάθε στοιχείο-παράγοντας προσδίδει κάποια συνεισφορά στην τελική βαθμολογία κινδύνου πτώσης. Προσθέτοντας τα στοιχεία, εάν το άθροισμα ξεπερνάει τις προσδιορισμένες τιμές που ορίζονται για το κάθε εργαλείο, τότε οι ηλικιωμένοι ανήκουν στην ομάδα χαμηλού, μεσαίου ή υψηλού κινδύνου πτώσης.

Ωστόσο, τα παραδοσιακά εργαλεία εκτίμησης των πτώσεων έχουν περιορισμούς σχετικά με την συλλογή των δεδομένων και την ανάγκη μετάλλαξής τους προκειμένου να προσαρμόζονται σε διαφορετικούς πληθυσμούς και κλινικές καταστάσεις [6]. Ένα σημαντικό μειονέκτημα είναι ότι απαιτούν την συλλογή των

δειγμάτων και την κλινική εκτίμηση από επαγγελματίες του χώρου, γεγονός που επιβαρύνει χρονικά την τεκμηρίωση των αποτελεσμάτων, και συνεπώς τις τελικές παρεμβάσεις που πρέπει να ασκηθούν στους ηλικιωμένους [6]. Τα δείγματα βασίζονται σε στατικά δεδομένα και συχνά τα αποτελέσματα των εργαλείων υστερούν σε προγνωστική ακρίβεια και εξειδίκευση λόγω μη γραμμικής φύσης των παραγόντων κινδύνου πτώσης με τις πραγματικές συνθήκες πτώσεων της καθημερινής ζωής. Το MFS παρουσιάζει ασυνέπειες στην διαγνωστική ακρίβεια των τιμών του σε διαφορετικές υποδομές και πληθυσμούς, επηρεάζοντας τα αποτελέσματα και προκαλώντας εσφαλμένες βαθμολογίες κινδύνου πτώσης [11]. Έτσι λοιπόν, αυτού του είδους τα εργαλεία, συχνά εμφανίζουν καθυστερήσεις της εξαγωγής τους και χαρακτηρίζονται από υποκειμενικές βαθμολογίες, ως προς την προσωπική άποψη του φυσικοθεραπευτή ή γιατρού, οδηγώντας σε δυσκολίες καθορισμού λήψης αποφάσεων για την αντιμετώπιση και ανάπτυξη τακτικών παρεμβάσεων σε υγειονομικές υποδομές, για την κατάλληλη φροντίδα των ηλικιωμένων. Ως αποτέλεσμα, δεν μπορούν να αποτελέσουν αποδοτικές μέθοδοι για την πρόβλεψη και αποτροπή των πτώσεων.

Τα τελευταία χρόνια, με την ανάπτυξη της τεχνολογίας δημιουργήθηκαν ολοένα και περισσότερες προσεγγίσεις για την ανίχνευση και αποτροπή ή μείωση των πτώσεων σε ηλικιωμένα κυρίως άτομα, καθώς αυτά είναι που επιβαρύνονται περισσότερο και πιο έντονα στο μεγαλύτερο ποσοστό του πληθυσμού. Οι προσεγγίσεις αυτές βασίζονται σε συστήματα αισθητήρων. Ένας αισθητήρας είναι μία συσκευή που συλλέγει πληροφορίες και καταστάσεις από το περιβάλλον του και στη συνέχεια τις αποστέλλει σε άλλες ηλεκτρονικές συσκευές για αποθήκευση και επεξεργασία τους [14]. Η χρήση ενός ή πολλών αισθητήρων και η διασύνδεσή του/ς με μία τέτοια ηλεκτρονική συσκευή καθιστά την διαδικασία να αποτελεί ένα σύστημα. Στην βιβλιογραφία, δεν υπάρχει ξεκάθαρη αναφορά των διακεκριμένων συστημάτων καθώς σε μία ποικιλία επιστημονικών κειμένων υπάρχουν διαφορετικές θεωρίες των συστημάτων, προσδίδοντας σύγχυση. Ωστόσο, ύστερα από εκτενή έρευνα, ο συγγραφέας του κειμένου κατέληξε τα συστήματα αυτά στο να διακρίνονται σε τέσσερις ευρύς πυλώνες, με βάση την τοποθέτηση των αισθητήρων.

a) Συστήματα βασισμένα σε φορετές συσκευές
b) Συστήματα βασισμένα σε περιβαλλοντικές συσκευές
c) Συστήματα με συσκευές καταγραφής οπτικής κίνησης
d) Υβριδικά συστήματα

Στην πρώτη κατηγορία εμφανίζονται συστήματα που βασίζονται σε φορετές συσκευές. Αποτελούν την πιο διαδεδομένη κατηγορία, και αυτό διαπιστώνεται εύκολα από την χρήση τους στα περισσότερα σύνολα δεδομένων που επικρατούν, μερικά από τα οποία συναντώνται σε επόμενα κεφάλαια της εργασίας. Οι αισθητήρες των συσκευών αυτών ενσωματώνουν τεχνολογίες με τριαξονικά επιταχυνσιόμετρα και γυροσκόπια όπου ευρύτερα ονομάζονται Αδρανειακές Μονάδες Μέτρησης (Inertial Measurement Units – IMU). Η αξιοποίησή του επιτρέπει την καταγραφή γραμμικών επιταχύνσεων, γωνιακών ταχυτήτων και προσανατολισμού στον χώρο, παρέχοντας λεπτομερείς πληροφορίες για την βάδιση, τη στάση του σώματος, τη συμμετρία και τις μεταβάσεις κίνησης [15]. Επίσης, τα IMU μπορεί να περιέχουν μαγνητόμετρα, επεκτείνοντας τις δυνατότητες συλλογής δεδομένων. Οι αισθητήρες αυτοί μπορεί να βρίσκονται είτε σε έξυπνα κινητά τηλέφωνα είτε να αποτελούν ανεξάρτητες μικρές φορετές συσκευές ειδικού σκοπού [13]. Συνηθέστερα, τοποθετούνται στον καρπό του χεριού, γύρω από τη μέση και χαμηλά στο σώμα (λόγου χάριν στα άκρα του ποδιού), για την καλύτερη ευαισθησία στην ανίχνευση και ποιότητα των βαθμολογιών των παραμέτρων που σχετίζονται με τις πτώσεις [17]. Για παράδειγμα, υπάρχουν αισθητήρες πίεσης και έξυπνα πέλματα που αντικαθιστούν τα κλασσικά πέλματα παπουτσιού για την λήψη μετρήσεων βάδισης,

ισορροπίας και κατανομής του φορτίου του σώματος [18].Τα χαρακτηριστικά που εξάγονται από δεδομένα μέσω IMU, δεν ανιχνεύονται από τις παραδοσιακές κλινικές κλίμακες, γεγονός που καθιστά τις IMU ιδιαίτερα χρήσιμες για ακριβέστερη αξιολόγηση κινδύνου πτώσεων. Ωστόσο η χρήση φορετών συσκευών περιλαμβάνει προκλήσεις [19]. Τα ηλικιωμένα άτομα, συχνά δεν είναι εξοικειωμένα με τις ηλεκτρονικές συσκευές. Μπορεί να αντιμετωπίσουν σημαντικό πρόβλημα αποδοχής και ορθής χρήσης των συσκευών αυτών [20]. Παράλληλα, πολλές φορές απαιτείται η απομάκρυνσή των συσκευών από το σημείο εφαρμογής και η επανατοποθέτησή τους αργότερα διότι δυσκολεύουν τους χρήστες σε καταστάσεις όπως ο ύπνος ή άλλες δραστηριότητες. Τέλος, σημαντικό μειονέκτημα ακόμη αποτελεί η ανάγκη φόρτισης, αλλαγής ή αντικατάστασής τους ανά τακτά χρονικά διαστήματα, καθώς συνήθως η πηγή τροφοδοσίας τους η μπαταρία.

Η δεύτερη κατηγορία σχετίζεται με συστήματα τα οποία εστιάζουν στην παρακολούθηση των κινήσεων του χώρου. Αυτές τοποθετούνται σε ένα σημείο του χώρου και καταγράφουν τιμές σύμφωνα με την αλληλεπίδρασή τους με τον χρήστη. Τέτοιες συσκευές είναι αισθητήρες πίεσης δαπέδου, υπερήχων και ραντάρ [15], [19]. Η προσέγγιση αυτή επιτρέπει την παθητική παρακολούθηση της κίνησης χωρίς την άμεση επαφή με τον χρήστη. Βασικό πλεονέκτημα σε σχέση με την προηγούμενη κατηγορία αφού δεν χρειάζεται να συμμορφωθεί ο χρήστης ως προς τους αισθητήρες, (να τους φοράει κατάλληλα).

Τα συστήματα με συσκευές καταγραφής οπτικής κίνησης αποτελούν την τρίτη κατηγορία. Τοποθετούνται επίσης στο περιβάλλον που απαιτείται καταγραφή δεδομένων, επομένως είναι ένα ειδικό είδος της δεύτερης κατηγορίας. Η διαφορά προέρχεται στο τρόπο εισαγωγής της πληροφορίας των συσκευών αυτών, ο οποίος για πρώτη φορά, καταγράφει οπτικά τις κινήσεις στο πεδίο που είναι τοποθετημένες, αποθηκεύοντας τα δεδομένα σε μορφή εικόνας ή βίντεο. Αυτά βρίσκουν εφαρμογή σε συστήματα που εμπλέκουν το πεδίο της Μηχανικής Μάθησης, την Μηχανική Όραση για την επεξεργασία και ανάλυση των δεδομένων. Συνηθέστερα τέτοιες συσκευές είναι RGB κάμερες, κάμερες βάθους και κάμερες με θερμικές υπέρυθρες κάμερες [21]. Μία από τις δυνατότητες των συσκευών αυτών είναι η παράλληλη και σε πραγματικό χρόνο καταγραφή πολλαπλών γεγονότων. Όσων αφορά τις θερμικές κάμερες, αποτελούν εργαλεία φιλικά ως προς την ιδιωτικότητα των ατόμων που καταγράφονται, σε αντίθεση με τις RGB. Ωστόσο, τίθεται θέμα προσοχής η οπτική γωνία που εστιάζουν, ώστε να μην εκλαμβάνουν άσχετες πληροφορίες (π.χ. λόγω υψηλών θερμοκρασιών άσχετων αντικειμένων). Ένα μειονέκτημα που χαρακτηρίζει όλες τις συσκευές αυτής της κατηγορίας είναι ο περιορισμός σε κλειστές περιοχές, καθώς συνήθως τοποθετούνται σε πολύ συγκεκριμένα πεδία, όπως στο σπίτι και στη κλινική. Επιπλέον, συνήθως το κόστος τους είναι υψηλό καθιστώντας τα οικονομικά ασύμφορα για απλά νοικοκυριά [19].

Η τέταρτη προσέγγιση συστημάτων περιλαμβάνει έναν συνδυασμό μερικών ή όλων των αναφερθέντων κατηγοριών. Η χρήση τέτοιων συστημάτων σκοπεύει στην αξιοποίηση των πλεονεκτημάτων της κάθε μίας κατηγορίας, βελτιώνοντας την ανθεκτικότητα σε περιβαλλοντικούς παράγοντες (όπως υψηλές θερμοκρασίες, μεταβολές φωτισμού κλπ) [19]. Βασικό μειονέκτημα όμως είναι η ρητή αύξηση της πολυπλοκότητας και της διαχείρισής τους.

Τα έξυπνα κινητά τηλέφωνα όπως αναφέρθηκε παραπάνω μπορούν να χρησιμοποιηθούν για διαφορετικό σκοπό καταγραφής μετρήσεων καταστάσεων ανάλογα με τον τρόπο την άμεσης επαφής ή μη αλληλεπίδρασης του τηλεφώνου με τον χρήστη [19]. Λαμβάνουν ρόλο πολυαισθητηριακών πλατφορμών και οι μετρήσεις τους εξαρτάται από την εφαρμογή.

Συμπερασματικά, τον πρόσφατο καιρό έχουν αναπτυχθεί πολλές προσεγγίσεις μεταβαίνοντας από την παραδοσιακή στη σύγχρονη αξιολόγηση των δεδομένων. Τείνουν όχι μόνο στη μείωση των πτώσεων αλλά και στην ανίχνευσή τους ώστε να αποφευχθούν οι ανεπιθύμητες συνέπειές τους. Ενώ τα κλινικά εργαλεία

προσφέρουν μία αρχική εικόνα για τα δεδομένα, οι νέες τεχνολογίες είτε με φορετές συσκευές είτε χωρίς, επιτρέπουν τη συνεχή συλλογή δεδομένων υψηλής ανάλυσης αποτελώντας κρίσιμο τμήμα για την αντιμετώπιση του υψηλού αριθμού πτώσεων των ηλικιωμένων.

2.2 Μηχανική Μάθηση

Οι ραγδαίες εξελίξεις της τεχνολογίας έχει οδηγήσει στην εύκολη πρόσβαση και συλλογή μεγάλου όγκου δεδομένων υψηλής ανάλυσης. Η αυξημένη ποσότητα, η πολυπλοκότητα και ο ταχύτατος ρυθμός συλλογής των δεδομένων καθιστά ανεπαρκείς τις κλασσικές στατιστικές μεθόδους [22].

Στο πλαίσιο αυτό, η συνεχής ανάπτυξη της ML και της εξόρυξης δεδομένων αποτελούν τον ακρογωνιαίο λίθο για την υποστήριξη συστημάτων με ογκώδεις δεδομένα και την εξαγωγή αποτελεσμάτων με μεγαλύτερη ακρίβεια. Ο όρος ML αποτελεί κεντρικό κλάδο της Τεχνητής Νοημοσύνης (ΤΝ), εστιάζοντας στην ανάπτυξη αλγορίθμων και μαθηματικών μοντέλων με δυνατότητα να μαθαίνουν από δεδομένα χωρίς τον ρητό προκαθορισμένο προγραμματισμό τους για κάθε περίπτωση [23]. Τις τελευταίες δεκαετίες το ML χαρακτηρίζεται ως βασικό εργαλείο εξαγωγής γνώσης, αναγνώρισης προτύπων και υποστήριξης λήψης αποφάσεων σε ποικίλα επιστημονικά πεδία όπως η τεχνολογία, η ιατρική και η βιολογία [22]. Η υπεροχή της έναντι των παραδοσιακών στατιστικών μοντέλων βασίζεται στην ικανότητά της να διαχειρίζεται πολυπαραμετρικά δεδομένα και δεδομένα που έχουν μη γραμμικές συσχετίσεις.

Στο ίδιο οικοσύστημα μεθόδων εντάσσονται και τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ), τα οποία αποτελούν μοντέλα μάθησης εμπνευσμένα από τη λειτουργία του ανθρώπινου νευρικού συστήματος. Τα μοντέλα αυτά εκπαιδεύονται σε παραδείγματα εισόδου–εξόδου, μαθαίνοντας σύνθετες συσχετίσεις μεταξύ μεταβλητών και χρησιμοποιούνται εκτενώς σε προβλήματα ταξινόμησης και πρόβλεψης [23]. Μια ειδικότερη κατηγορία ΤΝΔ, είναι η Βαθιά Μάθηση (Deep Learning - DL), όπου τα νευρωνικά δίκτυα περιλαμβάνουν πολλαπλά ενδιάμεσα επίπεδα επιτρέποντας την αυτόματη εξαγωγή αναπαραστάσεων υψηλού επιπέδου από την αξιοποίηση σύνθετων δεδομένων όπως εικόνες, βίντεο και χρονοσειρές [24]. Μολονότι, αυτά τα μοντέλα απαιτούν μεγαλύτερα σύνολα δεδομένων και υπολογιστικούς πόρους, έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά και ικανά σε μη γραμμικά και σύνθετα προβλήματα.

Στον τομέα της υγειονομικής πληροφορικής, οι τεχνικές ML, συμπεριλαμβανομένων των ειδικών κατηγοριών ΤΝΔ και DL, χρησιμοποιούνται για την υποστήριξη κλινικών αποφάσεων, την πρόγνωση σημαντικών κινδύνων και την εξατομίκευση παρεμβάσεων, με στόχο την βελτίωση της ποιότητας ζωής και την μείωση ανεπιθύμητων καταστάσεων, όπως είναι οι πτώσεις. Αυτό επιτρέπεται χάρη στη συμβολή των τεχνικών της ευρύτερης κατηγορίας ML και της εξόρυξης δεδομένων· δύο προσεγγίσεων που σχετίζονται και χρησιμοποιούνται εγγενώς μεταξύ τους.

Στο πεδίο ενδιαφέροντος της εργασίας, αξιοποιούνται τεχνικές ML, κυρίως για συστήματα που επικεντρώνονται σε τρεις βασικούς σκοπούς: (α) ανίχνευση πτώσης, (β) πρόβλεψη/αξιολόγηση προφίλ κινδύνου πτώσης και (γ) υποστήριξη εξατομικευμένων παρεμβάσεων. Η κάθε μία περίπτωση αποφέρει διαφορετικά αποτελέσματα, όμως ο απώτερος σκοπός όλων είναι κοινός: η αποφυγή και μείωση νέων πτώσεων.

Η Μηχανική Μάθηση διακρίνεται σε τρεις βασικές κατηγορίες ανάλογα με τον σκοπό χρήσης της εφαρμογής και την διαθεσιμότητα των ετικετών στα δεδομένα [24]. Αυτές είναι η μη Επιβλεπόμενη Μάθηση (Unsupervised Learning - UL), η Επιβλεπόμενη Μάθηση (Supervised Learning - SL) και η Ενισχυτική Μάθηση (Reinforcement Learning - RL). Στο παρών ακαδημαϊκό κείμενο συναντώνται μόνο οι

δύο πρώτες προσεγγίσεις, καθώς αυτές χρησιμοποιήθηκαν στο πειραματικό σκέλος, απλώς αναφέρονται όλες για λόγους πληρότητας.

2.2.1 Μη Επιβλεπόμενη Μάθηση

Το UL χρησιμοποιείται όταν εκλείπουν οι ετικέτες από τα δεδομένα. Στοχεύει στην ανακάλυψη προτύπων και δομών στα δεδομένα. Είδη UL αποτελούν η συσταδοποίηση, οι τεχνικές μείωσης των πολλών διαστάσεων και οι τεχνικές κανόνων συσχέτισης. Η συσταδοποίηση είναι η διαδικασία ομαδοποίησης των αντικειμένων δεδομένων σύμφωνα μόνο με τις πληροφορίες που βρίσκονται στα δεδομένα και τις σχέσεις μεταξύ των αντικειμένων [22]. Στόχος είναι τα αντικείμενα που είναι όμοια μεταξύ τους να τοποθετηθούν στην ίδια ομάδα, ενώ αυτά που είναι διαφορετικά να βρίσκονται σε διαφορετικές ομάδες [22]. Πιο ειδικά στο πλαίσιο της εργασίας, το UL επιτρέπει την ομαδοποίηση ατόμων με παρόμοια χαρακτηριστικά, διευκολύνοντας την αναγνώριση διαφορετικών προφίλ κινδύνου πτώσης. Οι τεχνικές μείωσης των διαστάσεων βοηθούν στην εξάλειψη γνωρισμάτων που προσφέρουν μικρή σημασία στο σύνολο των δεδομένων αυξάνοντας την ερμηνευσιμότητά τους και διατηρώντας το μεγαλύτερο ποσοστό της χρήσιμης πληροφορίας. Παράδειγμα αποτελεί η τεχνική Ανάλυσης Κυρίων Συνιστωσών (Principal Component Analysis - PCA), η οποία χρησιμοποιείται για τη μείωση των δεδομένων με το να δημιουργεί νέα ορθογώνια - ασυσχέτιστα χαρακτηριστικά από τα ήδη υπάρχοντα, που διαδοχικά μεγιστοποιούν την διακύμανση των δεδομένων [25], [26]. Οι νέες ασυσχέτιστες μεταβλητές ονομάζονται και Κύριες Συνιστώσες (Principal Components). Συχνά, τέτοιες τεχνικές χρησιμοποιούνται και για την απεικόνιση των δεδομένων σε δισδιάστατη ή τρισδιάστατη μορφή που σε διαφορετική περίπτωση δε θα ήταν εφικτή λόγω των πολλών διαστάσεων. Τέλος η ανάλυση συσχέτισης είναι χρήσιμη για την ανακάλυψη ενδιαφερόντων σχέσεων μεταξύ των αντικειμένων δεδομένων και αναπαριστώνται με τη μορφή συνόλων στοιχείων [22]. Και τα τρία είδη UL χρησιμοποιούνται όταν δεν υπάρχουν ετικέτες στο σύνολο δεδομένων ή όταν απαιτείται η ανακάλυψη κρυμμένης πληροφορίας ανάμεσα στα δεδομένα. Σε κάθε περίπτωση όμως, η χρήση της αντίστοιχης τεχνικής εξυπηρετεί διαφορετικούς στόχους. Στην περίπτωση της εργασίας παρουσιάζονται μόνο τεχνικές συσταδοποίησης και μείωσης διαστασιμότητας.

Στη επόμενη ενότητα, αναλύονται οι αλγόριθμοι UL που εξετάστηκαν κατά την εκπόνηση του πειραματικού τμήματος της εργασίας, οι οποίοι είναι οι K-Means, Ιεραρχική Συσταδοποίηση (Hierarchical Clustering Analysis - HCA) και DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Η επιλογή των συγκεκριμένων αλγορίθμων έγινε με βάση τη χρήση διαφορετικών κατηγοριών συσταδοποίησης που μοντελοποιούν μοναδικά τη δομή των δεδομένων και την διασφάλιση της αξιοπιστίας της ανάλυσης, μειώνοντας την εξάρτηση από έναν και μόνο αλγόριθμο και διευκολύνοντας στην κατανόηση της ερμηνείας των αποτελεσμάτων.

2.2.2 Αλγόριθμοι Συσταδοποίησης που εξετάστηκαν

2.2.2.1 K-Means

Ο αλγόριθμος K-Means αποτελεί μία από τις πιο διαδεδομένες τεχνικές συσταδοποίησης λόγω της υπολογιστικής του αποδοτικότητας και της εννοιολογικής του απλότητας. Αποτελεί έναν από τους πρώτους αλγορίθμους συσταδοποίησης, καθώς πρωτοσυστάθηκε το 1955 [27]. Στόχος του είναι η κατάτμηση ενός συνόλου παρατηρήσεων σε k συστάδες, βελτιστοποιώντας την διασπορά εντός κάθε συστάδας με την ελαχιστοποίηση του αθροίσματος των τετραγωνικών αποστάσεων του κάθε σημείου από το κέντρο της

συστάδας του [27], [28]. Ο K-Means μοντελοποιεί τις συστάδες ως σφαιρικές περιοχές στον χώρο χαρακτηριστικών και προϋποθέτει ότι οι ομάδες έχουν παρόμοια πυκνότητα και μέγεθος.

Οι κύριες παράμετροι του K-Means είναι ο αριθμός συστάδων k, το κριτήριο απόστασης (συνήθως η Ευκλείδεια απόσταση) και η στρατηγική αρχικοποίησης των κέντρων των συστάδων (π.χ. k-means++). Η επιλογή του πλήθους συστάδων αποτελεί κρίσιμο βήμα για τον αλγόριθμο, ενώ θεωρείται και μειονέκτημά του. Διερευνητικές τεχνικές όπως το 'Elbow method', που εξετάζει την καμπύλη της ενδοσυσταδικής διασποράς, δηλαδή το πόσο συμπαγείς είναι τα σημεία των συστάδων ως προς τα κέντρα τους, μέσω του WCSS (Within Cluster Sum of Squares), ως προς το k και την οπτική επιβεβαίωση της σταθερότητας των συστάδων, είναι μία αξιόλογη μέθοδος εύρεσης του βέλτιστου k ως προς τα δεδομένα στα οποία χρησιμοποιείται [29]. Πλεονεκτήματα του αλγορίθμου είναι η ταχύτητα και η κλιμάκωση σε μεγάλα σύνολα δεδομένων. Αντιθέτως, είναι ευαίσθητο σε ακραίες τιμές, εξαρτάται από την αρχικοποίηση των κέντρων του και αδυνατεί να μοντελοποιήσει μη-σφαιρικές, άνισες και με διαφορετική πυκνότητα συστάδες.

2.2.2.2 Ιεραρχική Συσταδοποίηση

Η Ιεραρχική Συσταδοποίηση (HCA) αποτελεί μία οικογένεια αλγορίθμων που κατασκευάζουν ιεραρχική δομή συστάδων είτε με συσσωρευτική (agglomerative) είτε με διαιρετική (divisive) στρατηγική προσέγγιση [22]. Στην παρούσα εργασία εξετάστηκε η συσσωρευτική στρατηγική, στην οποία κάθε παρατήρηση του dataset ξεκινά ως αυτόνομη συστάδα και σταδιακά συγχωνεύεται με άλλες βάσει ενός κριτηρίου απόστασης και συνάρτησης σύνδεσης (linkage) [22]. Ο HCA μοντελοποιεί τις σχέσεις ομοιότητας μεταξύ των εγγραφών σε πολλαπλά επίπεδα ανάλυσης και δεν απαιτεί εκ των προτέρων καθορισμό του αριθμού συστάδων, γεγονός που τον καθιστά ιδιαίτερα χρήσιμο σε διερευνητικά πλαίσια, όπως η αναγνώριση προφίλ κινδύνου.

Βασικές παράμετροι του HCA είναι η μετρική απόστασης (π.χ. Ευκλείδεια, Manhattan) και η μέθοδος σύνδεσης (όπως single, complete, average, ward). Μία τεχνική αξιολόγησης των συστάδων για το HCA, είναι μέσω δενδρογράμματος που επιτρέπει την οπτική εξέταση της ιεραρχίας και τη διερεύνηση πιθανών σημείων τομής για την επιλογή του κατάλληλου αριθμού συστάδων [30],[31]. Πλεονεκτήματα της μεθόδου είναι η ευελιξία ως προς τη γεωμετρία των συστάδων, η απουσία ανάγκης προκαθορισμού πλήθος k συστάδων και η υψηλή ερμηνευσιμότητα της ιεραρχικής δομής, χάρη στην αναπαράστασή της μέσω του δενδρογράμματος. Ωστόσο, εμφανίζει αυξημένο υπολογιστικό κόστος για μεγάλα σύνολα δεδομένων, ευαισθησία σε θόρυβο και ακραίες τιμές, καθώς και ισχυρή εξάρτηση από την επιλογή μετρικής και συνάρτησης σύνδεσης [22].

2.2.2.3 DBSCAN

Σε αντίθεση με τις προηγούμενες δύο τεχνικές συσταδοποίησης, ο DBSCAN αποτελεί αλγόριθμο συσταδοποίησης βασισμένο στην πυκνότητα και είναι νεότερος καθώς παρουσιάστηκε επίσημα το 1996 [32]. Ορίζει τις συστάδες ως περιοχές υψηλής συγκέντρωσης σημείων που διαχωρίζονται από περιοχές χαμηλής πυκνότητας. Δεν απαιτεί εκ των προτέρων καθορισμό αριθμού συστάδων, και έχει την ιδιαιτερότητα να αναγνωρίζει σημεία θορύβου, αγνοώντας τα, χωρίς να τα αναθέτει σε κάποια συστάδα. Μοντελοποιεί συστάδες αυθαίρετου σχήματος, γεγονός που τον καθιστά ιδιαίτερα κατάλληλο για δεδομένα με μη γραμμική δομή ή με άνισες πυκνότητες [32], [33].

Κρίσιμες παράμετροι του αλγορίθμου αποτελούν η ακτίνα γειτνίασης (γνωστή ως 'eps') και ο ελάχιστος αριθμός σημείων (γνωστός ως 'minPts') που απαιτούνται για τον χαρακτηρισμό ενός σημείου, ως σημείο:

- πυρήνα
- ορίου
- θορύβου ή ακραίας τιμής

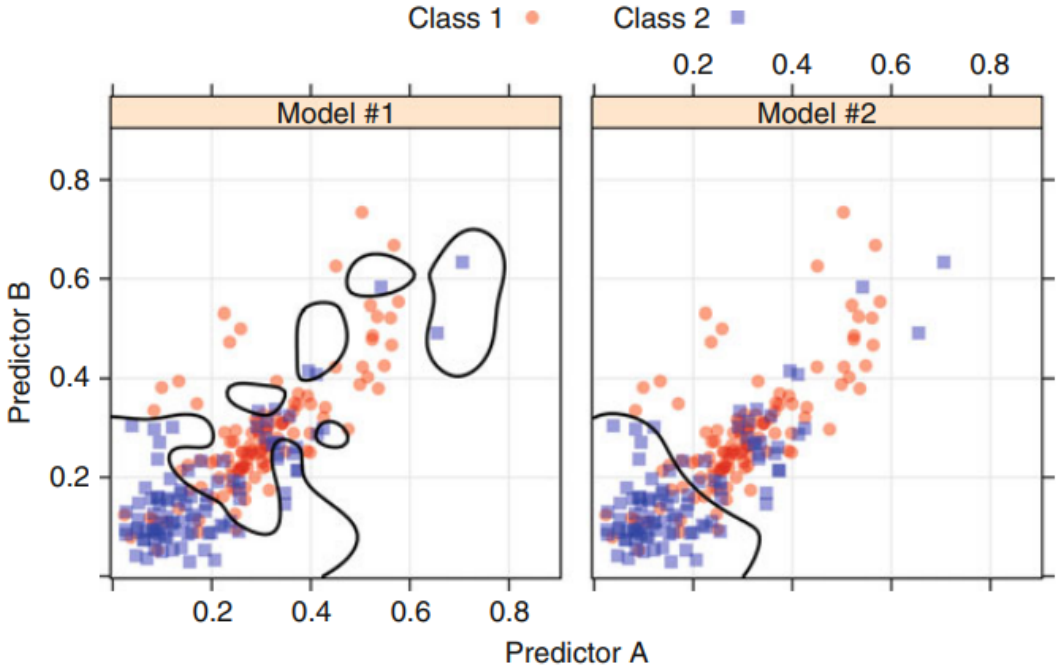
Η επιλογή των παραπάνω παραμέτρων συχνά υποστηρίζεται από τη γραφική παράσταση k-dist (γράφος k-distance), όπου εξετάζεται η κατανομή των αποστάσεων προς τον k-οστό γείτονα για τον εντοπισμό κατάλληλου σημείου καμπής, βρίσκοντας την κατάλληλη παράμετρο eps [32]. Πλεονεκτήματα του DBSCAN είναι η σχετική ανθεκτικότητα σε θόρυβο (ακραίες τιμές), η ικανότητα εντοπισμού συστάδων μη κανονικού σχήματος και μεγέθους και η απουσία ανάγκης για προκαθορισμένο k. Αντίθετα, παρουσιάζει ευαισθησία στην επιλογή των παραμέτρων και μειωμένη απόδοση σε μικρά σύνολα δεδομένων και με έντονα μεταβαλλόμενες πυκνότητες [32], [34]. Τέλος αντιμετωπίζει δυσκολίες σε δεδομένα υψηλής διαστασιμότητας, διότι η πυκνότητα είναι πιο δύσκολο να οριστεί για αυτά τα δεδομένα [22].

2.2.3 Επιβλεπόμενη Μάθηση

Το SL, βασίζεται σε σύνολα δεδομένων που συνοδεύονται από ετικέτες, αντιπροσωπεύοντας τον σωστό τύπο που ανήκει η κάθε παρατήρηση. Ανάλογα με την φύση της μεταβλητής στόχου, οι τύποι SL διακρίνονται σε προβλήματα ταξινόμησης (classification) και προβλήματα παλινδρόμησης (regression). Στην ταξινόμηση, στόχος είναι η ανάθεση κάθε παρατήρησης σε μία ή περισσότερες διακριτές κλάσεις, ενώ στην παλινδρόμηση επιδιώκεται η πρόβλεψη συνεχών αριθμητικών τιμών. Η διάκριση αυτή καθορίζει σε μεγάλο βαθμό την επιλογή της κατάλληλης οικογένειας αλγορίθμων. Οι αλγόριθμοι SL μπορούν να κατηγοριοποιηθούν περαιτέρω βάσει της θεωρητικής τους προσέγγισης και του τρόπου με τον οποίο μοντελοποιούν τη σχέση μεταξύ εισόδου και εξόδου. Ενδεικτικά, διακρίνονται σε γραμμικά μοντέλα, όπως η γραμμική και η λογιστική παλινδρόμηση, σε μοντέλα βασισμένα σε δέντρα αποφάσεων και σύνολα δέντρων, καθώς και σε αλγορίθμους βασισμένους σε περιθώρια και πυρήνες.

Σημαντικό κριτήριο επιλογής αλγορίθμων αφορά η ερμηνευσιμότητα και η γενίκευση του μοντέλου. Με τον όρο ερμηνευσιμότητα, υπονοείται ο βαθμός στον οποίο το αποτέλεσμα της εξόδου και της λογικής απόφασης ενός μοντέλου μπορεί να εξηγηθεί από τον χρήστη. Περαιτέρω ανάλυση της ερμηνευσιμότητας καλύπτεται στην ενότητα 2.4.4. Από την άλλη μεριά, ο όρος γενίκευση αναφέρεται στην ικανότητα ενός μοντέλου να ταξινομεί με ακρίβεια τόσο τα δεδομένα στα οποία εκπαιδεύεται όσο και στα άγνωστα, τα οποία χρησιμοποιούνται για να κάνει εκτίμηση και να αποδώσει την κατάλληλη τιμή κλάσης. Η γενίκευση αποτελεί θεμελιώδη στόχο της ML, καθώς διασφαλίζει ότι το μοντέλο δεν εξαρτάται υπερβολικά από συγκεκριμένα δείγματα εκπαίδευσης [24]. Συχνά προβλήματα που επηρεάζουν αρνητικά τη γενίκευση είναι η υπερπροσαρμογή (overfitting) και η υποπροσαρμογή (underfitting). Η υπερπροσαρμογή αφορά την περίπτωση που το μοντέλο έχει πολύ μεγάλη πολυπλοκότητα, μαθαίνοντας σε μεγάλο βαθμό το μοτίβο των δεδομένων που εκπαιδεύεται, ακόμη και θόρυβο, με αποτέλεσμα να αποτυγχάνει στην ορθή εκτίμηση άγνωστων δεδομένων [35]. Αντίθετα, η υποπροσαρμογή είναι η περίπτωση που το μοντέλο είναι τόσο απλό, που δεν καταφέρνει να αναπαραστήσει πλήρως την σχέση που υπάρχει μεταξύ των γνωρισμάτων και των τιμών ετικετών και επομένως έχει μικρή ολική ακρίβεια [22]. Για να γίνει πιο κατανοητή η διαφορά τους, υιοθετείται ένα παράδειγμα ταξινόμησης από την βιβλιογραφία [35]. Στο συγκεκριμένο παράδειγμα, η διακριτοποίηση των κλάσεων επιτυγχάνεται μέσω γραμμών οριοθέτησης. Στους άξονες χρησιμοποιούνται δύο ανεξάρτητες μεταβλητές (δηλαδή μεταβλητές που χρησιμοποιούνται για την πρόβλεψη). Σύμφωνα με το σχήμα 2.2, οι κλάσεις του μοντέλου που διακρίνονται είναι δύο και είναι σχετικά ισορροπημένες ως προς τον αριθμό των δειγμάτων που τις αντιπροσωπεύουν. Μολαταύτα, τα σημεία τους στον δισδιάστατο χώρο

είναι αρκετά επικαλυπτόμενα καθιστώντας δύσκολη την ταξινόμηση, αν και προσεγγίζουν δεδομένα που συναντώνται στον πραγματικό κόσμο. Όπως μπορεί να αντιληφθεί κανείς, στο αριστερό τμήμα του σχήματος 2.2, οι οριοθετημένες γραμμές είναι ιδιαίτερα πολύπλοκες, καθώς προσπαθούν να περιβάλλουν κάθε στιγμιότυπο της δεύτερης κλάσης (που υποδηλώνεται με χρώμα μπλε). Αφετέρου, το δεξί τμήμα της εικόνας 2.2, δείχνει τις γραμμές ορίου να είναι αρκετά ομαλές και να περιλαμβάνουν μόνο ένα τμήμα της κλάσης που στοχεύουν να περιβάλλουν, αγνοώντας πολλά στιγμιότυπα της κλάσης εκτός. Προφανώς, το αριστερό τμήμα αντιστοιχεί στην υπερπροσαρμογή, ενώ το δεξί στην υποπροσαρμογή ενός μοντέλου.



Σχήμα 2.2: Παράδειγμα απεικόνισης γραμμών οριοθέτησης μοντέλου [35]

Παρακάτω, περιγράφονται οι αλγόριθμοι SL που χρησιμοποιήθηκαν στην εργασία, αναλυτικά, προσδιορίζοντας τις ιδιότητές τους. Πρόκειται για τους εξής: Λογιστική Παλινδρόμηση, Δέντρα Απόφασης, Τυχαία Δάση, XGBoost, Μηχανές Διανυσμάτων Υποστήριξης και MLP ταξηνομητή.

2.2.4 Μοντέλα Επιβλεπόμενης Μάθησης που εξετάστηκαν

2.2.4.1 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση αποτελεί ένα από τα βασικότερα και περισσότερο μελετημένα μοντέλα επιβλεπόμενης μάθησης για δυαδική ταξινόμηση. Αν και η ονομασία της περιέχει τον όρο “παλινδρόμηση”, στην πραγματικότητα χρησιμοποιείται για προβλήματα όπου ο στόχος είναι η πρόβλεψη κατηγοριών, και όχι συνεχών τιμών. Στην ουσία, η Λογιστική Παλινδρόμηση μοντελοποιεί την πιθανότητα κατανομής ενός γεγονότος μέσω της λογιστικής συνάρτησης (sigmoid), η οποία μετασχηματίζει το γραμμικό συνδυασμό των εισόδων σε ένα πεδίο τιμών μεταξύ 0 και 1 (σχήμα Error: Reference source not found), επιτρέποντας την ανάθεση παρατηρήσεων σε διαφορετικές κλάσεις με βάση ένα κατώφλι πιθανότητας [36]

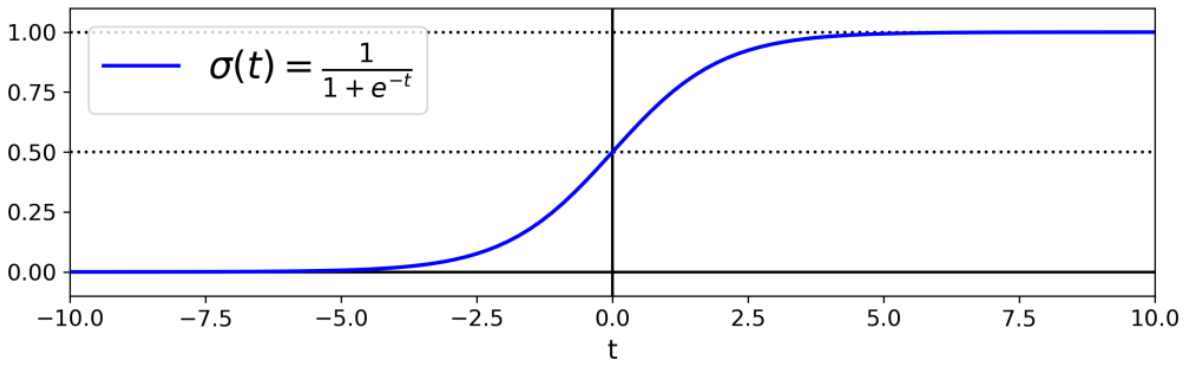
Η μορφή της ορίζεται όπως φαίνεται στην ισότητα**:**

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{2.1}$$

Όπου η μεταβλητή t αντιπροσωπεύει το γραμμικό συνδυασμό των χαρακτηριστικών εισόδου.

Ένα από τα βασικά πλεονεκτήματα του μοντέλου είναι η ερμηνευσιμότητά της, διότι οι παράμετροι που εκτιμώνται έχουν άμεση σχέση με το πώς η μεταβολή στην τιμή ενός χαρακτηριστικού επηρεάζει τις πιθανότητες κατηγοριοποίησης, δίνοντας στους ερευνητές και τους χρήστες του μοντέλου μια σαφή εικόνα για τις συσχετίσεις μεταξύ εισόδων και εξόδων [36]. Επιπλέον, η μέθοδος μπορεί να προσαρμοστεί ώστε να αντιμετωπίζει την πολυταξική ταξινόμηση μέσω χρήσης στρατηγικών πολυωνυμικής (ή Softmax) λογιστικής παλινδρόμησης ή One-vs-Rest, καθιστώντας την ευέλικτη σε διαφορετικά πλαίσια εφαρμογής [36].

Παρά τα πλεονεκτήματα, η Λογιστική Παλινδρόμηση παρουσιάζει και περιορισμούς. Η κύρια υπόθεση του μοντέλου είναι η γραμμικότητα στο λογιστικό χώρο, κάτι που σημαίνει ότι οι μη γραμμικές σχέσεις μεταξύ χαρακτηριστικών και εξόδου δεν μπορούν να συλληφθούν επαρκώς χωρίς την εφαρμογή προηγμένων τεχνικών μετασχηματισμού ή επέκτασης των χαρακτηριστικών. Επιπλέον το μοντέλο προϋποθέτει την μη ύπαρξη συγγραμικότητας μεταξύ εισόδων χαρακτηριστικών, γεγονός που μπορεί να επηρεάσει την σταθερότητα των παραμέτρων. Τέλος, δεν είναι ιδιαίτερα αποτελεσματική σε περιπτώσεις όπου οι σχέσεις μεταξύ μεταβλητών και αποτελέσματος είναι μη γραμμικές [37].



Σχήμα 2.3: Λογιστική ή Σιγμοειδής συνάρτηση [36]

2.2.4.2 Δέντρα Απόφασης

Τα Δέντρα Απόφασης (Decision Trees) αποτελούν μία από τις πιο διαισθητικές και ευρέως χρησιμοποιούμενες μεθόδους επιβλεπόμενης μάθησης, τόσο για ταξινόμηση όσο και για παλινδρόμηση, καθώς βασίζονται σε μια ιεραρχική διαδικασία λήψης αποφάσεων που προσομοιάζει τον ανθρώπινο τρόπο συλλογισμού. Η έννοια πρωτοσυστάθηκε το 1963 από τους Charles J. Clopper και Egon S. Pearson, ωστόσο η σύγχρονη υλοποίηση των Δέντρων Απόφασης, όπως τη γνωρίζουμε σήμερα, αναπτύχθηκε για πρώτη φορά το 1984, με τον αλγόριθμο CART [38]. Με το πέρασμα των ετών, κι άλλες διαφοροποιήσεις του αλγορίθμου συστάθηκαν, όπως οι ID3 [39] και C4.5 [40] μέχρι τα τέλη του 20ου αιώνα, οι οποίοι είναι ιδιαίτερα γνωστοί ακόμη και σήμερα. Ουσιαστικά, το γενικότερο μοντέλο αναπαριστά μία συνάρτηση που αντιστοιχίζει ένα διάνυσμα τιμών εισόδου χαρακτηριστικών, σε μία τιμή-στόχο εξόδου [41]. Η βασική ιδέα είναι η επαναλαμβανόμενη διάσπαση του χώρου των δεδομένων σε υποσύνολα, με σκοπό τη μεγιστοποίηση της ομοιογένειας των παρατηρήσεων σε κάθε τελικό κόμβο.

δομή ενός δέντρου αποτελείται από έναν ριζικό κόμβο, ενδιάμεσους κόμβους απόφασης και τελικούς κόμβους, όπου κάθε εσωτερικός κόμβος αντιστοιχεί σε έναν έλεγχο πάνω σε κάποιο χαρακτηριστικό και κάθε κλαδί αντιπροσωπεύει το αποτέλεσμα αυτού του ελέγχου [41]. Η επιλογή του χαρακτηριστικού σε κάθε κόμβο βασίζεται σε κριτήρια βελτιστοποίησης, όπως η μείωση της εντροπίας, η αύξηση της καθαρότητας των κλάσεων, το κέρδος πληροφορίας και ο δείκτης Gini, και όλα αυτά με στόχο τη βελτίωση της διαχωριστικής ικανότητας του μοντέλου.

Ένα από τα σημαντικότερα πλεονεκτήματα των Δέντρων Απόφασης είναι η ερμηνευσιμότητά τους [30]. Εύκολα μπορεί να ερμηνευθεί μία απόφαση, διασταυρώνοντας τους κόμβους από τη ρίζα έως και το φύλλο, χάρη στους κανόνες απόφασης που τα χαρακτηρίζουν. Μέσω των κριτηρίων βελτιστοποίησης, σε κάθε κόμβο επιλέγεται το χαρακτηριστικό με την υψηλότερη σπουδαιότητα [41], και μάλιστα επιλέγεται η κατάλληλη διάσπαση, εάν πρόκειται για αριθμητικό χαρακτηριστικό. Επιπλέον, πλεονέκτημα αποτελεί και το γεγονός ότι εφαρμόζεται εύκολα τόσο σε συνεχή όσο και σε κατηγορικά δεδομένα, χωρίς να απαιτείται η διαδικασία χρήσης μετασχηματισμών των γνωρισμάτων [22]. Για τον ίδιο λόγο δεν είναι απαραίτητη ούτε η κλιμάκωση των δεδομένων στο στάδιο της προεπεξεργασίας, παρ' όλα αυτά δε θα υπάρχει κάποια ιδιαίτερη επίπτωση εάν κλιμακωθούν.

Τα Δέντρα Απόφασης όμως, παρουσιάζουν και περιορισμούς, όπως η υψηλή ευαισθησία σε μικρές μεταβολές των δεδομένων, οι οποίες μπορούν να οδηγήσουν σε σημαντικά διαφορετική δομή δέντρου [30]. Επιπλέον, τα μεμονωμένα δέντρα συχνά αδυνατούν να συλλάβουν σύνθετες μη γραμμικές σχέσεις με την ίδια σταθερότητα που επιτυγχάνουν πιο σύνθετα σύνολα μοντέλων.

Για τον λόγο αυτό, τα Δέντρα Απόφασης χρησιμοποιούνται συχνά είτε ως μοντέλα αναφοράς είτε ως δομικά στοιχεία πιο ισχυρών μεθόδων συνόλων, όπως τα Τυχαία Δάση και τα σύνολα ενίσχυσης (Gradient Boosting) μοντέλα, τα οποία επιδιώκουν να μετριάσουν τα εγγενή τους μειονεκτήματα [24], [36].

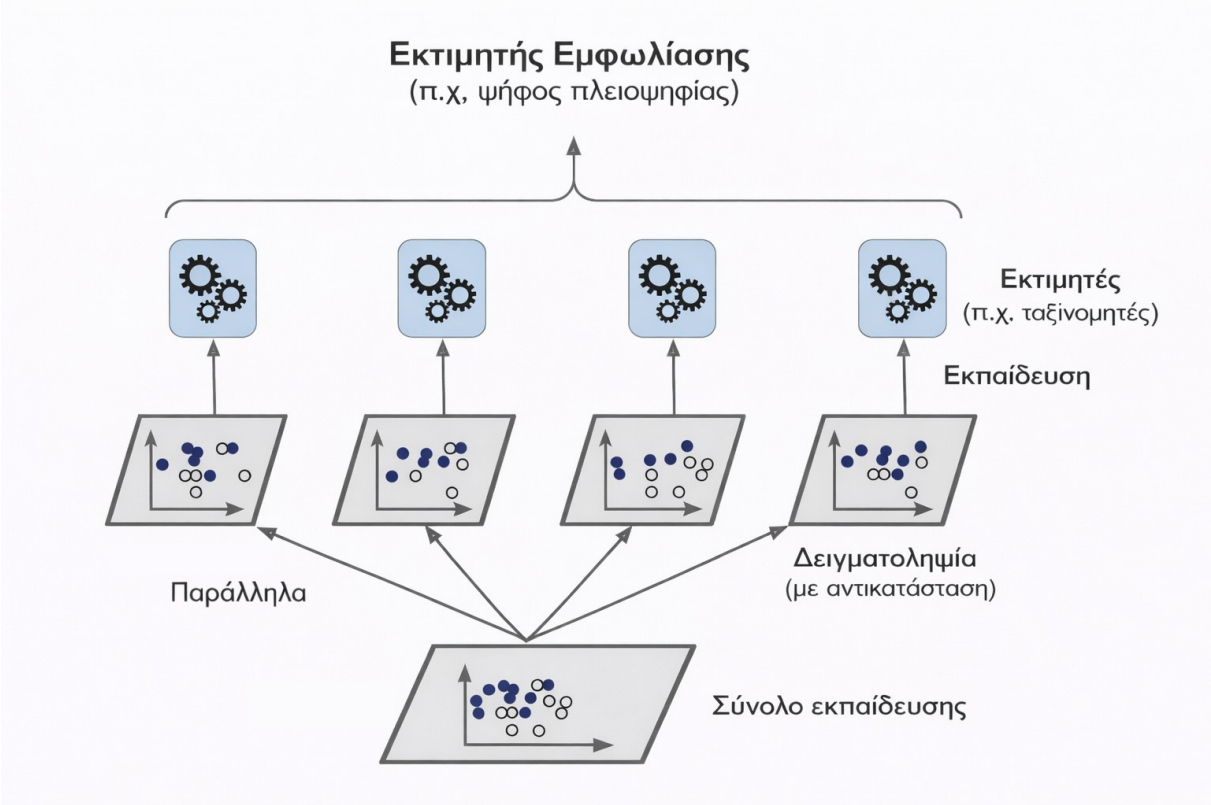
2.2.4.3 Τυχαία Δάση

Τα Τυχαία Δάση αποτελούν εξέλιξη των Δέντρων Απόφασης. Το μοντέλο αυτό είναι σχεδιασμένο για να αντιμετωπίσει τις εγγενείς αδυναμίες των μεμονωμένων δέντρων, όπως η υψηλή διακύμανση και η ευαισθησία στον θόρυβο των δεδομένων [36]. Ουσιαστικά, μειώνει την αστάθεια που χαρακτηρίζουν τα Δέντρα Απόφασης, καθώς μικρές αλλαγές στο σύνολο δεδομένων εκπαίδευσης, μπορεί να οδηγήσουν σε αισθητά διαφορετική δομή και απόδοση του εν λόγο μεμονωμένου δέντρου απόφασης.

Η βασική ιδέα της τεχνικής είναι ότι αντί να βασίζεται σε ένα μόνο δέντρο απόφασης, κατασκευάζεται ένα σύνολο από δέντρα όπου η τελική πρόβλεψη προκύπτει από τη συνάθροιση των αποτελεσμάτων των επιμέρους προβλέψεων των εκτιμητών [42]. Με αυτό τον τρόπο, συνδυάζει τις προβλέψεις πολλών διαφορετικών δέντρων, επιτυγχάνοντας πιο σταθερά και αξιόπιστα αποτελέσματα.

Η παραπάνω ιδέα αξιοποιεί την έννοια της εμφωλίασης (bagging) ή αλλιώς συνάθροιση αυτοδυναμίας (bootstrap aggregating). Πρόκειται για μια διαδικασία κατά την οποία δημιουργούνται παράλληλα πολλαπλά "νέα" σύνολα εκπαίδευσης μέσω μιας επαναληπτικής δειγματοληψίας (με αντικατάσταση) από το αρχικό σύνολο εκπαίδευσης. Σε κάθε bootstrap δείγμα εκπαιδεύεται ένας "ασθενής" εκτιμητής, και οι προβλέψεις συνδυάζονται με το μέσο όρο (παλινδρόμηση) ή τη πλειοψηφική ψήφο (ταξινόμηση) (βλέπε σχήμα 2.4). Ο τελικός εκτιμητής είναι μία συνάθροιση όλων των "ασθενών" εκτιμητών [22], [43]. Να σημειωθεί ότι, "ασθενής" εκτιμητής θεωρείται αυτός που αποδίδει σχετικά καλύτερα από μία τυχαία ταξινόμηση (συνήθως πάνω από 50% απόδοση) [44]. Δίνοντας ένα παράδειγμα ταξινόμησης, κάθε δέντρο του Τυχαίου Δάσους είναι ανεξάρτητο, άρα παράγει μια ανεξάρτητη πρόβλεψη κλάσης για κάθε δείγμα. Η τελική απόφαση

λαμβάνεται με βάση την κλάση που συγκεντρώνει τις περισσότερες ψήφους από το σύνολο των δέντρων [43]. Καθόσον μεγαλώνει ο αριθμός προβλέψεων και σε συνδυασμό με την αξιοποίηση διαφορετικών συνόλων εκπαίδευσης σε κάθε βασικό μοντέλο, το τελικό μοντέλο καθίσταται πιο αξιόπιστο, μειώνοντας τη διακύμανση του σφάλματος μεταξύ των εκτιμητών, ιδιαίτερα σε σχέση ενός μόνο εκτιμητή, όπως ενός μεμονωμένου Δέντρου Απόφασης. Συνεπώς, βελτιώνεται η σταθερότητα του τελικού μοντέλου.



Σχήμα 2.4: Επισκόπηση διαδικασίας εμφωλίασης (πηγή έμπνευσης: [36])

Στα Τυχαία Δάση όμως, η εμφωλίαση αποτελεί το βασικό συστατικό, και δεν είναι αφορά όλη τη διαδικασία. Η τεχνική των Τυχαίων Δασών περιλαμβάνει ένα επιπλέον επίπεδο τυχαιότητας, την διαχείριση των γνωρισμάτων εισόδου [22], [43]. Στη πράξη, σε κάθε διάσπαση κόμβου επιλέγονται προς εξέταση είτε τυχαία είτε με κάποια υπόδειξη εντός του πεδίου εργασίας, υποσύνολα χαρακτηριστικών για τον σχηματισμό του αντίστοιχου συνόλου εκπαίδευσης του βασικού κατηγοριοποιητή Δέντρου Απόφασης.

