ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗ ΝΟΣΟ ΤΟΥ PARKINSON

Όλγα Ελευθεράκου

Διπλωματική Εργασία

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

Πειραιάς
Μάρτιος 2023

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

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗ ΝΟΣΟ ΤΟΥ PARKINSON

Όλγα Ελευθεράκου

Διπλωματική Εργασία

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

Πειραιάς
Μάρτιος 2023

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

Τα μέλη της Επιτροπής ήταν:

- Σωτήριος Μπερσίμης (Επιβλέπων, Αναπληρωτής Καθηγητής)
- Κωνσταντίνος Πολίτης (Αναπληρωτής Καθηγητής)
- Σωτήριος Τασουλής (Επίκουρος Καθηγητής)

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

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

APPLICATIONS OF MACHINE LEARNING IN PARKINSON'S DISEASE

By

Olga Eleftherakou

MSc Dissertation

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

Piraeus, Greece
March 2023

Ευχαριστίες

Η παρούσα διπλωματική εργασία ολοκληρώνει έναν πολύ σημαντικό κύκλο της ζωής μου, ο οποίος έχει ξεκινήσει από τη στιγμή που εισήχθηκα στο τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης στο Πανεπιστήμιο Πειραιώς. Από την πρώτη κιόλας στιγμή αγάπησα τις σπουδές και το τμήμα μου και ονειρευόμουν ήδη το μεταπτυχιακό μου.

Κατά τη διάρκεια όλων αυτών των χρόνων άλλαξαν πολλά, γνώρισα πολύ σημαντικά άτομα αλλά και τους καθηγητές από τους οποίους πήρα πολύτιμες γνώσεις.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Σωτήριο Μπερσίμη, τόσο για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα και την εμπιστοσύνη που μου έδειξε για την εκπόνησή του, όσο και για τις γνώσεις που απεκόμισα από εκείνον όλα αυτά τα χρόνια. Επιπλέον, θα ήθελα να ευχαριστήσω τον κ. Κωνσταντίνο Πολίτη και τον κ. Σωτήριο Τασουλή για τις συμβουλές τους και την υποστήριξή τους στην εκπόνηση της διπλωματικής μου εργασίας.

Τέλος, θέλω να ευχαριστήσω τον κ. Παναγιώτη Ζήκο, νευρολόγο με εξειδίκευση στην επεμβατική αντιμετώπιση της Νόσου του Πάρκινσον, που μου έδωσε τα ανωνυμοποιημένα δεδομένα από τους ασθενείς του με Νόσο Πάρκινσον ώστε να μπορέσω να πραγματοποιήσω την παρούσα μελέτη.

Περίληψη

Στις μέρες μας, η μηχανική μάθηση έχει γίνει βασικό εργαλείο για την καταπολέμηση της νόσου του Πάρκινσον. Με την ικανότητά της να αναλύει τεράστιες ποσότητες δεδομένων, η μηχανική μάθηση μπορεί να βοηθήσει στην έγκαιρη ανίχνευση και την ακριβή διάγνωση της νόσου του Πάρκινσον. Μπορεί επίσης να βοηθήσει στην παρακολούθηση της εξέλιξης της νόσου και στην πρόβλεψη των αποτελεσμάτων των ασθενών. Οι αλγόριθμοι μηχανικής μάθησης μπορούν επίσης να χρησιμοποιηθούν για τον εντοπισμό πιθανών νέων φαρμακευτικών στόχων και την αξιολόγηση της αποτελεσματικότητας των υφιστάμενων θεραπειών. Επιπλέον, η μηχανική μάθηση μπορεί να βοηθήσει στην ανάπτυξη εξατομικευμένων σχεδίων θεραπείας για τους ασθενείς, λαμβάνοντας υπόψη τις ατομικές διαφορές στην παρουσίαση της νόσου και την ανταπόκριση στις θεραπείες. Εν γένει, η μηχανική μάθηση έχει τη δυνατότητα να φέρει επανάσταση στην κατανόηση και τη θεραπεία της νόσου του Πάρκινσον, βελτιώνοντας τελικά τα αποτελέσματα και την ποιότητα ζωής των ασθενών. Στην παρούσα εργασία, γίνεται μια εκτενής περιγραφή σε προβλήματα που εφάπτονται στην αντιμετώπιση της Νόσου του Πάρκινσον, ενώ δίνεται περισσότερη έμφαση στη διαχείριση δεδομένων από Holter ώστε να παρθεί απόφαση, χρησιμοποιώντας μεθόδους μηχανικής μάθησης, για το ποιοι ασθενείς είναι κατάλληλοι για «εν τω βάθει εγκεφαλική διέγερση».

Abstract

Nowadays, machine learning has become an essential tool in the fight against Parkinson's disease. With its ability to analyse vast amounts of data, machine learning can assist in the early detection and accurate diagnosis of Parkinson's disease. It can also help track disease progression and predict patient outcomes. Machine learning algorithms can also be used to identify potential new drug targets and evaluate the efficacy of existing treatments. Moreover, machine learning can aid in the development of personalized treatment plans for patients, taking into account individual differences in disease presentation and response to therapies. Overall, machine learning has the potential to revolutionize our understanding and treatment of Parkinson's disease, ultimately improving patient outcomes and quality of life. In this thesis, an extensive description is given of problems that are relevant to the treatment of Parkinson's Disease, with more emphasis on Holter data management to make a decision, using machine learning methods, about which patients are suitable for deep brain stimulation (DBS).

Περιεχόμενα

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Η σημαντικότητα της επιστήμης δεδομένων και της μηχανικής μάθησης στη Νόσο του Πάρκινσον

Η επιστήμη των δεδομένων έχει φέρει επανάσταση στον τομέα των επιστημών υγείας, παρέχοντας ισχυρά εργαλεία για την ανάλυση μεγάλων ποσοτήτων σύνθετων δεδομένων, τα οποία μπορούν να βοηθήσουν τους ερευνητές και τους επαγγελματίες του ιατρικού κλάδου να κατανοήσουν καλύτερα τις ασθένειες και να αναπτύξουν αποτελεσματικότερες θεραπείες. Ένας τομέας όπου η επιστήμη των δεδομένων είναι ιδιαίτερα σημαντική είναι η νόσος του Πάρκινσον (ΝΠ), μια χρόνια και προοδευτική νευρολογική διαταραχή που επηρεάζει εκατομμύρια ανθρώπους παγκοσμίως. Η ΝΠ χαρακτηρίζεται από ποικίλα συμπτώματα, όπως τρέμουλο, δυσκαμψία και δυσκολία στην κίνηση, και μπορεί να είναι δύσκολο να διαγνωστεί και να αντιμετωπιστεί αποτελεσματικά.

Τεχνικές της επιστήμης των δεδομένων, όπως η μηχανική μάθηση, η εξόρυξη δεδομένων και η προγνωστική μοντελοποίηση, μπορούν να εφαρμοστούν σε μεγάλα σύνολα δεδομένων με πληροφορίες ασθενών για τον εντοπισμό μοτίβων και πληροφοριών που μπορεί να είναι χρήσιμες για την ανάπτυξη ακριβέστερων διαγνωστικών εργαλείων και αποτελεσματικότερων θεραπειών για τη ΝΠ. Για παράδειγμα, οι ερευνητές μπορούν να χρησιμοποιήσουν την επιστήμη των δεδομένων για να αναλύσουν αρχεία ασθενών και να εντοπίσουν παράγοντες κινδύνου που μπορεί να συμβάλλουν στην ανάπτυξη της νόσου ή να αναπτύξουν αλγορίθμους που μπορούν να προβλέψουν την εξέλιξη της νόσου σε μεμονωμένους ασθενείς.

Γενικά, η σημασία της επιστήμης των δεδομένων στις επιστήμες της υγείας και οι δυνατότητές της για τη βελτίωση της κατανόησης και της θεραπείας της ΝΠ είναι τεράστιες. Αξιοποιώντας τη δύναμη της επιστήμης των δεδομένων, μπορούμε να ξεκλειδώσουμε νέες γνώσεις σχετικά με αυτή την πολύπλοκη ασθένεια και να αναπτύξουμε πιο αποτελεσματικές στρατηγικές για την καταπολέμησή της.

1.2 Η επιστήμη των δεδομένων, η μηχανική μάθηση, και η στατιστική

Η επιστήμη των δεδομένων, η μηχανική μάθηση και η στατιστική είναι όλα βασικά στοιχεία της σύγχρονης εποχής της λήψης αποφάσεων με βάση τα δεδομένα. Η επιστήμη των δεδομένων είναι ένας διεπιστημονικός τομέας που συνδυάζει υπολογιστικές και στατιστικές μεθόδους για την εξαγωγή συμπερασμάτων από τα δεδομένα. Η μηχανική μάθηση είναι ένα υποσύνολο της επιστήμης δεδομένων που περιλαμβάνει την κατασκευή αλγορίθμων και μοντέλων που μπορούν να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις ή αποφάσεις με βάση αυτή τη μάθηση. Η στατιστική, από την άλλη πλευρά, είναι ένας μαθηματικός κλάδος που περιλαμβάνει τη συλλογή, την ανάλυση και την ερμηνεία των δεδομένων.

Μαζί, αυτοί οι τομείς διαδραματίζουν κρίσιμο ρόλο στο να βοηθούν οργανισμούς και άτομα να λαμβάνουν τεκμηριωμένες αποφάσεις αξιοποιώντας τη δύναμη των δεδομένων. Η επιστήμη των δεδομένων παρέχει το πλαίσιο για τη συλλογή, την επεξεργασία και την ανάλυση δεδομένων, ενώ η μηχανική μάθηση προσφέρει ισχυρές τεχνικές για την ανακάλυψη μοτίβων και την πραγματοποίηση προβλέψεων. Η στατιστική, εν τω μεταξύ, παρέχει τα θεμέλια για την κατανόηση και την ερμηνεία των αποτελεσμάτων της ανάλυσης δεδομένων.

Καθώς ο όγκος των δεδομένων που παράγονται από άτομα και οργανισμούς συνεχίζει να αυξάνεται με πρωτοφανή ρυθμό, η σημασία της επιστήμης των δεδομένων, της μηχανικής μάθησης και της στατιστικής έχει μόνο αυξηθεί. Με αυτά τα εργαλεία, οι επιχειρήσεις μπορούν να ανακαλύψουν νέες γνώσεις σχετικά με τους πελάτες, τα προϊόντα και τις δραστηριότητές τους, ενώ οι ερευνητές μπορούν να αποκτήσουν νέες προοπτικές σε πολύπλοκα επιστημονικά προβλήματα. Εν ολίγοις, η επιστήμη των δεδομένων, η μηχανική μάθηση και η στατιστική αποτελούν βασικά συστατικά της επανάστασης που καθοδηγείται από τα δεδομένα και μεταμορφώνει τον κόσμο μας σήμερα.

1.3 Η Νόσος του Πάρκινσον και η επέμβαση DBS

Η ΝΠ είναι μια χρόνια και προοδευτική νευρολογική διαταραχή που πλήττει εκατομμύρια ανθρώπους παγκοσμίως. Χαρακτηρίζεται από μια ποικιλία συμπτωμάτων, όπως τρέμουλο, δυσκαμψία και δυσκολία στην κίνηση. Αν και δεν υπάρχει θεραπεία για τη ΝΠ, υπάρχουν διάφορες διαθέσιμες θεραπείες που μπορούν να βοηθήσουν στη διαχείριση των συμπτωμάτων της.

Μία από τις πιο υποσχόμενες θεραπείες για τη ΝΠ είναι η χειρουργική επέμβαση βαθιάς εγκεφαλικής διέγερσης (Deep Brain Stimulation, εφεξής DBS). Η DBS περιλαμβάνει την

εμφύτευση ηλεκτροδίων σε συγκεκριμένες περιοχές του εγκεφάλου και τη χρήση τους για την παροχή ηλεκτρικών ερεθισμάτων που μπορούν να βοηθήσουν στη μείωση του τρέμουλου και άλλων συμπτωμάτων της ΝΠ. Η χειρουργική επέμβαση DBS έχει αποδειχθεί ότι είναι αποτελεσματική στη βελτίωση των κινητικών συμπτωμάτων, στη μείωση των αναγκών σε φάρμακα και στη βελτίωση της ποιότητας ζωής των ασθενών με ΝΠ.

Αν και η χειρουργική επέμβαση DBS δεν αποτελεί θεραπεία για τη ΝΠ, μπορεί να προσφέρει σημαντική ανακούφιση στους ασθενείς που παλεύουν με τα συμπτώματα αυτής της εξουθενωτικής πάθησης. Ως αποτέλεσμα, έχει γίνει μια όλο και πιο δημοφιλής θεραπευτική επιλογή για άτομα με ΝΠ που δεν ανταποκρίνονται σε άλλες θεραπείες ή που αντιμετωπίζουν σημαντικές παρενέργειες από τη φαρμακευτική αγωγή. Συνολικά, η χειρουργική επέμβαση DBS αποτελεί σημαντική πρόοδο στη θεραπεία της ΝΠ και προσφέρει ελπίδα στα άτομα που παλεύουν με αυτή τη δύσκολη κατάσταση.

Κ Ε Φ Α Λ Α Ι Ο 2

Η νόσος του Πάρκινσον

2.1 Εισαγωγή στη νόσο του Πάρκινσον

2.1.1 Η ιστορία της νόσου του Πάρκινσον

Παρόλο που τα στοιχεία για την πιθανή ΝΠ υπάρχουν σε πολύ πρώιμα έγγραφα, η πρώτη σαφής ιατρική περιγραφή γράφτηκε το 1817 από τον James Parkinson. Στα μέσα της δεκαετίας του 1800, ο Jean-Martin Charcot άσκησε ιδιαίτερη επιρροή στη βελτίωση, στην επέκταση αλλά και στη διάδοση των πληροφοριών διεθνώς σχετικά με τη ΝΠ. Ακόμη, διαχώρισε τη ΝΠ από την πολλαπλή σκλήρυνση και άλλες διαταραχές που χαρακτηρίζονται από τρέμουλο, και αναγνώρισε περιπτώσεις που αργότερα θα ταξινομούνταν -πιθανότατα- μεταξύ των συνδρόμων Parkinsonism-plus.

Οι πρώτες θεραπείες της ΝΠ βασίστηκαν στην εμπειρική παρατήρηση, ενώ τα αντιχολινεργικά φάρμακα χρησιμοποιούνταν ήδη από τον 19ο αιώνα. Η ανακάλυψη της ανεπάρκειας των ντοπαμινεργικών στη ΝΠ και τα συνθετικά παράγωγα (synthetic paths) της ντοπαμίνης, οδήγησαν στις πρώτες δοκιμές της λεβοντόπας σε ανθρώπους. Περαιτέρω ιστορικά σημαντικές, ανατομικές, βιοχημικές και φυσιολογικές μελέτες προσδιόρισαν πρόσθετους φαρμακολογικούς και νευροχειρουργικούς στόχους για τη ΝΠ και επιτρέπουν στους σύγχρονους κλινικούς ιατρούς να προσφέρουν μια σειρά θεραπειών με στόχο τη βελτίωση της λειτουργίας σε αυτή την, ακόμη, ανίατη ασθένεια.

2.1.2 Εισαγωγή και επιδημιολογικά στοιχεία

Η ΝΠ εμφανίζεται κυρίως λόγω της προοδευτικής εκφύλισης των ντοπαμινεργικών νευρώνων της μέλαινας ουσίας των βασικών γαγγλίων. Σε μεγάλο βαθμό, οι ασθενείς με ΝΠ υποφέρουν από μη κινητικά συμπτώματα, όπως κατάθλιψη, άγχος, κόπωση και διαταραχές του ύπνου, που χρειάζονται περαιτέρω διερεύνηση. Η κατάθλιψη στη ΝΠ είναι ένα κυρίαρχο και πολύπλοκο σύμπτωμα και η παθολογία της είναι εξωγενής στο μελαινοραβδωτό σύστημα. Η

νόσος αυτή μπορεί, τελικά, να αντιμετωπιστεί με συνδυασμό τακτικής φυσικοθεραπείας και κατάλληλης φαρμακευτικής αγωγής. Γενικά, η ΝΠ χρειάζεται πρόοδο στην έρευνα και την ανάπτυξη, ώστε οι ασθενείς με ΝΠ να μπορούν να ζήσουν μια φυσιολογική ζωή.

Αναλυτικότερα, η ΝΠ είναι μια συχνή νευροεκφυλιστική διαταραχή - μια συνουκλεϊνοπάθεια - με επιπολασμό 160/100.000 στη Δυτική Ευρώπη, ενώ ανέρχεται στο 4% του πληθυσμού άνω των 80 ετών. Είναι η δεύτερη πιο συχνή νευροεκφυλιστική νόσος, με 10.000.000 άτομα παγκοσμίως να έχουν διαγνωστεί με τη νόσο, και ο επιπολασμός της έχει προβλεφθεί να διπλασιαστεί τα επόμενα 30 χρόνια. Ο αριθμός αυτός προκύπτει από μια αύξηση του επιπολασμού κατά 2-5 φορές τα τελευταία 30 χρόνια, καθιστώντας τη ΝΠ μία από τις κύριες αιτίες νευρολογικής αναπηρίας. Να σημειωθεί ότι η παγκόσμια επιβάρυνση από τη ΝΠ - από την άποψη των θανάτων και της αναπηρίας - έχει διπλασιαστεί τις τελευταίες δύο δεκαετίες.

Το 2016, η Μελέτη Παγκόσμιας Νοσοεπιβάρυνσης της Δημόσιας Υγείας (Global Burden of Disease Study, Dorsey, E. et al.) υπολόγισε ότι περίπου 6,1 εκατομμύρια άνθρωποι παγκοσμίως πάσχουν από Πάρκινσον. Αυτός ο αριθμός αποτελεί σημαντική αύξηση από τους 2,5 εκατομμύρια πάσχοντες το 1990. Στην ΕΕ, εκτιμάται ότι η πάθηση προσβάλλει 1,2 εκατομμύρια ανθρώπους. Σύμφωνα με την ίδια έκθεση, υπολογίστηκε ότι μέχρι το 2040, ο αριθμός των ανθρώπων που πάσχουν από Πάρκινσον θα πλησιάζει τα 13 εκατομμύρια.

Όσον αφορά τις Ηνωμένες Πολιτείες, σχεδόν 90.000 άτομα διαγιγνώσκονται με ΝΠ κάθε χρόνο, ενώ 1.000.000 άνθρωποι ζουν με αυτή. Να σημειωθεί ότι η επίπτωση της ΝΠ αυξάνεται με την ηλικία, και εκτιμάται ότι το 4% των ατόμων με ΝΠ, διαγιγνώσκονται με αυτή πριν την ηλικία των 50 ετών. Το συνδυασμένο άμεσο και έμμεσο κόστος της ΝΠ, συμπεριλαμβανομένης της θεραπείας, των πληρωμών κοινωνικής ασφάλισης και του απωλεσθέντος εισοδήματος, εκτιμάται ότι ανέρχεται σε σχεδόν 52 δισεκατομμύρια δολάρια ετησίως μόνο στις ΗΠΑ. Μόνο τα φάρμακα κοστίζουν κατά μέσο όρο 2.500 δολάρια ετησίως και η θεραπευτική χειρουργική επέμβαση μπορεί να κοστίσει έως και 100.000 δολάρια ανά άτομο.

Παγκοσμίως	**Ηνωμένες Πολιτείες**	**Ευρώπη**	**Ελλάδα**
6.100.000	1.000.000	1.200.000	~30.000

Πίνακας 2.1: Επιπολασμός της ΝΠ

2.2 Παράγοντες κινδύνου και αιτίες

Με τη γήρανση του πληθυσμού, η διαχείριση της νόσου είναι πιθανό να αποδειχθεί φλέγον ζήτημα και να αποτελέσει μεγάλη πρόκληση για τους νευρολόγους και τους γενικούς ιατρούς.

Ο αυξανόμενος επιπολασμός της παγκοσμίως, θυμίζει τα χαρακτηριστικά που συνήθως παρατηρούνται κατά τη διάρκεια μιας πανδημίας, με σημαντική διαφορά ότι η ΝΠ δεν είναι μολυσματική. Στους περισσότερους πληθυσμούς, το 3-5% των ασθενών της ΝΠ εξηγείται από γενετικά αίτια που συνδέονται με γνωστά γονίδιά της, συνιστώντας έτσι τη μονογονιδιακή ΝΠ, ενώ 90 γενετικές παραλλαγές κινδύνου εξηγούν συνολικά το 16-36% του κληρονομικού κινδύνου της μη μονογονιδιακής ΝΠ. Πρόσθετες αιτιώδεις συσχετίσεις περιλαμβάνουν την ύπαρξη συγγενικού προσώπου που πάσχει από τη νόσο.

Επιπλέον, τα φυτοφάρμακα, οι περιβαλλοντικοί ρύποι, η μολυσμένη περιβαλλοντική ατμόσφαιρα αλλά και άλλοι παράγοντες όπως το κάπνισμα, ο καφές, και η άσκηση έχουν βρεθεί ότι παίζουν ρόλο στην παθογένεια της ΝΠ σε διάφορους πληθυσμούς.

Οι δύο τύποι παραγόντων κινδύνου περιλαμβάνουν γενετικούς και μη γενετικούς παράγοντες κινδύνου. Δεν υπάρχει κάποια συσχέτιση μεταξύ του καπνίσματος και της ΝΠ, ενώ η κατανάλωση του καφέ βρέθηκε ότι μειώνει τον κίνδυνο εμφάνισής της. Ακόμη και οι διατροφικοί παράγοντες όπως τα λιπαρά οξέα και τα αντιοξειδωτικά είναι υπό διερεύνηση. Επιπροσθέτως, η παρουσία γενετικής μετάλλαξης θεωρείται ότι ενέχει κίνδυνο για την ανάπτυξη της ΝΠ. Μία πρόσφατη μελέτη έδειξε, επίσης, ότι ο κίνδυνος εμφάνισης ΝΠ ήταν αυξημένος μετά από εγκεφαλικό επεισόδιο. Συνεπώς, η ισχαιμία παίζει και αυτή ρόλο στην ανάπτυξη της γνωστικής εξασθένησης.

Η ΝΠ είναι μια ασθένεια που σχετίζεται με την ηλικία, με τη συχνότητα και τον επιπολασμό να αυξάνονται σταθερά όσο αυξάνεται και η ηλικία. Ωστόσο, η εσφαλμένη αντίληψη ότι η ΝΠ προσβάλλει αποκλειστικά τους ηλικιωμένους πρέπει να απορριφθεί. Η ηλικία έναρξης της νόσου για σχεδόν το 25% των προσβεβλημένων ατόμων είναι μικρότερη των 65 ετών και για το 5-10% είναι μικρότερη των 50 ετών. Βέβαια, η ηλικία είναι ο σημαντικότερος παράγοντας κινδύνου για την ανάπτυξη της νόσου, και οι άνδρες είναι πιο επιρρεπείς από ό,τι οι γυναίκες, με αναλογία επιπολασμού περίπου 3:2.

Η κατανόηση της παθογένειάς της έχει προχωρήσει την τελευταία δεκαετία, με την ταυτοποίηση αρκετών γονιδιακών μεταλλάξεων, που μπορούν να ρίξουν φως στους μηχανισμούς της παθογένειας σε σποραδικές περιπτώσεις ΝΠ.

2.2.1 Κληρονομικότητα

Οι πρώιμες μελέτες διδύμων και ο εντοπισμός πολλών οικογενειών που παρουσίαζαν μοτίβο Μεντελικής κληρονομικότητας (επικρατές και υπολειπόμενο) παρείχαν ενδείξεις για γενετικές αιτίες της νόσου, οι οποίες, το 1997, κατέληξαν στην ανακάλυψη της α-συνουκλεΐνης (SNCA), του πρώτου γονιδίου που σχετίζεται με την ΝΠ. Ένα χρόνο αργότερα, εντοπίστηκε η μετάλλαξη στην Parkin (PRKN), που συνδέεται με την αυτοσωμική υπολειπόμενη μορφή της ΝΠ. Η ονοματολογία της αντιστοίχισης ενός αριθμού "PARK" σε αυτά τα γονίδια έχει προκαλέσει σύγχυση και, ως εκ τούτου, προτιμούμε την ταξινόμηση που προτάθηκε από τη Διεθνή Εταιρεία Πάρκινσον και Διαταραχών της Κίνησης χρησιμοποιώντας τα ονόματα των γονιδίων.

2.3 Συμπτώματα

2.3.1 Κύρια συμπτώματα

Το κλινικό σήμα κατατεθέν της ΝΠ είναι ένα κινητικό σύνδρομο που, εκτός από αλλαγές στη στάση του σώματος και τη βάδιση, χαρακτηρίζεται από βραδυκινησία, τρέμουλο ηρεμίας και δυσκαμψία.

Αν και τα κλασικά κινητικά συμπτώματα εμφανίζονται νωρίς και αποτελούν τους πυλώνες των διαγνωστικών κριτηρίων, η ανάπτυξη της αστάθειας και των αυξανόμενων δυσκολιών βάδισης, όπως επίσης και η δυσφαγία και η δυσαρθρία, οδηγούν στην εξέλιξη της κινητικής αναπηρίας.

Η ΝΠ θεωρείται κινητική διαταραχή, αλλά σχετίζεται με μια ποικιλία μη κινητικών συμπτωμάτων σε όλους σχεδόν τους ασθενείς, συμπεριλαμβανομένης της υποσμίας, της δυσκοιλιότητας, της δυσλειτουργίας της ούρησης, την ορθοστατική υπόταση, την απώλεια μνήμης, την κατάθλιψη, τον πόνο και τις διαταραχές του ύπνου. Τα κλασικά κινητικά συμπτώματα της ΝΠ συνδέονται με τον εκφυλισμό του νευρικού συστήματος και την εξάντληση της ντοπαμίνης του ραβδωτού σώματος, ενώ τα μη κινητικά συμπτώματα -πιθανώς- συνδέονται με νευροεκφυλισμό άλλων δομών, συμπεριλαμβανομένου του περιφερικού αυτόνομου νευρικού συστήματος.

2.3.2 Το πρώιμο στάδιο

Όπως αναφέρθηκε, είναι αρκετά τα μη κινητικά συμπτώματα που σχετίζονται με τη ΝΠ, όπως η απώλεια της όσφρησης ή η δυσκοιλιότητα και αναφέρονται συνήθως από τους ασθενείς πριν από την εμφάνιση των κλασικών κινητικών συμπτωμάτων - μερικές φορές πριν από την εμφάνιση των κινητικών χαρακτηριστικών κατά χρόνια ή ακόμη και δεκαετίες. Η περίοδος κατά την οποία εμφανίζονται αυτά τα συμπτώματα έχει χαρακτηριστεί ως προδρομική φάση της ΝΠ που αντιστοιχεί σε ένα στάδιο της νόσου κατά το οποίο οι νευροεκφυλιστικές αλλαγές αφορούν εξωεγκεφαλικές περιοχές, όπως το κατώτερο εγκεφαλικό στέλεχος, τον οσφρητικό βολβό και τις οσφρητικές οδούς, και το περιφερικό αυτόνομο νευρικό σύστημα.

2.4 Διάγνωση

Τα τελευταία χρόνια έχουν επικυρωθεί κλινικά διαγνωστικά κριτήρια, σχεδιασμένα για την ενίσχυση της διαγνωστικής ακρίβειας της νόσου. Η διάγνωση της ιδιοπαθούς ΝΠ μπορεί να είναι μια απλή κλινική εξέταση σε περιπτώσεις με κλασικό ιστορικό και αποκλεισμό εναλλακτικών αιτιών. Ωστόσο, στην κλινική πρακτική ρουτίνας η διαγνωστική λανθασμένη ταξινόμηση είναι συχνή, με ποσοστά σφάλματος που κυμαίνονται από 15% έως 24% σε διάφορες χρονικές στιγμές και μελέτες.

Είναι γεγονός ότι η διάγνωση παραμένει μια πρόκληση και ο χαρακτηρισμός των πρώιμων σταδίων της νόσου βρίσκεται σε εξέλιξη επειδή τα κλινικά χαρακτηριστικά μπορεί να συγχέονται με εκείνα άλλων νευροεκφυλιστικών παθήσεων και οι εξετάσεις ή οι βιοδείκτες να μην επιτρέπουν την οριστική διάγνωση από τα πρώτα στάδια. Ως αποτέλεσμα, η κλινική διαγνωστική ακρίβεια παραμένει υποβαθμισμένη, ακόμη και όταν η νόσος εκδηλώνεται κλινικά πλήρως. Γενικά, όμως, η διάγνωση της νόσου βασίζεται σε κλινικές μεθόδους - οι συμπληρωματικές εξετάσεις προορίζονται για άτομα με άτυπη εικόνα.

Η ΝΠ είναι εντυπωσιακά ετερογενής όσον αφορά την ηλικία έναρξης, την κλινική εικόνα, τον ρυθμό εξέλιξης και την ανταπόκριση στη θεραπεία και γι’ αυτό, έχουν προταθεί αρκετοί κλινικοί υποτύποι. Η ανακάλυψη γενετικά καθορισμένων μορφών της νόσου, οι οποίες μπορεί να διαφέρουν από σποραδικές μορφές σε διάφορες κλινικές μεταβλητές, έχουν αμφισβητήσει την ενιαία άποψη για τη ΝΠ και εισήγαγαν έναν βιολογικό ορισμό των υποπεριπτώσεων που την απαρτίζουν.

Για την υποκατηγοριοποίηση της ΝΠ έχουν χρησιμοποιηθεί είτε εμπειρικές εκτιμήσεις μεμονωμένων κλινικών χαρακτηριστικών είτε πιο αντικειμενικές και χωρίς υποθέσεις

μεθοδολογίες ιεραρχικής ανάλυσης κατά συστάδες και άλλες μορφές μηχανικής μάθησης. Τα κλινικά χαρακτηριστικά που χρησιμοποιήθηκαν για την υποκατηγοριοποίηση και με τις δύο προσεγγίσεις περιλάμβαναν την ηλικία κατά την έναρξη (πρώιμη έναρξη εναντίον όψιμης έναρξης), τον επικρατή κινητικό φαινότυπο (περιπτώσεις με κυρίαρχο τρέμουλο έναντι περιπτώσεων χωρίς τρέμουλο), τις κινητικές επιπλοκές ως προς την ανταπόκριση στη θεραπεία με λεβοντόπα, τα μη κινητικά χαρακτηριστικά, ιδίως αυτόνομη δυσλειτουργία, γνωσιακή δυσλειτουργία και διαταραχή της συμπεριφοράς στον ύπνο REM, καθώς και τον ρυθμό εξέλιξής τους.

2.5 Θεραπεία

Οι στόχοι της θεραπείας διαφέρουν από άτομο σε άτομο, γεγονός που ενισχύει την ανάγκη για εξατομικευμένη διαχείριση. Η λεβοντόπα είναι το πιο συνηθισμένο φάρμακο που χρησιμοποιείται ως θεραπεία πρώτης γραμμής. Η βέλτιστη διαχείριση πρέπει να ξεκινά από τη διάγνωση και απαιτεί μια διεπιστημονική ομαδική προσέγγιση, που περιλαμβάνει ένα αυξανόμενο φάσμα μη φαρμακευτικών παρεμβάσεων. Προς το παρόν, καμία θεραπεία δεν μπορεί να επιβραδύνει ή να ανακόψει την εξέλιξη της ΝΠ, αλλά με βάση τις νέες γνώσεις σχετικά με τα γενετικά αίτια και τους μηχανισμούς του νευρωνικού θανάτου, δοκιμάζονται διάφορες υποσχόμενες στρατηγικές για τη δυνατότητα τροποποίησης της νόσου.

2.5.1 Φαρμακοθεραπεία

Επί του παρόντος, δεν υπάρχει μόνιμη θεραπεία κατά της ΝΠ που να είναι διαθέσιμη. Μόνο η φαρμακευτική αγωγή και η χειρουργική επέμβαση παρέχουν ανακούφιση από τα συμπτώματα της νόσου.

Πριν από τριάντα χρόνια, περίπου 50.000 άτομα διαγιγνώσκονταν ετησίως με ΝΠ. Το μόνο φάρμακο που ήταν διαθέσιμο εκείνη την εποχή ήταν η λεβοντόπα - μια χημική ένωση που το σώμα μπορεί να μετατρέψει σε ντοπαμίνη. Αυτό βοήθησε πολλούς από τους ασθενείς με ΝΠ να επιβιώσουν, αλλά η μακροχρόνια χρήση αυτού του φαρμάκου είχε ως αποτέλεσμα ανεξέλεγκτες κινήσεις στο σώμα του ασθενούς. Η χειρουργική επέμβαση στον εγκέφαλο των κατεστραμμένων περιοχών είναι επίσης ένας εναλλακτικός τρόπος διαθέσιμης θεραπείας εκτός από τη θεραπεία με φάρμακα.