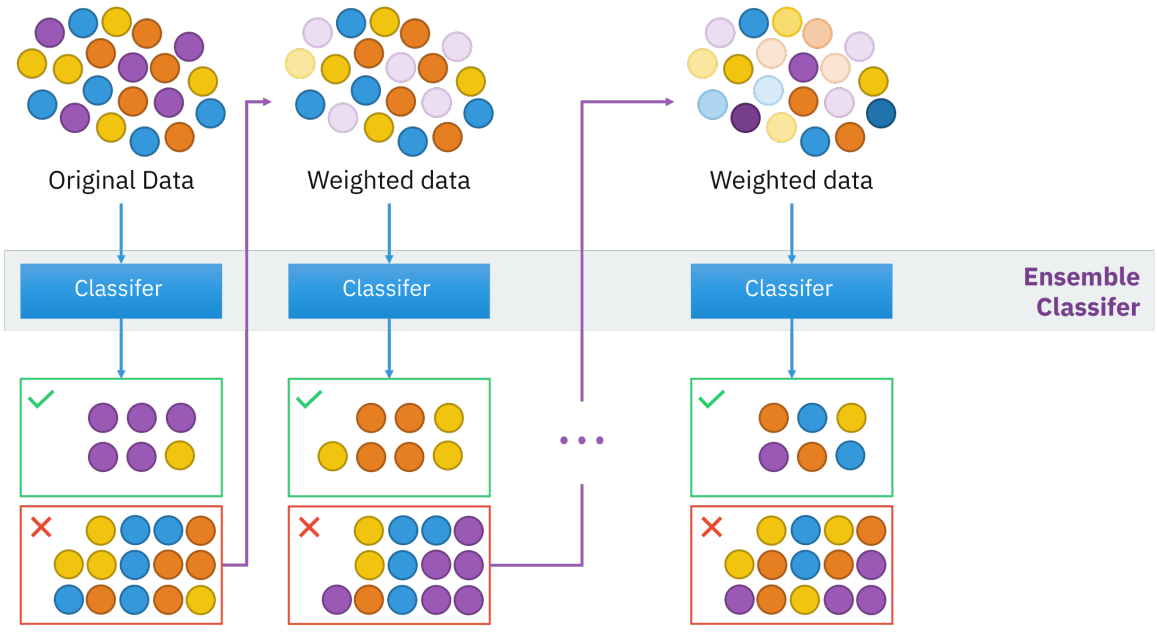
Ένα από τα κύρια πλεονεκτήματα του εκτιμητή λοιπόν, είναι η υψηλή ακρίβεια και η ανθεκτικότητα στον θόρυβο και στην υπερπροσαρμογή [42]. Επιπλέον, μπορούν να χειριστούν σύνολα δεδομένων υψηλής διαστασιμότητας και να παρέχουν εκτιμήσεις σημαντικότητας χαρακτηριστικών, οι οποίες είναι χρήσιμες για ανάλυση και ερμηνεία [36].

Τα μειονεκτήματα των Τυχαίων Δασών είναι το δαπανηρό υπολογιστικό κόστος σε σύγκριση σε ένα απλό Δέντρο Απόφασης τόσο σε χρόνο εκπαίδευσης όσο και σε απαιτήσεις μνήμης. Η υπολογιστική απαίτηση αυξάνεται λόγω του πλήθους δέντρων, ιδιαίτερα όταν γίνεται εκτεταμένη αναζήτηση υπερπαραμέτρων ή όταν ο αριθμός χαρακτηριστικών είναι μεγάλος [36]. Ακόμη, το μοντέλο παρουσιάζει δυσκολότερη άμεση ερμηνεία σε σχέση με ένα απλό δέντρο απόφασης.

Εν κατακλείδι, τα Τυχαία δάση επιτυγχάνουν τη μείωση της διακύμανσης σφαλμάτων και την αξιόπιστη απόδοση του τελικού μοντέλου, το οποίο έχει δοκιμαστεί και αποδοθεί, μέσα από μία γκάμα συνόλου ομάδων εκτιμητών Δέντρων Απόφασης, σε διαφορετικά σύνολα εκπαίδευσης κάθε φορά. Χάρη στην εγγενή χρήση της εμφωλίασης και της περαιτέρω ενίσχυσης τυχαιότητας στα χαρακτηριστικά εισόδου, μειώνεται η συσχέτιση μεταξύ των επιμέρους δέντρων, ενισχύοντας την γενίκευση της τεχνικής συνόλου ομάδας. Συνεπώς, επιλέγεται εκείνο το μοντέλο, που παρέχει μεγαλύτερο αντίκτυπο στο πλήρες σύνολο δεδομένων. Όμως, έχει πιο δύσκολη ερμηνεία των γνωρισμάτων του και πιο μεγάλο υπολογιστικό κόστος εκπαίδευσης σε σχέση με ένα απλό μεμονωμένο δέντρο.

2.2.4.4 XGBoost

Μία εναλλακτική φιλοσοφία των συνόλων ομάδων αποτελούν οι μέθοδοι ενίσχυσης (boosting). Σε αντίθεση με την εμφωλίαση, οι μέθοδοι αυτοί βασίζονται στην διαδοχική εκπαίδευση των "ασθενών" εκτιμητών. Ενώ η εμφωλίαση, όπως τα Τυχαία Δέντρα, στοχεύει κυρίως στη μείωση της διακύμανσης μέσω ανεξάρτητων προβλεπτών [43], η ενίσχυση επικεντρώνεται στη σταδιακή μείωση του σφάλματος μαθαίνοντας από τα λάθη των προηγούμενων μοντέλων. Έτσι, επικεντρώνεται σε δείγματα τα οποία είναι πιο δύσκολο να εκτιμηθούν, κατά τη διάρκεια της εκπαίδευσης, όπως απεικονίζεται στη εικόνα 2.5. Η πραγματοποίηση αυτή, συμβαίνει χάρη στην απόδοση κάποιου βάρους σε κάθε δείγμα εκπαίδευσης, σταθμίζοντας με αυτό τον τρόπο κατάλληλα τα βάρη των δειγμάτων μετά το πέρας της κάθε επανάληψης [22].



Σχήμα 2.5: Επισκόπηση διαδικασίας ενίσχυσης [45]

Έχουν αναπτυχθεί διάφορες μορφές μοντέλων ενίσχυσης. Η κύρια διαφορά τους οφείλεται στον τρόπο που αποδίδονται τα βάρη μεταξύ των επαναλήψεων (σταθμίζοντας τα σφάλματα) και τον τρόπο που συνδυάζονται οι προβλέψεις του κάθε εκτιμητή [22], [46]. Μερικοί από αυτούς είναι οι AdaBoost, Gradient Boosting Machines (GBM), XGBoost, CatBoost. Ο αλγόριθμος ενίσχυσης που χρησιμοποιείται στην παρούσα εργασία είναι ο XGBoost.

Ο XGBoost (eXtreme Gradient Boosting), είναι μία από τις πιο αποδοτικές και ευρέως χρησιμοποιούμενες υλοποιήσεις ενίσχυσης στη σύγχρονη πρακτική ML, χάρη στην επεκτασιμότητα, την αποδοτικότητα και τον έλεγχο υπερπροσαρμογής του [47]. Υπάρχει υλοποιημένη βιβλιοθήκη ανοικτού κώδικα λογισμικού, η οποία έχει το ίδιο όνομα με τον αλγόριθμο. Η βασική λειτουργία του αλγορίθμου συνίσταται στην εκπαίδευση μιας ακολουθίας δέντρων, όπου κάθε νέο δέντρο προσεγγίζει το υπολειπόμενο σφάλμα του συνόλου των προηγούμενων μοντέλων [46]. Σε αντίθεση με απλούστερες προσεγγίσεις ενίσχυσης, ο XGBoost ενσωματώνει ρητά όρους κανονικοποίησης στη συνάρτηση κόστους, επιτρέποντας καλύτερο έλεγχο της πολυπλοκότητας των δέντρων [47]

Ο XGBoost εμφανίζει ιδιαίτερα υψηλή προβλεπτική απόδοση σε πληθώρα προβλημάτων, γεγονός που επιβεβαιώνεται από την εκτεταμένη χρήση του σε ανταγωνιστικά περιβάλλοντα και επιστημονικές εφαρμογές [47]. Υπάρχουν ακόμη δυνατότητες, υποστήριξης παραλληλισμού, αποδοτικής διαχείρισης της μνήμης καθώς και αυτόματος χειρισμός των απουσών τιμών ενός συνόλου δεδομένων. Παράλληλα, η βιβλιοθήκη της, παρέχει συμβατό API με αυτό της βιβλιοθήκης scikit-learn, που είναι η βασική βιβλιοθήκη για μηχανισμούς ML, διευκολύνοντας την ομαλή ενσωμάτωσή της σε ροές ML.

Στα μειονεκτήματα της τεχνικής ανήκει αυξημένη πολυπλοκότητα στη ρύθμιση υπερπαραμέτρων και το πλήθος των παραμέτρων που την αποτελούν, γεγονός που χρήζει συστηματική πειραματική διαδικασία. Επίσης, η ερμηνευσιμότητα του μοντέλου είναι περιορισμένη σε σύγκριση με απλούστερα μοντέλα ή μεμονωμένα δέντρα, καθώς η τελική απόφαση προκύπτει από μεγάλο αριθμό αλληλεπιδρώντων δέντρων [24].

Συνεπώς, ο XGBoost βασίζεται σε ακολουθιακή μάθηση, όπου κάθε επόμενο (νέο) δέντρο εξαρτάται από το προηγούμενο. Αποτελεί ισχυρότερο αλλά και πιο απαιτητικό εργαλείο, από τα Τυχαία Δάση, ιδίως σε εφαρμογές όπου η μέγιστη δυνατή απόδοση υπερισχύει της απλότητας και της ερμηνευσιμότητας.

2.2.4.5 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs) αποτελούν μια κατηγορία επιβλεπόμενων μοντέλων ML και χρησιμοποιούνται τόσο για προβλήματα ταξινόμησης όσο και παλινδρόμησης. Από τη δεκαετία του 1990 ο αλγόριθμος έλαβε σημαντική δημοσιότητα χάρη στην εφεύρεση του μη-γραμμικού μοντέλου, από τους Vapnik και Cortes [48] και πλέον αποτελεί ένα από τα πιο μελετημένα μοντέλα όσον αφορά τα πλαίσια στατιστικής μάθησης. Στην εργασία χρησιμοποιήθηκε ο αλγόριθμος SVC του πακέτου ‘svm’ της βιβλιοθήκης scikit-learn.

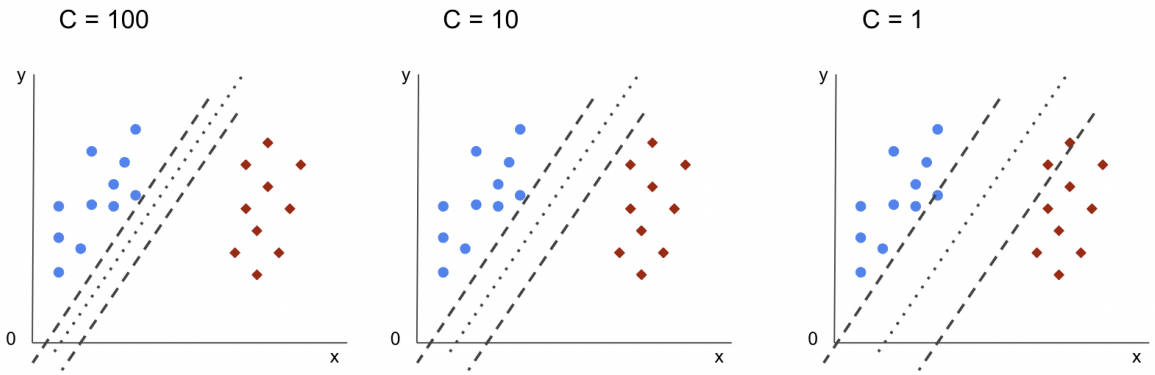
Μερικές από τις βασικές ιδιότητες των SVMs είναι οι παρακάτω:

- Κατασκευάζει έναν διαχωριστή μέγιστου περιθωρίου, επιδιώκοντας να βρει τον καλύτερο. Ο διαχωριστής στη ουσία είναι ένα σύνορο απόφασης μεταξύ των σημείων δεδομένων διαφορετικών κλάσεων, και σκοπός είναι η εύρεση αυτού που εμφανίζει τη μεγαλύτερη απόσταση μεταξύ των σημείων αυτών [48], [30]. Τα σημεία που βρίσκονται πιο κοντά στο περιθώριο, ονομάζονται διανύσματα υποστήριξης (support vectors).
- Η βασική μορφή του αφορά γραμμικά διαχωρίσιμα δεδομένα, όπου το σύνορο απόφασης είναι γραμμικό (π.χ. μία ευθεία σε δύο διαστάσεις ή υπερεπίπεδο σε περισσότερες διαστάσεις) και μπορεί εύκολα να διαχωρίσει πλήρως τα σημεία στον χώρο, χωρίς σφάλματα, ως προς τις κλάσεις που υπάρχουν. Ωστόσο, επεκτείνεται και σε μη γραμμικές περιπτώσεις μέσω του τεχνάσματος πυρήνα (kernel trick) [41], [49], το οποίο επιτρέπει την αποδοτική αντιστοίχιση των δεδομένων σε χώρο

υψηλότερης διάστασης όπου μπορεί να χρησιμοποιηθεί γραμμικός διαχωριστής (π.χ. μία παραβολή). Τεχνικές για μη γραμμικά διαχωρίσιμα προβλήματα αποτελούν οι συναρτήσεις: 'Radial Basis Function' (RBF), 'Polynomial' και 'Sigmoid'.

- Περιέχει παραμέτρους για την καταπολέμηση της υπερπροσαρμογής. Υπάρχουν αρκετές παράμετροι που χρησιμοποιούνται σαν κανονικοποίηση (regularization), όπως οι: C, gamma, degree. Οι τιμές των παραμέτρων εξαρτώνται από το σύνολο δεδομένων και χρειάζεται να προσαρμόζονται κατάλληλα στην κάθε περίσταση. Αναλύονται οι χρήσεις μερικών απ'αυτών, παρακάτω.

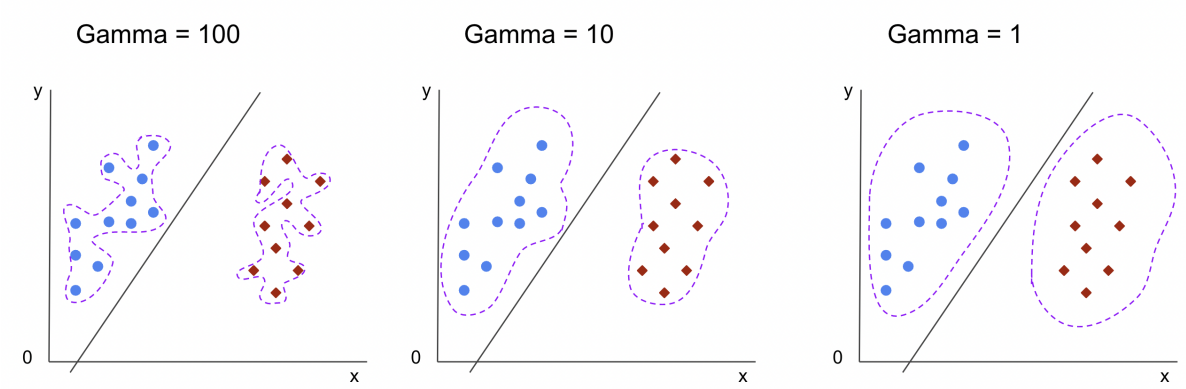
Η παράμετρος C είναι μία από τις κύριες, και ελέγχει τον συμβιβασμό μεταξύ μεγιστοποίησης του περιθωρίου και ελαχιστοποίησης του σφάλματος κατά την διάρκεια εκπαίδευσης του μοντέλου [36]. Σύμφωνα με το σχήμα 2.6 για μικρές τιμές του C, τα σύνορα απόφασης είναι πιο απλά δημιουργώντας μεγαλύτερο περιθώριο μεταξύ των δεδομένων των κλάσεων, ενώ για μεγάλες τιμές του C, το μοντέλο προσαρμόζεται καλύτερα πάνω στα δεδομένα εκπαίδευσης, αυξάνοντας την πιθανότητα κινδύνου υπερπροσαρμογής.



Σχήμα 2.6: Απεικόνιση επίδρασης της παραμέτρου C στα σύνορα απόφασης του μοντέλου SVM [50]

Αντίστοιχα λειτουργεί η παράμετρος που είναι γνωστή με την ονομασία, gamma ή sigma. Η συγκεκριμένη παράμετρος επηρεάζει την κλίμακα επιλογής του συνόρου απόφασης. Όπως εμφανίζεται στην εικόνα 2.7, για μεγαλύτερη τιμή του gamma, το σύνορο απόφασης εφαρμόζεται πιο στενά πάνω στα σημεία δεδομένων, ενώ για μικρότερη τιμή είναι πιο απλή η εφαρμογή. Εύκολα μπορεί να διαπιστωθεί ότι για μεγάλη τιμή του gamma, το μοντέλο είναι πιο επιρρεπές στην υπερπροσαρμογή.

Τα SVMs αρχικά σχεδιάστηκαν ως δυαδικοί ταξινομητές, όπου το ζητούμενο είναι ο διαχωρισμός των δεδομένων σε δύο κλάσεις μέσω ενός μοναδικού υπερεπιπέδου (συνόρου) απόφασης [48]. Η απόφαση λαμβάνεται βάσει του προσήμου της συνάρτησης απόφασης, η οποία εκφράζει την απόσταση ενός σημείου από το όριο του διαχωρισμού [51]. Λόγω της ύπαρξης προβλημάτων όπως η αναγνώριση χαρακτήρων κειμένου, η αναγνώριση προσώπων, η ταξινόμηση σε Αργότερα όμως, επεκτάθηκαν σε προβλήματα με περισσότερες από δύο κλάσεις, με βάση δύο προσεγγίσεων, τους One-vs-Rest (OvR) και One-vs-One (OvO) [22].



Σχήμα 2.7: Απεικόνιση επίδρασης της παραμέτρου gamma στα σύνορα απόφασης του μοντέλου SVM [50]

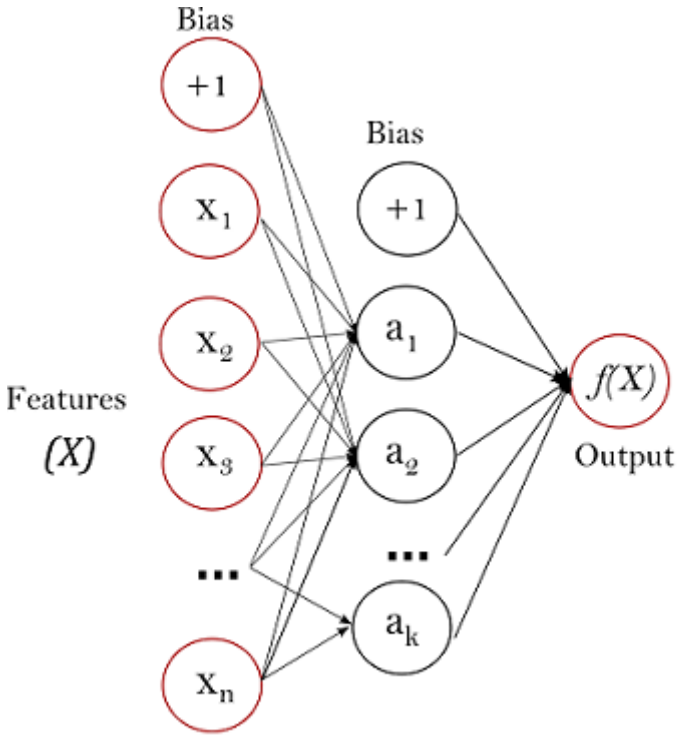
Η προσέγγιση One-vs-Rest κατασκευάζει έναν δυαδικό ταξινομητή για κάθε κλάση έναντι όλων των υπολοίπων, με την τελική απόφαση να προκύπτει από τη μέγιστη τιμή της συνάρτησης απόφασης [52]. Αντίθετα, η προσέγγιση One-vs-One εκπαιδεύει ταξινομητές για κάθε ζεύγος κλάσεων, με την τελική ετικέτα να καθορίζεται μέσω πλειοψηφικής ψήφου μεταξύ των επιμέρους μοντέλων. [52], [30] Η επιλογή μεταξύ One-vs-Rest και One-vs-One επηρεάζει τόσο το υπολογιστικό κόστος όσο και τη σταθερότητα της ταξινόμησης, ιδιαίτερα σε προβλήματα με μεγάλο αριθμό κλάσεων [53]. Επιπλέον, για τα προβλήματα με πολλές κατηγορίες, υπάρχει η δυνατότητα εξαγωγής της πιθανότητας της κάθε κλάσης [54].

Συνολικά, τα μοντέλα SVMs αποτελούν αποτελούν ισχυρά μοντέλα ταξινόμησης, με υψηλή ικανότητα γενίκευσης και ευελιξία μέσω των πυρήνων και των παραμέτρων κανονικοποίησης. Η χρήση του κατάλληλου πυρήνα και των παραμέτρων (όπως και των τιμών τους), εξαρτάται απόλυτα από το είδος των δεδομένων και την αναπαράστασή τους στον χώρο.

2.2.4.6 MLP

Ο Multi-Layer Perceptron Classifier (MLPClassifier ή MLP), είναι ακόμη ένας αλγόριθμος SL ο οποίος όμως ανήκει στην κατηγορία των Νευρωνικών Δικτύων. Αποτελεί επέκταση του αρχικού perceptron, που προτάθηκε στα τέλη της δεκαετίας του 1950 ως γραμμικός ταξινομητής, με στόχο την προσομοίωση βασικών λειτουργιών του ανθρώπινου νευρικού συστήματος [55]. Η επέκταση υφίσταται στην εισαγωγή πολλαπλών κρυφών επιπέδων και μη γραμμικών συναρτήσεων ενεργοποίησης, που επέτρεψε στον αλγόριθμο MLP να υπερβεί τους περιορισμούς του γραμμικού διαχωρισμού, καθιστώντας τον ικανό να προσεγγίζει σύνθετες μη γραμμικές συναρτήσεις [24].

Ο αλγόριθμος MLP αποτελείται από ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα, και ένα επίπεδο εξόδου, όπου κάθε επίπεδο περιλαμβάνει τεχνητούς νευρώνες πλήρως συνδεδεμένους με το επόμενο επίπεδο [23].



Σχήμα 2.8: Νευρωνικό Δίκτυο MLP με 1 κρυφό επίπεδο [56]

Κάθε νευρώνας έχει βάρη, που υπολογίζονται από το σταθμισμένο άθροισμα των εισόδων του με την εφαρμογή μιας μη γραμμικής συνάρτησης ενεργοποίησης, όπως οι ReLU, tanh και logistic. Έτσι, επιτυγχάνεται η μοντελοποίηση σύνθετων σχέσεων μεταξύ των χαρακτηριστικών του κάθε στιγμιότυπου του συνόλου δεδομένων [24], [30]. Η εκπαίδευση του MLP πραγματοποιείται μέσω του αλγορίθμου backpropagation [23], [57], ο οποίος προσαρμόζει τα βάρη ελαχιστοποιώντας μια συνάρτηση κόστους με χρήση μεθόδων βαθμιαίας καθόδου.

Στην εργασία, χρησιμοποιείται ο αλγόριθμος MLPClassifier της βιβλιοθήκης scikit-learn, η οποία υλοποιεί έναν πλήρως συνδεδεμένο MLP για προβλήματα ταξινόμησης, παρέχοντας μια σχετικά απλή αλλά ισχυρή διεπαφή για πειραματισμό με νευρωνικά δίκτυα. Η υλοποίηση υποστηρίζει διαφορετικές αρχιτεκτονικές (πλήθος και μέγεθος κρυφών επιπέδων), συναρτήσεις ενεργοποίησης και αλγορίθμους βελτιστοποίησης, οι οποίοι είναι οι ‘stochastic gradient descent’ (‘sgd’), ‘adam’ και ‘lbfgs’. Ο κάθε αλγόριθμος, βελτιστοποιεί τα βάρη μεταξύ των επιπέδων νευρωνικών δικτύων με διαφορετικό τρόπο. Η κατάλληλη επιλογή βελτιστοποιητή εξαρτάται από το σύνολο των δεδομένων. Γενικώς, ο MLPClassifier της βιβλιοθήκης scikit-learn προορίζεται κυρίως για μικρά και μεσαία σύνολα δεδομένων, όπου η απλότητα και η ενσωμάτωση σε κλασικές ροές ML υπερισχύουν της μέγιστης απόδοσης [36]. Η μελέτη και εφαρμογή της εργασίας αυτής, όπως προκύπτει από τα παρακάτω κεφάλαια, αντιστοιχούν σε αυτή τη περίπτωση, η οποία συμβιβάζεται με τη βιβλιοθήκη που επιλέχθηκε.

Όπως και στα υπόλοιπα μοντέλα, έτσι και ο MLP κατέχει μηχανισμούς αποφυγής της υπερπροσαρμογής μέσω τεχνικών κανονικοποίησης (regularization) με κυριότερο την παράμετρο ‘alpha’, η οποία περιορίζει το μέγεθος των βαρών του δικτύου. Επιπροσθέτως, η επιλογή του αριθμού των κρυφών επιπέδων και των

νευρώνων ανά επίπεδο επηρεάζει την πολυπλοκότητα του μοντέλου και συνεπώς ρυθμίζει την υπερπροσαρμογή. Επιπλέον, υπάρχουν κι άλλες παράμετροι, όπως το 'early stopping' η οποία διακόπτει την εκπαίδευση όταν η απόδοση παύει να βελτιώνεται, και το 'learning_rate' που καθορίζει τον βαθμό που γίνονται αλλαγές στα βάρη, βελτιώνοντας την γενίκευση και καθιστώντας ταχύτερη την σύγκλιση αντίστοιχα.

Ο MLPClassifier μπορεί να χρησιμοποιηθεί τόσο για δυαδική όσο και για πολυταξική ταξινόμηση, ανάλογα με τη μορφή του επιπέδου εξόδου και της συνάρτησης κόστους. Στη δυαδική περίπτωση, το επίπεδο εξόδου περιλαμβάνει συνήθως έναν νευρώνα με συνάρτηση ενεργοποίησης 'sigmoid', ενώ στη περίπτωση πολλαπλών τάξεων, χρησιμοποιείται η συνάρτηση softmax, η οποία παράγει πιθανότητες για κάθε κλάση [24],[30]. Σε αντίθεση με μοντέλα όπως οι SVMs, εδώ για την πολυταξική ταξινόμηση δε χρησιμοποιούνται στρατηγικές One-vs-Rest ή One-vs-All, καθώς πραγματοποιείται εγγενώς μέσω της συνάρτησης ενεργοποίησης [30].

Ένα μοντέλο MLP απαιτεί μία διαδικασία ρύθμισης των υπερπαραμέτρων του, όπως αυτό του πλήθος νευρώνων και των επιπέδων που αποτελείται αλλά και της παραμέτρου 'alpha', , καθώς περιλαμβάνει μία μεγάλη συλλογή από υπερπαραμέτρους. Παράλληλα, είναι ιδιαίτερα ευαίσθητο σε διαφορετικές κλίμακες μεταξύ των δεδομένων των χαρακτηριστικών του συνόλου δεδομένων, με αποτέλεσμα, επίσης να απαιτεί την χρήση τεχνικών κλιμάκωσης γνωρισμάτων (feature scaling) κατά τη φάση της προεπεξεργασίας των δεδομένων. Ακόμη, χαρακτηρίζεται από χαμηλή ερμηνευσιμότητα, διότι η εξαγωγή των αποτελεσμάτων του, προκύπτει από την σύνθετη αλληλεπίδραση πολλαπλών επιπέδων νευρώνων και μη γραμμικών συναρτήσεων ενεργοποίησης, πράγμα που δυσχεραίνει την κατανόηση της συνεισφοράς των επιμέρους χαρακτηριστικών [24].

Από τα παραπάνω προκύπτει ότι ο MLPClassifier αποτελεί έναν ευέλικτο επιβλεπόμενο ταξινομητή, ικανό να μοντελοποιήσει μη γραμμικές σχέσεις εντός των δεδομένων του. Παρότι υστερεί σε κλίμακα και εξειδίκευση σε σχέση με πιο σύγχρονα μοντέλα βαθιάς μάθησης, προσφέρει τη δυνατότητα να χρησιμοποιείται για εφαρμογές με περιορισμένο μέγεθος δεδομένων και ανάγκη ενσωμάτωσης σε κλασικές ροές ML.

2.3 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί θεμελιώδες στάδιο στη διαδικασία της ML, καθώς καθορίζει σε μεγάλο βαθμό την ποιότητα της πληροφορίας που παρέχεται στα μοντέλα και κατ' επέκταση, την αξιοπιστία των αποτελεσμάτων τους. Τα δεδομένα που συλλέγονται από πραγματικά συστήματα, ιδίως στον χώρο της υγείας και της ανάλυσης ανθρώπινης κίνησης, είναι συχνά ατελή, ετερογενή, θορυβώδη και ασύμβατα με τις μαθηματικές υποθέσεις των αλγορίθμων. Στη βιβλιογραφία αναφέρεται ότι στις περισσότερες περιπτώσεις πραγματικών εφαρμογών το στάδιο της προεπεξεργασίας είναι το πιο χρονοβόρο, προσεγγίζοντας το 60-80% του συνολικού κύκλου ζωής ενός έργου [58 - 60]. Η προεπεξεργασία λειτουργεί ως ενδιάμεσο στάδιο μετασχηματισμού, μετατρέποντας τα ακατέργαστα δεδομένα σε μορφές κατάλληλες για ανάλυση και μάθηση.

Περιλαμβάνεται ένα σύνολο τεχνικών που στοχεύουν στη βελτίωση της ποιότητας, της αναπαράστασης και της πληροφορικής πυκνότητας των δεδομένων. Οι τεχνικές αυτές επιλέγονται ανάλογα με το είδος των μεταβλητών, τη δομή του προβλήματος και τον αλγόριθμο που πρόκειται να χρησιμοποιηθεί [36].

Είναι σημαντικό να αναφερθεί η ανάγκη της μείωσης των διαστάσεων ενός συνόλου δεδομένων. Ένα συχνό φαινόμενο που παρατηρείται όταν ο αριθμός των χαρακτηριστικών ενός συνόλου δεδομένων είναι αρκετά μεγάλος σε σχέση με το πλήθος των διαθέσιμων δειγμάτων, ονομάζεται «κατάρα των πολλών διαστάσεων» [22], [23], [30]. Σε μία τέτοια περίπτωση, ο χώρος των δεδομένων γίνεται πολύ αραιός με αποτέλεσμα τα στιγμιότυπα δεδομένων να μην είναι αντιπροσωπευτικά όλων των εφικτών αντικειμένων. Έτσι, δυσχεραίνεται η αξιόπιστη εκτίμηση των αποστάσεων, πυκνοτήτων και στατιστικών συσχετίσεων [61]. Για τον λόγο αυτό απαιτείται κατάλληλη προεπεξεργασία με στόχο την σταθερότητα των μεταγενέστερων μοντέλων.

Υπάρχουν δύο κύριες φάσεις προεπεξεργασίας που περιγράφονται σε αυτή την ενότητα. Το πρώτο σχετίζεται με την διόρθωση των συλλεγμένων δεδομένων ενός συνόλου δεδομένων, δηλαδή επρόκειτο για την εφαρμογή τεχνικών εξασφάλισης της καλής ποιότητας των δεδομένων. Το δεύτερο τμήμα, που έπεται του πρώτου, είναι η μηχανική γνωρισμάτων (feature engineering) που συγκροτείται για την τροποποίηση ήδη υπαρχόντων γνωρισμάτων αλλά και τη δημιουργία νέων, προκειμένου να βελτιωθεί η απόδοση ενός μοντέλου ML.

2.3.1 Καθαρισμός Δεδομένων

Στην πράξη τα δεδομένα ενός συνόλου δεδομένων, σπάνια συλλέγονται με στόχο την αναλυτική επεξεργασία ή την εκπαίδευση αλγορίθμων. Για το λόγο αυτό, και όχι μόνο, συχνά τα δεδομένα χρήζουν αντιμετώπισης θεμάτων ποιότητας. Έτσι απαιτείται ο εντοπισμός και η διόρθωση σφαλμάτων που εμπεριέχονται στα δεδομένα, για την ομαλή μετάβαση στους αλγορίθμους ML [22]. Αυτού του είδους η διαδικασία ονομάζεται καθαρισμός δεδομένων (data cleaning). Το σύνολο δεδομένων πρέπει να ελέγχεται για:

- Ελλιπείς τιμές
- Ακραίες τιμές
- Διπλότυπα αντικείμενα δεδομένων
- Ασυνέπειες τιμών δεδομένων
- Θόρυβος

Οι **ελλιπείς τιμές** αποτελούν ένα από τα συχνότερα προβλήματα ποιότητας δεδομένων και προκύπτουν όταν μία ή περισσότερες μεταβλητές δεν έχουν καταγραφεί για ορισμένες παρατηρήσεις. Οι αιτίες μπορεί να περιλαμβάνουν σφάλματα συλλογής, αποτυχία αισθητήρων ή παραλείψεις χρηστών. Τεχνικές αντιμετώπισης είναι η εύρεση και συμπλήρωση μέσης,ενδιάμεσης ή πιο συχνής τιμής ενός χαρακτηριστικού, σύμφωνα με τον τύπο του. Επιπλέον εάν στις περισσότερες τιμές μίας στήλης εμφανίζονται ελλιπείς τιμές τότε συχνά εφαρμόζεται εξάλειψη της στήλης. Αντίστοιχα εάν μία παρατήρηση περιλαμβάνει σημαντικό αριθμό ελλιπών τιμών, γεγονός που αποδυναμώνει την πληροφορία του στιγμιότυπου, τότε αφαιρείται ολόκληρη η παρατήρηση.

Οι **ακραίες τιμές** αναφέρονται σε παρατηρήσεις που αποκλίνουν σημαντικά από τη γενική κατανομή των δεδομένων. Μπορεί να προκύπτουν είτε λόγω σφαλμάτων καταγραφής (π.χ. λανθασμένες μονάδες μέτρησης) είτε να αντιπροσωπεύουν σπάνια αλλά έγκυρα φαινόμενα. Η διάκριση μεταξύ αυτών των δύο περιπτώσεων είναι κρίσιμη, καθώς η αδιάκριτη αφαίρεση ακραίως τιμών μπορεί να οδηγήσει σε απώλεια σημαντικής πληροφορίας.

Τα **διπλότυπα ή σχεδόν διπλότυπα** αντικείμενα δεδομένων εμφανίζονται όταν η ίδια οντότητα καταγράφεται περισσότερες από μία φορές στο σύνολο δεδομένων. Το φαινόμενο αυτό είναι συχνό σε συστήματα που συνδυάζουν δεδομένα από πολλαπλές πηγές ή δεν διαθέτουν μοναδικά αναγνωριστικά.

Οι **ασυνέπειες τιμών** προκύπτουν όταν τα δεδομένα παραβιάζουν λογικούς, σημασιολογικούς ή δομικούς περιορισμούς, συνήθως λόγω σφάλματος κατά την καταγραφή τους. Για παράδειγμα, το ύψος ενός ατόμου δεν μπορεί να λάβει αρνητική τιμή [22]. Ακόμη, εάν τα δεδομένα προέρχονται από διαφορετικές πηγές τότε ενδέχεται να υπάρχει διαφοροποίηση της μορφής αναπαράστασης μίας μεταβλητής. Παραδείγματος χάριν, τα δεδομένα ενός γνωρίσματος σχετικά με τον χρόνο, μπορεί να εμφανίζονται άλλες φορές σε μορφή λεπτών και άλλες σε μορφή δευτερολέπτων. Η ύπαρξη ασυνεπειών στα δεδομένα μπορεί να υπονομεύσει τη λογική συνοχή του συνόλου δεδομένων και να οδηγήσει σε εσφαλμένες συσχετίσεις.

Ο **θόρυβος** αναφέρεται σε τυχαίες διακυμάνσεις ή σφάλματα μέτρησης που αλλοιώνουν το πραγματικό σήμα των δεδομένων. Ιδιαίτερα συχνή περίπτωση εμφανίζεται σε δεδομένα αισθητήρων, λόγω ελαττωματικότητας ή περιβαλλοντικών συνθηκών, με αποτέλεσμα να επηρεάζει αρνητικά την ποιότητα των δεδομένων, καταγράφοντας μη ρεαλιστικές πληροφορίες.

Συνοψίζοντας, ο καθαρισμός δεδομένων συνιστά αναπόσπαστο στάδιο της προεπεξεργασίας, καθώς επηρεάζει άμεσα την εγκυρότητα, τη σταθερότητα και τη γενικευσιμότητα των μοντέλων ML. Η συστηματική αντιμετώπιση ελλιπών τιμών, ακραίων και διπλοτύπων παρατηρήσεων, ασυνεπειών και θορύβου, καθώς και άλλων ζητημάτων ποιότητας, δεν αποτελεί απλώς τεχνική διαδικασία, αλλά προϋποθέτει εννοιολογική κατανόηση των δεδομένων και του πεδίου εφαρμογής. Η παράλειψη ή η εσφαλμένη εφαρμογή σταδίων καθαρισμού δύναται να οδηγήσει σε μεροληπτικά ή μη αξιόπιστα αποτελέσματα. Κατά συνέπεια, η επένδυση χρόνου και προσοχής στον καθαρισμό των δεδομένων αποτελεί θεμελιώδη προϋπόθεση για την ομαλή μετάβαση στα επόμενα στάδια της προεπεξεργασίας και, τελικά, για την επιτυχή εφαρμογή αλγορίθμων ML.

2.3.2 Κωδικοποίηση Μεταβλητών

Ένα από τα συχνότερα ζητήματα στην προεπεξεργασία αφορά τη διαχείριση κατηγορικών μεταβλητών, που δεν μπορούν να χρησιμοποιηθούν απευθείας από τα περισσότερα μοντέλα ML, καθώς αυτά βασίζονται σε αριθμητικές πράξεις και μετρικές αποστάσεων [22]. Η κωδικοποίηση αποσκοπεί στη μετατροπή των κατηγορικών τιμών σε αριθμητικές αναπαραστάσεις, με τρόπο που να διατηρεί, το δυνατόν περισσότερο, τη σημασιολογική πληροφορία των δεδομένων.

Η καλύτερη τεχνική που εφαρμόζεται για την τροποποίηση των κατηγοριών των μεταβλητών σε αριθμητικές παραστάσεις, είναι αυτή που παράγει τα καλύτερα αποτελέσματα για τον εκάστοτε αλγόριθμο ML [22]. Μερικές από τις πιο σύνηθες προσεγγίσεις παρουσιάζονται παρακάτω.

- **Δυαδική κωδικοποίηση:** Μετατρέπει τις κατηγορικές τιμές μίας μεταβλητής σε αριθμητικές παραστάσεις μέσω δυαδικών ψηφίων (0,1). Συγκριτικά με άλλες μεθόδους, μειώνει τον αριθμό των παραγόμενων χαρακτηριστικών, όμως εισάγει μια πιο έμμεση αναπαράσταση, η οποία μπορεί να δυσχεράνει την ερμηνεία των αποτελεσμάτων και να αλλοιώσει τις αποστάσεις μεταξύ παρατηρήσεων. Επιχειρεί να αντιμετωπίσει το πρόβλημα της διαστασιμότητας [22].
- **Τακτική κωδικοποίηση:** Εφαρμόζεται όταν οι κατηγορίες διαθέτουν εγγενή διάταξη, δηλαδή ο τύπος της μεταβλητής είναι τακτικό αριθμητικό. Σε αυτή την περίπτωση, η αριθμητική αναπαράσταση μπορεί να αποτυπώσει τη σχετική σειρά των τιμών, απαιτεί όμως ιδιαίτερη προσοχή,

καθώς η αυθαίρετη ανάθεση αποστάσεων μεταξύ κατηγοριών μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα. Ένα παράδειγμα αποτελεί η ιεραρχική αναπαράσταση τιμών: ‘Κακό’, ‘Μέτριο’, ‘Καλό’ [36].

- **Κωδικοποίηση μεταβλητής-στόχου:** Βασίζεται στη χρήση στατιστικών πληροφοριών της μεταβλητής-στόχου για την κωδικοποίηση των κατηγοριών. Χρησιμοποιείται τόσο για δυαδικές όσο και για πολλαπλές κατηγορίες μεταβλητής στόχου [62]. Η μέθοδος αυτή μπορεί να προσφέρει ισχυρή προβλεπτική πληροφορία, ιδιαίτερα σε προβλήματα SL, ενέχει όμως αυξημένο κίνδυνο υπερπροσαρμογής, εάν δεν εφαρμοστεί με κατάλληλες τεχνικές κανονικοποίησης ή διασταυρούμενης επικύρωσης.
- **One-Hot Κωδικοποίηση (OHE):** Αποτελεί μία από τις πιο διαδεδομένες τεχνικές, κατά την οποία κάθε κατηγορία αναπαρίσταται ως ξεχωριστή δυαδική μεταβλητή [36]. Το κύριο πλεονέκτημά της είναι ότι αποφεύγει την εισαγωγή τεχνητής ιεραρχίας μεταξύ κατηγοριών. Ωστόσο, σε περιπτώσεις μεταβλητών με πολλές κατηγορίες, οδηγεί σε σημαντική αύξηση της διαστασιμότητας, γεγονός που μπορεί να επιβαρύνει τόσο την υπολογιστική πολυπλοκότητα όσο και την απόδοση των μοντέλων.

Εκτός από την τροποποίηση της αναπαράστασης των κατηγορικών τιμών, συχνά, ανάλογα με το πεδίο του προβλήματος, απαιτείται η μετατροπή μιας αριθμητικής μεταβλητής σε κατηγορική, με μεθόδους διακριτοποίησης με επίβλεψη ή χωρίς. Ωστόσο δεν καλύπτεται αναλυτικά στη παρούσα εργασία.

2.3.3 Συνάθροιση Δεδομένων

Η συνάθροιση δεδομένων αποτελεί τεχνική κατά την οποία πολλαπλές παρατηρήσεις ή εγγραφές συνδυάζονται σε μία ενιαία αναπαράσταση, συνήθως με βάση κάποια λογική οντότητα ή χρονική περίοδο [22]. Μέσω στατιστικών συναρτήσεων (π.χ. μέσος όρος, διακύμανση, συχνότητα), οι παρατηρήσεις συμπυκνώνονται σε αναπαραστάσεις υψηλότερου επιπέδου. Η πρακτική αυτή είναι ιδιαίτερα χρήσιμη όταν οι αρχικές παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους, Το πλεονέκτημα της συνάθροισης είναι η μείωση του θορύβου και η ενίσχυση της σταθερότητας, καθώς και η αποφυγή της ιδιότητας εξάρτησης μεταξύ των παρατηρήσεων, που αποτελεί κριτικό παράγοντα για τη μεροληψία των μοντέλων.

Παρά τα πλεονεκτήματά της, η συνάθροιση ενδέχεται να οδηγήσει σε απώλεια πληροφορίας, δυναμικών μη γραμμικών μοτίβων. Ως εκ τούτου, η επιλογή της κατάλληλης στρατηγικής συνάθροισης απαιτεί προσεκτική ανάλυση του προβλήματος και των στόχων της μελέτης.

2.3.4 Δημιουργία Γνωρισμάτων

Η δημιουργία γνωρισμάτων περιλαμβάνει δύο είδη, τόσο την εξαγωγή γνωρισμάτων (feature extraction) από ακατέργαστα δεδομένα, όσο και την κατασκευή νέων γνωρισμάτων (feature creation) μέσω συνδυασμών ή μετασχηματισμών υπαρχόντων μεταβλητών.

Η εξαγωγή γνωρισμάτων επιδιώκει τη συμπύκνωση της πληροφορίας σε πιο περιεκτικές μορφές, για το λόγο αυτό είναι γνωστή και ως τεχνική μείωσης της διαστασιμότητας. Στην πράξη, δεδομένα υψηλής διαστασιμότητας (δηλαδή σύνολο δεδομένων με πολλά χαρακτηριστικά) μετατρέπονται σε τέτοια μορφή ώστε τόσο η πληροφορία των αρχικών δεδομένων να διατηρείται, όσο να μειώνεται το πλήθος των χαρακτηριστικών [63]. Έχουν αναπτυχθεί ποικίλες μέθοδοι, μία εκ των οποίων είναι και το PCA, που έχει αναφερθεί στη αρχή του κεφαλαίου.

Η κατασκευή γνωρισμάτων συνδυάζει ή μετασχηματίζει ήδη υπάρχοντες μεταβλητές ενός συνόλου δεδομένων, αναγνωρίζοντας πιο εκφραστικά γνωρίσματα και καλύπτοντας πληροφορία σχετική με το πεδίο εφαρμογής και τη μορφή που επιδιώκεται. Ένα απλό παράδειγμα κατασκευής γνωρισμάτων αποτελεί η συλλογή δεδομένων ενός ατόμου σχετικά με το ύψος, το βάρος και την ηλικία του. Μέσω αυτών των χαρακτηριστικών, μπορεί να εξαχθεί ο δείκτης μάζας σώματος του ατόμου. Παρ' όλα αυτά, στις περισσότερες περιπτώσεις τα προβλήματα είναι αρκετά πιο πολύπλοκα.

2.3.5 Επιλογή Γνωρισμάτων

Ένας άλλος τρόπος μείωσης των πολλών διαστάσεων των δεδομένων είναι η επιλογή χαρακτηριστικών ή επιλογή υποσυνόλου χαρακτηριστικών (feature subset selection) στοχεύοντας στη διατήρηση μόνο των πλέον χρήσιμων μεταβλητών, μειώνοντας τον θόρυβο και τη διαστασιμότητα του προβλήματος. Όπως και στην εξαγωγή χαρακτηριστικών, έτσι και εδώ, τεχνικές μείωσης διαστασιμότητας, όπως η ανάλυση κύριων συνιστωσών, επιτρέπουν τη μετατροπή των δεδομένων σε χαμηλότερης διάστασης χώρους, με δυνατότητα μίας έμμεσης επιλογής χαρακτηριστικών. Πιθανό κόστος βέβαια, καθίσταται η μειωμένη ερμηνευσιμότητα.

Επιπλέον, συχνά στα δεδομένα εμφανίζονται άσχετα γνωρίσματα τα οποία δεν συνεισφέρουν στην επίλυση του προβλήματος και συνεπώς στην καλύτερη απόδοση των αλγορίθμων, παρά μόνο επηρεάζουν αρνητικά προσφέροντας σύγχυση στα τελικά αποτελέσματα. Μία τέτοια περίπτωση είναι η μεταβλητή αναγνωριστικού παρατήρησης, η οποία δεν έχει καμία σχέση με την εξόρυξη των δεδομένων. Παράλληλα, συχνό φαινόμενο αποτελούν και τα περιττά γνωρίσματα, τα οποία διπλασιάζουν την πληροφορία που περιέχει/ουν μία ή περισσότερες μεταβλητές, προκαλώντας επικράτηση έναντι άλλων και μεροληψία στα μοντέλα ML [22]. Υπάρχουν τρεις βασικές προσεγγίσεις [22], [64] για την επιλογή γνωρισμάτων:

Μέθοδοι φίλτρου (filter methods): Βασίζονται στη χρήση στατιστικών κριτηρίων για την αξιολόγηση της σχέσης ζευγαριών χαρακτηριστικών, πριν το στάδιο χρήσης μοντέλου ML. Η επιλογή πραγματοποιείται μέσω κατάταξης των χαρακτηριστικών βάσει μέτρων όπως συσχέτιση Pearson και Spearman, στατιστικοί έλεγχοι ή πληροφοριακά κριτήρια σε σχέση με το πεδίο ενδιαφέροντος, γεγονός που τις καθιστά υπολογιστικά αποδοτικές και ανθεκτικές στην υπερπροσαρμογή.

Ενσωματωμένες προσεγγίσεις (embedded methods): Η επιλογή χαρακτηριστικών εδώ, πραγματοποιείται ως τμήμα της φάσης εκπαίδευσης του μοντέλου ML. Ο αλγόριθμος είναι αυτός που αποφασίζει, μέσω κριτηρίων, ποια γνωρίσματα θα κρατήσει και ποια θα αγνοήσει. Τέτοιοι είναι οι αλγόριθμοι που βασίζονται σε δέντρα αποφάσεων, και αλγόριθμοι που περιλαμβάνουν την παράμετρο κανονικοποίησης (regularization) [65]. Η κανονικοποίηση, επιβάλει ποινή στα πιο πολύπλοκα μοντέλα, ενθαρρύνοντας τα πιο απλά, προκειμένου να βελτιωθεί η γενίκευση του μοντέλου σε νέα δεδομένα. Μειονέκτημα αποτελεί η άμεση εξάρτηση της προσέγγισης με το επιλεγμένο μοντέλο ML.

Μέθοδοι περιτυλίγματος (wrapper methods): Αξιολογούν διαφορετικά υποσύνολα χαρακτηριστικών εκπαιδεύοντας επαναληπτικά το μοντέλο και επιλέγοντας εκείνο που μεγιστοποιεί την απόδοση. Όμως, λειτουργούν αυθαίρετα, και βασίζονται αποκλειστικά στη μεταβλητή-στόχο. Για αυτό, απαιτείται ιδιαίτερη προσοχή κατά την χρήση τους [22].

Η διαδικασία επιλογής γνωρισμάτων είναι σίγουρα μία από τις πιο σημαντικές φάσεις της προεπεξεργασίας δεδομένων. Με τη σειρά της, συμβάλει στην αποφυγή της κατάρας διαστασιμότητας, στη βελτίωση της υπολογιστικής αποδοτικότητας, στην απλούστευση των μοντέλων, στην μείωση της μεροληψίας και υπερπροσαρμογής, και φυσικά στην αύξηση της ικανότητας γενίκευσης του τελικού μοντέλου.

2.3.6 Μετασχηματισμός Γνωρισμάτων

Ο μετασχηματισμός μεταβλητών αποτελεί ακόμη μία βασική πρακτική στην προεπεξεργασία δεδομένων και αποσκοπεί στη βελτίωση της στατιστικής και υπολογιστικής συμπεριφοράς των χαρακτηριστικών πριν αυτά χρησιμοποιηθούν ως είσοδοι σε μοντέλα ML. Στην ουσία είναι ένα από τα τελευταία βήματα της προεπεξεργασίας δεδομένων, αν όχι το τελευταίο. Πολλά πραγματικά σύνολα δεδομένων εμφανίζουν ασυμμετρικές κατανομές, ακραίες τιμές ή μη γραμμικές σχέσεις, οι οποίες δυσχεραίνουν την εκπαίδευση και την ερμηνεία των μοντέλων. Μέσω κατάλληλων μετασχηματισμών, οι μεταβλητές μπορούν να αποκτήσουν πιο ευνοϊκές ιδιότητες, όπως μεγαλύτερη κανονικότητα ή σταθερότερη διακύμανση, διευκολύνοντας τόσο την εκτίμηση παραμέτρων όσο και τη γενίκευση.

Συχνά, λογαριθμικοί, τετραγωνικής ρίζας, απόλυτης τιμής ή γενικότεροι μετασχηματισμοί χρησιμοποιούνται, με στόχο την αντιμετώπιση της ασυμμετρίας και της ετερογένειας.

Ωστόσο, ένας από τους πιο σημαντικούς μετασχηματισμούς που είναι απαραίτητοι να εφαρμόζονται στα περισσότερα είδη αλγορίθμων ML, τα οποία δεν αποδίδουν καλά σε χαρακτηριστικά με πολύ διαφορετικές κλίμακες, είναι η κλιμάκωση γνωρισμάτων. Υπάρχουν δύο κύρια είδη κλιμάκωσης, η τυποποίηση (standardization) και η κανονικοποίηση (normalization ή Min-Max) [36]. Και οι δύο χρησιμοποιούνται για την ευθυγράμμιση των κλιμάκων μεταξύ των τιμών των χαρακτηριστικών, ιδιαίτερα σε αλγορίθμους που βασίζονται σε αποστάσεις και αυτούς που περιέχουν αριθμητικές τιμές με μεγάλες αποκλίσεις μεταξύ των χαρακτηριστικών των δεδομένων.

Πιο συγκεκριμένα, η τυποποίηση αναφέρεται στον μετασχηματισμό των δεδομένων, έτσι ώστε κάθε χαρακτηριστικό να έχει μηδενικό μέσο όρο και μοναδιαία τιμή τυπικής απόκλισης. Η διαδικασία αυτή επιτυγχάνεται αφαιρώντας από κάθε τιμή τη μέση τιμή της αντίστοιχης μεταβλητής και διαιρώντας το αποτέλεσμα ως προς την τυπική απόκλιση. Με αυτό τον τρόπο, τα χαρακτηριστικά περιορίζονται σε συγκεκριμένο διάστημα τιμών, ανάλογα με τις κατηγορίες τιμών της μεταβλητής. Η τεχνική θεωρείται λιγότερο ευαίσθητη σε ακραίες τιμές, σε σχέση με την Min-Max και βοηθάει στη διαδικασία κανονικοποίησης (regularization) αποτρέποντας την υπερπροσαρμογή του μοντέλου [30]. Γενικώς, η τυποποίηση θεωρείται ιδιαίτερα κατάλληλη σε περιπτώσεις όπου τα δεδομένα ακολουθούν ή προσεγγίζουν κανονική κατανομή και χρησιμοποιείται ευρέως σε αλγορίθμους που βασίζονται σε αποστάσεις.

Ο μαθητικός τύπος της τυποποίησης είναι:

$$z = \frac{x - \mu}{\sigma} \tag{2.2}$$

όπου, μ είναι η μέση τιμή και σ η τυπική απόκλιση της μεταβλητής προτού μετασχηματιστεί. Το x είναι η τιμή της παρατηρούμενης εγγραφής, ενώ το z είναι η τυποποιημένη πλέον τιμή.

Απεναντίας, η τεχνική της κανονικοποίησης αποτελεί μία απλούστερη τεχνική κλιμάκωσης, καθώς λαμβάνει τις αρχικές τιμές και τις επαναδατιατυπώνει σε μία προκαθορισμένη κλίμακα, συνήθως εντός των τιμών 0 και 1. Η μέθοδος αυτή διατηρεί την σχετική διάταξη των τιμών. Επιπροσθέτως, η κανονικοποίηση χρησιμοποιείται συχνά σε εφαρμογές όπου απαιτείται σαφής οριοθέτηση των τιμών εισόδου, όπως σε νευρωνικά δίκτυα. Πέρα από τα πλεονέκτηματά της, παρουσιάζει αυξημένη ευαισθησία σε ακραίες τιμές, καθώς αυτές καθορίζουν τα όρια της κλίμακας, σε αντίθεση με την περίπτωση της τυποποίησης.

Ο μαθηματικός τύπος της κανονικοποίησης δίνεται ως εξής:

$$x\,scaled = \frac{x - min(X)}{max(X) - min(X)} \tag{2.3}$$

όπου X είναι η μεταβλητή-χαρακτηριστικό, ενώ min και max είναι η ελάχιστη και η μέγιστη τιμή της μεταβλητής στο σύνολο δεδομένων. Το x, και εδώ, αποτελεί την τιμή της παρατηρούμενης εγγραφής.

Συνολικά, ο μετασχηματισμός μεταβλητών συνιστά ισχυρό εργαλείο για τη βελτίωση τόσο της ποιότητας των δεδομένων όσο και της απόδοσης των μοντέλων, εφόσον εφαρμόζεται ορθά πάνω στο αντικείμενο ενδιαφέροντος. Έτσι καταφέρνει να ενισχύει τη σταθερότητα, τη γενίκευση και την αξιοπιστία των προκυπτόντων μοντέλων.

2.4 Διαδικασία Ανάπτυξης και Αξιολόγησης Μοντέλων

2.4.1 Μετρικές

Ένα μοντέλο ML, αφού εκπαιδευτεί στη συνέχεια πρέπει να αξιολογηθεί η απόδοσή του σε δεδομένα τα οποία του είναι άγνωστα. Η αξιολόγηση της απόδοσης ενός μοντέλου ML, επιτρέπει την αντικειμενική αποτίμηση της ποιότητας των αποτελεσμάτων και τη σύγκριση εναλλακτικών προσεγγίσεων. Τη διαδικασία αξιολόγησης εκπληρώνουν οι μετρικές των αλγορίθμων ML. Οι μετρικές αυτές λειτουργούν ως ποσοτικοί δείκτες που συνδέουν τη θεωρητική λειτουργία των αλγορίθμων με την πρακτική τους αποτελεσματικότητα, παρέχοντας παρακάτω κριτήρια για επιλογή μοντέλου, ρύθμιση παραμέτρων και ερμηνεία των αποτελεσμάτων [30].

Στο πλαίσιο της Μηχανικής Μάθησης που ενδιαφέρουν την παρούσα διατριβή, οι μετρικές διακρίνονται σε δύο βασικές κατηγορίες: εκείνες που χρησιμοποιούνται σε αλγορίθμους συσταδοποίησης και εκείνες που εφαρμόζονται σε μοντέλα ταξινόμησης. Η διάκριση αυτή δεν είναι τυπική, καθώς οι δύο κατηγορίες βασίζονται σε διαφορετικές πηγές πληροφορίας και εξυπηρετούν διαφορετικούς στόχους αξιολόγησης.

Πριν από την ανάλυση των επιμέρους μετρικών, είναι κρίσιμο να διατυπωθεί ότι στις μη επιβλεπόμενες μεθόδους δεν υπάρχουν προκαθορισμένες ετικέτες (κλάσεων), άρα είναι αδύνατη η χρήση μετρικών που συγκρίνουν προβλέψεις με “πραγματικές” τιμές (ground-truth) [66]. Βέβαια, θα πρέπει να διασαφηνιστεί ότι υπάρχουν περιπτώσεις συσταδοποίησης, όπου περιέχονται ετικέτες στα δεδομένα, όμως αυτές εξυπηρετούν διαφορετικό σκοπό απ’ αυτό της παρούσας εργασίας. Αντίθετα, η αξιολόγηση βασίζεται σε εσωτερικά κριτήρια συνοχής, διαχωρισιμότητας και δομής των δεδομένων, συχνά εξαρτώμενα από τα ίδια τα μοντέλα ML, κριτήρια που αποτυπώνουν το πόσο καλά οι συστάδες που προκύπτουν αντανακλούν τα υποκείμενα μοτίβα των δεδομένων [66]. Στα μοντέλα ταξινόμησης, οι μετρικές εστιάζουν κυρίως στη συμφωνία μεταξύ προβλεπόμενων και πραγματικών ετικετών, επιτρέποντας άμεση ποσοτικοποίηση της γενίκευσης του μοντέλου [67].

Παρακάτω, αναλύονται οι μετρικές που χρησιμοποιούνται για την αξιολόγηση αλγορίθμων ομαδοποίησης, και στη συνέχεια οι μετρικές για τα επιβλεπόμενα μοντέλα ταξινόμησης, οι οποίες εφαρμόστηκαν στο πλαίσιο της παρούσας εργασίας.

<u>Μετρικές συσταδοποίησης</u>

Elbow Method

Η μέθοδος ‘Elbow’ αποτελεί μια ευρετική τεχνική που χρησιμοποιείται κυρίως για την επιλογή του βέλτιστου αριθμού συστάδων σε αλγορίθμους όπως ο K-Means, στους οποίους πρέπει ο χρήστης να επιλέξει εκ των προτέρων το πλήθος των συστάδων ομαδοποίησης. Βασίζεται στη συμπεριφορά της ενδο-συσταδικής (intra-cluster) διασποράς (άθροισμα τετραγωνικού σφάλματος) ως συνάρτηση του πλήθους των συστάδων [27], [30]. Η βασική ιδέα είναι ότι η μείωση του σφάλματος επιβραδύνεται μετά από ένα συγκεκριμένο σημείο (η γνωστή αναπαράσταση “αγκώνα”), το οποίο θεωρείται ενδεικτικό του βέλτιστου αριθμού συστάδων [27].

Πλεονέκτημα της μεθόδου είναι η απλότητα και η διαισθητική-οπτική ερμηνεία, ωστόσο μειονέκτημα αποτελεί η υποκειμενικότητα στον εντοπισμό του “αγκώνα”, ιδιαίτερα σε δεδομένα χωρίς σαφή δομή, όπου εν τέλει δεν είναι εύκολα ορατό το σημείο αναπαράστασης “αγκώνα” [30].
Η μέθοδος “Elbow” είναι κατάλληλη κυρίως για αλγορίθμους που βασίζονται στη βελτιστοποίηση απόστασης, όπως ο K-Means, και όχι για προσεγγίσεις βασισμένες σε πυκνότητα ή άλλες [66].

Silhouette Score

Ο δείκτης αξιολόγησης Silhouette score αποτυπώνει την ποιότητα της ομαδοποίησης συγκρίνοντας τη συνοχή των σημείων εντός της ίδιας συστάδας με τη διαχωρισιμότητά τους από άλλες συστάδες [68]. Για κάθε παρατήρηση υπολογίζεται μια τιμή μέσα στο διάστημα [−1,1]. Οι υψηλότερες τιμές, κοντά στην τιμή 1, υποδηλώνουν καλή αντιστοίχιση στην συστάδα και σαφέστερο διαχωρισμό, οι τιμές κοντά στο 0 παρέχουν την πληροφορία ότι η παρατήρηση είναι κοντά στο σύνορο απόφασης μεταξύ δύο γειτονικών συστάδων, ενώ οι αρνητικές τιμές υποδεικνύουν ότι μάλλον η ανάθεση της παρατηρούμενης εγγραφής στη συστάδα είναι εσφαλμένη [69]. Επομένως στην περίπτωση του δείκτη αξιολόγησης Silhouette, η μεγαλύτερη τιμή αποσκοπεί στον καλύτερο διαχωρισμό των συστάδων.

Πλεονέκτημα του Silhouette score είναι ότι παρέχει τόσο συνολική όσο και τοπική πληροφόρηση για την ποιότητα των συστάδων, ενώ μειονέκτημά του αποτελεί η αυξημένη υπολογιστική απαίτηση σε μεγάλα σύνολα δεδομένων [66]. Η μετρική εφαρμόζεται ευρέως σε συστάδες βασισμένες στο κέντρο και ιεραρχική συσταδοποίηση, υπό την προϋπόθεση ορισμού κατάλληλης μετρικής απόστασης [68].

Davies-Bouldin Index

Ο δείκτης Davies–Bouldin χρησιμοποιείται για την αξιολόγηση της ποιότητας της ομαδοποίησης ως λόγο της ενδο-συσταδικής διασποράς προς την μεταξύ-συστάδων (inter-cluster) απόσταση [70]. Με πιο απλά λόγια, εξετάζεται το πόσο συγκεντρωμένα είναι τα σημεία εντός κάθε συστάδας με το πόσο κοντά βρίσκεται η πιο παρόμοια γειτονική της συστάδα. Μικρότερες τιμές του δείκτη υποδηλώνουν καλύτερη δομή συστάδων, καθώς σημαίνουν ότι τα δεδομένα μέσα σε κάθε συστάδα είναι ομοιογενή και ταυτόχρονα δείχνουν ότι είναι καλύτερα διαχωρισμένα από τις υπόλοιπες συστάδες [71]. Η μετρική αυτή είναι πλήρως εσωτερική και δεν απαιτεί εξωτερική πληροφορία, γεγονός που την καθιστά κατάλληλη για μη επιβλεπόμενα σενάρια [66]. Πλεονέκτημά της αποτελεί η υπολογιστική αποδοτικότητα, ενώ βασικός περιορισμός της είναι η ευαισθησία της σε συστάδες μη σφαιρικού σχήματος ή διαφορετικής πυκνότητας. Για το λόγο αυτό εφαρμόζεται κυρίως σε βασισμένη στο κέντρο συσταδοποίηση, όπου τα σχήματα των συστάδων είναι σφαιρικά. Επιπλέον, χρησιμοποιείται για τον καθορισμό του βέλτιστου αριθμού συστάδων σε αλγορίθμους όπως ο K-Means και για συγκριτική αξιολόγηση διαφορετικών αλγορίθμων συσταδοποίησης.

DBCV

Ο δείκτης αξιολόγησης DBCV (Density-Based Clustering Validation) αποτελεί μία μετρική που είναι σχεδιασμένη για αλγορίθμους συσταδοποίησης βασισμένοι στην πυκνότητα, όπως ο DBSCAN [72]. Σε αντίθεση με κλασικούς δείκτες που στηρίζονται αποκλειστικά σε αποστάσεις μεταξύ κέντρων συστάδων ή σημείων, ο δείκτης DBCV αξιολογεί την ποιότητα των συστάδων λαμβάνοντας υπόψη τη δομή της τοπικής πυκνότητας των δεδομένων. Έτσι, αποτυπώνει τόσο τη συνοχή εντός κάθε συστάδας όσο και τον βαθμό διαχωρισμού μεταξύ διαφορετικών συστάδων. Σε κάθε παρατήρηση, αποδίδει μία τιμή στο διάστημα [-1, 1], όπου υψηλότερες θετικές τιμές υποδηλώνουν καλά διαχωρισμένες και συνεκτικές συστάδες, ενώ αρνητικές τιμές υποδηλώνουν κακή ομαδοποίηση ή ακατάλληλη παραμετροποίηση του αλγορίθμου [72]. Ένα από τα βασικά πλεονεκτήματα της μετρικής είναι ότι μπορεί να χειριστεί καλά συστάδες αυθαίρετου σχήματος και διαφορετικής πυκνότητας. Όμως, παρουσιάζει υψηλή υπολογιστική πολυπλοκότητα, γεγονός που μπορεί να αποτελέσει περιορισμό χρήσης της μετρικής σε μεγάλα σύνολα δεδομένων, ενώ ακόμη επηρεάζεται από την αραιότητα των δεδομένων σε υψηλές διαστάσεις, καθώς η εκτίμηση της πυκνότητας γίνεται λιγότερο διακριτική. Παρ' όλα αυτά, το DBCV θεωρείται κατάλληλη μετρική αξιολόγησης αλγορίθμων που βασίζονται στην πυκνότητα, αφού εφαρμόζεται καλύτερα στους στόχους των αλγορίθμων αυτών.

Μετρικές ταξινόμησης

Accuracy

Η μετρική accuracy (ή ακρίβεια) εκφράζει το ποσοστό των συνολικών σωστών προβλέψεων επί του συνόλου των δειγμάτων παρέχοντας μια γενική εικόνα της απόδοσης ενός μοντέλου. Πρόκειται για μία θεμελιώδη μετρική ταξινόμησης η οποία είναι ιδιαίτερα ερμηνεύσιμη. Ωστόσο, ανάλογα με το πρόβλημα ταξινόμησης, η μετρική μπορεί να αποδειχθεί παραπλανητική ως προς τη γενίκευση του μοντέλου. Αυτό συμβαίνει σε περιπτώσεις συνόλων δεδομένων με ανισόρροπες κλάσεις, στις οποίες τα δείγματα από τις μειοψηφικές κλάσεις είναι πολύ λίγα, με αποτέλεσμα η μετρική να εμφανίζει υψηλή ακρίβεια πρόβλεψης της πλειοψηφικής κλάσης, χωρίς όμως να αποδίδει το ίδιο καλά με τις μειοψηφικές [22], [73].

Precision

Η μετρική precision (ή ακρίβεια) αποτυπώνει το ποσοστό των σωστών θετικών προβλέψεων επί του συνόλου των προβλέψεων που χαρακτηρίστηκαν ως θετικές. Η συγκεκριμένη μετρική είναι χρήσιμη σε περιπτώσεις που το κόστος ψευδώς θετικών αποτελεσμάτων είναι υψηλό, όπως συμβαίνει σε περιπτώσεις προβλημάτων υγείας. Επιπλέον, δεν παρέχει πληροφορία σχετικά με το πραγματικό ποσοστό θετικών περιπτώσεων, παρά μόνο προβλέψεων [74].

Recall

Η μετρική recall (ή ευαισθησία) αποδίδει το ποσοστό των πραγματικών θετικών περιπτώσεων που αναγνωρίστηκαν σωστά από ένα μοντέλο ML. Θεωρείται κρίσιμη τεχνική σε εφαρμογές όπου η απουσία ανίχνευσης θετικών περιστατικών (δηλαδή της θετικής κλάσης) έχει σοβαρές συνέπειες, όπως σε προβλήματα υγείας ή ασφάλειας [73], [74].

Πίνακας Σύγχυσης

Ο πίνακας σύγχυσης αποτελεί βασικό εργαλείο ανάλυσης της απόδοσης ενός μοντέλου ταξινόμησης, καθώς παρέχει αναλυτικά τη σχέση μεταξύ των πραγματικών κλάσεων και των προβλεπόμενων από το μοντέλο [74]. Μέσω του πίνακα κατανοείται καλύτερα τα σφάλματα του μοντέλου. Συνδυάζει μία ποικιλία από μετρικές που μπορούν να εξαχθούν από αυτόν, μερικές από τις οποίες είναι οι accuracy, precision, recall που

αναλύθηκαν παραπάνω. Με τον τρόπο αυτό, προσφέρει λεπτομερείς πληροφορίες τόσο για τα θετικά όσο για τα αρνητικά αποτελέσματα του μοντέλου. Παράλληλα, μπορεί να παρέχει την εικόνα της συμπεριφοράς του κάθε μοντέλου ανά κλάση, εάν το πρόβλημα ταξινόμησης είναι πολυταξικό.

F1-score

Η μετρική F1-score ή F-measure αποτελεί έναν συνδυαστικό δείκτη αξιολόγησης, ενσωματώνοντας τις μετρικές precision και recall, μέσω του αρμονικού μέσου τους [73], [75]. Χρησιμοποιείται ιδιαίτερα σε περιπτώσεις ύπαρξης πολλών κλάσεων ή όταν απαιτείται ισορροπημένη αξιολόγηση του μοντέλου μεταξύ προβλέψεων και πραγματικών αποτελεσμάτων ενός μοντέλου.

$$F1\,score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.4}$$

Σε προβλήματα ταξινόμησης πολλών τάξεων, ο δείκτης χρησιμοποιεί διάφορους τύπους υπολογισμού του αποτελέσματός του, εξαρτώμενος από την περίσταση του προβλήματος. Δύο συνηθισμένες τεχνικές είναι οι “macro” και η ”weight” [75]. Η πρώτη, υπολογίζει τη μετρική F1-score για κάθε κλάση ανεξάρτητα, και στο τέλος λαμβάνει τον μέσο όρο τους, αποδίδοντας ίσο βάρος σε όλες. Βρίσκει χρήση όταν όλες οι κλάσεις θεωρούνται εξίσου σημαντικές. Αντίθετα, η δεύτερη τεχνική λαμβάνει υπόψη το μέγεθος της κάθε κλάσης, υπολογίζοντας έναν σταθμισμένο μέσο όρο της αξιολόγησης, όπου το βάρος κάθε κλάσης αντιστοιχεί στο ποσοστό των δειγμάτων της. Αυτή εκδοχή της F1-score αποδίδει καλύτερη συνολική απόδοση σε ανισόρροπα σύνολα δεδομένων, όμως εμπεριέχει τον κίνδυνου να ευνοεί τις κλάσεις με μεγαλύτερο ποσοστό δειγμάτων. Συνολικά, η επιλογή της καλύτερης εκδοχής εξαρτάται αποκλειστικά από το πρόβλημα εφαρμογής. Κατά συνέπεια, η μετρική αυτή χρησιμοποιείται για δυαδικές αλλά και πολυταξικές περιπτώσεις ταξινόμησης, λαμβάνοντας υπόψη τόσο την ακρίβεια όσο και την πληρότητα των προβλέψεων.

2.4.2 Εκτίμηση Μοντέλων

Γενικά το μέγεθος του συνόλου δεδομένων διαδραματίζει καθοριστικό ρόλο στην ικανότητα γενίκευσης ενός μοντέλου ML σε στιγμιότυπα χωρίς ετικέτα, καθώς επηρεάζει άμεσα το πόσο αντιπροσωπευτικά είναι τα δεδομένα εκπαίδευσης σε σχέση με τον πραγματικό πληθυσμό. Αυτό συμβαίνει διότι το μοντέλο δεν είναι ικανό αναπαραστήσει πλήρως τα πραγματικά υποδείγματα του συνόλου δεδομένων λόγω του περιορισμένου πλήθος στιγμιοτύπων του. Μεγαλύτερα σύνολα δεδομένων αυξάνουν την πιθανότητα το μοντέλο να εκτεθεί σε ποικιλία προτύπων και περιπτώσεων, μειώνοντας τον κίνδυνο υπερπροσαρμογής και επιτρέποντας την εκμάθηση πιο σταθερών και γενικεύσιμων σχέσεων. Αυτό όμως δεν αποκλείει το γεγονός και πάλι να μην υπάρχει καλή γενίκευση. Ιδιαίτερα σε εφαρμογές υγείας, η γενίκευση μπορεί να είναι δυσκολότερη καθώς μπορεί το πρόβλημα να χαρακτηρίζεται από μεγάλη ανισορροπία μεταξύ των στιγμιοτύπων των κλάσεων. Για παράδειγμα, σε ένα δυαδικό πρόβλημα ανίχνευσης μίας σπάνιας ασθένειας, τα δεδομένα της κλάσης που την ανιχνεύει είναι πιθανό να είναι αρκετά περιορισμένα. Για τον σκοπό αυτό, έχουν αναπτυχθεί πολλές προσεγγίσεις για την αντιμετώπιση των παραπάνω φαινομένων, ώστε να περιοριστεί η ευαισθησία ενός μοντέλου.

Διάφορες τεχνικές προεπεξεργασίας δεδομένων του συνόλου δεδομένων είναι απαραίτητο πολλές φορές να εφαρμοστούν ώστε τα δεδομένα να βρεθούν σε ώριμη μορφή, για να αξιοποιηθούν κατάλληλα από τις τεχνικές ML. Η απόδοση των μοντέλων επηρεάζεται άμεσα από την ποιότητα των τιμών εισόδου που τροφοδοτούνται. Πολλά μοντέλα απαιτούν την χρήση μίας τεχνικής κανονικοποίησης ή προτυποποίησης των χαρακτηριστικών τους, δεδομένου ότι είναι ευαίσθητα στη διαφορετική κλίμακα των μεταβλητών. Η απουσία κανονικοποίησης μπορεί να οδηγήσει σε υπέρμετρη επίδραση χαρακτηριστικών με μεγάλες

αριθμητικές τιμές, αλλοιώνοντας τη διαδικασία εκπαίδευσης και τις αποστάσεις στο χώρο των χαρακτηριστικών [36]. Αντιθέτως, η κατάλληλη κανονικοποίηση εξασφαλίζει συγκρίσιμες αναπαραστάσεις των δεδομένων, βελτιώνει τη σύγκλιση των αλγορίθμων και συμβάλλει στην ενίσχυση της γενίκευσης του μοντέλου σε νέα δεδομένα.

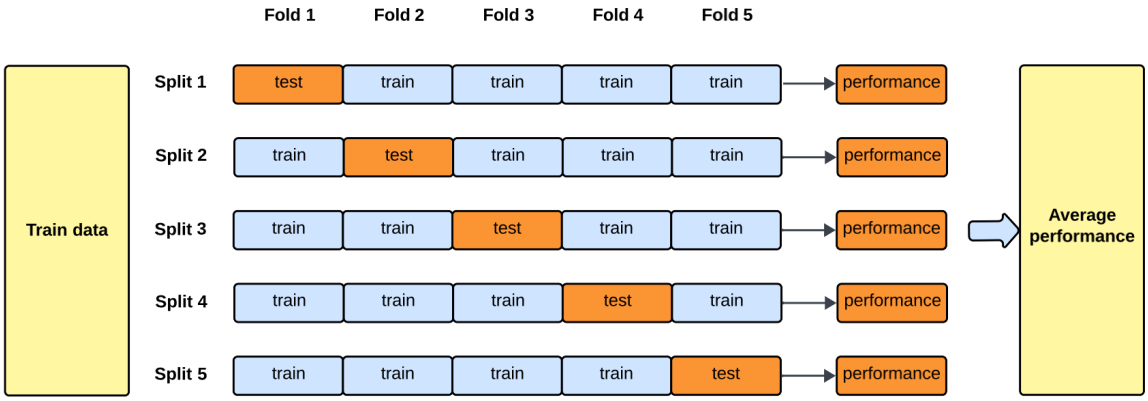
Επιπλέον τεχνικές προεπεξεργασίας όπως η επαύξηση (data augmentation / over-sampling) ή η μείωση (under-sampling) δεδομένων μπορούν να συμβάλλουν σημαντικά στη βελτίωση της γενίκευσης των μοντέλων ML, ιδίως σε περιπτώσεις ανισοκατανομής των κλάσεων. Οι τεχνικές αυτές στοχεύουν στην εξισορρόπηση του πλήθους των δειγμάτων μεταξύ των κλάσεων ενός συνόλου δεδομένων, μειώνοντας την πιθανότητα μεροληψίας του μοντέλου υπέρ της επικρατούσας κλάσης. Η επαύξηση δεδομένων αναφέρεται σε μεθόδους που αυξάνουν τεχνητά το πλήθος των δειγμάτων των υποεκπροσωπούμενων κλάσεων [76]. Μία προσέγγιση βασίζεται στη δημιουργία συνθετικών δεδομένων, τα οποία δεν προέρχονται από νέες πραγματικές παρατηρήσεις αλλά παράγονται μέσω συνδυασμών ή τροποποιήσεων μεταξύ υπαρχόντων δειγμάτων. Χαρακτηριστικό παράδειγμα αποτελεί η τεχνική SMOTE η οποία δημιουργεί νέα δείγματα στο χώρο των χαρακτηριστικών, αξιοποιώντας πληροφορία από γειτονικές παρατηρήσεις της μειονοτικής κλάσης [77]. Μία εναλλακτική προσέγγιση επαύξησης δεδομένων αφορά τον συνδυασμό ή τη διασταύρωση δεδομένων από διαφορετικές πηγές, για την αντιμετώπιση της ανισοκατανομής. Η προσέγγιση αυτή είναι ιδιαίτερα συνδεδεμένη σε εφαρμογές υγείας. Αντιθέτως, η μείωση δεδομένων βοηθάει στις περιπτώσεις που η επικρατούμενη κλάση υπερτερεί σημαντικά σε αριθμό δειγμάτων έναντι των υπολοίπων. Στη περίπτωση αυτή, αφαιρούνται επιλεκτικά δείγματα από την κλάση πλειοψηφίας, υπό την προϋπόθεση ότι διατηρείται επαρκής αριθμός αντιπροσωπευτικών παρατηρήσεων, ώστε να μην αλλοιωθεί η πληροφοριακή της περιεκτικότητα. Η επιλογή της επαύξησης ή μείωσης δεδομένων εξαρτάται από το μέγεθος και την ιδιαιτερότητα του συνόλου δεδομένων, την φύση του προβλήματος και τον κίνδυνο εισαγωγής θορύβου ή απώλειας πληροφορίας.

Η επιλογή κατάλληλου μοντέλου ML αποτελεί κρίσιμο στάδιο σε κάθε αναλυτική διαδικασία, καθώς επηρεάζει άμεσα τόσο την προβλεπτική απόδοση όσο και τη δυνατότητα γενίκευσης σε νέα δεδομένα. Η διαδικασία αυτή δεν περιορίζεται στην επιλογή ενός συγκεκριμένου αλγορίθμου, αλλά περιλαμβάνει την αξιολόγηση πολλαπλών υποψηφίων μοντέλων σε συνάρτηση με τη φύση των δεδομένων, το μέγεθος του δείγματος και τον εκάστοτε ερευνητικό στόχο.

Μία καλή πρακτική εκτίμησης της γενίκευσης των δεδομένων είναι διάσπασή τους σε επιμέρους σύνολα. Αυτό επιτρέπει την εκπαίδευση, την αξιολόγηση και σε πιο σύνθετες περιπτώσεις την επικύρωση των μοντέλων με τρόπο που περιορίζει τη μεροληψία και ενισχύει τη δυνατότητα γενίκευσης. Κατά περιόδους αναπτύχθηκαν διάφορες μορφές διάσπασης, όπως είναι οι εξής:

- **Διαχωρισμός εκπαίδευσης-ελέγχου** (train-test split): Ο απλούστερος και ευρύτερα χρησιμοποιούμενος τρόπος διάσπασης, όπου το σύνολο δεδομένων χωρίζεται σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Συνήθως η αναλογία διακυμαίνεται σε 70–30, 80–20 ή 90-10. Η μέθοδος αυτή προσφέρει γρήγορη εκτίμηση της απόδοσης.
- **Κ-φορών Διασταυρωμένη Επικύρωση** (K-Fold Cross Validation): Εδώ, τα δεδομένα χωρίζονται σε Κ ισομεγέθη υποσύνολα, εκ των οποίων το ένα χρησιμοποιείται ως σύνολο ελέγχου και τα υπόλοιπα ως σύνολο εκπαίδευσης, επαναλαμβάνοντας τη διαδικασία Κ φορές, με τέτοιο τρόπο ώστε κάθε υποσύνολο να λαμβάνει ρόλο συνόλου ελέγχου ακριβώς μία φορά (βλέπε σχήμα 2.9). Έτσι, επιτυγχάνεται η αξιολόγηση του μοντέλου σε διαφορετικά τμήματα δεδομένων. Η τεχνική αυτή παρέχει μία πιο αξιόπιστη εκτίμηση της γενίκευσης σε σύγκριση με έναν απλό διαχωρισμό [22].

- **Στρωματοποιημένη Διασταυρούμενη Επικύρωση**: Η στρωματοποιημένη εκδοχή της K-φορών Διασταυρούμενης Επικύρωσης, διατηρεί την αναλογία των κλάσεων σε κάθε υποσύνολο, γεγονός ιδιαίτερα σημαντικό σε προβλήματα με ανισορροπία κλάσεων. Η μέθοδος αυτή χρησιμοποιείται εκτενώς σε εφαρμογές υγείας και ταξινόμησης κινδύνου [78].
- **Leave-One-Out Διασταυρούμενη Επικύρωση:** Αποτελεί επίσης μία ειδική περίπτωση της K-φορών Διασταυρούμενης Επικύρωσης, όπου κάθε παρατήρηση χρησιμοποιείται διαδοχικά ως σύνολο ελέγχου. Παρέχει σχεδόν αμερόληπτη εκτίμηση, αλλά είναι υπολογιστικά δαπανηρή και παρουσιάζει υψηλή διακύμανση [78]. Συνήθως χρησιμοποιείται σε μικρά σύνολα δεδομένων.



Σχήμα 2.9: Απεικόνιση διαδικασίας 5-fold Cross Validation [79]

Ιδιαίτερη προσοχή πρέπει να δίνεται κατά τη διαδικασία τροφοδοσίας των δεδομένων διάσπασης στα μοντέλα ML. Τα δεδομένα του συνόλου ελέγχου θα πρέπει πάντοτε να είναι διαφορετικά από τα δεδομένα εκπαίδευσης, και να μην εξετάζονται ποτέ κατά τη διάρκεια επιλογής μοντέλου, παρά μόνο στη φάση αξιολόγησης του μοντέλου [22]. Έτσι διασφαλίζεται η ορθή εκτίμηση του μοντέλου σε άγνωστα δεδομένα, που δεν έχουν χρησιμοποιηθεί κατά τη διαδικασία εκπαίδευσης, αποφεύγοντας κάθε πιθανή μεροληψία.

2.4.3 Υπερπαράμετροι

Κάθε μοντέλο Μηχανικής Μάθησης χαρακτηρίζεται από δύο βασικές κατηγορίες παραμέτρων: (α) τις εσωτερικές παραμέτρους, οι οποίες εκτιμώνται κατά τη φάση εκπαίδευσης και (β) τις υπερπαραμέτρους ή αλλιώς γενικές παραμέτρους, οι οποίες καθορίζονται εκ των προτέρων και ρυθμίζουν τη συμπεριφορά του αλγορίθμου. Η διαδικασία ρύθμισης υπερπαραμέτρων, αποτελεί μέρος της επιλογής μοντέλου και αποσκοπεί στην βελτιστοποίηση της απόδοσης με βάση προκαθορισμένα κριτήρια [24]. Παραδείγματα υπερπαραμέτρων αποτελούν ο βαθμός κανονικοποίησης σε SVMs μοντέλα ή το βάθος ενός δέντρου αποφάσεων. Η επιλογή ακατάλληλων ή η χρήση προκαθορισμένων υπερπαραμέτρων, μπορεί να οδηγήσει είτε σε υποπροσαρμογή είτε σε υπερπροσαρμογή του μοντέλου. Κάθε κατηγορία μοντέλων έχει διαφορετικές υπερπαραμέτρους. Η επιλογή του καλύτερου αλγορίθμου εξαρτάται κατά βάση από την ποιότητα των δεδομένων.

Μία αρχική προσέγγιση για την επιλογή κατάλληλων τιμών γενικών παραμέτρων είναι η χειρονακτική δοκιμή διαφορετικών συνδυασμών για την εύρεση του καλύτερου δυνατού. Ωστόσο, μία τέτοια προσέγγιση

καθίσταται μη πρακτική. Καθώς το πλήθος των υπερπαραμέτρων προς εξέταση μεγαλώνει, αυξάνεται εκθετικά το πλήθος των τιμών εξέτασης. Για τον λόγο αυτό, έχουν αναπτυχθεί αυτοματοποιημένοι αλγόριθμοι που τακτοποιούν την αναζήτηση, ανακαλύπτοντας και επιλέγοντας τον καλύτερο συνδυασμό υπερπαραμέτρων για το αντίστοιχο μοντέλο στα δεδομένα που εκπαιδεύεται [36].

Μία από τις πιο διαδεδομένες τεχνικές είναι η αναζήτηση πλέγματος σε συνδυασμό με διασταυρούμενη επικύρωση (GridSearchCV). Πρόκειται για μία εξαντλητική αναζήτηση πλέγματος η οποία εξετάζει κάθε συνδυασμό τιμών υπερπαραμέτρων που τροφοδοτείται, χρησιμοποιώντας την τεχνική διασταυρούμενης επικύρωσης που αναφέρθηκε παραπάνω. Για παράδειγμα, εάν εξετάζονται τρεις υπερπαράμετροι με τέσσερις πιθανές τιμές η καθεμία, τότε υπάρχουν $4^3 = 64$ συνδυασμοί, και αντίστοιχα 64 ξεχωριστά μοντέλα προς εκτέλεση. Κάθε ατομική εκτέλεση αποτελείται από την εκπαίδευση και την αξιολόγηση του μοντέλου με χρήση της διασταυρούμενης επικύρωσης στα διαθέσιμα δεδομένα. Για ένα πιο πολύπλοκο μοντέλο ή ένα μεγάλο σύνολο δεδομένων η ολοκλήρωση εκτέλεσης του αλγορίθμου μπορεί να λάβει από μερικές ώρες μέχρι και ημέρες. Για το σκοπό αυτό έχουν κατασκευαστεί εναλλακτικές προσεγγίσεις.

Η τυχαία αναζήτηση υπερπαραμέτρων (RandomSearchCV), είναι μία πιο σύγχρονη και αποτελεσματικότερη μέθοδος η οποία προτάθηκε για πρώτη φορά το 2012 [80]) και αντιμετωπίζει το πρόβλημα που μπορεί να προκληθεί σε μεγάλα σύνολα δεδομένων ή για μεγάλο πλήθος υπερπαραμέτρων. Συγκεκριμένα εφαρμόζει την ίδια λογική με την προηγούμενη τεχνική, με την διαφορά ότι οι συνδυασμοί υπερπαραμέτρων επιλέγονται τυχαία από προκαθορισμένες κατανομές, και ο αριθμός των δοκιμών ορίζεται εκ των προτέρων [80].

Η βασική διαφορά μεταξύ των δύο μεθόδων έγκειται στον αριθμό και στον τρόπο εξερεύνησης των συνδυασμών υπερπαραμέτρων. Η εξαντλητική αναζήτηση πλέγματος αποφέρει σίγουρα το βέλτιστο μοντέλο, με τις ευρεθείσες κατάλληλες τιμές γενικών παραμέτρων, με τίμημα τον αυξημένο χρόνο αναζήτησης. Απεναντίας, η τυχαία αναζήτηση υπερπαραμέτρων εκτελείται σε περιορισμένο πλήθος συνδυασμών και καταλήγει στο καλύτερο μοντέλο, μεταξύ αυτών που εξετάστηκαν. Τις περισσότερες φορές επιτυγχάνει ικανοποιητικά αποτελέσματα που προσεγγίζουν εκείνα της εξαντλητικής αναζήτησης υπερπαραμέτρων και μάλιστα σε μικρότερο χρονικό διάστημα. Η επιλογή της κατάλληλης προσέγγισης εξαρτάται καθαρά από τις απαιτήσεις της εφαρμογής, τους διαθέσιμους πόρους και τα δεδομένα.

Συνοψίζοντας, η ρύθμιση υπερπαραμέτρων αποτελεί αναπόσπαστο μέρος της διαδικασίας ανάπτυξης αξιόπιστων μοντέλων ML, καθώς επηρεάζει καθοριστικά τόσο την απόδοση όσο και τη δυνατότητα γενίκευσης των μοντέλων. Η επιλογή μεταξύ εξαντλητικών και στοχαστικών μεθόδων αναζήτησης, όπως η Grid Search και η Random Search, συνεπάγεται ένα συμβιβασμό μεταξύ υπολογιστικού κόστους και πληρότητας εξερεύνησης του χώρου υπερπαραμέτρων. Και οι δύο προσεγγίσεις υποστηρίζονται από τη βιβλιοθήκη scikit-learn [62], η οποία παρέχει τυποποιημένες υλοποιήσεις που διευκολύνουν την ενσωμάτωσή τους σε σύγχρονες ροές εργασίας ML. Στο πλαίσιο εφαρμογών με αυξημένες απαιτήσεις αξιοπιστίας, όπως αυτές που αφορούν την ανάλυση δεδομένων υγείας, η ορθολογική επιλογή και εφαρμογή μεθόδων υπερπαραμετροποίησης συμβάλλει ουσιαστικά στη βελτιστοποίηση των μοντέλων και στην ενίσχυση της επιστημονικής εγκυρότητας των αποτελεσμάτων.

2.4.4 Ερμηνευσιμότητα Μοντέλων

Η ερμηνευσιμότητα των μοντέλων ML αποτελεί έναν κρίσιμο παράγοντα στον κλάδο της ΤΝ, συμβάλλοντας στην ποιοτική και ολοκληρωμένη εικόνα ενός καλά σχεδιασμένου συστήματος. Η ερμηνευσιμότητα προϋποθέτει ότι οι προβλέψεις των μοντέλων δεν επαρκούν για το αποτέλεσμα των

μεθόδων. Ειδικότερα, όταν η ML εφαρμόζεται σε τομείς όπως η υγεία, οι απαιτήσεις ενός σύγχρονου μοντέλου είναι μεγαλύτερες. Δεν αρκεί απλώς η βελτιστοποίηση της ακρίβειας των προβλέψεων. Για παράδειγμα, μπορεί να υπάρχει δυσκολία όταν τα αντικείμενα του πραγματικού κόσμου πρέπει να κωδικοποιηθούν και ερμηνευθούν ορθά σε τιμές συναρτήσεων από τα μοντέλα. Κατά κόρον, σε αυτές τις εφαρμογές υπαγορεύεται μία πιο απαιτητική και συγκροτημένη ανάλυση των αποτελεσμάτων, εκπροσωπώντας το δυνατόν περισσότερο τα αντικείμενα πραγματικού κόσμου, καθώς στην πραγματικότητα αφορούν τους τρόπους αντιμετώπισης και διαχείρισης αληθινών καταστάσεων. Επομένως, η ερμηνευσιμότητα των μοντέλων εξυπηρετεί την εφαρμογή στα δεδομένα πραγματικού κόσμου τα οποία θεωρούνται σημαντικά αλλά συχνά τα μοντέλα είναι πολύπλοκα και δυσκολεύονται να κατανοηθούν με σαφή τρόπο από τους χρήστες [81].

Η αξιοποίηση της ερμηνευσιμότητας παρέχει πολλά οφέλη στα συστήματα τεχνικών ML, ανάλογα με την ανάγκη της εφαρμογής. Συχνά αιτιολογείται ως προϋπόθεση για την οικοδόμηση εμπιστοσύνης προς αυτά. Ωστόσο, η έννοια της εμπιστοσύνης δεν είναι μονοσήμαντη και χρήζει περαιτέρω αποσαφήνιση. Σε αυτές τις περιπτώσεις δεν επαρκεί απλώς η πεποίθηση το μοντέλο να αποδώσει με υψηλή ακρίβεια στις προβλέψεις του, αλλά μπορεί να οριστεί και με πιο υποκειμενικούς όρους διασφαλίζοντας την πρακτική εφαρμογή [81]. Αυτό συμβαίνει όταν οι στόχοι εκπαίδευσης ενός μοντέλου αποκλίνουν από τους πραγματικούς στόχους ή τα σενάρια εφαρμογής του. Σε τέτοιες περιπτώσεις, η εμπιστοσύνη δεν αφορά μόνο την ακρίβεια των προβλέψεων, αλλά και τη βεβαιότητα ότι το μοντέλο θα συμπεριφερθεί κατάλληλα σε πραγματικές, ενδεχομένως απρόβλεπτες συνθήκες.

Παρότι τα μοντέλα SL βελτιστοποιούνται πρωτίστως για την ανίχνευση και μοντελοποίηση συσχετίσεων μεταξύ μεταβλητών, τα αποτελέσματά τους μπορούν να χρησιμοποιηθούν για την εξαγωγή ενδείξεων ή τη διατύπωση υποθέσεων σχετικά με φαινόμενα του πραγματικού κόσμου. Φυσικά είναι σημαντικό να τονιστεί ότι οι συσχετίσεις που προκύπτουν δεν συνεπάγονται αποκλειστικά αιτιώδη σχέση. Ωστόσο, δύναται η δυνατότητα μέσω της ερμηνευσιμότητας, τα μοντέλα να λειτουργήσουν ως αφετηρία για τη διατύπωση επιστημονικών υποθέσεων και να εξεταστούν και επαληθευτούν μέσω κλινικών μελετών [81]. Μερικές από τις τεχνικές ερμηνευσιμότητας εξετάζονται στο κεφάλαιο 5 της παρούσας εργασίας, όπου έχει τεκμηριωθεί και επιλεγεί η χρήση του βέλτιστου μοντέλου.

Τελευταίο αλλά εξίσου σημαντικό, αποτελεί η σημασία της πληροφορίας που προσδίδει στην αιτιολόγηση των αποτελεσμάτων εξόδου. Σε πολλές εφαρμογές, τα μοντέλα ML δεν αποσκοπούν στη λήψη αυτόνομων αποφάσεων, αλλά λειτουργούν υποστηρικτικά προς τον άνθρωπο, παρέχοντας πληροφορίες που διευκολύνουν τη διαδικασία λήψης αποφάσεων. Σε τέτοιες καταστάσεις, ο ρόλος του μοντέλου δεν είναι απαραίτητα να αντικαταστήσει τον ανθρώπινο παράγοντα, αλλά να προσφέρει χρήσιμα τεκμήρια και ενδείξεις που μπορούν να αξιοποιηθούν από ειδικούς ή κλινικούς χρήστες, γεγονός που υιοθετείται στη παρούσα εργασία.

2.5 Ανασκόπηση Σχετικών Εργασιών

Τα τελευταία χρόνια, έχει παρατηρηθεί έντονη ερευνητική δραστηριότητα στην ανάπτυξη προσεγγίσεων για την ανίχνευση πτώσεων, την πρόβλεψη μελλοντικών συμβάντων και την εκτίμηση του κινδύνου πτώσης. Οι προσεγγίσεις αυτές αξιοποιούν δεδομένα από αισθητήρες, κλινικές δοκιμασίες και ιστορικό πτώσεων των ατόμων, περιβαλλοντικές παρατηρήσεις και δημογραφικά χαρακτηριστικά. Παρά την συστηματική πρόοδο που έχει επιτευχθεί μέσω του ML στην αντιμετώπιση των πτώσεων, εξακολουθούν να υφίστανται σημαντικές προκλήσεις των αντίστοιχων τεχνικών. Οι δυσκολίες απορρέουν κυρίως από την ετερογένεια και

τις ιδιαιτερότητες του πληθυσμού στον οποίο εφαρμόζονται τα μοντέλα, τις διαφοροποιήσεις των συμμετεχόντων από τους οποίους συλλέγονται τα δεδομένα και η μη πολυπαραγοντική χρήση παραγόντων κινδύνου.

Παρακάτω παρουσιάζεται η σχετική βιβλιογραφική ανασκόπηση, γύρω από τις ήδη υπάρχουσες μελέτες. Οι θεματικές με τις οποίες περιγράφονται, σχετίζονται με το πρωτεύων ερευνητικό ζητούμενο της κάθε μελέτης:

1) ανίχνευση πτώσης,
2) πρόβλεψη και εκτίμηση κινδύνου πτώσης,
3) υποστήριξη παρεμβάσεων/λήψη αποφάσεων για πρόληψη από πτώσεις

Στη μελέτη των Hussain et al. [16], στόχος ήταν η ανάπτυξη ενός αποδοτικού αλγορίθμου ανίχνευσης πτώσης με κλασικές μεθόδους SL, με σκοπό τη μείωση του χρόνου απόκρισης/ειδοποίησης μετά από ένα συμβάν πτώσης. Χρησιμοποιήθηκε το σύνολο δεδομένων SisFall, το οποίο περιέχει τόσο δραστηριότητες καθημερινής ζωής (ADLs) όσο και πτώσεις. Σε επίπεδο χαρακτηριστικών, η μελέτη βασίστηκε κυρίως σε μετρήσεις IMU, παρέχοντας κλινικούς και δημογραφικούς παράγοντες. Χρησιμοποιήθηκαν οι αλγόριθμοι SVM, KNN (k=1), Δέντρο Απόφασης και Λογιστική Παλινδρόμηση, από τους οποίους την υψηλότερη επίδοση είχε ο SVM (99.98% ακρίβεια). Κρίσιμο παράγοντα της μελέτης, αποτελεί ότι τα δεδομένα συλλέχθηκαν από προσομοιωμένες δοκιμές, γεγονός που περιορίζει την μεταφορά των αποτελεσμάτων σε πραγματικά περιστατικά.

Στην έρευνα των Bargiotal et al. [82], διερευνήθηκε ο διαχωρισμός των ασθενών με σύνδρομο Πάρκινσον που έχουν πέσει έναντι αυτών που δεν έχουν πέσει, με στόχο την ανάδειξη πολυπαραγοντικών διαφορών στη στάση και την ισορροπία που σχετίζονται με αυξημένη συχνότητα πτώσεων. Το σύνολο δεδομένων περιλαμβάνει 123 ασθενείς, όπου τα δεδομένα συλλέχθηκαν από έναν αισθητήρα πίεσης δαπέδου. Βασίστηκε σε χαρακτηριστικά με κλινικές δοκιμασίες, ενώ ακόμη περιλαμβάνει κλινικές και περιβαλλοντικές μετρήσεις. Το πρακτικό τμήμα της έρευνας, αξιοποίησε τα δεδομένα με τη χρήση του αλγορίθμου Τυχαίου Δάσους ενώ χρησιμοποιήθηκαν και παραδοσιακές μέθοδοι στατιστικής ανάλυσης. Τα αποτελέσματα παρουσίασαν ότι η πολυμεταβλητή προσέγγιση εντοπίζει σημαντικές διαφορές μεταξύ των ομάδων. Ακόμη αναδεικνύθηκαν χαρακτηριστικά που διακρίνουν τις ομάδες μεταξύ τους. Ως περιορισμοί αναφέρονται ζητήματα σχετικά με την ανισοκατανομή του πλήθους των ομάδων (μικρότερο ποσοστό αντιπροσώπευσης της ομάδας των ατόμων που είχαν υποστεί πτώση) και την εστίαση σε συγκεκριμένη κλινική δοκιμασία.

Το άρθρο των Lindberg et al. [83] είχε ως στόχο την βελτίωση και μείωση των συμβάντων πτώσεων εντός νοσοκομειακών εγκαταστάσεων και συγχρόνως την κατανόηση των πιο σημαντικών παραγόντων που οδηγούν στην πτώση. Εφαρμόστηκαν τεχνικές SL, και συγκεκριμένα: Δέντρο Απόφασης, Bagging, Τυχαία Δάση και Adaptive Boosting, με τη χρήση 10-φορών Διασταυρούμενη Επικύρωση. Η κύρια μετρική αξιολόγησης αποτέλεσε η AUROC αλλά αξιοποιήθηκαν και δείκτες ευαισθησίας και εξειδίκευσης. Οι αλγόριθμοι εφαρμόστηκαν πάνω σε ένα "κλειστό" σύνολο δεδομένων από EHR δεδομένα 814 ασθενών, που είτε έπεσαν είτε όχι κατά την εισαγωγή τους σε νοσοκομείο. Λήφθηκαν μέσα από 14 υγειονομικές μονάδες νοσοκομείων, καθώς παράλληλα περιελάμβαναν 38 γνωρίσματα σχετικά με κλινικά και υγειονομικά δεδομένα, ιστορικό κ.α. Τα αποτελέσματα ανέδειξαν ότι τα μοντέλα ομάδων υπερέχουν των απλών Δέντρων Απόφασης, όπου τα Τυχαία Δάση υπερίσχυσαν καταφθάνοντας 0.9 τιμή αξιολόγησης AUROC. Σημαντικοί προγνωστικοί παράγοντες αποτέλεσαν το ιστορικό πτώσεων, η ηλικία, η ποιότητα βάδισης, η λειτουργική

μεταβλητή αξιολόγησης MFS, η νοητική κατάσταση, ο τύπος μονάδας και ο αριθμός φαρμάκων που αυξάνουν τον κίνδυνο πτώσεων. Σύμφωνα με το άρθρο, οι παραπάνω παράγοντες και η συνολική μελέτη, έγκειται στην αξιοποίηση των ευρημάτων από τα νοσοκομεία και το υγειονομικό προσωπικό και στην βελτιστοποίηση της διαχείρισης του κινδύνου πτώσης λαμβάνοντας πιο ορθές μελλοντικές αποφάσεις.

Στην τελευταία αλλά εξίσου σημαντική μελέτη, αυτή των Ye et al. [84], αναπτύχθηκε ένα εργαλείο πρόβλεψης πτώσης εντός ενός έτους, κυρίως για την πρόληψη ηλικιωμένων υψηλού κινδύνου, αξιοποιώντας μεγάλης κλίμακας EHR δεδομένα, με τελικό στόχο την παραγωγή έγκαιρων προειδοποιήσεων και την ανάδειξη εξατομικευμένων παραγόντων ώστε να διευκολυνθούν στοχευμένες παρεμβάσεις πρόληψης. Το σύνολο δεδομένων διαμορφώθηκε από 265 225 συμμετέχοντες ηλικίας άνω των 65 ετών που επισκέφτηκαν υγειονομικές εγκαταστάσεις της πολιτείας Maine των Η.Π.Α., εντός δύο ετών. Οι μεταβλητές του συνόλου δεδομένων χαρακτηρίζονται από κλινικά και επιχειρησιακά δεδομένα συμπεριλαμβανομένων ιστορικό πτώσεων, διαταραχές βάδισης και φαρμακευτικής αγωγής. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο XGBoost, ο οποίος αξιολογήθηκε με βάση την μετρική C-statistic (δείκτης διάκρισης επιπέδων κινδύνου όπως η πτώση) και την καμπύλη ROC. Το αποτέλεσμα της μετρικής ήταν 0.807 γεγονός που αποδεικνύει ότι η αξιολόγηση κινδύνου πτώσης των ατόμων ήταν αρκετά καλή. Σημαντικό εύρημα της μελέτης αποτελεί ότι το 50% των ατόμων που εκτιμήθηκαν ότι ανήκουν στο προφίλ υψηλού κινδύνου πτώσης, πράγματι έπεσαν εντός του επόμενου τριμήνου. Οι πέντε πιο σημαντικοί παράγοντες πτώσης, μεταξύ 137 παραγόντων, αναδείχθηκαν οι γνωστικές διαταραχές, προβλήματα ισορροπίας και βάδισης, διάγνωση με Πάρκινσον, ιστορικό πτώσεων και η οστεοπόρωση. Με αυτό τον τρόπο επιτρέπεται η διευκόλυνση των παρεμβάσεων που πρέπει να γίνουν ανάλογα με το αντίστοιχο προφίλ κινδύνου του κάθε ατόμου. Πιθανά μειονεκτήματα της έρευνας αποτελούν η εξάρτηση από συγκεκριμένες πρακτικές τεκμηρίωσης των πτώσεων που μπορεί να διαφέρουν μεταξύ των υγειονομικών συστημάτων καταγραφής, καθώς και η πιθανή μεροληψία που μπορεί να προκύψει εάν το μοντέλο εφαρμοστεί σε διαφορετικούς πληθυσμούς. Ωστόσο οι συγγραφείς, υποστηρίζουν ότι το μοντέλο μπορεί να εφαρμοστεί σε υγειονομικά πληροφοριακά συστήματα για την παραγωγή αυτοματοποιημένων ειδοποιήσεων στους ασθενείς υψηλού κινδύνου πτώσης και την ανάδειξη εξατομικευμένων προφίλ κινδύνου για την διευκόλυνση των παρεμβάσεων.

Συνολικά η βιβλιογραφία καταδεικνύει ότι οι τεχνικές ML έχουν αξιοποιηθεί εκτενώς για την ανίχνευση πτώσεων, την αξιολόγηση κινδύνου αλλά και την υποστήριξη παρεμβάσεων πρόληψης. Οι αλγόριθμοι που χρησιμοποιούνται συχνά είναι οι SVM, Τυχαία Δάση, αλγόριθμοι ενίσχυσης ομάδων. Παρά τα υψηλά επίπεδα απόδοσης που αναφέρονται, οι μελέτες εμφανίζουν περιορισμούς όπως είναι η έμφαση στην ανίχνευση συμβάντων πτώσης αντί πρόληψης, η επιλογή μη γηραιού πληθυσμού συμμετεχόντων, η εξάρτηση από μονοδιάστατους τύπους δεδομένων (όπως μόνο κλινικά δεδομένα), περιορισμένη γενίκευση και έλλειψη σε αναφερόμενες εξατομικευμένες παρεμβάσεις. Επομένως, απαιτείται η χρήση υβριδικών προσεγγίσεων, που να συνδυάζουν τόσο την ανακάλυψη προφίλ κινδύνου πτώσης, την πρόβλεψη ατόμων σε αυτές και ουσιαστικά εξατομικευμένες παρεμβάσεις για την πρόληψη από τις πτώσεις, μέσω πολυδιάστατων δεδομένων.

2.6 Επίλογος

Στο κεφάλαιο αυτό περιγράφηκε το θεωρητικό υπόβαθρο της εργασίας, παρουσιάζοντας την ερευνητική μελέτη που διαδραματίστηκε. Στην αρχή έγινε μία ανάλυση σχετικά με τη σημασία των πτώσεων, τις συνέπειες και την αιτιότητά τους. Στη συνέχεια περιγράφηκαν οι παραδοσιακές και οι σύγχρονες τεχνικές αξιολόγησης του κινδύνου πτώσεων και το παραγόντων τους, συγκρίνοντας τις δύο κατηγορίες τεχνικών. Ακόμη, επιδείχθηκαν θεωρητικά οι αλγόριθμοι (συσταδοποίησης) και τα μοντέλα (ταξινόμησης) ML που

χρησιμοποιήθηκαν κατά τη διάρκεια εκπόνησης της εργασίας, εστιάζοντας στα πιο σημαντικά χαρακτηριστικά, τα πλεονεκτήματα και τα μειονεκτήματά τους. Κατόπιν, αναδείχθηκαν οι τεχνικές προεπεξεργασίας δεδομένων που χρησιμοποιήθηκαν στην εργασία, ενώ αργότερα περιγράφηκαν οι μέθοδοι που συνέβαλαν στην ανάπτυξη και αξιολόγηση των αλγορίθμων. Τελικώς, παρουσιάστηκε η ανασκόπηση σχετικών εργασιών.

Κεφάλαιο 3ο: Δεδομένα και Μεθοδολογία

Η ποιότητα, η δομή και η καταλληλότητα των δεδομένων αποτελούν καθοριστικούς παράγοντες για την επιτυχία κάθε προσέγγισης ML, ιδίως σε εφαρμογές που ασχολούνται με τον τομέα της υγείας. Στο πεδίο της ανίχνευσης και πρόληψης πτώσεων, τα δεδομένα δεν λειτουργούν απλώς ως είσοδοι αλγορίθμων, αλλά ενσωματώνουν πολύ σημαντική πληροφορία σχετικά με τη λειτουργική κατάσταση, την κινητικότητα και γενικώς τα χαρακτηριστικά των ατόμων. Ως εκ τούτου, η συστηματική κατανόηση των διαθέσιμων συνόλων δεδομένων και των ιδιαιτεροτήτων τους αποτελεί αναγκαία προϋπόθεση για την επιλογή του κατάλληλου συνόλου και τη σχεδίαση αξιόπιστων και γενικεύσιμων μοντέλων.

Το παρόν κεφάλαιο εστιάζει τόσο στη φύση των δεδομένων που αξιοποιούνται στην εργασία όσο και στη μεθοδολογία που ακολουθήθηκε στο πρακτικό σκέλος. Αρχικά, παρουσιάζονται αντιπροσωπευτικά δημόσια διαθέσιμα σύνολα δεδομένων τα οποία αποτέλεσαν υποψήφια προς επιλογή. Έχουν χρησιμοποιηθεί εκτενώς στη διεθνή βιβλιογραφία για τη μελέτη των πτώσεων, τόσο σε θεωρητικό όσο και σε πρακτικό κομμάτι. Αναδεικνύονται τα βασικά χαρακτηριστικά τους, οι τύποι δεδομένων που περιλαμβάνουν (όπως δεδομένα αισθητήρων ή βίντεο), καθώς και τα πλεονεκτήματα και οι περιορισμοί τους. Στο παράρτημα Α εμφανίζονται με συμπαγή κατάσταση όλα τα σύνολα δεδομένων που μελετήθηκαν κατά την εκπόνηση της έρευνας της διπλωματικής. Στη συνέχεια το κεφάλαιο επικεντρώνεται στο σύνολο δεδομένων που επιλέχθηκε για την παρούσα εργασία, τεκμηριώνοντας τους λόγους της επιλογής του έναντι εναλλακτικών και αναλύοντας τη δομή και το περιεχόμενό του. Ακολουθεί η ανάδειξη των βασικών θεμάτων και προκλήσεων που σχετίζονται με τα δεδομένα, και αποτέλεσαν σημαντική επιρροή για τον σχεδιασμό της προεπεξεργασίας τους. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση του μεθοδολογικού πλαισίου που εφαρμόστηκε, συνδέοντας τα δεδομένα, τις τεχνικές ML και τα στάδια της πειραματικής ροής που υλοποιήθηκαν στο πλαίσιο της εργασίας.