Προσφάτως, εισήχθησαν διάφορα νέα φάρμακα εκτός από τη λεβοντόπα ως θεραπευτικό αποτέλεσμα για τους ασθενείς με ΝΠ. Τα φάρμακα για την ΝΠ υπάγονται σε τρεις διακριτές

κατηγορίες που βοηθούν τον έλεγχο της νόσου και μετριάζουν τις παρενέργειές της. Φάρμακα που δρουν άμεσα ή έμμεσα για να αυξήσουν το επίπεδο της ντοπαμίνης στον εγκέφαλο και περιλαμβάνουν πρόδρομες ουσίες της ντοπαμίνης, όπως η λεβοντόπα, αποτελούν το πρώτο είδος φαρμάκων για την ΝΠ. Το δεύτερο είδος φαρμάκων, επηρεάζει άλλους νευροδιαβιβαστές στο σώμα προκειμένου να ελέγξει τη νόσο. Αυτά τα φάρμακα βοηθούν στη μείωση του τρέμουλου και της μυϊκής δυσκαμψίας, τα οποία μπορεί να προκύψουν από την ύπαρξη περισσότερης ακετυλοχολίνης από ό,τι ντοπαμίνης στο σύστημα. Ο τρίτος τύπος φαρμάκων που συνταγογραφούνται για την ΝΠ περιλαμβάνει φάρμακα που βοηθούν στον έλεγχο των μη κινητικών συμπτωμάτων της νόσου.

Η συνήθης θεραπευτική προσέγγιση είναι από του στόματος φαρμακοθεραπεία και πρόσφατα επικεντρώθηκε στην χειρουργική τροποποίηση της περιοχής του εγκεφάλου που σχετίζεται με τη ΝΠ. Γενικά, οι φαρμακοθεραπείες σχετίζονται με τις διαταραχές των νευροδιαβιβαστών μονοαμίνης. Υπάρχουν διάφορα φάρμακα διαθέσιμα για τη θεραπεία των κινητικών διαταραχών στη ΝΠ όπως η καρβιντόπα/λεβοντόπα, η πραμιπεξόλη και η ροπινιρόλη. Αυτά τα φάρμακα είναι σε θέση να τροποποιήσουν τις ανισορροπίες στους νευρώνες που παράγουν ντοπαμίνη στους ασθενείς με ΝΠ. Οι φαρμακοθεραπείες είναι πολύ αποτελεσματικές σε σύντομο χρονικό διάστημα. Τα άνω και κάτω άκρα ανταποκρίνονται στη ντοπαμινεργική θεραπεία, όπου υπάρχει περιορισμένη ανταπόκριση από τα αξονικά συμπτώματα. Πέραν τούτου, η ντοπαμίνη ασκεί μεγάλη δράση στα βασικά συμπτώματα (τρέμουλο, βραδυκινησία, δυσκαμψία), ενώ ασκεί περιορισμένη στα κορμικά συμπτώματα. Όσον αφορά τις χειρουργικές θεραπείες, η επιτυχία άνω του 72% είναι μέθοδος εκλογής σε όλους τους ασθενείς που φτάνουν σε προχωρημένο στάδιο.

Κατά τη διάρκεια του πρώιμου σταδίου εξέλιξης της νόσου, προκύπτουν τα σπονδυλικά συμπτώματα και το μη ντοπαμινεργικό αξονικό σύμπτωμα. Σε εκτεταμένα στάδια επηρεάζονται περαιτέρω τα μη ντοπαμινεργικά συστήματα (μετωπιαίος φλοιός και παρεγκεφαλίδα).

2.5.2 Φυσικοθεραπεία

Υπάρχουν αρκετές στρατηγικές φυσικοθεραπείας που είναι ευεργετικές. Είναι σημαντικό να κατανοηθούν οι ειδικές πρακτικές στην φυσικοθεραπεία με εξειδίκευση τη ΝΠ – οι φυσικοθεραπευτές που έλαβαν εκπαίδευση ειδικά για τη ΝΠ και αντιμετωπίζουν μεγάλο αριθμό ασθενών με ΝΠ καθημερινά, μπορούν να προσφέρουν στους ασθενείς τους καλύτερα

αποτελέσματα και με μικρότερο κόστος σε σχέση με τους φυσικοθεραπευτές χωρίς εξειδίκευση. Πολλές στρατηγικές φυσικοθεραπείας εκμεταλλεύονται την ικανότητα των ατόμων με ΝΠ να αντισταθμίζουν τις κινητικές τους αναπηρίες, με εναλλακτικά κινητικά προγράμματα που παρακάμπτουν τα ελαττωματικά κυκλώματα των βασικών γαγγλίων. Ένα πολύ γνωστό παράδειγμα είναι αυτό ενός ατόμου με ΝΠ το οποίο πάσχει από σοβαρό πάγωμα της βάδισης (FoG), αλλά μπορεί ακόμα να κάνει αβίαστα ποδήλατο.

Κ Ε Φ Α Λ Α Ι Ο 3

Βιβλιογραφική ανασκόπηση σε εφαρμογές της νόσου του Πάρκινσον

3.1 Εισαγωγή

Ενώ οι μετρητές Holter παρέχουν πληθώρα δεδομένων, η ανάλυσή τους μπορεί να είναι χρονοβόρα και πολύπλοκη διαδικασία. Σε αυτό το σημείο έρχεται η μηχανική μάθηση. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να εκπαιδευτούν ώστε να αναλύουν αυτόματα μεγάλες ποσότητες δεδομένων από μόνιτορ Holter και να εντοπίζουν μοτίβα που μπορεί να είναι ενδεικτικά της ΝΠ. Με την αυτοματοποίηση αυτής της διαδικασίας, οι αλγόριθμοι μηχανικής μάθησης μπορούν να εξοικονομήσουν πολύτιμο χρόνο για τους επαγγελματίες υγείας και να βελτιώσουν την ακρίβεια της διάγνωσης της ΝΠ. Επιπροσθέτως, μπορεί να πραγματοποιηθεί και περαιτέρω ανάλυση χρησιμοποιώντας μεταβλητές που παρέχει το Holter, χωρίς να είναι απαραίτητο να χρησιμοποιηθούν οι χρονοσειρές, χωρίς δηλαδή να γίνει χρήση των δεδομένων απευθείας. Με αυτόν τον τρόπο, παρέχεται η εγκυρότητα του μηχανήματος αλλά υπάρχει και ελευθερία εκτέλεσης μεθόδων που θα χρησιμοποιούνταν και σε συμβατικά δεδομένα (π.χ. συλλογή από νοσοκομείο, ερωτηματολόγια, κλπ.).

Εκτός από τη βοήθεια στη διάγνωση, η μηχανική μάθηση μπορεί επίσης να χρησιμοποιηθεί για την παρακολούθηση της εξέλιξης της ΝΠ με την πάροδο του χρόνου. Αναλύοντας τις αλλαγές στα δεδομένα με Holter σε πολλαπλές καταγραφές, οι αλγόριθμοι μηχανικής μάθησης μπορούν να παρέχουν πληροφορίες σχετικά με το πώς η νόσος επηρεάζει την καρδιακή λειτουργία του ασθενούς αλλά και οτιδήποτε άλλο μπορεί να χρειαστεί να μελετηθεί. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για την προσαρμογή των θεραπειών και τη βελτίωση των αποτελεσμάτων των ασθενών.

Συνολικά, η μηχανική μάθηση έχει τη δυνατότητα να φέρει επανάσταση στην ανάλυση των δεδομένων από Holter για τη ΝΠ. Με την αυτοματοποίηση της διαδικασίας ανάλυσης και την παροχή πολύτιμων πληροφοριών, η μηχανική μάθηση μπορεί να βελτιώσει την ακρίβεια της διάγνωσης και της θεραπείας αυτής της εξουθενωτικής πάθησης.

Έπειτα από διερευνητική ανασκόπηση στη βιβλιογραφία, εντοπίστηκαν μελέτες οι οποίες χρησιμοποιούν μεθόδους μηχανικής και βαθιάς μάθησης σε δεδομένα από Holter.

3.2 Εντοπισμός παγώματος βάδισης με χρήση Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs)

Στην εργασία των Rodríguez et al. (2017) παρουσιάζεται ένας νέος αλγόριθμος για την ανίχνευση της FoG με μια προσέγγιση μηχανικής μάθησης που βασίζεται σε Μηχανές Διανυσμάτων Υποστήριξης (SVM) και ένα τριαξονικό επιταχυνσιόμετρο που φοριέται στη μέση.

Η μέθοδος αξιολογείται μέσω των σημάτων επιτάχυνσης από δεδομένα που έχουν συλλεχθεί από 21 ασθενείς, που βρίσκονται στο σπίτι τους, με ΝΠ και αξιολογείται υπό δύο διαφορετικές συνθήκες· πρώτον, δοκιμάζεται ένα γενικό μοντέλο με τη χρήση της προσέγγισης leave-one-out και, δεύτερον, ένα εξατομικευμένο μοντέλο που χρησιμοποιεί επίσης μέρος του συνόλου δεδομένων από κάθε ασθενή.

Τα αποτελέσματα δείχνουν σημαντική βελτίωση της ακρίβειας του εξατομικευμένου μοντέλου σε σύγκριση με το γενικό μοντέλο, παρουσιάζοντας βελτίωση της ειδικότητας και της ευαισθησίας γεωμετρικού μέσου όρου (GM) κατά 7.2%. Επιπλέον, η προσέγγιση SVM που υιοθετήθηκε συγκρίθηκε με την πιο ολοκληρωμένη μέθοδο ανίχνευσης FoG που χρησιμοποιείται επί του παρόντος (η οποία στην παρούσα εργασία αναφέρεται ως MBFA). Τα αποτελέσματα της νέας γενικής μεθόδου παρέχουν βελτίωση 11.2% στο GM σε σύγκριση με το γενικό μοντέλο MBFA και, στην περίπτωση του εξατομικευμένου μοντέλου, βελτίωση 10% σε σχέση με το εξατομικευμένο μοντέλο MBFA.

Έτσι, τα αποτελέσματα της μελέτης δείχνουν ότι μια προσέγγιση μηχανικής μάθησης μπορεί να χρησιμοποιηθεί για την παρακολούθηση της FoG κατά τη διάρκεια της καθημερινής ζωής των ασθενών με PD και, επιπλέον, τα εξατομικευμένα μοντέλα για την ανίχνευση της FoG μπορούν να χρησιμοποιηθούν για τη βελτίωση της ακρίβειας παρακολούθησης.

3.3 Βαθιά μάθηση για ανίχνευση επεισοδίων παγώματος βάδισης στη ΝΠ

Στην εργασία των Julià Camps et al. (2017) παρουσιάζεται μια μέθοδος για την ανίχνευση FOG που βασίζεται σε τεχνικές βαθιάς μάθησης και επεξεργασίας σήματος. Αυτή είναι και η πρώτη φορά που η ανίχνευση FoG αντιμετωπίζεται με βαθιά μάθηση.

Η αξιολόγηση του μοντέλου έγινε με βάση τα δεδομένα από 15 ασθενείς με ΝΠ που εκδήλωσαν FoG. Μια αδρανειακή μονάδα μέτρησης που τοποθετήθηκε στην αριστερή πλευρά της μέσης κατέγραψε σήματα από τριαξονικό επιταχυνσιόμετρο, γυροσκόπιο και μαγνητόμετρο.

Εξαγωγή χαρακτηριστικών (feature extraction) με συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network, CNN)

Ένα CNN είναι ένας τύπος βαθιού νευρωνικού δικτύου τροφοδότησης προς τα εμπρός, το οποίο συνήθως συνδυάζει συνελικτικά στρώματα με παραδοσιακά πυκνά στρώματα για να μειώσει τον αριθμό των βαρών που συνθέτουν το μοντέλο.

Ενώ τα παραδοσιακά μοντέλα βαθιάς μάθησης αποτελούνται από στοιβαγμένα πυκνά στρώματα, τα οποία οδηγούν σε υπερβολικό αριθμό βαρών, τα CNN εφαρμόζουν μια ισχυρή και αποδοτική εναλλακτική λύση εάν τα δεδομένα-στόχοι παρουσιάζουν υποκείμενα χωρικά μοτίβα.

Η προσέγγιση που παρουσιάζεται είναι ένα μονοδιάστατο CNN, το οποίο περιγράφεται ως εξής:

$$C(16|3) - C(16|3) - C(16|3) - C(16|3) - C(16|3) - D(32) - D(32) - L$$

Όπου, το C(x|y) αντιστοιχεί σε ένα συνελικτικό στρώμα x πυρήνων μήκους y, το D(z) αντιστοιχεί σε ένα πυκνό στρώμα z νευρώνων και το L είναι το τελευταίο στρώμα του νευρωνικού δικτύου.

Η ανίχνευση συμβάντων FoG αντιμετωπίστηκε ως πρόβλημα διπλής ταξινόμησης (binary classification), έτσι ώστε οι περιπτώσεις FoG να χαρακτηρίζονται ως θετικές τιμές (δηλαδή 1), ενώ οι περιπτώσεις μη FoG χαρακτηρίζονται ως αρνητικές τιμές (δηλαδή -1). Ως εκ τούτου, εφαρμόστηκε μια γραμμική συνάρτηση με συντελεστή ποινής κανονικοποίησης (weight regularisation penalty coefficient) βάρους L2 που είχε οριστεί στο 0.01 ως συνάρτηση ενεργοποίησης του τελευταίου στρώματος, το οποίο είχε οριστεί ως πυκνό στρώμα ενός νευρώνα.

Αναπαράσταση και επεξεργασία δεδομένων

Μια συνήθης πρακτική για τη βελτίωση της ποιότητας εκπαίδευσης των μοντέλων μηχανικής μάθησης είναι η κανονικοποίηση στα δεδομένα. Τα δεδομένα κανονικοποιήθηκαν

με την προϋπολογισμένη δειγματική τυπική απόκλιση από το συνολικό σύνολο δεδομένων εκπαίδευσης.

Όσον αφορά την αντιμετώπιση των χρονοσειρών, η πιο συνηθισμένη τεχνική για την αντιμετώπιση προβλημάτων ταξινόμησης σε δεδομένα χρονοσειρών είναι η χρήση μιας στρατηγικής γνωστή ως windowing. Το windowing συνίσταται στη διάσπαση των δεδομένων σε διαδοχικά τμήματα ίσου μεγέθους για την αντιμετώπιση του προβλήματος ταξινόμησης κατά παράθυρο αντί κατά περίπτωση. Οι στρατηγικές επαύξησης δεδομένων που εφαρμόστηκαν ήταν οι εξής:

- Τυχαία μετατόπιση των σημείων εκκίνησης του παραθύρου.
- Περιστροφή κάθε παραθυροποιημένου σήματος προσομοιώνοντας μια περιστροφή στον αισθητήρα μέσης μέσω ενός πίνακα περιστροφής που παράγεται από γωνίες δειγματοληψίας σε μια κατανομή που ορίζεται από τις ακόλουθες περιοχές:
    - άξονας x: [-30◦, +30◦]
    - άξονας y: [-45◦, +45◦] και
    - άξονας z: [-10◦, 10◦], οι οποίες σχεδιάστηκαν για να μοιάζουν με τις φυσικές κινήσεις της μέσης του ασθενούς.

Είναι γνωστό ότι τα μοντέλα μηχανικής μάθησης μπορούν να αποτύχουν να επιλύσουν προβλήματα ταξινόμησης, αλλά και άλλα προβλήματα, σε περίπτωση που τα δεδομένα είναι ανεπαρκή. Για την εξασφάλιση όσο το δυνατόν πιο έγκυρης και ακριβούς πληροφορίας, οι Rodríguez et al. (2017) εκτελούν μια εξαγωγή χαρακτηριστικών (feature extraction) που υπολογίζεται από τα τρέχοντα και από τα προηγούμενα παράθυρα. Η χρήση μεταβατικών πληροφοριών των παραθύρων υποδηλώνει ότι τα δεδομένα από ένα μόνο παράθυρο μπορεί να είναι ανεπαρκή για την επιτυχή ανίχνευση της FoG. Για να δοθεί στα μοντέλα βαθιάς μάθησης μια αναπαράσταση δείγματος από την οποία θα μπορούσαν να μάθουν ποιο μέρος των δεδομένων του δείγματος άνηκε στο τρέχον παράθυρο και ποιο στη μετάβαση παραθύρων, εφαρμόστηκε μία νέα στρατηγική, η οποία στο εξής θα αναφέρεται ως στοίβαξη. Η στρατηγική στοίβαξης μπορεί να θεωρηθεί ως μια συνάρτηση η οποία εξάγει το παράθυρο που πρέπει να τροφοδοτήσει το μοντέλο από τα δεδομένα του τρέχοντος και του προηγούμενου παραθύρου συν μια τιμή για την παράμετρο στοίβαξης p. Έτσι,

$$\mathrm{Sn} = \mathrm{stacking}(\mathrm{Wn}, \mathrm{Wn} - 1, \mathrm{p})$$

Όπου, η Sn αναφέρεται στο νιοστό στοιβαγμένο παράθυρο που χρησιμοποιείται για την τροφοδοσία του μοντέλου, τα W_n και W_{n-1} αναφέρονται στα τρέχοντα και τα προηγούμενα

παραθυροποιημένα δεδομένα, αντίστοιχα, και p είναι μια παράμετρος αντιστάθμισης της συνάρτησης στοίβαξης η οποία ορίζεται στη συνέχεια:

1. Το προηγούμενο και το τρέχον παράθυρο χωρίζονται σε p μέρη ίσου μεγέθους.
2. Κάθε i-οστό μέρος του τρέχοντος παραθύρου αντιστοιχίζεται με το (i-1)-οστό μέρος, ακόμη και αν αυτό ανήκει στο προηγούμενο ή στο ίδιο παράθυρο, οπότε, σε αυτή την περίπτωση, το κομμάτι αυτό επαναλαμβάνεται. Τα μέρη από το προηγούμενο παράθυρο που δεν είχαν αντιστοιχιστεί, αφαιρούνται.
3. Κάθε εναπομείναν μέρος μετασχηματίζεται με την εφαρμογή του γρήγορου μετασχηματισμού Fourier (FFT).
4. Κάθε τμήμα του τρέχοντος παραθύρου συμπληρώνεται από το προηγούμενο τμήμα του με απλή συνένωση των στηλών του τρέχοντος τμήματος με τη διαφορά μεταξύ των δύο τμημάτων, η οποία παράγει ένα νέο εκτεταμένο τμήμα που αποτελείται από 18 στήλες.
5. Τέλος, όλα τα μέρη στοιβάζονται εκ νέου μαζί διατηρώντας την αρχική χρονική σειρά.

Αποτελέσματα

Συνοψίζοντας, η συγκεκριμένη μελέτη παρουσιάζει μια προσέγγιση η οποία υλοποιεί ένα νευρωνικό δίκτυο συνελικτικού τύπου 8 επιπέδων, το οποίο αποτελείται από δύο πυκνά (δηλ. πλήρως συνδεδεμένα) στρώματα και ένα στρώμα εξόδου. Στα πειράματα που πραγματοποιήθηκαν, αυτό το μοντέλο πέτυχε συγκρίσιμες επιδόσεις με τις πλέον ενδεδειγμένες και σύγχρονες μεθόδους. Συγκεκριμένα, τα μοντέλα που παρουσιάστηκαν ήταν σε θέση να επιτύχουν επιδόσεις 88.6% και 78% για την ευαισθησία και την ειδικότητα αντίστοιχα. Η κύρια συμβολή της παρούσας μελέτης είναι να χρησιμοποιηθεί ως βάση από την οποία να σχεδιάζονται μοντέλα βαθιών δικτύων ικανά να κυριαρχήσουν στην ανίχνευση FoG από αδρανειακά δεδομένα που συλλέγονται σε φυσιολογικά περιβάλλοντα.

3.4 Εκτίμηση της σοβαρότητας της βραδυκινησίας στη ΝΠ

Στην μελέτη των A. Samà et al. (2017) αξιολογείται μια μέθοδος μηχανικής μάθησης που αναλύει τα σήματα που παρέχονται από ένα τριαξονικό επιταχυνσιόμετρο τοποθετημένο στη μέση των ασθενών με ΝΠ, προκειμένου να αξιολογείται αυτόματα η βραδυκινησία.

Η μέθοδος αυτή χρησιμοποιεί μηχανές διανυσμάτων υποστήριξης (SVM) για τον προσδιορισμό των τμημάτων των σημάτων που αντιστοιχούν στη βάδιση. Η συχνότητα των βημάτων χρησιμοποιείται στη συνέχεια για τον προσδιορισμό των περιόδων βραδυκινησίας

και για την εκτίμηση της σοβαρότητας της βραδυκινησίας με βάση ένα μοντέλο epsilon Support Vector Regression. Η μέθοδος εφαρμόζεται σε 12 ασθενείς με ΝΠ, γεγονός που οδηγεί σε δύο βασικά συμπεράσματα. Πρώτον, η συχνότητα των βηματισμών επιτρέπει την διχοτομική ανίχνευση της βραδυκινησίας με ακρίβεια μεγαλύτερη από 90%. Η διαδικασία αυτή απαιτεί τη χρήση ενός κατωφλίου που εξαρτάται από τον ασθενή και το οποίο εκτιμάται με βάση ένα μοντέλο παλινδρόμησης leave-one-patient-out. Δεύτερον, η σοβαρότητα της βραδυκινησίας που μετράται μέσω των σκορ UPDRS προσεγγίζεται μέσω ενός μοντέλου παλινδρόμησης με σφάλματα κάτω του 10%.

Παρόλο που η μέθοδος πρέπει να επικυρωθεί περαιτέρω σε περισσότερους ασθενείς, τα αποτελέσματα που προέκυψαν υποδηλώνουν ότι η συγκεκριμένη προσέγγιση μπορεί να χρησιμοποιηθεί με επιτυχία για την αξιολόγηση της βραδυκινησίας στην καθημερινή ζωή των ασθενών με ΝΠ.

3.5 Ανάλυση μετάβασης στάσης με βαρόμετρα

Σε μια ακόμη μελέτη των Rodríguez et al. (2018), παρουσιάζεται ένας αλγόριθμος που συνδυάζει τις πληροφορίες ενός βαρομέτρου και ενός επιταχυνσιόμετρου για την ανίχνευση μεταβάσεων στάσης και πτώσεων.

Σε αντίθεση με άλλες εργασίες, δοκιμάζονται διαφορετικές δραστηριότητες (στις οποίες εμπλέκεται το υψόμετρο) προκειμένου να επιτευχθεί ένας αξιόπιστος ταξινομητής έναντι ψευδώς θετικών αποτελεσμάτων (false positive). Επιπλέον, μέσω μεθόδων επιλογής χαρακτηριστικών, επιτυγχάνονται βέλτιστα υποσύνολα χαρακτηριστικών για τους αισθητήρες επιταχυνσιόμετρου και βαρόμετρου για την πλαισίωση αυτών των δραστηριοτήτων. Τα επιλεγμένα χαρακτηριστικά δοκιμάζονται μέσω διαφόρων ταξινομητών (classifiers) μηχανικής μάθησης, οι οποίοι αξιολογούνται με ένα σύνολο δεδομένων αξιολόγησης (validation set). Τα αποτελέσματα δείχνουν ότι η προσθήκη χαρακτηριστικών του βαρομέτρου πέραν εκείνων που λαμβάνονται για το επιταχυνσιόμετρο βελτιώνει σαφώς την ακρίβεια ανίχνευσης έως και 11%, από την άποψη του γεωμετρικού μέσου μεταξύ ευαισθησίας και ειδικότητας, σε σύγκριση με αλγορίθμους όπου χρησιμοποιείται μόνο το επιταχυνσιόμετρο.

Τέλος, αναλύεται επίσης η επιβάρυνση του υπολογιστή και, υπό αυτή την έννοια, η χρήση βαρομέτρων, εκτός από την αύξηση της ακρίβειας, μειώνει επίσης τους υπολογιστικούς πόρους που απαιτούνται για την ταξινόμηση ενός νέου μοτίβου, όπως φαίνεται από τη μείωση στον αριθμό των διανυσμάτων υποστήριξης.

Ταξινομητές που χρησιμοποιήθηκαν (Classifiers)

Σύμφωνα με προηγούμενες μελέτες, οι ταξινομητές κοντινότεροι γείτονες (k-NN), δένδρα απόφασης (decision trees) και Naïve Bayes, δεν παρέχουν πολύ υψηλά αποτελέσματα επίδοσης και γι' αυτό τον λόγο δεν εφαρμόστηκαν στη συγκεκριμένη μελέτη. Οι ταξινομητές που εφαρμόστηκαν στην παρούσα μελέτη είναι οι:

- Μηχανές Διανυσμάτων Υποστήριξης
- Λογιστική Παλινδρόμηση
- Πολυστρωματικό perceptron
- Τυχαίο δάσος

Κ Ε Φ Α Λ Α Ι Ο 4

Μηχανική Μάθηση

4.1 Εισαγωγή

Η Μηχανική Μάθηση ως έννοια υπάρχει εδώ και αρκετό καιρό. Ο όρος "μηχανική μάθηση" επινοήθηκε από τον Arthur Samuel το 1959, έναν επιστήμονα πληροφορικής στην IBM, και πρωτοπόρο στην τεχνητή νοημοσύνη και τα ηλεκτρονικά παιχνίδια. Ο Samuel σχεδίασε ένα πρόγραμμα υπολογιστή για να παίζει ντάμα. Όσο περισσότερο έπαιζε το πρόγραμμα, τόσο περισσότερο μάθαινε από την εμπειρία, χρησιμοποιώντας αλγορίθμους για να κάνει προβλέψεις.

Ως επιστημονικός κλάδος, η μηχανική μάθηση διερευνά την ανάλυση και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις σε αυτά.

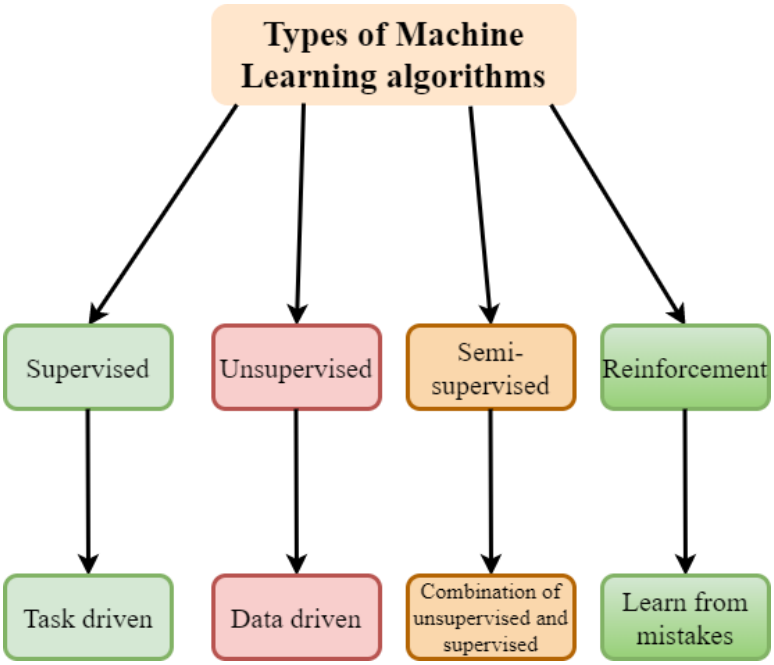
Η μηχανική μάθηση έχει αποδειχθεί πολύτιμη επειδή μπορεί να επιλύσει προβλήματα με ταχύτητα, και σε τέτοια κλίμακα, που δεν μπορεί να αναπαραχθεί μόνο από το ανθρώπινο μυαλό. Με τεράστιες ποσότητες υπολογιστικών ικανοτήτων, οι μηχανές μπορούν να εκπαιδευτούν ώστε να εντοπίζουν μοτίβα και σχέσεις μεταξύ των δεδομένων εισόδου και να αυτοματοποιούν διαδικασίες ρουτίνας.

Η μηχανική μάθηση, διακρίνεται, γενικά, σε τέσσερις κατηγορίες, οι οποίες θα αναλυθούν εκτενώς παρακάτω. Η πρώτη κατηγορία είναι η Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning), η δεύτερη κατηγορία είναι η Μη Εποπτευόμενη Μηχανική Μάθηση (Unsupervised Machine Learning), η τρίτη κατηγορία είναι η Μηχανική Μάθηση με Ημι-επίβλεψη (Semi-supervised Machine Learning) και, τέλος, η τέταρτη κατηγορία, είναι η Ενισχυτική Μάθηση (Reinforcement Learning).

4.2 Είδη μηχανικής μάθησης

Αν και, γενικότερα, μπορεί να θεωρηθεί ότι η μηχανική μάθηση χωρίζεται σε δύο μεγάλες κατηγορίες, την εποπτευόμενη μηχανική μάθηση (Supervised learning) και τη μη

εποπτευόμενη μηχανική μάθηση (Unsupervised learning), στην παρούσα εργασία θα αναφερθούν και οι τέσσερις κατηγορίες, οι οποίες είναι οι προαναφερθείσες και, επιπλέον, η ημι-εποπτευόμενη μηχανική μάθηση (semi-supervised learning) και η ενισχυτική μάθηση (Reinforcement learning).



Σχήμα 4.1: Είδη μηχανικής μάθησης

4.2.1 Εποπτευόμενη μάθηση (Supervised learning)

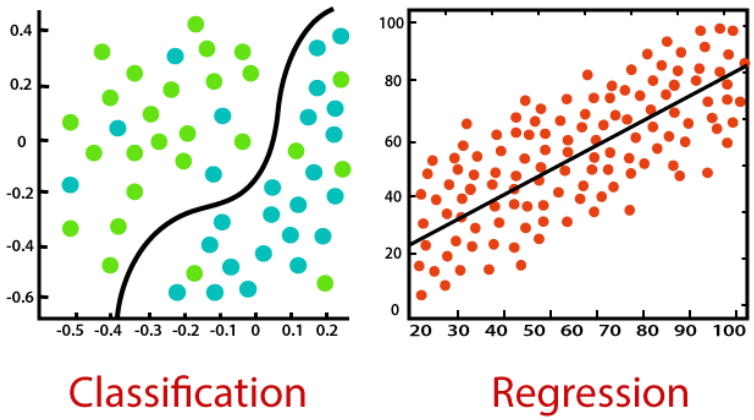
Το καθοριστικό χαρακτηριστικό της μάθησης με επίβλεψη είναι η διαθεσιμότητα σχολιασμένων ή, αλλιώς, χαρακτηρισμένων (labelled) δεδομένων εκπαίδευσης (training dataset). Το όνομα επικαλείται την ιδέα ενός "επόπτη" που δίνει οδηγίες στο σύστημα μάθησης σχετικά με τις ετικέτες που πρέπει να συσχετιστούν με τα παραδείγματα εκπαίδευσης. Συνήθως αυτές οι ετικέτες είναι ετικέτες κλάσεων σε προβλήματα ταξινόμησης. Οι αλγόριθμοι μάθησης με επίβλεψη δημιουργούν μοντέλα από τα δεδομένα εκπαίδευσης και τα μοντέλα αυτά μπορούν να χρησιμοποιηθούν για την ταξινόμηση άλλων δεδομένων χωρίς ετικέτες.

Η εποπτευόμενη μάθηση χρησιμοποιείται στους παρακάτω δύο τύπους προβλημάτων:

- Σε προβλήματα ταξινόμησης (classification) που χρησιμοποιούν έναν αλγόριθμο για την ακριβή ανάθεση δεδομένων ελέγχου (test data) σε συγκεκριμένες κατηγορίες. Αναγνωρίζει συγκεκριμένες οντότητες μέσα στο σύνολο δεδομένων και προσπαθεί να βγάλει κάποια συμπεράσματα σχετικά με το πώς αυτές οι οντότητες θα πρέπει να

επισημανθούν ή να οριστούν. Οι συνήθεις αλγόριθμοι ταξινόμησης είναι οι γραμμικοί ταξινομητές (linear classifiers), οι μηχανές διανυσμάτων υποστήριξης (SVM), τα δέντρα απόφασης (decision trees), ο k-κοντινότερος γείτονας (k-nearest neighbour) και το τυχαίο δάσος (random forest). Ένα παράδειγμα πραγματικού προβλήματος ταξινόμησης είναι όταν κάποιος ιατρός θέλει να αποφασίσει, βάσει κάποιων συγκεκριμένων χαρακτηριστικών, εάν οι ασθενείς είναι κατάλληλοι για να κάνουν μια επέμβαση.

- Σε προβλήματα παλινδρόμησης (regression) που χρησιμοποιούνται για την κατανόηση της σχέσης μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Ο πιο δημοφιλής αλγόριθμος παλινδρόμησης είναι η γραμμική παλινδρόμηση. Ένα παράδειγμα πραγματικού προβλήματος παλινδρόμησης στον κλάδο της υγείας, είναι όταν οι ιατροί προσπαθούν να κατανοήσουν τη σχέση μεταξύ της δοσολογίας των φαρμάκων και της αρτηριακής πίεσης.