3.1 Δημόσια Διαθέσιμα Datasets

Τα τελευταία χρόνια έχουν διερευνηθεί και πραγματοποιηθεί πολλές μελέτες γύρω από το πλαίσιο των πτώσεων των ατόμων, δημιουργώντας μία πληθώρα από datasets, τα οποία διαφέρουν ως προς το είδος των δεδομένων, τον σκοπό, το πληθυσμιακό δείγμα, και τις μεθόδους καταγραφής των δεδομένων. Σε αυτή την ενότητα παρουσιάζονται τέσσερα γνωστά και ευρέως χρησιμοποιούμενα datasets,

SisFall Dataset

Το SisFall αποτελεί ένα από τα πιο γνωστά και εκτενώς χρησιμοποιούμενα datasets στον τομέα της ανίχνευσης πτώσεων, με έμφαση στα δεδομένα αισθητήρων IMU [85]. Το dataset περιλαμβάνει καταγραφές από 2 επιταχυνσιόμετρα και 1 γυροσκόπιο, με συχνότητα δειγματοληψίας 200Hz, τα οποία τοποθετούνται στη περιφέρεια της μέσης του ατόμου, καταγράφοντας τρισδιάστατα σήματα επιτάχυνσης και γωνιακής ταχύτητας.

Οι καταγραφές αφορούν προσομοιωμένα σενάρια πτώσεων και καθημερινών δραστηριοτήτων (Activities of Daily Living – ADL), σε ειδικούς πειραματικούς χώρους. Ειδικότερα, περιέχονται 19 διαφορετικά ADL και 15 τύποι πτώσης εκτελεσμένα από 23 νεαρά άτομα, 15 ADL τύποι από 14 υγιείς ανθρώπους έως την ηλικία των 62 ετών, ενώ 1 μόλις συμμετέχων σε ηλικία 60 ετών εκτέλεσε όλους τους διαφορετικούς τύπους ADL και πτώσεων. Επομένως το πείραμα περιλάμβανε 38 άτομα (κάθε φύλου). Τα δεδομένα αποτελούν 1,789

καταγραφές πτώσεων και 2,707 ADL, και παράλληλα παρέχονται σε μορφή χρονοσειρών, καθιστώντας το dataset κατάλληλο για μεθόδους ανάλυσης χρονικών σημάτων και εξαγωγής χαρακτηριστικών.

Βασικό πλεονέκτημα του SisFall είναι ο μεγάλος όγκος δεδομένων και η σαφής δομή του, ωστόσο ένας συχνά αναφερόμενος περιορισμός είναι ότι μεγάλο μέρος των πτώσεων έχει προσομοιωθεί από νεότερα άτομα, γεγονός που μπορεί να επηρεάσει τη γενίκευση των αποτελεσμάτων σε πραγματικό ηλικιωμένο πληθυσμό [85]. Παράλληλα με αυτό, δεν περιλαμβάνονται δεδομένα ηλικιωμένων ατόμων άνω των 65 ετών, ηλικία που θεωρείται ορόσημο σε εφαρμογές πρόληψης ηλικιωμένων ατόμων από τον κίνδυνο πτώσης (εδώ ανώτατο όριο είναι τα 62 έτη) [86].

e**HomeSeniors Dataset**

Το eHomeSeniors dataset αποτελεί ένα δημόσια προσβάσιμο σύνολο δεδομένων που έχει σχεδιαστεί ειδικά για έρευνα στην ανίχνευση πτώσεων και τη παρακολούθηση κινητικών συμπεριφορών ηλικιωμένων σε οικιακό περιβάλλον, μέσω περιβαλλοντικών αισθητήρων [21]. Σε αντίθεση με τους αισθητήρες που χρησιμοποιούνται στα περισσότερα datasets, το eHomeSeniors παρέχει δεδομένα από δύο διαφορετικά είδη υπέρυθρωμ θερμικών αισθητήρων, οι οποίοι είναι χαμηλού κόστους. Στο πείραμα χρησιμοποιήθηκαν τέσσερις “Omron D6T-8L-06” και ένας “Melexis MLX90640” (βλέπε κεφάλαιο 2.1.3). Το πρώτο είδος αισθητήρα τοποθετήθηκε στο πάτωμα (ή πολύ κοντά στο ύψος του πατώματος) ενώ το δεύτερο στον τοίχο, με συχνότητα δειγματοληψίας στα 5Hz.

Το dataset περιλαμβάνει μόνο πτώσεις, και πιο συγκεκριμένα 15 διαφορετικούς τύπους πτώσεων. Για την προσομοίωση των πτώσεων, χρησιμοποιήθηκαν δύο κατηγορίες συμμετεχόντων: 3 υγιείς νεαροί και 3 κατάλληλα εκπαιδευμένοι ηθοποιοί, έτσι ώστε να μιμούνται τον τρόπο που πέφτουν οι ηλικιωμένοι [21]. Με αυτό τον τρόπο, το πείραμα επιδιώκει να γεφυρώσει τις φυσικές πτώσεις μεταξύ των νεαρών και των ηλικιωμένων. Υπάρχουν 180 αρχεία σε .csv μορφή, τα οποία έχουν καταγράψει 448 πτώσεις.

Ένα από τα πλεονεκτήματα του eHomeSeniors dataset είναι η προστασία της ιδιωτικότητας χάρη στη χρήση θερμικών καμερών για τη συλλογή των δεδομένων, χωρίς να παραβιάζει την ιδιωτικότητα των χρηστών. Σχετικά με τα μειονεκτήματα, τα άτομα που κάνουν προσομοιώσεις των πτώσεων είναι λίγα σε αριθμό (μόλις 6), γεγονός που υπονοεί πολύ μικρή αντιπροσώπευση. Ομοίως, από το dataset εκλείπουν δεδομένα από πραγματικούς ηλικιωμένους, περιορίζοντας έτσι την πραγματική εικόνα των χαρακτηριστικών πτώσεων των ηλικιωμένων.

MobiFall dataset

Το MobiFall dataset σχεδιάστηκε με στόχο την αξιοποίηση των αισθητήρων που είναι ενσωματωμένοι σε έξυπνα κινητά τηλέφωνα, καθιστώντας το ιδιαίτερα σχετικό με φορητές και χαμηλού κόστους προσεγγίσεις για ανίχνευση πτώσεων [87]. Τα δεδομένα συλλέχθηκαν από ενσωματωμένους αισθητήρες, όπως επιταχυνσιόμετρα και γυροσκόπια, ενώ οι συσκευές τοποθετήθηκαν σε διαφορετικά σημεία του σώματος.

Το dataset περιλαμβάνει καταγραφές από 11 νεαρούς ενήλικες συμμετέχοντες (22-36 ετών), οι οποίοι εκτέλεσαν τόσο ADL δραστηριότητες (9) όσο και προσομοιωμένες πτώσεις (5), με τα δεδομένα να καταγράφονται σε μορφή χρονικών σειρών (.txt).

Το MobiFall παρέχει ετικέτες δραστηριοτήτων και πτώσεων, επιτρέποντας την εφαρμογή επιβλεπόμενων αλγορίθμων ταξινόμησης. Χρησιμοποιείται κυρίως σε προβλήματα ανίχνευσης πτώσης και αναγνώρισης δραστηριότητας [87]. Βασικός περιορισμός του είναι η απουσία συλλογής δεδομένων από πραγματικούς ηλικιωμένους συμμετέχοντες, περιορίζοντας τη γενίκευση των αποτελεσμάτων σε πληθυσμούς νεαρής ηλικίας.

Σχήμα 3.1: Προσομοίωση πλάγιας πτώσης σε μία ακολουθία πλαισίων [87]

KINECAL

Το KINECAL dataset διαμορφώνει ένα εξειδικευμένο σύνολο δεδομένων που έχει σχεδιαστεί για την ανάλυση των κινητικών λειτουργιών των ανθρώπων στο πλαίσιο της αξιολόγησης του κινδύνου πτώσεων και της ισορροπίας [88]. Σε αντίθεση με τα περισσότερα datasets αυτής της ειδικότητας, το KINECAL εστιάζει στην ποσοτικοποίηση της κινητικής δυσλειτουργίας μέσω της καταγραφής 11 κλινικών κινητικών λειτουργιών που χρησιμοποιούνται ευρέως στην αξιολόγηση της ευθραυστότητας και της στατικής ή δυναμικής ισορροπίας. Παράλληλα, παρέχονται δεδομένα σχετικά με το ιστορικό πτώσεων των ηλικιωμένων καθώς και άλλες κλινικές αξιολογήσεις (όπως μετρικές ταλάντωσης σώματος). Η συλλογή τους επιτεύχθηκε είτε μέσω ερωτηματολογίου είτε μέσω του αισθητήρα (κάμερα) “Microsoft Kinect V2”, προσφέροντας δεδομένα βάθους και σκελετικές αναπαραστάσεις της κίνησης των ατόμων, διευκολύνοντας τόσο την ποσοτική όσο και την ποιοτική ανάλυση των δεδομένων. Σε πρακτική σκοπιμότητα, τα δεδομένα μπορούν να εφαρμοστούν για την εκτίμηση του κινδύνου πτώσης των ηλικιωμένων.

Το KINECAL, περιλαμβάνει καταγραφές 90 συμμετεχόντων από ένα ευρύ ηλικιακό φάσμα (18-92) και κατανεμημένο ποσοστό φύλου, διασφαλίζοντας με αυτό τον τρόπο την αντιπροσωπευτικότητα των δεδομένων για διαφορετικές πληθυσμιακές ομάδες [88]. Τα δεδομένα παρέχονται σε ακατέργαστη μορφή (για δεδομένα βάθους και σκελετικά) (σε μορφή csv).

Συνεπώς, μία σημαντική ιδιαιτερότητα του KINECAL dataset αποτελεί ο εμπλουτισμός του με δεδομένα κινητικών λειτουργιών. Ακόμη, οι ηλικιωμένοι μπορούν να διακριθούν σε ομάδες μέσω του ιστορικού πτώσεων ή της ηλικιακής ομάδας που ανήκουν, ενώ τα ακατέργαστα δεδομένα μπορούν να

χρησιμοποιηθούν για τη κατασκευή νέων ειδικών χαρακτηριστικών. Έτσι, το dataset μπορεί να χρησιμοποιηθεί για ανάπτυξη προληπτικών μοντέλων πτώσεων.

3.2 Dataset Διπλωματικής Εργασίας

Το σύνολο δεδομένων που αξιοποιείται στην παρούσα εργασία προέρχεται από μια διατομεακή μελέτη των D. Lytras et al. [89], η οποία επικεντρώνεται στην καταγραφή πτώσεων αποκλειστικά σε ηλικιωμένα άτομα και στη διερεύνηση τόσο ενδογενών όσο και εξωγενών παραγόντων κινδύνου που σχετίζονται με αυτές. Η μελέτη πραγματοποιήθηκε στην Ελλάδα, πιο συγκεκριμένα στο Βόρειο τμήμα της, και η συλλογή των δεδομένων διαδραματίστηκε σε 5 διαφορετικές πόλεις της Κεντρικής Μακεδονίας, από μέλη Κέντρων Ανοικτής Προστασίας Ηλικιωμένων (ΚΑΠΗ). Παρέχονται δημογραφικά, κοινωνικά, κλινικά δεδομένα και δεδομένα τρόπου ζωής και καθώς ακόμη πληροφορίες λειτουργικής κατάστασης των συμμετεχόντων και πληροφορίες που αντιστοιχούν σε κάθε πτώση (όπως ο χώρος πτώσης). Σκοπός του συγκεκριμένου dataset είναι η συστηματική καταγραφή και πολυπαραγοντική ανάλυση των πτώσεων σε ηλικιωμένα άτομα, με στόχο την κατανόηση των παραγόντων κινδύνου που συμβάλλουν στην εμφάνισή τους και την υποστήριξη παρεμβάσεων πρόληψης.

Το dataset περιλαμβάνει 150 ηλικιωμένα άτομα, ηλικίας 65-80 ετών, τα οποία είχαν υποστεί τουλάχιστον μία πτώση εντός των τελευταίων 12 μηνών από την έρευνα. Κάθε πτώση που προκλήθηκε στον κάθε συμμετέχοντα εντός του χρονικού διαστήματος, καταγράφηκε ως ξεχωριστή εγγραφή στο dataset, με αποτέλεσμα το σύνολο των καταγεγραμμένων συμβάντων πτώσεων να αποτελεί 304 εγγραφές. Το dataset δεν είναι δημόσια διαθέσιμο, αλλά στο άρθρο που έχει δημοσιευθεί περιγράφονται τα δεδομένα που περιέχονται [89].

Η συνδυαστική δομή των πολυπαραγοντικών δεδομένων, καθιστά το dataset ιδιαίτερα πλούσιο καθώς επιτρέπει τη συσχέτιση κλινικών, λειτουργικών και περιβαλλοντικών παραγόντων με την συχνότητα και την σοβαρότητα των πτώσεων. Εκτός απ' αυτό, η χρήση λειτουργικών δεδομένων εμπλουτίζει με ρεαλιστικές μετρήσεις τα δεδομένα για ανάλυση κινδύνου των ατόμων. Επιπλέον, περιλαμβάνονται μόνο ηλικιωμένα άτομα, γεγονός που αποσκοπεί στην εστίαση και κατανόηση των παραγόντων που οδηγούν στις πτώσεις για τα άτομα της τρίτης ηλικίας. Από την άλλη, ως προς τα μειονεκτήματα του dataset, το πλήθος των εγγραφών δεν είναι ιδιαίτερα μεγάλο, ενώ ακόμη δεν περιλαμβάνεται μεταβλητή ετικέτας, καθώς όλες οι εγγραφές ανήκουν στην ίδια κατηγορία, αυτή των ατόμων που έχουν υποστεί τουλάχιστον μία πτώση, χωρίς κάποιου χαρακτηριστικού ένδειξης επικινδυνότητας.

3.2.1 Αιτιολόγηση Επιλογής του Dataset

Τα datasets που περιγράφηκαν στις ενότητες 3.1 και 3.2 είναι μόλις εκείνα που επιλέχθηκαν ως υποψήφια μεταξύ του συνόλου των datasets που εξετάστηκαν για την ανάπτυξη της εργασίας σχετικά με τις πτώσεις (βλέπε παράρτημα Α). Γενικότερα, διαφέρουν σημαντικά μεταξύ τους, ως προς τη φύση των δεδομένων, το επίπεδο ανάλυσης, και τον ερευνητικό τους προσανατολισμό.

Τα datasets MobiFall και SisFall είναι από τα πιο διαδεδομένα στον χώρο της ανίχνευσης πτώσεων και βασίζονται κυρίως σε δεδομένα αισθητήρων IMU, όπως επιταχυνσιόμετρα και γυροσκόπια, που συλλέγονται μέσω φορητών συσκευών. Ο σχεδιασμός τους εστιάζει κυρίως στην ανίχνευση της στιγμής που πραγματοποιείται η πτώση. Τις περισσότερες φορές τα δεδομένα προέρχονται από προσομοιωμένες πτώσεις ή από νεαρότερους πληθυσμούς, γεγονός που περιορίζει τη γενίκευση των ευρημάτων σε πραγματικούς ηλικιωμένους πληθυσμούς και σε σενάρια πρόληψης.

Αντίστοιχα, το eHomeSeniors dataset περιλαμβάνει δεδομένα περιβαλλοντικού περιεχομένου, αξιοποιώντας αισθητήρες εντός της οικίας για την παρακολούθηση προσομοίωσης καθημερινών δραστηριοτήτων και του τρόπου πτώσης των ατόμων. Ομοίως με τα υπόλοιπα σύνολα δεδομένων, οι συμμετέχοντες είναι νεαρής ηλικίας ή ηθοποιοί, πράγμα που οδηγεί σε μη ορθολογική γενίκευση των αποτελεσμάτων, όσον αφορά πραγματικούς ηλικιωμένους. Σαφώς, προσφέρει σημαντικά πλεονεκτήματα για συνεχή παρακολούθηση, ωστόσο το επίπεδο πληροφορίας είναι συχνά έμμεσο και λιγότερο συνδεδεμένο με κλινικές ή λειτουργικές μετρήσεις, περιορίζοντας τη δυνατότητα συσχέτισης των πτώσεων με συγκεκριμένους βιοϊατρικούς παράγοντες. Σε σύγκριση με αυτό, το KINECAL dataset, βασίζεται σε δεδομένα κίνησης από κάμερες βάθους εστιάζοντας στη λεπτομερή ανάλυση κίνησης κατά τη διάρκεια κλινικών/λειτουργικών δοκιμασιών ισορροπίας.

Σε αντίθεση με τα παραπάνω, το dataset που επιλέχθηκε περιλαμβάνει πολυπαραγοντικά γεγονότα, συμπεριλαμβανομένου λειτουργικές δοκιμασίες. Δεν περιορίζεται στην ανίχνευση στιγμιαίας πτώσης, όμως είναι ικανό να χρησιμοποιηθεί για την ανίχνευση κινδύνου πτώσης στα άτομα. Οι παρατηρήσεις του, αφορούν Έλληνες ηλικιωμένους ανθρώπους, και εστιάζει στη σημασία του ιστορικού των πτώσεων και όχι σε μεμονωμένες περιπτώσεις. Κατόπιν, η επιλογή του συγκεκριμένου dataset, αιτιολογείται από την συμβατότητά του με τον στόχο πρόβλεψης και πρόληψης πτώσεων, και όχι απλώς ανίχνευσής τους, όπως συμβαίνει στις περισσότερες περιπτώσεις συνόλων δεδομένων. Τέλος, ίσως μία από τις σημαντικότερες παραμέτρους επιλογής του συγκεκριμένου dataset, είναι ότι δεν έχει αξιοποιηθεί από άλλες μελέτες με χρήση τεχνικών ML, γεγονός που απεικονίζει την πρωτοτυπία της παρούσας εργασίας.

3.3 Περιγραφή Δεδομένων

Πρώτο βήμα σε κάθε ροή που περιλαμβάνει προσεγγίσεις ML είναι η ανάλυση του συνόλου δεδομένων. Σκοπός είναι η κατανόηση της δομής του περιεχομένου και των χαρακτηριστικών του επιλεγμένου συνόλου δεδομένων. Αυτό επηρεάζει καθοριστικά τις επιλογές προεπεξεργασίας, τη διαμόρφωση των χαρακτηριστικών και τελικώς την απόδοση και τη γενίκευση των μοντέλων.

Στις επόμενες υποενότητες παρουσιάζονται οι 39 μεταβλητές που αποτελούν το επιλεγμένο dataset και δομούνται σύμφωνα με την θεματική κατηγορία που αντιπροσωπεύουν. Η κατηγορίες χαρακτηριστικών είναι οι εξής:

- Δημογραφικά δεδομένα
- Κλινικά, ιατρικά και λειτουργικά δεδομένα
- Λειτουργικές δοκιμασίες
- Περιβαλλοντικά δεδομένα και Στοιχεία πτώσης/εων

Στην αρχική μορφή του dataset, όλες οι μεταβλητές είχαν αριθμητική τιμή, ανεξάρτητα από τον στατιστικό τύπου που αποτελούσαν. Τα δεδομένα που παρουσιάζονται αποτελούν την αρχική μορφή του dataset.

Πρέπει να διευκρινιστεί ότι, τα χαρακτηριστικά με τιμή στατιστικού τύπου ‘λογική-δυαδική’, λαμβάνουν απάντηση που αντιστοιχίζεται σε ‘όχι/ναι’, δηλαδή ‘0/1’.

3.3.1 Δημογραφικά Δεδομένα

Τα δημογραφικά δεδομένα, είναι είτε δυαδικής ή λογικής-δυαδικής μορφής είτε ποσοτικής.

Η μεταβλητή ‘Age’ επικροτείται εντός του εύρους τιμών 65-80, υποδηλώνοντας την ελάχιστη και την μέγιστη ηλικία εμφάνισης των συμμετεχόντων.

Η μεταβλητή ‘Gender’ λαμβάνει τις τιμές ‘Male/Female’ σε αριθμητική αναπαράσταση, δηλαδή ‘0/1’.

Πίνακας 3.1: Δημογραφικά Δεδομένα του dataset

Χαρακτηριστικό	**Περιγραφή**	**Τύπος τιμής**	**Στατιστικός τύπος**
Age	Ηλικία	Αριθμητική	Ποσοτική
Gender	Φύλλο	Αριθμητική	Ποιοτική/Δυαδική
Leavealone	Ζει μόνος	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Careprov	Φροντίζει άλλον ηλικιωμένο	Αριθμητική	Ποιοτική/Λογική-Δυαδική

3.3.2 Ιατρικά, Κλινικά και Λειτουργικά Δεδομένα

Τα κλινικά χαρακτηριστικά αποτελούν την μεγαλύτερη ομάδα δεδομένων του dataset. Όπως παρουσιάζονται στον 3.2, διακρίνονται κυρίως σε λογικές-δυαδικές τιμές, αλλά υπάρχουν και μεταβλητές που σχετίζονται με την καθημερινή φυσική δραστηριότητα των ατόμων.

Πίνακας 3.2: Κλινικά, Ιατρικά και Λειτουργικά δεδομένα του dataset

Χαρακτηριστικό	**Περιγραφή**	**Τύπος τιμής**	**Στατιστικός τύπος**
FALLSNUMBER	Αριθμός πτώσεων τελευταίων 12 μηνών	Αριθμητική	Ποσοτική
PILSSPERDAY	Αριθμός χαπιών την ημέρα	Αριθμητική	Ποσοτική
PHYCOTROPILLS	Χορήγηση ηρεμιστικών ψυχοτρόπων	Αριθμητική	Ποιοτική/Λογική-Δυαδική
EYESCHECK	Έλεγχος ματιών τα τελευταία δύο χρόνια	Αριθμητική	Ποιοτική/Λογική-Δυαδική
BLOODTEST	Ετήσιες εξετάσεις αίματος (ρουτίνας)	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Visionimpair	Προβλήματα όρασης	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Incontinence	Ακράτεια ούρων	Αριθμητική	Ποιοτική/Λογική-Δυαδική
BLOODHYPEPTEN	Διαταραχές πίεσης	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Osteoporosis	Οστεοπόρωση	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Pecemaker	Βηματοδότης	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Cardioprob	Καρδιακά προβλήματα	Αριθμητική	Ποιοτική/Λογική-Δυαδική

diabetes	Σακχαρώδης διαβήτης	Αριθμητική	Ποιοτική/Λογική-Δυαδική
OTHERPROBLEMS	Άλλα προβλήματα υγείας	Αριθμητική	Ποιοτική/Λογική-Δυαδική
v1	Ίλιγγος	Αριθμητική	Ποιοτική/Λογική-Δυαδική
Funcionalability	Βαθμός αυτοεξυπηρέτησης (κατά τον/την συμμετέχων/ουσα)	Αριθμητική	Ποιοτική/Τακτική
Balancedeficits	Αστάθεια στη βάδιση και στη στάση	Αριθμητική	Ποιοτική/Λογική-Δυαδική
kneehipproblems	Μυοσκελετικά προβλήματα που δυσχαιρένουν τη βάδιση	Αριθμητική	Ποιοτική/Λογική-Δυαδική
physactivhoursperweek	Χρόνος σωματικής δραστηριότητας την εβδομάδα	Αριθμητική	Ποιοτική/Τακτική
Physacttype	Είδος σωματικής δραστηριότητας	Αριθμητική	Ποιοτική

Η αριθμητική μεταβλητή ‘FALLSNUMBER’, αποτελεί την μεταβλητή που αποτυπώνει την ιστορικότητα των πτώσεων των τελευταίων 12 μηνών. Το μέγιστο πλήθος πτώσεων που εμφανίζεται στο σύνολο δεδομένων είναι 5 πτώσεις.

Το διατακτικό χαρακτηριστικό ‘Funcionalability’ σχετίζεται με τον βαθμό εξυπηρέτησης που θεωρεί ο ίδιος ο ηλικιωμένος ότι ανήκει, σύμφωνα με την κλίμακα Likert [90]. Στο dataset περιλαμβάνει τις τιμές:

- Πάρα πολύ
- Πολύ
- Μέτρια
- Λίγο
- Καθόλου

Η μεταβλητή ‘physactivhoursperweek’, παρέχει επίσης διάταξη στις κατηγορίες τιμών της. Αυτές είναι οι εξής:

- Καμία
- 1-2 ώρες
- 3-5 ώρες
- >5 ώρες

Επιπλέον, η ποιοτική μεταβλητή ‘Physacttype’ αποτελείται από τις εξής κατηγορίες τιμών:

- Κανένα
- Περπάτημα
- Χορός
- Ποδήλατο

- Γυμναστική
- Άλλο

3.3.3 Δεδομένα από Λειτουργικές Δοκιμασίες

Εμφανίζονται 7 γνωρίσματα από αξιολογήσεις λειτουργικών δοκιμασιών στο dataset, οι οποίες όλες αποτελούν ποσοτική στατιστική τιμή. Αυτές διακρίνονται στις εξής:

‘TUG’ (Timed Up and Go): μία διαδικασία εκτίμησης της κινητικότητας και της ισορροπίας των ηλικιωμένων, με μονάδα μέτρησης τον χρόνο σε δευτερόλεπτα [91]. Στο συγκεκριμένο dataset, ανώτατο όριο τιμής εμφανίζονται ρητά τα 15 δευτερόλεπτα. Κατά κανόνα, μικρότερη τιμή της μεταβλητής σημαίνει καλύτερη επίδοση του ηλικιωμένου.

‘FOURSBT’ (Four-Stage Balance Test): ένα εργαλείο εκτίμησης της στατικής ισορροπίας ενός ηλικιωμένου, με την αξιολόγηση τεσσάρων δοκιμασιών. Το εύρος κλίμακας είναι μεταξύ των τιμών 0-28. Μεγαλύτερη τιμή δείχνει καλύτερη ισορροπία [92].

‘CHAIRSTANDTEST’ (30-Second Chair Stand Test): μία δοκιμασία μέτρησης της μυϊκής δύναμης και αντοχής των κάτω άκρων, καταγράφοντας τον αριθμό επαναλήψεων καθίσματος-ανόρθωσης που μπορεί να εκτελέσει το άτομο μέσα σε 30 δευτερόλεπτα [93]. Γενικά, το εύρος τιμών κυμαίνεται από 0 έως 30+ επαναλήψεις.

‘BBS’ (Berg Balance Scale): μία κλίμακα αξιολόγησης της δυναμικής και στατικής ισορροπίας μέσω 14 λειτουργικών δοκιμασιών. Το συνολικό σκορ κυμαίνεται από 0 έως 56, με την μεγαλύτερη βαθμολογία να αντιστοιχεί σε καλύτερη σταθερότητα του ατόμου [94].

‘MMSE’ (Mini-Mental State Examination): ένα εργαλείο αξιολόγησης της γνωστικής λειτουργίας του ατόμου. Γενικώς, το σκορ κυμαίνεται μεταξύ των τιμών 0-30 [95], ωστόσο στο dataset υπάρχει ρητό κατώτατο όριο στην τιμή 23. Κάτω από την τιμή αυτή, σηματοδοτείται διάγνωση με άνοια.

‘CONFbal’ (Confidence in Balance): μία εκτίμηση υποκειμενικού επιπέδου εμπιστοσύνης του ατόμου στην ικανότητά του να διατηρεί την ισορροπία κατά την εκτέλεση καθημερινών δραστηριοτήτων (λήψη μέσω ερωτηματολογίου). Πρόκειται για μία Ελληνική διατύπωση του ερωτηματολογίου, το οποίο λαμβάνει τιμές 10-30 [96]. Χαμηλότερη τιμή του δείκτη, απεικονίζει υψηλότερη αυτοπεποίθηση του ατόμου στο να διαπράττει με σταθερότητα τις καθημερινές του δραστηριότητες.

‘ShortFES-I’ (Short Falls Efficacy Scale – International): μία αξιολόγηση του φόβου πτώσης κατά την εκτέλεση καθημερινών δραστηριοτήτων (λήψη μέσω ερωτηματολογίου). Αντίστοιχα με τοην προυγούμενη μεταβλητή, στο dataset χρησιμοποιήθηκε μία Ελληνική εκδοχή [96]. Λαμβάνει τιμές στο εύρος 7-28, όπου μία χαμηλή τιμή απεικονίζει μικρότερο φόβο πτώσης.

Πίνακας 3.3: Δεδομένα Λειτουργικών Δοκιμασιών του dataset

Χαρακτηριστικό	Περιγραφή	Τύπος τιμής	Στατιστικός τύπος
TUG	Αξιολόγηση δοκιμασίας Timed Up and Go	Αριθμητική/ Κινητής υποδιαστολής	Ποσοτική
FOURSBT	Αξιολόγηση δοκιμασίας 4-Stage Balance	Αριθμητική	Ποσοτική

CHAIRSTANDTEST	Αξιολόγηση δοκιμασίας 30-Second Chair Stand	Αριθμητική	Ποσοτική
BBS	Αξιολόγηση δοκιμασίας Berg Balance Scale	Αριθμητική	Ποσοτική
MMSE	Αξιολόγηση δοκιμασίας Mini-Mental State Exam	Αριθμητική	Ποσοτική
CONFbal	Αξιολόγηση ερωτηματολογίου CONFbal	Αριθμητική	Ποσοτική
ShortFES-I	Αξιολόγηση ερωτηματολογίου Short FES-I	Αριθμητική	Ποσοτική

3.3.4 Περιβαλλοντικά Δεδομένα και Δεδομένα Στοιχείων Πτώσης

Μία εξίσου σημαντική κατηγορία μεταβλητών αποτελεί στοιχεία σχετικά με τις πτώσεις. Σχετίζονται με τον χώρο, τον χρόνο και τον τρόπο που υπέστη πτώση το ηλικιωμένο άτομο, καθώς και πληροφορίες σχετικά με την αιτία και την σοβαρότητά της.

Πίνακας 3.4: Περιβαλλοντικά Δεδομένα του dataset

Χαρακτηριστικό	**Περιγραφή**	**Τύπος τιμής**	**Στατιστικός τύπος**
FallInjury	Τραυματισμός κατά την πτώση	Αριθμητική	Ποιοτική/Λογική-Δυαδική
HospDays	Ημέρες νοσηλείας μετά την πτώση	Αριθμητική	Ποσοτική
FallSite	Χώρος πτώσης	Αριθμητική	Ποιοτική/Δυαδική
HomeFallSite	Σημείο πτώσης σε κλειστό χώρο	Αριθμητική	Ποιοτική
OutFallSite	Σημείο πτώσης σε ανοικτό χώρο	Αριθμητική	Ποιοτική
FallCause	Τρόπος πτώσης	Αριθμητική	Ποιοτική
FootwearCause	Πτώση λόγω ακατάλληλου υποδήματος	Αριθμητική	Ποιοτική/Τακτική
VisionCause	Πτώση λόγω μη καλής ορατότητας	Αριθμητική	Ποιοτική/Τακτική
FallTime	Χρονική στιγμή πτώσης	Αριθμητική	Ποιοτική

Η δυαδική μεταβλητή ‘FallSite’ λαμβάνει τιμές ‘ανοικτός χώρος/κλειστός χώρος’ σε αριθμητική αναπαράσταση, που αντιστοιχίζεται σε ‘0/1’.

Στα περιβαλλοντικά χαρακτηριστικά διακρίνονται αρκετά που αποτελούν ποιοτικά με πολλές κατηγορίες τιμών. Αρχικά η μεταβλητή ‘HomeFallSite’ κατοχυρώνει την τοποθεσία που πραγματοποιήθηκε το συμβάν, εντός οικείας. Διακρίνεται στις τιμές:

- Κρεβατοκάμαρα
- Μπάνιο
- Κουζίνα
- Εσωτερικές Σκάλες
- Σαλόνι

- Άλλο

Η μεταβλητή 'OutFallSite' είναι αντίστοιχη της προηγούμενης, με την διαφορά ότι αφορά μόνο τοποθεσίες πτώσης σε εξωτερικό χώρο, όπως διακρίνεται:

- Αυλή/Μπαλκόνι
- Πεζοδρόμιο
- Διάσχιση Δρόμου
- Εξωτερική Σκάλα
- Πάρκο
- Άλλο

Η μεταβλητή 'FallCause' περιλαμβάνει τις κατηγορίες:

- Γλίστρα
- Σκόνταμμα
- Έγερση/Κάθισμα
- Ζάλη
- Προσπάθεια ανάκτησης αντικειμένου σε υπερυψωμένο σημείο
- Άλλο

Εν συνεχεία, η μεταβλητή 'FallTime' αποτελείται από τις κατηγορίες:

- Πρωί
- Εντός ημέρας
- Βράδυ (πριν τον ύπνο)
- Κατάκλιση (έγερση από τον ύπνο σε βραδινή ώρα)

Τέλος, οι μεταβλητές 'FootwearCause' και 'VisionCause' είναι δύο απόλυτα διατακτικές μεταβλητές, καθώς ακολουθούν την κλίμακα Likert, αντίστοιχα με την μεταβλητή 'FuncionalAbility' που εξετάστηκε παραπάνω.

3.4 Θέματα και Προκλήσεις του Dataset

Ένα από τα πιο βασικά ζητήματα που ανακύπτουν από τη χρήση του συγκεκριμένου συνόλου δεδομένων αφορά το μέγεθος του δείγματος. Μολονότι το dataset περιλαμβάνει πλούσια πολυπαραγοντική πληροφορία, ο αριθμός των διαθέσιμων παρατηρήσεων (304) παραμένει σχετικά περιορισμένος. Επιπλέον, η συνύπαρξη μεγάλου πλήθους χαρακτηριστικών (39), με σχετικά μικρό αριθμό δειγμάτων ενισχύει το φαινόμενο της "κατάρας της διαστασιμότητας", αυξάνοντας την πολυπλοκότητα του χώρου χαρακτηριστικών και καθιστώντας δυσχερή την αξιόπιστη εκμάθηση σταθερών προτύπων, αυξάνοντας τον κίνδυνο υπερπροσαρμογής. Ένα παράδειγμα αποτελεί η πιθανή ανισορροπία της κατανομής των κατηγοριών που αντιστοιχούν στα ποιοτικά χαρακτηριστικά και πρωτίστως σε εκείνα με υψηλό πλήθος μοναδικών τιμών. Ομοίως, λόγω του μικρού αριθμού δείγματος, είναι πιθανό να εμφανιστεί ανισορροπία στα παραγόμενα προφίλ κινδύνου πτώσης εξαιτίας της υποαντιπροσώπευσής τους, καθιστώντας αδύνατη την ερμηνεία και τη γενίκευσή τους. Έτσι, η εκπαίδευση αλγορίθμων συσταδοποίησης και επιβλεπόμενων μοντέλων απαιτεί τη

χρήση πρακτικών προσεκτικής προεπεξεργασίας (όπως η επιλογή γνωρισμάτων), διασταυρωμένης επικύρωσης και υπερπαραμετροποίησης.

Παράλληλα, η ποιότητα και η ετερογένεια των δεδομένων αποτελούν κρίσιμα ζητήματα. Η παρουσία ελλιπών τιμών σε επιμέρους μεταβλητές δημιουργεί ανάγκη για τεχνικές χειρισμού απουσών δεδομένων. Το σύνολο δεδομένων περιλαμβάνει ανομοιογενείς τύπους μεταβλητών (δυαδικές, κατηγορικές, διάταξης και συνεχείς), γεγονός που καθιστά απαραίτητες τις διαδικασίες κωδικοποίησης, κανονικοποίησης και τελικής ενοποίησής τους. Σχετικά με ορισμένες μεταβλητές, πρόκληση αποτελεί η υποκειμενικότητα των τιμών τους, καθώς σχετίζονται με ψυχολογικούς και κοινωνικούς παράγοντες των ηλικιωμένων (όπως φόβος πτώσης ή αντιλαμβανόμενη ισορροπία των συμμετεχόντων σε καθημερινές τους δραστηριότητες). Τέλος, η διακριτή απουσία διαχρονικών δεδομένων σίγουρα περιορίζει τη δυνατότητα ανάλυσης της εξέλιξης του κινδύνου στο χρόνο και την εξαγωγή αιτιοκρατικών συμπερασμάτων σύμφωνα με τη διαχρονική λειτουργική κατάσταση των ατόμων.

Όσον αφορά την δομή και την κατανομή των δεδομένων του dataset, παρουσιάζει ιδιαιτερότητες που επηρεάζουν τη μοντελοποίηση. Αρχικά, οι εγγραφές είναι οργανωμένες σε επίπεδο πτώσης, με αποτέλεσμα να αντιστοιχούν πολλαπλές παρατηρήσεις στο ίδιο άτομο. Ως συνέπεια, συμμετέχοντες με πολλαπλές εγγεγραμμένες πτώσεις, μπορεί να επηρεάσουν αρνητικά τα αποτελέσματα των μοντέλων οδηγώντας σε μεροληψία λόγω της υπεραντιπροσώπευσής τους στο dataset. Μάλιστα, όπως έχει ήδη αναφερθεί, εμφανίζονται μέχρι και 5 παρατηρήσεις του ίδιου ατόμου στο dataset.

Εξίσου σημαντικό ζήτημα είναι η ύπαρξη ετικέτας. Σε αντίθεση με τα περισσότερα δημόσια datasets, στην προκειμένη περίπτωση όλα τα δεδομένα αφορούν μόνο ένα είδος ατόμων, αυτών που έχουν υποστεί πτώση τουλάχιστον μία φορά εντός 12 μηνών. Όμως, απουσιάζει και οποιαδήποτε ετικέτα κινδύνου, που να προσδιορίζει τον βαθμό κινδύνου πτώσης των ατόμων. Το γεγονός αυτό, καθιστά αναγκαία τη χρήση μη επιβλεπόμενων μεθόδων για την ανάδειξη προφίλ κινδύνου, έπειτα την ερμηνεία τους και τη χρήση επιβλεπόμενων μοντέλων, διαπράττοντας την προγνωστική ικανότητά τους για την πρόληψη νέων ηλικιωμένων από τις πτώσεις.

Προχωρώντας στα ηθικά θέματα του συνόλου δεδομένων, οπωσδήποτε η διεξαγωγή ερευνητικών μελετών που αξιοποιούν δεδομένα υγείας και αφορούν ευάλωτους πληθυσμούς, όπως οι ηλικιωμένοι, καθιστά τα ηθικά ζητήματα κεντρικό άξονα της επιστημονικής δεοντολογίας. Η συλλογή, επεξεργασία και ανάλυση τέτοιων δεδομένων επέχει σοβαρές ευθύνες ως προς την προστασία της ιδιωτικότητας.

Στο dataset που χρησιμοποιείται για ανάλυση και επεξεργασία στη παρούσα εργασία, δεν υπάρχει τέτοιο μέλημα. Οι ηλικιωμένοι συμμετέχοντες, προτού ενταχθούν στην έρευνα συμπλήρωσαν την συγκατάθεσή τους για τη συλλογή και επεξεργασία των δεδομένων [89]. Παράλληλα τα δεδομένα, είναι ανωνυμοποιημένα, γεγονός που καλύπτει τον κίνδυνο ηθικών θεμάτων.

Δεν ισχύει το ίδιο όμως με την δημοσίευση των πρωτογενών δεδομένων του συνόλου. Συγκεκριμένα, λόγω του ότι το dataset δεν είναι δημοσίως διαθέσιμο, δεν επιτρέπεται η προβολή ακατέργαστων στοιχείων. Για το λόγο αυτό, η παρούσα εργασία δεν περιλαμβάνει παρουσίαση των ακατέργαστων δεδομένων, σεβόμενη τις αρχές της ιδιωτικότητας και τα πνευματικά δικαιώματα των κατόχων τους.

3.5 Μεθοδολογία Πρακτικού Μέρους

Η μεθοδολογία που υιοθετείται στο πρακτικό σκέλος της παρούσας εργασίας διαμορφώθηκε με άμεσο αντίκτυπο στη φύση των δεδομένων και στις προκλήσεις που αυτά παρουσιάζουν. Η μεθοδολογική

προσέγγιση σχεδιάστηκε ώστε να αντιμετωπίσει τα θέματα και τις προκλήσεις που αποκαλύφτηκαν στην προηγούμενη ενότητα.

Κεντρικός στόχος της εργασίας είναι η ανάπτυξη ενός συστήματος υβριδικής προσέγγισης μέσω ML, το οποίο επιτρέπει την αναγνώριση προτύπων και μοτίβων των ηλικιωμένων, την επακόλουθη ταξινόμησή τους σε ένα από τα προφίλ κινδύνου πτώσης που έχουν την μεγαλύτερη ομοιότητα, με τελικό σκοπό την υποστήριξη εξατομικευμένων παρεμβάσεων πρόληψης, και την παρουσίαση της ικανότητας του μοντέλου σε μία διαδραστική εφαρμογή. Η μεθοδολογία εστιάζει στην κατανόηση της υποκείμενης δομής των δεδομένων σε επίπεδο ατόμου και στη μετατροπή της πληροφορίας αυτής σε λειτουργικά αξιοποιήσιμη γνώση για την πρόληψη από μελλοντικές πτώσεις.

Για την επίτευξη του ανωτέρω στόχου, ακολουθήθηκε μία εκτεταμένη ροή εργασιών, η οποία οργανώνεται σε διακριτά αλλά αλληλοεξαρτώμενα στάδια, όπως απεικονίζονται στο σχήμα 3.2. Αρχικά, η επιλογή του κατάλληλου συνόλου δεδομένων αποτέλεσε το πρώτο στάδιο της εργασίας, καθώς από αυτή εξαρτώνται όλα τα παρακάτω στάδια. Το αμέσως επόμενο, αφορά την εκτενή προεπεξεργασία δεδομένων, συμπεριλαμβανομένου τεχνικών όπως: χειρισμός ελλιπών τιμών, ενοποίηση ετερογενών τύπων μεταβλητών, κωδικοποίηση κατηγορικών χαρακτηριστικών, κανονικοποίηση αριθμητικών μεταβλητών, επιλογή υποσυνόλου χαρακτηριστικών και άλλα. Στη συνέχεια, διερευνήθηκε η ανάλυση συσταδοποίησης για την ανάδειξη κοινών προτύπων στα δεδομένα και τη διερευνητική ομαδοποίησή τους βάσει των πολυπαραγοντικών χαρακτηριστικών τους. Η αξιολόγηση και η συγκριτική ανάλυση ήταν το επακόλουθο βήμα, για την επιλογή της καλύτερης τεχνικής συσταδοποίησης για τα δεδομένα. Οι συστάδες που δημιουργήθηκαν υποβλήθηκαν σε διαδικασία ετικετοποίησης (labeling), κατά την οποία αποδόθηκαν ονομασίες σε κάθε ομάδα βάση των επικρατέστερων χαρακτηριστικών τους.

Επόμενο βήμα της διαδικασίας, αποτελεί η εκπαίδευση SL μοντέλων ταξινόμησης, τα οποία αξιοποίησαν τις ετικέτες που προέκυψαν από το στάδιο της συσταδοποίησης, προκειμένου να μάθουν τα αντίστοιχα όρια απόφασης και να επιτυγχάνουν γενίκευση σε νέα, άγνωστα δεδομένα ηλικιωμένων. Η απόδοση των μοντέλων αξιολογήθηκε μέσω κατάλληλων μετρικών αξιολόγησης και ύστερα μιας συγκριτικής ανάλυσης επιλέχθηκε το μοντέλο που απέδιδε καλύτερα αποτελέσματα, λαμβάνοντας υπόψιν τους κινδύνους που ελλοχεύουν. Παράλληλα εξετάστηκε η ερμηνεία των προφίλ κινδύνου πτώσης, μέσω τεχνικών ερμηνευσιμότητας, αλλά όχι μόνο, ώστε τα αποτελέσματα να μην περιορίζονται απλώς σε αριθμητικούς δείκτες ακρίβειας αλλά να συνδέονται με κλινικά λογικές κατηγορίες. Τελευταίο στάδιο σχετικά με τη διαδικασία ερμηνείας των προφίλ κινδύνου, αποτέλεσε η δημιουργία κατάλληλων παρεμβάσεων πρόληψης νέων πτώσεων, σύμφωνα με τον κίνδυνο πτώσης κάθε ηλικιωμένου, αντλώντας πληροφορίες από πιστοποιημένους οργανισμούς στο πεδίο των πτώσεων. Τον κύκλο της ροής εργασιών ολοκληρώνει η ανάπτυξη μίας διαδραστικής εφαρμογής για την χρήση του τελικού μοντέλου από τους χρήστες για τους οποίους προορίζεται.

Σχήμα 3.2: Διάγραμμα ροής εργασιών του πειραματικού σκέλους

3.6 Επίλογος

Σε αυτό το κεφάλαιο παρουσιάστηκε με συστηματικό τρόπο το πλαίσιο δεδομένων και η μεθοδολογική προσέγγιση που θεμελιώνουν το πρακτικό σκέλος της εργασίας. Στην αρχή αναδείχθηκαν τα υποψήφια προς χρήση δημόσια διαθέσιμα σύνολα δεδομένων που έχουν χρησιμοποιηθεί στη βιβλιογραφία για τη μελέτη των πτώσεων μέσω της χρήσης ML. Ακολούθησε η παρουσίαση του επιλεγμένου dataset, της παρούσας εργασίας, και τεκμηριώθηκε η επιλογή του. Στη συνέχεια, αναλύθηκε η δομή και το περιεχόμενο των δεδομένων, διασπώντας σε επιμέρους υποενότητες το κείμενο, με βάση τη θεματική κατηγορία που ανήκει το κάθε χαρακτηριστικό. Παράλληλα, επισημάνθηκαν τα κύρια θέματα και οι προκλήσεις που αποτελούν το αρχικό επιλεγμένο dataset, όπως ο περιορισμένος αριθμός δειγμάτων. Σε συνάρτηση με τα χαρακτηριστικά, τα δεδομένα και τους περιορισμούς, διατυπώθηκε η κατανόηση του dataset, η οποία επέφερε ως αποτέλεσμα την μεθοδολογική προσέγγιση που περιγράφτηκε αμέσως μετά, η οποία σχετίζεται με την υιοθέτηση της υβριδικής ροής εργασιών του πρακτικού μέρους. Με αυτό τον τρόπο, το κεφάλαιο αυτό θέτει το τελικό υπόβαθρο για την ανάλυση του πειράματος και την εξαγωγή αποτελεσμάτων που εφαρμόζονται στα επόμενα κεφάλαια της εργασίας.

Κεφάλαιο 4ο: Πειράματα και Αξιολογήσεις

Το παρόν κεφάλαιο επικεντρώνεται στη πειραματική διερεύνηση και την εφαρμογή της λογικής αλληλουχίας του πειραματικού σχεδιασμού, που παρουσιάστηκε στο προηγούμενο κεφάλαιο. Αφού καθορίστηκε το πλαίσιο των δεδομένων, οι προκλήσεις τους και η υβριδική ροή εργασιών, ακολουθεί η ανάλυση του πειράματος προεπεξεργασίας και των τεχνικών ML, με στόχο την εμπειρική τεκμηρίωση της αποτελεσματικότητάς της.

Η ανάλυση του κεφαλαίου διαρθρώνεται σε διακριτά στάδια. Αρχικά, παρουσιάζεται το υπολογιστικό περιβάλλον και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση και εκτέλεση των πειραμάτων. Ακολούθως, περιγράφονται αναλυτικά οι τεχνικές προεπεξεργασίας των δεδομένων που εφαρμόστηκαν, καθώς και τα αποτελέσματά τους ως προς τη διαμόρφωση του τελικού συνόλου χαρακτηριστικών. Στη συνέχεια, εξετάζονται οι τεχνικές συσταδοποίησης που υλοποιήθηκαν, μέσω ποσοτικών και ποιοτικών αποτελεσμάτων (μέσω δεικτών) και τεκμηριώνεται η επιλογή της καταλληλότερης μεθόδου, ως προς τον στόχο της εργασίας. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση της διαδικασίας ταξινόμησης, όπου αναλύονται συστηματικά τα στάδια που απαιτήθηκαν για την ανάπτυξη, βελτιστοποίηση και αξιολόγηση των μοντέλων, η συγκριτική αποτίμηση της απόδοσής τους και η τελική επιλογή του βέλτιστου εκτιμητή, με βάση την κύρια μετρική αξιολόγησης που υιοθετείται.

4.1 Περιβάλλον Εργασίας

Η υλοποίηση των πειραμάτων πραγματοποιήθηκε στο περιβάλλον Python 3.12, αξιοποιώντας το Jupyter Notebook ως διαδραστική πλατφόρμα ανάπτυξης του κώδικα. Το Jupyter Notebook επιτρέπει την σταδιακή εκτέλεση των πειραμάτων και τη συστηματική καταγραφή των αποτελεσμάτων, διευκολύνοντας την αναπαραγωγιμότητα και την ερμηνευτική ανάλυση. Για την υλοποίηση των αλγορίθμων ML χρησιμοποιήθηκε κυρίως η βιβλιοθήκη scikit-learn, η οποία παρέχει ένα ολοκληρωμένο και αξιόπιστο σύνολο εργαλείων, για προεπεξεργασία δεδομένων, μη επιβλεπόμενη μάθηση, επιβλεπόμενη ταξινόμηση, αξιολόγηση μοντέλων και υπερπαραμετροποίηση. Άλλες κύριες βιβλιοθήκες της python που χρησιμοποιήθηκαν είναι οι:

- pandas, για την διαχείριση και την ανάλυση δεδομένων
- numpy, για μαθηματικούς υπολογισμούς
- scipy, για μαθηματικούς υπολογισμούς και ανάπτυξη αλγορίθμων ML
- matplotlib, για την οπτικοποίηση των δεδομένων
- seaborn, για τη στατιστική οπτικοποίηση των δεδομένων βασισμένη στη βιβλιοθήκη matplotlib, παρέχοντας μία πιο ‘υψηλού επιπέδου’ διεπαφή.

Η βιβλιοθήκη XGBoost (έκδοση 3.1.2) χρησιμοποιήθηκε επίσης, για την υλοποίηση του ομώνυμου αλγορίθμου, η οποία περιλαμβάνει ένα συμβατό API με εκείνο της scikit-learn, για την διευκόλυνση και χρήση των προτερημάτων της.

4.2 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί τη συστηματική προσέγγιση μετασχηματισμού του συνόλου δεδομένων, προκειμένου να μεταβεί σε κατάλληλη μορφή για τη χρήση των μοντέλων ML. Στόχος είναι η

μετατροπή των αρχικών δεδομένων για ανάλυση, λαμβάνοντας όμως υπόψη τις ιδιαιτερότητες και τους περιορισμούς που παρουσιάστηκαν στο προηγούμενο κεφάλαιο. Η προεπεξεργασία αντιμετωπίζεται ως κρίσιμο στάδιο που καθορίζει τόσο τη σταθερότητα των μοντέλων όσο και τη δυνατότητα ερμηνείας των αποτελεσμάτων.

Ένας από τους πιο σημαντικούς παράγοντες των ζητημάτων του αρχικού dataset, αποτελεί η δομή του. Όπως έχει ήδη αναφερθεί, η μορφή του βασίζεται στις πτώσεις, δηλαδή κάθε παρατήρηση του dataset αφορά μία πτώση. Έτσι για κάθε εγγραφή που αφορά τον ίδιο συμμετέχοντα, οι μοναδικές πληροφορίες που είναι διαφορετικές είναι αυτές που σχετίζονται με το περιβάλλον του συμβάντος πτώσης (βλέπε πίνακα 3.4). Για την αποφυγή μεροληψίας λοιπόν από τα μοντέλα ML, αναπτύχθηκαν δύο ροές προεπεξεργασίας δεδομένων, που αντανακλούν σε διαφορετικές οπτικές προσέγγισης του φαινομένου των πτώσεων. Αυτές είναι οι (α) επιπέδου-πτώσης που διατηρεί την αρχική αναπαράσταση των δεδομένων, επιτρέποντας την ανάλυση των χαρακτηριστικών και των συνθηκών που συνοδεύουν μεμονωμένα περιστατικά πτώσης. Από την άλλη η επιπέδου-ατόμου οπτική βασίζεται στη συνάθροιση των αρχικών εγγραφών, ώστε κάθε παρατήρηση να αντιστοιχεί σε ένα μοναδικό άτομο και να αποτυπώνει συνοπτικά το προφίλ του. Ωστόσο, ύστερα από έρευνα και πειραματισμούς επιλέχθηκε η δεύτερη προσέγγιση καθώς εφαρμόζεται καλύτερα στην μελέτη και αποκλείει προκλήσεις που παρουσιάζει η πρώτη μέθοδος. Παρακάτω αναλύονται συστηματικά τα στάδια της προεπεξεργασίας του συνόλου δεδομένων, διασπώμενα σε βήματα.

Διόρθωση εσφαλμένης τιμής της μεταβλητής ‘FootwearCause’

Πρώτο βήμα αποτελεί η διόρθωση ενός σφάλματος που υπήρχε στα δεδομένα. Η μεταβλητή ‘FootwearCause’ ενώ σε κανονικές συνθήκες εμφανίζει εύρος τιμών 0-5, ακολουθώντας την κλίμακα Likert όπως αναφέρθηκε στο κεφάλαιο 3, στα δεδομένα του dataset εμφανίστηκε σε 4 παρατηρήσεις με την τιμή 6. Οι τιμές αντικαταστάθηκαν από την τιμή 6 στην τιμή 5, ύστερα από επικοινωνία με τον διευθύνων άτομο του dataset.

Διαχείριση Ελλιπών Τιμών

Με χρήση κατάλληλου φίλτρου, παρατηρείται ότι δύο στήλες του dataset έχουν ελλιπές τιμές (‘HomeFallSite’, ‘OutFallSite’). Μελετώντας το άρθρο του dataset [89] αλλά και το ίδιο το dataset αντιλαμβάνεται κανείς ότι οι δύο στήλες είναι αμοιβαία αποκλειόμενες, δηλαδή, σε κάθε γραμμή όταν η μία στήλη έχει τιμή τότε η άλλη δεν έχει (λαμβάνει τιμή ‘NaN’) και αντιστρόφως. Πράγματι, εάν αθροιστούν οι τιμές του πλήθους ελλιπών τιμών της κάθε μεταβλητής τότε το αποτέλεσμα ισούται με το πλήθος των εγγραφών (68 + 236 = 304).

```
df[df.columns[df.isnull().any()]].isnull().sum()

HomeFallSite     68
OutFallSite     236
dtype: int64
```

Σχήμα 4.1: Εμφάνιση μεταβλητών με ελλιπείς τιμές

Σε συνδυασμό με την στήλη FallSite, που συγγενεύει εγγενώς με τις δύο προαναφερθείσες στήλες, δημιουργήθηκε μία νέα στήλη, η ‘FallSiteMerged’. Οι αρχικές στήλες που παρήγαγαν την καινούρια, δε θα χρησιμοποιηθούν περαιτέρω, εκτός κι αν χρειαστεί για περιπτώσεις κατανόησης των δεδομένων. Η τεχνική αυτή ονομάζεται «Μηχανική Γνωρισμάτων».

Κωδικοποίηση τιμών της νέας στήλης (‘FallSiteMerged’) σε ακέραια μορφή

Α νέα στήλη, περιλαμβάνει ένα συνδυασμό των αντίστοιχων τιμών κατηγοριών των στηλών ‘HomeFallSite’, ‘OutFallSite’. Σε αυτό το βήμα, οι τιμές της στήλης κωδικοποιήθηκαν από συμβολοσειρά σε ακέραια αριθμητική μορφή.

```
# Dictionary mapping
mapping_FallSiteMerged = {
    'Home_1': 0,
    'Home_2': 1,
    'Home_3': 2,
    'Home_4': 3,
    'Home_5': 4,
    'Home_6': 5,
    'Out_1': 6,
    'Out_2': 7,
    'Out_3': 8,
    'Out_4': 9,
    'Out_5': 10,
    'Out_6': 11
}

df['FallSiteMerged'] = df['FallSiteMerged'].map(mapping_FallSiteMerged)
```

Σχήμα 4.2: Κωδικοποίηση μεταβλητής "FallSiteMerged"

Διαχείριση εσφαλμένα αναθετημένων τύπων μεταβλητών

Ρίχνοντας μια ματιά στο dataset, παρατηρείται ότι όλες οι μεταβλητές είναι τύπου float64, (προφανώς εκτός από τη νέα μεταβλητή). Σύμφωνα όμως με το άρθρο του dataset και την περιγραφή των χαρακτηριστικών, που διατυπώθηκε στους πίνακες του 3ου κεφαλαίου, εύκολα διαπιστώνεται ότι οι μεταβλητές δεν είναι όλες στατιστικά float64, αλλά ποικίλουν. Επομένως οι περισσότεροι τύποι των μεταβλητών είναι εσφαλμένως αναθετημένοι και πρέπει να τροποποιηθούν όπως αρμόζει. Στην πραγματικότητα, στατιστικά, κάποιες από αυτές αντιστοιχούν σε κατηγορικές μεταβλητές και άλλες είτε σε ακέραιες είτε σε κινητή υποδιαστολή μεταβλητές.

Τροποποίηση/Αντιστοίχιση δυαδικών τιμών

Όσες μεταβλητές είναι δίτιμες, είναι δυαδικές. Από αυτές οι μεταβλητές ‘Gender’ και ‘FallSite’ είναι δυαδικές κατηγορικές μεταβλητές. Όλες οι υπόλοιπες είναι λογικές-δυαδικές μεταβλητές, δηλαδή οι τιμές τους αντιστοιχούν σε ‘Αληθής’ ή ’Ψευδής’. Στο σύνολο δεδομένων όλες οι δυαδικές μεταβλητές είναι ήδη κωδικοποιημένες σε ακέραια μορφή. Συγκεκριμένα, δίνονται με τιμές 1 για ‘True’ και 2 για ‘False’. Για λόγους συνέπειας με την επιστημονική κοινότητα για τις δυαδικές τιμές, μετατρέπονται σε 1 για ‘True’ (καμία μεταβολή) και 0 για ‘False’ (από την τιμή 2), ή εναλλακτικά πραγματοποιείται η ορθή αντιστοίχιση των δυαδικών κατηγορικών μεταβλητών, σε 1 και 0 αντίστοιχα.

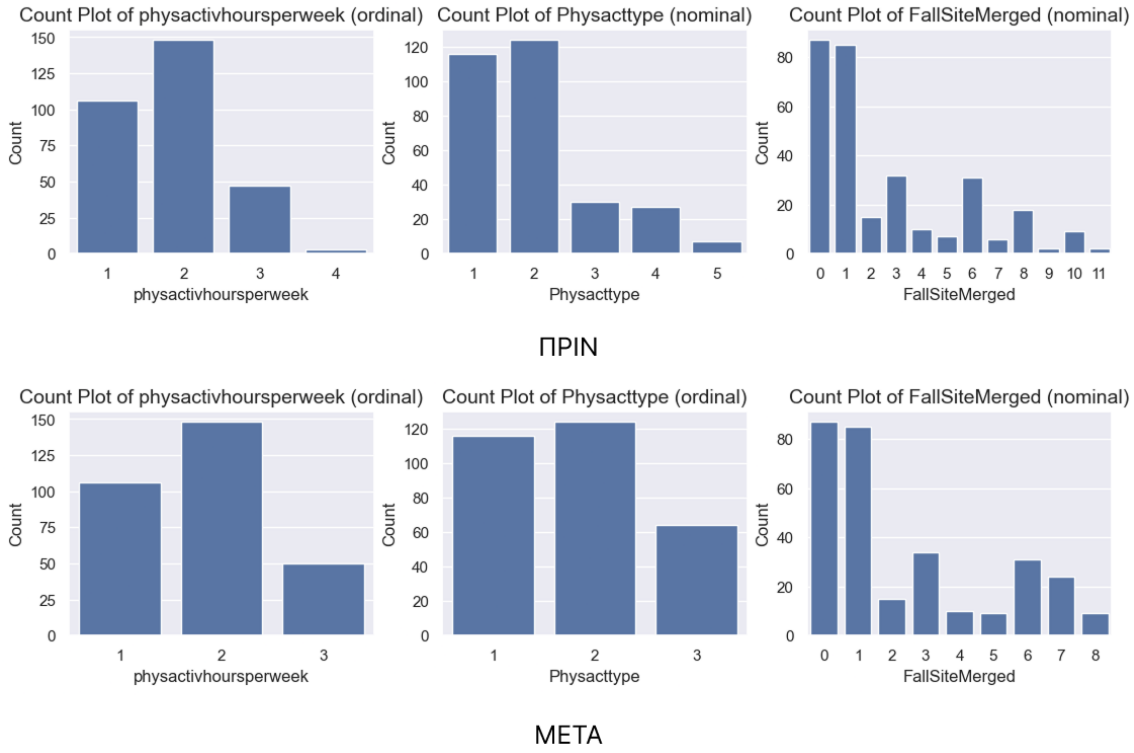
Δημιουργία στήλης αναγνωριστικού ατόμου

Για την ορθή και αμερόληπτη εκτέλεση των επόμενων πρακτικών βημάτων της εργασίας, κλήθηκε απαραίτητο να (διακριθούν) γνωστοποιηθούν οι πτώσεις που αντιστοιχούν στο ίδιο άτομο, καθώς απουσίαζε μία τέτοια στήλη στο αρχικό σύνολο δεδομένων. Όπως έχει αναφερθεί, η διάκριση των συμμετεχόντων εντός του dataset, συμβαίνει ως προς συγκεκριμένες στήλες, εκείνες που σχετίζονται με δεδομένα του συμβάντος πτώσης. Η ανάθεση τιμών της νέας στήλης, πραγματοποιήθηκε με βάση των υπόλοιπων στηλών του συνόλου δεδομένων (δεδομένα δημογραφικά, κλινικά, κινητικά). Έτσι, επιτεύχθηκε η δημιουργία της στήλης αναγνωριστικού, ονόματι ‘Id’, με αρχική τιμή 1, και τελική τιμή 150, αριθμό ίσο με το πλήθος των συμμετεχόντων που αναγράφεται στο άρθρο του συνόλου δεδομένων.

Περιγραφική ανάλυση ποιοτικών μεταβλητών και συγχώνευση

Ως συμπληρωματικό μέτρο αντιμετώπισης της υψηλής διαστασιμότητας αποτελεί η συγχώνευση των τιμών των ποιοτικών μεταβλητών, με γνώμονα τη διατήρηση της πληροφορίας και τη μικρότερη δυνατή απώλεια. η συγχώνευση κατηγοριών συνέβαλε στη μερική εξισορρόπηση των συχνοτήτων μεταξύ των τιμών κάθε μεταβλητής, καθώς εφαρμόστηκε κυρίως σε κατηγορίες με χαμηλή εκπροσώπηση, οι οποίες ενδέχεται να εισάγουν θόρυβο και αστάθεια στη διαδικασία εκμάθησης των μοντέλων.

Η προσέγγιση αυτή εφαρμόστηκε στα γνωρίσματα ‘physactivhoursperweek’, ‘Physacttype’ και ‘FallSiteMerged’, καθώς σε αυτές ήταν δυνατή η συγχώνευση των τιμών δεδομένων. Ιδιαίτερα για το γνώρισμα ‘Physacttype’, η συγχώνευση των κατηγοριών επέτρεψε τον μετασχηματισμό του από ονομαστικό σε τακτικό αριθμητικό τύπο. Κατ’ αυτόν τον τρόπο, το γνώρισμα δεν απαιτεί πλέον διάσπαση σε πολλαπλές δυαδικές μεταβλητές μέσω OHE, αποφεύγοντας την περαιτέρω αύξηση της διαστασιμότητας.



Σχήμα 4.3: Σχήμα απεικόνισης συγχώνευσης κατηγορικών τιμών (πριν και μετά)

Οι συγχωνεύσεις που εφαρμόστηκαν, τροποποίησαν τις έννοιες των τιμών της κάθε μεταβλητής.

- ‘physactivhoursperweek’: 1:’Καμία’, 2:’1-2 ώρες’, 3:’>3 ώρες’
- ‘Physacttype’: 1:’Κανένα’, 2:’Περπάτημα/Ελαφριά δραστηριότητα’, 3:’Έντονη δραστηριότητα’
- ‘FallSiteMerged’: 0:Κρεβατοκάμαρα’, 1:’Μπάνιο’, 2:’Κουζίνα’, 3:’Εσωτερικές/Εξωτερικές σκάλες’, 4:‘Σαλόνι’, 5:’Άλλο εξωτερικού/εσωτερικού χώρου’, 6: ‘Αυλή/Μπαλκόνι’, 7:‘Πεζοδρόμιο/Διάσχιση δρόμου’, 8:’Πάρκο’

Η μεγαλύτερη μετατροπή ορίζεται στην μεταβλητή ‘FallSiteMerged’, η οποία πλέον περιλαμβάνει τιμές που οι ίδιες σχετίζονται τόσο για εσωτερικό όσο και για εξωτερικό χώρο, που μέχρι πριν δεν συνέβαινε (τιμή 4 και 5).

Με την συγχώνευση ορισμένων ποιοτικών κατηγοριών, εξασφαλίζεται η αποφυγή της περαιτέρω αύξησης της διαστασιμότητας, συμβάλλοντας ουσιαστικά στον περιορισμό των επιπτώσεων της «κατάρας των πολλών διαστάσεων» και της καλύτερης γενίκευσης των μοντέλων.

Συνάθροιση

Η συνάθροιση αποτελεί μία ακόμη τεχνική μετασχηματισμού δεδομένων που βασίζεται στην εφαρμογή μαθηματικών συναρτήσεων με σκοπό τη συμπύκνωση και τη σύνοψη της πληροφορίας που περιέχεται σε πολλαπλές εγγραφές. Ο κύριος στόχος της συνάθροισης στην προκειμένη περίπτωση, είναι ο μετασχηματισμός της αναπαράστασης των δεδομένων από επίπεδο συμβάντος πτώσης σε επίπεδο ατόμου. Η μετάβαση αυτή υλοποιήθηκε μέσω της στήλης μοναδικού αναγνωριστικού ('Id') που δημιουργήθηκε σε προηγούμενο βήμα, η οποία αντιστοιχεί στον κάθε συμμετέχοντα. Με την ολοκλήρωση της διαδικασίας της συνάθροισης, κάθε παρατήρηση του νέου συνόλου δεδομένων πρόκειται να αναπαριστά ένα μοναδικό ηλικιωμένο άτομο. Η αναπαράσταση αυτή ενισχύει την ακεραιότητα του dataset και διασφαλίζει την ανεξαρτησία των παρατηρήσεων, προϋπόθεση κρίσιμη για την ορθή εφαρμογή των τεχνικών ML.

Για την μετάβαση στο σύνολο δεδομένων επιπέδου-ατόμου μέσω συνάθροισης, διακρίθηκαν δύο κατηγορίες μεταβλητών. Στην πρώτη κατηγορία εντάσσονται οι μεταβλητές που δεν σχετίζονται άμεσα με το επεισόδιο της πτώσης και εμφανίζουν σταθερές τιμές ανεξαρτήτως του αριθμού πτώσεων του ίδιου ατόμου. Σε αυτές περιλαμβάνονται δεδομένα που αφορούν την κατάσταση υγείας, την λειτουργικότητα και τα δημογραφικά χαρακτηριστικά των ηλικιωμένων. Ο συνολικός αριθμός των μεταβλητών αυτής της κατηγορίας ανέρχεται σε 31, εξαιρουμένης της μεταβλητής ('Id'), που χρησιμοποιείται αποκλειστικά για σκοπούς ομαδοποίησης της συνάθροισης. Οι τιμές τους διατηρήθηκαν αναλλοίωτες κατά τη διαδικασία της συνάθροισης, καθώς αποτυπώνουν σταθερά χαρακτηριστικά σε επίπεδο ατόμου.

```
untouchable_cols = sorted_df.iloc[:, 1:31].columns.values

untouchable_cols
```

Η δεύτερη κατηγορία περιλαμβάνει τις υπόλοιπες 7 μεταβλητές που σχετίζονται άμεσα με τα επιμέρους επεισόδια πτώσης. Οι μεταβλητές αυτές δεν διατηρήθηκαν αυτούσιες, αλλά αξιοποιήθηκαν για τη δημιουργία νέων χαρακτηριστικών μέσω κατάλληλων συναθροιστικών συναρτήσεων, με βασική αρχή την ελαχιστοποίηση της απώλειας πληροφορίας από τα αρχικά δεδομένα. Κάθε γνώρισμα εξετάστηκε ξεχωριστά και διερευνήθηκαν εναλλακτικοί τρόποι αποτύπωσης της πληροφορίας του. Η επιλογή των τελικών μεθόδων συνάθροισης βασίστηκε σε στατιστικά κριτήρια και στις ιδιότητες του κάθε ατομικού γνωρίσματος. Ακολούθως, παρουσιάζεται ένα απόσπασμα κώδικα που απεικονίζει τις συναθροίσεις για κάθε χαρακτηριστικό, ενώ στη συνέχεια εξηγείται η σημασία και τεκμηριώνεται η χρήση τους.

```
agg_numerical_dict = {

        'FallInjury': ['sum'],
        'HospDays': ['mean', lambda x: (x>0).sum(), 'min'],
        'FallCause': ['nunique'],
        'FootwearCause': [lambda x: ((x>=4).sum()) / len(x)],
        'VisionCause' : [lambda x: ((x>=4).sum()) / len(x)],
        'FallTime': ['nunique'],
        'FallSiteMerged': ['nunique']
}
# Aggregation grouped by 'Id' and store in a new DataFrame

df_numerical_agg = sorted_df.groupby(by=['Id']).agg( agg_numerical_dict )
```

Για τη μεταβλητή ‘FallInjury’ υπολογίστηκε το άθροισμα των τιμών, το οποίο εκφράζει τον συνολικό αριθμό τραυματισμών που σχετίζονται με πτώσεις ανά άτομο.

Για τη μεταβλητή ‘HospDays’ εφαρμόστηκαν τρεις συμπληρωματικές συναρτήσεις:

- ο μέσος όρος ημερών νοσηλείας,
- το πλήθος εισαγωγών σε νοσοκομείο,
- ο ελάχιστος αριθμός ημερών νοσηλείας

Οι μεταβλητές ‘FallCause’, ‘FallTime’ και ‘FallSiteMerged’ υποβλήθηκαν σε υπολογισμό του πλήθους διακριτών τιμών (ποικιλία διαφορετικών περιπτώσεων), ώστε να αποτυπωθεί η ποικιλία των αιτιών, των χρονικών πλαισίων και των χωρικών συνθηκών πτώσης αντίστοιχα. Ωστόσο, για να μη χαθεί πολύτιμη πληροφορία ως προς τα ακριβή δεδομένα των γνωρισμάτων αυτών, οι ίδιες οι μεταβλητές δεν εξαλείφθηκαν αμέσως, όπως έγινε με τις υπόλοιπες κατά τη διαδικασία συνάθροισης, αλλά μετασχηματίστηκαν σε κατάλληλη μορφή, η οποία αναφέρεται στο επόμενο βήμα προεπεξεργασίας. Έτσι διατηρείται τόσο η πληροφορία σχετικά με την ετερογένεια των συμβάντων όσο και οι ακριβείς πληροφορίες για κάθε ένα χαρακτηριστικό.

Για τις μεταβλητές ‘FootwearCause’ και ‘VisionCause’ υπολογίστηκε το ποσοστό των επεισοδίων στα οποία η τιμή υπερέβαινε προκαθορισμένο κατώφλι υψηλής επικινδυνότητας (τιμές ≥ 4), καθώς θεωρήθηκε ότι το πλήθος των πιο σημαντικών πτώσεων είναι η πιο κρίσιμη πληροφορία που πρέπει να διατηρηθεί για τις διατακτικές μεταβλητές.

Στη συνέχεια οι νέες μεταβλητές μετονομάστηκαν κατάλληλα βάση της μετατροπής ή παραγωγής τους. Η συνδυαστική χρήση αθροιστικών, μέσων και ελάχιστων τιμών, ποικιλομορφίας και αναλογικών δεικτών επέτρεψε την δημιουργία ενός νέου και πλούσιου συνόλου χαρακτηριστικών, που καθίσταται κατάλληλο για τη χρήση μεθόδων ML, πλέον σε επίπεδο ατόμου.

Τεχνική κωδικοποίησης (OHE) στα ονομαστικά δεδομένα

Τα γνωρίσματα ‘FallCause’, ‘FallTime’ και ‘FallSiteMerged’ εν τέλει μετασχηματίστηκαν μέσω της τεχνικής OHE, με αποτέλεσμα την επέκτασή τους από 3 αρχικές κατηγορικές μεταβλητές σε συνολικά 19 δυαδικά χαρακτηριστικά. Ο μετασχηματισμός αυτός επέτρεψε την αναπαράσταση κάθε κατηγορίας τους στον χώρο χαρακτηριστικών, διασφαλίζοντας ότι η κατηγορική πληροφορία ενσωματώνεται στο σύνολο δεδομένων με τρόπο κατάλληλο για αριθμητική επεξεργασία από τις τεχνικές ML.

Εφαρμογή συνάθροισης και κανονικοποίησης των One Hot κωδικοποιημένων γνωρισμάτων

Η εφαρμογή της τεχνικής OHE επέφερε ένα ουσιώδες πρόβλημα: ανεξαρτήτως του πλήθους των εμφανίσεων μιας κατηγορίας, η μέγιστη τιμή που αποδιδόταν στο αντίστοιχο χαρακτηριστικό ήταν ίση με 1.0. Κατά συνέπεια, περιπτώσεις με διαφορετική συχνότητα εμφάνισης αντιμετωπίζονταν ισοδύναμα. Ενδεικτικά, εάν ένα άτομο είχε υποστεί τρεις πτώσεις στον ίδιο χώρο (π.χ. στο μπάνιο), η τιμή του αντίστοιχου χαρακτηριστικού παρέμενε ίση με 1.0, όπως ακριβώς για ένα άλλο άτομο με μία μόνο πτώση στο ίδιο σημείο. Συγχρόνως, το γεγονός ότι τα OHE γνωρίσματα προέρχονταν από το αρχικό σύνολο δεδομένων, δηλαδή σε επίπεδο συμβάντος, όφειλαν να μετασχηματιστούν ώστε να αποτυπώνουν την πληροφορία σε επίπεδο ατόμου.

Η αντιμετώπιση του ζητήματος βασίστηκε στην εισαγωγή κανονικοποίησης των τιμών των χαρακτηριστικών σε συνδυασμό με την απόδοση βαρών στις μεταβλητές OHE, λαμβάνοντας υπόψη τόσο το πλήθος πτώσεων που αντιστοιχεί σε κάθε άτομο όσο και τη μέγιστη τιμή πτώσεων που παρατηρήθηκε στο

σύνολο δεδομένων. Συγκεκριμένα, για κάθε χαρακτηριστικό παρατήρησης υπολογίστηκε συντελεστής βάρους ως λόγος του αριθμού πτώσεων του ατόμου προς τον μέγιστο αριθμό πτώσεων που καταγράφηκε στο dataset, που είναι οι 5 πτώσεις. Εν ολίγοις, η σχέση αποτίμησης της τιμής βάρους προκύπτει από τον εξής μαθηματικό τύπο:

$$weight = \frac{πλήθος\ πτώσεων\ ατόμου}{μέγιστο\ πλήθος\ πτώσεων\ στο\ dataset} \quad (4.1)$$

Η στάθμιση αυτή επέτρεψε τη κλιμάκωση των δυαδικών τιμών OHE σε συνεχή διαστήματα εύρους [0.0-1.0], αποτυπώνοντας την πραγματική ένταση εμφάνισης κάθε κατηγορίας σε σχέση με το συνολικό ιστορικό πτώσεων του ατόμου.

Ως αποτέλεσμα, τα παραγόμενα χαρακτηριστικά απέκτησαν αναλογικό χαρακτήρα, καθώς οι τιμές τους αντανακλούν πλέον το ποσοστό συμμετοχής κάθε κατηγορίας στο σύνολο των καταγεγραμμένων πτώσεων για κάθε άτομο. Στην πράξη, η εμφάνιση κάθε κατηγορίας κατά μία φορά ισοδυναμεί με την τιμή 0.2. Επομένως, εάν για παράδειγμα σε ένα άτομο παρατηρηθούν δύο πτώσεις στην ίδια κατηγορία (π.χ. ίδιος τρόπος πτώσης), η τιμή του αντίστοιχου χαρακτηριστικού ανέρχεται σε 0.4, ως γινόμενο του βάρους επί τον αριθμό εμφανίσεων του χαρακτηριστικού (weight x value_frequency).

Με την παραπάνω μετατροπή, τα δυαδικά OHE γνωρίσματα (0,1) μετασχηματίστηκαν σε κανονικο-ποιημένες αναλογίες, απαντώντας στην ερώτηση "ποια είναι η συμμετοχή της αναλογίας συγκεκριμένων χωρικών, χρονικών ή αιτιολογικών παραγόντων πτώσης για έναν ηλικιωμένο".

Συνένωση γνωρισμάτων συνάθροισης σε ένα ενιαίο σύνολο δεδομένων

Κλείνοντας με το τμήμα της συνάθροισης, τα νέα γνωρίσματα που κατασκευάστηκαν (δεύτερο είδος μεταβλητών) συνενώθηκαν με τα μη τροποποιημένα γνωρίσματα (του πρώτου είδους μεταβλητών) διαμορφώνοντας το νέο σύνολο δεδομένων συνάθροισης, που είναι εστιασμένο στα άτομα που έχουν πέσει.

Επιλογή Υποσυνόλου Γνωρισμάτων

Λόγω του μεγάλου αριθμού χαρακτηριστικών σε αντίθεση με το περιορισμένο πλήθος δειγμάτων του συνόλου δεδομένων, υπάρχει μεγάλη πιθανότητα αρκετοί από τους αλγόριθμους ML να μην καταφέρουν να ανακαλύψουν πρότυπα που υπάρχουν μέσα στα δεδομένα λόγω της υψηλής αραιότητας των δεδομένων. Το φαινόμενο αυτό που είναι γνωστό ως "κατάρα των πολλών διαστάσεων" καθιστά κρίσιμη την εφαρμογή τεχνικών μείωσης της πολυπλοκότητας και επιλογής πληροφορίας. Η αντιμετώπιση των προβλημάτων αυτών συνέβει με χρήση μεθόδων προσέγγισης φίλτρου και τεχνικές γραμμικής άλγεβρας. Συγκεκριμένα, αξιοποιήθηκαν οι:

- εξάλειψη περιττών και άσχετων γνωρισμάτων για τη διαδικασία εκπαίδευσης των μοντέλων
- έλεγχος συσχέτισης ανά-δύο χαρακτηριστικών
- PCA

Βασικό όφελος των μεθόδων αυτών, είναι η εξάλειψη άσχετων για τον τομέα χρήσης των μοντέλων γνωρισμάτων και η μείωση του θορύβου. Ακόμη ένα όφελος είναι ότι μειώνουν την πολυπλοκότητα των μοντέλων, καθιστώντας τα πιο κατανοητά διότι περιέχουν λιγότερα χαρακτηριστικά και άρα τα δεδομένα είναι λιγότερο αραιά από ότι προηγουμένως. Επιπλέον, χρησιμοποιούνται για τον εντοπισμό πλεοναζόντων γνωρισμάτων με στόχο την μείωση της διαστασιμότητας. Επιπροσθέτως, λόγω της μείωσης του πλήθους των γνωρισμάτων, η οπτικοποίηση των δεδομένων μπορεί επίσης να γίνει πιο εύκολη. Μπορεί τα δεδομένα να μη μειώνονται σε 2 ή 3 διαστάσεις, ωστόσο ο συνολικός αριθμός συνδυασμών χαρακτηριστικών που

μπορούν να παρουσιασθούν καθίσταται μικρότερος, σύμφωνα με τη βιβλιογραφία [22]. Όλες οι προαναφερθείσες τεχνικές, χρησιμοποιήθηκαν σε αυτή τη φάση. Συγκεκριμένα, παρακάτω περιγράφονται οι μέθοδοι που χρησιμοποιήθηκαν και τα αποτελέσματά τους.

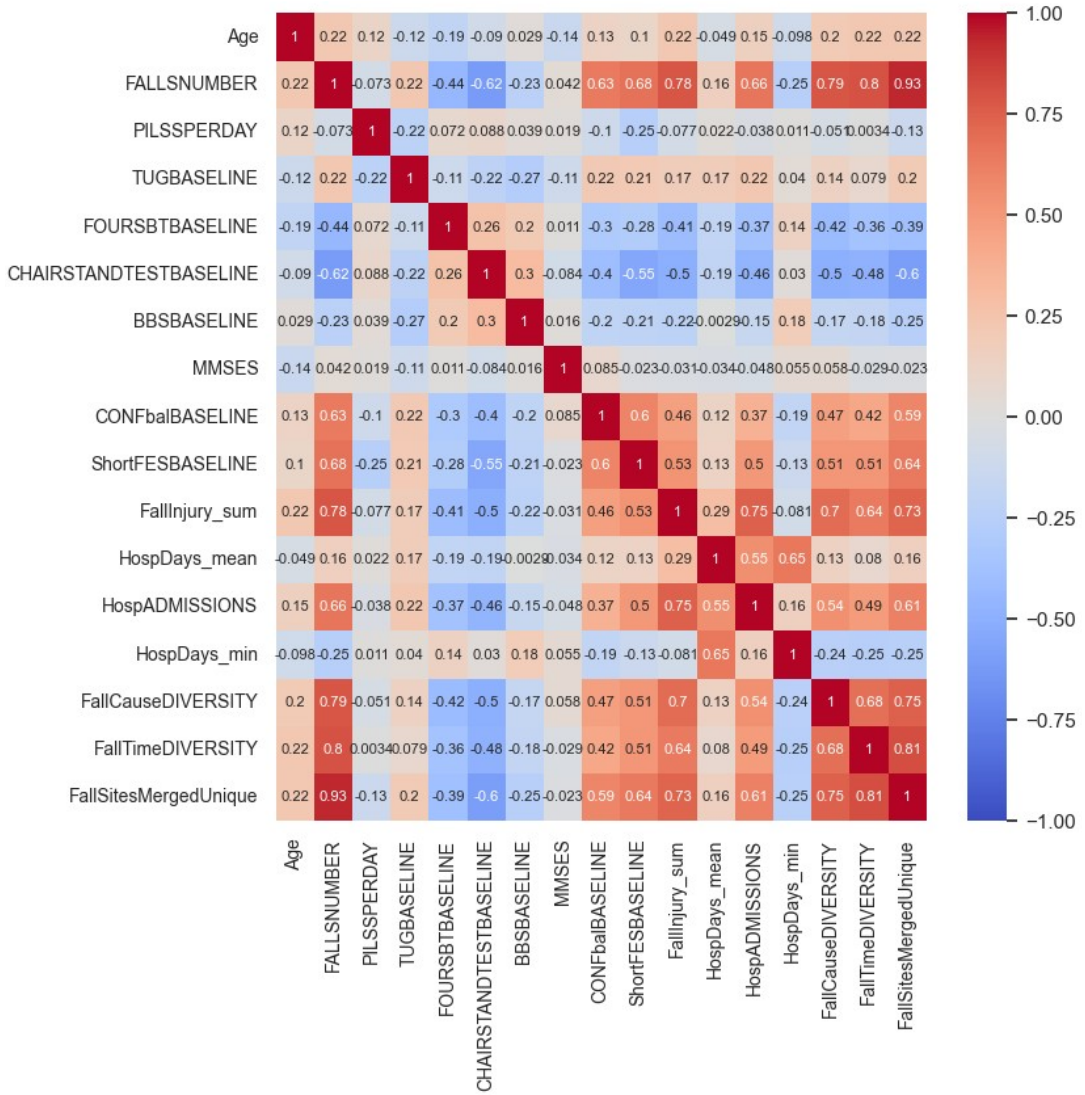
i. Εξάλειψη περιττών γνωρισμάτων

Η μοναδική στήλη που δε συμβάλει στην εκπαίδευση των μοντέλων ML, είναι η στήλη αναγνωριστικού (‘Id’) που δημιουργήθηκε σε ένα από τα προηγούμενα βήματα προεπεξεργασίας. Φυσικά, η εξαρχής χρήση της στήλης, δεν προοριζόταν για την εκπαίδευση στα μοντέλα, αλλά για την ενίσχυση της κατανόησης των αποτελεσμάτων των μοντέλων.

ii. Εξάλειψη γνωρισμάτων λόγω υψηλής συσχέτισης

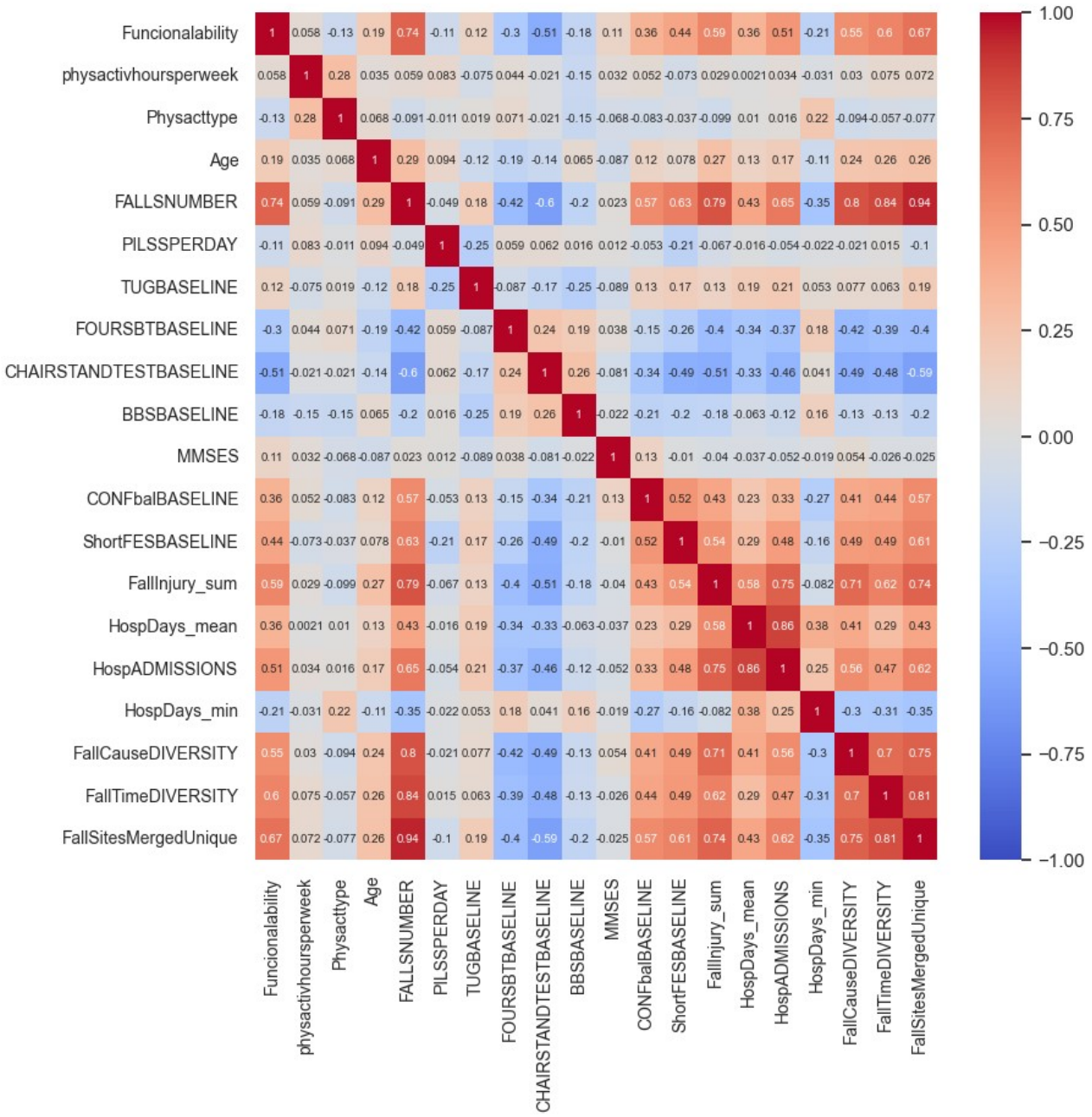
Η τεχνική αυτή αξιοποίησε τα στατιστικά εργαλεία Pearson (για σύγκριση ανά-δύο αριθμητικών μεταβλητών) και Spearman (για σύγκριση ανά-δύο αριθμητικών και τακτικών μεταβλητών). Η οπτικοποίηση των αποτελεσμάτων διακρίνεται με την αξιοποίηση σχημάτων heatmap. Σχετικά με τα heatmaps, πρέπει να διευκρινιστεί ότι είναι συμμετρικά ως προς την κύρια διαγώνιό τους, και επιπλέον τα κελιά που εμφανίζουν πιο έντονη απόχρωση του κόκκινου ή του μπλε, λαμβάνουν τιμή πιο κοντά στην τιμή 1 (όπως απεικονίζει η δεξιά μπάρα), και επομένως αντικατοπτρίζεται υψηλή συσχέτιση μεταξύ των διασταυρούμενων χαρακτηριστικών.

Συσχέτιση Pearson:



Σχήμα 4.4: Απεικόνιση heatmap συσχέτισης Pearson

Συσχέτιση Spearman:



Σχήμα 4.5: Απεικόνιση heatmap συσχέτισης Spearman

Η τιμές κατωφλίου που αποκαλύπτουν υψηλή συσχέτιση μεταξύ δύο μεταβλητών για μικρότερη και μεγαλύτερη τιμή αντίστοιχα για τις 2 τεχνικές συσχέτισης, είναι οι -0.8 και +0.8.

Με βάση τα heatmaps, τα ζεύγη στα οποία υπάρχει υψηλή συσχέτιση είναι τα εξής:

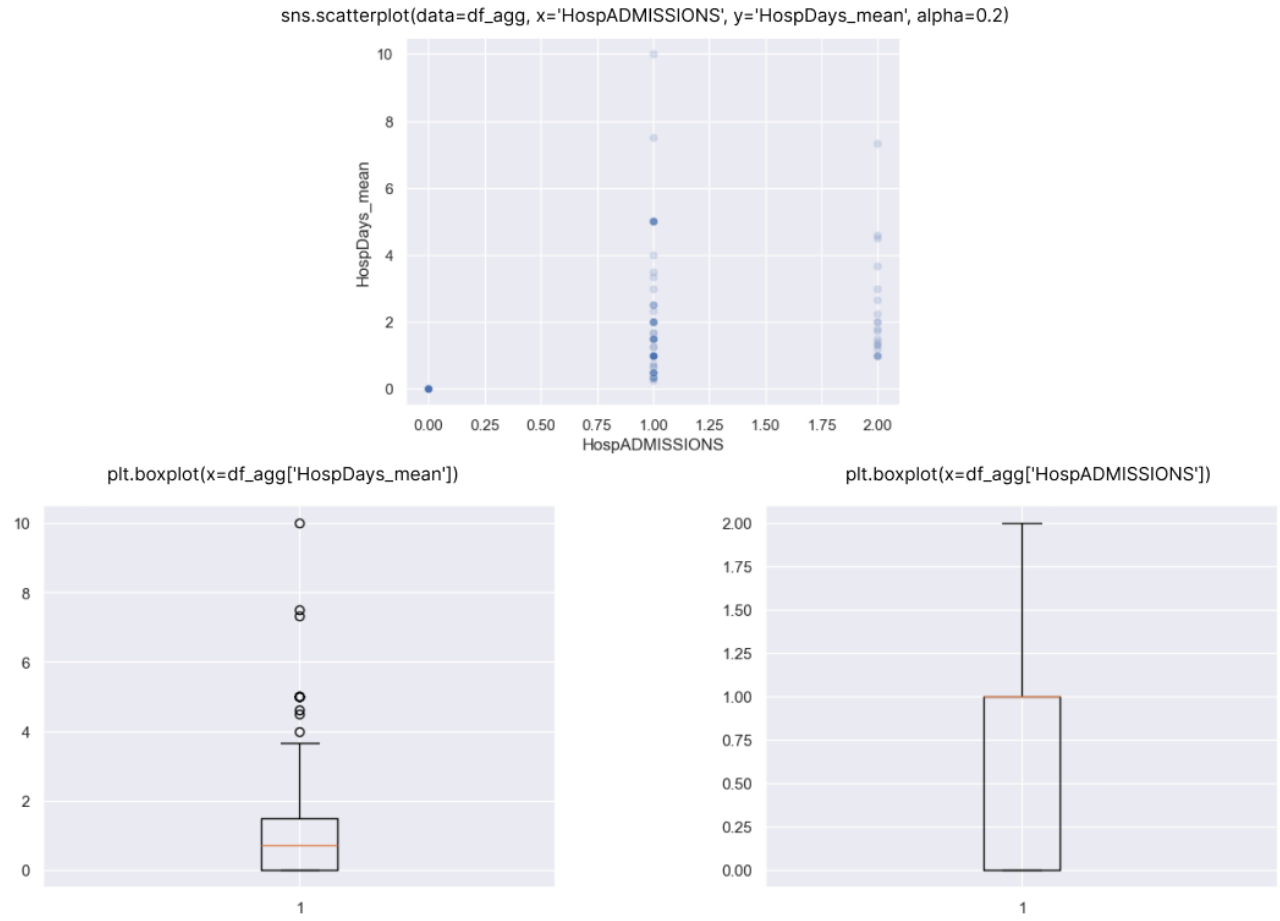
- ‘FALLSNUMBER - FallCauseDIVERSITY’
- ‘FALLSNUMBER - FallTimeDIVERSITY’
- ‘FALLSNUMBER - FallSitesMergedUnique’

- ‘FallTimeDIVERSITY - FallSitesMergedUnique’
- ‘HospADMISSIONS – HospDays_mean’

Τα γνωρίσματα που επιλέχθηκαν να αφαιρεθούν από το σύνολο δεδομένων είναι τα: ‘FallCauseDIVERSITY’, ‘FallTimeDIVERSITY’, ‘FallSitesMergedUnique’, ‘HospDays_mean’.

Η αιτία επιλογής του γνωρίσματος ‘FALLSNUMBER’, σχετίζεται με το γεγονός ότι παρέχει παρόμοια πληροφορία με τις υπόλοιπες τρεις με τις οποίες εμφανίζει υψηλή συσχέτιση, και επομένως προτιμάται να τις αντικαταστατήσει.

Όσον αφορά το ζεύγος ‘HospADMISSIONS – HospDays_mean’, έπειτα από διερευνητική ανάλυση κατανομών ως προς τις 2 μεταβλητές αποφασίστηκε η επιλογή της πρώτης, καθώς όπως φαίνεται στην παρακάτω εικόνα, η μεταβλητή ‘HospADMISSIONS’ είναι πιο «σταθερή» και λιγότερο επηρεαζόμενη από ακραίες τιμές που οδηγούν ιδιαίτερες περιπτώσεις πολλών ημερών νοσηλείας σε νοσοκομείο, σε σχέση με την συγκρινόμενη.



Σχήμα 4.6: Σύγκριση του ζεύγους γνωρισμάτων ‘HospADMISSIONS - HospDays_mean’ με χρήση scatterplot και boxplot

iii. PCA

Η τεχνική PCA συνήθως χρησιμοποιείται για την εξαγωγή χαρακτηριστικών. Όμως, υπάρχουν περιπτώσεις στις οποίες μπορεί να αξιοποιηθεί και για την επιλογή τους. Μία τέτοια περίπτωση αποτελεί η παρούσα

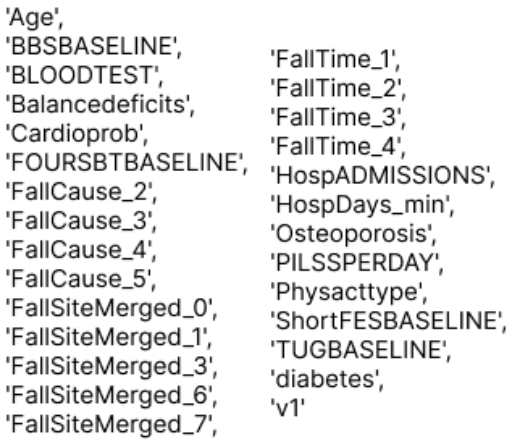
φάση της εργασίας. Το κύριο όφελος της χρήσης της σχετίζεται με την διατήρηση μόνο των γνωρισμάτων που συνεισφέρουν περισσότερο μεταξύ των δεδομένων, μειώνοντας το αυξημένο πλήθος διαστασιμότητας. Ακολουθήθηκε μία διαδικασία η οποία είναι ένας συνδυασμός των μεθόδων που αναφέρονται στις βιβλιογραφικές πηγές [97 - 99]. Πιο συγκεκριμένα:

1. Κατασκευή πλαισίου δεδομένων αποκλειστικά για χρήση της τεχνικής PCA
2. Προτυποποίηση όλων των χαρακτηριστικών, καθώς η τεχνική PCA είναι ιδιαίτερη ευαίσθητη σε δεδομένα με μη διαφορετικές κλίμακες
3. Υπολογισμός του πίνακα συνδιακύμανσης, ο οποίος περιγράφει τις γραμμικές συσχετίσεις μεταξύ των μεταβλητών
4. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης (εντοπισμός των Κύριων Συνιστωσών)
5. Επιλογή των Κύριων Συνιστωσών βάσει του 80% της αθροιστικής εξηγούμενης διακύμανσης. Επιλέχθηκε το συγκεκριμένο ποσοστό αυθαίρετα διότι δεν υπάρχει καθολικό όριο και η τιμή αυτή είναι ικανοποιητική
6. Υπολογισμός της συνεισφοράς κάθε χαρακτηριστικού στις Κύριες Συνιστώσες. Για κάθε αρχικό χαρακτηριστικό υπολογίζεται η συνολική του συμβολή (Cj) στο αποτέλεσμα της τεχνικής PCA. Για την δίκαιη αποτίμηση των συνεισφορών των γνωρισμάτων, χρησιμοποιήθηκαν ως βάρη τα αποτελέσματα από τα ποσοστά διακύμανσης για την κάθε Κύρια Συνιστώσα. Έτσι, η συνολική συμβολή αποδίδεται βάσει του γινομένου των βαρών επί των απόλυτων τιμών των συντελεστών, σύμφωνα με τον παρακάτω μαθηματικό τύπο:

$$Cj = \sum_{p=1}^{m} W\,p \cdot |Vpj|, \tag{4.2}$$

   όπου p={1, 2, …, m} αποτελεί τις Κύριες Συνιστώσες, Wp αφορά το «βάρος» για την p-ιοστή Κύρια Συνιστώσα, j είναι το j-ιοστό αρχικό χαρακτηριστικό και Vpj είναι το ιδιοδιάνυσμα της p-ιοστής Κύριας Συνιστώσας.
7. Ταξινόμηση των χαρακτηριστικών κατά φθίνουσα σειρά συνεισφοράς. Τα χαρακτηριστικά κατατάσσονται με βάση τη συνολική τους συμβολή στις κύριες συνιστώσες.
8. Επιλογή των χαρακτηριστικών με τη μεγαλύτερη συνεισφορά. Χρήση τιμής κατωφλίου η οποία διατηρεί τα χαρακτηριστικά που η συνολική συνεισφορά τους είναι ανώτερη της μέσης τιμής συνεισφοράς, σύμφωνα με το άρθρο των F. Rahmat et al. [97].

Με τον τρόπο αυτό, το αποτέλεσμα του πλήθους των μεταβλητών μειώθηκε από 54 σε μόλις 28, οι οποίες μάλιστα έχουν την μεγαλύτερη συνεισφορά με βάση την PCA τεχνική επιλογής γνωρισμάτων. Το ποσοστό κατά το οποίο μειώθηκαν τα γνωρίσματα φθάνει σχεδόν το 50% των αρχικών προτού χρησιμοποιηθεί η τεχνική. Τα γνωρίσματα που επικράτησαν απεικονίζονται στο παρακάτω σχήμα.



Σχήμα 4.7: Τελικά γνωρίσματα μέσω PCA επιλογής υποσυνόλου γνωρισμάτων

Στο τέλος, διατηρήθηκαν στο αρχικό πλαίσιο δεδομένων μόνο τα γνωρίσματα που κατέληξαν να αποτελούν τα πιο χρήσιμα.

Μετονομασία των επιλεγέντων γνωρισμάτων

Ύστερα από όλη τη διαδικασία της προεπεξεργασίας, χρήσιμο βήμα αποτελεί η μετονομασία των τελικών επιλεγμένων γνωρισμάτων, με ονομασίες που παρέχουν την ορθή αντιστοίχιση στην ιδιότητά τους.

```
columns_rename = {

  'Age': 'Age', # same
  'BBSBASELINE': 'BBS_Score',
  'BLOODTEST': 'has_BloodTest',
  'Balancedeficits': 'has_BalanceDeficitis',
  'Cardioprob': 'has_CardiovascularProblems',
  'FOURSBTBASELINE': 'FICSIT4_Score',
  'FallCause_2': 'FallCause_TrippedOnSomething',
  'FallCause_3': 'FallCause_UpDownSitting',
  'FallCause_4': 'FallCause_Dizzines',
  'FallCause_5': 'FallCause_ReachingHighObject',
  'FallSiteMerged_0': 'FallSiteMerged_Bedroom',
  'FallSiteMerged_1': 'FallSiteMerged_Bathroom',
  'FallSiteMerged_3': 'FallSiteMerged_Stairs',
  'FallSiteMerged_6': 'FallSiteMerged_YardBalcony',
  'FallSiteMerged_7': 'FallSiteMerged_PavementRoad',
  'FallTime_1': 'FallTime_InMorning',
  'FallTime_2': 'FallTime_DuringDay',
  'FallTime_3': 'FallTime_NightBeforeBed',
  'FallTime_4': 'FallTime_NightAtBed',
```

```
  'HospADMISSIONS': 'HospitalAdmissions',
  'HospDays_min': 'HospDays_min', # same
  'Osteoporosis': 'has_Osteoporosis',
  'PILSSPERDAY': 'PillsPerDay',
  'Physacttype': 'PhysicalActivity',
  'ShortFESBASELINE': 'ShortFESI_Score',
  'TUGBASELINE': 'TUG_Score',
  'diabetes': 'has_Diabetes',
  'v1': 'has_Vertigo'
}

df_agg_fs = df_agg_fs.rename(columns_rename, axis=1)
```

Προτυποποίηση μεταβλητών

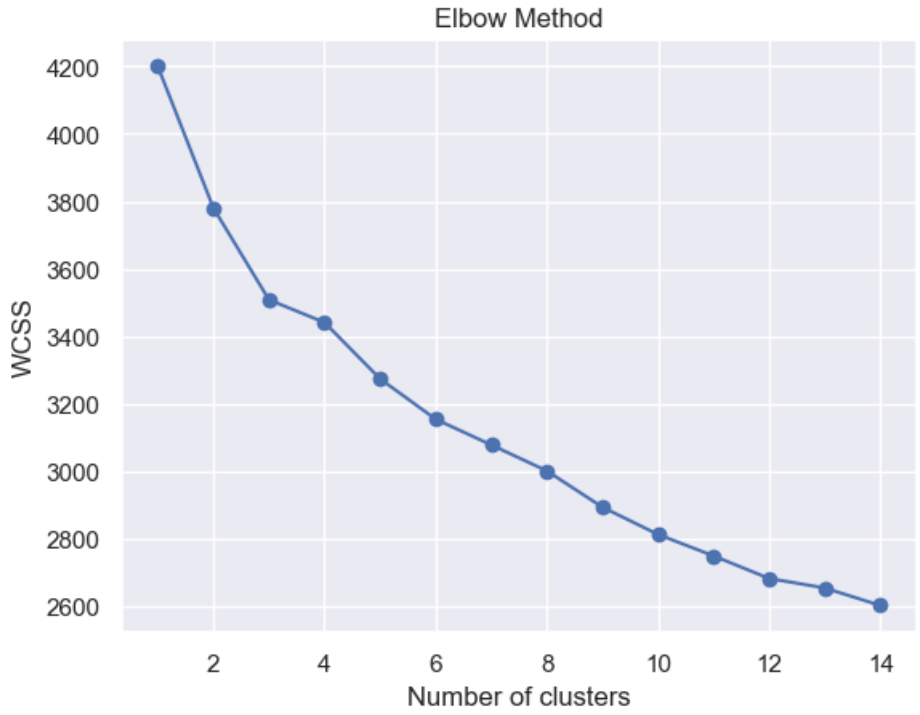
Τελευταίο βήμα της προεπεξεργασίας ήταν η υλοποίηση προτυποποίησης των δεδομένων, ώστε τα δεδομένα να είναι έτοιμα για χρήση εισόδου στις τεχνικές ML.

4.3 Συσταδοποίηση

Στο παρόν στάδιο της μελέτης, τα τελικά προεπεξεργασμένα δεδομένα αξιοποιούνται για την εφαρμογή τεχνικών UL, και ειδικότερα τεχνικών συσταδοποίησης, με σκοπό την ανακάλυψη υποκείμενων δομών και την ομαδοποίηση ατόμων με παρόμοια χαρακτηριστικά για τη δημιουργία προφίλ κινδύνου πτώσης ηλικιωμένων ατόμων, τα οποία έχουν ήδη υποστεί πτώση. Οι αλγόριθμοι που εφαρμόζονται είναι αυτοί που αναλύθηκαν στο θεωρητικό υπόβαθρο: K-Means, HCA και DBSCAN. Ο καθένας, ενσωματώνει διαφορετικές υποθέσεις και δομές γύρω από τα δεδομένα, επιτρέποντας τη συγκριτική διερεύνηση της καταλληλότητάς τους στο συγκεκριμένο πρόβλημα. Η αξιολόγηση των αποτελεσμάτων πραγματοποιείται συνδυαστικά, μέσω ποσοτικών μετρικών ποιότητας συστάδων και ποιοτικών κριτηρίων.

4.3.1 K-Means

Αρχικά, για την εκπαίδευση του μοντέλου K-Means εξετάστηκε ένα εύρος τιμών του αριθμού συστάδων k, προκειμένου να διερευνηθεί η επίδραση της παραμέτρου στη δομή των αποτελεσμάτων. Επιλέχθηκε να εξεταστούν 15 διαφορετικά μοντέλα, καθένα με διαφορετική παράμετρο k. Η μετρική απόστασης που χρησιμοποιήθηκε είναι η Ευκλείδεια απόσταση, καθώς μόνο αυτή υποστηρίζεται από το scikit-learn για τον συγκεκριμένο αλγόριθμο. Επιπλέον, χρησιμοποιήθηκε η μέθοδος αρχικοποίησης κέντρων των συστάδων ‘kmeans++’ και ένας μοναδικός αριθμός παραγωγής τυχαιότητας (παράμετρος ‘random_state’) των μοντέλων για λόγους αναπαραγωγικότητας.



Σχήμα 4.8: Αναπαράσταση Elbow method για αξιολόγηση του αλγορίθμου K-Means

Το αποτέλεσμα του 'Elbow method' δεν είναι ιδιαίτερα ικανοποιητικό καθώς δεν υπάρχει καλά ορισμένη καμπύλη η οποία να διαμορφώνει 'αγκώνα' μεταξύ των k πλήθος συστάδων, αλλά υπάρχει σχετικά ομαλή μετάβαση από το ένα k στο άλλο. Όπως φαίνεται στο σχήμα 4.8, η μεγαλύτερη μείωση του αθροίσματος τετραγωνικού σφάλματος μεταξύ των συστάδων (WCSS) εμφανίζεται μεταξύ των συστάδων 1-3, και λιγότερη μεταξύ των 4-5. Η καμπύλη γίνεται σχεδόν γραμμική έπειτα της τιμής k=5 γεγονός που σημαίνει ότι κάθε συστάδα προσφέρει μικρή μείωση του WCSS από εκεί και πέρα, άρα δεν παρέχει επιπλέον χρήσιμη πληροφορία για τη δομή των δεδομένων. Επομένως, σύμφωνα με την μέθοδο αυτή, το βέλτιστο μοντέλο μπορεί να βρεθεί μεταξύ των συστάδων 3-5.

Οι μετρικές Silhouette και Davies-Bouldin scores μπορούν να συμβάλουν στην εύρεση του καλύτερου πλήθους συστάδων για τα δεδομένα της εργασίας. Σύμφωνα με τον πίνακα 4.1 οι συστάδες δεν είναι έντονα διαχωρισμένες και πιθανώς υπάρχει επικάλυψη μεταξύ των ομάδων λόγω του μικρού μεγέθους των δεικτών. Ωστόσο, υπάρχει ασθενής δομή συσταδοποίησης. Η μετρική Silhouette παρουσιάζει μέγιστη τιμή για k=2, ενώ μερικώς ικανοποιητικές είναι οι τιμές για k=3, 5 και 6. Από την άλλη, η μετρική Davies-Bouldin παρουσιάζει μείωση των τιμών καθώς αυξάνονται οι συστάδες, γεγονός φυσιολογικό, ενώ καλές τιμές εμφανίζονται για πλήθος συστάδων k=2, 3 και 6.

Πίνακας 4.1: Μετρικές αξιολόγησης K-Means

k	Silhouette Score	Davies-Bouldin Score
2	0.1146	2.9200
3	0.0639	2.7985

4	0.0537	3.1529
5	0.0700	2.7773
6	0.0679	2.4659
7	0.0550	2.6151
8	0.0534	2.5719
9	0.0593	2.4599
10	0.0605	2.3296
11	0.0623	2.2232
12	0.0649	2.2155
13	0.0606	2.1671
14	0.0622	2.0801
15	0.0627	2.0193

Κατά συνέπεια, τα αποτελέσματα των μετρικών δε μπορούν να χρησιμοποιηθούν ατομικά για την εξαγωγή της βέλτιστης τιμής k, παρά μόνο συνδυαστικά, διότι η δομή συσταδοποίησης είναι ιδιαίτερα ασθενής. Λόγω του περιορισμένου πλήθους δειγμάτων, πιθανώς το πλήθος των συστάδων να πρέπει να παραμείνει μικρό. Ωστόσο, για την πλήρη κατανόηση της εικόνας των συστάδων, η αξιολόγησή τους θα πρέπει να γίνει με την προσθήκη μίας ακόμη εκτίμησης, η οποία αφορά την ποιοτική αξιολόγησή τους, σχετικά με το πλήθος των δειγμάτων τους. Όπως απεικονίζει ο πίνακας 4.2, το πλήθος συστάδων για τα οποία υπάρχει επαρκής αριθμός δειγμάτων αντιπροσώπευσης όλων των συστάδων είναι για k=2 και k=3. Το επαρκές πλήθος δειγμάτων των συστάδων αποτελεί σημαντικό κριτήριο επιλογής του k, διότι όσο περισσότερα είναι τα δεδομένα σε μία συστάδα, τόσο πιο γενικεύσιμα και αξιόπιστα θα είναι τα αποτελέσματα που θα εξαχθούν από τους αλγορίθμους επιβλεπόμενης μάθησης.