Σχήμα 4.2: Διαγραμματική απεικόνιση των προβλημάτων ομαδοποίησης και παλινδρόμησης (Πηγή: https://www.researchgate.net/figure/Classification-vs-Regression_fig2_350993856)

4.2.2 Μη εποπτευόμενη μάθηση (Unsupervised learning)

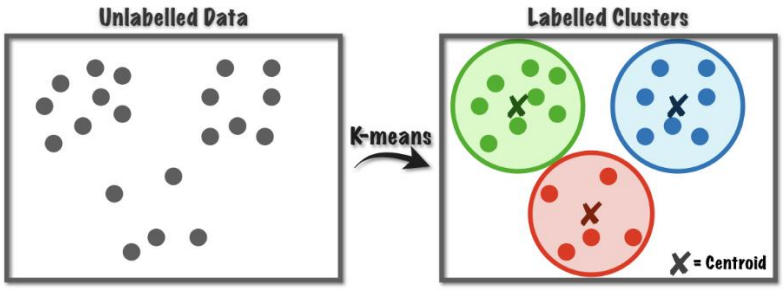
Η μη επιβλεπόμενη μάθηση, χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για την ανάλυση και ομαδοποίηση μη επισημασμένων (unlabelled) συνόλων δεδομένων. Αυτοί οι αλγόριθμοι ανακαλύπτουν κρυμμένα μοτίβα ή ομαδοποιήσεις δεδομένων χωρίς την ανάγκη ανθρώπινης παρέμβασης.

Η μη εποπτευόμενη μάθηση χρησιμοποιείται στις ακόλουθες περιπτώσεις προβλημάτων:

- Σε προβλήματα κατηγοριοποίησης ή, αλλιώς, ομαδοποίησης (clustering) που χρησιμοποιούνται για την επεξεργασία ακατέργαστων, μη ταξινομημένων δεδομένων

σε ομάδες που αντιπροσωπεύονται από δομές ή μοτίβα (patterns). Ένας από τους δημοφιλέστερους αλγορίθμους ομαδοποίησης είναι ο αλγόριθμος ομαδοποίησης k μέσων (k-means clustering). Ένα παράδειγμα πραγματικού προβλήματος ομαδοποίησης, θα μπορούσε να είναι η κατηγοριοποίηση δεδομένων που έχουν παρθεί απευθείας από Holter, ώστε να βρεθούν χρήσιμες πληροφορίες μέσω μοτίβων που θα αναγνωριστούν από τον αλγόριθμο.

- Σε προβλήματα μείωσης των διαστάσεων (Dimensionality Reduction) των οποίων οι τεχνικές χρησιμοποιούνται όταν ο αριθμός των χαρακτηριστικών ή διαστάσεων σε ένα σύνολο δεδομένων είναι πολύ υψηλός. Μειώνει τον αριθμό των εισερχόμενων δεδομένων σε ένα διαχειρίσιμο μέγεθος, διατηρώντας παράλληλα όσο το δυνατόν περισσότερο την ακεραιότητα του συνόλου δεδομένων. Χρησιμοποιείται συνήθως στο στάδιο της προεπεξεργασίας δεδομένων και υπάρχουν μερικές διαφορετικές μέθοδοι μείωσης διαστάσεων που μπορούν να χρησιμοποιηθούν. Μία από τις πιο γνωστές μεθόδους είναι η ανάλυση κύριων συνιστωσών (principal components analysis).



Σχήμα 4.3: Διαγραμματική απεικόνιση της διαδικασίας ομαδοποίησης με k-means
(Πηγή: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c)

4.2.3 Ημι-εποπτευόμενη μάθηση (semi-supervised learning)

Η μάθηση με ημι-επίβλεψη είναι ένα πρόβλημα μάθησης που περιλαμβάνει μικρό αριθμό επισημασμένων δεδομένων και μεγάλο αριθμό μη επισημασμένων δεδομένων.

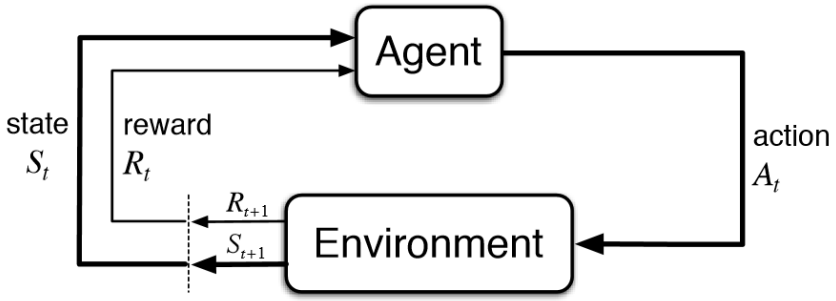
Τα προβλήματα μάθησης αυτού του τύπου αποτελούν πρόκληση, καθώς ούτε οι αλγόριθμοι μάθησης με επίβλεψη ούτε οι αλγόριθμοι μάθησης χωρίς επίβλεψη είναι σε θέση να αξιοποιήσουν αποτελεσματικά τα μείγματα επισημασμένων και μη επισημασμένων δεδομένων (labelled and unlabelled data). Ως εκ τούτου, απαιτούνται εξειδικευμένοι αλγόριθμοι μάθησης με ημιεπίβλεψη.

Ένα συνηθισμένο παράδειγμα εφαρμογής της μάθησης με ημιεπίβλεψη είναι ένας ταξινομητής εγγράφων κειμένου. Αυτό είναι ένα παράδειγμα όπου η μάθηση με ημιεπίβλεψη είναι ιδανική, επειδή θα ήταν σχεδόν αδύνατο να βρεθεί ένας μεγάλος αριθμός εγγράφων κειμένου με ετικέτες. Αυτό συμβαίνει απλώς επειδή δεν είναι αποδοτικό, από άποψη χρόνου, να διαβάζει ένα άτομο ολόκληρα έγγραφα κειμένου μόνο και μόνο για να του αναθέσει μια απλή ταξινόμηση.

Έτσι, η μάθηση με ημι-επίβλεψη επιτρέπει στον αλγόριθμο να μαθαίνει από έναν μικρό αριθμό εγγράφων κειμένου με ετικέτες, ενώ παράλληλα ταξινομεί έναν μεγάλο αριθμό εγγράφων κειμένου χωρίς ετικέτες στα δεδομένα εκπαίδευσης.

4.2.4 Ενισχυτική μάθηση (Reinforcement learning)

Πρόκειται για τη λήψη κατάλληλων αποφάσεων για τη μεγιστοποίηση της ανταμοιβής σε μια συγκεκριμένη κατάσταση. Χρησιμοποιείται από διάφορα λογισμικά και μηχανές για να βρουν την καλύτερη δυνατή συμπεριφορά ή πορεία που πρέπει να ακολουθήσουν σε μια συγκεκριμένη κατάσταση. Η ενισχυτική μάθηση διαφέρει από την επιβλεπόμενη μάθηση με την έννοια ότι, στην επιβλεπόμενη μάθηση τα δεδομένα εκπαίδευσης έχουν μαζί τους το κλειδί της απάντησης, οπότε το μοντέλο εκπαιδεύεται με την ίδια τη σωστή απάντηση, ενώ στην ενισχυτική μάθηση δεν υπάρχει απάντηση αλλά ο ενισχυτικός παράγοντας αποφασίζει τι πρέπει να κάνει για να εκτελέσει τη δεδομένη εργασία. Ελλείψει συνόλου δεδομένων εκπαίδευσης, είναι υποχρεωμένος να μάθει από την εμπειρία του.



Σχήμα 4.4: Διαγραμματική απεικόνιση της διαδικασίας της ενισχυτικής μάθησης
(Πηγή: https://www.researchgate.net/figure/Figura-1-Reinforcement-Learning-Illustration-Tratto-da-Sutton-e-Barto-2014_fig1_342866258)

Παρακάτω θα αναφερθούν δύο αναλυτικά παραδείγματα για την καλύτερη κατανόηση της ενισχυτικής μάθησης.

Έστω μια εταιρεία που διαθέτει στόλο μη επανδρωμένων αεροχημάτων παράδοσης (drones). Η εταιρεία θέλει να βελτιστοποιήσει τις διαδρομές των μη επανδρωμένων αεροσκαφών της για να ελαχιστοποιήσει τους χρόνους παράδοσης και να μεγιστοποιήσει την αποδοτικότητα. Για την επίτευξη αυτού του στόχου, η εταιρεία θα μπορούσε να χρησιμοποιήσει ενισχυτική μάθηση.

Σε αυτό το σενάριο, τα μη επανδρωμένα αεροσκάφη είναι οι πράκτορες (agents), το περιβάλλον (environment) είναι το δίκτυο παράδοσης και οι ενέργειες των μη επανδρωμένων αεροσκαφών είναι οι κινήσεις τους μέσα στο δίκτυο. Τα μη επανδρωμένα αεροσκάφη μπορούν να παρατηρούν την τρέχουσα θέση τους και τη θέση των πακέτων που πρέπει να παραδώσουν. Ο στόχος είναι να μάθουν μια πολιτική που μεγιστοποιεί τον συνολικό αριθμό των πακέτων που παραδίδονται ανά μονάδα χρόνου.

Για να εκπαιδεύσει τα μη επανδρωμένα αεροσκάφη χρησιμοποιώντας ενισχυτική μάθηση, η εταιρεία θα ξεκινούσε αρχικοποιώντας τις πολιτικές τους τυχαία. Στη συνέχεια, θα έβαζαν τα drones να πετούν επανειλημμένα μέσα στο δίκτυο παράδοσης, λαμβάνοντας ενέργειες και λαμβάνοντας ανατροφοδότηση με τη μορφή σήματος ανταμοιβής (reward). Το σήμα ανταμοιβής θα μπορούσε να βασίζεται σε παράγοντες όπως ο χρόνος παράδοσης, η διανυθείσα απόσταση και η αποδοτικότητα των καυσίμων.

Με την πάροδο του χρόνου, τα μη επανδρωμένα αεροσκάφη θα μάθαιναν ποιες ενέργειες οδηγούν στις υψηλότερες ανταμοιβές σε κάθε κατάσταση και θα ενημέρωναν τις πολιτικές τους αναλόγως. Αυτή η διαδικασία ενημέρωσης της πολιτικής με βάση την ανατροφοδότηση ονομάζεται ενισχυτική μάθηση. Καθώς τα μη επανδρωμένα αεροσκάφη συνεχίζουν να πετούν στο δίκτυο παράδοσης και να λαμβάνουν ανατροφοδότηση, οι πολιτικές τους θα συγκλίνουν σταδιακά σε μια βέλτιστη πολιτική που μεγιστοποιεί τον αριθμό των πακέτων που παραδίδονται ανά μονάδα χρόνου.

Μόλις τα μη επανδρωμένα αεροσκάφη μάθουν τη βέλτιστη πολιτική, μπορούν να τη χρησιμοποιήσουν για να περιηγηθούν στο δίκτυο παράδοσης και να παραδώσουν πακέτα χωρίς ανθρώπινη παρέμβαση. Αυτό μπορεί να οδηγήσει σε σημαντική εξοικονόμηση κόστους και βελτίωση των χρόνων παράδοσης για την εταιρεία.

Εμβαθύνοντας περισσότερο στο αντικείμενο της μελέτης της παρούσας εργασίας, θα αναφερθεί ένα ακόμη πραγματικό παράδειγμα ενισχυτικής μάθησης που έχει να κάνει με τη χρήση βαθιάς εγκεφαλικής διέγερσης κλειστού βρόχου (DBS) σε ασθενείς με ΝΠ.

Ένα πραγματικό παράδειγμα ενισχυτικής μάθησης στη ΝΠ, λοιπόν, είναι η χρήση της βαθιάς εγκεφαλικής διέγερσης κλειστού βρόχου (DBS) για τη μείωση των κινητικών συμπτωμάτων. Η DBS είναι μια χειρουργική επέμβαση κατά την οποία εμφυτεύονται ηλεκτρόδια στον εγκέφαλο για την παροχή ηλεκτρικής διέγερσης σε στοχευμένες περιοχές, με στόχο τη μείωση των συμπτωμάτων όπως το τρέμουλο και η ακαμψία. Ωστόσο, οι βέλτιστες παράμετροι διέγερσης μπορεί να διαφέρουν από ασθενή σε ασθενή και να αλλάζουν με την πάροδο του χρόνου, καθιστώντας δύσκολη την επίτευξη σταθερού ελέγχου των συμπτωμάτων.

Για την αντιμετώπιση αυτής της πρόκλησης, οι ερευνητές έχουν διερευνήσει τη χρήση της ενισχυτικής μάθησης για τη βελτιστοποίηση των παραμέτρων διέγερσης σε πραγματικό χρόνο. Σε αυτή την προσέγγιση, η συσκευή DBS συνδέεται με ένα σύστημα υπολογιστή που χρησιμοποιεί αλγόριθμους ενισχυτικής μάθησης για να προσαρμόζει τις παραμέτρους διέγερσης με βάση την ανταπόκριση του ασθενούς. Στόχος είναι η εκμάθηση μιας πολιτικής διέγερσης που μεγιστοποιεί τον έλεγχο των συμπτωμάτων και ελαχιστοποιεί τις παρενέργειες.

Σε αρκετές μελέτες, οι ερευνητές χρησιμοποίησαν την ενισχυτική μάθηση για τη βελτιστοποίηση των παραμέτρων DBS σε μια ομάδα ασθενών με ΝΠ. Οι ασθενείς φορούσαν έναν αισθητήρα που παρακολουθούσε τα κινητικά τους συμπτώματα και ο αλγόριθμος ενισχυτικής μάθησης προσάρμοζε τις παραμέτρους διέγερσης σε απόκριση στις αλλαγές της σοβαρότητας των συμπτωμάτων. Τα αποτελέσματα έδειξαν ότι η προσέγγιση της ενισχυτικής μάθησης μπόρεσε να επιτύχει αποτελεσματικότερο έλεγχο των συμπτωμάτων από τις παραδοσιακές μεθόδους προγραμματισμού DBS και ήταν επίσης σε θέση να προσαρμοστεί στις αλλαγές των συμπτωμάτων των ασθενών με την πάροδο του χρόνου.

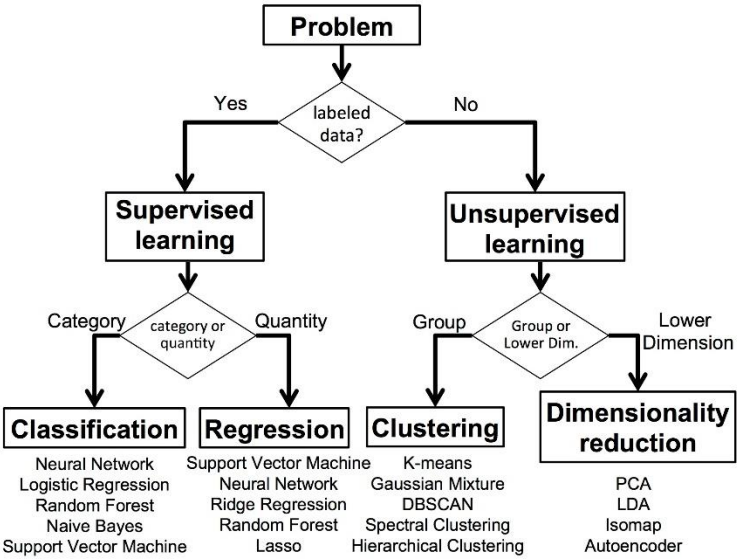
Συνολικά, η χρήση της ενισχυτικής μάθησης για τη βελτιστοποίηση των παραμέτρων του DBS υπόσχεται ένα πιθανό εργαλείο για τη βελτίωση της θεραπείας της ΝΠ και θα μπορούσε να οδηγήσει σε πιο εξατομικευμένες και αποτελεσματικές θεραπείες για τους ασθενείς με αυτή την πάθηση.

4.3 Αλγόριθμοι μηχανικής μάθησης

Ο τομέας της μηχανικής μάθησης περιλαμβάνει ένα ευρύ φάσμα αλγορίθμων, καθένας από τους οποίους έχει σχεδιαστεί για να εξυπηρετεί έναν διαφορετικό τύπο προβλήματος. Η επιλογή του κατάλληλου αλγορίθμου και η βελτιστοποίηση της απόδοσής του εξαρτάται από τον τύπο και τα χαρακτηριστικά των διαθέσιμων δεδομένων, καθώς και από τη φύση του εκάστοτε προβλήματος. Ως εκ τούτου, η κατανόηση των βασικών αλγορίθμων τόσο στην

επιβλεπόμενη όσο και στη μη επιβλεπόμενη μηχανική μάθηση είναι ζωτικής σημασίας για κάθε επιστήμονα δεδομένων ή επαγγελματία.

Με την κατανόηση των θεμελιωδών εννοιών και αρχών πίσω από αυτούς τους αλγορίθμους, οι επιστήμονες δεδομένων και οι επαγγελματίες μπορούν να επιλέξουν τον κατάλληλο αλγόριθμο για την επίλυση ενός συγκεκριμένου προβλήματος και να βελτιστοποιήσουν την απόδοσή του για μέγιστη ακρίβεια και αποτελεσματικότητα.



Σχήμα 4.5: Διαγραμματική απεικόνιση των αλγορίθμων της εποπτευόμενης και της μη εποπτευόμενης μάθησης

(Πηγή: https://subashsigdel.com.np/Re/research.html)

4.3.1 Τεχνικές ταξινόμησης (Classification methods)

Η ταξινόμηση είναι ένα θεμελιώδες πρόβλημα στη μηχανική μάθηση και περιλαμβάνει την πρόβλεψη της κλάσης ή της κατηγορίας των δεδομένων. Οι μέθοδοι ταξινόμησης είναι αλγόριθμοι που χρησιμοποιούν επισημασμένα δεδομένα (labelled data) για να μάθουν ένα όριο απόφασης που μπορεί να χρησιμοποιηθεί για την ταξινόμηση νέων, μη επισημασμένων δεδομένων (unlabelled data) σε μία ή περισσότερες προκαθορισμένες κλάσεις.

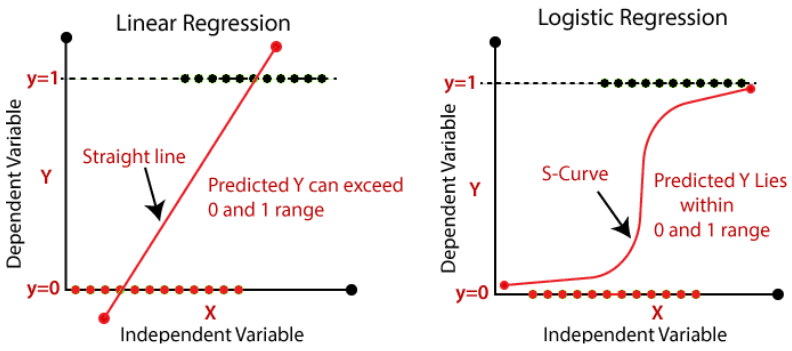
Συνολικά, οι μέθοδοι ταξινόμησης είναι ισχυρά εργαλεία για την επίλυση ποικίλων προβλημάτων του πραγματικού κόσμου και μπορούν να παράσχουν πολύτιμες πληροφορίες για πολύπλοκα δεδομένα.

Στην παρούσα εργασία θα αναλυθούν οι παρακάτω μέθοδοι ταξινόμησης:

- Λογιστική παλινδρόμηση (Logistic Regression)
- Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis, LDA)
- Κ Κοντινότεροι Γείτονες (K-Nearest Neighbours, KNN)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM)
- Extreme Gradient Boosting (XGBoost)
- Δένδρα απόφασης (Decision Trees)
- Τυχαίο Δάσος (Random Forests)

4.3.1.1 Λογιστική παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση (Logistic Regression) είναι μια στατιστική μέθοδος που χρησιμοποιείται για την ανάλυση της σχέσης μεταξύ μιας κατηγορικής εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Χρησιμοποιείται συνήθως στη μηχανική μάθηση και την ανάλυση δεδομένων για προβλήματα δυαδικής ταξινόμησης (binary classification), όπου ο στόχος είναι να προβλεφθεί η πιθανότητα εμφάνισης ενός γεγονότος (π.χ. εάν ένας ασθενής είναι κατάλληλος για επέμβαση ή όχι).



Σχήμα 4.6: Διαγραμματική απεικόνιση της διαφοράς μεταξύ Γραμμικής και Λογιστικής Παλινδρόμησης (Πηγή: https://www.researchgate.net/figure/Liner-regrassionand-logistic-regression-H-Logistic-regression-Formula_fig5_348705771)

Το μοντέλο λογιστικής παλινδρόμησης λειτουργεί εφαρμόζοντας έναν μετασχηματισμό σε έναν γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών. Αυτός ο μετασχηματισμός είναι γνωστός ως λογιστική συνάρτηση, η οποία αντιστοιχίζει οποιαδήποτε είσοδο πραγματικής τιμής σε μια τιμή πιθανότητας μεταξύ 0 και 1. Η λογιστική συνάρτηση ορίζεται ως εξής:

$$f(z) = \frac{1}{1 + e^{-z}}$$

όπου f(z) είναι η πιθανότητα να συμβεί το γεγονός, z είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών και e είναι η βάση του φυσικού λογαρίθμου.

Ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών παριστάνεται ως εξής:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

όπου β_0 είναι η τεταγμένη, $\beta_1, \beta_2, \ldots, \beta_n$ είναι οι συντελεστές των ανεξάρτητων μεταβλητών $x_1, x_2, \ldots, x_n$, αντίστοιχα.

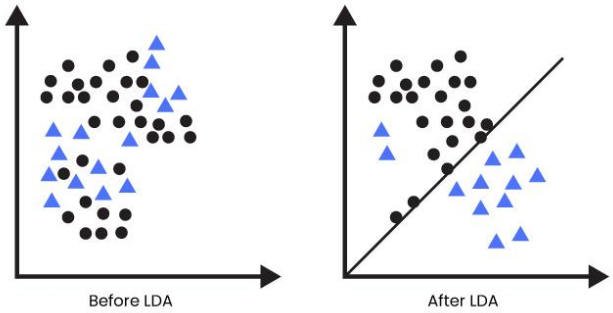
Το μοντέλο λογιστικής παλινδρόμησης εκτιμά τις τιμές των συντελεστών $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ μεγιστοποιώντας την πιθανότητα των παρατηρούμενων δεδομένων. Η συνάρτηση πιθανότητας μετρά την πιθανότητα παρατήρησης των δεδομένων με δεδομένες τις τιμές των συντελεστών. Ο στόχος είναι να βρεθούν οι τιμές των συντελεστών που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας, κάτι που μπορεί να επιτευχθεί με τη χρήση διαφόρων αλγορίθμων βελτιστοποίησης.

Μόλις εκτιμηθούν οι συντελεστές, το μοντέλο λογιστικής παλινδρόμησης μπορεί να χρησιμοποιηθεί για την πρόβλεψη της πιθανότητας εμφάνισης του γεγονότος για νέα δεδομένα. Εάν η προβλεπόμενη πιθανότητα είναι μεγαλύτερη ή ίση με μια τιμή κατωφλίου (συνήθως 0.5), το συμβάν προβλέπεται να συμβεί. Διαφορετικά, προβλέπεται ότι δεν θα συμβεί.

Η λογιστική παλινδρόμηση έχει πολλά πλεονεκτήματα, όπως η απλότητα, η ερμηνευσιμότητα και η ικανότητά της να χειρίζεται τόσο αριθμητικές όσο και κατηγορικές ανεξάρτητες μεταβλητές. Ωστόσο, έχει επίσης ορισμένους περιορισμούς, όπως η παραδοχή της γραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών και των λογαριθμικών αποδόσεων της εξαρτημένης μεταβλητής, καθώς και η ευαισθησία της σε ακραίες τιμές και πολυσυγγραμμικότητα. Παρ' όλα αυτά, η λογιστική παλινδρόμηση παραμένει ένα ευρέως χρησιμοποιούμενο και χρήσιμο εργαλείο στην ανάλυση δεδομένων και τη μηχανική μάθηση.

4.3.1.2 Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis, LDA)

Η γραμμική διακριτική ανάλυση (Linear Discriminant Analysis) είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη που χρησιμοποιείται για ταξινόμηση και μείωση διαστάσεων. Πρόκειται για μια στατιστική μέθοδο που αποσκοπεί στον εντοπισμό των υποκείμενων παραγόντων που διακρίνουν δύο ή περισσότερες κατηγορίες δεδομένων. Η LDA είναι μια παραμετρική μέθοδος, που σημαίνει ότι κάνει ορισμένες υποθέσεις σχετικά με τα δεδομένα και την κατανομή των κλάσεων.



Σχήμα 4.7: Διαγραμματική απεικόνιση της Γραμμικής Διακριτικής Ανάλυσης (LDA)
(Πηγή: http://pen.ius.edu.ba/index.php/pen/article/viewFile/2507/1023)

Υπάρχουν δύο τύποι γραμμικής διακριτικής ανάλυσης:

- LDA δύο κλάσεων: Χρησιμοποιείται όταν υπάρχουν μόνο δύο κλάσεις δεδομένων προς ταξινόμηση. Στην περίπτωση αυτή, η LDA μειώνει τη διαστατικότητα των δεδομένων σε μία διάσταση και βρίσκει το καλύτερο κατώφλι για την ταξινόμηση των δεδομένων.
- LDA πολλαπλών κλάσεων: Χρησιμοποιείται όταν υπάρχουν περισσότερες από δύο κλάσεις δεδομένων προς ταξινόμηση. Στην περίπτωση αυτή, η LDA μειώνει τη διαστατικότητα των δεδομένων στον αριθμό των κλάσεων μείον μία.

Ο αλγόριθμος LDA λειτουργεί υπολογίζοντας πρώτα το μέσο διάνυσμα και τον πίνακα συνδιακύμανσης για κάθε κλάση σημείου. Το μέσο διάνυσμα αντιπροσωπεύει τη μέση τιμή κάθε χαρακτηριστικού για κάθε κλάση, ενώ ο πίνακας συνδιακύμανσης μετρά τη διακύμανση μεταξύ των κλάσεων και τη διακύμανση εντός των κλάσεων.

Ο πίνακας διασποράς υπολογίζεται στη συνέχεια ως το άθροισμα των πινάκων συνδιακύμανσης για κάθε κλάση. Αυτός ο πίνακας μετρά τη διακύμανση μεταξύ των κλάσεων

και μπορεί να χρησιμοποιηθεί για την εύρεση των κατευθύνσεων μέγιστης διακύμανσης, γνωστών ως ιδιοδιανύσματα.

Τα ιδιοδιανύσματα του πίνακα διασποράς ταξινομούνται στη συνέχεια σε φθίνουσα σειρά με βάση τις αντίστοιχες ιδιοτιμές τους, οι οποίες αντιπροσωπεύουν τα μεγέθη των κατευθύνσεων της μέγιστης διακύμανσης. Τα ιδιοδιανύσματα με τις υψηλότερες ιδιοτιμές επιλέγονται ως οι κατευθύνσεις του υποχώρου προβολής, ο οποίος χρησιμοποιείται για τον μετασχηματισμό των δεδομένων σε έναν χώρο χαμηλότερης διάστασης.

Τα μετασχηματισμένα δεδομένα χρησιμοποιούνται στη συνέχεια για την εκπαίδευση ενός ταξινομητή, για την πρόβλεψη της κλάσης των νέων δεδομένων.

Η LDA έχει αρκετά πλεονεκτήματα, συμπεριλαμβανομένης της ικανότητάς της να χειρίζεται πολλαπλές κλάσεις και της τάσης της να παράγει πιο σταθερά και ερμηνεύσιμα αποτελέσματα σε σύγκριση με άλλες μεθόδους μείωσης της διαστατικότητας. Υποθέτει επίσης ότι τα δεδομένα ακολουθούν κανονική κατανομή, η οποία αποτελεί κοινή υπόθεση σε πολλά στατιστικά μοντέλα.

Ωστόσο, η LDA έχει ορισμένους περιορισμούς, όπως η ευαισθησία της σε ακραίες τιμές και η παραδοχή της για ίσους πίνακες συνδιακύμανσης σε όλες τις κλάσεις. Μπορεί επίσης να είναι λιγότερο αποτελεσματική όταν ο αριθμός των κλάσεων είναι μεγάλος ή όταν οι κλάσεις δεν είναι καλά διαχωρισμένες στον αρχικό χώρο χαρακτηριστικών.

Παρ' όλα αυτά, η LDA παραμένει ένα ευρέως χρησιμοποιούμενο και χρήσιμο εργαλείο στη μηχανική μάθηση και την ανάλυση δεδομένων για εργασίες ταξινόμησης και μείωσης διαστάσεων.

Η Διακριτική Διαχωριστική Ανάλυση στην πράξη

Το πρώτο βήμα είναι ο υπολογισμός της διαχωριστικότητας μεταξύ των διαφόρων κλάσεων (δηλαδή της απόστασης μεταξύ των μέσων τιμών των διαφόρων κλάσεων) που ονομάζεται και διακύμανση μεταξύ των κλάσεων:

$$S_b = \sum_{i=1}^{d} N_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$$

Το δεύτερο βήμα είναι ο υπολογισμός της απόστασης μεταξύ της μέσης τιμής και του δείγματος κάθε κλάσης, η οποία ονομάζεται διακύμανση εντός της κλάσης:

$$S_w = \sum_{i=1}^{d} (N_i - 1) S_i = \sum_{i=1}^{d} \sum_{j=1}^{N_i} \left(\bar{X}_{i,j} - \bar{X}\right)\left(\bar{X}_{i,j} - \bar{X}\right)^T$$

Το τρίτο βήμα είναι η κατασκευή του χώρου χαμηλότερης διάστασης που μεγιστοποιεί τη διακύμανση μεταξύ των κλάσεων και ελαχιστοποιεί τη διακύμανση εντός των κλάσεων. Έστω, λοιπόν, P η προβολή του χώρου χαμηλότερης διάστασης, η οποία είναι γνωστή ως κριτήριο του Fisher:

$$J(P) = \frac{P^T S_b P}{P^T S_W P}$$

4.3.1.3 Κ Κοντινότεροι Γείτονες (K Nearest Neighbours)

Ο αλγόριθμος Κ πλησιέστεροι γείτονες (K Nearest Neighbours) είναι ένας αλγόριθμος ταξινόμησης που προβλέπει την κλάση ενός σημείου (data point) με βάση την κλάση των k πλησιέστερων γειτόνων του στα δεδομένα εκπαίδευσης, όπου k είναι μια υπερπαράμετρος που ορίζεται από τον χρήστη. Ο KNN είναι μια μη παραμετρική μέθοδος, που σημαίνει ότι δεν κάνει υποθέσεις σχετικά με την κατανομή των δεδομένων ή τη σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου.

Ο αλγόριθμος KNN λειτουργεί υπολογίζοντας πρώτα την απόσταση μεταξύ του νέου σημείου των δεδομένων και κάθε σημείου των δεδομένων εκπαίδευσης. Τα συνήθη μέτρα απόστασης περιλαμβάνουν την Ευκλείδεια απόσταση, την απόσταση Manhattan και την απόσταση Minkowski. Το μέτρο απόστασης που χρησιμοποιείται μπορεί να έχει σημαντικό αντίκτυπο στην απόδοση του αλγορίθμου.

Αφού υπολογιστούν οι αποστάσεις, προσδιορίζονται οι k πλησιέστεροι γείτονες του νέου σημείου με βάση τις αποστάσεις τους. Στη συνέχεια, η κλάση του νέου σημείου δεδομένων καθορίζεται με ψηφοφορία πλειοψηφίας μεταξύ των κλάσεων των k πλησιέστερων γειτόνων του. Εάν k = 1, το νέο σημείο δεδομένων λαμβάνει απλώς την κλάση του πλησιέστερου γείτονά του.

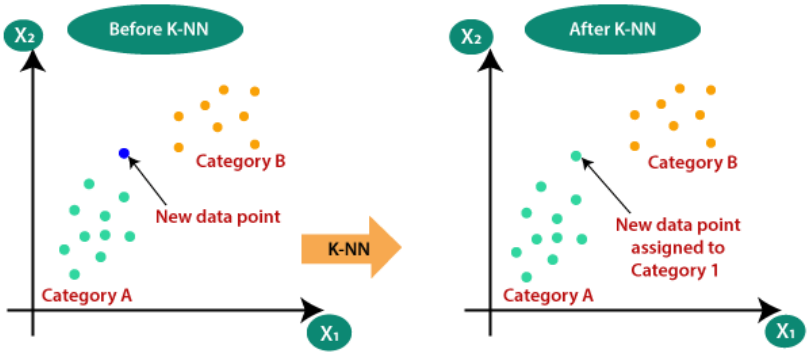
Η επιλογή του k είναι μια σημαντική υπερπαράμετρος στο KNN, καθώς επηρεάζει την αντιστάθμιση μεροληψίας-διακύμανσης του μοντέλου. Μια μικρή τιμή του k μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting), όπου το μοντέλο είναι πολύ πολύπλοκο και προσαρμόζεται στο θόρυβο (noise) των δεδομένων εκπαίδευσης. Μια μεγάλη τιμή του k μπορεί να οδηγήσει σε υποπροσαρμογή (underfitting), όπου το μοντέλο είναι πολύ απλό και

αποτυγχάνει να συλλάβει τα υποκείμενα μοτίβα στα δεδομένα. Η βέλτιστη τιμή του k εξαρτάται από το συγκεκριμένο σύνολο δεδομένων και το πρόβλημα που επιλύεται.