Πίνακας 4.2: Ποιοτική εκτίμηση πλήθους συστάδων αλγορίθμου K-Means

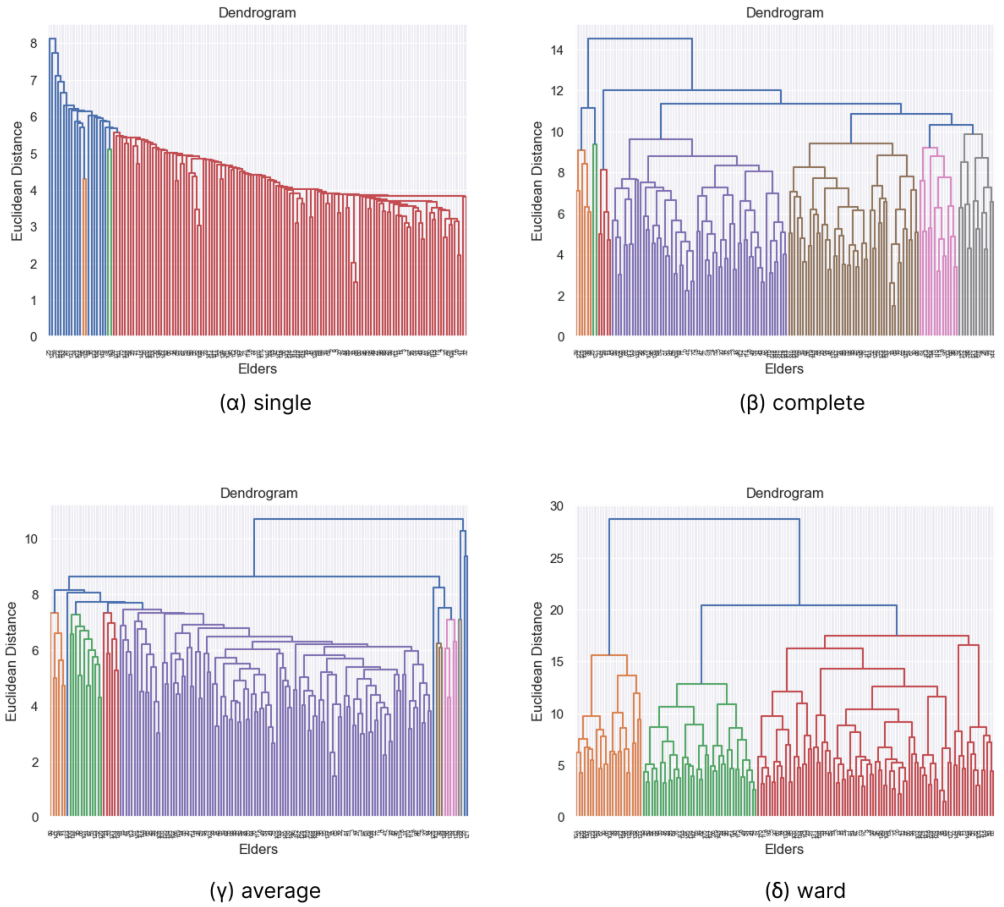
k	Κατανομή Δειγμάτων ανά συστάδα
2	0: 92, 1: 58
3	0: 72, 1: 47, 2: 31
4	0: 56, 1: 53, 2: 22, 3: 19
5	4: 58, 1: 38, 2: 28, 3: 16, 0: 10

6	1: 34, 4: 60, 0: 10, 2: 25, 3: 17, 5: 4

Επομένως, οι καταλληλότερες τιμές του k, σύμφωνα τόσο με τους ποσοτικούς δείκτες αξιολόγησης όσο και με τους ποιοτικούς, είναι μεταξύ των τιμών k=2 και k=3.

4.3.2 Ιεραρχική Συσταδοποίηση

Η Ιεραρχική Συσταδοποίηση, σε αντίθεση με τους υπόλοιπους αλγορίθμους, υλοποιήθηκε μέσω της βιβλιοθήκης scipy.cluster. Εξετάστηκαν πολλαπλές μέθοδοι σύνδεσης όπως οι single, complete, average και ward, προκειμένου να αξιολογηθεί η επίδραση του τρόπου συγχώνευσης των συστάδων στη δομή των αποτελεσμάτων. Σύμφωνα με την παρακάτω εικόνα, η μέθοδος (δ) ward παρήγαγε τις πιο διακριτές συστάδες, γεγονός που είναι συμβατό με τον στόχο της ελαχιστοποίησης της ενδοσυσταδικής διασποράς. Με βάση το δενδρόγραμμα, οι καλύτερες επιλογές για επίπεδα αποκοπής του, αποτελούν οι τιμές απόστασης 22 και 20, που οδηγούν αντίστοιχα σε k=2 και k=3 συστάδες.



Σχήμα 4.9: Αξιολόγηση μεθόδων 'linkage' για τον αλγόριθμο HCA, μέσω δενδρογραμμάτων

Για την ποσοτική και ποιοτική αποτίμηση της ποιότητας των συστάδων που προέκυψαν από τη μέθοδο ward, υπολογίστηκαν οι δείκτες Silhouette και Davies–Bouldin για τις διαμορφώσεις με k=2 και k=3 πλήθος συστάδων, όπως απεικονίζονται στους παρακάτω δύο πίνακες.

Πίνακας 4.3: Μετρικές αξιολόγησης HCA

k	Silhouette Score	Davies-Bouldin Score
2	0.1856	2.304
3	0.0495	2.9212

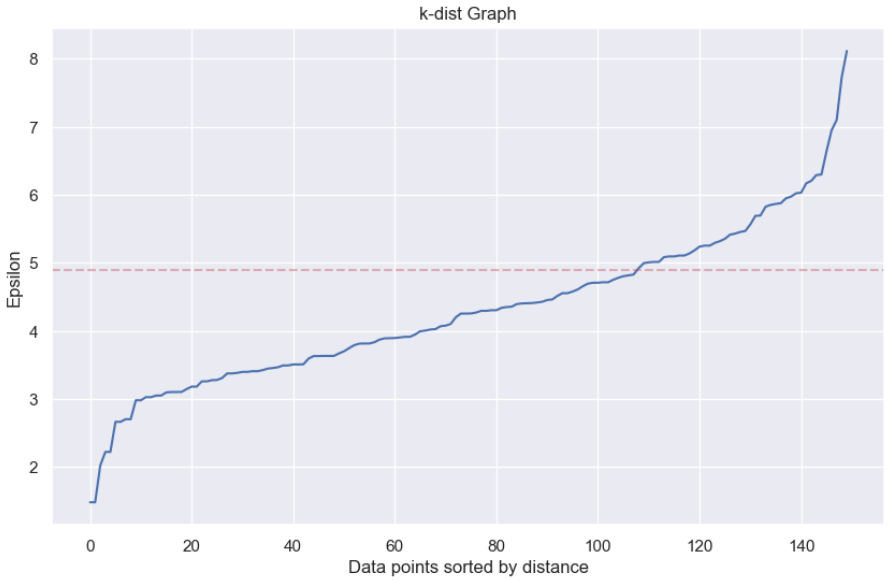
Πίνακας 4.4: Ποιοτική εκτίμηση πλήθους συστάδων αλγορίθμου HCA

k	Κατανομή Δειγμάτων ανά συστάδα
2	2: 126, 1: 24
3	0: 85, 1: 41, 2: 24

Στη σύγκριση μεταξύ των δύο διαφορετικών k στην περίπτωση του HCA, παρουσιάζεται επίσης χαμηλή διαχωρισιμότητα μεταξύ των συστάδων σε κάθε περίσταση, όμως τα αποτελέσματα είναι αρκετά καλύτερα σε σχέση με τον K-Means για k=2, καταφθάνοντας μέχρι τα 18.5% Silhouette score. Όμως σε αυτή τη περίπτωση τα δείγματα είναι λιγότερο ισορροπημένα, έχοντας και για τις δύο περιπτώσεις k για την συστάδα μειονότητας 24 δείγματα αντιπροσώπευσης της συστάδας τους. Αυτό δυσχεραίνει τη γενίκευση για τις τεχνικές αλγορίθμων SL.

4.3.3 DBSCAN

Στον αλγόριθμο DBSCAN πρώτο μέλημα αποτελεί η εύρεση της παραμέτρου ‘eps’ με τη βοήθεια του γραφήματος k-dist. Στο παρακάτω γράφημα, στον οριζόντιο άξονα απεικονίζονται τα σημεία του dataset ταξινομημένα ως προς την απόστασή τους από τον k-οστό πλησιέστερο γείτονά τους (συγκεκριμένα για k=3), ενώ στον κατακόρυφο άξονα η αντίστοιχη απόσταση (που υποδηλώνει την ακτίνα γειτνίασης ‘eps’).



Σχήμα 4.10: Γράφημα k-dist του αλγορίθμου DBSCAN

Παρατηρείται ομαλή αλλά αργή αύξηση των αποστάσεων μέχρι περίπου την τιμή eps=5, ενώ στη συνέχεια υπάρχει απότομη κλίση και αύξηση των τιμών της καμπύλης. Έτσι, ως κατάλληλη τιμή για την παράμετρο ‘eps’ ορίστηκε η τιμή 4.9, όπως φαίνεται από την διακεκομμένη γραμμή επάνω στο γράφημα. Το ότι δεν υπάρχει ιδιαίτερα απότομη κλίση κατά το μεγαλύτερο τμήμα της καμπύλης υποδηλώνει ότι υπάρχουν δεδομένα τα οποία έχουν σχετικά αυξημένη πυκνότητα, παριστάνοντας αραιά σημεία τα οποία είναι δύσκολα διαχωρίσιμα από το αλγόριθμο.

Πράγματι, κοιτώντας τον πινάκα 4.5, ο αλγόριθμος είναι μη ικανός να διαχωρίσει ορθά τα δεδομένα σε συστάδες, καθώς σχηματίζει μία συστάδα με 117 εγγραφές (συστάδα ‘0’), ενώ τις υπόλοιπες τις εντάσσει στην ομάδα θορύβου (επισημαίνεται ως συστάδα ‘-1’). Ακόμη, χρησιμοποιήθηκαν τεχνικές υπερπαραμετροποίησης του αλγορίθμου με βάση τον δείκτη Silhouette score, αλλά ούτε αυτές κατάφεραν να αποφέρουν ικανά αποτελέσματα, με συστάδες που να ικανοποιούν τις προϋποθέσεις για την παραγωγή προφίλ κινδύνου. Το γεγονός δύναται να οφείλεται στο μικρό πλήθος δειγμάτων που χαρακτηρίζεται το προεπεξεργασμένο dataset (μόλις 150 παρατηρήσεις), και στο ότι τα δεδομένα έχουν διαφορετικές πυκνότητες μεταξύ τους. Παράλληλα χρησιμοποιήθηκε μία τεχνική αξιολόγησης για αλγορίθμους πυκνότητας, η DBCV, ωστόσο απέφερε παρόμοια αποτελέσματα με αυτή του Silhouette score.

Πίνακας 4.5: Ποιοτική εκτίμηση πλήθους συστάδων αλγορίθμου DBSCAN

Συστάδα	Πλήθος δειγμάτων
0	117
-1	33

4.3.4 Αξιολόγηση και Επιλογή Βέλτιστης Τεχνικής

Η επιλογή της καταλληλότερης τεχνικής συσταδοποίησης βασίστηκε σε συνδυαστική αξιολόγηση ποσοτικών μετρικών και ποιοτικών κριτηρίων, με γνώμονα τόσο την ποιότητα των συστάδων, την ευκολία

χρήσης τους στην εκπαίδευση εποπτευόμενων μοντέλων όσο και τη χρησιμότητά τους σε κλινικό πλαίσιο. Στη συνέχεια γίνεται μία σύνοψη των αποτελεσμάτων των αλγορίθμων.

K-Means: Παρέχει συνεπή αποτελέσματα και υπολογιστική αποδοτικότητα, ωστόσο δεν παρουσίασε ορατό σημείο απότομης κλίσης της μεθόδου 'Elbow', ενώ οι χαμηλές τιμές Silhouette και η σταδιακή βελτίωση του δείκτη Davies–Bouldin χωρίς σαφές σημείο βέλτιστου k, υποδήλωσαν ασθενή αλλά υπαρκτή δομή. Οι καλύτερες τιμές k αποδείχτηκαν να είναι οι k=3, 4 και 5.

HCA: Η αξιοποίηση της μεθόδου σύνδεσης 'ward', ανέδειξε σαφέστερα διαχωρισμένες και εσωτερικά συνεκτικές ομάδες για συγκεκριμένα επίπεδα αποκοπής, επιτρέποντας διερευνητική κατανόηση της δομής μέσω του δενδρογράμματος. Οι βέλτιστες τιμές k, αποδείχτηκαν να είναι οι k=2 και k=3.

DBSCAN: Μολονότι τεκμηριωμένα προσδιόρισε κατάλληλη τιμή παραμέτρου 'eps' μέσω του k-dist graph και των τεχνικών υπερπαραμέτρων, ακόμη και ύστερα διάφορων πειραμάτων, εμφάνισε περιορισμένη ικανότητα παραγωγής ισορροπημένων συστάδων, με κύριο αποτέλεσμα την κατασκευή μίας μοναδικής συστάδας η οποία δεν αποτελεί θόρυβο. Δεν αποτέλεσε καλό διαχωριστικό μοντέλο για τα δεδομένα του dataset.

Για τον λόγο της δυσκολίας επιλογής της βέλτιστης τεχνικής, η τελική επιλογή δεν στηρίχθηκε σε μία ή δύο μετρικές, αλλά σε ένα συμβιβασμό μεταξύ (α) της διακριτότητας και συνοχής των δεικτών αξιολόγησης, (β) της επάρκειας μεγέθους συστάδων, (γ) της εύρεσης βέλτιστων υπερπαραμέτρων, (δ) της ανθεκτικότητας στο θόρυβο, (ε) της χρησιμότητας στα επόμενα στάδια του πειραματικού σκέλους και (στ) της κλινικής ερμηνείας των εξαγόμενων ομάδων/προφίλ κινδύνου. Για τους αναφερθέντες λόγους επιλέχθηκε η χρήση του αλγορίθμου K-Means, για k=3. Η κυριότερη αιτία επιλογής αυτής της μεθόδου και για k=3, σχετίζεται με την ικανοποίηση των περισσότερων κριτηρίων συμβιβασμού. Ειδικότερα:

1. Ο αλγόριθμος K-Means με 3 συστάδες, παρουσιάζει συγκριτικά ικανοποιητικές τιμές στους δείκτες αξιολόγησης (Silhouette, Davies–Bouldin) σε σχέση με τις εναλλακτικές τεχνικές συσταδοποίησης που εξετάστηκαν.

2. Η κατανομή των παρατηρήσεων ανά συστάδα είναι ισορροπημένη, με κάθε συστάδα να περιλαμβάνει τουλάχιστον 30 δείγματα, γεγονός που δεν παρατηρείται σε άλλες τεχνικές για τρεις ή περισσότερες συστάδες.

3. Δεν κρίθηκε αναγκαία περαιτέρω βελτιστοποίηση υπερπαραμέτρων.

4. Παρότι η τεχνική είναι ευαίσθητη στον θόρυβο και σε ακραίες τιμές, δεν παρατηρήθηκαν έντονες αποκλίσεις που να επηρεάζουν δυσανάλογα τη δομή των συστάδων.

5. Η ύπαρξη επαρκούς αριθμού δειγμάτων ανά συστάδα (τουλάχιστον 30 δείγματα), αναμένεται να συμβάλει ουσιαστικά στην εκπαίδευση και στην αξιόπιστη αξιολόγηση των επιβλεπόμενων μοντέλων ML.

6. Η κλινική ερμηνευσιμότητα των συστάδων διασφαλίζεται μέσω της ανάλυσης των κυρίαρχων χαρακτηριστικών κάθε ομάδας. Η επιλογή τριών συστάδων επιτρέπει τη διαμόρφωση διακριτών προφίλ κινδύνου σε διαβαθμίσεις (π.χ. χαμηλού, μέτριου και υψηλού), προσφέροντας λεπτομερέστερη και πιο ρεαλιστική αποτύπωση, σε σύγκριση με τη χρήση δύο συστάδων.

Έτσι λοιπόν, το dataset διαμορφώνεται με την προσθήκη μίας επιπλέον στήλης, την στήλη 'ετικέτα' ή 'label', που προσδιορίζει την συστάδα που ανήκει η κάθε παρατήρηση. Σε διαφορετική διατύπωση, η προσθήκη της νέας στήλης καθορίζει το προφίλ κινδύνου που ανατέθηκε ο κάθε ηλικιωμένος του dataset, μέσω του αλγορίθμου K-Means.

4.4 Ταξινόμηση

Στο παρόν τμήμα του πειραματικού σκέλους της εργασίας εξετάζεται το στάδιο του SL, με σκοπό την ανάπτυξη και αξιολόγηση μοντέλων ταξινόμησης που προβλέπουν το προφίλ κινδύνου πτώσης των ηλικιωμένων. Η διαδικασία αυτή βασίζεται στις ετικέτες που προέκυψαν από το προηγούμενο στάδιο, αυτό της συσταδοποίησης. Έτσι, το αρχικό σύνολο δεδομένων το οποίο δεν περιελάμβανε ετικέτα, μετασχηματίζεται σε πρόβλημα πολυκατηγορικής ταξινόμησης, επιτρέποντας τη μοντελοποίηση των ορίων απόφασης μεταξύ των διαφορετικών προφίλ. Στόχος της ενότητας είναι η εκπαίδευση, η σύγκριση και η επιλογή του καταλληλότερου ταξινομητή, ο οποίος θα μπορεί να γενικεύει αποτελεσματικά σε νέα δεδομένα ηλικιωμένων, αναθέτοντάς τους στο αντίστοιχο επίπεδο κινδύνου πτώσης που τους αναλογεί.

Η μεταβλητή στόχος που χρησιμοποιείται είναι η ετικέτα συστάδας που προέκυψε από το προηγούμενο βήμα, ενώ ως είσοδοι των μοντέλων, αξιοποιείται το τελικό σύνολο χαρακτηριστικών μετά την προεπεξεργασία. Η διαδικασία αυτή συνιστά μορφή επιβλεπόμενης μάθησης με ψευδο-ετικέτες (pseudo-labels), καθώς οι κλάσεις δεν προέρχονται από τα ίδια τα δεδομένα του αρχικού dataset, αλλά από χρήση τεχνικής μη επιβλεπόμενης ομαδοποίησης των δεδομένων.

Το σύνολο ταξινομητών που εξετάστηκε, καλύπτει διαφορετικές δομές επιβλεπόμενης μάθησης, επιτυγχάνοντας συγκριτική αξιολόγηση και μειώνοντας την εξάρτηση από υποθέσεις ενός μόνο μοντέλου. Συνολικά παρέχουν δυνατότητες για γραμμικούς και μη γραμμικούς διαχωρισμούς καθώς και νευρωνικές προσεγγίσεις, όπως έχουν ήδη παρουσιαστεί στο κεφάλαιο 2. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι οι: Δέντρα Απόφασης, Τυχαία Δάση, XGBoost, Λογιστική Παλινδρόμηση, SVC και MLP. Οι υλοποιήσεις προέρχονται κυρίως από τη βιβλιοθήκη scikit-learn, με εξαίρεση τον XGBoost που υλοποιήθηκε μέσω της ομώνυμης βιβλιοθήκης. Η αξιολόγησή τους λαμβάνεται μέσω της μετρικής ‘f1_macro’ λόγω των ιδιαιτεροτήτων του dataset, όπως αναλύεται παρακάτω. Στη συνέχεια αναλύονται τα βήματα ανάπτυξης και αξιολόγησης των μοντέλων.

4.4.1 Προετοιμασία Δεδομένων

Φόρτωση βιβλιοθηκών

Το πρώτο βήμα της διαδικασίας αποτελεί η εισαγωγή των κατάλληλων βιβλιοθηκών και υλοποιήσεων που θα αξιοποιηθούν στη συνέχεια.

Διάσπαση του συνόλου δεδομένων σε σύνολα χαρακτηριστικών και στόχου

Το σύνολο δεδομένων αναδιαμορφώθηκε σε πίνακα εισόδων (X), ο οποίος περιλαμβάνει το τελικό σύνολο χαρακτηριστικών μετά την προεπεξεργασία και την επιλογή γνωρισμάτων, και σε διάνυσμα στόχου (y), το οποίο αντιστοιχεί στις ψευδο-ετικέτες που προέκυψαν από το στάδιο της συσταδοποίησης, όπως φαίνεται στο παρακάτω απόσπασμα κώδικα. Η διάκριση αυτή αποτελεί το θεμέλιο για τον επακόλουθο πειραματικό σχεδιασμό και την αντικειμενική σύγκριση των μοντέλων.

```
X = df.iloc[:, :-1]
y = df.iloc[:, -1]
```

Διάσπαση συνόλου δεδομένων

Επόμενο βήμα αποτελεί η διάσπαση του dataset, σε σύνολα εκπαίδευσης και ελέγχου, μέσω της αξιοποίησης της τεχνικής train_test_split που προσφέρει η βιβλιοθήκη scikit-learn. Προτού όμως συμβεί αυτό, τα δεδομένα ανακατατάχθηκαν τυχαία, ώστε να αποφευχθεί οποιαδήποτε συστηματική μεροληψία που θα

μπορούσε να προκύψει από την αρχική διάταξη των παρατηρήσεων. Στη συνέχεια εφαρμόστηκε στρωματοποιημένη διάσπαση, στα δύο σύνολα με αναλογία 80:20, εξασφαλίζοντας ότι η κατανομή των κλάσεων του στόχου διατηρείται αναλογικά στα επιμέρους σύνολα. Επίσης χρησιμοποιήθηκε η παράμετρος random_state, για λόγους αναπαραγωγικότητας, ώστε τα δύο σύνολα να παραμένουν σταθερά σε κάθε εκτέλεση.

```
X_train, X_test, y_train, y_test = train_test_split(
       X, y,
       test_size=0.2,
       random_state=42,
       stratify=y,
       shuffle=True
)
```

Η στρωματοποίηση είναι ιδιαίτερα κρίσιμη στο πλαίσιο του συγκεκριμένου dataset, καθώς οι κλάσεις έχουν ανισόρροπες κατανομές μεταξύ τους. Έτσι διατηρείται η αναλογία των κλάσεων, τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο ελέγχου διασφαλίζοντας αντιπροσωπευτική μάθηση και αξιόπιστη εκτίμηση της γενίκευσης των μοντέλων.

4.4.2 Ανάπτυξη Ταξινομητών

Ανάπτυξη ταξινομητών με προεπιλεγμένες παραμέτρους

Όλοι οι ταξινομητές εκπαιδεύτηκαν αρχικά με τις προεπιλεγμένες παραμέτρους τους, ώστε να επιτευχθεί αμερόληπτη σύγκριση της βασικής τους συμπεριφοράς. Η εκπαίδευση υλοποιήθηκε μέσω pipeline που περιλαμβάνει στάδιο προτυποποίησης των αριθμητικών μεταβλητών πριν από την εφαρμογή του εκάστοτε ταξινομητή, με στόχο την αποφυγή ακούσιας επικράτησης χαρακτηριστικών με μεγαλύτερες κλίμακες και τη διασφάλιση ισότιμης συνεισφοράς των μεταβλητών, σύμφωνα με το παρακάτω απόσπασμα κώδικα. Η χρήση της παραμέτρου random_state λαμβάνει χώρα και εδώ για λόγους αναπαραγωγικότητας.

```
models = {
    "Logistic_Regression": LogisticRegression(random_state=42),
    "Decision Tree": DecisionTreeClassifier(random_state=42),
    "Random Forest": RandomForestClassifier(random_state=42),
    "XGBoost": XGBClassifier(random_state=42),
    "SVC": SVC(random_state=42),
    "MLP": MLPClassifier(max_iter=100, random_state=42)
}

pipeline = Pipeline([
        ('scaler', StandardScaler()),
        ('clf', model)
    ])
```

Η αξιολόγηση πραγματοποιήθηκε με διασταυρωμένη επικύρωση πέντε πτυχών (5-fold CV), ώστε τα αποτελέσματα να είναι πιο αξιόπιστα. Ως κύρια μετρική επιλέχθηκε η f1_macro, καθώς αποδίδει ίσο βάρος σε κάθε κλάση ανεξαρτήτως μεγέθους, στοιχείο κρίσιμο σε περιβάλλον με ανισορροπία κλάσεων και με απαιτήσεις ισότιμης σημασίας όλων των προφίλ κινδύνου, όπως στη προκειμένη περίπτωση. Πέραν της κύριας μετρικής, καταγράφηκαν συμπληρωματικά και ο δείκτης accuracy καθώς και ο πίνακας σύγχυσης.

```
scoring = ['f1_macro', 'accuracy']
cv_results = cross_validate(
        pipeline,
        X_train,
        y_train,
        cv=5,
        scoring=scoring,
        n_jobs=-1
)
```

Ανάπτυξη και Βελτιστοποίηση Ταξινομητών

Ακολούθως, για κάθε ταξινομητή δημιουργήθηκε ξεχωριστό περιβάλλον πειραματισμού μέσω ατομικού Jupyter Notebook, με σκοπό την καθαρή και εστιασμένη υλοποίηση των επιμέρους σταδίων εκπαίδευσης, αξιολόγησης και βελτιστοποίησης.

Αρχικά, κάθε ταξινομητής αξιολογήθηκε στο σύνολο ελέγχου χρησιμοποιώντας τις προεπιλεγμένες παραμέτρους του. Η επίδοση αποτυπώθηκε μέσω βασικών μετρικών, όπως η ακρίβεια, ο πίνακας σύγχυσης και η κύρια μετρική f1_macro, προκειμένου να αποκτηθεί μια αρχική εικόνα της συμπεριφοράς του μοντέλου πριν από οποιαδήποτε διαδικασία βελτιστοποίησης.

Στη συνέχεια προσδιορίστηκαν, για κάθε ταξινομητή, οι υπερπαράμετροι που επηρεάζουν ουσιωδώς την απόδοση και τη γενίκευση του μοντέλου. Ως παράδειγμα, παρουσιάζεται η διαδικασία του αλγορίθμου SVC.

```
param_grid_svc = {
        'clf__C': [0.001, 0.01, 0.1, 0.5, 1.0, 10],
        'clf__kernel': ['linear', 'poly', 'rbf'],
        'clf__gamma': [10, 1, 0.1, 0.01, 0.001],
        'clf__class_weight': [None, 'balanced'],
        'clf__degree': [1,2,3,4,5]
}
```

Για τη συστηματική διερεύνηση των συνδυασμών υπερπαραμέτρων εφαρμόστηκε ο αλγόριθμος GridSearchCV, ενσωματωμένος σε διοχέτευση που περιλαμβάνει στάδιο προτυποποίησης των δεδομένων, όπως φαίνεται στον κώδικα παρακάτω.

```
pipe = Pipeline([
    ('scaler', StandardScaler()),
    ('clf', SVC(random_state=42, probability=True))
])

skf = StratifiedKFold(n_splits=5, shuffle=False)

grid = GridSearchCV(
       pipe,
       param_grid_svc,
       scoring='f1_macro',
       n_jobs=-1,
       cv=skf,
       verbose=2
)
grid.fit(X_train, y_train)

best_clf = grid.best_estimator_
```

Στην προσέγγιση αυτή, ο εκτιμητής δεν τοποθετείται απευθείας στη ροή, αλλά ενσωματώνεται στον μηχανισμό αναζήτησης, ο οποίος εκτελεί εξαντλητική αξιολόγηση όλων των προκαθορισμένων συνδυασμών υπερπαραμέτρων μέσω διασταυρούμενης επικύρωσης (skf). Για κάθε συνδυασμό υπολογίζεται η κύρια μετρική αξιολόγησης, και επιλέγεται ο εκτιμητής με τη βέλτιστη μέση επίδοση, μέσω του γνωρίσματος ‘best_estimator_’.

Εσωτερικά στην GridSearchCV συνάρτηση υλοποιείται το εξής: μετά τον προσδιορισμό των βέλτιστων υπερπαραμέτρων του εκτιμητή, ο τελικός εκτιμητής επανεκπαιδεύεται αποκλειστικά στο σύνολο εκπαίδευσης με προτυποποιημένα δεδομένα, χωρίς χρήση διασταυρούμενης επικύρωσης, ώστε να αξιοποιηθεί το σύνολο της διαθέσιμης πληροφορίας για την τελική μάθηση. Ο τελικός εκτιμητής στη συνέχεια χρησιμοποιείται για την εκτίμηση των άγνωστων δεδομένων του συνόλου ελέγχου και την εξαγωγή των τελικών μετρικών απόδοσης.

```
y_pred = best_clf.predict(X_test)

macro_f1 = f1_score(y_test, y_pred, average='macro')
```

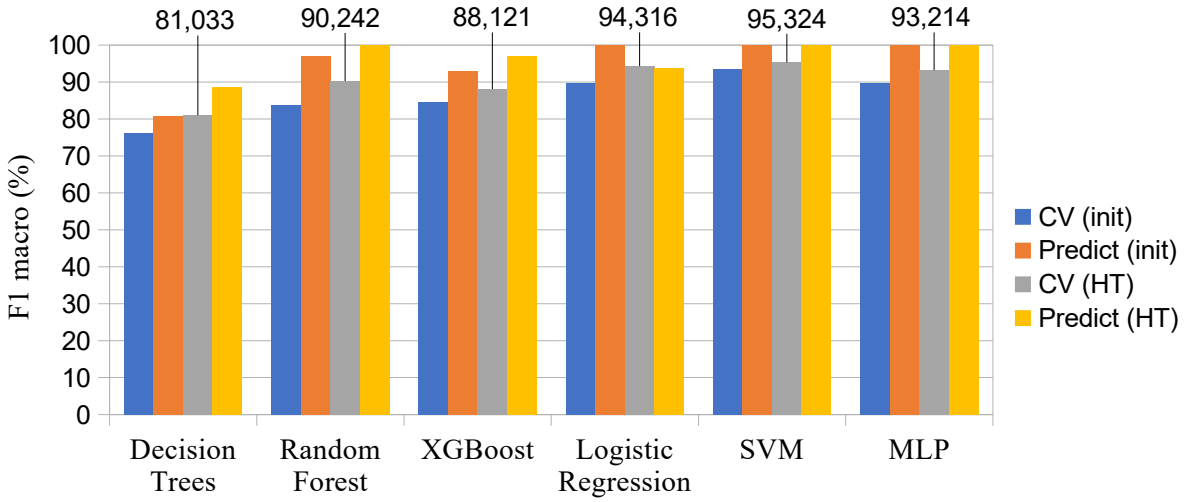
Για την περίπτωση του εκτιμητή MLP, όπου το πλήθος των δυνατών συνδυασμών υπερπαραμέτρων ήταν ιδιαίτερα μεγάλο, χρησιμοποιήθηκε ο αλγόριθμος RandomizedSearchCV ως εναλλακτική του GridSearchCV.

4.4.3 Αξιολόγηση Ταξινομητών και Επιλογή Βέλτιστου

Τα αποτελέσματα των παραπάνω πειραμάτων συγκεντρώθηκαν σε ένα συνολικό συγκριτικό διάγραμμα, το οποίο περιλαμβάνει την κύρια μετρική απόδοσης (f1_macro) για κάθε ταξινομητή τόσο πριν όσο και μετά τη βελτιστοποίηση. Η παρουσίαση αυτή επιτρέπει την άμεση σύγκριση των αλγορίθμων και τεκμηριώνει τη διαδικασία επιλογής του τελικού μοντέλου.

Στο διάγραμμα που ακολουθεί, το υπόμνημα υποδηλώνει τον τρόπο που διαχωρίζονται οι επιδόσεις των μοντέλων ανάλογα με το στάδιο εκπαίδευσης ή ελέγχου. Συγκεκριμένα οι ενδείξεις σημαίνουν:

- **CV (init)**: μέση επίδοση που προκύπτει από τη διαδικασία 5-CV κατά την εκπαίδευση, ως προς το σύνολο επικύρωσης. Μοντέλο με χρήση προκαθορισμένων παραμέτρων.
- **Predict (init)**: επίδοση που προκύπτει από το σύνολο ελέγχου. Μοντέλο με χρήση προκαθορισμένων υπερπαραμέτρων.
- **CV (HT)**: μέση επίδοση που προκύπτει από τη διαδικασία 5-CV κατά την εκπαίδευση, ως προς το σύνολο επικύρωσης. Μοντέλο μετά τη διαδικασία βελτιστοποίησης υπερπαραμέτρων.
- **Predict (HT)**: επίδοση που προκύπτει από το σύνολο ελέγχου. Μοντέλο μετά τη διαδικασία βελτιστοποίησης υπερπαραμέτρων.



Σχήμα 4.11: Σύγκριση μετρικής F1_macro μεταξύ των ταξινομήτων

Σύμφωνα με το διάγραμμα, τα Δέντρα Απόφασης εμφανίζουν αρκετά χαμηλή απόδοση σε όλα τα στάδιο της αξιολόγησης. Από την άλλη τα Τυχαία Δάση υπερέχουν του XGBoost σχεδόν σε κάθε φάση τους πειράματος, επιτυγχάνοντας τιμές άνω των 90%. Ωστόσο οι καλύτεροι ταξινομητές εμφανίζεται να είναι οι τρεις τελευταίοι κατά σειρά.

Η πιο αξιόπιστη βάση αξιολόγησης προκύπτει από τον συνδυασμό της μέσης απόδοσης των CVs και της απόδοσης πρόβλεψης στο σύνολο ελέγχου κατά το στάδιο βελτιστοποίησης των υπερπαραμέτρων. Η διαφορές των αποτελεσμάτων μεταξύ των μοντέλων Λογιστικής Παλινδρόμησης, SVC και MLP είναι πολύ

μικρές. Παρ' όλα αυτά, μπορεί ο SVC αναδεικνύεται ως ο καλύτερος ταξινομητής για το συγκεκριμένο σύνολο δεδομένων, παρουσιάζοντας ελαφρώς ανώτερη και σταθερότερη συμπεριφορά ως προς τη γενίκευση σε άγνωστα δεδομένα, με ποσοστό 95.32% μέση απόδοση από τη διασταυρούμενη επικύρωση και 100% στις προβλέψεις άγνωστων δεδομένων.

4.5 Επίλογος

Στο παρόν κεφάλαιο εξετάστηκαν συνδυαστικά οι αλγόριθμοι συσταδοποίησης και τα μοντέλα επιβλεπόμενης μάθησης για την ανάπτυξη του πειραματικού μέρους της εργασίας. Η μεθοδολογική προσέγγιση που υιοθετήθηκε βασίστηκε στο «συσταδοποίηση-μετά-ταξινόμηση», σύμφωνα με το οποίο οι ομάδες που προέκυψαν από τη διαδικασία συσταδοποίησης αξιοποιήθηκαν ως ψευδο-ετικέτες και χρησιμοποιήθηκαν ως μεταβλητή-στόχος στο στάδιο της ταξινόμησης.

Μετά την ολοκλήρωση της προεπεξεργασίας και της κατάλληλης διαμόρφωσης των δεδομένων, πραγματοποιήθηκε συγκριτική αξιολόγηση πολλαπλών τεχνικών συσταδοποίησης, με τον αλγόριθμο K-Means να αναδεικνύεται ως η καταλληλότερη επιλογή, καθώς πέτυχε σαφή δομή των δεδομένων και ικανοποίησε τα κριτήρια επιλογής βάσει ενός ισορροπημένου συμβιβασμού μεταξύ ποιότητας συστάδων, ερμηνευσιμότητας και των γενικών στόχων της εργασίας. Ακολούθως, τα δεδομένα οργανώθηκαν σε κατάλληλη μορφή για είσοδο σε μοντέλα SL, λαμβάνοντας υπόψη τους περιορισμούς του συνόλου δεδομένων και τις απαιτήσεις γενίκευσης. Η διαδικασία ταξινόμησης περιλάμβανε συστηματική σύγκριση διαφορετικών αλγορίθμων μέσα από διαδοχικά στάδια αξιολόγησης στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου. Μέσα από τη διαδικασία αυτή, ο ταξινομητής SVC αναδείχθηκε ως η βέλτιστη μέθοδος, καθώς επέδειξε ικανότητα εύρεσης σταθερών ορίων μεταξύ κλάσεων, αποτελεσματικό χειρισμό μη γραμμικών διαχωρισμών και αυξημένη ικανότητα γενίκευσης σε άγνωστα δεδομένα.

Με την ολοκλήρωση του κεφαλαίου, έχει αναπτυχθεί ένα μοντέλο ML το οποίο επιτρέπει την αυτόματη αντιστοίχιση ενός νέου ηλικιωμένου ατόμου που έχει υποστεί πτώση σε προφίλ κινδύνου, όπως αυτά προέκυψαν από τη μη επιβλεπόμενη ομαδοποίηση.

Κεφάλαιο 5ο: Ερμηνεία και Ανάλυση Μοντέλου

Η ερμηνευσιμότητα αναφέρεται στη δυνατότητα κατανόησης του τρόπου με τον οποίο ένα μοντέλο καταλήγει στις αποφάσεις του και της συνεισφοράς των επιμέρους χαρακτηριστικών στην τελική πρόβλεψη. Στο πλαίσιο της παρούσας εργασίας, η ερμηνευσιμότητα δεν αντιμετωπίζεται ως συμπληρωματικό στοιχείο, αλλά ως αναγκαία προϋπόθεση για την αξιοπιστία, τη διαφάνεια και τη μεταφορά των αποτελεσμάτων σε κλινικά ή εφαρμοσμένα περιβάλλοντα. Κατά συνέπεια, το παρόν κεφάλαιο εστιάζει στην ανάλυση του εκπαιδευμένου μοντέλου, με στόχο την ανάδειξη των παραγόντων που επηρεάζουν τις προβλέψεις, την τεκμηρίωση της λογικής πίσω από την ανάθεση κάθε ατόμου σε συγκεκριμένο προφίλ κινδύνου καθώς και την παροχή εξατομικευμένων οδηγιών πρόληψης.

Το κεφάλαιο αυτό ξεκινάει με την ανάδειξη της ανάγκης χρήσης της ερμηνευσιμότητας. Συνεχίζει με την περιγραφή των πιο σημαντικών χαρακτηριστικών του κάθε προφίλ κινδύνου μέσω μίας τεχνικής ερμηνευσιμότητας, των SHAP values. Αργότερα, απλώνεται στην ανάλυση των υποομάδων (προφίλ) προσφέροντας μία ερμηνεία για το καθένα ανάλογα με τις ιδιαιτερότητές τους. Ύστερα, σημειώνονται οι παρατηρήσεις των παραπάνω ερμηνειών. Το κεφάλαιο ολοκληρώνεται με την προσφορά παρεμβάσεων πρόληψης πτώσεων με βάση τα χαρακτηριστικά του κάθε προφίλ κινδύνου, προκειμένου να συμβάλει στην λήψη αποφάσεων πρόληψης πτώσεων.

5.1 Ο ρόλος της ερμηνευσιμότητας

Η πρόληψη πτώσεων, ιδιαίτερα σε ηλικιωμένους ανθρώπους, αποτελεί πεδίο όπου οι αποφάσεις που βασίζονται σε δεδομένα οφείλουν να είναι όχι μόνο ακριβείς, αλλά και κατανοητές από τους επαγγελματίες υγείας που καλούνται να τις αξιοποιήσουν. Η εκτίμηση του κινδύνου πτώσης συνδέεται άμεσα με την επιλογή των παρεμβάσεων, τη διαχείριση πόρων και την επικοινωνία με τον ίδιο τον ασθενή. Ως εκ τούτου, η δυνατότητα εξήγησης της αιτίας πίσω από την πρόβλεψη αποτελεί κρίσιμο παράγοντα για την αποδοχή, την εμπιστευτικότητα και την υπεύθυνη χρήση των υπολογιστικών μοντέλων.

Το μοντέλο SVC εκπαιδεύτηκε να αναθέτει κάθε άτομο σε προφίλ κινδύνου που προέκυψαν από μη επιβλεπόμενη ομαδοποίηση, λαμβάνοντας υπόψη πολυπαραγοντικά χαρακτηριστικά (λειτουργικά, κλινικά, δημογραφικά και περιβαλλοντικά). Η προσέγγιση αυτή επιτρέπει τη δημιουργία ομάδων ατόμων με κοινά χαρακτηριστικά, αλλά δεν παρέχει εκ των προτέρων κλινική ερμηνεία του κινδύνου πτώσης σχετικά με το ίδιο το άτομο. Συνεπώς, η ερμηνευσιμότητα καθίσταται απαραίτητη για τη μετάφραση των αποτελεσμάτων, ώστε να είναι γνωστοί οι παράγοντες που αυξάνουν ή μειώνουν τον εκτιμώμενο κίνδυνο, η κατεύθυνση και η βαρύτητα που αντιστοιχεί στην κάθε ομάδα.

Επιπλέον, η ερμηνευτική ανάλυση επιτρέπει την αξιολόγηση από κλινική άποψη των προβλέψεων και τη διασταύρωσή τους με την υπάρχουσα γνώση για τους παράγοντες των πτώσεων. Με τον τρόπο αυτό, διασφαλίζεται ότι το μοντέλο δεν βασίζεται σε τυχαίες συσχετίσεις ή σε μεροληπτικά μοτίβα των δεδομένων, αλλά σε χαρακτηριστικά με ουσιαστικό νόημα για την πρόληψη.

5.2 SHAP values

Μία δημοφιλής τεχνική ερμηνείας των προβλέψεων των ML μοντέλων, είναι τα SHAP (Shapley Additive exPlanations) values, που βασίζονται στις τιμές Shapley της θεωρίας παιγνίων. Τα SHAP values αξιοποιούν την τελική πρόβλεψη ενός μοντέλου, παρέχοντας ένα συνολικό άθροισμα συνεισφοράς των χαρακτηριστικών του, όπου κάθε χαρακτηριστικό «αμείβεται» αναλογικά με τη συμβολή του στη μεταβολή της πρόβλεψης σε σχέση με μια τιμή αναφοράς [100].

Ένα κρίσιμο πλεονέκτημα των SHAP values είναι ότι ικανοποιούν ιδιότητες που θεωρούνται θεμελιώδεις για αξιόπιστη ερμηνεία, όπως είναι η τοπική ακρίβεια (η άθροιση των αποδόσεων του κάθε γνωρίσματος αναπαράγει ακριβώς την τιμή πρόβλεψης), και η απουσία (χαρακτηριστικά χωρίς επίδραση σε κάποια κλάση, λαμβάνουν μηδενική συνεισφορά) [100]. Το γεγονός ότι οι αποδόσεις είναι προσθετικές επιτρέπει την άμεση ποσοτική σύγκριση της σχετικής συμβολής διαφορετικών χαρακτηριστικών, διευκολύνοντας την ερμηνεία σε εφαρμοσμένα πλαίσια όπου απαιτείται σαφής αιτιολόγηση των αποφάσεων, όπως συμβαίνει σε περιπτώσεις προβλημάτων υγείας. Ακόμη, το SHAP δεν εξαρτάται από συγκεκριμένα είδη μοντέλων (model-agnostic), επιτρέποντας έτσι την εφαρμογή του τόσο σε γραμμικά όσο και σε μη γραμμικά μοντέλα, γεγονός που είναι σημαντικό για την κατανόηση μέχρι και σύνθετων μοντέλων. Ωστόσο, παρουσιάζονται συγκεκριμένοι περιορισμοί που πρέπει να ληφθούν υπόψη. Στην παρουσία ισχυρά συσχετισμένων χαρακτηριστικών, η κατανομή της συνεισφοράς μπορεί να διαμοιραστεί μεταξύ μεταβλητών με τρόπο που δυσχεραίνει την αιτιακή ερμηνεία, παρότι διατηρείται η προσθετική ακρίβεια [101]. Ο υπολογισμός των τιμών SHAP μπορεί να είναι υπολογιστικά απαιτητικός σε κάποιες περιπτώσεις, καθιστώντας αναγκαία την επιλογή αντιπροσωπευτικών υποσυνόλων δεδομένων ή προσεγγιστικών μεθόδων. Συνεπώς, το SHAP παρέχει ισχυρή ερμηνευτική ένδειξη, αλλά δεν υποκαθιστά την κλινική κρίση των ατόμων, παραμένοντας ένα χρήσιμο εργαλείο για συμπληρωματική χρήση και συμβολή στη λήψη αποφάσεων.

Επομένως, το SHAP, λειτουργεί ως γέφυρα μεταξύ των προβλέψεων των μοντέλων ML, με την κλινική ερμηνεία των αποτελεσμάτων τους. Στην περίπτωση της διαχείρισης των προφίλ κινδύνου της εργασίας, έχουν την δυνατότητα να ποσοτικοποιήσουν την θετική ή αρνητική συμβολή κάθε χαρακτηριστικού σε επίπεδο ομάδας (προφίλ κινδύνου) ή ατόμου (ατομική πρόβλεψη ανάθεσης σε προφίλ κινδύνου).

5.2.1 Αποτελέσματα των SHAP values

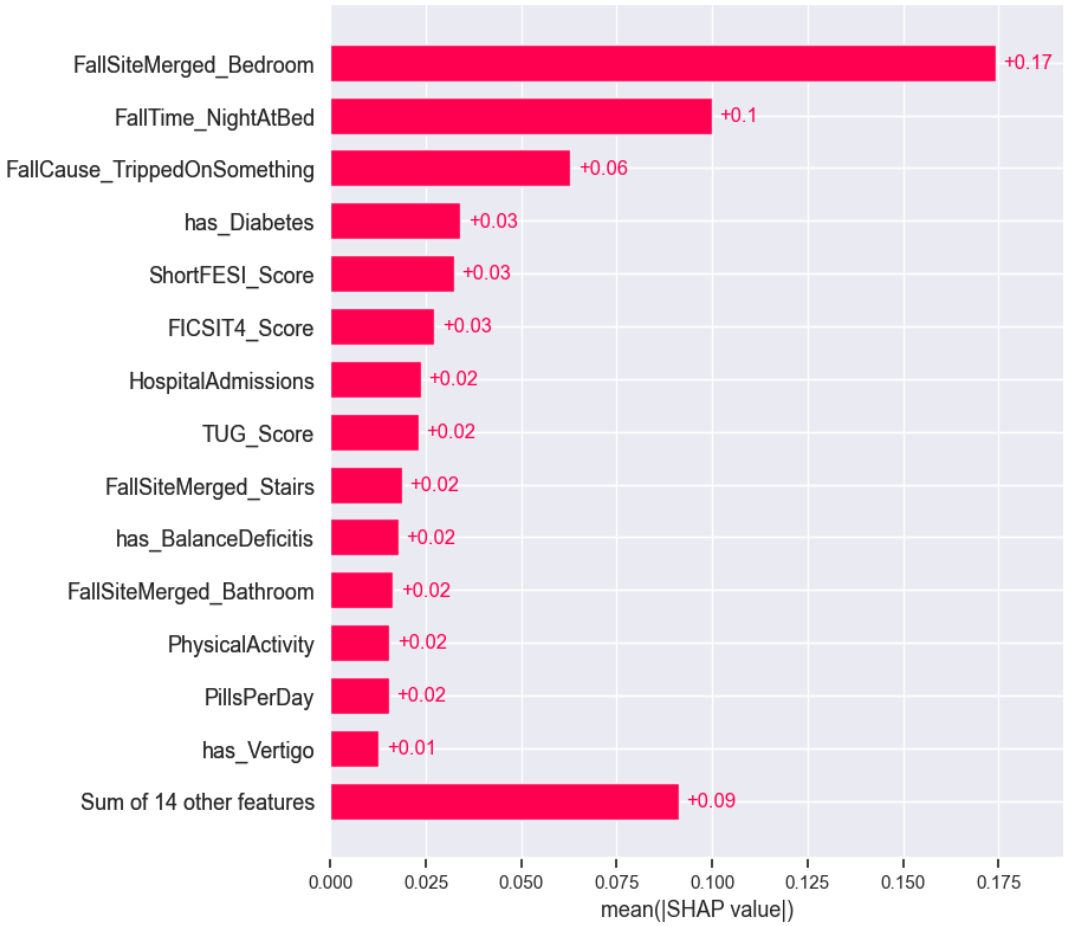
Οι τεχνικές SHAP ενσωματώθηκαν στην εργασία με σκοπό την ερμηνεία των προβλέψεων του τελικού ταξινομητή και την τεκμηρίωση των παραγόντων που καθορίζουν την ανάθεση κάθε ατόμου σε συγκεκριμένο προφίλ κινδύνου. Η υλοποίηση πραγματοποιήθηκε επί του εκπαιδευμένου μοντέλου που επιλέχθηκε (SVC), διατηρώντας σταθερή τη διαδικασία προεπεξεργασίας (προτυποποίηση) ώστε οι αποδόσεις SHAP να αντανακλούν ακριβώς τον χώρο εισόδου του μοντέλου. Έτσι, τα SHAP values ερμηνεύουν την συμπεριφορά του μοντέλου που χρησιμοποιείται για την παραγωγή των προβλέψεων.

Η ερμηνεία του μοντέλου οργανώθηκε σε δύο επίπεδα. Το πρώτο αφορά την παγκόσμια (global) ανάλυση, η οποία αποκαλύπτει την ανάδειξη της συνολικής σημαντικότητας των χαρακτηριστικών ως προς τις νέες προβλέψεις του μοντέλου, συγκρίνοντας έτσι τη σχετική συνεισφορά λειτουργικών, κλινικών, δημογραφικών και περιβαλλοντικών μεταβλητών. Το δεύτερο επίπεδο αφορά την τοπική (local) ανάλυση, η οποία απεικονίζει τα αίτια που μία πρόβλεψη κατέληξε στην ανάθεση ενός συγκεκριμένου προφίλ κινδύνου, αποτυπώνοντας ποια χαρακτηριστικά ενίσχυσαν ή αποδυνάμωσαν την ανάθεση στο αντίστοιχο προφίλ. Και στις δύο περιπτώσεις, οι αποδόσεις ερμηνεύτηκαν ως θετικές ή αρνητικές συνεισφορές ως προς την προβλεπόμενη κλάση. Σημαντική παρατήρηση αποτελεί ότι οι αποδόσεις υπολογίστηκαν ανά κλάση, επιτρέποντας με αυτό τον τρόπο την απλούστευση της παρουσίασης των διαγραμμάτων και την αποσαφήνιση των παραγόντων που διαφοροποιούν το κάθε προφίλ.

Το σύνολο δεδομένων αναφοράς που χρησιμοποιήθηκε είναι το αντιπροσωπευτικό υποσύνολο δεδομένων εκπαίδευσης (X_train) του τελικού μοντέλου, ώστε να αντανακλά την εμπειρική κατανομή του πληθυσμού της μελέτης και να αποφεύγεται μεροληψία στις αποδόσεις.

Όσον αφορά την ερμηνεία του μοντέλου σε παγκόσμιο επίπεδο, η ανάλυση παρουσιάζεται μέσω διαγραμμάτων bar plot, τα οποία αποτυπώνουν τη μέση απόλυτη τιμή των SHAP αποδόσεων για κάθε χαρακτηριστικό. Τα διαγράμματα αυτά επιτρέπουν την άμεση ιεράρχηση των μεταβλητών ως προς τη συνολική συνεισφορά τους στις προβλέψεις του μοντέλου (από πάνω προς τα κάτω), αναδεικνύοντας τους πλέον καθοριστικούς παράγοντες για τη διάκριση των προφίλ κινδύνου σε επίπεδο πληθυσμού. Παρακάτω

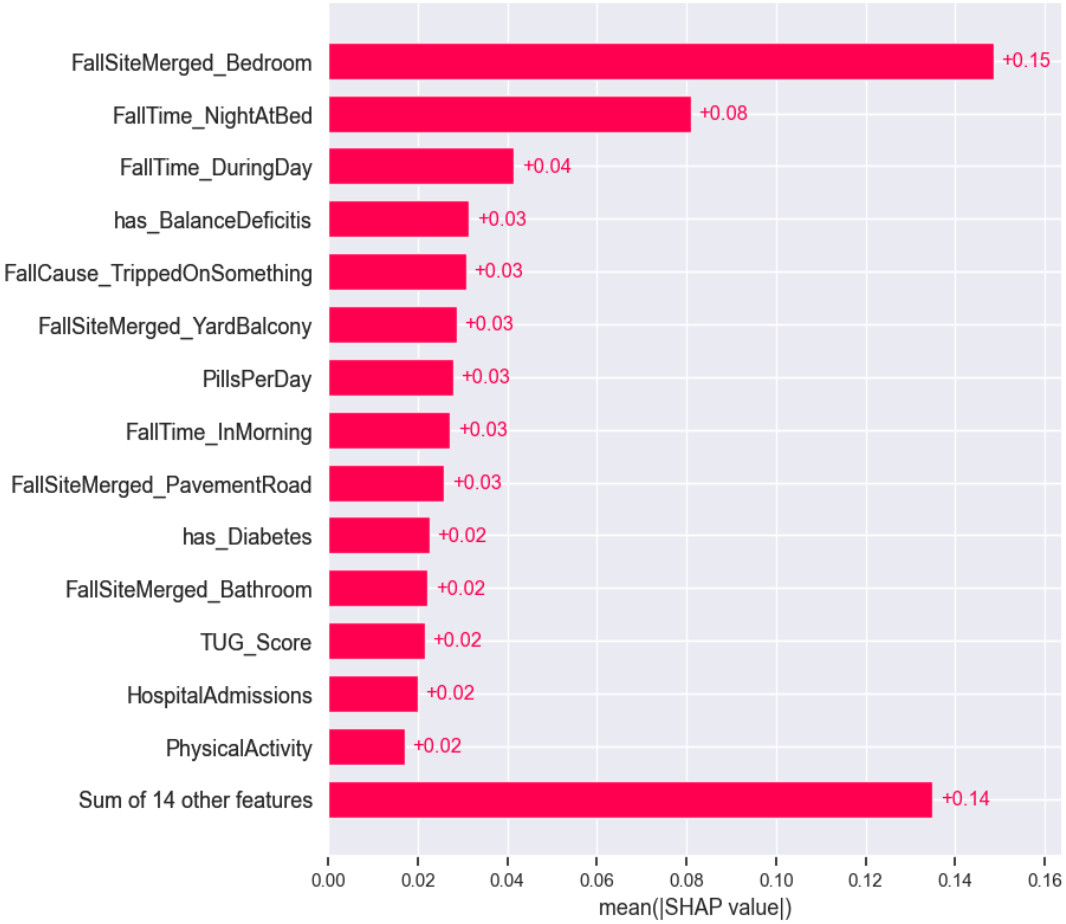
απεικονίζονται τα bar plots για κάθε προφίλ κινδύνου, παρουσιάζοντας την συνεισφορά των πρώτων 14 πιο σημαντικών παραγόντων ανά προφίλ.