Το KNN έχει πολλά πλεονεκτήματα για εργασίες ταξινόμησης, συμπεριλαμβανομένης της απλότητας, της ευελιξίας και της ικανότητάς του να χειρίζεται μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου.

Ωστόσο, ο KNN έχει επίσης ορισμένους περιορισμούς, όπως η ευαισθησία του στην επιλογή του k, η οποία μπορεί να επηρεάσει την απόδοση του αλγορίθμου. Είναι επίσης υπολογιστικά δαπανηρός και απαιτητικός σε μνήμη, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων.

Παρ' όλα αυτά, ο KNN παραμένει ένα ευρέως χρησιμοποιούμενο και χρήσιμο εργαλείο στη μηχανική μάθηση και την ανάλυση δεδομένων για εργασίες ταξινόμησης. Είναι ιδιαίτερα χρήσιμο όταν οι υποκείμενες σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου είναι πολύπλοκες ή δεν είναι καλά καθορισμένες.



Σχήμα 4.8: Διαγραμματική απεικόνιση της μεθόδου KNN

(Πηγή: https://www.researchgate.net/figure/Illustration-of-KNN-technique-adapted-from-JavaTpoint_fig5_361156312)

Οι Κ Κοντινότεροι Γείτονες στην πράξη

Προκειμένου να προσδιοριστεί ποια σημεία δεδομένων βρίσκονται πλησιέστερα σε ένα δεδομένο σημείο, θα πρέπει να υπολογιστεί η απόσταση μεταξύ του σημείου και των άλλων σημείων των δεδομένων. Αυτές οι μετρικές απόστασης βοηθούν στη διαμόρφωση ορίων απόφασης, τα οποία χωρίζουν τα σημεία που μελετώνται σε διαφορετικές περιοχές.

Υπάρχουν διάφορα μέτρα απόστασης που μπορούν να χρησιμοποιηθούν, μερικά από τα οποία παρατίθενται παρακάτω:

- **Ευκλείδεια απόσταση**: Περιορίζεται σε διανύσματα πραγματικών τιμών. Χρησιμοποιώντας τον παρακάτω τύπο, μετρά μια ευθεία γραμμή μεταξύ του σημείου που μελετάται και του άλλου σημείου που μετράται.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **Απόσταση Manhattan**: Αυτή είναι επίσης μια άλλη δημοφιλής μετρική απόστασης, η οποία μετρά την απόλυτη τιμή μεταξύ δύο σημείων. Αναφέρεται επίσης ως «απόσταση ταξί» ή «απόσταση οικοδομικού τετραγώνου», καθώς συνήθως απεικονίζεται με ένα πλέγμα, απεικονίζοντας τον τρόπο με τον οποίο μπορεί κανείς να πλοηγηθεί από μια διεύθυνση σε μια άλλη μέσω των δρόμων της πόλης.

$$d(x,y) = (\sum_{i=1}^{m}|x_i - y_i|)$$

- **Απόσταση Minkowski**: Αυτό το μέτρο απόστασης είναι η γενικευμένη μορφή των μετρικών της Ευκλείδειας απόστασης και της απόστασης Μανχάταν. Η παράμετρος, p, στον παρακάτω τύπο, επιτρέπει τη δημιουργία άλλων μετρικών απόστασης. Η ευκλείδεια απόσταση παριστάνεται με αυτόν τον τύπο όταν το p=2, ενώ η απόσταση Μανχάταν συμβολίζεται με p=1.

$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|\right)^{1/p}$$

- **Απόσταση Hamming**: Αυτή η τεχνική χρησιμοποιείται συνήθως με διανύσματα Boolean ή συμβολοσειρών, εντοπίζοντας τα σημεία όπου τα διανύσματα δεν ταιριάζουν. Ως εκ τούτου, αναφέρεται επίσης ως μετρική «επικάλυψης».

$$D_H = (\sum_{i=1}^{k}|x_i - y_i|)$$

4.3.1.4 Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines) είναι ένας δημοφιλής αλγόριθμος μηχανικής μάθησης με επίβλεψη που χρησιμοποιείται για προβλήματα ταξινόμησης. Είναι ένας γραμμικός και μη γραμμικός αλγόριθμος διάκρισης που βασίζεται στην εύρεση ενός υπερεπιπέδου σε έναν χώρο χαρακτηριστικών υψηλής διάστασης που

διαχωρίζει τα δεδομένα σε διαφορετικές κλάσεις με τον μέγιστο δυνατό τρόπο. Το υπερεπίπεδο επιλέγεται έτσι ώστε να μεγιστοποιεί το περιθώριο, το οποίο είναι η απόσταση μεταξύ του υπερεπιπέδου και των πλησιέστερων σημείων των δεδομένων από κάθε κλάση. Τα σημεία των δεδομένων που βρίσκονται πλησιέστερα στο υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης και ορίζουν το υπερεπίπεδο.

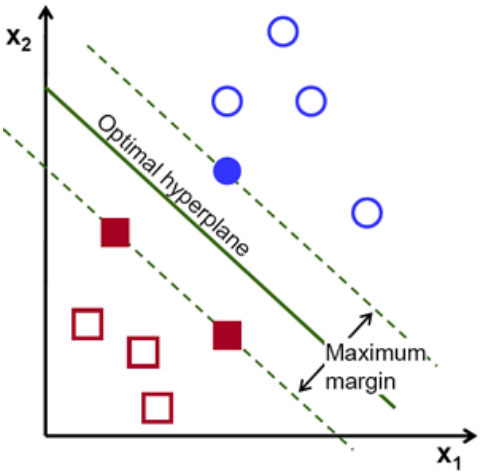
Το SVM λειτουργεί μετασχηματίζοντας πρώτα τα δεδομένα σε έναν χώρο χαρακτηριστικών υψηλότερων διαστάσεων, όπου είναι πιο πιθανό να είναι γραμμικά διαχωρίσιμα. Αυτός ο μετασχηματισμός γίνεται με τη χρήση μιας συνάρτησης πυρήνα, η οποία απεικονίζει τα δεδομένα στον χώρο υψηλότερων διαστάσεων χωρίς στην πραγματικότητα να υπολογίζει τις συντεταγμένες των δεδομένων στον εν λόγω χώρο. Οι συνήθεις συναρτήσεις πυρήνα περιλαμβάνουν γραμμικούς, πολυωνυμικούς και πυρήνες συναρτήσεων ακτινωτής βάσης (RBF).

Αφού μετασχηματιστούν τα δεδομένα, το SVM βρίσκει το υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων. Αυτό γίνεται με την επίλυση ενός προβλήματος βελτιστοποίησης με περιορισμούς, όπου ο στόχος είναι η μεγιστοποίηση του περιθωρίου με τον περιορισμό ότι τα σημεία των δεδομένων ταξινομούνται σωστά.

Η SVM έχει πολλά πλεονεκτήματα για προβλήματα ταξινόμησης. Μπορεί να χειριστεί τόσο γραμμικά όσο και μη γραμμικά διαχωρίσιμα δεδομένα, χάρη στη χρήση συναρτήσεων πυρήνα. Ο SVM είναι επίσης λιγότερο επιρρεπής σε υπερπροσαρμογή από άλλους αλγορίθμους ταξινόμησης, καθώς επιδιώκει να μεγιστοποιήσει το περιθώριο μεταξύ των κλάσεων αντί να προσαρμόζει στενά τα δεδομένα.

Ωστόσο, ο SVM έχει επίσης ορισμένους περιορισμούς, όπως η ευαισθησία του στην επιλογή της συνάρτησης πυρήνα και στις υπερπαραμέτρους του αλγορίθμου, όπως η παράμετρος κανονικοποίησης και οι παράμετροι του πυρήνα. Η SVM μπορεί επίσης να είναι υπολογιστικά δαπανηρή, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων ή πολύπλοκες συναρτήσεις πυρήνα. Τέλος, ο SVM μπορεί να είναι δύσκολο να ερμηνευτεί, καθώς το υπερεπίπεδο στον χώρο χαρακτηριστικών υψηλής διάστασης μπορεί να μην αντιστοιχεί άμεσα σε ένα απλό όριο απόφασης στον αρχικό χώρο χαρακτηριστικών.

Παρόλα αυτά, η SVM παραμένει ένα ευρέως χρησιμοποιούμενο και χρήσιμο εργαλείο στη μηχανική μάθηση και την ανάλυση δεδομένων για εργασίες ταξινόμησης. Είναι ιδιαίτερα χρήσιμο όταν οι υποκείμενες σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου είναι πολύπλοκες ή δεν είναι καλά καθορισμένες.

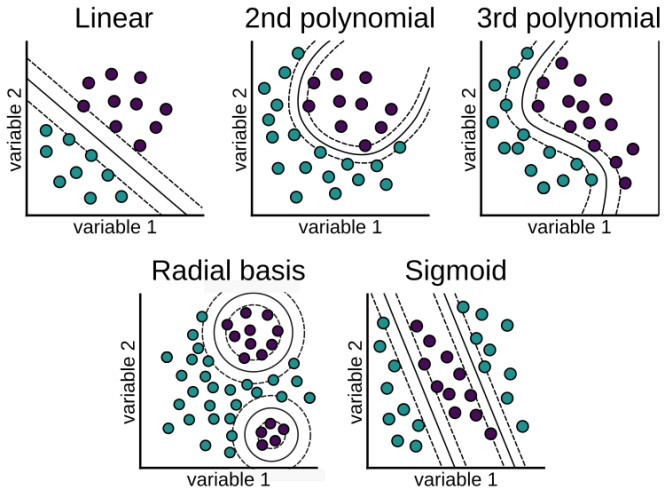


Σχήμα 4.9: Γραφική απεικόνιση της επιλογής του μέγιστου περιθωρίου μεταξύ των σημείων δεδομένων των κατηγοριών

(Πηγή: https://www.researchgate.net/figure/Optimal-Hyperplane-and-Margin-of-SVM_fig3_338698374)

Το τέχνασμα του πυρήνα (The Kernel Trick)

Είναι μέθοδοι τις οποίες οι γραμμικοί ταξινομητές, όπως ο SVM, χρησιμοποιούν για να ταξινομήσουν μη γραμμικά διαχωρίσιμα σημεία δεδομένων. Αυτό γίνεται με την αναπαράσταση των σημείων δεδομένων σε έναν χώρο υψηλότερης διάστασης από τον αρχικό του. Για παράδειγμα, ένα 1D δεδομένο μπορεί να αναπαρασταθεί ως 2D δεδομένο στο χώρο, ένα 2D δεδομένο μπορεί να αναπαρασταθεί ως 3D δεδομένο κ.λπ. Το SVM αναπαριστά έξυπνα μη γραμμικά σημεία δεδομένων χρησιμοποιώντας οποιαδήποτε από τις συναρτήσεις πυρήνα με τρόπο που να φαίνεται ότι τα δεδομένα έχουν μετασχηματιστεί, και στη συνέχεια βρίσκει το βέλτιστο διαχωριστικό υπερεπίπεδο. Ωστόσο, στην πραγματικότητα, τα σημεία δεδομένων εξακολουθούν να παραμένουν τα ίδια, δεν έχουν μετασχηματιστεί στην πραγματικότητα. Γι' αυτό το λόγο ονομάζεται "τέχνασμα του πυρήνα".



Σχήμα 4.10: Γραφική απεικόνιση των τεχνασμάτων των πυρήνων του SVM

(Πηγή: https://assets.researchsquare.com/files/rs-2117151/v1_covered.pdf?c=1668738576)

Υπάρχουν διάφορες λειτουργίες που χρησιμοποιεί το SVM για να εκτελέσει αυτό το έργο. Μερικές από τις πιο συνηθισμένες είναι οι εξής:

- **Πολυωνυμική συνάρτηση πυρήνα (Polynomial Kernel Function):** Αυτή μετασχηματίζει τα σημεία δεδομένων χρησιμοποιώντας το γινόμενο τελείας και μετασχηματίζοντας τα δεδομένα σε μια "n-διάσταση", το n θα μπορούσε να είναι οποιαδήποτε τιμή από 2, 3 κ.ο.κ., δηλαδή ο μετασχηματισμός θα είναι είτε τετραγωνικό γινόμενο είτε υψηλότερο. Συνεπώς, τα δεδομένα αναπαρίστανται σε χώρο υψηλότερων διαστάσεων χρησιμοποιώντας τα νέα μετασχηματισμένα σημεία.
- **Η συνάρτηση ακτινωτής βάσης (Radial Basis Function, RBF):** Αυτή η συνάρτηση συμπεριφέρεται σαν ένα "σταθμισμένο μοντέλο κοντινότερου γείτονα". Μετασχηματίζει τα δεδομένα αναπαριστώντας τα σε άπειρες διαστάσεις και στη συνέχεια χρησιμοποιεί τον σταθμισμένο κοντινότερο γείτονα (παρατήρηση με τη μεγαλύτερη επιρροή στο νέο σημείο δεδομένων) για ταξινόμηση. Η ακτινωτή συνάρτηση μπορεί να είναι είτε Gaussian είτε Laplace. Αυτό εξαρτάται από μια υπερπαράμετρο γνωστή ως γάμμα. Αυτός είναι ο πιο συχνά χρησιμοποιούμενος πυρήνας.
- **Η σιγμοειδής συνάρτηση (Sigmoid Function):** επίσης γνωστή ως συνάρτηση σιγμοειδούς εφαπτομένης, βρίσκει περισσότερη εφαρμογή στα νευρωνικά δίκτυα ως συνάρτηση ενεργοποίησης.

- **Ο γραμμικός πυρήνας (Linear Kernel):** Χρησιμοποιείται για γραμμικά δεδομένα. Αυτός, απλά αναπαριστά τα σημεία δεδομένων χρησιμοποιώντας μια γραμμική σχέση.

Όνομα πυρήνα	Μαθηματικός τύπος	Παράμετροι
Γραμμικός πυρήνας	$k(x_i, x_j) = w^T x_j$	-
Πολυωνυμικός πυρήνας	$k(x_i, x_j) = (\gamma w^T x_j + r)^d$	γ, r, d
RBF πυρήνας	$k(x_i, x_j) = exp(\gamma \|x_i - x_j\|^2)$	γ>0
Σιγμοειδής πυρήνας	$k(x_i, x_j) = tanh(\gamma w^T x_j + r)$	γ>0, r

Πίνακας 4.1: Πυρήνες (kernels) της μεθόδου SVM

4.3.1.5 Extreme Gradient Boosting (XGBoost)

Ο XGBoost (Extreme Gradient Boosting) είναι ένας ισχυρός και ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης για προβλήματα ταξινόμησης. Είναι μια επέκταση του αλγορίθμου Gradient Boosting, ο οποίος συνδυάζει πολλαπλούς αδύναμους ταξινομητές για τη δημιουργία ενός ισχυρού ταξινομητή. Ο XGBoost βελτιώνει τον παραδοσιακό αλγόριθμο Gradient Boosting προσθέτοντας ορισμένα βασικά χαρακτηριστικά, όπως η παράλληλη επεξεργασία, η κανονικοποίηση και ο χειρισμός των ελλειπόντων δεδομένων.

Ο αλγόριθμος XGBoost λειτουργεί προσθέτοντας επαναληπτικά «αδύναμους μαθητές» στο μοντέλο, με κάθε μαθητή να επικεντρώνεται στη διόρθωση των σφαλμάτων του προηγούμενου μαθητή. Οι αδύναμοι μαθητές που χρησιμοποιούνται στον XGBoost είναι δέντρα απόφασης, τα οποία είναι απλά μοντέλα που χωρίζουν αναδρομικά τα δεδομένα σε υποσύνολα με βάση τις τιμές των χαρακτηριστικών εισόδου. Η δομή του δέντρου μαθαίνεται μέσω μιας διαδικασίας που ονομάζεται gradient boosting, η οποία περιλαμβάνει την ελαχιστοποίηση μιας συνάρτησης απώλειας που μετρά το σφάλμα των προβλέψεων του μοντέλου.

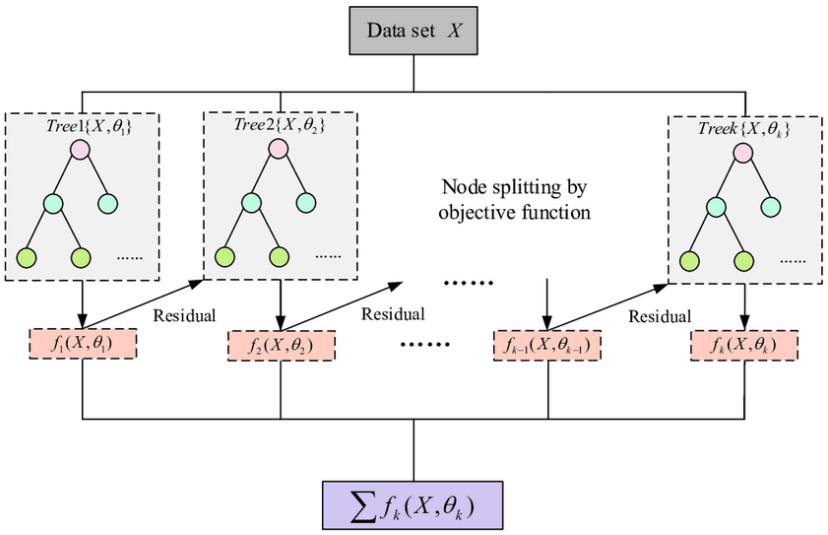
Ένα από τα βασικά πλεονεκτήματα του XGBoost είναι η ικανότητά του να χειρίζεται μεγάλα και πολύπλοκα σύνολα δεδομένων. Το XGBoost χρησιμοποιεί παράλληλη επεξεργασία για να επιταχύνει τη διαδικασία εκπαίδευσης, καθιστώντας εφικτή την εργασία με πολύ μεγάλα σύνολα δεδομένων. Επιπλέον, το XGBoost περιλαμβάνει διάφορες τεχνικές κανονικοποίησης για την αποτροπή της υπερπροσαρμογής, η οποία μπορεί να συμβεί όταν το μοντέλο γίνεται

υπερβολικά πολύπλοκο και προσαρμόζεται στο θόρυβο των δεδομένων και όχι στο υποκείμενο μοτίβο.

Ένα άλλο σημαντικό χαρακτηριστικό του XGBoost είναι η ικανότητά του να χειρίζεται τις ελλείπουσες τιμές (missing data). Πολλά σύνολα δεδομένων του πραγματικού κόσμου περιέχουν ελλείπουσες τιμές, γεγονός που μπορεί να προκαλέσει προβλήματα σε ορισμένους αλγορίθμους μηχανικής μάθησης. Ο XGBoost έχει σχεδιαστεί για να χειρίζεται τις ελλείπουσες τιμές μαθαίνοντας αυτόματα πώς να κάνει προβλέψεις όταν λείπουν ορισμένα χαρακτηριστικά.

Εκτός από τα τεχνικά χαρακτηριστικά του, ο XGBoost είναι επίσης γνωστός για την υψηλή ακρίβεια και την ευελιξία του. Έχει εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα προβλημάτων ταξινόμησης, συμπεριλαμβανομένης της ταξινόμησης εικόνων, της επεξεργασίας φυσικής γλώσσας και της ανίχνευσης απάτης.

Συνοψίζοντας, ο XGBoost είναι ένας ισχυρός και ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης για προβλήματα ταξινόμησης. Συνδυάζει τα δυνατά σημεία των δέντρων απόφασης, της ενίσχυσης κλίσης, της παράλληλης επεξεργασίας, της κανονικοποίησης και του χειρισμού ελλειπόντων δεδομένων για τη δημιουργία ενός εξαιρετικά ακριβούς και ευέλικτου μοντέλου.



Σχήμα 4.11: Διαγραμματική απεικόνιση της μεθόδου XGBoost

(Πηγή: https://www.researchgate.net/figure/Flow-chart-of-XGBoost_fig3_345327934)

4.3.1.6 Δένδρα απόφασης (Decision Trees)

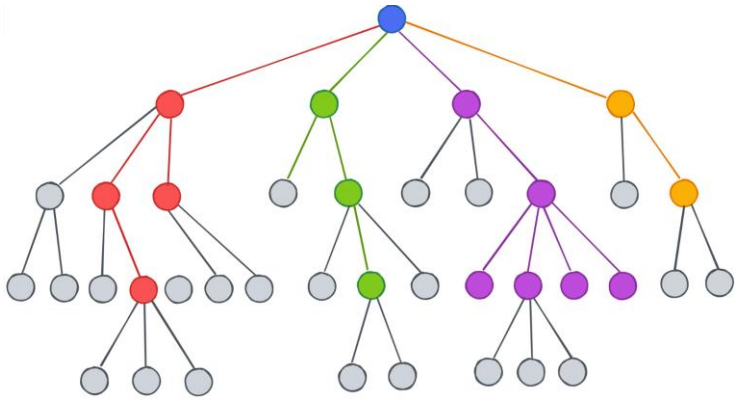
Ο ταξινομητής δέντρων απόφασης είναι ένας δημοφιλής αλγόριθμος που χρησιμοποιείται στη μηχανική μάθηση για προβλήματα ταξινόμησης. Είναι ένας τύπος αλγορίθμου μάθησης με επίβλεψη που χρησιμοποιεί ένα δενδροειδές μοντέλο αποφάσεων και των πιθανών συνεπειών

τους για να προβλέψει την τάξη ή την κατηγορία μιας παρατήρησης με βάση τα χαρακτηριστικά εισόδου της.

Σε ένα δέντρο αποφάσεων, οι εσωτερικοί κόμβοι αναπαριστούν χαρακτηριστικά ή ιδιότητες των δεδομένων και οι άκρες αναπαριστούν τις πιθανές τιμές ή τα εύρη αυτών των χαρακτηριστικών. Τα φύλλα του δέντρου αναπαριστούν τις κλάσεις ή τις κατηγορίες που πρέπει να προβλεφθούν. Το δέντρο κατασκευάζεται με αναδρομική διάσπαση των δεδομένων σε μικρότερα υποσύνολα με βάση τις τιμές των χαρακτηριστικών, έως ότου τα υποσύνολα είναι καθαρά, δηλαδή περιέχουν μόνο παρατηρήσεις μιας μόνο κλάσης.

Για να χρησιμοποιήσουμε ένα δέντρο απόφασης για ταξινόμηση, ξεκινάμε από τον κόμβο-ρίζα και ακολουθούμε τη διαδρομή προς τα κάτω στο δέντρο με βάση τις τιμές των χαρακτηριστικών εισόδου. Σε κάθε εσωτερικό κόμβο, λαμβάνουμε μια απόφαση με βάση το χαρακτηριστικό και την τιμή που αντιπροσωπεύει ο συγκεκριμένος κόμβος και ακολουθούμε την αντίστοιχη ακμή προς τον επόμενο κόμβο. Συνεχίζουμε να κατεβαίνουμε το δέντρο μέχρι να φτάσουμε σε έναν κόμβο φύλλου, ο οποίος αντιπροσωπεύει την προβλεπόμενη κλάση για την παρατήρηση.

Ο αλγόριθμος του δέντρου αποφάσεων μπορεί να χειριστεί τόσο κατηγορικά όσο και αριθμητικά δεδομένα και μπορεί να χειριστεί αυτόματα τις ελλείπουσες τιμές επιλέγοντας την πιο κοινή τιμή ή προβλέποντας την τιμή με βάση άλλα χαρακτηριστικά.



Σχήμα 4.12: Διαγραμματική απεικόνιση του Δένδρου Απόφασης

4.3.1.7 Τυχαία Δάση (Random Forests)

Η ταξινόμηση τυχαίου δάσους είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης. Πρόκειται για μια τεχνική μάθησης συνόλου που συνδυάζει πολλαπλά δέντρα αποφάσεων για να βελτιώσει την ακρίβεια των προβλέψεων.

Σε ένα τυχαίο δάσος, κάθε δέντρο απόφασης εκπαιδεύεται σε ένα τυχαίο υποσύνολο των δεδομένων εισόδου και σε κάθε κόμβο του δέντρου, μόνο ένα τυχαίο υποσύνολο των χαρακτηριστικών λαμβάνεται υπόψη για διαχωρισμό. Αυτή η τυχαιότητα συμβάλλει στη μείωση της υπερπροσαρμογής και στην αύξηση της ποικιλομορφίας των δέντρων στο δάσος.

Κατά την πραγματοποίηση μιας πρόβλεψης, το τυχαίο δάσος συνδυάζει τις προβλέψεις όλων των μεμονωμένων δέντρων απόφασης για να παράγει μια τελική πρόβλεψη. Η τελική πρόβλεψη καθορίζεται συνήθως με ψηφοφορία πλειοψηφίας, όπου η πιο συχνά προβλεπόμενη κλάση σε όλα τα δέντρα απόφασης επιλέγεται ως τελική έξοδος.

4.3.2 Τεχνικές παλινδρόμησης (Regression methods)

Οι μέθοδοι παλινδρόμησης είναι μια κατηγορία τεχνικών μηχανικής μάθησης που χρησιμοποιούνται για τη μοντελοποίηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής (που ονομάζεται επίσης μεταβλητή απόκρισης) και μιας ή περισσότερων ανεξάρτητων μεταβλητών (που ονομάζονται επίσης επεξηγηματικές μεταβλητές).

Στόχος της ανάλυσης παλινδρόμησης είναι η εκτίμηση των παραμέτρων του μοντέλου παλινδρόμησης που προβλέπουν καλύτερα την τιμή της εξαρτημένης μεταβλητής για ένα δεδομένο σύνολο μεταβλητών εισόδου. Οι μέθοδοι παλινδρόμησης χρησιμοποιούνται συνήθως για την πρόβλεψη, την πρόγνωση και την κατανόηση της σχέσης μεταξύ των μεταβλητών σε ένα σύνολο δεδομένων.

Στην παρούσα εργασία, θα αναλυθούν οι παρακάτω μέθοδοι παλινδρόμησης:

- Γραμμική παλινδρόμηση (Linear Regression)
- Παλινδρόμηση LASSO (LASSO Regression)
- Παλινδρόμηση Κορυφογραμμής (Ridge Regression)
- Παλινδρόμηση με Τυχαία Δάση (Random Forest Regression)
- Παλινδρόμηση με Δένδρα Απόφασης (Decision Trees Regression)
- Παλινδρόμηση Gradient Boosting

4.3.2.1 Γραμμική παλινδρόμηση (Linear Regression)

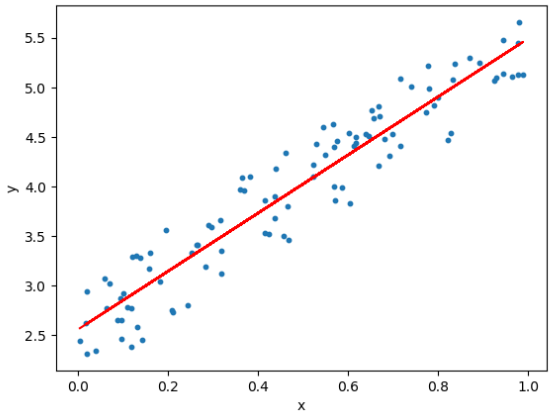
Η γραμμική παλινδρόμηση (Linear Regression) είναι μια στατιστική τεχνική που χρησιμοποιείται για τη μοντελοποίηση της γραμμικής σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών, όπου όταν υπάρχει μία

ανεξάρτητη μεταβλητή τότε πραγματοποιείται απλή γραμμική παλινδρόμηση, ενώ όταν υπάρχουν παραπάνω, πραγματοποιείται πολλαπλή γραμμική παλινδρόμηση. Είναι μια απλή και ευρέως χρησιμοποιούμενη μέθοδος παλινδρόμησης που υποθέτει γραμμική σχέση μεταξύ των μεταβλητών. Στη γραμμική παλινδρόμηση, η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής μοντελοποιείται χρησιμοποιώντας μια γραμμική εξίσωση της μορφής:

$$Y = a + \beta X_i$$

όπου Y είναι η εξαρτημένη μεταβλητή, x_1, x_2, ..., x_n είναι οι ανεξάρτητες μεταβλητές, β_0 είναι ο σταθερός όρος, και β_1, β_2, ..., β_n είναι οι συντελεστές παλινδρόμησης που αντιπροσωπεύουν τη μεταβολή στο y για μια μοναδιαία μεταβολή σε καθεμία από τις ανεξάρτητες μεταβλητές.

Στόχος της γραμμικής παλινδρόμησης είναι η εκτίμηση των τιμών των συντελεστών που ελαχιστοποιούν το άθροισμα των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών της εξαρτημένης μεταβλητής. Αυτό είναι γνωστό ως η μέθοδος των ελαχίστων τετραγώνων.



Σχήμα 4.13: Διαγραμματική απεικόνιση της προσαρμογής της ευθείας ελάχιστων τετραγώνων στην απλή γραμμική παλινδρόμηση

Η εξίσωση που πρέπει να ελαχιστοποιηθεί είναι η:

$$\sum_{i=1}^{\nu} \varepsilon_i^2 = \sum_{i=1}^{\nu} (y_i - \alpha \pm \beta x_i)^2$$

Οι τιμές α και β που ελαχιστοποιούν την παραπάνω εξίσωση, ονομάζονται αμερόληπτες εκτιμήτριες ελαχίστων τετραγώνων και υπολογίζονται από τις παρακάτω σχέσεις:

$$\hat{\beta} = \frac{\nu \sum_{i=1}^{\nu} x_i y_i - (\sum_{i=1}^{\nu} x_i)(\sum_{i=1}^{\nu} y_i)}{\nu \sum_{i=1}^{\nu} x_i^2 - (\sum_{i=1}^{\nu} x_i)^2}$$, όπου $\bar{y} = \frac{1}{\nu}(\sum_{i=1}^{\nu} y_i)$, $\bar{x} = \frac{1}{\nu}(\sum_{i=1}^{\nu} x_i)$

και $\hat{a} = \bar{y} - \hat{\beta}\bar{x}$

Η γραμμική παλινδρόμηση έχει αρκετά πλεονεκτήματα, όπως η απλότητα, η ερμηνευσιμότητα και η ανθεκτικότητα σε ακραίες τιμές. Είναι επίσης ένα χρήσιμο εργαλείο για την πραγματοποίηση προβλέψεων και την κατανόηση της σχέσης μεταξύ μεταβλητών. Ωστόσο, έχει ορισμένους περιορισμούς, όπως η παραδοχή της γραμμικότητας, η οποία μπορεί να μην ισχύει για ορισμένα σύνολα δεδομένων, και η αδυναμία της να αποτυπώσει μη γραμμικές σχέσεις μεταξύ μεταβλητών.

Η γραμμική παλινδρόμηση χρησιμοποιείται ευρέως σε διάφορους τομείς, συμπεριλαμβανομένων των οικονομικών, των χρηματοοικονομικών, της μηχανικής και των κοινωνικών επιστημών. Μπορεί να χρησιμοποιηθεί για πρόβλεψη, πρόγνωση και έλεγχο υποθέσεων.

4.3.2.2 Παλινδρόμηση LASSO (LASSO Regression)

Η παλινδρόμηση Lasso (Least absolute Shrinkage and Selection Operator) είναι ένας τύπος γραμμικής παλινδρόμησης που χρησιμοποιεί συρρίκνωση. Με τη συρρίκνωση οι τιμές των δεδομένων συρρικνώνονται προς ένα κεντρικό σημείο, όπως η μέση τιμή. Η διαδικασία Lasso ενθαρρύνει απλά, αραιά μοντέλα (δηλαδή μοντέλα με λιγότερες παραμέτρους). Αυτός ο συγκεκριμένος τύπος παλινδρόμησης ενδείκνυται για μοντέλα που παρουσιάζουν υψηλά επίπεδα μεταβλητότητας ή όταν χρειάζεται να γίνει επιλογή μεταβλητών ή εξάλειψη παραμέτρων.

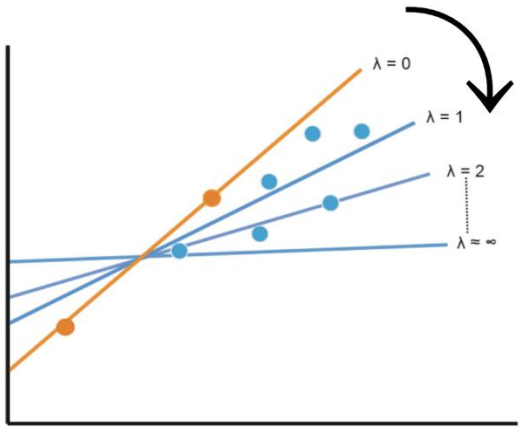
Η παλινδρόμηση Lasso εκτελεί κανονικοποίηση L1, η οποία προσθέτει μια ποινή ίση με την απόλυτη τιμή του μεγέθους των συντελεστών. Αυτός ο τύπος μπορεί να οδηγήσει σε αραιά μοντέλα με λίγους συντελεστές- ορισμένοι συντελεστές μπορούν να μηδενιστούν και να εξαλειφθούν από το μοντέλο. Μεγαλύτερες ποινές έχουν ως αποτέλεσμα τιμές συντελεστών πιο κοντά στο μηδέν, το οποίο είναι το ιδανικό για την παραγωγή απλούστερων μοντέλων. Από την άλλη πλευρά, η κανονικοποίηση L2 (π.χ. παλινδρόμηση Ridge, που θα δούμε παρακάτω) δεν οδηγεί σε εξάλειψη συντελεστών ή αραιά μοντέλα. Αυτό καθιστά την Lasso πολύ πιο εύκολη στην ερμηνεία από την Ridge.