Σχήμα 5.1: Bar plot προφίλ κινδύνου 0

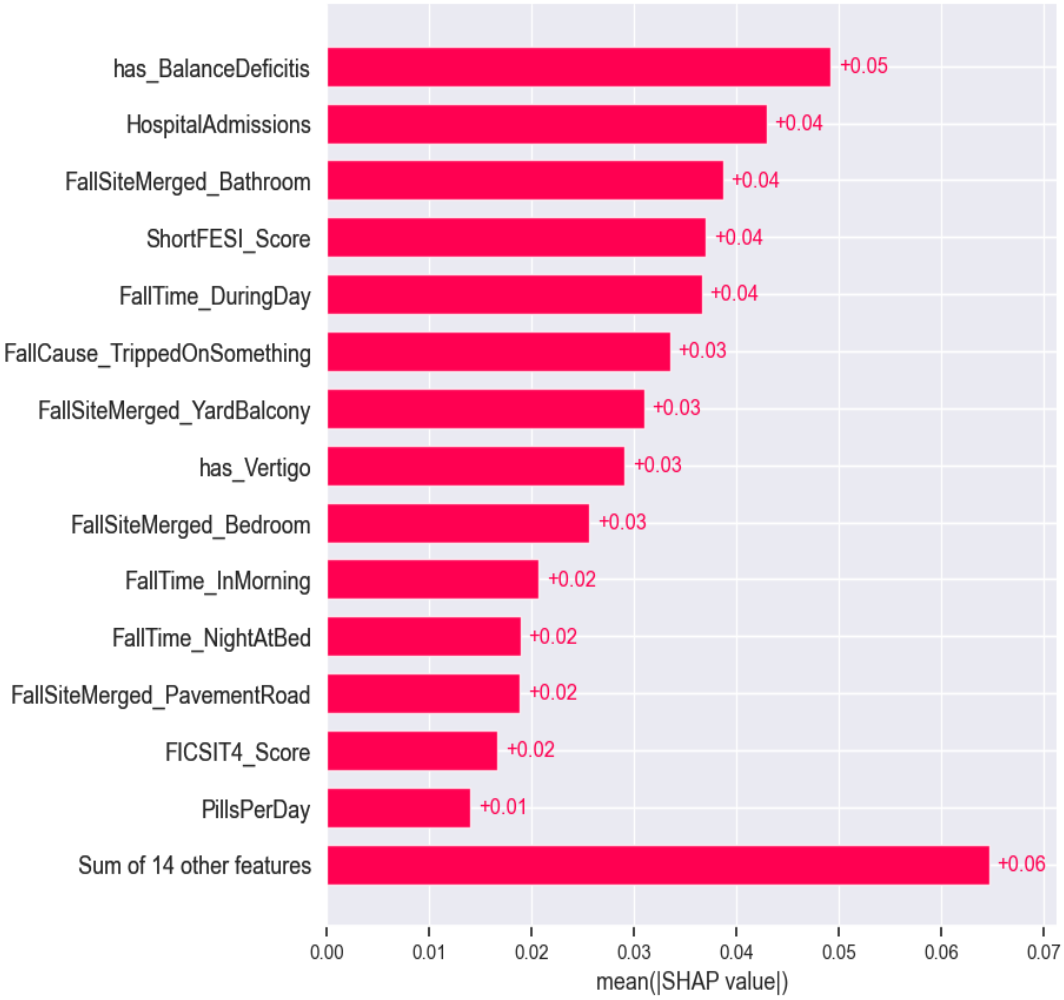
Για το πρώτο προφίλ κινδύνου, τη μεγαλύτερη επίδραση καταλαμβάνουν οι παράγοντες: πτώση στο υπνοδωμάτιο, πτώση κατά την κατάκλιση, τρόπος πτώσης σκόνταμμα, ύπαρξη διαβήτη, και οι λειτουργικές αξιολογήσεις Short FES-I και FICSIT-4. Μάλιστα, ο παράγοντας ‘πτώση στο υπνοδωμάτιο’, επηρεάζει σε μεγάλο βαθμό της ανάθεση μίας πρόβλεψης στο συγκεκριμένο προφίλ, αποκαλύπτοντας τη μεγάλη επιρροή του. Γενικά, η μεγαλύτερη επίδραση εμφανίζεται να υπάρχει στα περιβαλλοντικά χαρακτηριστικά του συνόλου, έπειτα στα κλινικά και λειτουργικά.

Στο δεύτερο προφίλ κινδύνου, οι πέντε πρώτοι παράγοντες που καταλαμβάνουν την μεγαλύτερη επίδραση



Σχήμα 5.2: Bar plot προφίλ κινδύνου 1

είναι οι: πτώση στο υπνοδωμάτιο, πτώση κατά την κατάκλιση και την ημέρα, έλλειψη ισορροπίας και τρόπος πτώσης σκόνταμμα. Και σε αυτή τη περίπτωση, την υψηλότερη επίδραση την έχει η τοποθεσία πτώσης 'υπνοδωμάτιο' γεγονός που σημαίνει ότι συμμετέχει σημαντικά στη συνεισφορά των χαρακτηριστικών. Σχετικά με τις κατηγορίες μεταβλητών, πρώτη και με μεγάλη διαφορά από τις υπόλοιπες κατηγορίες σε αυτό το προφίλ κινδύνου αποτελεί η κατηγορία των περιβαλλοντικών παραγόντων, ενώ επόμενες παρουσιάζονται οι κλινικές και οι λειτουργικές μεταβλητές.



Σχήμα 5.3: Bar plot προφίλ κινδύνου 2

Τέλος, το τρίτο προφίλ κινδύνου παρουσιάζει μία πιο αναμειγμένη πολυπαραγοντική επιρροή, σε αντίθεση με τα προηγούμενα δύο. Τις πέντε πρώτες θέσεις κατά υπολογιστική επιρροή καταλαμβάνουν οι παράγοντες: έλλειψη ισορροπίας, πλήθος εισαγωγών σε νοσοκομείο, πτώση στο μπάνιο, αξιολόγηση Short FES-I και πτώση κατά τη διάρκεια της ημέρας. Εδώ ο παράγοντας ‘πτώση στο υπνοδωμάτιο’ έχει πολύ μικρή επιρροή καθώς δεν περιλαμβάνεται ούτε στις πρώτες 14 ισχυρότερες μεταβλητές συνεισφοράς διαμόρφωσης του προφίλ. Όσον αφορά τις κατηγορίες μεταβλητών, κατά σειρά μεγαλύτερη επιρροή παρουσιάζουν οι περιβαλλοντικοί, οι κλινικοί και οι λειτουργικοί παράγοντες.

Συνοψίζοντας, οι τιμές SHAP ερμηνεύουν τα αποτελέσματα απεικονίζοντας ιδιαίτερα μεγάλη επιρροή από τα περιβαλλοντικά χαρακτηριστικά των πτώσεων (για κάθε προφίλ) και ύστερα ακολουθούν οι υπόλοιπες κατηγορίες χαρακτηριστικών.

Από την τοπική ερμηνεία των προβλέψεων με χρήση των SHAP values, χρησιμοποιούνται διαγράμματα καταρράκτη (waterfall), τα οποία απεικονίζουν τη διαδοχική συνεισφορά των χαρακτηριστικών από την τιμή αναφοράς (baseline) έως την τελική πρόβλεψη για ένα συγκεκριμένο άτομο. Η απεικόνιση αυτή καθιστά σαφές ποια χαρακτηριστικά ενίσχυσαν ή αποδυνάμωσαν την ανάθεση στο αντίστοιχο προφίλ κινδύνου, παρέχοντας αιτιολόγηση της απόφασης σε επίπεδο μεμονωμένου ατόμου. Παράδειγμα τοπικής ερμηνείας παρουσιάζεται στο επόμενο κεφάλαιο.

5.3 Ανάλυση Κλάσεων

Η παρούσα ενότητα εστιάζει στην αναλυτική ερμηνεία των προφίλ κινδύνου που προέκυψαν από την εφαρμογή του υβριδικού πλαισίου συσταδοποίησης και ταξινόμησης. Εφόσον τεκμηριώθηκε η λειτουργία του τελικού μοντέλου και η ερμηνευσιμότητα των προβλέψεών του μέσω τεχνικών SHAP, στο εξής επιχειρείται η μετάβαση από την υπολογιστική αναπαράσταση σε μία εννοιολογική και κλινική διατύπωση των ομάδων. Κατά συνέπεια, και με τη συμβολή των τεχνικών SHAP, ολοκληρώνεται η αποσαφήνιση των χαρακτηριστικών που διαφοροποιούν τις κλάσεις μεταξύ τους και αναδεικνύονται τα συνεκτικά προφίλ σε φυσική γλώσσα, τα οποία μπορούν να υποστηρίξουν την κατανόηση του κινδύνου πτώσης και τη στοχευμένη πρόληψη.

Η ερμηνεία των προφίλ βασίζεται σε συστηματική ανάλυση των δεδομένων που προηγήθηκε σε επίπεδο κώδικα, όπου για κάθε κλάση υπολογίστηκαν στατιστικά περιγραφικά μέτρα (μέσες τιμές και διασπορές) των χαρακτηριστικών. Η προσέγγιση αυτή επιτρέπει τη σύγκριση των ομάδων ως προς τις κεντρικές τάσεις και τη μεταβλητότητά τους, προσφέροντας μια πιο εύρωστη και σε βάθος κατανόηση της δομής των προφίλ κινδύνου. Στις επόμενες υποενότητες αναδεικνύονται για κάθε κλάση ξεχωριστά, η ερμηνεία τους με βάση τα κυρίαρχα χαρακτηριστικά τους, και αποδίδεται ένας συνοπτικός χαρακτηρισμός που περιγράφει σε γενικές γραμμές το περιεχόμενο και τη φύση του αντίστοιχου προφίλ κινδύνου. Η διαδικασία αυτή στοχεύει στη συγκρίσιμη και κλινικά αξιοποιήσιμη αποτύπωση των ομάδων που αναδείχθηκαν από το μοντέλο.

5.3.1 Ανάλυση Προφίλ Κινδύνου 0

Η κλάση 0 περιλαμβάνει άτομα με διατηρημένη λειτουργική ικανότητα, η οποία εκδηλώνεται μέσω καλών επιπέδων ισορροπίας και κινητικότητας, χαμηλού φόβου πτώσης και περιορισμένης κλινικής επιβάρυνσης. Η πλειονότητα των ατόμων της ομάδας έχει πραγματοποιήσει τις ετήσιες καρδιολογικές εξετάσεις (87.5%), γεγονός που υποδηλώνει αυξημένη συμμόρφωση σε προληπτικές πρακτικές υγείας. Η παρουσία παραγόντων που σχετίζονται με αυξημένο κίνδυνο πτώσεων είναι περιορισμένη, καθώς μόλις το 1/5 εμφανίζει οστεοπόρωση, ενώ περίπου το 1/3 αναφέρει προβλήματα βάδισης ή ισορροπίας. Οι πτώσεις εμφανίζονται κυρίως ως περιστασιακά και περιβαλλοντικά γεγονότα, και όχι ως αποτέλεσμα συστηματικής αστάθειας ή παθολογικών αιτιών. Το μεγαλύτερο ποσοστό των ηλικιωμένων της ομάδας έχει υποστεί ακριβώς μία πτώση (μέση τιμή 1.43). Παράλληλα, τα περιστατικά είναι συνδεδεμένα κυρίως με χώρους εντός της οικίας των ατόμων (~75% των περιπτώσεων) και συμβαίνουν μέσα στο πλαίσιο καθημερινών δραστηριοτήτων, χρονικά τόσο κατά τη διάρκεια της ημέρας όσο και της νύχτας. Επιπλέον, η χαμηλή πολυφαρμακία (μέση τιμή ~1.5 ανά άτομο) και η περιορισμένη παρουσία ιλίγγου, ενισχύουν την εικόνα ενός προφίλ χαμηλού συνολικού κινδύνου.

Εν ολίγοις, η κλάση 0 αντιπροσωπεύει άτομα που:

- διαθέτουν καλή ισορροπία και κινητικότητα (γενικά καλή λειτουργική κατάσταση, μέση τιμή δραστηριότητας: 2.04),
- έχουν χαμηλό φόβο πτώσης,
- παρουσιάζουν ελάχιστη κλινική επιβάρυνση (περιορισμένα προβλήματα συννοσηροτήτων),
- υφίστανται πτώσεις κυρίως χαμηλής σοβαρότητας,
- υποτροπιάζουν σε περιστασιακές πτώσεις που σχετίζονται κυρίως με το περιβάλλον.

Επομένως, ένας χαρακτηρισμός που μπορεί να δοθεί στο προφίλ είναι: «Προφίλ χαμηλού κινδύνου πτώσης με διατηρημένη λειτουργική ικανότητα».

5.3.2 Ανάλυση Προφίλ Κινδύνου 1

Η κλάση 1 παρουσιάζει σχετικά ικανοποιητικό επίπεδο λειτουργικής ικανότητας των ατόμων, ωστόσο συνυπάρχουν επιβαρυντικοί παράγοντες που διαφοροποιούν την ομάδα από το χαμηλού κινδύνου προφίλ. Ειδικότερα, παρατηρούνται αυξημένος φόβος πτώσης, μειωμένη φυσική δραστηριότητα και μεγαλύτερη φαρμακευτική επιβάρυνση. Οι ετήσιες καρδιολογικές εξετάσεις πραγματοποιούνται από μικρότερο ποσοστό των ηλικιωμένων (66%), περίπου το 1/3 εμφανίζει οστεοπόρωση και προβλήματα βάδισης ή ισορροπίας, ενώ σχεδόν το 19% παρουσιάζει διαβήτη. Αυτά τα στοιχεία ενισχύουν το συνολικό κλινικό φόρτο της ομάδας. Ιδιαίτερο ενδιαφέρον έχει η πλήρης απουσία ιλίγγου (0%) σε συνδυασμό με το υψηλό μέσο πλήθος κατανάλωσης χαπιών ανά ημέρα (~2.3). Το κυρίαρχο χαρακτηριστικό της κλάσης αποτελεί το έντονο νυχτερινό μοτίβο πτώσεων, το οποίο παρατηρείται στο 84.5% των περιστατικών. Συχνό φαινόμενο πτώσεων εμφανίζεται στο υπνοδωμάτιο (60%) και στις μετακινήσεις κατά τη διάρκεια της νύχτας (58%) ύστερα από έγερση από τον ύπνο (π.χ. μετάβαση στη τουαλέτα). Η επικρατέστερη αιτία πτώσης είναι το παραπάτημα (53.33%), στοιχείο που ερμηνεύει και τη σχεδόν καθολική εμφάνιση των περιστατικών εντός της οικίας των ηλικιωμένου (~94.5%). Τέλος, σε σύγκριση με την κλάση 0, διαπιστώνεται αυξημένος αριθμός πτώσεων ανά άτομο (μέση τιμή ~1.9), γεγονός που υποδηλώνει μεγαλύτερη τάση επαναληπτικότητας των πτώσεων.

Σε γενικές γραμμές η κλάση 1, χαρακτηρίζεται από:

- σχετικά διατηρημένη λειτουργική ικανότητα,
- αυξημένο φόβο πτώσης και μειωμένη δραστηριότητα (μέση τιμή: 1.70),
- μέτρια κλινική επιβάρυνση (οστεοπόρωση, προβλήματα βάδισης/ισορροπίας, διαβήτης, πολυφαρμακία),
- σαφές νυχτερινό μοτίβο πτώσεων εντός οικίας, κυρίως στο υπνοδωμάτιο,
- υψηλή αιτία πτώσης το ‘παραπάτημα’ σε κάποιο αντικείμενο.

Σύμφωνα με όλα τα παραπάνω, αυτό το προφίλ κινδύνου ερμηνεύεται ως: «Προφίλ μέτριου κινδύνου πτώσης, με νυχτερινή ευαλωτότητα και αυξημένο κλινικό φορτίο».

Αξίζει να επισημανθεί, η επιβεβαίωση της δήλωσης του άρθρου του dataset [89], το οποίο αναφέρει ότι οι πτώσεις στο χώρο του υπνοδωματίου συσχετίζονται με πτώσεις που συνέβησαν κατά τη διάρκεια της νύχτας, όπως ακριβώς αποτυπώνει η ερμηνεία αυτού του προφίλ κινδύνου.

5.3.3 Ανάλυση Προφίλ Κινδύνου 2

Η κλάση 2 αντιπροσωπεύει άτομα με γενικευμένη αστάθεια, έντονα μειωμένη λειτουργική ικανότητα και υψηλό φόβο πτώσης, συγκροτώντας το προφίλ της πιο ευάλωτης ομάδας του συνόλου δεδομένων. Χαρακτηρίζεται από πολύ υψηλή παρουσία προβλημάτων αστάθειας και στάσης (>90%), καθώς και από υψηλή συχνότητα ιλίγγου (~48.5%) σε σχέση με τα υπόλοιπα προφίλ. Το γεγονός ότι εμφανίζεται πιο ομοιόμορφη κατανομή των πτώσεων ως προς τον χώρο πτώσης και τη χρονική διασπορά σε όλο το εικοσιτετράωρο, συντελεί σε πιο απρόβλεπτες πτώσεις. Ακόμη, αν και εξακολουθεί να επικρατεί το μοτίβο των πτώσεων εντός οικείας, το ποσοστό είναι μειωμένο σε σχέση με τις υπόλοιπες κλάσεις (<79.5%), πράγμα που υποδηλώνει ότι ο κίνδυνος δεν περιορίζεται σε συγκεκριμένα περιβάλλοντα ή δραστηριότητες. Η ιδιαίτερα υψηλή συχνότητα επαναλαμβανόμενων πτώσεων (μέση τιμή ~3.5 ανά άτομο) και οι αυξημένες εισαγωγές σε νοσοκομείο (~1.5 ανά άτομο) καταδεικνύουν ένα χρόνιο και εξελισσόμενο πρόβλημα για τους ηλικιωμένους που εντάσσονται στο προφίλ αυτό. Συνολικά, η κλάση, αποτυπώνει ένα μοτίβο

πολυπαραγοντικής αστάθειας, με αυξημένες λειτουργικές αξιολογήσεις, επιβάρυνση μέσω κλινικών δεικτών και επαναληπτικότητα των πτώσεων.

Με λίγα λόγια, τα κύρια ζητήματα του προφίλ αποτελούν την/τον:

- σοβαρή λειτουργική αστάθεια,
- πολύ υψηλή παρουσία προβλημάτων αστάθειας, βάδισης και ιλίγγου,
- έντονο φόβο πτώσης,
- χαμηλή κινητικότητα,
- υψηλή συχνότητα επαναλαμβανόμενων πτώσεων, ανεξαρτήτως χώρου ή χρόνου.

Το συγκεκριμένο προφίλ μπορεί να διακριθεί ως: «Προφίλ υψηλού κινδύνου πτώσης, με γενικευμένη αστάθεια και επαναλαμβανόμενες πτώσεις».

5.3.4 Παρατηρήσεις και Συμπεράσματα

Η διαδικασία ερμηνείας των κλάσεων επέφερε αξιοπρεπείς διαχωρισμούς μεταξύ των γενικών χαρακτηρισμών που τις αντιπροσωπεύουν. Μερικά χρήσιμα στοιχεία που παρατηρήθηκαν και αξίζουν αναφοράς είναι τα εξής:

1. Απουσία αναφοράς στη μεταβλητή “Age” (η οποία αποτελεί τη μοναδική αντιπρόσωπο των δημογραφικών μεταβλητών, ύστερα από την επιλογή υποσυνόλου δεδομένων στο στάδιο της προεπεξεργασίας)

Πράγματι, σε κανένα προφίλ δεν εμφανίστηκε κάποιο ενδιαφέρον ως προς το χαρακτηριστικό ‘ηλικία’. Το παραπάνω συνεπάγεται και μέσω της συγκριτικής ανάλυσης των τιμών της κάθε κλάσης, όπου αποκαλύφθηκε ότι η μέση ηλικία για τις τρεις κλάσεις κυμαίνεται μεταξύ 69.80–71.10, γεγονός που δεν αποδίδει σημαντική διαφοροποίηση ως προς αυτό το χαρακτηριστικό.

2. Σημαντικότητα μεταβλητής “has_BloodTest”

Στο προφίλ κινδύνου 1 εμφανίζεται σχετικά χαμηλό ποσοστό διενέργειας εξέτασης αίματος (66%). Από την άλλη μεριά, στα προφίλ 0 και 2, η αντίστοιχη μεταβλητή λαμβάνει αρκετά υψηλότερη συχνότητα (87.5% και 93.5% αντίστοιχα). Όμως, η ερμηνεία των αποτελεσμάτων, υποδηλώνει ότι η αυξημένη παρουσία της εξέτασης στις δύο αυτές ομάδες ενδέχεται να απορρέει από διαφορετικά αίτια. Στο προφίλ 0 (χαμηλού κινδύνου), η υψηλή συχνότητα πιθανόν να σχετίζεται με το ότι τα άτομα αυτά κάνουν συχνά εξετάσεις αίματος στο πλαίσιο πρόληψης, ενώ στο προφίλ 2 (υψηλού κινδύνου) η συχνή διενέργεια των εξετάσεων φαίνεται να σχετίζεται περισσότερο με κλινική αναγκαιότητα και παρακολούθηση της κατάστασης των ηλικιωμένων.

3. Συνεισφορά μεταβλητής “HospitalADMISSIONS”

Το συγκεκριμένο χαρακτηριστικό φέρεται να αποτυπώνει, σε κάποιο βαθμό, την κλινική και λειτουργική κατάσταση των ηλικιωμένων που έχουν υποστεί πτώση, καθώς υψηλότερες τιμές, υποδηλώνουν αντίστοιχα επαναλαμβανόμενες πτώσεις και άρα πιθανώς σηματοδοτεί σε προβλήματα υγείας και μειωμένη λειτουργική ικανότητα. Ως εκ τούτου, το χαρακτηριστικό μπορεί να ερμηνευθεί ως δείκτης σοβαρότητας και επιβάρυνσης που συνεπάγονται τα πτωτικά επεισόδια των ηλικιωμένων.

Θα πρέπει να προστεθεί επίσης, η επιβεβαίωση της δήλωσης του άρθρου του dataset, το οποίο αναφέρει ότι οι πτώσεις στο χώρο του υπνοδωματίου συσχετίζεται με πτώσεις που συνέβησαν κατά τη διάρκεια της νύχτας.

Συνεπώς, η διαβάθμιση των τριών προφίλ κινδύνου σε χαμηλό, μέτριο και υψηλό επίπεδο αποτυπώνει μια συνεπή και προοδευτική κλιμάκωση της ευαισθησίας, τόσο σε λειτουργικό όσο και σε κλινικό επίπεδο. Το προφίλ χαμηλού κινδύνου χαρακτηρίζεται από διατηρημένη λειτουργική ικανότητα, περιορισμένη κλινική επιβάρυνση και κυρίως περιστασιακές και περιβαλλοντικές πτώσεις. Αντιθέτως, το προφίλ μέτριου κινδύνου ενσωματώνει επιβαρυντικούς παράγοντες όπως αυξημένη φαρμακευτική αγωγή, μειωμένη φυσική δραστηριότητα και νυχτερινό μοτίβο πτώσεων, υποδηλώνοντας μεταβατική ευαλωτότητα και μεγαλύτερη πιθανότητα επαναληπτικότητας πτώσεων. Τέλος, το προφίλ υψηλού κινδύνου συγκεντρώνει χαρακτηριστικά γενικευμένης αστάθειας, έντονης λειτουργικής αδυναμίας και συχνών νοσηλειών σε νοσοκομειακές εγκαταστάσεις, αποτυπώνοντας ένα πρότυπο αυξημένου κινδύνου. Η ιεράρχηση αυτή ευθυγραμμίζεται πλήρως με τα συναφή αποτελέσματα της ανάλυσης SHAP που παρουσιάστηκαν στην προηγούμενη ενότητα, καθώς τα χαρακτηριστικά με τη μεγαλύτερη συνεισφορά στις προβλέψεις του μοντέλου διαφοροποιούνται μεταξύ των κλάσεων, ενισχύοντας την ερμηνευσιμότητα και την κλινική εγκυρότητα των εξαγόμενων προφίλ κινδύνου.

5.4 Παρεμβάσεις πρόληψης

Μετά την ολοκλήρωση της ερμηνείας των κλάσεων και τη διατύπωσή τους σε κλινικά, περιβαλλοντικά και λειτουργικά κατανοητά προφίλ κινδύνου, το τελικό στάδιο της παρούσας υλοποίησης αφορά την παροχή εξατομικευμένων παρεμβάσεων πρόληψης από νέες πτώσεις. Κάθε προφίλ κινδύνου αντιμετωπίζεται ως διακριτή κατηγορία, στην οποία αντιστοιχίζονται στοχευμένες συστάσεις που αποσκοπούν στη μείωση της πιθανότητας μελλοντικών πτωτικών επεισοδίων.

Οι προτεινόμενες παρεμβάσεις δεν αποτελούν αυθαίρετες συστάσεις του μοντέλου, αλλά βασίζονται σε τεκμηριωμένες οδηγίες και πρακτικές που προέρχονται από έγκυρα και πιστοποιημένα επιστημονικά περιοδικά και επίσημους οργανισμούς πρόληψης πτώσεων [1 - 2], [102 - 112]. Με τον τρόπο αυτό διασφαλίζεται ότι οι παρεχόμενες παρεμβάσεις είναι σύμφωνες με τη διεθνή βιβλιογραφία και τις καθιερωμένες κλινικές πρακτικές. Η ενσωμάτωση τεκμηριωμένων πηγών προσδίδει αξιοπιστία στις προτάσεις και ενισχύει τη δυνατότητα υιοθέτησής τους σε πραγματικά περιβάλλοντα φροντίδας.

Θα πρέπει να σημειωθεί ότι, ότι όλες οι ομάδες κινδύνου, ανεξαρτήτως επιπέδου διαβάθμισης, εξακολουθούν να διατρέχουν πιθανότητα πτώσης. Συνεπώς, οι παρεμβάσεις δεν αντιμετωπίζονται ως απόλυτες ή δεσμευτικές οδηγίες, αλλά ως πλαίσιο υποστήριξης της κλινικής κρίσης. Οι προτάσεις δύνανται να τροποποιηθούν, να εξειδικευθούν ή να εμπλουτιστούν από τον εκάστοτε επαγγελματία υγείας που τις εφαρμόζει, λαμβάνοντας υπόψη τις ιδιαιτερότητες του ατόμου, το ιατρικό του ιστορικό, το περιβάλλον διαβίωσης και τους διαθέσιμους πόρους. Η προσέγγιση αυτή αναγνωρίζει τον κεντρικό ρόλο ειδικών όπως φυσικοθεραπευτές, ιατροί, και επαγγελματίες αποκατάστασης στον σχεδιασμό και την εφαρμογή αποτελεσματικών παρεμβάσεων για τον τρόπο συμπεριφοράς και τροποποίησης του χώρου διαμονής των ηλικιωμένων-ασθενών.

Για κάθε προφίλ κινδύνου έχουν διαμορφωθεί προκαθορισμένες οδηγίες πρόληψης, οι οποίες αντανακλούν τα κυρίαρχα χαρακτηριστικά της κάθε ομάδας. Οι οδηγίες αυτές λειτουργούν ως βασικό επίπεδο παρέμβασης σε ομαδικό επίπεδο, επιτρέποντας την ταχεία αντιστοίχιση προτάσεων σε άτομα με παρόμοια πρότυπα κινδύνου.

Συνολικά, το στάδιο της παροχής παρεμβάσεων συνιστά το πρακτικό αποτύπωμα της παρούσας προσέγγισης, καθώς μετασχηματίζει τα προφίλ κινδύνου σε συγκεκριμένες, κλινικά ερμηνεύσιμες και προσαρμόσιμες συστάσεις. Με τον τρόπο αυτό, η μελέτη της εργασίας δεν περιορίζεται στην πρόβλεψη ή

την κατηγοριοποίηση του κινδύνου πτώσης, αλλά υποστηρίζει την πρόληψη πτώσεων, προάγοντας μια εξατομικευμένη και τεκμηριωμένη στρατηγική φροντίδας για τον ηλικιωμένο πληθυσμό.

5.5 Επίλογος

Στο κεφάλαιο αυτό αναδείχθηκε ο ρόλος της ερμηνευσιμότητας ως προϋπόθεση για την αξιόπιστη αξιοποίηση μοντέλων ML σε εφαρμογές υγείας. Με τη χρήση των SHAP values πραγματοποιήθηκε η ποσοτικοποίηση της συμβολής των χαρακτηριστικών τόσο σε παγκόσμιο όσο και σε τοπικό επίπεδο, επιτρέποντας με αυτό τον τρόπο την κατανόηση των παραγόντων που καθορίζουν τις προβλέψεις του τελικού μοντέλου. Τα ευρήματα αυτά συνδέθηκαν άμεσα με την ερμηνευτική ανάλυση των προφίλ κινδύνου που προέκυψαν, επιβεβαιώνοντας τη σταδιακή κλιμάκωση από το χαμηλό προς το υψηλό επίπεδο. Μέσω των αναλύσεων, τα προφίλ κινδύνου μετασχηματίστηκαν σε κλινικά, λειτουργικά και περιβαλλοντικά ερμηνεύσιμα προφίλ. Ολοκληρώνοντας, το κεφάλαιο παρουσίασε τη σύνδεση της ερμηνείας των προφίλ κινδύνου με την παροχή εξατομικευμένων παρεμβάσεων πρόληψης δευτερευόντων πτώσεων των ηλικιωμένων ατόμων, βασισμένο σε έγκυρες βιβλιογραφικές πηγές.

Κεφάλαιο 6ο: Εφαρμογή Streamlit

Ολόκληρη η μελέτη και τα πειράματα που διαδραματίστηκαν στα προηγούμενα κεφάλαια και στάδια της εφαρμογής είχαν ως σκοπό την ανάδειξή τους σε ένα ολοκληρωμένο περιβάλλον για την αξιοποίηση του προτεινόμενου μοντέλου. Η ερμηνεία και η εννοιολογική αποτύπωση του μοντέλου διασφαλίστηκε δίνοντας έμφαση στη διαφάνεια και την αξιοπιστία των αποτελεσμάτων. Στο πλαίσιο αυτό, αναπτύχθηκε η εφαρμογή της παρούσας εργασίας, η οποία βασίστηκε στη βιβλιοθήκη ανοικτού κώδικα Streamlit (έκδοση 1.52) της Python [115]. Με την αξιοποίηση των εξαγόμενων αποτελεσμάτων των προηγούμενων πειραματικών φάσεων, το κεφάλαιο αυτό δύναται να συσσωρεύσει και παρουσιάσει το τελικό αποτέλεσμα της μελέτης, το οποίο είναι έτοιμο να χρησιμοποιηθεί για τους σκοπούς δημιουργίας του.

Το κεφάλαιο αρχικά οργανώνεται με την ανάδειξη της λειτουργικότητας της πλατφόρμας Streamlit για εφαρμογές όπως αυτή που παρουσιάζεται στη συγκεκριμένη εργασία. Παρακάτω αναλύεται η χρησιμότητα της πρόβλεψης του μοντέλου από τον χρήστη. Ακολούθως, ενδεικνύονται και περιγράφονται οι σελίδες της εφαρμογής που αναπτύχθηκαν μέσω της βιβλιοθήκης Streamlit, αναλύοντας κάθε αξιοσημείωτο στοιχείο της. Έπειτα προσφέρονται προτάσεις βελτίωσης και μελλοντικής χρήσης της εφαρμογής προκειμένου να εξελιχθεί σε επίπεδο λειτουργικότητας και αποτελεσματικότητας. Τέλος αποτυπώνεται η συνολική εικόνα του κεφαλαίου.

6.1 Λειτουργικότητα και Περιγραφή Βιβλιοθήκης Streamlit

Το Streamlit έχει σχεδιαστεί για την ταχεία ανάπτυξη διαδραστικών εφαρμογών ανάλυσης δεδομένων και μηχανικής μάθησης, επιτρέποντας τη μετατροπή υπολογιστικών ροών και πειραματικών σεναρίων σε λειτουργικές web εφαρμογές. Η λειτουργία του βασίζεται σε έναν απλό, δηλωτικό προγραμματιστικό μηχανισμό, όπου η διεπαφή χρήστη δημιουργείται δυναμικά μέσω Python εντολών. Η βασική του φιλοσοφία έγκειται στην άμεση σύνδεση της ανάλυσης δεδομένων με την παρουσίαση των αποτελεσμάτων. Κάθε μεταβολή στις εισόδους πεδία του χρήστη (π.χ. συμπλήρωση τιμών γνωρισμάτων) οδηγεί σε αυτόματη επανεκτέλεση της εφαρμογής, εξασφαλίζοντας ότι τα αποτελέσματα αντανακλούν πάντα την τρέχουσα κατάσταση των δεδομένων και του μοντέλου [115]. Στην εργασία, η δυνατότητα αυτή καθιστά εφικτή τη διαδραστική αξιοποίηση του εκπαιδευμένου μοντέλου, επιτρέποντας την εισαγωγή δεδομένων νέων ηλικιωμένων ατόμων από τους χρήστες (π.χ. φυσικοθεραπευτής) και την άμεση ανάθεση τους σε προφίλ κινδύνου, καθώς και την προβολή ερμηνευτικών αποτελεσμάτων.

Ένα από τα κύρια πλεονεκτήματα του Streamlit είναι η υψηλή ευχρηστία και η ταχύτητα ανάπτυξης εφαρμογών, γεγονός που το καθιστά ιδιαίτερα κατάλληλο για ερευνητικά και πιλοτικά έργα. Επιπλέον, παρέχει ενσωματομένη την δυνατότητα διασύνδεσης με βιβλιοθήκες ML και επιστήμης δεδομένων (όπως οι scikit-learn, numpy, pandas και SHAP), επιτρέποντας την άμεση ενσωμάτωση μοντέλων, μετρικών αξιολόγησης και εργαλείων ερμηνευσιμότητας. Παράλληλα, η δυνατότητα εύκολης ανάπτυξης και διαμοιρασμού της εφαρμογής διευκολύνει την επικοινωνία των αποτελεσμάτων σε μη τεχνικούς χρήστες.

Εντούτοις, η βιβλιοθήκη παρουσιάζει και ορισμένους περιορισμούς. Η αρχιτεκτονική της δεν έχει σχεδιαστεί για ιδιαίτερα σύνθετες ή βαριάς κλίμακας εφαρμογές, ενώ ο τρόπος διαχείρισης της κατάστασης των αντικειμένων της, καθίσταται λιγότερα ευέλικτος σε σύγκριση με web frameworks της γλώσσας Python. Επιπλέον, υπάρχουν περιορισμένες υπηρεσίες τροποποίησης της διεπαφής, γεγονός που μπορεί να αποτελέσει μειονέκτημα σε εφαρμογές με αυξημένες απαιτήσεις σχεδιασμού, αν και σε κάθε νέα έκδοσή της, εμπλουτίζεται με ολοένα και περισσότερα χρήσιμα εργαλεία.

Παρά τους περιορισμούς, το Streamlit κρίνεται ιδιαίτερα κατάλληλο για την παρούσα εργασία, καθώς εξυπηρετεί τον βασικό στόχο της, τη μεταφορά των αποτελεσμάτων της μελέτης σε ένα γρήγορο, λειτουργικό και κατανοητό εργαλείο υποστήριξης αποφάσεων.

6.2 Χρησιμότητα Πρόβλεψης από τον Χρήστη

Η εφαρμογή παρέχει στον χρήστη τη δυνατότητα εισαγωγής δεδομένων που αντιστοιχούν στα χαρακτηριστικά ενός νέου ατόμου, ακολουθώντας τη δομή και τους περιορισμούς του τελικού συνόλου δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου. Μέσα από την διεπαφή, ο χρήστης μπορεί να συμπληρώσει τις απαιτούμενες μεταβλητές, που αντιπροσωπεύουν λειτουργικές, κλινικές, ιστορικές και περιβαλλοντικές παραμέτρους σχετικά με τις πτώσεις. Τα δεδομένα εισόδου αποτελούν μία φόρμα εισαγωγής. Αυτά εισάγονται από την διεπαφή μέσω του χρήστη και υποβάλλονται στην ίδια διαδικασία προεπεξεργασίας που εφαρμόστηκε στο σύνολο εκπαίδευσης, διασφαλίζοντας τη συνέπεια της πρόβλεψης και την εγκυρότητα της εξαγόμενης εκτίμησης κινδύνου. Έτσι, το σύστημα εγγυάται την αφοσίωσή του στην μεθοδολογία της εργασίας που ακολουθήθηκε, αποφεύγοντας την εισαγωγή μεροληψιών λόγω ασυμβατότητας των δεδομένων εισόδου.

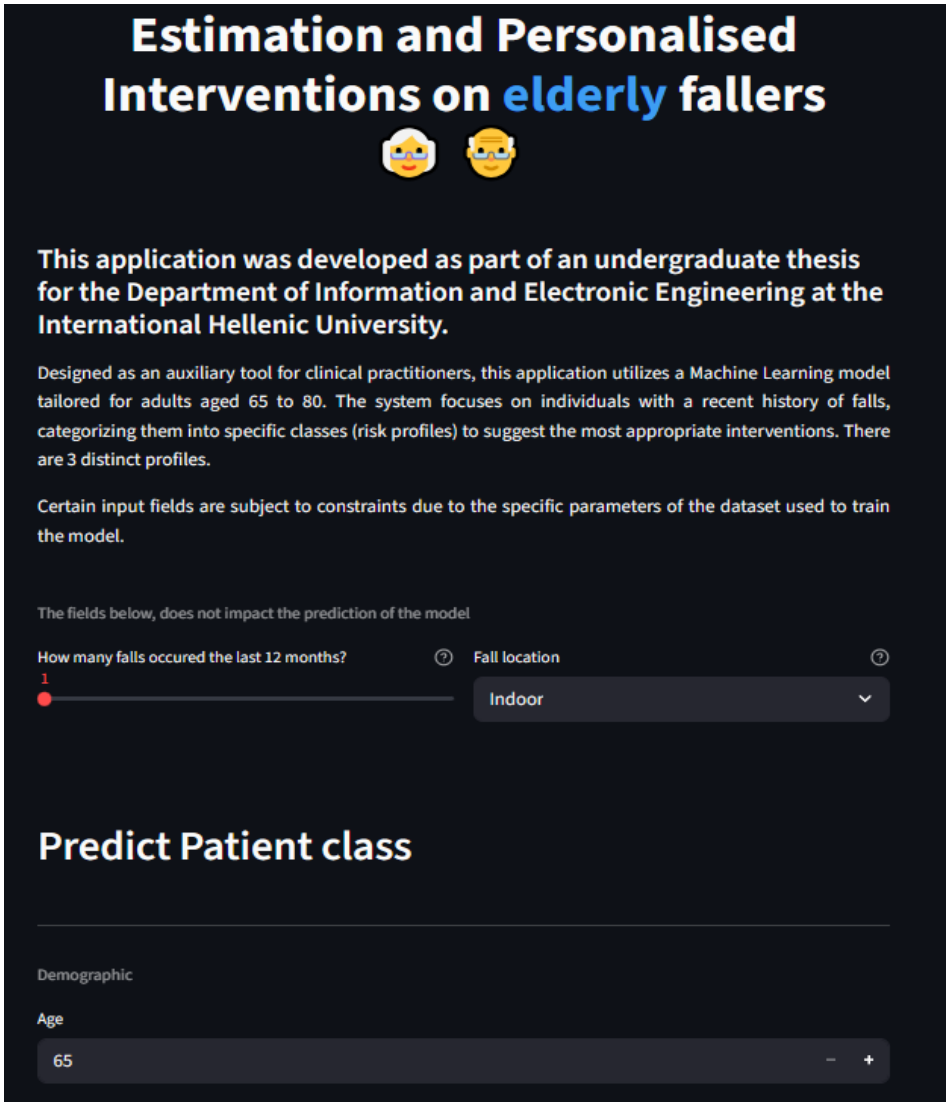
Αφού πραγματοποιηθεί η υποβολή της φόρμας δεδομένων, το εκπαιδευμένο μοντέλο ταξινόμησης χρησιμοποιείται για την πρόβλεψη του προφίλ κινδύνου στο οποίο εντάσσεται το νέο άτομο. Το αποτέλεσμα παρουσιάζεται με σαφή και κατανοητό τρόπο, επιτρέποντας στον χρήστη να αντιληφθεί άμεσα το επίπεδο κινδύνου που αποδίδεται, μέσα από τεκμηρίωση των αποτελεσμάτων του. Παράλληλα, η εφαρμογή παρέχει συμπληρωματικές πληροφορίες και οδηγίες/συστάσεις πρόληψης πτώσης για τον ηλικιωμένο, με βάση το προφίλ κινδύνου που αναγνωρίστηκε και συγκεκριμένα χαρακτηριστικά του χώρου πτώσης/εων. Η διαδικασία αυτή ενισχύει τον υποστηρικτικό ρόλο της εφαρμογής στη λήψη αποφάσεων, χωρίς όμως να υποκαθιστά τη κρίση ενός ειδικού, αλλά λειτουργώντας ως εργαλείο συμπληρωματικής αξιολόγησης (πρόβλεψης κινδύνου) και παροχής συστάσεων στο πλαίσιο της πρόληψης πτώσεων.

6.3 Γραφική Παρουσίαση

Η γραφική διεπαφή της εφαρμογής σχεδιάστηκε με στόχο την καθαρή, δομημένη και κλινικά κατανοητή παρουσίαση των αποτελεσμάτων του προτεινόμενου μοντέλου, αξιοποιώντας τις βέλτιστες πρακτικές του εγγράφου της εφαρμογής Streamlit [115]. Αρχικά, η εφαρμογή οργανώνεται σε τέσσερις διακριτές σελίδες καθεμία από τις οποίες εξυπηρετεί διαφορετικό λειτουργικό σκοπό και απευθύνεται σε συγκεκριμένες ανάγκες του χρήστη. Η διάρθρωση αυτή επιτρέπει τη σαφή διάκριση των θεματικών πρόβλεψης, ερμηνείας, συνοπτικής ανάλυσης και τεκμηρίωσης, ενισχύοντας τη χρηστικότητα και τη διαφάνεια της εφαρμογής.

6.3.1 Κύρια Σελίδα Εφαρμογής

Το πρώτο και βασικότερο τμήμα της εφαρμογής σχετίζεται με την πρόβλεψη του κινδύνου πτώσης, που αφορά την κεντρική σελίδα της. Ο χρήστης καλείται να συμπληρώσει τα πεδία που αντιστοιχούν στα χαρακτηριστικά εισόδου του μοντέλου, τα οποία αντικατοπτρίζουν λειτουργικούς, κλινικούς και περιβαλλοντικούς (σχετικά με την κάθε περίπτωση πτώσης) παράγοντες, όπως φαίνεται στο σχήμα 6.1.

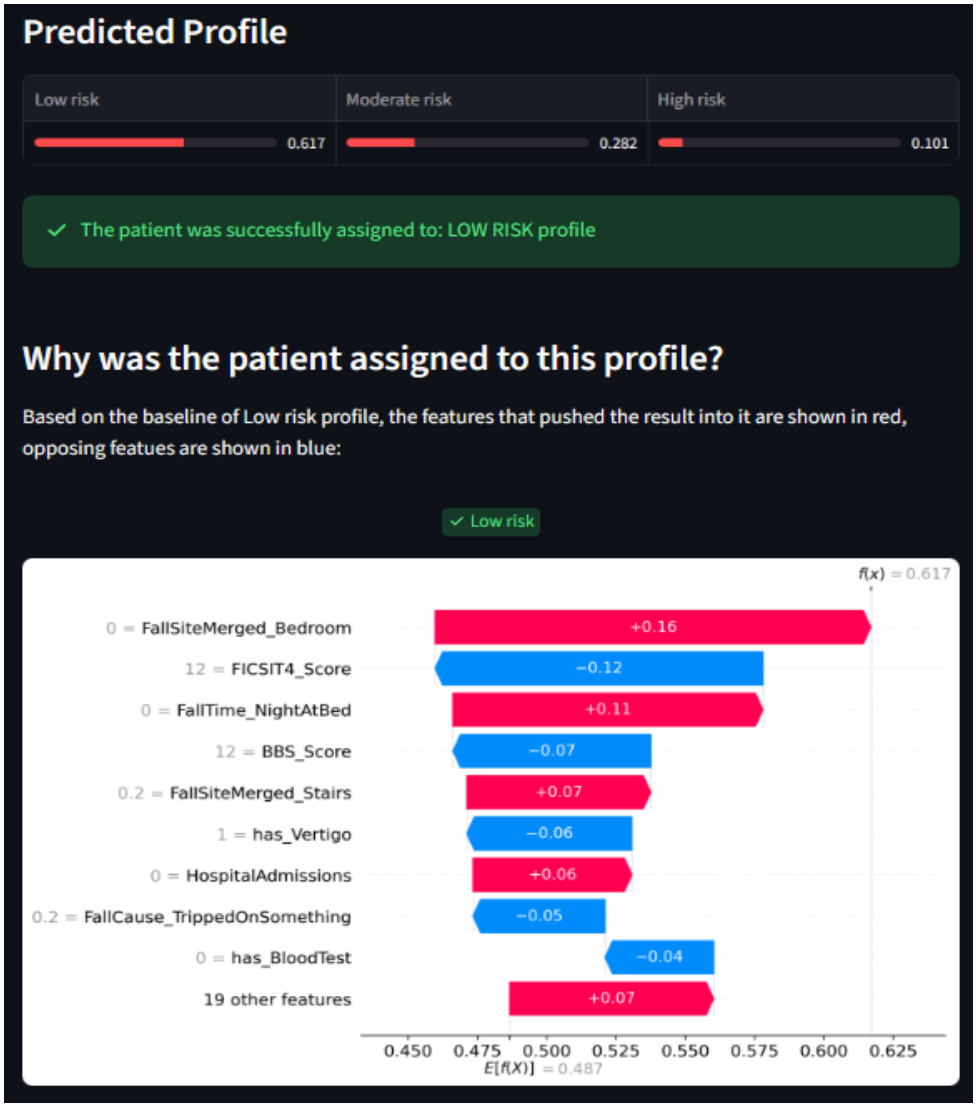


Σχήμα 6.1: Κεντρική σελίδα εφαρμογής Streamlit

Τα πρώτα δύο πεδία, όπως αναγράφεται και στη λεζάντα (που απεικονίζεται με γκρι γράμματα), δεν επηρεάζουν ούτε εισάγονται ως δεδομένα στην πρόβλεψη του μοντέλου. Αντιθέτως, το αριστερό πεδίο, προσδιορίζει το πλήθος των πτώσεων που έχει υποστεί ο ασθενής, προκειμένου να δημιουργηθεί αντίστοιχος αριθμός πεδίων σχετικών με το περιβάλλον και το συμβάν της/των πτώσης/εων. Το δεξί πεδίο αφορά τον τύπο της τοποθεσίας πτώσης. Λαμβάνει τρεις πιθανές τιμές, οι οποίες είναι οι “εσωτερικού χώρου”, “εξωτερικού χώρου”, “και τα δύο”. Ειδικότερα, η επιλογή της επηρεάζει τις ειδικές συστάσει που θα προσφερθούν στο τέλος της διαδικασίας πρόβλεψης για το άτομο.

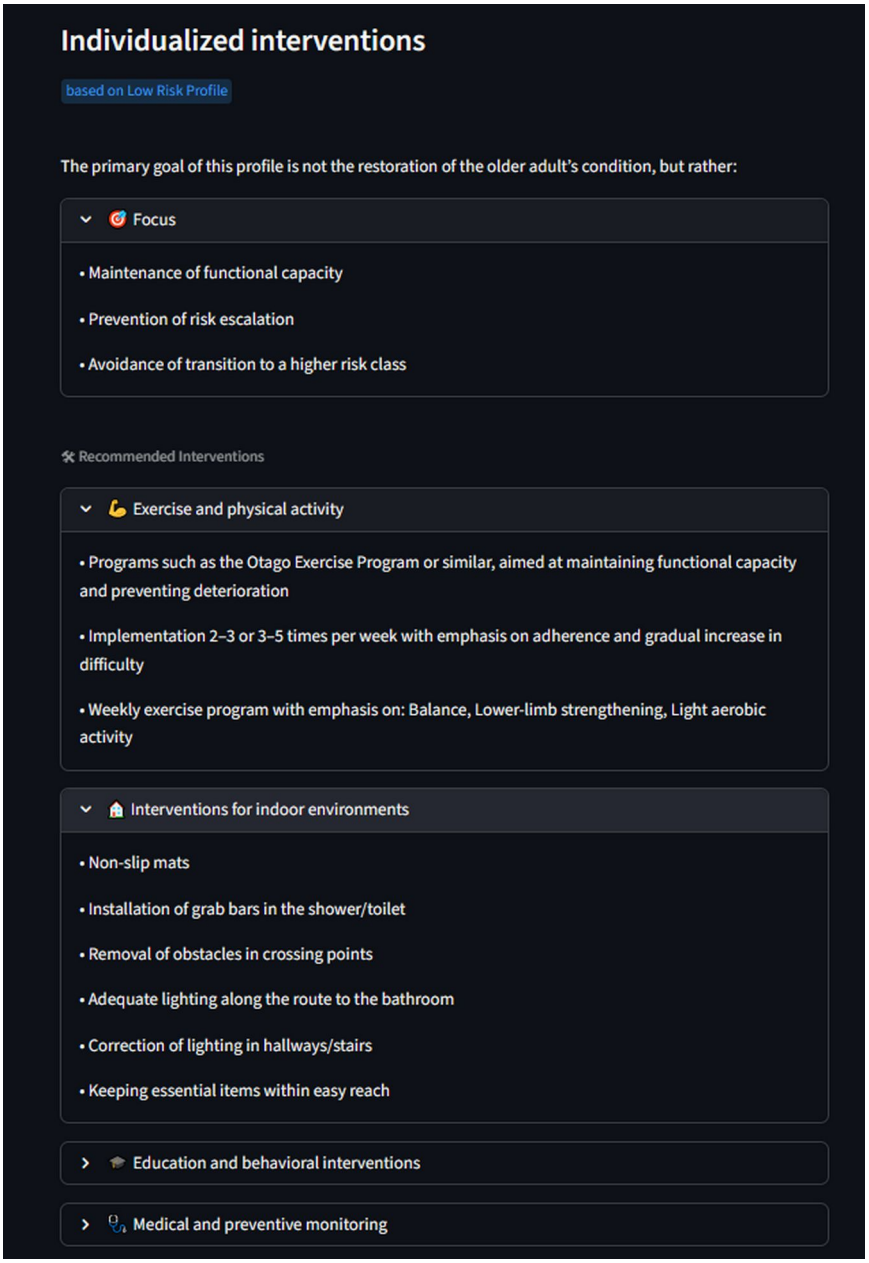
Αφού συμπληρωθούν όλα τα πεδία πρόβλεψης και ο χρήστης ενεργοποιήσει το κουμπί πρόβλεψης, τότε το εκπαιδευμένο μοντέλο ταξινόμησης (SVC) αποδίδει πιθανοτικά την ένταξη του ατόμου σε ένα από τα τρία προφίλ κινδύνου. Η τελική κλάση επιλέγεται βάσει της μέγιστης εκτιμώμενης πιθανότητας. Να σημειωθεί ότι, σε περίπτωση που κάποιο πεδίο δεν έχει συμπληρωθεί, τότε δεν γίνεται πρόβλεψη αλλά εμφανίζεται ένα μήνυμα προτροπής συμπλήρωσης του αντίστοιχου πεδίου εισόδου. Ταυτόχρονα με την ανακοίνωση του προφίλ κινδύνου, παρουσιάζεται η αιτιολόγηση της απόφασης μέσω τοπικής ερμηνείας χαρακτηριστικών με χρήση SHAP values σε μορφή διαγράμματος τύπου waterfall (βλέπε σχήμα 6.2). Στην περίπτωση που η διαφορά της μέγιστης εκτιμώμενης πιθανότητας κλάσης από οποιαδήποτε από τις υπόλοιπες, είναι

μικρότερη από 10%, τότε αποδίδεται αυτόματα η τοπική ερμηνεία για κάθε κλάση απ' αυτές, έτσι ώστε ο χρήστης να κατανοήσει τους λόγους που βρέθηκαν τόσο κοντά οι πιθανότητες των κλάσεων, και να συνειδητοποιήσει την αιτία που αποδόθηκε ως μέγιστη η τελική.



Σχήμα 6.2: Παράδειγμα αποτελέσματος πρόβλεψης με ερμηνεία μέσω της εφαρμογής Streamlit

Επιπροσθέτως, ανάλογα με τα χαρακτηριστικά του ατόμου (κυρίως περιβαλλοντικά στοιχεία των πτώσεων) και το προφίλ στο οποίο εντάσσεται, παρέχονται προκαθορισμένες συστάσεις πρόληψης, οργανωμένες σε θεματικές κατηγορίες, με σκοπό την υποστήριξη της κλινικής λήψης αποφάσεων από τους ειδικούς ιατρούς. Συγχρόνως, παρουσιάζονται τα βασικά σημεία γύρω από τα οποία επικεντρώνονται οι συστάσεις πρόληψης, ως προς το αντίστοιχο προφίλ κινδύνου. Οι προτάσεις αυτές, έχουν συζητηθεί στην ενότητα 5.4 ως προς την επιλογή τους. Για την διευκόλυνση του χρήστη, μπορεί να επεκτείνει ή να συστέλλει τις θεματικές των συστάσεων, προκειμένου να επικεντρωθεί σε αυτή που τον ενδιαφέρει, όπως φαίνεται στο παρακάτω σχήμα.

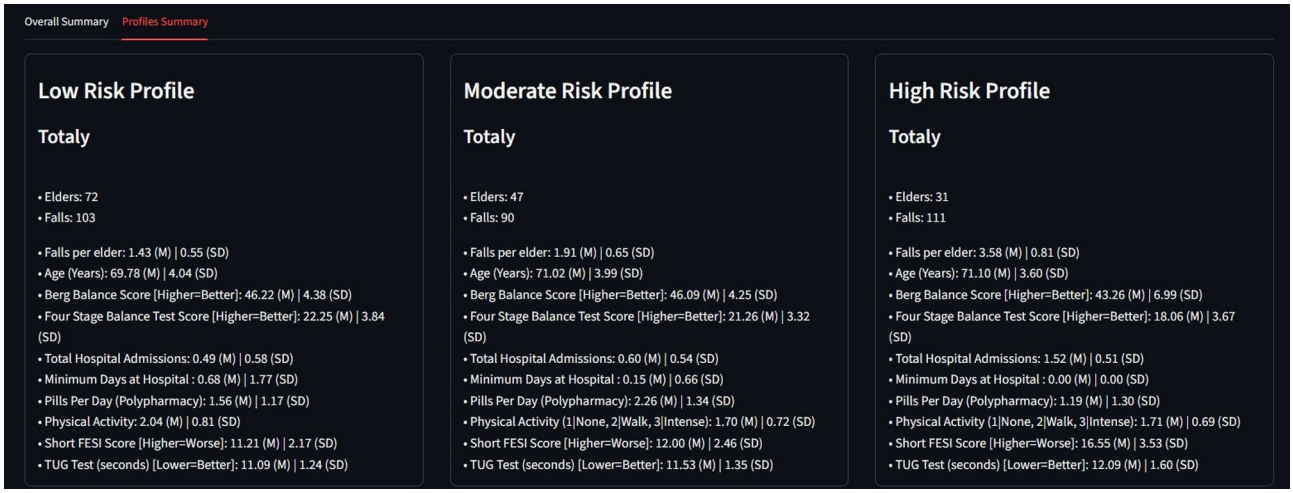


Σχήμα 6.3: Απεικόνιση ενδεικτικών συστάσεων μέσω της εφαρμογής Streamlit

6.3.2 Σελίδα Συνοπτικής Ανάλυσης

Η δεύτερη σελίδα της εφαρμογής, παρέχει μια συνολική εικόνα τόσο των προφίλ κινδύνου όσο και του τελικού συνόλου δεδομένων (βλέπε σχήμα 6.4). Σε αυτή παρουσιάζονται στατιστικά μέτρα όπως μέσες τιμές και τυπικές αποκλίσεις για τα αριθμητικά χαρακτηριστικά, καθώς και ποσοστά συμμετοχής για τα κατηγορικά γνωρίσματα. Παράλληλα, εμφανίζονται τα διαγράμματα SHAP bar plots που αναλύθηκαν στο προηγούμενο κεφάλαιο επιτρέποντας την κατανόηση της παγκόσμιας σημαντικότητας των χαρακτηριστικών για κάθε προφίλ κινδύνου (σχήμα 6.5). Η σελίδα αυτή λειτουργεί ως εργαλείο ερμηνευτικής επισκόπησης, ιδιαίτερα χρήσιμο για τον χρήστη, είτε είναι νέος και ενδιαφέρεται να μάθει τις ιδιαιτερότητες του μοντέλου, είτε γνωρίζει την εφαρμογή αλλά επιθυμεί να κάνει μία ανασκόπηση. Επομένως, προσφέρεται στον χρήστη

μια γενική, όμως ουσιαστική εικόνα της συμπεριφοράς του μοντέλου και των παραγόντων που επηρεάζουν περισσότερο τις προβλέψεις του.



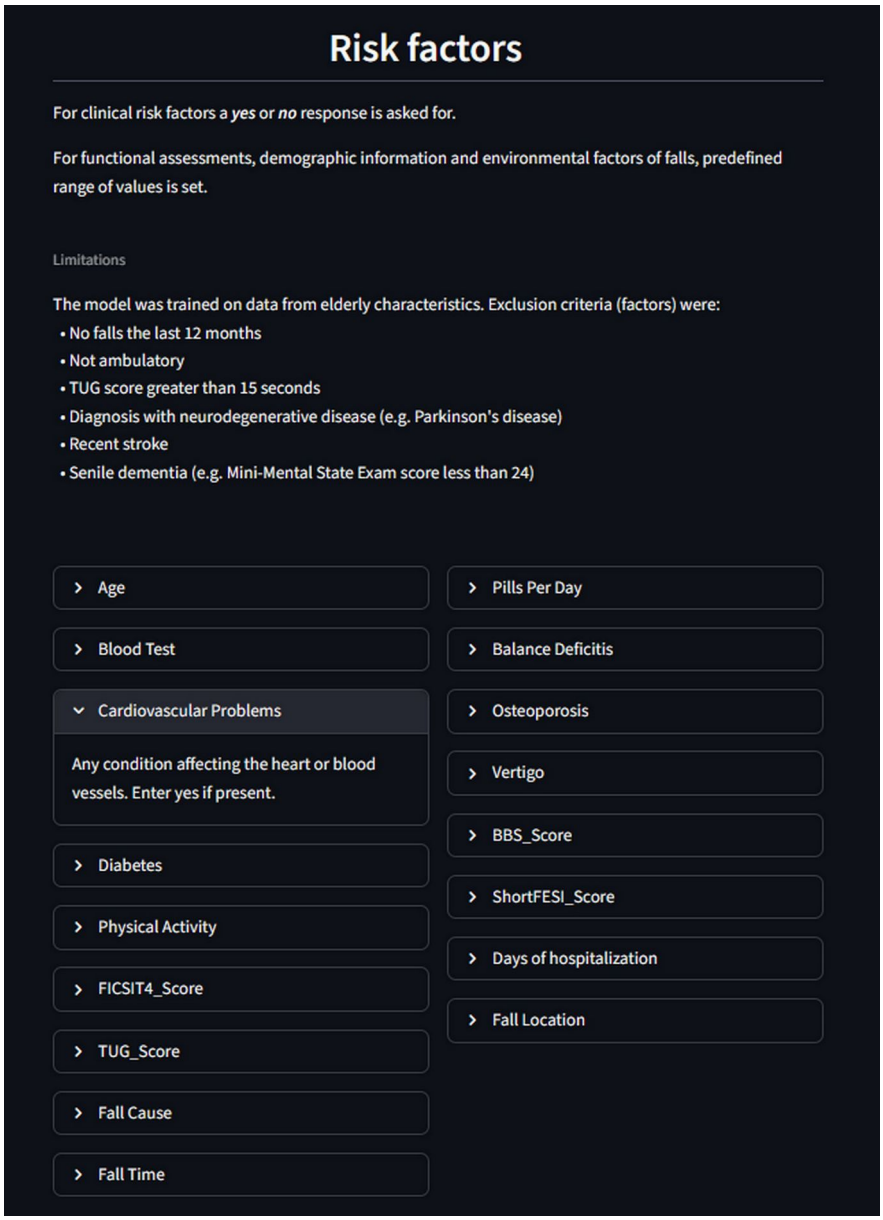
Σχήμα 6.4: Απόσπασμα στατιστικών μέτρων ανά προφίλ κινδύνου

Σχήμα 6.5: Απεικόνιση SHAP bar plots ανά προφίλ κινδύνου στην εφαρμογή στην εφαρμογή Streamlit

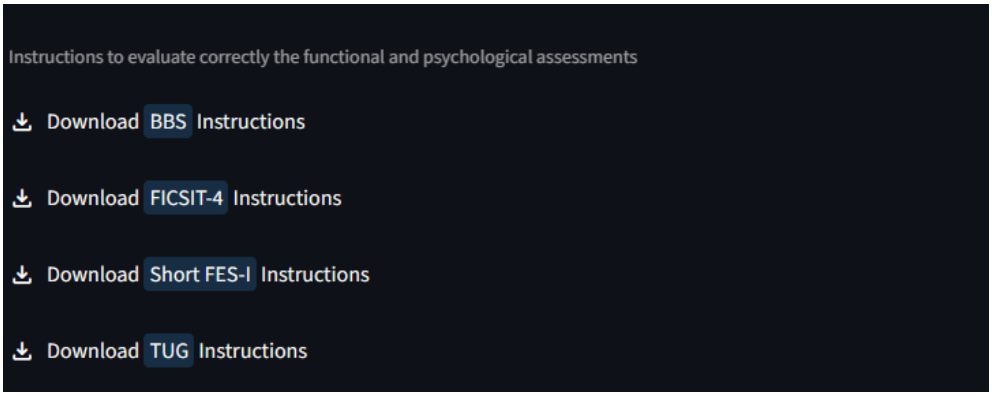
6.3.3 Σελίδα Παραγόντων Κινδύνου

Η τρίτη σελίδα, επικεντρώνεται στην τεκμηρίωση των παραγόντων κινδύνου και την εγκυρότητα της εφαρμογής. Περιλαμβάνει επεξηγήσεις για κάθε πεδίο εισόδου της εφαρμογής (ως εκ τούτου και του μοντέλου), καθώς και σαφή αναφορά στους περιορισμούς που αφορούν τον πληθυσμό για τον οποίο η εφαρμογή θεωρείται έγκυρη (σχήμα 6.6). Επιπλέον, παρέχεται η δυνατότητα λήψης αρχείων που σχετίζονται

με τις λειτουργικές δοκιμασίες και τις αξιολογήσεις που χρησιμοποιήθηκαν στην εργασία, ενισχύοντας τη διαφάνεια και τη δυνατότητα κατανόησης της μεθοδολογίας από τον χρήστη (σχήμα 6.7). Το τμήμα αυτό της εφαρμογής Streamlit παρέχει στον ιατρό ή φυσικοθεραπευτή πλήρη πληροφόρηση για τη σωστή και υπεύθυνη χρήση της.



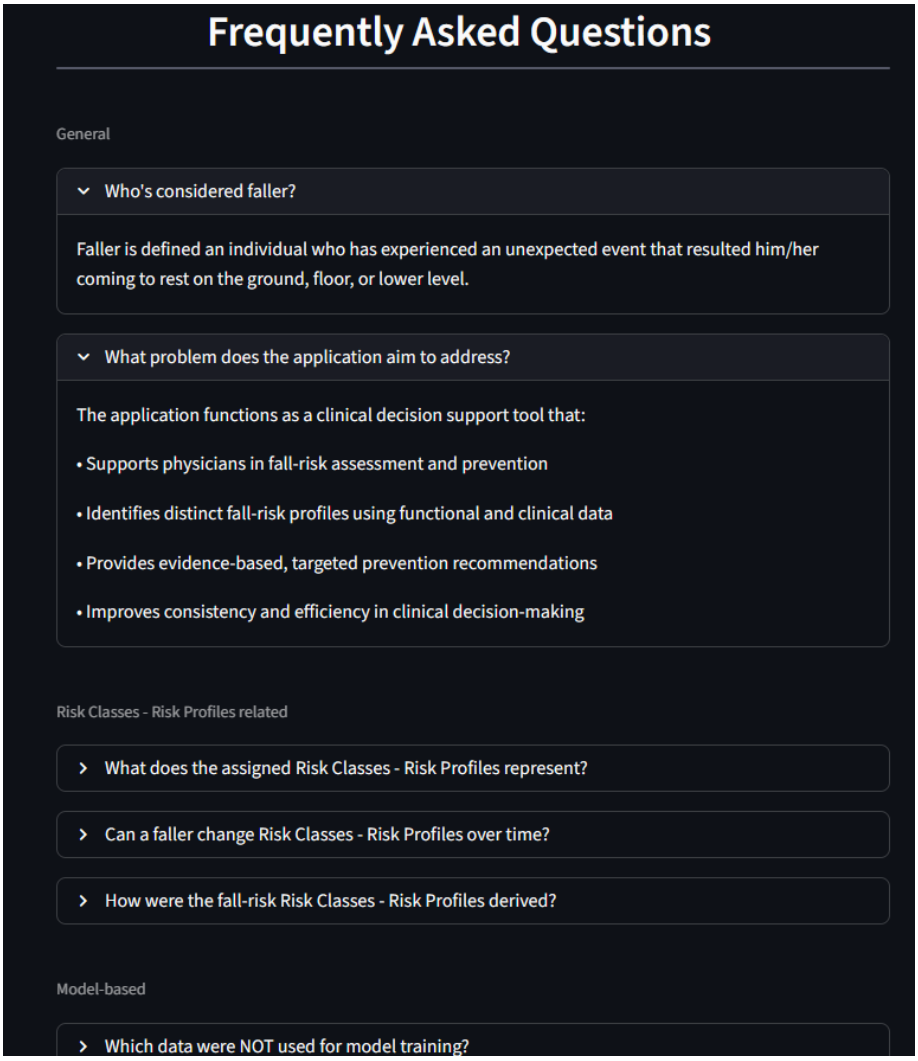
Σχήμα 6.6: Απεικόνιση σελίδας "Παραγόντων Κινδύνου" της εφαρμογής Streamlit (1)



Σχήμα 6.7: Απεικόνιση σελίδας "Παραγόντων Κινδύνου" της εφαρμογής Streamlit (2)

6.3.4 Σελίδα FAQ

Η σελίδα “FAQ” (Frequently Asked Questions), συγκεντρώνει απαντήσεις συχνών ερωτημάτων που μπορεί να παρουσιάζονται στους χρήστες σχετικά με τη λειτουργία της εφαρμογής, τη διαδικασία ανάπτυξης του μοντέλου και της εγκυρότητάς της. Ενδεικτικό απόσπασμα εμφανίζεται στο σχήμα 6.8.



Σχήμα 6.8: Σελίδα FAQ εφαρμογής Streamlit

6.4 Προτάσεις Βελτίωσης και Μελλοντική Χρήση

Η εφαρμογή που αναπτύχθηκε, αποτέλεσε το ορόσημο της εργασίας, επιτυγχάνοντας τον βασικό στόχο της υποστήριξης της πρόβλεψης και της ερμηνείας του κινδύνου πτώσης σε νέα ηλικιωμένα άτομα. Παρ' όλα αυτά, υπάρχουν σημαντικά περιθώρια περαιτέρω βελτίωσης και επέκτασης της λειτουργικότητάς της.

Ένα σημαντικό βήμα επέκτασης της εφαρμογής είναι ο εμπλουτισμός του μοντέλου με νέα δεδομένα, ώστε να καταστεί δυνατή η παρακολούθηση της εξέλιξης του κινδύνου των ατόμων σε βάθος χρόνου. Με αυτό το τρόπο, η εφαρμογή θα επιτρέπει την πρόβλεψη μεταβολών του προφίλ κινδύνου του ατόμου, ενισχύοντας το προληπτικό χαρακτήρα της. Διαφορετικά, ένα άλλο βήμα επέκτασής της είναι η χρήση ενός νέου συνόλου δεδομένων το οποίο να περιλαμβάνει μεγαλύτερου μεγέθους πολυπαραγοντικά δεδομένα, το οποίο να μην εστιάζει απόλυτα σε ηλικιωμένο πληθυσμό, αλλά να εφαρμόζεται και σε μικρότερες ηλικιακές ομάδες, συμμορφώνοντας, ευαισθητοποιώντας και εκπαιδεύοντάς τες, προτού φθάσουν στη τελευταία ηλικιακή κλίμακα.

Επιπλέον, μία νέα προσθήκη για την εφαρμογή θα ήταν η επιλογή προκαθορισμένης τιμής για την εισαγωγή τιμής σε κάποια λειτουργική αξιολόγηση (π.χ. TUG), όπως η μέση τιμή των δεδομένων εκπαίδευσης για την εκάστοτε αξιολόγηση. Αυτό θα διευκολύνει τον χρήστη, σε περίπτωση που δεν είναι δυνατή η διενέργεια

μίας ή περισσότερων τεχνικής αξιολόγησης, να εισάγει μία προκαθορισμένη τιμή ώστε να μπορεί να αξιοποιήσει το μοντέλο με τη μεγαλύτερη δυνατή αποφυγή μεροληψίας.

Από πλευράς λειτουργικότητας ακόμη, μελλοντικές εκδόσεις της εφαρμογής θα μπορούσαν να ενισχύσουν τον βαθμό εξατομίκευσης των παρεμβάσεων πρόληψης, συνδυάζοντας τα προφίλ κινδύνου με πρόσθετες πληροφορίες και παρέχοντας πιο σύνθετες και εξατομικευμένες παρεμβάσεις με βάση το ίδιο το άτομο. Ακόμη, η παροχή διαγραμμάτων σύγκρισης πολύπλοκων εννοιών των χαρακτηριστικών του κάθε προφίλ, θα μπορούσε να συμβάλει στην κατανόηση των ιδιαιτεροτήτων τους.

Καίριας σημασίας για την βελτίωση της εφαρμογής αποτελεί η αξιολόγησή της από τους τελικούς χρήστες, τους οποίους προορίζεται να αξιοποιηθεί. Μέσα από αυτή τη διαδικασία, δύναται να αποδοθεί ουσιαστική ανατροφοδότηση σε θεωρητικό, λειτουργικό και γραφικό επίπεδο, η οποία θα οδηγήσει στην εξέλιξή της. Μέσω της ανατροφοδότησης αυτής, η εφαρμογή μπορεί να διασφαλίσει ένα αξιόπιστο και διαφανές περιβάλλον παροχής πληροφορίας, ενισχύοντας τόσο την εγκυρότητα των αποτελεσμάτων όσο και την εμπιστοσύνη των χρηστών της.

Σε επίπεδο μελλοντικής χρήσης, η εφαρμογή μπορεί να χρησιμοποιηθεί ως εκπαιδευτικό εργαλείο για την ευαισθητοποίηση επαγγελματιών υγείας και φοιτητών σχετικά με τους παράγοντες κινδύνου πτώσεων και τη σημασία της ερμηνευσιμότητας σε συστήματα ΤΝ με χρήση ML. Ευρύτερα, μπορεί να αποτελέσει τη βάση για πιλοτικές εφαρμογές σε δομές πρωτοβάθμιας φροντίδας ή αποκατάστασης. Με όλα τα παραπάνω, η εφαρμογή έχει τη δυναμική να εξελιχθεί σε ένα ευέλικτο και επεκτάσιμο σύστημα, εστιάζοντας πάντα στη βελτίωση της πρόληψης πτώσεων και της ποιότητας ζωής των ανθρώπων.

6.5 Επίλογος

Στο παρόν κεφάλαιο παρουσιάστηκε η ανάπτυξη και η υλοποίηση της εφαρμογής που βασίστηκε στα αποτελέσματα της παρούσας μελέτης. Αρχικά, περιγράφηκε η βιβλιοθήκη Streamlit και τεκμηριώθηκε η καταλληλότητά της για την ανάπτυξη της παρούσας εφαρμογής, λαμβάνοντας υπόψη τα πλεονεκτήματα και τους περιορισμούς της. Στη συνέχεια, αναδείχθηκε η χρησιμότητα της πρόβλεψης από τον χρήστη, μέσω της οποίας καθίσταται δυνατή η ένταξη νέων ατόμων σε προφίλ κινδύνου, με ταυτόχρονη παροχή ερμηνευτικών στοιχείων και εξειδικευμένων προτάσεων πρόληψης πτώσεων. Παράλληλα, η γραφική παρουσίαση των αποτελεσμάτων και λειτουργιών της εφαρμογής, επέτρεψαν τη σαφή απεικόνιση της λειτουργίας του μοντέλου και την τεκμηρίωση και σύνοψη των παραγόμενων προφίλ κινδύνου.

Το κεφάλαιο ολοκληρώθηκε με την παρουσίαση προτάσεων βελτίωσης και μελλοντικής χρήσης της εφαρμογής, επισημαίνοντας τις δυνατότητες επέκτασης, του εμπλουτισμού των δεδομένων και της ενίσχυσης της αξιοπιστίας μέσω αξιολόγησης από τελικούς χρήστες. Το επόμενο και τελευταίο κεφάλαιο της εργασίας, εστιάζει στη συζήτηση των συνολικών ευρημάτων της μελέτης, εξετάζοντας τα συμπεράσματα που προκύπτουν, τους περιορισμούς της προτεινόμενης προσέγγισης και τις προοπτικές μελλοντικής βελτίωσης.

Κεφάλαιο 7ο: Συζήτηση

7.1 Συνοπτικά Ευρήματα

Στα πλαίσια της παρούσας διπλωματικής εργασίας, πραγματοποιήθηκε εκτεταμένη μελέτη γύρω από το πεδίο πτώσεων ηλικιωμένων ατόμων, εφαρμόζοντας τεχνικές ML για την ανίχνευση και ερμηνεία προτύπων κινδύνου πτώσεων και τελικώς την παροχή προτάσεων πρόληψης. Η προεπεξεργασία δεδομένων, αποτέλεσε το πρώτο και πιο χρονοβόρο τμήμα της εργασίας, δεδομένου ότι καθιέρωσε τη βάση της, καθώς και λόγω της απουσίας ετικέτας στο επιλεγμένο σύνολο δεδομένων. Στη συνέχεια, μέθοδοι μη εποπτευόμενης συσταδοποίησης και εποπτευόμενης ταξινόμησης χρησιμοποιήθηκαν για την παραγωγή ψευδο-ετικετών, μέσω της στρατηγικής συσταδοποίησης και μεταγενέστερα ταξινόμησης. Η προσέγγιση αυτή, συνέβαλε στην ομαδοποίηση των δεδομένων ως προς τα κοινά τους χαρακτηριστικά, διαμορφώνοντας τρία διακριτά προφίλ κινδύνου. Τα προφίλ, χαρακτηρίστηκαν για την διαφοροποίησή τους ως προς τη λειτουργική ικανότητα, την κλινική επιβάρυνση, τα μοτίβα πτώσεων και τη συχνότητα επαναληπτικών συμβάντων πτώσεων, γεγονός που ενισχύει τη μεθοδολογική εγκυρότητα της προσέγγισης. Η επιλογή του αλγορίθμου K-Means για τη συσταδοποίηση και του SVC για τη ταξινόμηση αποδείχθηκε αποτελεσματική για την ανάδειξη καλής γενίκευσης σε άγνωστα δεδομένα.

Η ανωτέρω διαδικασία ενισχύθηκε χάρη στην ερμηνεία που εφαρμόστηκε στα προφίλ κινδύνου. Η διαβάθμιση σε χαμηλό, μέτριο και υψηλό επίπεδο κινδύνου συγκροτεί ένα κατανοητό πλαίσιο υποστήριξης αποφάσεων, λειτουργώντας ως σημείο αναφοράς για τη διαφανή και αξιόπιστη κατανόηση της ιδεολογίας των αποτελεσμάτων από τους επαγγελματίες υγείας στο πλαίσιο πρόληψης πτώσεων. Η ερμηνεία των ευρημάτων ανέδειξε την κρισιμότητά της για την διαμόρφωση κατάλληλων και στοχευμένων συστάσεων ανά προφίλ κινδύνου, πλαισιώνοντας και ενισχύοντας περαιτέρω τον ρόλο της υποστήριξης λήψης αποφάσεων. Αυτό ακολούθως, επέτρεψε την ανάπτυξη μίας διαδραστικής εφαρμογής οπτικοποίησης και αλληλεπίδρασης με το μοντέλο ταξινόμησης SVC, διατηρώντας την ευχρηστία, την ευελιξία και την απαραίτητη ερμηνεία των χαρακτηριστικών που καλείται να γνωρίζει ο χρήστης της εφαρμογής.

Συνολικά, τα ευρήματα υποστηρίζουν ότι η συνδυαστική χρήση εποπτευόμενων και μη εποπτευόμενων τεχνικών ML, σε συνδυασμό με τη χρήση εργαλείων ερμηνείας των αποτελεσμάτων, μπορεί να διατελέσει προστιθέμενη αξία στη κατανόηση και την πρόληψη των πτώσεων.

7.2 Περιορισμοί και Προκλήσεις

Κατά τη διαδικασία έρευνας, μελέτης και ανάπτυξης της εργασίας παρουσιάστηκαν ορισμένοι περιορισμοί που σχετίζονται τόσο με τα δεδομένα όσο και με τη μεθοδολογική προσέγγιση που οφείλουν να ληφθούν υπόψη κατά την ερμηνεία των ευρημάτων. Πρώτα απ' όλα, το σύνολο δεδομένων χαρακτηρίζεται από σχετικά περιορισμένο μέγεθος και απουσία διαχρονικής πληροφορίας, γεγονός που περιορίζει τη δυνατότητα γενίκευσης των αποτελεσμάτων σε ευρύτερους πληθυσμούς ηλικιωμένων. Χρειάστηκε να εξεταστεί το επίπεδο της εικόνας του dataset, είτε μετασχηματίζοντάς το σε επίπεδο ατόμου, είτε διατηρώντας το σε επίπεδο πτώσης. Η επιλογή μετασχηματισμού του, αποτελεί μία από τις πιο σημαντικές προκλήσεις της εργασίας, διότι αφενός αποφεύχθηκε η περαιτέρω πιθανή μεροληψία γνωρισμάτων λόγω επικράτησης, αφετέρου τα στιγμιότυπα μειώθηκαν περίπου στο μισό των αρχικών (μόλις 150) Σε συνδυασμό με το υψηλό αριθμό χαρακτηριστικών αυξάνεται ο κίνδυνος υπερπροσαρμογής. Παρά τις τεχνικές μείωσης διαστάσεων, επιλογής χαρακτηριστικών και στρωματοποιημένης διασταυρωμένης επικύρωσης, η επίδραση της υψηλής διαστασιμότητας δεν μπορεί να αποκλειστεί πλήρως. Παράλληλα, γίνανε προσπάθειες συγχώνευσης του dataset με άλλα παρόμοια, όμως η πολυτροπική φύση των δεδομένων,

ιδιαίτερα οι λειτουργικές και ψυχολογικές αξιολογήσεις, κατέστησαν ιδιαίτερα δύσκολη την εύρεση παρόμοιων συνόλων δεδομένων, τα οποία ακόμη να σχετίζονται με δεδομένα ηλικιωμένων ατόμων.

Επιπλέον, η έλλειψη άμεσων ετικετών κατέστησε αναγκαία την εξαγωγή και αργότερα χρήση ψευδο-ετικετών που προέκυψαν μέσω συσταδοποίησης. Η τεχνική αυτή, αν και επέτρεψε τη μοντελοποίηση του κινδύνου σε συνθήκες πραγματικής αβεβαιότητας, εισάγει ταυτόχρονα ένα βαθμό εξάρτησης από τους αλγορίθμους συσταδοποίησης. Συνεπώς, τα προφίλ κινδύνου θα πρέπει να ερμηνεύονται ως υπολογιστικές αναπαραστάσεις προτύπων δεδομένων και όχι ως απόλυτες κλινικές κατηγορίες.

Επιπροσθέτως, σε επίπεδο μοντέλων και ερμηνευσιμότητας, οι περιορισμοί σχετίζονται με τις παραδοχές που ισχύουν στους αλγορίθμους που χρησιμοποιήθηκαν. Για παράδειγμα, η συσταδοποίηση εξαρτάται από την επιλογή παραμέτρων και μετρικών απόστασης. Το γεγονός αυτό, επηρέασε τη δομή και τον αριθμό των εξαγόμενων ομάδων, ακόμη και αν αξιολογήθηκαν τα αποτελέσματα των τεχνικών. Αντίστοιχα, τα SHAP values παρότι προσφέρουν ένα ισχυρό πλαίσιο ερμηνείας, ενδέχεται να μην αντανακλούν πλήρως την πολυπλοκότητα των πραγματικών κλινικών δεδομένων. Για τους ανωτέρω λόγους, τα αποτελέσματα θα πρέπει να αξιοποιούνται συμπληρωματικά για την κλινική λήψη αποφάσεων.

Οι περιορισμοί και οι προκλήσεις που αναφέρθηκαν, δεν αναιρούν τη χρησιμότητα της προτεινόμενης προσέγγισης, αλλά υπογραμμίζουν την ανάγκη προσεκτικής ερμηνείας των αποτελεσμάτων και την επικύρωσή τους σε μεγαλύτερα, πολυπαραγοντικά πραγματικά σύνολα δεδομένων. Η αναγνώριση των περιορισμών, αποτελεί κρίσιμο βήμα για την σαφή κατανόηση και αξιολόγηση της εργασίας, θέτοντας τα θεμέλια για κατάλληλες βελτιώσεις σε μελλοντικές επεκτάσεις.

7.3 Προτάσεις Βελτίωσης

Η ανάπτυξη της εργασίας θέτει ένα σταθερό και μεθοδολογικό υπόβαθρο για την ανάλυση και πρόβλεψη του κινδύνου πτώσεων, ωστόσο υπάρχουν πολλαπλές κατευθύνσεις στις οποίες μπορεί να βελτιωθεί.

Βασικό πεδίο εξέλιξης αφορά την επέκταση των διαθέσιμων δεδομένων, είτε με την επέκταση του ίδιου του συνόλου δεδομένων από τους δημιουργούς του, είτε πιθανώς με συγχώνευση με άλλο dataset, ή ακόμη την αντικατάσταση αυτού από άλλο dataset. Η ενσωμάτωση διαχρονικών δεδομένων επίσης, θα επέτρεπε τη μελέτη της εξέλιξης του κινδύνου πτώσης σε βάθος χρόνου.

Παράλληλα, η μελέτη θα μπορούσε να εστιάσει στη διερεύνηση εναλλακτικών ή πιο σύνθετων μοντέλων ML, όπως μοντέλα DL ή πιο σύνθετα ΤΝΔ από αυτά που χρησιμοποιήθηκαν στη παρούσα μελέτη.

Σε επίπεδο εφαρμογής, σημαντικές προοπτικές επέκτασης σχετίζονται με τη λειτουργικότητα της εφαρμογής στο πεδίο εισόδου. Αυτές αφορούν επεκτάσεις όπως:

- Προσθήκη κουμπιού σε κάθε πεδίο λειτουργικής αξιολόγησης, το οποίο να παραθέτει μία προκαθορισμένη τιμή (όπως μέση τιμή των δεδομένων εκπαίδευσης του χαρακτηριστικού), ώστε να διευκολύνει τον χρήστη να συμπληρώσει όλα τα πεδία και να εκτελέσει την πρόβλεψη του μοντέλου, σε περίπτωση που δεν διαθέτει κάποια δεδομένα ή δεν είναι δυνατή η εκτέλεση μίας ή παραπάνω λειτουργικής αξιολόγησης των ατόμων.
- Περαιτέρω ανάπτυξη και διαμόρφωση των συστάσεων πρόληψης πτώσεων, επικεντρώνοντας το ενδιαφέρον κυρίως στο άτομο παρά στο προφίλ κινδύνου που εντάσσεται.

- Προσθήκη κουμπιού για την αυτόματη αποθήκευση των αποτελεσμάτων (συστάσεων, προφίλ κινδύνου).
- Ενεργοποίηση διαδικασίας αναφοράς προβλημάτων της εφαρμογής.
- Ανάπτυξη και φιλοξενία της εφαρμογής σε διαδικτυακό περιβάλλον, διευκολύνοντας την απομακρυσμένη πρόσβαση από τους επαγγελματίες υγείας.

Επιπλέον, εξαιρετική συμβολή πρόκειται να αποδώσει η χρήση και αξιολόγηση της εφαρμογής από επαγγελματίες υγείας, σε πραγματικά κλινικά ή κοινοτικά περιβάλλοντα, παρέχοντας πολύτιμη ανατροφοδότηση για τη βελτίωση της χρηστικότητας, της εγκυρότητας και της αποδοχής της.

Οι προοπτικές βελτίωσης που αναδείχθηκαν, υπογραμμίζουν τις κατευθύνσεις στις οποίες μπορεί να εξελιχθεί τόσο η ερευνητική προσέγγιση όσο και η εφαρμογή της παρούσας εργασίας, συμβάλλοντας στη διαμόρφωση μίας πιο ολοκληρωμένης και τεκμηριωμένης αντίληψης στον τομέα της πρόληψης πτώσεων.

7.4 Επίλογος

Ανακεφαλαιώνοντας, συζητήθηκε μία συνολική αποτίμηση της μελέτης και της αναπτυγμένης εφαρμογής, εστιάζοντας στα βασικά ευρήματα, τους περιορισμούς της προτεινόμενης προσέγγισης, τις προκλήσεις που αντιμετωπίσθηκαν κατά τη διάρκεια της ανάπτυξης του πειραματικού σκέλους και τις προοπτικές μελλοντικής εξέλιξης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] World Health Organization, “Falls,” World Health Organization, Apr. 26, 2021. https://www.who.int/news-room/fact-sheets/detail/falls

[2] WHO, “WHO Global report on falls Prevention in older Age,” 2007. Available: https://extranet.who.int/agefriendlyworld/wp-content/uploads/2014/06/WHo-Global-report-on-falls-prevention-in-older-age.pdf

[3] M. Musci, D. De Martini, T. Facchinetti, M. Piastra, and N. Blago, “Online Fall Detection using Recurrent Neural Networks,” in arXiv preprint arXiv:1804.04976, Apr. 2018.

[4] I. Bargiotas et al., “Preventing falls: the use of machine learning for the prediction of future falls in individuals without history of fall,” Journal of Neurology, vol. 270, Jul. 2022, doi: https://doi.org/10.1007/s00415-022-11251-3.

[5] Y. K. Haddad et al., “Healthcare spending for non-fatal falls among older adults, USA,” Injury prevention, vol. 30, no. 4, pp. 272–276, Jul. 2024, doi: https://doi.org/10.1136/ip-2023-045023.

[6] S. Jahandideh et al., “Using machine learning models to predict falls in hospitalised adults,” International journal of medical informatics, vol. 187, pp. 105436–105436, Jul. 2024, doi: https://doi.org/10.1016/j.ijmedinf.2024.105436.

[7] “Fingertips | Department of Health and Social Care,” Phe.org.uk, 2026. https://fingertips.phe.org.uk/search/hip%20fractures#page/6

[8] N. El-Bendary, Q. Tan, F. C. Pivot, and A. Lam, “FALL DETECTION AND PREVENTION FOR THE ELDERLY: A REVIEW OF TRENDS AND CHALLENGES,” International Journal on Smart Sensing and Intelligent Systems, vol. 6, no. 3, pp. 1230–1266, 2013, doi: https://doi.org/10.21307/ijssis-2017-588.

[9] M. Montero-Odasso et al., “World guidelines for falls prevention and management for older adults: A global initiative,” Age and Ageing, vol. 51, no. 9, 2022, doi: https://doi.org/10.1093/ageing/afac205.

[10] GOV.UK, “Falls: Applying All Our Health,” Gov.uk, Feb. 25, 2022. https://www.gov.uk/government/publications/falls-applying-all-our-health/falls-applying-all-our-health

[11] S. Baek, J. Piao, Y. Jin, and S.-M. Lee, “Validity of the Morse Fall Scale implemented in an electronic medical record system,” Journal of Clinical Nursing, vol. 23, no. 17–18, pp. 2434–2441, Sep. 2013, doi: https://doi.org/10.1111/jocn.12359.

[12] K. Milisen et al., “Fall Prediction in Inpatients by Bedside Nurses Using the St. Thomas’s Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) Instrument: A Multicenter Study,” Journal of the American Geriatrics Society, vol. 55, no. 5, pp. 725–733, May 2007, doi: https://doi.org/10.1111/j.1532-5415.2007.01151.x.

[13] A. L. Hendrich, P. S. Bender, and A. Nyhuis, “Validation of the Hendrich II Fall Risk Model: A large concurrent case/control study of hospitalized patients,” Applied Nursing Research, vol. 16, no. 1, pp. 9–21, Feb. 2003, doi: https://doi.org/10.1053/apnr.2003.016009.

[14] Wikipedia Contributors, “Sensor,” Wikipedia, Sep. 25, 2019. https://en.wikipedia.org/wiki/Sensor

[15] G. Zhao, L. Chen, and H. Ning, “Sensor-Based Fall Risk Assessment: A Survey,” Healthcare, vol. 9, no. 11, p. 1448, Oct. 2021, doi: https://doi.org/10.3390/healthcare9111448.

[16] F. Hussain et al., “An Efficient Machine Learning-based Elderly Fall Detection Algorithm,” arXiv.org, Nov. 27, 2019. https://arxiv.org/abs/1911.11976

[17] M. Chen et al., “A Systematic Review of Wearable Sensor-Based Technologies for Fall Risk Assessment in Older Adults,” Sensors, vol. 22, no. 18, p. 6752, Jan. 2022, doi: https://doi.org/10.3390/s22186752.

[18] A. H. Abdul Razak, A. Zayegh, R. K. Begg, and Y. Wahab, “Foot Plantar Pressure Measurement System: A Review,” Sensors, vol. 12, no. 7, pp. 9884–9912, Jul. 2012, doi: https://doi.org/10.3390/s120709884.

[19] R. Modak et al., “An Analysis of Current Advancements: Elderly Fall Detection Systems Using Machine Learning Techniques,” Communications in Computer and Information Science, vol. 1921, pp. 45–69, 2023, doi: https://doi.org/10.1007/978-3-031-45124-9_5.

[20] R. Igual, C. Medrano, and I. Plaza, “Challenges, issues and trends in fall detection systems,” BioMedical Engineering OnLine, vol. 12, no. 1, p. 66, 2013, doi: https://doi.org/10.1186/1475-925x-12-66.

[21] F. Riquelme, C. Espinoza, T. Rodenas, J.-G. Minonzio, and C. Taramasco, “eHomeSeniors Dataset: An Infrared Thermal Sensor Dataset for Automatic Fall Detection Research,” Sensors, vol. 19, no. 20, p. 4565, Oct. 2019, doi: https://doi.org/10.3390/s19204565.

[22] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Εισαγωγή στην Εξόρυξη Δεδομένων, 2nd edition. 2021.

[23] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[24] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, Massachusetts: The MIT Press, 2016. Available: https://www.deeplearningbook.org/

[25] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: https://doi.org/10.1098/rsta.2015.0202.

[26] H. Abdi and L. J. Williams, “Principal component analysis,” Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433–459, Jun. 2010, doi: https://doi.org/10.1002/wics.101.

[27] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: https://doi.org/10.1016/j.patrec.2009.09.011.

[28] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, vol. 5.1, pp. 281–298, 1967, Available: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

[29] Muhammad Ali Syakur, B.K. Khotimah, Eka Malasari Rohman, and Budi Dwi Satoto, “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster,” ResearchGate, Apr. 2018. https://www.researchgate.net/publication/324553963_Integration_K-

Means_Clustering_Method_and_Elbow_Method_For_Identification_of_The_Best_Customer_Profile_Cluste
r

[30] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY: Springer New York, 2009. doi: https://doi.org/10.1007/978-0-387-84858-7.

[31] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 86–97, Dec. 2011, doi: https://doi.org/10.1002/widm.53.

[32] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996. Available: https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf

[33] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 169–194, 1998, doi: https://doi.org/10.1023/a:1009745219419.

[34] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited," ACM Transactions on Database Systems, vol. 42, no. 3, pp. 1–21, Aug. 2017, doi: https://doi.org/10.1145/3068335.

[35] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY: Springer New York, 2013. doi: https://doi.org/10.1007/978-1-4614-6849-3.

[36] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. "O'Reilly Media, Inc.," 2019.

[37] GeeksforGeeks, "Logistic Regression in Machine Learning," GeeksforGeeks, May 09, 2017. https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/.

[38] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," IEEE access, vol. 4, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3416838.

[39] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: https://doi.org/10.1007/bf00116251.

[40] John Ross Quinlan, C4.5 programs for machine learning. San Mateo, Calif. M. Kaufmann, 1993. Available: https://doi.acm.org/10.1145/152181

[41] S. Russell and P. Norvig, Τεχνητή Νοημοσύνη: Μία Σύγχρονη Προσέγγιση, 4η αμερικανική έκδοση. Αθήνα, Ελλάδα: ΚΛΕΙΔΑΡΙΘΜΟΣ, 2021.

[42] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[43] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, Aug. 1996, Available: https://link.springer.com/article/10.1007/BF00058655

[44] I. B. Djordjevic, "Quantum Machine Learning," Quantum Information Processing, Quantum Computing, and Quantum Error Correction, pp. 619–701, 2021, doi: https://doi.org/10.1016/b978-0-12-821982-9.00007-1.

[45] "Boosting (machine learning)," Wikipedia, Sep. 15, 2021. https://en.wikipedia.org/wiki/Boosting_(machine_learning)

[46] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: https://doi.org/10.1214/aos/1013203451.

[47] T. Chen and C. Guestrin, “XGBoost: a Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: https://doi.org/10.1145/2939672.2939785.

[48] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: https://doi.org/10.1007/BF00994018.

[49] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” The Annals of Statistics, vol. 36, no. 3, pp. 1171–1220, Jun. 2008, doi: https://doi.org/10.1214/009053607000000677.

[50] C. Sampaio, “Understanding SVM Hyperparameters,” Stack Abuse, Apr. 21, 2023. https://stackabuse.com/understanding-svm-hyperparameters/

[51] V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer New York, 2000. doi: https://doi.org/10.1007/978-1-4757-3264-1.

[52] Chih-Wei Hsu and Chih-Jen Lin, “A comparison of methods for multiclass support vector machines,” IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 415–425, Mar. 2002, doi: https://doi.org/10.1109/72.991427.

[53] J. Weston and C. Watkins, “Support Vector Machines for Multi-Class Pattern Recognition,” ResearchGate, pp. 219–224, 1999, Accessed: Jan. 22, 2026. [Online]. Available: https://www.researchgate.net/publication/221166057_Support_Vector_Machines_for_Multi-Class_Pattern_Recognition

[54] T.-F. Wu, C.-J. Lin, and R. Weng, “Probability Estimates for Multi-class Classification by Pairwise Coupling,” Journal of Machine Learning Research, vol. 5, pp. 975–1005, 2004, Available: https://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf

[55] Wikipedia Contributors, “Multilayer perceptron,” Wikipedia, Apr. 07, 2019. https://en.wikipedia.org/wiki/Multilayer_perceptron

[56] scikit-learn, “1.17. Neural Network Models (supervised) — scikit-learn 0.23.1 Documentation,” scikit-learn.org, 2025. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[57] “Unsupervised Feature Learning and Deep Learning Tutorial,” Stanford.edu, 2019. http://ufldl.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/

[58] G. Press, “Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says,” Forbes, Mar. 23, 2016. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

[59] H. M. Yasin and A. K. Khorsheed, “Automated Data Cleaning in Large Databases Using Machine Learning Methods,” Asian Journal of Research in Computer Science, vol. 18, no. 5, pp. 364–386, Apr. 2025, doi: https://doi.org/10.9734/ajrcos/2025/v18i5661.

[60] S. Kandel et al., “Research directions in data wrangling: Visualizations and transformations for usable and credible data,” Information Visualization, vol. 10, no. 4, pp. 271–288, Sep. 2011, doi: https://doi.org/10.1177/1473871611415994.

[61] I. Guyon and A. De, “An Introduction to Variable and Feature Selection André Elisseeff,” Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003, Available: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

[62] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot Edouard Duchesnay,” Journal of Machine Learning Research, vol. 12, no. 12, pp. 2825–2830, 2011, Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[63] M. Kang and J. Tian, “Machine Learning: Data Pre-processing,” Prognostics and Health Management of Electronics, pp. 111–130, Aug. 2018, doi: https://doi.org/10.1002/9781119515326.ch5.

[64] K. Pal, S. Ari, Arindam Bit, and S. Bhattacharyya, Advanced Methods in Biomedical Signal Processing and Analysis. Elsevier, 2022.

[65] J. Gillariose, J. Joseph, and C. Chesneau, “Lasso and Ridge regression: a comprehensive review of applications and developments in machine learning,” International Journal of Data Science and Analytics, vol. 21, no. 1, Nov. 2025, doi: https://doi.org/10.1007/s41060-025-00957-y.

[66] J. Han, M. Kamber, and J. Pei, “Cluster Analysis,” Data Mining, pp. 443–495, 2012, doi: https://doi.org/10.1016/b978-0-12-381479-1.00010-1.

[67] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning with Applications in R Second Edition,” 2021. Available: https://www.casact.org/sites/default/files/2022-12/James-G.-et-al.-2nd-edition-Springer-2021.pdf

[68] P. J. Rousseeuw, “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis,” Journal of Computational and Applied Mathematics, vol. 20, no. 0377–0427, pp. 53–65, Nov. 1987, doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[69] scikit-learn, “Selecting the number of clusters with silhouette analysis on KMeans clustering — scikit-learn 0.21.2 documentation,” Scikit-learn.org, 2019. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

[70] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: https://doi.org/10.1109/TPAMI.1979.4766909.

[71] “sklearn.metrics.davies_bouldin_score — scikit-learn 0.22.2 documentation,” scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[72] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, “Density-Based Clustering Validation,” Proceedings of the 2014 SIAM International Conference on Data Mining, Apr. 2014, doi: https://doi.org/10.1137/1.9781611973440.96.

[73] D. Powers and Ailab, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION.” Available: https://arxiv.org/pdf/2010.16061

[74] Evidently AI Team, “Accuracy vs. precision vs. recall in machine learning: what’s the difference?,” www.evidentlyai.com, Oct. 01, 2024. https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall.

[75] R. Kundu, “F1 Score in Machine Learning: Intro & Calculation,” V7, Dec. 16, 2022. https://www.v7labs.com/blog/f1-score-guide.

[76] H. He and E. A. Garcia, “Learning from Imbalanced Data - IEEE Journals & Magazine,” Ieee.org, 2009. https://ieeexplore.ieee.org/document/5128907.

[77] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, no. 16, pp. 321–357, Jun. 2002, doi: https://doi.org/10.1613/jair.953.

[78] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” 1995. Available: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf.

[79] H. Pelletier, “How-To: Cross Validation with Time Series Data | Towards Data Science,” Towards Data Science, Dec. 29, 2023. https://towardsdatascience.com/how-to-cross-validation-with-time-series-data-9802a06272c6/.

[80] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012, Available: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

[81] Zachary Chase Lipton, “The Mythos of Model Interpretability,” ResearchGate, Oct. 06, 2016. https://www.researchgate.net/publication/303942775_The_Mythos_of_Model_Interpretability

[82] I. Bargiotas, A. Kalogeratos, M. Limnios, P.-P. Vidal, D. Ricard, and N. Vayatis, “Revealing posturographic features associated with the risk of falling in patients with Parkinsonian syndromes via machine learning,” arXiv.org, 2019. https://arxiv.org/abs/1907.06614

[83] D. S. Lindberg et al., “Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: A machine-learning approach,” International Journal of Medical Informatics, vol. 143, p. 104272, Nov. 2020, doi: https://doi.org/10.1016/j.ijmedinf.2020.104272.

[84] C. Ye et al., “Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm,” International Journal of Medical Informatics, vol. 137, p. 104105, May 2020, doi: https://doi.org/10.1016/j.ijmedinf.2020.104105.

[85] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, “SisFall: A Fall and Movement Dataset,” Sensors (Basel, Switzerland), vol. 17, no. 1, p. 198, Jan. 2017, doi: https://doi.org/10.3390/s17010198.

[86] NICE, “Overview | Falls: Assessment and Prevention in Older People and in People 50 and over at Higher Risk | Guidance | NICE,” Nice.org.uk, Apr. 29, 2025. https://www.nice.org.uk/guidance/ng249.

[87] G. Vavoulas, M. Pediaditis, E. G. Spanakis, and Manolis Tsiknakis, “The MobiFall dataset: An initial evaluation of fall detection algorithms using smartphones,” IEEE Transactions on Biomedical Engineering, Nov. 2013, doi: https://doi.org/10.1109/bibe.2013.6701629.

[88] S. Maudsley-Barton and M. H. Yap, “KINECAL,” Physionet.org, Jun. 08, 2023. https://physionet.org/content/kinecal/1.0.3/

[89] D. Lytras, E. Sykaras, P. Iakovidis, K. Kasimis, I. Myrogiannis, and A. Kottaras, “Recording of Falls in Elderly Fallers in Northern Greece and Evaluation of Aging Health-Related Factors and Environmental

Safety Associated with Falls: a Cross-Sectional Study," Occupational Therapy International, vol. 2022, pp. 1–11, Jan. 2022, doi: https://doi.org/10.1155/2022/9292673.

[90] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," ResearchGate, 2021. https://www.researchgate.net/publication/276394797_Likert_Scale_Explored_and_Explained.

[91] D. Podsiadlo and S. Richardson, "The Timed 'Up & Go': a Test of Basic Functional Mobility for Frail Elderly Persons," Journal of the American Geriatrics Society, vol. 39, no. 2, pp. 142–8, 1991, doi: https://doi.org/10.1111/j.1532-5415.1991.tb01616.x.

[92] J. E. Rossiter-Fornoff, S. L. Wolf, L. I. Wolfson, and D. M. Buchner, "A Cross-sectional Validation Study of the FICSIT Common Data Base Static Balance Measures," The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, vol. 50A, no. 6, pp. M291–M297, Nov. 1995, doi: https://doi.org/10.1093/gerona/50a.6.m291.

[93] R. E. Rikli and C. J. Jones, "Development and Validation of a Functional Fitness Test for Community-Residing Older Adults," Journal of Aging and Physical Activity, vol. 7, no. 2, pp. 129–161, Apr. 1999, doi: https://doi.org/10.1123/japa.7.2.129.

[94] K. O. Berg, S. L. Wood-Dauphinee, J. I. Williams, and B. Maki, "Measuring balance in the elderly: validation of an instrument," Canadian Journal of Public Health = Revue Canadienne De Sante Publique, vol. 83 Suppl 2, no. 2, pp. S7-11, Jul. 1992, Available: https://pubmed.ncbi.nlm.nih.gov/1468055/.

[95] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician," Journal of Psychiatric Research, vol. 12, no. 3, pp. 189–198, Nov. 1975, doi: https://doi.org/10.1016/0022-3956(75)90026-6.

[96] E. Billis et al., "Cross-cultural validation of the Falls Efficacy Scale International (FES-I) in Greek community-dwelling older adults," Disability and Rehabilitation, vol. 33, no. 19–20, pp. 1776–1784, Jan. 2011, doi: https://doi.org/10.3109/09638288.2010.546937.

[97] Fariq Rahmat et al., "Supervised feature selection using principal component analysis," Knowledge and Information Systems, Nov. 2023, doi: https://doi.org/10.1007/s10115-023-01993-5.

[98] F. Song, Z. Guo, and D. Mei, "Feature Selection Using Principal Component Analysis," IEEE Xplore, 2010. https://ieeexplore.ieee.org/abstract/document/5640135.

[99] Zaid Mundher Radeef, S. H. Hashem, and Ekhlas Khalaf Gbashi, "New Feature Selection Using Principal Component Analysis," Journal of soft computing & computer applications., vol. 1, no. 2, Dec. 2024, doi: https://doi.org/10.70403/3008-1084.1012.

[100] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," arXiv.org, Nov. 24, 2017. https://arxiv.org/abs/1705.07874v2.

[101] Christoph Molnar, "Interpretable Machine Learning," Github.io, Aug. 27, 2019. https://christophm.github.io/interpretable-ml-book/.

[102] M. Montero-Odasso et al., "World guidelines for falls prevention and management for older adults: A global initiative," Age and Ageing, vol. 51, no. 9, 2022, doi: https://doi.org/10.1093/ageing/afac205.

[103] M. Kim, E. Shin, S. Kim, and S. Sok, "The Effectiveness of Multicomponent Intervention on Daily Functioning among the Community-Dwelling Elderly: A Systematic Review," International Journal of

Environmental Research and Public Health, vol. 19, no. 12, p. 7483, Jun. 2022, doi: https://doi.org/10.3390/ijerph19127483.

[104] S. Trudeau, "18 Steps to Fall Proofing Your Home," Ncoa.org, Sep. 2017. https://www.ncoa.org/article/18-steps-to-fall-proofing-your-home/.

[105] E. Nabors, "How to Prevent Falls with Home Safety Modifications," Ncoa.org, Apr. 17, 2025. https://www.ncoa.org/article/how-to-prevent-falls-with-home-safety-modifications/.

[106] CDC, "Facts About Falls," Older Adult Fall Prevention, May 09, 2024. https://www.cdc.gov/falls/data-research/facts-stats/index.html.

[107] CDC, "Falls Compendium," Older Adult Fall Prevention, Jul. 28, 2024. https://www.cdc.gov/falls/interventions/falls-compendium.html.

[108] GOV.UK, "Falls: Applying All Our Health," Gov.uk, Feb. 25, 2022. https://www.gov.uk/government/publications/falls-applying-all-our-health/falls-applying-all-our-health.

[109] CDC, "Clinical Resources," STEADI - Older Adult Fall Prevention, May 16, 2024. https://www.cdc.gov/steadi/hcp/clinical-resources/?CDC_Aaref_Val=https://www.cdc.gov/steadi/materials.html.

[110] National Institute on Aging, "Falls and fractures in older adults: Causes and prevention," National Institute on Aging, Sep. 12, 2022. https://www.nia.nih.gov/health/falls-and-falls-prevention/falls-and-fractures-older-adults-causes-and-prevention.

[111] "Six Tips To Help Prevent Falls," National Institute on Aging, Jun. 05, 2025. https://www.nia.nih.gov/health/falls-and-falls-prevention/six-tips-help-prevent-falls.

[112] NICE, "Recommendations | Falls: assessment and prevention in older people and in people 50 and over at higher risk | Guidance | NICE," Nice.org.uk, Apr. 29, 2025. https://www.nice.org.uk/guidance/ng249/chapter/Recommendations#interventions-to-reduce-the-risk-of-falls

[113] Streamlit, "Streamlit • The fastest way to build and share data apps," streamlit.io, 2025. https://streamlit.io/.

ΠΑΡΑΡΤΗΜΑ Α: Σύνολα Δεδομένων Μελέτης

Public Dataset	Year	Age of subjects	No of Subjects (Male/ Female)	No of Types of Emulated ADLs/Falls	No of samples of ADLs/Falls	Total No of samples	Sensors	Placement	Data
Tai Chi, Physiological Complexity, and Healthy Aging - Gait	2021	50-79	60	5/0	5000/0	5000	1x IMU (A, G)	Lower back	Time series (.csv)
SisFall	2017	19-30 & 60-75	38 Young:(11/12,) Elders:(8/7)	19/15	2707/1789	4505	2x A, 1x G (custom device) – 200Hz	Waist	Time series (.txt)
HuGaDB	2017	20-27	18 (14/4)	6/0	2 111 962/0	2 111 962	6x IMU (A, G), 2x EMG	Thighs, shins, feet, front thigh	Time series (.csv) (10h of data)
GaitData	2019	-	230 (141/89)	1/0	1020/0	1020	2x IMUs	Ankles	Time series (.csv) (8.5h gait data)
eHomeSeniors	2019	-	6 (0/6)	0/15	0/448	448	1x MLX90640, 4x Omron D6T-8L-06	Wall, floor	Thermal images (.csv)
PlantPre	2022	Older adults	48	5/0	500/0	500	1x IMU (A, G)	Waist	Time series (.csv)
Le2i	2013	-	6 (4/2)	5/5	48/143	191	1x RGB camera	Fixed positions in room	Video (.avi)
UR Fall	2014	-	6 (4/2)	5/5	40/30	70	2x Microsoft Kinect	Fixed positions in room	RGB-D

							Cameras		video (.avi)
Activity Recognition in Senior Citizens	2021	70-95	18 (8/10)	8/0	2 196 986/0	2 196 986	2x 3-axial A	Right thigh & lower back	Time series (.csv)
sy3kcttdtx	2021	-	20 (10/10)	5/0	600 000/0	600 000	3-axial A, 3-axial G	Waist	Time series (.csv)
MobiFall	2016	22-36	11 (6/5)	9/4	342/288	630	1x Smartphone (A, G)	Waist	Time series (.txt)
UMA Fall	2017	18-55	17 (10/7)	13/1	322/209	746	1x 3-axial A, 1x 3-axial G, 1x M	Ankles, wrists, waist, chest	Time series (.csv)
KFall	2021	Average 25	32 (32/0)	21/15	2729/2346	5076 αρχεία κίνησης	1x 9-axial (A, G, M (LPMS-B2))	Waist	Time series (.csv)
KINECAL	2023	18-92	90 (45/45)	11/0	1000/0	1000	1x Microsoft Kinect V2	Fixed position in room	RGB-D video (.avi), clinical metadata (.csv)
UniMiB SHAR	2016	18 - 60	30 (15/15)	9/8	5427/6344	11771	Smartphone (Bosh BMA220) (A: ±2g to ±16g)	Front trouser pocket	Time series (.txt)

A	Επιταχυνσιόμετρο
G	Γυροσκόπιο
M	Μαγνητόμετρο
B	Βαρόμετρο