Η συνάρτηση που πρέπει να ελαχιστοποιηθεί, είναι η παρακάτω:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Μια παράμετρος ρύθμισης, το λ, ελέγχει την ισχύ της ποινής L1. Δηλαδή, το λ είναι ουσιαστικά το μέγεθος της συρρίκνωσης.

- Όταν το λ = 0, δεν εξαλείφεται καμία παράμετρος. Η εκτίμηση είναι ίση με εκείνη που θα βρισκόταν με με γραμμική παλινδρόμηση.
- Καθώς το λ αυξάνεται, όλο και περισσότεροι συντελεστές μηδενίζονται και εξαλείφονται (θεωρητικά, όταν λ = ∞, όλοι οι συντελεστές εξαλείφονται).
- Καθώς αυξάνεται το λ, η μεροληψία αυξάνεται.
- Καθώς μειώνεται το λ, αυξάνεται η διακύμανση.



Σχήμα 4.14: Διαγραμματική απεικόνιση της παραμέτρου λ (σε διάφορες περιπτώσεις) στην παλινδρόμηση LASSO

4.3.2.3 Παλινδρόμηση κορυφογραμμής (Ridge Regression)

Η παλινδρόμηση κορυφογραμμής (Ridge Regression) είναι μια στατιστική μέθοδος που χρησιμοποιείται για την ανάλυση δεδομένων στα οποία οι ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό. Πρόκειται για μια τεχνική που χρησιμοποιείται για την αποφυγή της υπερπροσαρμογής (overfitting) και τη βελτίωση της προβλεπτικής ακρίβειας ενός μοντέλου.

Στην παλινδρόμηση κορυφογραμμής, οι συντελεστές των ανεξάρτητων μεταβλητών εκτιμώνται ελαχιστοποιώντας μια συνάρτηση κόστους που περιλαμβάνει έναν όρο ποινής για μεγάλους συντελεστές. Ο όρος ποινής (penalty) ελέγχεται από μια υπερπαράμετρο που ονομάζεται λάμδα (λ), η οποία είναι μια παράμετρος ρύθμισης που καθορίζει το μέγεθος της κανονικοποίησης που εφαρμόζεται στο μοντέλο.
Η συνάρτηση που πρέπει να ελαχιστοποιηθεί, είναι:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

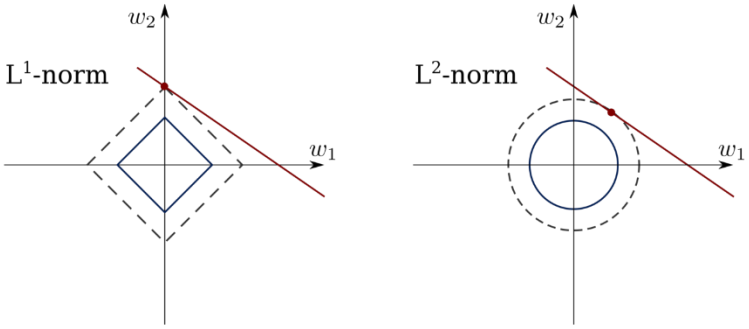
Ένα από τα πλεονεκτήματα της παλινδρόμησης κορυφογραμμής είναι ότι μπορεί να χειριστεί την πολυσυγγραμμικότητα, η οποία εμφανίζεται όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό μεταξύ τους. Η πολυσυγγραμμικότητα μπορεί να προκαλέσει προβλήματα στη γραμμική παλινδρόμηση, επειδή μπορεί να καταστήσει τους συντελεστές των ανεξάρτητων μεταβλητών ασταθείς και δύσκολα ερμηνεύσιμους. Η παλινδρόμηση Ridge μειώνει το μέγεθος των συντελεστών με την προσθήκη ενός όρου ποινής, ο οποίος συμβάλλει στον μετριασμό των επιπτώσεων της πολυσυγγραμμικότητας.

Η παλινδρόμηση Ridge είναι επίσης χρήσιμη όταν υπάρχουν περισσότεροι προγνωστικοί παράγοντες από τις παρατηρήσεις, γεγονός που αποτελεί κοινό πρόβλημα σε σύνολα δεδομένων υψηλής διάστασης. Στην περίπτωση αυτή, η συνήθης παλινδρόμηση ελαχίστων τετραγώνων (OLS) μπορεί να μην είναι σε θέση να εκτιμήσει με ακρίβεια τους συντελεστές, επειδή ο αριθμός των παραμέτρων που πρέπει να εκτιμηθούν είναι μεγαλύτερος από τον αριθμό των παρατηρήσεων. Η παλινδρόμηση Ridge μπορεί να βοηθήσει στην αντιμετώπιση αυτού του προβλήματος μειώνοντας τη διακύμανση των εκτιμήσεων.

Ένας περιορισμός της παλινδρόμησης κορυφογραμμής είναι ότι υποθέτει ότι όλες οι ανεξάρτητες μεταβλητές είναι σημαντικές για το μοντέλο. Στην πράξη, ορισμένες από τις μεταβλητές μπορεί να είναι άσχετες ή να έχουν αμελητέο αντίκτυπο στο αποτέλεσμα. Σε τέτοιες περιπτώσεις, μια πιο εξελιγμένη μέθοδος, όπως η παλινδρόμηση lasso που αναφέρθηκε προηγουμένως, μπορεί να είναι καταλληλότερη.

Συνολικά, η παλινδρόμηση κορυφογραμμής είναι ένα ισχυρό στατιστικό εργαλείο που μπορεί να βελτιώσει την ακρίβεια και τη σταθερότητα των μοντέλων γραμμικής

παλινδρόμησης, ιδίως όταν πρόκειται για πολυσυγγραμμικότητα και σύνολα δεδομένων υψηλών διαστάσεων.



Σχήμα 4.15: Διαγραμματική απεικόνιση των περιοχών περιορισμού των LASSO (αριστερά) και Ridge (δεξιά)

(Πηγή: https://en.wikipedia.org/wiki/Lasso_(statistics))

4.3.2.4 Παλινδρόμηση με τυχαία δάση (Random Forest Regression)

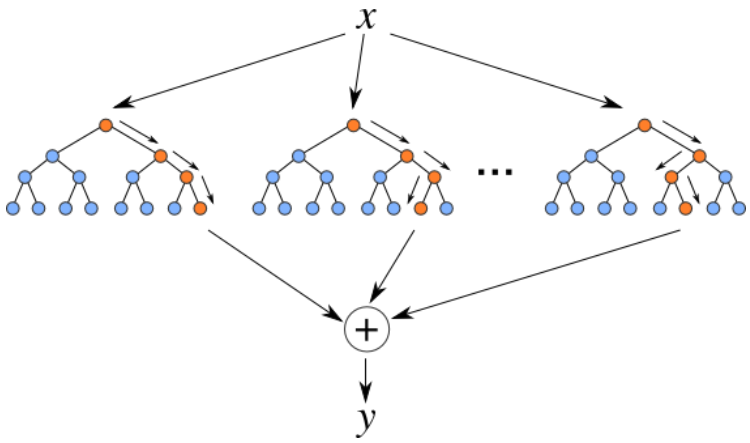
Η παλινδρόμηση τυχαίου δάσους (Random Forest Regression) είναι μια δημοφιλής τεχνική μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη συνεχών αποτελεσμάτων. Είναι μια τεχνική που συνδυάζει πολλαπλά δέντρα αποφάσεων για τη δημιουργία ενός πιο ακριβούς και σταθερού μοντέλου. Στην παλινδρόμηση τυχαίου δάσους, δημιουργείται ένας μεγάλος αριθμός δέντρων απόφασης και η τελική πρόβλεψη γίνεται με τη μέση τιμή των εξόδων όλων των δέντρων.

Η διαδικασία δημιουργίας ενός μοντέλου παλινδρόμησης τυχαίου δάσους περιλαμβάνει διάφορα βήματα. Πρώτον, επιλέγεται ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης για τη δημιουργία ενός δέντρου απόφασης. Σε κάθε κόμβο του δέντρου, επιλέγεται ένα τυχαίο υποσύνολο των χαρακτηριστικών για να βρεθεί ο καλύτερος διαχωρισμός. Η διαδικασία επαναλαμβάνεται έως ότου ικανοποιηθεί ένα κριτήριο διακοπής, όπως ένα μέγιστο βάθος δέντρου ή ένας ελάχιστος αριθμός δειγμάτων σε κάθε κόμβο φύλλου. Με αυτόν τον τρόπο δημιουργούνται πολλαπλά δέντρα απόφασης χρησιμοποιώντας διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης και των χαρακτηριστικών.

Αφού δημιουργηθούν τα δέντρα απόφασης, η τελική πρόβλεψη γίνεται με τη συγκέντρωση των αποτελεσμάτων όλων των δέντρων. Στην παλινδρόμηση τυχαίου δάσους, η έξοδος είναι η μέση τιμή των προβλεπόμενων τιμών όλων των δέντρων. Η χρήση πολλαπλών δέντρων

απόφασης και τυχαίων υποσυνόλων δεδομένων και χαρακτηριστικών βοηθά στη μείωση της υπερπροσαρμογής και στη βελτίωση της απόδοσης γενίκευσης του μοντέλου.

Ένα από τα πλεονεκτήματα της παλινδρόμησης τυχαίου δάσους είναι η ικανότητά της να χειρίζεται τόσο γραμμικές όσο και μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών εισόδου και της μεταβλητής-στόχου. Είναι επίσης ανθεκτική στις ακραίες τιμές και στο θόρυβο των δεδομένων και μπορεί να χειριστεί σύνολα δεδομένων υψηλής διάστασης με πολλά χαρακτηριστικά. Επιπλέον, η σημασία κάθε χαρακτηριστικού μπορεί να αξιολογηθεί εξετάζοντας τη συμβολή κάθε χαρακτηριστικού στην πρόβλεψη του μοντέλου.



Σχήμα 4.16: Διαγραμματική απεικόνιση τυχαίου δάσους σε προβλήματα παλινδρόμησης

4.3.2.5 Παλινδρόμηση με δένδρα απόφασης (Decision Tree Regression)

Η παλινδρόμηση δέντρων αποφάσεων (Decision Tree Regression) είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη συνεχών αποτελεσμάτων. Περιλαμβάνει τη δημιουργία ενός δενδροειδούς μοντέλου αποφάσεων και των πιθανών συνεπειών τους, με κάθε εσωτερικό κόμβο να αντιπροσωπεύει μια δοκιμή σε μια μεταβλητή εισόδου, κάθε κλάδο να αντιπροσωπεύει το αποτέλεσμα της δοκιμής και κάθε κόμβο φύλλου να αντιπροσωπεύει την τελική πρόβλεψη.

Η διαδικασία δημιουργίας ενός μοντέλου παλινδρόμησης δέντρου αποφάσεων περιλαμβάνει διάφορα βήματα. Πρώτον, τα δεδομένα εισόδου χωρίζονται σε υποσύνολα με βάση τις τιμές των μεταβλητών εισόδου. Στη συνέχεια, εκτελείται δοκιμή σε κάθε υποσύνολο για τον προσδιορισμό της μεταβλητής που διαχωρίζει καλύτερα τα δεδομένα στις κλάσεις εξόδου. Το δέντρο κατασκευάζεται με αναδρομική διάσπαση των δεδομένων σε μικρότερα υποσύνολα με βάση τα αποτελέσματα αυτών των δοκιμών, έως ότου ικανοποιηθεί ένα κριτήριο διακοπής, όπως ένα μέγιστο βάθος δέντρου ή ένας ελάχιστος αριθμός δειγμάτων σε κάθε κόμβο

φύλλου. Η τελική πρόβλεψη για μια νέα τιμή εισόδου γίνεται με τη διάσχιση του δέντρου από τη ρίζα στον κατάλληλο κόμβο φύλλου και την επιστροφή της τιμής εξόδου που σχετίζεται με αυτόν τον κόμβο.

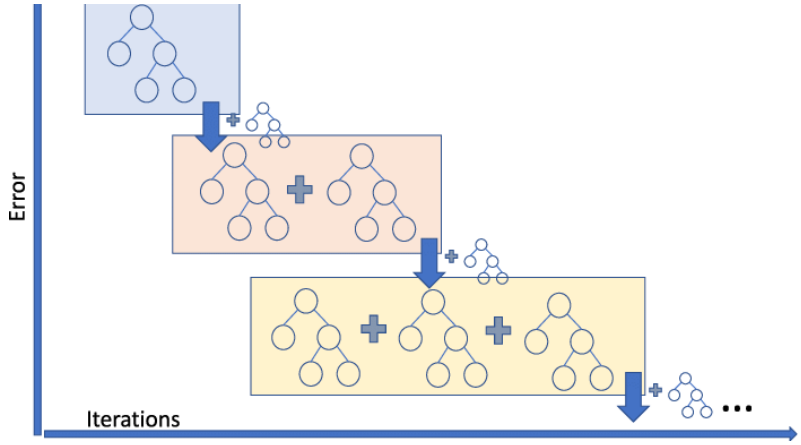
Ένα από τα πλεονεκτήματα της παλινδρόμησης με δέντρα αποφάσεων είναι η ερμηνευσιμότητά της. Το δέντρο απόφασης μπορεί να οπτικοποιηθεί και να γίνει εύκολα κατανοητό, γεγονός που μπορεί να προσφέρει πληροφορίες σχετικά με τη σχέση μεταξύ των μεταβλητών εισόδου και της μεταβλητής-στόχου. Επιπλέον, η παλινδρόμηση δέντρων αποφάσεων μπορεί να χειριστεί τόσο γραμμικές όσο και μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών εισόδου και της μεταβλητής-στόχου.

4.3.2.6 Παλινδρόμηση με Gradient Boosting

Η παλινδρόμηση Gradient Boosting περιλαμβάνει το συνδυασμό πολλαπλών αδύναμων μοντέλων, συνήθως δέντρων απόφασης, σε ένα ισχυρό μοντέλο που κάνει ακριβείς προβλέψεις.

Η διαδικασία δημιουργίας ενός μοντέλου παλινδρόμησης Gradient Boosting περιλαμβάνει διάφορα βήματα. Πρώτον, ένα αδύναμο μοντέλο, συνήθως ένα δέντρο απόφασης, εκπαιδεύεται στα δεδομένα εισόδου για να κάνει προβλέψεις. Στη συνέχεια υπολογίζεται η διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών και χρησιμοποιείται για την εκπαίδευση ενός νέου αδύναμου μοντέλου, με έμφαση στη βελτίωση των προβλέψεων όπου το προηγούμενο μοντέλο ήταν πιο αδύναμο. Αυτή η διαδικασία επαναλαμβάνεται, με κάθε νέο αδύναμο μοντέλο να επικεντρώνεται στα σφάλματα των προηγούμενων μοντέλων, μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής ή να επιτευχθεί ένας μέγιστος αριθμός μοντέλων. Η τελική πρόβλεψη για μια νέα τιμή εισόδου γίνεται με τη συγκέντρωση των προβλέψεων όλων των αδύναμων μοντέλων.

Ένα από τα πλεονεκτήματα της παλινδρόμησης Gradient Boosting είναι η ικανότητά της να χειρίζεται πολύπλοκες μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών εισόδου και της μεταβλητής-στόχου. Μπορεί επίσης να χειριστεί τις ελλείπουσες και τις ακραίες τιμές και είναι λιγότερο επιρρεπής στην υπερπροσαρμογή από άλλες τεχνικές μηχανικής μάθησης, όπως η παλινδρόμηση δέντρων απόφασης.



Σχήμα 4.17: Διαγραμματική απεικόνιση της παλινδρόμησης Gradient Boosting

(Πηγή: https://www.researchgate.net/figure/Schematical-representation-of-gradient-boosting-regression-in-regards-to-algorithm_fig3_340524896)

4.3.3 Τεχνικές μη εποπτευόμενης μάθησης

Ένα κοινό πρόβλημα στη μάθηση χωρίς επίβλεψη είναι η ομαδοποίηση, η οποία περιλαμβάνει την ομαδοποίηση παρόμοιων σημείων δεδομένων σε ομάδες ή τμήματα με βάση τις εγγενείς ιδιότητές τους. Η ομαδοποίηση μπορεί να χρησιμοποιηθεί σε ένα ευρύ φάσμα εφαρμογών, όπως η τμηματοποίηση της αγοράς, η τμηματοποίηση εικόνων και η ανίχνευση ανωμαλιών.

Υπάρχουν διάφοροι τύποι αλγορίθμων ομαδοποίησης, όπως η ιεραρχική ομαδοποίηση (hierarchical clustering), η ομαδοποίηση k-means και η ομαδοποίηση με βάση την πυκνότητα (density-based clustering). Κάθε τύπος αλγορίθμου ομαδοποίησης έχει τα δικά του πλεονεκτήματα και αδυναμίες και η επιλογή του αλγορίθμου εξαρτάται συχνά από το συγκεκριμένο πρόβλημα και τα χαρακτηριστικά των δεδομένων.

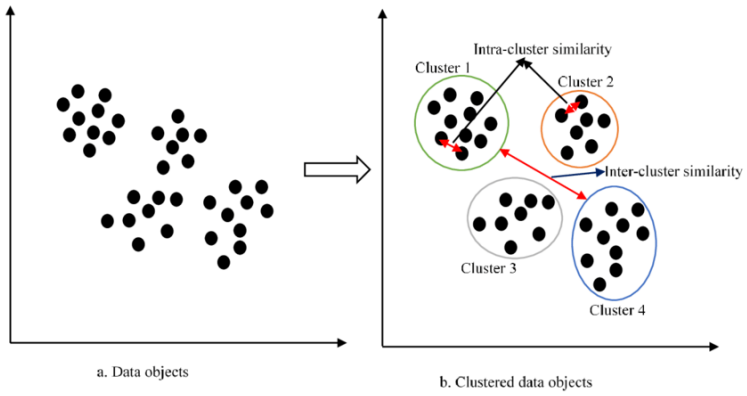
Ένα άλλο σημαντικό πρόβλημα στη μάθηση χωρίς επίβλεψη είναι η μείωση της διαστατικότητας, η οποία περιλαμβάνει τη μείωση του αριθμού των χαρακτηριστικών ή των μεταβλητών σε ένα σύνολο δεδομένων, διατηρώντας παράλληλα τις πιο σημαντικές πληροφορίες. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο για την οπτικοποίηση δεδομένων, καθώς επιτρέπει την αναπαράσταση δεδομένων υψηλής διάστασης σε χώρο χαμηλότερης διάστασης.

Υπάρχουν διάφορες τεχνικές για τη μείωση της διαστατικότητας, όπως η ανάλυση κύριων συνιστωσών (PCA), η t-κατανεμημένη στοχαστική ενσωμάτωση γειτόνων (t-SNE) και οι αυτοκωδικοποιητές. Κάθε τεχνική έχει τα δικά της πλεονεκτήματα και περιορισμούς και η επιλογή της τεχνικής εξαρτάται συχνά από τα συγκεκριμένα δεδομένα και το πρόβλημα που αντιμετωπίζουμε.

4.3.3.1 Αλγόριθμος k-means

Η ομαδοποίηση K-means είναι ένας δημοφιλής αλγόριθμος μη επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιείται για την ομαδοποίηση και την τμηματοποίηση δεδομένων. Ο αλγόριθμος αποσκοπεί στην κατάτμηση ενός δεδομένου συνόλου δεδομένων σε k διακριτές, μη επικαλυπτόμενες συστάδες με βάση την ομοιότητα των σημείων δεδομένων. Ο αλγόριθμος ξεκινά με την τυχαία επιλογή k αρχικών κεντροειδών, όπου κάθε κεντροειδές αντιπροσωπεύει το κέντρο μιας συστάδας. Στη συνέχεια, ο αλγόριθμος αναθέτει επαναληπτικά κάθε σημείο δεδομένων στο πλησιέστερο κεντροειδές, με βάση την ευκλείδεια απόσταση μεταξύ του σημείου δεδομένων και του κεντροειδούς. Αφού ανατεθούν όλα τα σημεία δεδομένων, το κεντροειδές κάθε συστάδας υπολογίζεται εκ νέου, με βάση τη μέση τιμή των ανατεθειμένων σημείων δεδομένων. Αυτή η διαδικασία ανάθεσης σημείων δεδομένων και επανυπολογισμού των κεντροειδών επαναλαμβάνεται έως ότου τα κεντροειδή συγκλίνουν και οι συστάδες σταθεροποιηθούν.

Η ομαδοποίηση K-means έχει πολλά πλεονεκτήματα, όπως η απλότητα, η επεκτασιμότητα και η ταχύτητά της. Ο αλγόριθμος είναι εύκολος στην υλοποίηση και μπορεί να χειριστεί μεγάλα σύνολα δεδομένων με πολλά χαρακτηριστικά. Επιπλέον, ο αριθμός των συστάδων k μπορεί να προσαρμοστεί ανάλογα με το εκάστοτε πρόβλημα, επιτρέποντας την ευέλικτη ομαδοποίηση με βάση το επιθυμητό επίπεδο λεπτομέρειας.



Σχήμα 4.18: Διαγραμματική απεικόνιση της ομαδοποίησης με k-means
(Πηγή: https://www.researchgate.net/figure/Clustering-example-with-intra-and-inter-clustering-illustrations_fig1_344590665)

Ωστόσο, η συσταδοποίηση k-means έχει επίσης ορισμένους περιορισμούς. Ένα σημαντικό μειονέκτημα είναι ότι ο αλγόριθμος είναι ευαίσθητος στην αρχική επιλογή των κεντροειδών,

γεγονός που μπορεί να οδηγήσει στο σχηματισμό διαφορετικών συστάδων ανάλογα με το σημείο εκκίνησης. Ως εκ τούτου, συχνά απαιτούνται πολλαπλές εκτελέσεις με διαφορετικά αρχικά κεντροειδή για να εξασφαλιστεί η σταθερότητα των συστάδων. Επιπλέον, η συσταδοποίηση k-means υποθέτει ότι οι συστάδες είναι σφαιρικές και έχουν παρόμοιες αποκλίσεις, κάτι που μπορεί να μην ισχύει πάντα στα σύνολα δεδομένων του πραγματικού κόσμου.

Ο αλγόριθμος Κ μέσων στην πράξη

Έστω ότι αναθέτουμε τυχαία τα κέντρα των ομάδων $K_1, K_2, K_3, \ldots, K_n$.

Για κάθε x_i υπολογίζουμε το κοντινότερο κέντρο K_j: $argmin_j D(x_i, K_j)$ και αναθέτουμε το x_i στη συστάδα K_j.

Στη συνέχεια, για κάθε συστάδα $K_1, K_2, K_3, \ldots, K_n$, υπολογίζουμε τα νέα γεωμετρικά τους κέντρα:

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

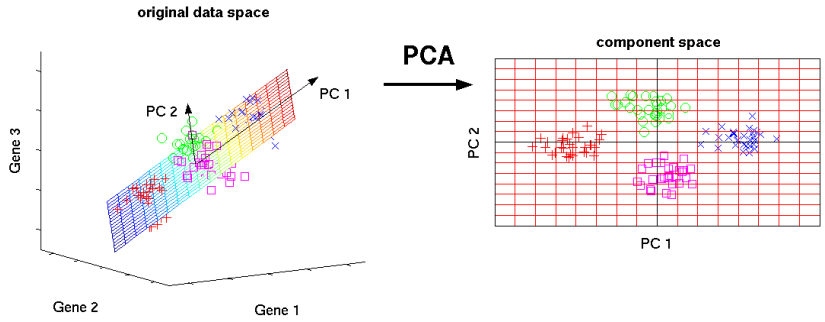
Η διαδικασία σταματάει όταν δεν υπάρχουν μεταβολές στις συστάδες, δηλαδή τα στοιχεία τους παραμένουν ίδια.

4.3.3.2 Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis)

Η Ανάλυση Κύριων Συνιστωσών (PCA) είναι μια δημοφιλής τεχνική που χρησιμοποιείται για τη μείωση της διαστατικότητας ενός συνόλου δεδομένων με τον εντοπισμό και την εξαγωγή των πιο σημαντικών μεταβλητών που αποτυπώνουν τη μεγαλύτερη διακύμανση στα δεδομένα. Με άλλα λόγια, η PCA μετασχηματίζει ένα σύνολο δεδομένων υψηλής διάστασης σε σύνολο δεδομένων χαμηλότερης διάστασης, διατηρώντας παράλληλα όσο το δυνατόν μεγαλύτερη μεταβλητότητα των δεδομένων.

Η PCA λειτουργεί με την εύρεση των κύριων συνιστωσών ενός συνόλου δεδομένων, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών που εξηγούν τη μεγαλύτερη διακύμανση στα δεδομένα. Η πρώτη κύρια συνιστώσα είναι ο γραμμικός συνδυασμός που εξηγεί τη μεγαλύτερη διακύμανση, η δεύτερη κύρια συνιστώσα είναι ο γραμμικός συνδυασμός που εξηγεί τη μεγαλύτερη διακύμανση μεταξύ των υπόλοιπων μεταβλητών μετά τη συνεκτίμηση της πρώτης κύριας συνιστώσας κ.ο.κ. Κάθε κύρια συνιστώσα είναι ορθογώνια προς τις άλλες, δηλαδή είναι ασυσχέτιστη.

Η διαδικασία εύρεσης των κύριων συνιστωσών περιλαμβάνει τον υπολογισμό του πίνακα συνδιακύμανσης των αρχικών μεταβλητών. Τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης είναι οι κύριες συνιστώσες και οι αντίστοιχες ιδιοτιμές τους αντιπροσωπεύουν το ποσοστό της διακύμανσης που εξηγείται από κάθε κύρια συνιστώσα.



Σχήμα 4.19: Διαγραμματική απεικόνιση της μεθόδου PCA

(Πηγή: https://www.researchgate.net/figure/Clustering-example-with-intra-and-inter-clustering-illustrations_fig1_344590665)

Η PCA χρησιμοποιείται συνήθως για την εξαγωγή χαρακτηριστικών στη μηχανική μάθηση, καθώς μπορεί να μειώσει τον αριθμό των μεταβλητών εισόδου, διατηρώντας παράλληλα τις περισσότερες από τις σημαντικές πληροφορίες των δεδομένων.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι η PCA υποθέτει ότι τα δεδομένα σχετίζονται γραμμικά και ενδέχεται να μην λειτουργεί καλά με μη γραμμικά δεδομένα. Επιπλέον, η ερμηνεία των κύριων συνιστωσών μπορεί να είναι δύσκολη, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων με πολλές μεταβλητές.

4.4 Τεχνικές αξιολόγησης των μοντέλων (Model evaluation techniques)

Οι τεχνικές αξιολόγησης μοντέλων χρησιμοποιούνται για την αξιολόγηση των επιδόσεων των μοντέλων μηχανικής μάθησης και για τον προσδιορισμό του πόσο καλά γενικεύονται σε νέα, αθέατα δεδομένα. Η διαδικασία αξιολόγησης μοντέλων περιλαμβάνει τη μέτρηση της ακρίβειας και της αποτελεσματικότητας του μοντέλου και τη σύγκρισή του με άλλα μοντέλα ή σημεία αναφοράς.

Υπάρχουν διάφορες κοινές τεχνικές αξιολόγησης μοντέλων που χρησιμοποιούνται στη μηχανική μάθηση, όπως:

- **Διασταυρούμενη επικύρωση (cross-validation):** Αυτή η τεχνική περιλαμβάνει τη διαίρεση των δεδομένων σε k ισομεγέθη υποσύνολα και τη χρήση μιας αναδίπλωσης ως συνόλου δοκιμής και των υπόλοιπων αναδιπλώσεων ως συνόλου εκπαίδευσης. Η διαδικασία επαναλαμβάνεται k φορές, με κάθε αναδίπλωση να χρησιμοποιείται μία φορά ως σύνολο δοκιμής. Στη συνέχεια, τα αποτελέσματα υπολογίζονται κατά μέσο όρο για να παρέχουν μια εκτίμηση της απόδοσης του μοντέλου.
- **Επικύρωση αναμονής (holdout validation):** Αυτή η τεχνική περιλαμβάνει τον τυχαίο διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, με τυπική κατανομή 70% για την εκπαίδευση και 30% για τη δοκιμή. Το μοντέλο εκπαιδεύεται στο σύνολο εκπαίδευσης και αξιολογείται στο σύνολο δοκιμών.
- **Μετρικές (metrics):** Οι μετρικές χρησιμοποιούνται για τη μέτρηση της απόδοσης ενός μοντέλου, όπως η ακρίβεια (accuracy), η ακρίβεια (precision), η ανάκληση (recall), το F1-score και η καμπύλη ROC. Αυτές οι μετρικές μπορούν να χρησιμοποιηθούν για τη σύγκριση διαφορετικών μοντέλων ή για τη ρύθμιση των υπερπαραμέτρων ενός μοντέλου.
- **Αναζήτηση σε πλέγμα (grid search):** Η αναζήτηση πλέγματος είναι μια τεχνική συντονισμού υπερπαραμέτρων, η οποία περιλαμβάνει τον ορισμό ενός πλέγματος πιθανών υπερπαραμέτρων και την αξιολόγηση της απόδοσης του μοντέλου για κάθε συνδυασμό υπερπαραμέτρων.
- **Καμπύλες μάθησης (learning curves):** Οι καμπύλες μάθησης χρησιμοποιούνται για την απεικόνιση της απόδοσης ενός μοντέλου καθώς αυξάνεται το μέγεθος του συνόλου εκπαίδευσης. Αυτό μπορεί να βοηθήσει στον εντοπισμό του κατά πόσον ένα μοντέλο προσαρμόζεται υπερβολικά ή υποπροσαρμόζεται στα δεδομένα.

Συνολικά, οι τεχνικές αξιολόγησης μοντέλων είναι απαραίτητες για να διασφαλιστεί ότι τα μοντέλα μηχανικής μάθησης είναι αποτελεσματικά και ακριβή.

4.4.1 Τεχνικές αξιολόγησης των μοντέλων ταξινόμησης

Ο πίνακας σύγχυσης (confusion matrix) είναι ένας πίνακας που χρησιμοποιείται συνήθως για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης. Βοηθά στην οπτικοποίηση των ποσοστών αληθώς θετικών, αληθώς αρνητικών, ψευδώς θετικών και ψευδώς αρνητικών ενός μοντέλου.

Αληθώς θετικό (TP): Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν σωστά ως θετικές από το μοντέλο.

Αληθινά αρνητικά (TN): Αυτό αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν σωστά ως αρνητικές από το μοντέλο.

Ψευδώς θετικό (FP): Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν ως θετικές από το μοντέλο, αλλά στην πραγματικότητα ήταν αρνητικές.

Ψευδώς αρνητικά (FN): Αναφέρεται στον αριθμό των περιπτώσεων που προβλέφθηκαν ως αρνητικές από το μοντέλο, αλλά ήταν στην πραγματικότητα θετικές.

Ο πίνακας σύγχυσης είναι συνήθως διατεταγμένος ως εξής:

	Πρόβλεψη	
Πραγματικότητα	**Θετικό**	**Αρνητικό**
Θετικό	Ορθά θετικό	Εσφαλμένα αρνητικό
Αρνητικό	Εσφαλμένα θετικό	Ορθά αρνητικό

Πίνακας 4.2: Πίνακας σύγχυσης (confusion matrix)

Τα διαγώνια στοιχεία του πίνακα αντιπροσωπεύουν τις σωστές προβλέψεις που έγιναν από το μοντέλο, ενώ τα στοιχεία εκτός διαγωνίου αντιπροσωπεύουν τις λανθασμένες προβλέψεις.

Με βάση τις τιμές στον πίνακα σύγχυσης, μπορούν να υπολογιστούν διάφορες μετρικές για την αξιολόγηση της απόδοσης του μοντέλου, όπως

Ορθότητα (Accuracy): Αυτή μετρά τη συνολική ορθότητα του μοντέλου και υπολογίζεται ως:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Ακρίβεια (Precision): Αυτό μετρά το ποσοστό των αληθώς θετικών μεταξύ όλων των θετικών προβλέψεων και υπολογίζεται ως:

$$\frac{TP}{TP + FP}$$

Ανάκληση (Recall): Αυτό μετρά την αναλογία των αληθώς θετικών μεταξύ όλων των πραγματικών θετικών προβλέψεων και υπολογίζεται ως:

$$\frac{TP}{TP + FN}$$

Βαθμολογία F1 (F-score): Πρόκειται για τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης και υπολογίζεται ως:

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Καμπύλη ROC (ROC Curve): Η καμπύλη ROC (Receiver Operating Characteristic) είναι μια γραφική αναπαράσταση της απόδοσης ενός δυαδικού μοντέλου ταξινόμησης. Παρουσιάζει το αληθώς θετικό ποσοστό (TPR) έναντι του ψευδώς θετικού ποσοστού (FPR) σε διαφορετικά κατώφλια ταξινόμησης. Η περιοχή κάτω από την καμπύλη ROC (AUC) είναι μια ευρέως χρησιμοποιούμενη μετρική για την ποσοτικοποίηση της συνολικής απόδοσης ενός μοντέλου ταξινόμησης, με ένα τέλειο μοντέλο να έχει AUC 1,0 και έναν τυχαίο ταξινομητή να έχει AUC 0,5. Η καμπύλη ROC και η AUC παρέχουν πολύτιμες πληροφορίες σχετικά με την ικανότητα του μοντέλου να διακρίνει μεταξύ θετικών και αρνητικών περιπτώσεων και μπορούν να βοηθήσουν στη βελτιστοποίηση του ορίου ταξινόμησης για μια συγκεκριμένη εφαρμογή.

Σχήμα 4.20: Καμπύλη ROC

(Πηγή: https://commons.wikimedia.org/wiki/File:Roc_curve.svg)

Ο πίνακας σύγχυσης και οι σχετικές μετρικές μπορούν να παρέχουν πολύτιμες πληροφορίες σχετικά με την απόδοση ενός μοντέλου ταξινόμησης και μπορούν να βοηθήσουν στον εντοπισμό περιοχών για βελτίωση.

4.4.2 Τεχνικές αξιολόγησης των μοντέλων παλινδρόμησης

Υπάρχουν διάφορες μετρικές αξιολόγησης που μπορούν να χρησιμοποιηθούν για την αξιολόγηση της απόδοσης των μοντέλων παλινδρόμησης, όπως:

Ο συντελεστής προσδιορισμού (R^2): Μετρά το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από τις ανεξάρτητες μεταβλητές του μοντέλου. Παίρνει μια τιμή μεταξύ 0 και 1, με υψηλότερες τιμές να υποδηλώνουν καλύτερη προσαρμογή μεταξύ του μοντέλου και των δεδομένων.

$$R^2 = 1 - \frac{SSR}{SST}$$

Όπου:

SSR = άθροισμα των τετραγώνων των υπολοίπων (προβλεπόμενες - πραγματικές τιμές)

SST = συνολικό άθροισμα τετραγώνων (πραγματικές τιμές - μέσος όρος πραγματικών τιμών)

Μέσο απόλυτο σφάλμα (MAE): Το MAE μετρά τη μέση απόλυτη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Υπολογίζεται ως ο μέσος όρος των απόλυτων διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Όπου:

n = αριθμός παρατηρήσεων

y_i = πραγματική τιμή της εξαρτημένης μεταβλητής

$\hat{y}_i$ = προβλεπόμενη τιμή της εξαρτημένης μεταβλητής

Μέσο τετραγωνικό σφάλμα (MSE): Το MSE μετρά τον μέσο όρο των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Επιβαρύνει τα μεγάλα σφάλματα περισσότερο από τα μικρότερα σφάλματα, καθιστώντας το χρήσιμο μέτρο όταν τα μεγάλα σφάλματα είναι ιδιαίτερα σημαντικά.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Ρίζα μέσου τετραγωνικού σφάλματος (RMSE): Το RMSE είναι η τετραγωνική ρίζα του MSE και χρησιμοποιείται συχνά ως μια ευκολότερη στην ερμηνεία εναλλακτική λύση του MSE.

$$RMSE = \sqrt{MSE}$$

Μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE): Το MAPE μετρά την ποσοστιαία διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Υπολογίζεται ως ο μέσος όρος των απόλυτων ποσοστιαίων διαφορών μεταξύ προβλεπόμενων και πραγματικών τιμών.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100$$

Συντελεστής προσδιορισμού (COD): Ο COD μετρά το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από τις ανεξάρτητες μεταβλητές. Είναι παρόμοιος με το R^2, αλλά παίρνει μια τιμή μεταξύ 0 και 100%.

$$COD = R^2 \cdot 100$$

Αυτές οι μετρικές παρέχουν έναν τρόπο ποσοτικοποίησης της απόδοσης των μοντέλων παλινδρόμησης και βοηθούν στον προσδιορισμό των μοντέλων που ταιριάζουν καλύτερα στα δεδομένα.

4.4.3 Διασταυρούμενη επικύρωση (Cross Validation)

Η διασταυρούμενη επικύρωση περιλαμβάνει τη διαίρεση των διαθέσιμων δεδομένων σε πολλαπλά υποσύνολα ή αναδιπλώσεις, την εκπαίδευση του μοντέλου σε ένα υποσύνολο των δεδομένων και τη δοκιμή του σε ένα άλλο υποσύνολο. Η διαδικασία επαναλαμβάνεται αρκετές φορές, με διαφορετικά υποσύνολα δεδομένων που χρησιμοποιούνται για την εκπαίδευση και τη δοκιμή, και τα αποτελέσματα υπολογίζονται κατά μέσο όρο για να προκύψει μια συνολική εκτίμηση της απόδοσης του μοντέλου.

Υπάρχουν διάφοροι τύποι διασταυρούμενης επικύρωσης, όπως:

Διασταυρούμενη επικύρωση Κ-πτυχών (K-fold cross-validation): Στην προσέγγιση αυτή, τα δεδομένα χωρίζονται σε Κ υποσύνολα ή αναδιπλώσεις ίσου μεγέθους. Το μοντέλο εκπαιδεύεται σε Κ-1 αναδιπλώσεις και δοκιμάζεται στις υπόλοιπες αναδιπλώσεις και η διαδικασία επαναλαμβάνεται Κ φορές, με κάθε αναδίπλωση να χρησιμοποιείται μία φορά ως σύνολο δοκιμής.

Διασταυρούμενη επικύρωση χωρίς αποκλεισμούς (Leave-one-out cross validation, LOOCV): Σε αυτή την προσέγγιση, κάθε παρατήρηση στο σύνολο δεδομένων χρησιμοποιείται ως σύνολο δοκιμής μία φορά, ενώ οι υπόλοιπες παρατηρήσεις χρησιμοποιούνται για την εκπαίδευση του μοντέλου.

Στρωματοποιημένη διασταυρούμενη επικύρωση k-πτυχών (stratified k-fold cross-validation): Αυτή είναι παρόμοια με τη διασταυρούμενη επικύρωση k-πτυχών, αλλά διασφαλίζει ότι κάθε πτυχή περιέχει ίση κατανομή των κλάσεων, καθιστώντας την χρήσιμη για ανισοβαρή σύνολα δεδομένων.

Η διασταυρούμενη επικύρωση περιλαμβάνει τα εξής οφέλη:

- Παρέχει μια πιο αξιόπιστη εκτίμηση της απόδοσης του μοντέλου από ό,τι μια απλή διαίρεση εκπαίδευσης-δοκιμής, καθώς χρησιμοποιεί πολλαπλές διαχωρίσεις των δεδομένων.
- Βοηθά στην αποφυγή της υπερπροσαρμογής (overfitting), καθώς το μοντέλο αξιολογείται σε δεδομένα στα οποία δεν έχει εκπαιδευτεί.
- Επιτρέπει την καλύτερη επιλογή μοντέλου, καθώς μπορεί να χρησιμοποιηθεί για τη σύγκριση της απόδοσης διαφορετικών μοντέλων στα ίδια δεδομένα.

Συνοπτικά, η διασταυρούμενη επικύρωση είναι μια πολύτιμη τεχνική για την αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης, παρέχοντας μια πιο αξιόπιστη εκτίμηση της απόδοσης και βοηθώντας στην αποφυγή της υπερπροσαρμογής.

4.5 Προεπεξεργασία δεδομένων (Data Pre-processing)

Η προεπεξεργασία δεδομένων είναι ένα κρίσιμο βήμα της διαδικασίας ανάλυσης δεδομένων που περιλαμβάνει τον καθαρισμό (data cleaning), τον μετασχηματισμό (transforming) και την προετοιμασία των ακατέργαστων δεδομένων (raw data) για ανάλυση. Τα ακατέργαστα δεδομένα είναι συχνά ακατάστατα, ελλιπή και ασυνεπή, γεγονός που τα καθιστά ακατάλληλα για άμεση χρήση σε αλγορίθμους μηχανικής μάθησης ή στατιστικά μοντέλα.

Η προεπεξεργασία δεδομένων περιλαμβάνει μια σειρά βημάτων για τον καθαρισμό και τον μετασχηματισμό των δεδομένων σε μια πιο εύχρηστη μορφή, συμπεριλαμβανομένης της αφαίρεσης των ελλειπόντων ή ασυνεπών δεδομένων, του μετασχηματισμού των μεταβλητών και της κλιμάκωσης (data scaling) των δεδομένων για να διασφαλιστεί ότι βρίσκονται στην ίδια κλίμακα.

Με την προεπεξεργασία δεδομένων, μπορούμε να βελτιώσουμε την ποιότητα των δεδομένων, καθιστώντας τα πιο ακριβή και αξιόπιστα, και να μειώσουμε την πιθανότητα σφαλμάτων στην ανάλυση που ακολουθεί. Αυτό, με τη σειρά του, μπορεί να οδηγήσει σε καλύτερες γνώσεις και αποτελέσματα από την ανάλυσή μας, καθιστώντας την προεπεξεργασία δεδομένων ένα κρίσιμο βήμα σε κάθε έργο ανάλυσης δεδομένων.

4.5.1 Κωδικοποίηση κατηγορικών μεταβλητών (Label Encoding)

Η κωδικοποίηση ετικέτας είναι μια τεχνική που χρησιμοποιείται στην προεπεξεργασία δεδομένων για τη μετατροπή κατηγορικών μεταβλητών σε αριθμητικές ετικέτες. Στους αλγορίθμους μηχανικής μάθησης και στα στατιστικά μοντέλα, είναι συχνά απαραίτητο να αναπαρασταθούν τα δεδομένα σε αριθμητική μορφή, και η κωδικοποίηση ετικέτας παρέχει έναν τρόπο για να γίνει αυτό για τις κατηγορικές μεταβλητές.

Στην κωδικοποίηση ετικέτας, σε κάθε μοναδική κατηγορία σε μια κατηγορική μεταβλητή αποδίδεται μια μοναδική ακέραια τιμή, ξεκινώντας από το 0 έως τον αριθμό των μοναδικών κατηγοριών μείον ένα. Για παράδειγμα, στην παρούσα εργασία, η ένδειξη για το ότι ο ασθενής είναι κατάλληλος για εγχείρηση λαμβάνει την τιμή 1, ενώ στην περίπτωση που δεν είναι κατάλληλος, λαμβάνει την τιμή 0.

Συνολικά, η κωδικοποίηση ετικέτας είναι μια χρήσιμη τεχνική για τη μετατροπή κατηγορικών μεταβλητών σε αριθμητικές ετικέτες και μπορεί να αποτελέσει ένα πολύτιμο βήμα της προεπεξεργασίας δεδομένων.

4.5.2 Διαχείριση ελλειπουσών τιμών (Handling Missing Values)

Ο χειρισμός των ελλειπουσών τιμών είναι ένα σημαντικό βήμα στην προεπεξεργασία των δεδομένων, καθώς τα ελλιπή δεδομένα μπορούν να επηρεάσουν την ακρίβεια και την αξιοπιστία της ανάλυσής μας. Υπάρχουν διάφοροι τρόποι χειρισμού των ελλιπών τιμών, όπως:

Αφαίρεση: Μια επιλογή είναι η αφαίρεση των γραμμών ή των στηλών που περιέχουν ελλείπουσες τιμές από το σύνολο δεδομένων. Ωστόσο, αυτό μπορεί να οδηγήσει σε απώλεια δεδομένων και μπορεί να μην είναι πρακτικό για σύνολα δεδομένων με πολλές ελλείπουσες τιμές.

Υπολογισμός: Μια άλλη προσέγγιση είναι ο καταλογισμός των ελλειπουσών τιμών με μια τιμή που βασίζεται στα διαθέσιμα δεδομένα. Υπάρχουν διάφορες μέθοδοι για τον καταλογισμό ελλειπουσών τιμών, όπως ο υπολογισμός της μέσης τιμής, ο υπολογισμός της διαμέσου, και άλλες.

Ελλείπουσα τιμή ως ξεχωριστή κατηγορία: Για κατηγορικά χαρακτηριστικά, οι ελλείπουσες τιμές μπορούν να αντιμετωπιστούν ως ξεχωριστή κατηγορία εάν η έλλειψη αυτή φέρει κάποια πληροφορία.

Η επιλογή του τρόπου χειρισμού των ελλιπουσών τιμών εξαρτάται από τα συγκεκριμένα χαρακτηριστικά των δεδομένων και τους στόχους της ανάλυσης. Θα πρέπει να εξετάζεται προσεκτικά ο τρόπος χειρισμού τους, καθώς μπορεί να έχει σημαντικό αντίκτυπο στα αποτελέσματα της ανάλυσης.

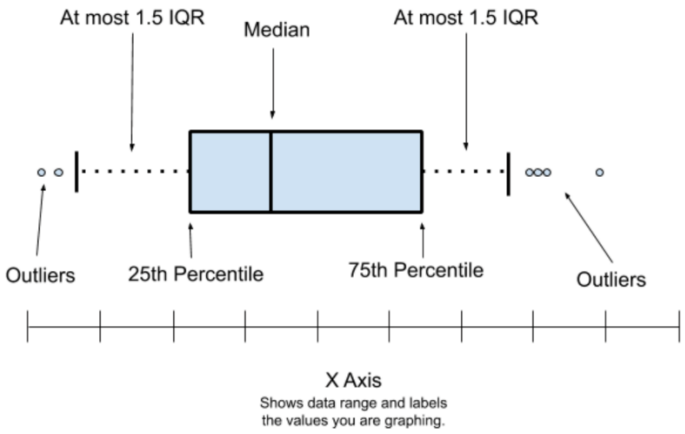
4.5.3 Εντοπισμός ακραίων τιμών (Outlier detection)

Η ανίχνευση ακραίων τιμών είναι η διαδικασία εντοπισμού σημείων δεδομένων ή παρατηρήσεων που διαφέρουν σημαντικά από άλλες παρατηρήσεις σε ένα σύνολο δεδομένων. Γενικά, οι ακραίες τιμές μπορεί να οφείλονται σε σφάλματα κατά τη συλλογή, μέτρηση ή καταχώρηση δεδομένων ή να αντιπροσωπεύουν πραγματικά ασυνήθιστα ή σπάνια γεγονότα.

Υπάρχουν διάφορες μέθοδοι για την ανίχνευση ακραίων τιμών, μεταξύ των οποίων:

Οπτικοποίηση (Visualization): Η απεικόνιση των δεδομένων και η αναζήτηση παρατηρήσεων που απέχουν πολύ από τα υπόλοιπα δεδομένα μπορεί να είναι ένας γρήγορος

τρόπος εντοπισμού πιθανών ακραίων τιμών. Ένα γράφημα που χρησιμοποιείται συχνά για αυτή τη δουλειά, είναι το θηκόγραμμα (boxplot).



Σχήμα 4.21: Διαγραμματική απεικόνιση του θηκογράμματος (boxplot)
(Πηγή: https://publiclab.org/notes/mimiss/06-18-2019/creating-a-boxplot-to-identify-outliers-using-codap)

Στατιστικές μέθοδοι: Στατιστικές μέθοδοι μπορούν να χρησιμοποιηθούν για τον εντοπισμό ακραίων τιμών με βάση τα μέτρα κεντρικής τάσης (π.χ. μέση τιμή, διάμεσος) και μεταβλητότητας (π.χ. τυπική απόκλιση, ενδοτεταρτημοριακό εύρος).

Τεχνικές μηχανικής μάθησης: Αλγόριθμοι μηχανικής μάθησης, όπως η ομαδοποίηση (clustering) ή τα δέντρα αποφάσεων (decision trees), μπορούν να χρησιμοποιηθούν για τον εντοπισμό ακραίων τιμών με βάση την απόκλισή τους.
Μόλις εντοπιστούν οι ακραίες τιμές, υπάρχουν διάφοροι τρόποι για τον χειρισμό τους, όπως:

Αφαίρεση: Οι ακραίες τιμές μπορούν να αφαιρεθούν από το σύνολο δεδομένων, εάν θεωρηθεί ότι είναι λανθασμένες ή ότι δεν αντιπροσωπεύουν τον πληθυσμό που μελετάται.

Μετασχηματισμός: Τα δεδομένα μπορούν να μετασχηματιστούν για να μειωθεί ο αντίκτυπος των ακραίων τιμών, όπως η λήψη του λογαρίθμου ή της τετραγωνικής ρίζας των δεδομένων.

Ανάλυση: Σε ορισμένες περιπτώσεις, οι ακραίες τιμές μπορεί να παρουσιάζουν ιδιαίτερο ενδιαφέρον και σπουδαιότητα, και ο αντίκτυπός τους στην ανάλυση θα πρέπει να εξετάζεται προσεκτικά.

Η ανίχνευση ακραίων τιμών είναι ένα σημαντικό βήμα στην ανάλυση δεδομένων, καθώς μπορεί να έχει σημαντικό αντίκτυπο στα αποτελέσματα των στατιστικών αναλύσεων και των αλγορίθμων μηχανικής μάθησης.

4.5.4 Διάσπαση συνόλου δεδομένων (Dataset split)

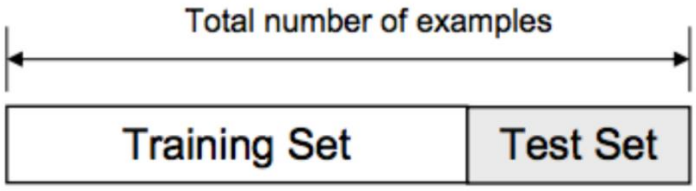
Ο διαχωρισμός συνόλων δεδομένων είναι μια κοινή πρακτική στη μηχανική μάθηση και την ανάλυση δεδομένων, όπου ένα σύνολο δεδομένων χωρίζεται σε δύο ή περισσότερα υποσύνολα για σκοπούς εκπαίδευσης και δοκιμής. Ο πιο συνηθισμένος τρόπος διαχωρισμού των δεδομένων είναι σε δύο υποσύνολα- ένα σύνολο εκπαίδευσης και ένα σύνολο δοκιμής.

Το σύνολο εκπαίδευσης χρησιμοποιείται για τη δημιουργία και την εκπαίδευση του μοντέλου μηχανικής μάθησης, ενώ το σύνολο δοκιμής χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου και την εκτίμηση της ικανότητάς του να γενικεύει σε νέα δεδομένα.

Ο σκοπός του διαχωρισμού των δεδομένων είναι να αποφευχθεί η υπερπροσαρμογή (overfitting). Αξιολογώντας το μοντέλο σε ένα ξεχωριστό σύνολο δοκιμών, μπορούμε να έχουμε μια καλύτερη εκτίμηση της απόδοσης του μοντέλου σε νέα, αθέατα δεδομένα.

Η αναλογία του συνόλου εκπαίδευσης προς το σύνολο δοκιμών μπορεί να ποικίλλει, ανάλογα με το μέγεθος του συνόλου δεδομένων και την πολυπλοκότητα του μοντέλου που κατασκευάζεται. Μια συνηθισμένη κατανομή είναι 80/20 ή 70/30, με το μεγαλύτερο μέρος των δεδομένων να χρησιμοποιείται για την εκπαίδευση.

Είναι σημαντικό να διασφαλιστεί ότι ο διαχωρισμός του συνόλου δεδομένων είναι τυχαίος, δηλαδή ότι τα σημεία δεδομένων ανατίθενται τυχαία είτε στο σύνολο εκπαίδευσης είτε στο σύνολο δοκιμής. Αυτό συμβάλλει στη διασφάλιση ότι το μοντέλο μηχανικής μάθησης που προκύπτει είναι γενικεύσιμο σε νέα δεδομένα και δεν είναι προκατειλημμένο προς κάποιο συγκεκριμένο υποσύνολο των δεδομένων.



Σχήμα 4.22: Διαχωρισμός δεδομένων σε σετ εκπαίδευσης (αριστερά) και σετ ελέγχου (δεξιά)

4.5.5 Μέθοδοι επαναδειγματοληψίας (Resampling methods)

Οι μέθοδοι επαναδειγματοληψίας είναι τεχνικές που χρησιμοποιούνται για την αντιμετώπιση ανισοβαρών συνόλων δεδομένων στη μηχανική μάθηση. Τα ανισόρροπα σύνολα δεδομένων εμφανίζονται όταν η κατανομή των κλάσεων ή των μεταβλητών-στόχων είναι ανισοβαρής, πράγμα που σημαίνει ότι μια κλάση ή κατηγορία είναι πολύ πιο διαδεδομένη από άλλες. Αυτό μπορεί να οδηγήσει σε μεροληπτικά μοντέλα μηχανικής μάθησης που δεν είναι σε θέση να προβλέψουν με ακρίβεια τις υποεκπροσωπούμενες κατηγορίες.

Μια δημοφιλής μέθοδος επαναδειγματοληψίας είναι η Τεχνική Συνθετικής Υπερδειγματοληψίας Μειονότητας (Synthetic Minority Over-sampling Technique - SMOTE), η οποία δημιουργεί νέα συνθετικά παραδείγματα της μειονοτικής κλάσης με παρεμβολή μεταξύ των υφιστάμενων παραδειγμάτων. Αυτή η μέθοδος μπορεί να βοηθήσει στην εξισορρόπηση του συνόλου δεδομένων και να αποτρέψει την υπερβολική προσαρμογή στην πλειοψηφική κλάση. Αυτή είναι και η μέθοδος που θα χρησιμοποιηθεί στην παρούσα εργασία.

Άλλες μέθοδοι επαναδειγματοληψίας περιλαμβάνουν την υποδειγματοληψία, όπου η πλειοψηφική κλάση υποδειγματοληπτείται τυχαία για να εξισορροπηθεί το σύνολο δεδομένων, και την υπερδειγματοληψία, όπου δημιουργούνται νέα παραδείγματα της μειοψηφικής κλάσης για να εξισορροπηθεί το σύνολο δεδομένων.

Μια άλλη προσέγγιση είναι η χρήση μεθόδων συνόλου, όπως τα τυχαία δάση ή η ενίσχυση, οι οποίες συνδυάζουν πολλαπλά μοντέλα για τη βελτίωση της απόδοσης σε ανισόρροπα σύνολα δεδομένων.

Η επιλογή της μεθόδου επαναδειγματοληψίας εξαρτάται από τα συγκεκριμένα χαρακτηριστικά των δεδομένων και τους στόχους της ανάλυσης. Η επαναδειγματοληψία μπορεί να συμβάλει στη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης σε ανισόρροπα σύνολα δεδομένων και θα πρέπει να εξετάζεται προσεκτικά κατά τη δημιουργία μοντέλων σε τέτοια σύνολα δεδομένων.

4.5.6 Κλιμάκωση και κανονικοποίηση δεδομένων (Data scaling and normalization)

Η κλιμάκωση δεδομένων είναι μια κοινή τεχνική προεπεξεργασίας δεδομένων που χρησιμοποιείται στην ανάλυση δεδομένων και τη μηχανική μάθηση για τη μετατροπή των δεδομένων σε μια κοινή κλίμακα. Αυτό γίνεται για να διασφαλιστεί ότι ο αντίκτυπος των μεταβλητών με διαφορετικές μονάδες ή εύρη δεν μεροληπτεί προς μια συγκεκριμένη μεταβλητή ή χαρακτηριστικό.

Οι τεχνικές κανονικοποίησης συνήθως περιλαμβάνουν την αλλαγή της κλίμακας των δεδομένων έτσι ώστε να εμπίπτουν σε ένα καθορισμένο εύρος, όπως 0 έως 1 ή -1 έως 1. Οι συνήθεις τεχνικές κανονικοποίησης περιλαμβάνουν:

Κλιμάκωση min-max: Αυτή η μέθοδος κλιμακώνει τα δεδομένα έτσι ώστε η ελάχιστη τιμή να μετασχηματίζεται σε 0, η μέγιστη τιμή σε 1 και όλες οι άλλες τιμές να αναπροσαρμόζονται αναλογικά.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Κανονικοποίηση Z-score: Αυτή η μέθοδος μετασχηματίζει τα δεδομένα έτσι ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Αυτό μπορεί να βοηθήσει στην τυποποίηση των δεδομένων και να διασφαλίσει ότι μεταβλητές με διαφορετικές μονάδες ή εύρη αντιμετωπίζονται ισότιμα.

$$x_{norm} = \frac{x - mean(x)}{std(x)}$$

Δεκαδική κλιμάκωση: Αυτή η μέθοδος περιλαμβάνει τη μετατόπιση του δεκαδικού σημείου των τιμών των δεδομένων, έτσι ώστε η μέγιστη απόλυτη τιμή να είναι μεταξύ 1 και 10. Αυτό μπορεί να βοηθήσει στη μείωση της επίδρασης των ακραίων τιμών και στη βελτίωση της ακρίβειας της ανάλυσης.

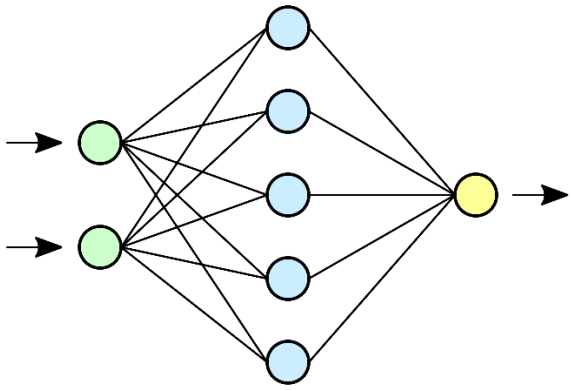
Η κλιμάκωση δεδομένων είναι ένα σημαντικό βήμα προεπεξεργασίας στην ανάλυση δεδομένων και τη μηχανική μάθηση και μπορεί να έχει σημαντικό αντίκτυπο στην ακρίβεια και την αποτελεσματικότητα των μοντέλων και των αλγορίθμων. Είναι σημαντικό να εξετάζεται προσεκτικά η κατάλληλη τεχνική κλιμάκωσης για το συγκεκριμένο σύνολο δεδομένων και την ανάλυση που εκτελείται.

ΚΕΦΑΛΑΙΟ 5

Νευρωνικά δίκτυα

5.1 Εισαγωγή

Τα νευρωνικά δίκτυα είναι ένα είδος μοντέλου τεχνητής νοημοσύνης που είναι εμπνευσμένο από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Έχουν σχεδιαστεί για να επεξεργάζονται μεγάλες ποσότητες δεδομένων και να εξάγουν από αυτά μοτίβα και σχέσεις με νόημα, παρόμοια με τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος επεξεργάζεται τα αισθητηριακά δεδομένα. Τα νευρωνικά δίκτυα αποτελούνται από στρώματα (layers) διασυνδεδεμένων κόμβων, γνωστά και ως νευρώνες, που συνεργάζονται για να αναλύουν και να επεξεργάζονται πληροφορίες. Κάθε νευρώνας δέχεται είσοδο από άλλους νευρώνες, εφαρμόζει μια μαθηματική συνάρτηση σε αυτή την είσοδο και παράγει μια έξοδο. Συνδέοντας πολλούς νευρώνες μεταξύ τους, τα νευρωνικά δίκτυα μπορούν να εκτελούν πολύπλοκους υπολογισμούς και να κάνουν προβλέψεις με βάση τα δεδομένα εισόδου. Τα νευρωνικά δίκτυα έχουν χρησιμοποιηθεί με επιτυχία σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της αναγνώρισης εικόνων, της αναγνώρισης ομιλίας, της επεξεργασίας φυσικής γλώσσας και της αυτόνομης οδήγησης. Με τη συνεχή έρευνα και ανάπτυξη, τα νευρωνικά δίκτυα αναμένεται να διαδραματίσουν ολοένα και σημαντικότερο ρόλο στον τομέα της τεχνητής νοημοσύνης και να συμβάλουν στην πρόοδο σε πολλούς διαφορετικούς τομείς.



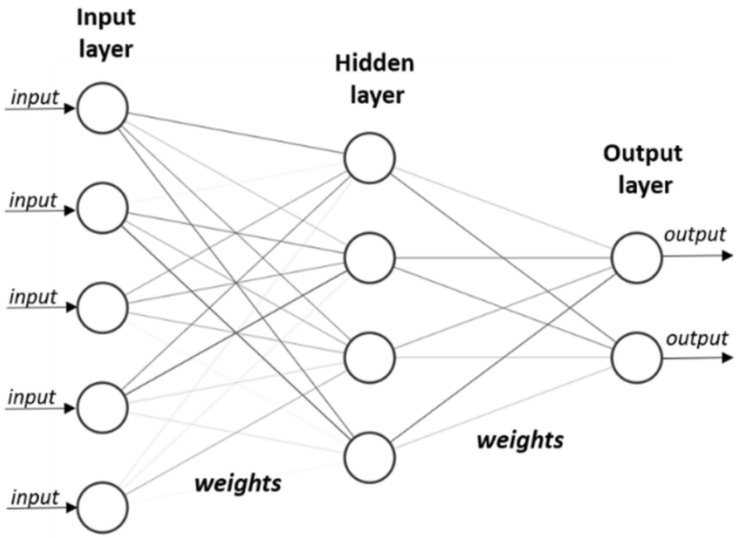
Σχήμα 5.1: Διαγραμματική απεικόνιση της λειτουργίας των νευρωνικών δικτύων

5.2 Τύποι Νευρωνικών Δικτύων

Υπάρχουν πολλοί διαφορετικοί τύποι νευρωνικών δικτύων, ο καθένας με τη δική του μοναδική αρχιτεκτονική και σύνολο εφαρμογών. Μερικοί από τους πιο συνηθισμένους τύπους νευρωνικών δικτύων περιλαμβάνουν:

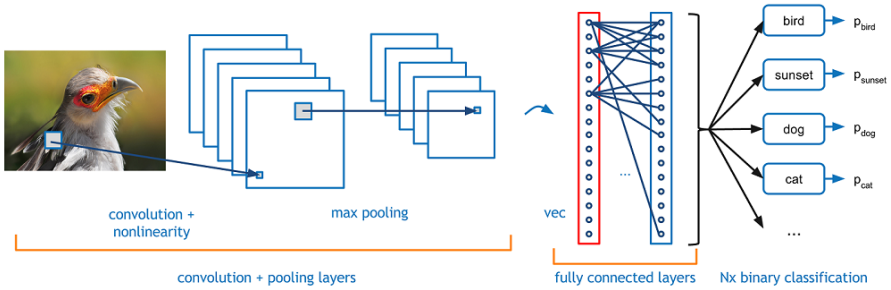
Νευρωνικά Δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks): Αυτά είναι ο απλούστερος τύπος νευρωνικού δικτύου και αποτελούνται από στρώματα διασυνδεδεμένων νευρώνων. Τα δεδομένα ρέουν προς μία κατεύθυνση μέσω του δικτύου, από το στρώμα εισόδου στο στρώμα εξόδου. Τα νευρωνικά δίκτυα τροφοδότησης χρησιμοποιούνται συνήθως

για εργασίες όπως η ταξινόμηση και η παλινδρόμηση. Σε αυτή την εργασία θα χρησιμοποιηθούν για την ταξινόμηση τον ασθενών ανάλογα με το εάν είναι κατάλληλοι για εγχείρηση ή όχι.



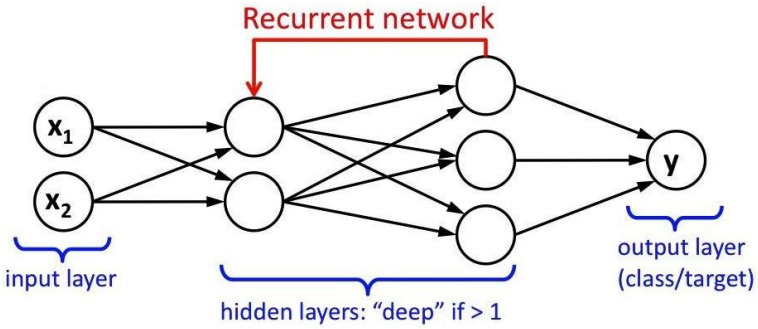
Σχήμα 5.2: Διαγραμματική απεικόνιση της λειτουργίας ενός Feedforward NN

Συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN): Πρόκειται για εξειδικευμένα νευρωνικά δίκτυα που έχουν σχεδιαστεί για την επεξεργασία δεδομένων εικόνας και βίντεο. Αποτελούνται από πολλαπλά στρώματα συνελικτικού τύπου που εκτελούν λειτουργίες όπως το φιλτράρισμα και η συγκέντρωση για την εξαγωγή χαρακτηριστικών από τα δεδομένα εισόδου.



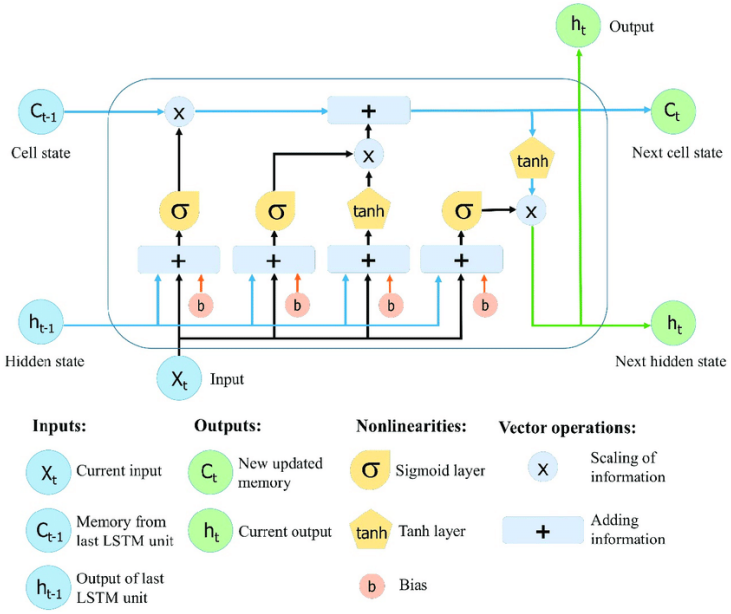
Σχήμα 5.3: Διαγραμματική απεικόνιση της λειτουργίας ενός CNN
(Πηγή: https://towardsdatascience.com/deep-learning-2-f81ebe632d5c)

Επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks - RNN): Πρόκειται για νευρωνικά δίκτυα που έχουν σχεδιαστεί για να επεξεργάζονται ακολουθίες δεδομένων, όπως δεδομένα κειμένου ή χρονοσειρές. Τα RNNs χρησιμοποιούν βρόχους στην αρχιτεκτονική τους για να επιτρέψουν τη διατήρηση των πληροφοριών με την πάροδο του χρόνου και να εκτελέσουν εργασίες όπως η μοντελοποίηση γλώσσας και η αναγνώριση ομιλίας.



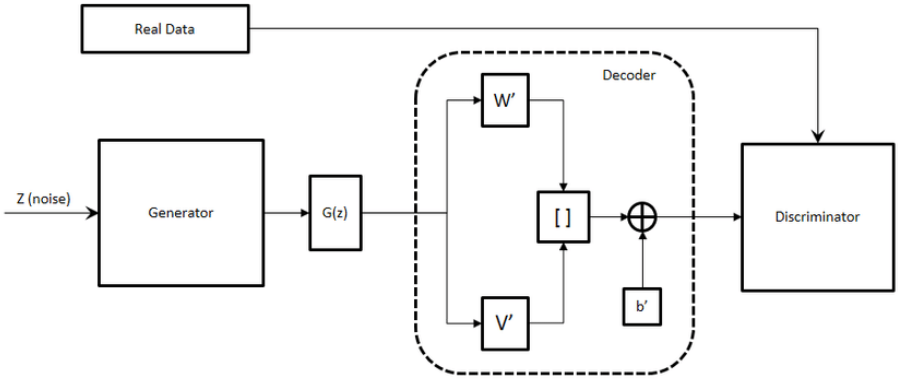
Σχήμα 5.4: Διαγραμματική απεικόνιση της λειτουργίας ενός RNN
(Πηγή: https://towardsdatascience.com/implementation-of-rnn-lstm-and-gru-a4250bf6c090)

Δίκτυα μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory Networks - LSTM): Πρόκειται για έναν τύπο αναδρομικού νευρωνικού δικτύου που έχει σχεδιαστεί για να ξεπεράσει το πρόβλημα της εξαφάνισης των διαβαθμίσεων στα παραδοσιακά RNN. Τα δίκτυα LSTM χρησιμοποιούν εξειδικευμένα κύτταρα μνήμης και πύλες για την επιλεκτική διατήρηση ή απόρριψη πληροφοριών, επιτρέποντάς τους να μοντελοποιούν αποτελεσματικά τις μακροπρόθεσμες εξαρτήσεις σε διαδοχικά δεδομένα.



Σχήμα 5.5: Διαγραμματική απεικόνιση της λειτουργίας ενός LSTM
(Πηγή: https://www.researchgate.net/figure/The-structure-of-the-Long-Short-Term-Memory-LSTM-neural-network-Reproduced-from-Yan_fig8_334268507)

Γενετικά αντιφατικά δίκτυα (Generative Adversarial Networks - GAN): Πρόκειται για νευρωνικά δίκτυα που έχουν σχεδιαστεί για να παράγουν νέα δεδομένα που είναι παρόμοια με ένα σύνολο δεδομένων εκπαίδευσης. Τα GAN αποτελούνται από ένα δίκτυο γεννήτριας που παράγει νέα δείγματα δεδομένων και ένα δίκτυο διάκρισης που διακρίνει μεταξύ πραγματικών και παραγόμενων δεδομένων. Τα GAN έχουν χρησιμοποιηθεί για εργασίες όπως η σύνθεση εικόνων και η παραγωγή φυσικής γλώσσας.



Σχήμα 5.6: Διαγραμματική απεικόνιση της λειτουργίας ενός GAN

Αυτά είναι μερικά μόνο παραδείγματα από τους πολλούς διαφορετικούς τύπους νευρωνικών δικτύων που υπάρχουν. Οι ερευνητές συνεχίζουν να αναπτύσσουν νέες αρχιτεκτονικές και τεχνικές για τα νευρωνικά δίκτυα και είναι πιθανό να παραμείνουν ένα σημαντικό εργαλείο για πολλούς διαφορετικούς τύπους εφαρμογών τεχνητής νοημοσύνης.

5.2.1 Νευρωνικά Δίκτυα πρόσθιας τροφοδότησης (Feedforward Neural Networks)

Τα Νευρωνικά Δίκτυα Πρόσθιας Τροφοδοσίας, γνωστά και ως πολυεπίπεδα perceptrons (MLPs), είναι ένας τύπος νευρωνικού δικτύου όπου τα δεδομένα ρέουν προς μία κατεύθυνση, από το επίπεδο εισόδου στο επίπεδο εξόδου, μέσω ενός ή περισσότερων κρυφών επιπέδων νευρώνων. Η έξοδος κάθε νευρώνα σε ένα στρώμα καθορίζεται από το σταθμισμένο άθροισμα των εισόδων από το προηγούμενο στρώμα, το οποίο περνά από μια συνάρτηση ενεργοποίησης. Τα βάρη και οι μεροληψίες (bias) των νευρώνων μαθαίνονται κατά τη διάρκεια της εκπαίδευσης με τη χρήση ενός αλγορίθμου.

Ο μαθηματικός τύπος για ένα νευρωνικό δίκτυο τροφοδότησης μπορεί να εκφραστεί ως εξής:

$$y = f(W_2 \cdot f(W_1 x + b_1) + b_2)$$

όπου x είναι το διάνυσμα εισόδου, y είναι το διάνυσμα εξόδου, W_1 και W_2 είναι οι πίνακες βαρών για το πρώτο και το δεύτερο επίπεδο, b_1 και b_2 είναι τα διανύσματα πόλωσης (bias) για το πρώτο και το δεύτερο επίπεδο και f() είναι η συνάρτηση ενεργοποίησης.

Η έξοδος του πρώτου στρώματος μπορεί να εκφραστεί ως εξής:

$$h_1 = f(W_1 x + b_1)$$

και η έξοδος του δεύτερου στρώματος μπορεί να εκφραστεί ως εξής:

$$h_2 = f(W_2 h_1 + b_2)$$

όπου h1 είναι η έξοδος του πρώτου στρώματος.

Η συνάρτηση ενεργοποίησης f() είναι συνήθως μια μη γραμμική συνάρτηση που εισάγει μη γραμμικότητα στο μοντέλο. Ορισμένες ευρέως χρησιμοποιούμενες συναρτήσεις ενεργοποίησης περιλαμβάνουν τη σιγμοειδή συνάρτηση, τη συνάρτηση υπερβολικής

εφαπτομένης και τη συνάρτηση διορθωμένης γραμμικής μονάδας (ReLU). Οι συναρτήσεις αυτές, παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 5.1).

Όνομα	Συνάρτηση
Σιγμοειδής συνάρτηση (Sigmoid function)	$f(z) = \frac{1}{1 + \exp(-z)}$
Συνάρτηση υπερβολικής εφαπτομένης	$f(z) = tahn(z)$
Συνάρτηση διορθωμένης γραμμικής μονάδας (ReLU)	$f(z) = \max(0, z)$

Πίνακας 5.1: Είδη συναρτήσεων ενεργοποίησης

Συνοπτικά, τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης είναι ένα ισχυρό εργαλείο για την επεξεργασία πολύπλοκων δεδομένων και την πραγματοποίηση προβλέψεων με βάση αυτά τα δεδομένα. Έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της αναγνώρισης εικόνας και ομιλίας, της επεξεργασίας φυσικής γλώσσας, της χρηματοοικονομικής και των επιστημών υγείας. Με την προσαρμογή του αριθμού των επιπέδων, του αριθμού των νευρώνων ανά επίπεδο και της συνάρτησης ενεργοποίησης, τα νευρωνικά δίκτυα τροφοδότησης μπορούν να προσαρμοστούν ώστε να ανταποκρίνονται σε μια ευρεία ποικιλία εφαρμογών.

5.2.2 Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks)

Τα συνεπτυγμένα νευρωνικά δίκτυα (CNN) είναι ένας τύπος νευρωνικού δικτύου που είναι ιδιαίτερα αποτελεσματικό στην επεξεργασία και ανάλυση οπτικών δεδομένων, όπως εικόνες και βίντεο. Αποτελούνται από πολλαπλά στρώματα, καθένα από τα οποία εκτελεί μια διαφορετική λειτουργία στα δεδομένα εισόδου. Η βασική καινοτομία στα CNN είναι η χρήση των επιπέδων συνελίξεων, τα οποία εφαρμόζουν ένα σύνολο φίλτρων στα δεδομένα εισόδου για την εξαγωγή τοπικών χαρακτηριστικών. Αυτά τα χαρακτηριστικά συνδυάζονται στη συνέχεια σε ανώτερα στρώματα για να σχηματίσουν πιο αφηρημένες αναπαραστάσεις των δεδομένων εισόδου.

Η δομή ενός τυπικού CNN αποτελείται από τρεις τύπους στρωμάτων: στρώματα συμβολής (convolutional layers), στρώματα συγκέντρωσης (pooling layers) και πλήρως συνδεδεμένα στρώματα (fully connected layers). Τα στρώματα συνελικτικού τύπου είναι υπεύθυνα για την εξαγωγή τοπικών χαρακτηριστικών από τα δεδομένα εισόδου, ενώ τα στρώματα συγκέντρωσης χρησιμοποιούνται για την υποδειγματοληψία των χαρτών χαρακτηριστικών (feature maps) και τη μείωση της διαστατικότητας των δεδομένων. Τα πλήρως συνδεδεμένα στρώματα χρησιμοποιούνται για την παραγωγή της τελικής εξόδου του δικτύου.

Το στρώμα συγκέντρωσης χρησιμοποιείται για την υποδειγματοληψία των χαρτών χαρακτηριστικών που παράγονται από το συνελικτικό στρώμα. Ο πιο συνηθισμένος τύπος συγκέντρωσης είναι η max pooling, η οποία επιλέγει τη μέγιστη τιμή από μια τοπική περιοχή του χάρτη χαρακτηριστικών.

Τα πλήρως συνδεδεμένα στρώματα χρησιμοποιούνται για την παραγωγή της τελικής εξόδου του δικτύου. Είναι παρόμοια με τα στρώματα σε ένα νευρωνικό δίκτυο τροφοδότησης και εφαρμόζουν ένα σύνολο βαρών στα δεδομένα εισόδου για να παράγουν την έξοδο.

Συνολικά, τα CNN είναι ένα ισχυρό εργαλείο για την επεξεργασία οπτικών δεδομένων και έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της αναγνώρισης εικόνων και βίντεο, της ανίχνευσης αντικειμένων και της τμηματοποίησης

(segmentation). Ρυθμίζοντας την αρχιτεκτονική του δικτύου, τον αριθμό και το μέγεθος των φίλτρων και τη συνάρτηση ενεργοποίησης, τα CNN μπορούν να προσαρμοστούν ώστε να ταιριάζουν σε μια ευρεία ποικιλία εφαρμογών.

5.2.3 Επαναλαμβανόμενα νευρωνικά δίκτυα (Recurrent Neural Networks)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) είναι ένας τύπος νευρωνικού δικτύου που είναι ιδιαίτερα αποτελεσματικός στην επεξεργασία διαδοχικών δεδομένων, όπως δεδομένα κειμένου και χρονοσειρών. Έχουν σχεδιαστεί για να χειρίζονται εισόδους διαφορετικού μήκους και μπορούν να χρησιμοποιηθούν για τη δημιουργία προβλέψεων και ταξινομήσεων με βάση ολόκληρη την ακολουθία εισόδου και όχι μόνο μεμονωμένα σημεία δεδομένων.

Η βασική καινοτομία στα RNN είναι η χρήση επαναλαμβανόμενων συνδέσεων, οι οποίες επιτρέπουν τη μετάδοση πληροφοριών από το ένα χρονικό βήμα στο επόμενο. Κάθε νευρώνας σε ένα RNN διατηρεί μια κρυφή κατάσταση η οποία ενημερώνεται σε κάθε χρονικό βήμα με βάση την είσοδο σε αυτό το χρονικό βήμα και την προηγούμενη κρυφή κατάσταση. Αυτό επιτρέπει στο δίκτυο να διατηρεί μια μνήμη των προηγούμενων εισόδων και να χρησιμοποιεί αυτές τις πληροφορίες για να κάνει προβλέψεις σχετικά με τις μελλοντικές εισόδους.

Η δομή ενός RNN αποτελείται από τρία είδη στρωμάτων: στρώματα εισόδου, κρυφά στρώματα και στρώματα εξόδου. Το στρώμα εισόδου είναι υπεύθυνο για την επεξεργασία της εισόδου σε κάθε χρονικό βήμα, ενώ το κρυφό στρώμα διατηρεί την κατάσταση του δικτύου σε όλα τα χρονικά βήματα. Το στρώμα εξόδου χρησιμοποιείται για την παραγωγή της τελικής εξόδου του δικτύου.

Ο μαθηματικός τύπος για ένα RNN μπορεί να εκφραστεί ως εξής:

$$h(t) = f(W \cdot x(t) + U \cdot h(t-1) + b)$$

όπου h(t) είναι η κρυφή κατάσταση στο χρονικό βήμα t, x(t) είναι η είσοδος στο χρονικό βήμα t, W και U είναι οι πίνακες βαρών για τις συνδέσεις εισόδου και τις αναδρομικές συνδέσεις, αντίστοιχα, και b είναι το διάνυσμα μεροληψίας. Η συνάρτηση f() είναι η συνάρτηση ενεργοποίησης, η οποία είναι συνήθως μια μη γραμμική συνάρτηση όπως η σιγμοειδής συνάρτηση ή η συνάρτηση υπερβολικής εφαπτομένης.

Η έξοδος του RNN μπορεί να υπολογιστεί εφαρμόζοντας ένα σύνολο βαρών στην κρυφή κατάσταση σε κάθε χρονικό βήμα και συνδυάζοντας τα αποτελέσματα χρησιμοποιώντας μια συνάρτηση softmax. Η συνάρτηση softmax χρησιμοποιείται για την παραγωγή μιας κατανομής πιθανότητας πάνω στις πιθανές κλάσεις εξόδου.

Εν κατακλείδι, τα RNN είναι ένα ισχυρό εργαλείο για την επεξεργασία διαδοχικών δεδομένων και έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένης της επεξεργασίας φυσικής γλώσσας, της αναγνώρισης ομιλίας και της πρόβλεψης χρονοσειρών. Με την προσαρμογή της αρχιτεκτονικής του δικτύου, του αριθμού των κρυφών μονάδων και της συνάρτησης ενεργοποίησης, τα RNN μπορούν να προσαρμοστούν ώστε να ανταποκρίνονται σε ένα ευρύ φάσμα εφαρμογών. Ωστόσο, τα RNN είναι επιρρεπή στο πρόβλημα της εξαφανιζόμενης κλίσης (vanishing gradient), το οποίο μπορεί να καταστήσει δύσκολη την εκπαίδευση βαθιών δικτύων με μεγάλες ακολουθίες εισόδων. Αυτό οδήγησε στην ανάπτυξη άλλων τύπων αναδρομικών δικτύων, όπως το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM) και το δίκτυο Gated Recurrent Unit (GRU), τα οποία έχουν σχεδιαστεί για την αντιμετώπιση αυτού του προβλήματος.

5.2.4 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory Networks)

Τα νευρωνικά δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) είναι ένας τύπος επαναλαμβανόμενου νευρωνικού δικτύου (RNN) που έχει σχεδιαστεί για να αντιμετωπίσει το πρόβλημα της εξαφάνισης των κλίσεων στα παραδοσιακά RNN. Τα LSTM εισήχθησαν το 1997 από τους Hochreiter και Schmidhuber και έχουν γίνει δημοφιλής επιλογή για την επεξεργασία διαδοχικών δεδομένων σε διάφορες εφαρμογές, όπως η επεξεργασία φυσικής γλώσσας, η αναγνώριση ομιλίας και οι λεζάντες εικόνων.

Η βασική καινοτομία στα LSTM είναι η χρήση κυττάρων με πύλες (gated cells) που επιτρέπουν στο δίκτυο να ξεχνά και να θυμάται επιλεκτικά πληροφορίες για μεγάλες χρονικές περιόδους. Κάθε κύτταρο LSTM περιέχει τρεις πύλες: την πύλη εισόδου (input gate), την πύλη λήθης (forget gate) και την πύλη εξόδου (output gate). Η πύλη εισόδου ελέγχει την ποσότητα των νέων πληροφοριών που προστίθενται στην κατάσταση του κελιού, η πύλη λήθης ελέγχει την ποσότητα των παλαιών πληροφοριών που απορρίπτονται από την κατάσταση του κελιού και η πύλη εξόδου ελέγχει την ποσότητα των πληροφοριών που εξάγονται από την κατάσταση του κελιού.

Η πύλη εισόδου υπολογίζεται με την εφαρμογή μιας σιγμοειδούς συνάρτησης στο σταθμισμένο άθροισμα της προηγούμενης κρυφής κατάστασης και της τρέχουσας εισόδου. Η πύλη λήθης υπολογίζεται με τον ίδιο τρόπο, αλλά ελέγχει την ποσότητα της πληροφορίας που διατηρείται στην κατάσταση του κελιού από το προηγούμενο χρονικό βήμα. Η πύλη εξόδου υπολογίζεται με παρόμοιο τρόπο, αλλά ελέγχει την ποσότητα της πληροφορίας που εξάγεται από την κατάσταση του κελιού. Η κατάσταση του κυττάρου ενημερώνεται με την προσθήκη νέων πληροφοριών στην κατάσταση του κυττάρου (ελέγχεται από την πύλη εισόδου) και τη λήθη των παλαιών πληροφοριών (ελέγχεται από την πύλη λήθης).

Συνοπτικά, τα LSTM είναι ένα ισχυρό εργαλείο για την επεξεργασία διαδοχικών δεδομένων και έχουν εφαρμοστεί με επιτυχία σε ένα ευρύ φάσμα εφαρμογών. Ρυθμίζοντας τον αριθμό των κελιών LSTM, τον αριθμό των στρωμάτων και την αρχιτεκτονική του δικτύου, τα LSTM μπορούν να προσαρμοστούν ώστε να ταιριάζουν σε μια μεγάλη ποικιλία εφαρμογών.

ΚΕΦΑΛΑΙΟ 6

Εφαρμογές

6.1 Σκοπός της ανάλυσης

Σκοπός της ανάλυσης των δεδομένων είναι η απόφαση για το πότε οι ασθενείς πρέπει να προβούν σε επέμβαση DBS. Βασιζόμενοι στην αξιοπιστία των δεδομένων του Holter και στα δεδομένα που μας παρέχονται από νευρολόγο με εξειδίκευση στην επεμβατική αντιμετώπιση της ΝΠ, τα οποία αφορούν το ποιοι από τους ασθενείς που έχουμε δεδομένα από Holter είναι κατάλληλοι στο να προχωρήσουν σε DBS (0 = DBS – No, 1 = DBS – Yes), πραγματοποιούμε αλγορίθμους ταξινόμησης, καθώς και νευρωνικού δικτύου ώστε να καταστήσουμε την απόφαση για το πότε ένας ασθενής θα προβεί σε DBS ή όχι, πιο ακριβή και αξιόπιστη μέσα από μοντέλα μηχανικής μάθησης.

6.2 Παρουσίαση δεδομένων Holter

Τα δεδομένα της παρούσας εργασίας έχουν ληφθεί απευθείας από STAT-ON Holter που φορούν οι ασθενείς οι οποίοι πάσχουν από ΝΠ.

Η Sense4Care είναι μια εταιρεία MedTech που αναπτύσσει και παράγει φορητές ιατρικές συσκευές που βασίζονται στην αναγνώριση των προτύπων ανθρώπινης κίνησης, ιδίως στη ΝΠ.

Η λύση STAT-ON™ επιτρέπει στους επαγγελματίες υγείας να αξιολογούν με μεγαλύτερη ακρίβεια τους ασθενείς τους, αναλύοντας την αντικειμενικότητα των συμπτωμάτων και προσφέροντας μια πλήρη διαχείριση της νόσου. Το STAT-ON™ χρησιμοποιείται σε κλινικές δοκιμές και κλινικές πρακτικές σε όλη την Ευρώπη.

Το Holter μπορεί να φορεθεί στον ασθενή όπως φαίνεται στο παρακάτω σχήμα (Σχήμα 6.1), δηλαδή στη μέση του ασθενούς.

Σχήμα 6.1: Απεικόνιση της τοποθέτησης του Holter

Οι αλγόριθμοι και η τεχνολογία STAT-ON™ έχουν επικυρωθεί στο πλαίσιο διαφόρων έργων και έχουν χρηματοδοτηθεί κυρίως από την Ευρωπαϊκή Ένωση τα τελευταία 10 χρόνια. Το αποτέλεσμα αυτής της μακράς πορείας κατέστησε το STAT-ON ™ την πιο αξιόπιστη συσκευή για την παρακολούθηση και την ανίχνευση των κινητικών συμπτωμάτων της ΝΠ με υψηλό επίπεδο χρησιμότητας.

Κάθε φορά που ο αισθητήρας συγχρονίζεται, δημιουργούνται δύο αρχεία στο φάκελο STAT-ON που έχει δημιουργηθεί στο κινητό. Ένα αρχείο PDF που περιέχει τις επεξεργασμένες πληροφορίες των ακατέργαστων (μη επεξεργασμένων) δεδομένων που αποστέλλονται από τον αισθητήρα στο κινητό και ένα αρχείο CSV, το οποίο περιλαμβάνει τα ακατέργαστα δεδομένα (raw data).

Το αρχείο CSV περιέχει τις πληροφορίες όλων των αλγορίθμων ανά λεπτό. Επιπλέον, περιλαμβάνει ορισμένους αθροιστές, φιλτραρισμένες μετρήσεις και ορισμένα δεδομένα που δεν βρίσκονται στο αρχείο PDF.

Τα δεδομένα που περιέχονται στο αρχείο CSV είναι οργανωμένα σε μορφή πίνακα. Οι στήλες αντιστοιχούν στις μεταβλητές εξόδου των διαφόρων αλγορίθμων και κάθε γραμμή αντιστοιχεί στην τιμή αυτών των αλγορίθμων ανά λεπτό. Με άλλα λόγια, κάθε λεπτό παράγεται ένα σύνολο εξόδων αλγορίθμων που σχετίζονται με τα κινητικά συμπτώματα και την κινητικότητα του ασθενούς. Στην παρούσα εργασία θα ασχοληθούμε μόνο με τα ακατέργαστα δεδομένα, δηλαδή τα αρχεία CSV.

Οι μεταβλητές που μας αφορούν, παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 6.1).

Όνομα	**Όνομα μεταβλητής**	**Μονάδα μέτρησης**	**Συχνότητα**	**Περιγραφή**
Stride fluidity (bradykinetic gait)	W_MEAN_FILT	m/s^2	1min	Είναι η συγκέντρωση ορισμένων χαρακτηριστικών συχνότητας των βημάτων που έχει εκτελέσει ο ασθενής στο αντίστοιχο λεπτό. Η μεταβλητή αυτή, προστίθεται κάθε μισή ώρα και δείχνει την ευχέρεια του βηματισμού. Είναι το κλειδί για τον υπολογισμό της μεταβλητής "BRADY10".
Activity level	SMA	m/s^2	1min	Είναι το άθροισμα των επιταχύνσεων σε ένα λεπτό και αναφέρεται στην κίνηση του

				ασθενούς που μετράται από τη μέση και η οποία έχει υψηλή συσχέτιση με την ενέργεια που δαπανάται σε ένα λεπτό. Θα πρέπει να ληφθεί υπόψη ότι οι ακούσιες κινήσεις που προκαλούνται από ορισμένες κινητικές διαταραχές, όπως η δυσκινησία, μπορούν να επηρεάσουν σημαντικά αυτή τη μεταβλητή.
Cadence	CAD	steps/s	1min	Μέσος ρυθμός των βημάτων που ανιχνεύονται σε ένα λεπτό.
Step length	LEN	m	1min	Μέσο μήκος βήματος σε μέτρα κατά τη διάρκεια ενός λεπτού που υπολογίζεται με τη χρήση ενός εκτιμητή που βασίζεται σε μοντέλο ανεστραμμένου εκκρεμούς. Σημειώνεται, ότι αυτός ο εκτιμητής έχει ως παράμετρο εισόδου το μήκος του ποδιού του ασθενούς. Εάν αυτή η παράμετρος δεν έχει εισαχθεί σωστά στη φάση διαμόρφωσης του αισθητήρα, μπορεί να προκαλέσει σημαντικό σφάλμα.

Stride speed	SPEED	m/s	1min	Μέση ταχύτητα βηματισμού, σύμφωνα με τον εκτιμητή που περιγράφεται στη μεταβλητή "LEN", εντός ενός λεπτού.
Number of steps	NUM_STEPS	steps	1min	Αριθμός βημάτων που ανιχνεύονται, σύμφωνα με τον εκτιμητή που περιγράφεται στη μεταβλητή "LEN", εντός ενός λεπτού.
Walking time	TW	s	1min	Ο χρόνος που περπατάει ο ασθενής.
Lying down time	TL	s	1min	Ο χρόνος που είναι ξαπλωμένος ο ασθενής.
Postural Transitions	TRANS	Transitions	1min	Αριθμός των αλλαγών στάσης που έχει πραγματοποιήσει ο ασθενής σε ένα λεπτό.
Falls	FALLS	Falls	1min	Ο αριθμός των πτώσεων του ασθενή σε ένα λεπτό.
FoG occurrence	FOG_EP	FoG detected	1min	Αριθμός επεισοδίων παγώματος βάδισης σε ένα λεπτό.
FoG Windows number	FOG_WIN	windows of FoG detected	1min	Η μέση διάρκεια των επεισοδίων παγώματος βάδισης σε ένα λεπτό.
Pressed button	BTN_PRESSED	Press	1min	Κουμπί ένδειξης συμβάντος. Αυτή η μεταβλητή μπορεί να έχει δύο τιμές: '0' ή '1', όπου '1' αντιστοιχεί στο κουμπί που είναι πατημένο εκείνη τη στιγμή.

Motor state	MOTOR10	-	10min	Κινητική κατάσταση του ασθενούς κάθε δέκα λεπτά. Αυτή η μεταβλητή μπορεί να έχει τέσσερις τιμές: 0 = OFF, 1 = ON, 2 = INT ή 3 = NaN.
Dyskinesia output	DYSK10	-	10min	Αποχώρηση από τον αλγόριθμο δυσκινησίας σε δέκα λεπτά. Η μεταβλητή αυτή μπορεί να έχει 4 τιμές: 0 = NaN, 1 = Dysk ναι, 2 = No Dysk ή 3 = NaN.
Motor state unfiltered	BRADY10	-	10min	Κινητική κατάσταση του ασθενούς κάθε 10 λεπτά χωρίς φίλτρο δυσκινησίας (MOTOR10 είναι με το φίλτρο δυσκινησίας). Αυτή η μεταβλητή μπορεί να έχει τέσσερις τιμές: 0 = OFF, 1 = ON, 2 = INT. ή 3 = NaN. Για τον υπολογισμό αυτής της μεταβλητής χρησιμοποιείται η μεταβλητή "W_MEAN_FILT", η οποία είναι μια προσέγγιση της σοβαρότητας του βραδυκίνητου βαδίσματος μέσω ενός μοντέλου παλινδρόμησης.
Upper flow threshold	TH_HI	m/s^2	-	-

Lower flow threshold	TH_LO	m/s^2	-	-
Age	Age	Years	-	Ηλικία ασθενή
Hoehn & Yahr Scale	H&Y	-	-	Τιμή Hoehn & Yahr που έχει ο ασθενής όταν ο ασθενής είναι σε OFF*.
Leg length	LL	cm	-	Το μήκος του ποδιού του ασθενούς είναι το μήκος από το έδαφος έως τη λαγόνια άκανθα του ασθενούς.
Patients identifier	PAT	-	-	Κωδικός ασθενή

*Ο χρόνος "εκτός λειτουργίας" είναι όταν τα κινητικά ή/και μη κινητικά συμπτώματα του Πάρκινσον εμφανίζονται μεταξύ των δόσεων φαρμάκων.

Πίνακας 6.1: Περιγραφή δεδομένων Holter

Από τα παραπάνω δεδομένα, αποφασίσαμε να αφαιρέσουμε από την ανάλυσή μας το μήκος του ποδιού (LL), καθώς έχει -συνήθως- μια συγκεκριμένη τιμή (100) και δεν μας δίνει χρήσιμη πληροφορία στην ανάλυση που θέλουμε να κάνουμε αφού δεν είναι κάποια ιατρική τιμή που επηρεάζει το εάν κάποιος ασθενής θα προβεί σε DBS.

6.3 Προετοιμασία και προεπεξεργασία δεδομένων

Τα δεδομένα που έχουμε αποτελούνται από 370 CSV αρχεία, κάθε ένα από τα οποία έχει στοιχεία για ξεχωριστό ασθενή. Στην παρούσα εργασία δεν θα ασχοληθούμε με χρονοσειρές, συνεπώς ο στόχος μας ήταν δημιουργήσουμε ένα νέο CSV αρχείο το οποίο θα αποτελείται από 370 γραμμές, η κάθε μία από τις οποίες θα αντιστοιχεί σε έναν μοναδικό ασθενή.

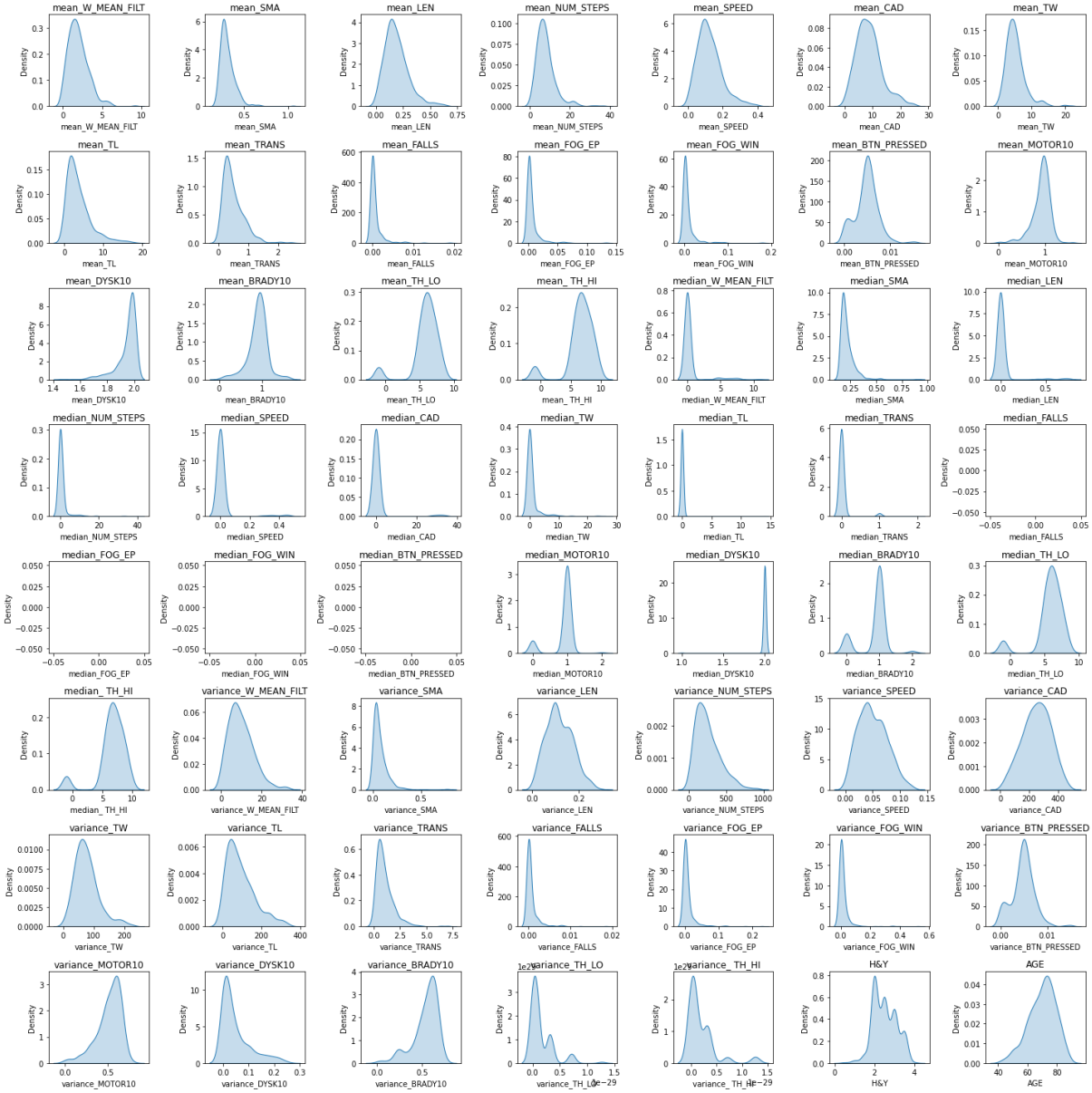
Για να επιτευχθεί το παραπάνω, υπολογίσαμε για κάθε μία μεταβλητή, σε κάθε ένα CSV, τη μέση τιμή, τη διάμεσο και τη διακύμανση ώστε να αντλήσουμε όσο περισσότερη πληροφορία είναι εφικτό, ώστε να μπορέσουμε να κατασκευάσουμε μοναδικές σειρές (rows) για κάθε ασθενή. Προφανώς, η μέση τιμή, η διάμεσος και η διακύμανση, δεν υπολογίζονται για τις μεταβλητές PAT, AGE, H&Y, αφού, όπως είναι λογικό, αυτές οι μεταβλητές δεν αλλάζουν όσο περνούν τα λεπτά ή δεν έχει νόημα να υπολογιστεί κάτι παραπάνω. Αφού πραγματοποιήσουμε αυτούς τους υπολογισμούς, συμπτύσσουμε όλα τα αρχεία CSV σε ένα με αποτέλεσμα να έχουμε 370 γραμμές με 370 ασθενείς και 58 στήλες οι οποίες περιέχουν τις μεταβλητές του παραπάνω πίνακα υπολογισμένες με μέση τιμή, διάμεσο και διακύμανση. Επιπλέον, περιέχεται η μεταβλητή DBS η οποία αποτελείται από τις τιμές 0 και 1, και διαδραματίζει τον ρόλο της μεταβλητής στόχου του προβλήματός μας. Η παραπάνω διαδικασία έγινε με ειδικές συναρτήσεις στην Python, για εξοικονόμηση χρόνου.

Να σημειωθεί ότι υπάρχουν μεταβλητές (MOTOR10, DYSK10, BRADY10) που περιέχονται NaN τιμές ως ακέραιοι. Οι τιμές αυτές έχουν αντικατασταθεί και περαστεί ως ελλείπουσες τιμές και στον υπολογισμό της μέσης τιμής, της διακύμανσης και της διαμέσου,

έχουν παραληφθεί και, κατά συνέπεια, οι υπολογισμοί αυτοί βασίζονται στις μη NaN values, ώστε -κατά προσέγγιση- να έχουμε πιο λογικό και αξιόπιστο αποτέλεσμα.

Όσον αφορά τις ελλείπουσες τιμές που δημιουργούνται στο τελικό dataset, που είναι 25 και δημιουργούνται στις τρεις μεταβλητές που προαναφέραμε, τις αντικαθιστούμε με την τιμή της ίδιας μεταβλητής του ασθενούς που έχει όσο πιο κοινό προφίλ γίνεται με τον ασθενή που έχει τις ελλείπουσες τιμές. Αυτό γίνεται αφαιρώντας όλες τις τιμές της σειράς του ασθενή με τις ελλείπουσες τιμές με τις τιμές των σειρών των άλλων ασθενών (εξαιρώντας προφανώς τη μεταβλητή PAT), και παίρνοντας το απόλυτο άθροισμα. Ο ασθενής με τις ελλείπουσες τιμές παίρνει την ίδια τιμή με τον κοντινότερο σε προφίλ ασθενή.

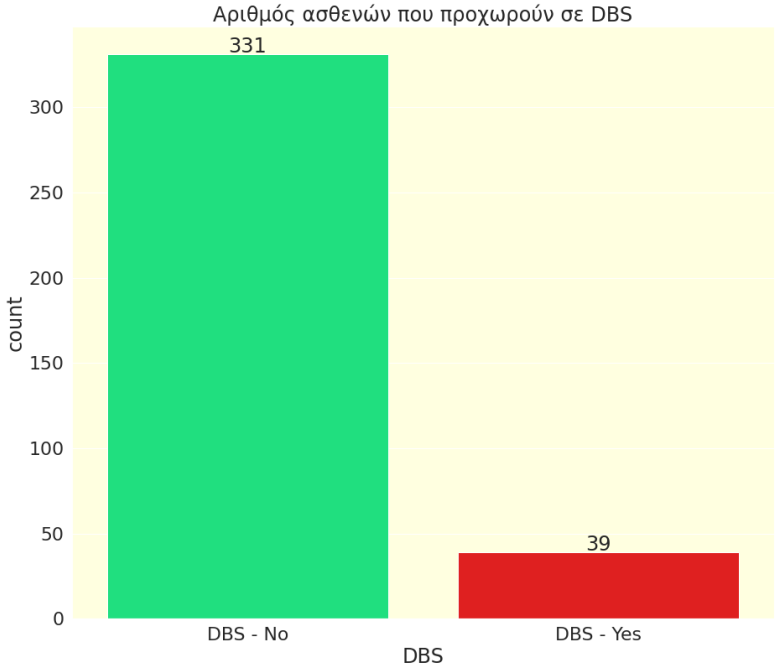
Στο παρακάτω σχήμα (Σχήμα 6.2) παρουσιάζονται οι κατανομές (density plots) όλων των μεταβλητών.



Σχήμα 6.2: Κατανομή των μεταβλητών

Τέλος, αφού τα δεδομένα έχουν πάρει την κατάλληλη μορφή και έχουμε καταλήξει στο τελικό μας dataset, πραγματοποιούμε τυχαία δειγματοληψία με υπερσυνθετική μειονότητα

(SMOTE) γιατί παρατηρούμε ότι έχουμε ανισορροπία στη μεταβλητή στόχο (Σχήμα 6.3), και, εν συνεχεία, την κανονικοποίηση που είναι απαραίτητη για τους αλγορίθμους μηχανικής μάθησης.



Σχήμα 6.3: Γραφική απεικόνιση των ασθενών που είναι κατάλληλοι για DBS

6.4 Εφαρμογή με αλγορίθμους ταξινόμησης

Οι αλγόριθμοι εποπτευόμενης ταξινόμησης που χρησιμοποιήθηκαν, είναι οι εξής:

- Λογιστική Παλινδρόμηση
- Μηχανές Διανυσματικής Υποστήριξης
- Κ Κοντινότεροι Γείτονες
- Γραμμική Διακριτική Ανάλυση
- Τυχαίο Δάσος
- Extreme Gradient Boosting

Τα δεδομένα έχουν χωριστεί σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου (80%-20%, αντίστοιχα) και οι μετρικές που χρησιμοποιούνται για την αξιολόγηση των μοντέλων είναι οι εξής:

- Accuracy
- Precision
- Recall
- F-score
- AUC

Η κλάση ενδιαφέροντος είναι η 1 (δηλαδή πότε οι ασθενείς πηγαίνουν σε εγχείρηση), συνεπώς οι μετρικές υπολογίζονται βάσει αυτής. Να σημειωθεί ότι έχει χρησιμοποιηθεί η τεχνική της διασταυρούμενης επικύρωσης 10-fold (10-fold Cross Validation) και στη συνέχεια υπολογίζεται η μέση τιμή και η τυπική απόκλιση των 10 επαναλήψεών της.

Ταξινόμηση χωρίς PCA
Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 6.2).

Model	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)	AUC(%)
Logistic Regression	85.2 (0.048)	82.3 (0.037)	90.9 (0.058)	86.3 (0.044)	85.1 (0.048)
SVM	86.6 (0.048)	82.4 (0.044)	94.1 (0.055)	87.8 (0.043)	86.3 (0.048)
KNN	83.6 (0.046)	76.3 (0.051)	99.6 (0.011)	86.3 (0.032)	83.1 (0.047)
LDA	83.5 (0.039)	80.4 (0.029)	90.1 (0.075)	84.8 (0.039)	83.4 (0.038)
RF	93.8 (0.021)	92.7 (0.034)	95.6 (0.036)	94 (0.020)	93.7 (0.021)
XGB	95.1 (0.023)	94.7 (0.035)	96 (0.035)	95.3 (0.022)	95.1 (0.023)

Πίνακας 6.2: Μετρικές μοντέλων ταξινόμησης χωρίς PCA

Παρατηρούμε ότι το μοντέλο που αποδίδει καλύτερα, λαμβάνοντας υπόψη όλες τις μετρικές, είναι το Extreme Gradient Boosting.

Η μέση ορθότητα μας λέει ότι, κατά μέσο όρο, το μοντέλο μας ταξινομεί σωστά το 95.1% όλων των παραδειγμάτων (τόσο των θετικών όσο και των αρνητικών).

Η μέση ακρίβεια μας λέει ότι, κατά μέσο όρο, όταν το μοντέλο μας προβλέπει ένα παράδειγμα ως θετικό (κλάση 1), είναι σωστό στο 94.7% των περιπτώσεων. Αυτό σημαίνει ότι από όλα τα παραδείγματα που το μοντέλο μας ταξινόμησε ως θετικά, μόνο το 94.7% από αυτά ανήκουν στην πραγματικότητα στην κλάση 1.

Η μέση ανάκληση μας λέει ότι, κατά μέσο όρο, το μοντέλο μας αναγνωρίζει σωστά το 96% όλων των θετικών παραδειγμάτων (κλάση 1). Αυτό σημαίνει ότι από όλα τα παραδείγματα που πραγματικά ανήκουν στην κλάση 1, το μοντέλο μας αναγνωρίζει σωστά το 96% αυτών.

Το μέσο F1-score είναι ένας αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης και μας δίνει ένα συνολικό μέτρο της απόδοσης του μοντέλου στην ταξινόμηση θετικών παραδειγμάτων. Σε αυτή την περίπτωση, το μέσο F_1-score είναι 95.3%, το οποίο υποδεικνύει ότι το μοντέλο μας έχει αρκετά καλή απόδοση στην ορθή ταξινόμηση θετικών παραδειγμάτων.

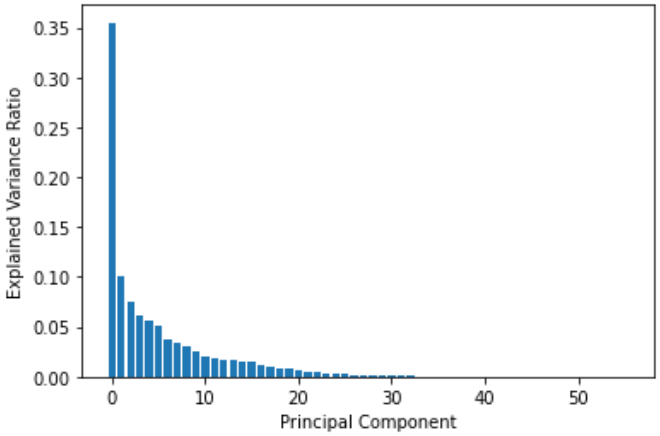
Συνολικά, αυτές οι μετρικές μας δίνουν μια ένδειξη του πόσο καλά το μοντέλο μας αποδίδει στην ορθή αναγνώριση θετικών παραδειγμάτων (κλάση 1). Ένας υψηλός μέσος όρος ανάκλησης σημαίνει ότι το μοντέλο μας αναγνωρίζει σωστά τα περισσότερα θετικά παραδείγματα, ενώ ένας υψηλός μέσος όρος ακρίβειας σημαίνει ότι το μοντέλο μας προβλέπει με ακρίβεια πότε ένα παράδειγμα ανήκει στην κλάση 1. Η μέση βαθμολογία F_1 παρέχει ένα συνολικό μέτρο της απόδοσης του μοντέλου, λαμβάνοντας υπόψη τόσο την ακρίβεια όσο και την ανάκληση.

Ταξινόμηση με PCA

Σε αυτή την περίπτωση, θα κάνουμε ό,τι ακριβώς κάναμε και παραπάνω, μόνο που τώρα, θα επιλέξουμε χαρακτηριστικά τόσα ώστε να εξηγείται το 95% της διακύμανσης, με τη μέθοδο ανάλυσης κυρίων συνιστωσών. Ο λόγος που επιλέγουμε τη μέθοδο ανάλυσης κυρίων συνιστωσών είναι για τη μείωση της διαστασιμότητας, μια και τα δεδομένα μας αποτελούνται από 370 ασθενείς (γραμμές) και 58 χαρακτηριστικά (στήλες), καθώς επίσης και για την

μετατροπή πιθανώς συσχετισμένων μεταβλητών σε ένα σύνολο τιμών γραμμικά μη συσχετισμένων μεταβλητών. Πέραν τούτου, στους αλγορίθμους που χρησιμοποιούμε δεν έχουμε κάνει ελέγχους σχετικά με το εάν πληρούνται οι υποθέσεις που έχει ο κάθε αλγόριθμος, ειδικά για τη γραμμική διακριτική ανάλυση που έχει την υπόθεση της γραμμικής ανεξαρτησίας, συνεπώς η εφαρμογή της ανάλυσης κυρίων συνιστωσών μπορεί να βοηθήσει τα αποτελέσματα της ανάλυσής μας.

Ο κατάλληλος αριθμός χαρακτηριστικών, είναι 19 (Σχήμα 6.4).



Σχήμα 6.4: Κατάλληλος αριθμός χαρακτηριστικών

Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 6.3).

Model	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)	AUC(%)
Logistic Regression	77.1 (0.045)	76.3 (0.049)	81.3 (0.056)	78.5 (0.039)	77 (0.046)
SVM	78.5 (0.039)	74.9 (0.041)	87.9 (0.050)	80.8 (0.033)	78.2 (0.040)
KNN	83.8 (0.055)	76.8 (0.062)	99.3 (0.015)	86.4 (0.039)	83.3 (0.056)
LDA	76.7 (0.045)	75.5 (0.052)	82 (0.046)	78.5 (0.036)	76.6 (0.045)
RF	91.1 (0.024)	90.3 (0.040)	95.4 (0.051)	92.6 (0.024)	92.4 (0.024)
XGB	92.1 (0.037)	88.3 (0.056)	96 (0.034)	91.8 (0.033)	91 (0.038)

Πίνακας 6.3: Μετρικές μοντέλων ταξινόμησης με PCA

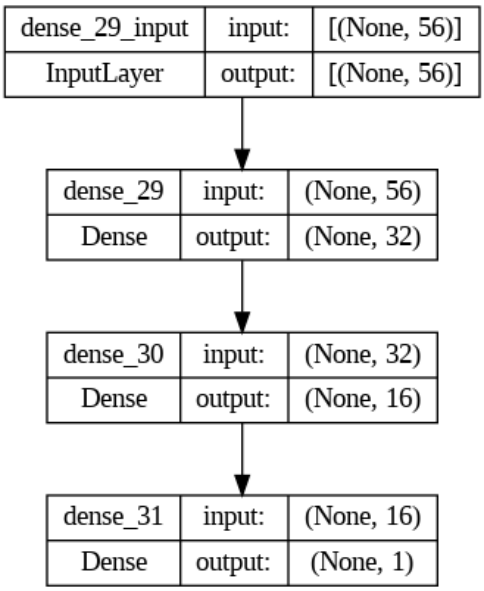
Γενικά, παρατηρούμε ότι τα μοντέλα ταξινόμησης χωρίς PCA στο συγκεκριμένο πρόβλημα ταξινόμησης, είχαν καλύτερη απόδοση αν και δεν παρατηρούνται μεγάλες διαφορές ανάμεσα στις μετρικές και, επιπλέον, τα μοντέλα μας εξακολουθούν και έχουν καλή απόδοση. Επιπλέον, και σε αυτή την περίπτωση, ο αλγόριθμος Extreme Gradient Boosting είναι αυτός με την καλύτερη απόδοση.

6.5 Εφαρμογή με νευρωνικό δίκτυο πρόσθιας τροφοδότησης

Τα δεδομένα έχουν χωριστεί σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου (80%-20%, αντίστοιχα) και το νευρωνικό δίκτυο «τρέχει» δύο φορές, μία με μόνη τεχνική τον διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου, και μία με την τεχνική της διασταυρούμενης επικύρωσης 10-fold. Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση του νευρωνικού δικτύου στο σύνολο δεδομένων ελέγχου (test), είναι οι εξής:

- Accuracy
- Precision
- Recall
- F-score
- AUC-ROC

Η αρχιτεκτονική του νευρωνικού δικτύου, αναπαρίσταται στο παρακάτω σχήμα (Σχήμα 6.5).



Σχήμα 6.5: Διαγραμματική απεικόνιση του MLP NN

Το Σχήμα 6.5 παρουσιάζει την αρχιτεκτονική του νευρωνικού δικτύου αυτού του μοντέλου. Δείχνει το επίπεδο εισόδου, το πρώτο κρυφό επίπεδο με 32 νευρώνες, το δεύτερο κρυφό επίπεδο με 16 νευρώνες και το επίπεδο εξόδου με έναν νευρώνα.

Το στρώμα εισόδου έχει 56 νευρώνες, οι οποίοι αντιστοιχούν στη διαστατικότητα των δεδομένων εισόδου. Το πρώτο κρυφό στρώμα έχει 32 νευρώνες, που σημαίνει ότι θα μάθει 32 διαφορετικές αναπαραστάσεις των δεδομένων εισόδου. Αυτές οι αναπαραστάσεις θα περάσουν στο επόμενο κρυφό επίπεδο, το οποίο έχει 16 νευρώνες. Το δεύτερο κρυφό στρώμα θα μάθει να συνδυάζει τις αναπαραστάσεις από το πρώτο κρυφό στρώμα για να δημιουργήσει πιο σύνθετες αναπαραστάσεις των δεδομένων εισόδου. Τέλος, το στρώμα εξόδου έχει έναν νευρώνα, ο οποίος χρησιμοποιείται για την πρόβλεψη της δυαδικής εξόδου του μοντέλου.

Η συνάρτηση ενεργοποίησης relu χρησιμοποιείται και στα δύο κρυφά στρώματα, η οποία εισάγει μη γραμμικότητα στο μοντέλο και του επιτρέπει να μαθαίνει πιο σύνθετα μοτίβα στα δεδομένα. Η σιγμοειδής συνάρτηση ενεργοποίησης χρησιμοποιείται στο στρώμα εξόδου, η οποία αντιστοιχίζει την έξοδο του δικτύου σε μια τιμή πιθανότητας μεταξύ 0 και 1, που αντιπροσωπεύει την πιθανότητα τα δεδομένα εισόδου να ανήκουν στη θετική κλάση.

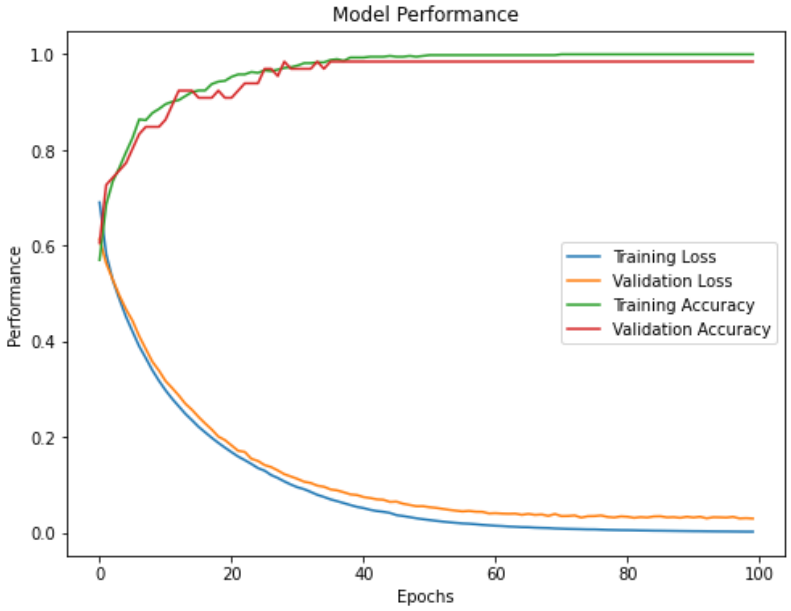
Συνολικά, αυτό το μοντέλο MLP έχει μια αρχιτεκτονική νευρωνικού δικτύου που μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων δυαδικής ταξινόμησης.

Μετρικές χωρίς την τεχνική της διασταυρούμενης επικύρωσης

Στον παρακάτω πίνακα (Πίνακας 6.4) παρουσιάζονται οι μετρικές του νευρωνικού δικτύου, ενώ στο σχήμα (Σχήμα 6.6) η συνολική απόδοσή του.

Accuracy	**98.48%**
Precision	97.05%
Recall	100%
F1-score	98.50%
AUC-ROC	98.48%

Πίνακας 6.4: Μετρικές MLP NN



Σχήμα 6.6: Απόδοση νευρωνικού δικτύου MLP

Μετρικές με την τεχνική της διασταυρούμενης επικύρωσης

Στον παρακάτω πίνακα (Πίνακας 6.5) παρουσιάζονται η ακρίβεια και η απώλεια του νευρωνικού δικτύου, χρησιμοποιώντας την τεχνική της διασταυρούμενης επικύρωσης 10-fold.

Cross-validation accuracy	**99.40%**
Cross-validation loss	0.02%

Πίνακας 6.5: 10-fold MLP NN

ΚΕΦΑΛΑΙΟ 7

Συμπεράσματα

Στην παρούσα εργασία, έγινε μια εκτενής περιγραφή της ΝΠ, των συμπτωμάτων της, των αιτιών, των παραγόντων κινδύνου, των πιθανών θεραπειών και των επιδημιολογικών στοιχείων. Ακόμη, έγινε μια εισαγωγή στην εν τω βάθει εγκεφαλική διέγερση, καθώς στην μετέπειτα ανάλυσή μας, μας αφορούν οι ασθενείς που προχωρούν σε αυτή. Επιπροσθέτως, μελετήθηκε η βιβλιογραφία σε εφαρμογές που έχουν γίνει σε δεδομένα από Holter, και μπορούμε να καταλήξουμε στο συμπέρασμα ότι, πλέον, με την πρόοδο της επιστήμης, με τη βοήθεια του Holter, των ιατρών και των επιστημόνων υγείας, υπάρχουν διαθέσιμα πιο αξιόπιστα δεδομένα, δίνοντας πληρέστερη και ακριβέστερη εικόνα για την κατάσταση των ασθενών, καθιστώντας πιο κατανοητή και αντιμετωπίσιμη τη θεραπεία της νόσου.

Παρουσιάστηκε, επίσης, το θεωρητικό υπόβαθρο των κατηγοριών της μηχανικής μάθησης, των αλγορίθμων της αλλά και των τρόπων αξιολόγησης αυτών. Συνοπτικά αλλά επεξηγηματικά, μελετήθηκαν και οι βασικότεροι τύποι νευρωνικών δικτύων.

Όσον αφορά το κομμάτι της εφαρμογής, χρησιμοποιήθηκαν δεδομένα από Holter ώστε να μπορέσει να γίνει ταξινόμηση, βάσει των στοιχείων που παρέχονται από αυτό, για το ποιοι ασθενείς είναι κατάλληλοι ώστε να προχωρήσουν σε επέμβαση DBS. Παρατηρήσαμε ότι το Holter παρέχει αρκετά δεδομένα, χρησιμοποιώντας κατάλληλους αλγορίθμους υπολογισμού, ώστε να παρέχονται όσο το δυνατόν περισσότερες πληροφορίες για την κατάσταση του ασθενούς κατά τη διάρκεια της ημέρας. Αξίζει να σημειωθεί ότι τα δεδομένα του Holter μπορούν να αξιοποιηθούν και από ανθρώπους που δεν είναι επαγγελματίες δεδομένων (data scientists, data analysts), καθώς, μαζί με τα δεδομένα, παράγεται και μία αναφορά (report) με γραφήματα και τις απαραίτητες τιμές που χρειάζεται να λάβει υπόψη του ένας ιατρός.

Στη μελέτη μας, χρησιμοποιήθηκαν αρκετοί αλγόριθμοι ταξινόμησης με εφαρμογή της ανάλυσης κυρίων συνιστωσών για την επιλογή χαρακτηριστικών αλλά και χωρίς. Παρατηρήθηκε ότι, στο συγκεκριμένο πρόβλημα, οι αλγόριθμοι αποδίδουν καλύτερα χωρίς την εφαρμογή της ανάλυσης κυρίων συνιστωσών αλλά με όχι και τόσο αξιοσημείωτες διαφορές. Επιπροσθέτως, ο Extreme Gradient Boosting αποδείχθηκε, και στις δύο περιπτώσεις, ότι δίνει τις καλύτερες μετρικές για την πρόβλεψη των ασθενών που πρέπει να προβούν σε εγχείρηση.

Τέλος, χρησιμοποιήθηκε νευρωνικό δίκτυο πρόσθιας τροφοδότησης το οποίο αποδείχθηκε επίσης πολύ χρήσιμο στην πρόβλεψη των ασθενών που προχωρούν σε DBS.

Πέραν τούτου, τα δεδομένα του Holter μπορούν να χρησιμοποιηθούν είτε για τη μελέτη των ασθενών μεμονωμένα είτε συγκεντρώνοντας ένα δείγμα ασθενών, όπως έγινε και στη δική μας μελέτη, ώστε όχι μόνο να γίνει κάποια ταξινόμηση ή παλινδρόμηση, αλλά να αναλυθούν και να μελετηθούν μοτίβα.

Αξιοποιώντας τα δεδομένα που παρέχονται από το Holter μπορούν να πραγματοποιηθούν ποικίλες μελέτες, όχι μόνο για την πρόβλεψη της επέμβασης αλλά και για οποιαδήποτε άλλη μεταβλητή χρειαστεί από επιστήμονες δεδομένων σε συνεργασία με τους επιστήμονες υγείας.

ΠΑΡΑΡΤΗΜΑΤΑ

Π1 Πηγαίος κώδικας στην Python

Εισαγωγή βιβλιοθηκών

```
# Import Libraries
import pandas as pd
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split, GridSearchCV
from collections import Counter

from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
import xgboost as xgb
from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import make_scorer, accuracy_score, precision_score, r
ecall_score, f1_score, roc_auc_score
from sklearn.model_selection import cross_validate
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

from keras.models import Sequential
from keras.layers import Dense
from keras.optimizers import Adam
from keras.metrics import binary_accuracy, Precision, Recall
```

```
from keras.utils import plot_model
from sklearn.model_selection import StratifiedKFold
```

Κώδικας για την εισαγωγή δεδομένων και επιλογή του X και y

```
df = pd.read_excel('MERGED_PD_DATA.xlsx')

X = df.drop(["DBS", "PAT"], axis = 1)
y = df["DBS"]
```

Κώδικας για τη διαχείριση ελλειπουσών τιμών

```
# find the indices of rows with missing values
missing_indices = X.index[df.isnull().any(axis=1)]

# create a new dataframe with rows where missing values are not present
X_new = X.dropna()

# iterate through each column and row with missing values
for col in X.columns:
    for index in missing_indices:
        if pd.isna(X.loc[index, col]):
            # compute the sum of absolute differences for each column
            abs_diff = X_new.sub(X.loc[index], axis=1).abs().sum(axis=1)
            # select the row with the smallest sum of absolute differences
            closest_row = abs_diff.idxmin()
            # replace the missing value with the corresponding value from the sele
cted row
            X.loc[index, col] = X_new.loc[closest_row, col]
```

Κώδικας για density plots

```
# Set the figure size
plt.figure(figsize=(20,20))

# Loop through each column and create a density plot
for i, column in enumerate(X.columns):
    plt.subplot(8, 7, i+1)
    sns.kdeplot(X[column], shade=True)
    plt.title(column)
```

```
# Display the plot
plt.tight_layout()
plt.show()
```

Κώδικας για SMOTE και scaling

```
# Balance dataset
sm = SMOTE(random_state = 42)
X, y = sm.fit_resample(X, y)

# Print the class distribution of the resampled dataset
print('Resampled class distribution: ', Counter(y))

# Scale the features
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

Κώδικας για PCA

```
# Perform PCA with all the components
pca_all = PCA().fit(X)
variance_ratio_all = pca_all.explained_variance_ratio_

# Plot the explained variance ratio of each component
plt.bar(range(len(variance_ratio_all)), variance_ratio_all)
plt.xlabel('Principal Component')
plt.ylabel('Explained Variance Ratio')
plt.show()

# Choose the best number of components based on the explained variance rati
o
cumulative_variance_ratio = np.cumsum(variance_ratio_all)
n_components = np.argmax(cumulative_variance_ratio >= 0.95) + 1
print(f"Best number of components: {n_components}")

# Perform PCA with the chosen number of components
pca_best = PCA(n_components=n_components).fit(X)
X = pca_best.transform(X)
```

Κώδικας για αλγορίθμους ταξινόμησης

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)

# Fit the logistic regression model on the training set
model = LogisticRegression(penalty="l2", C=1, solver="lbfgs", max_iter=300)
model.fit(X_train, y_train)

# Fit the LDA model on the training set
model = LinearDiscriminantAnalysis()
model.fit(X_train, y_train)

# Fit the SVM model on the training set
model = SVC(kernel='linear', C=1, random_state=42)
model.fit(X_train, y_train)

# Fit the KNN model on the training set
model = KNeighborsClassifier()
model.fit(X_train, y_train)

# Define XGBoost classifier model
params = {objective='binary:logistic',
n_estimators=300,
learning_rate=0.05,
max_depth=10,
min_child_weight=1,
gamma=0.1,
subsample=0.8,
colsample_bytree=0.8,
scale_pos_weight=1,
seed=42
}
model = xgb.XGBClassifier(**params)
model.fit(X_train, y_train)

# Define Random Forest classifier model
model = RandomForestClassifier(n_estimators=300, max_depth=10, random_state
=42)
model.fit(X_train, y_train)
```

```
# Models performance
# Predict on the test data
y_pred = model.predict(X_test)

# Print the classification report and AUC score
print(classification_report(y_test, y_pred, digits=4))
print('AUC score:', roc_auc_score(y_test, y_pred))

scorers = {'accuracy': make_scorer(accuracy_score),
           'precision': make_scorer(precision_score, pos_label=1),
           'recall': make_scorer(recall_score, pos_label=1),
           'f1': make_scorer(f1_score, pos_label=1),
           'auc': make_scorer(roc_auc_score)}

cv_results = cross_validate(model, X_train, y_train, cv=10, scoring=scorers
)

print('Accuracy: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_accuracy'
]), np.std(cv_results['test_accuracy'])))
print('Precision: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_precisio
n']), np.std(cv_results['test_precision'])))
print('Recall: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_recall']),
np.std(cv_results['test_recall'])))
print('F1 Score: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_f1']), np
.std(cv_results['test_f1'])))
print('AUC: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_auc']), np.std
(cv_results['test_auc'])))
# Assuming y_true and y_pred are the true and predicted labels respectively
cm = confusion_matrix(y_test, y_pred)
print("Confusion matrix:")
print(cm)
print("Total samples:", np.sum(cm))
print("True positives:", cm[1, 1])
print("False positives:", cm[0, 1])
print("True negatives:", cm[0, 0])
print("False negatives:", cm[1, 0])
```

Κώδικας για νευρωνικό δίκτυο

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)


#Χωρίς 10-fold cross-validation
# Define the model
model = Sequential()
model.add(Dense(32, input_dim=56, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
# Compile the model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accur
acy'])
# Train the model
history = model.fit(X_train, y_train, epochs=100, batch_size=32, verbose=1,
 validation_data=(X_test, y_test))
# Evaluate the model on test data
y_pred = model.predict(X_test)
y_pred = (y_pred > 0.5).astype(int)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
auc_roc = roc_auc_score(y_test, y_pred)

print('Test accuracy:', accuracy)
print('Test precision:', precision)
print('Test recall:', recall)
print('Test F1-score:', f1)
print('Test AUC-ROC:', auc_roc)

plot_model(model, show_shapes=True)

# Plot loss and accuracy over epochs
plt.figure(figsize=(8, 6))
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.plot(history.history['accuracy'], label='Training Accuracy')
```

```
plt.plot(history.history['val_accuracy'], label='Validation Accuracy')
plt.title('Model Performance')
plt.xlabel('Epochs')
plt.ylabel('Performance')
plt.legend()
plt.show()

#Με 10-fold cross-validation
# Define 10-fold cross-validation
kfold = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)

# Initialize lists to store cross-validation results
acc_scores = []
loss_scores = []

# Perform 10-fold cross-validation
for train_index, test_index in kfold.split(X, y):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    # Train the model on the current fold
    history = model.fit(X_train, y_train, epochs=100, batch_size=32, verbos
e=1, validation_data=(X_test, y_test))

    # Evaluate the model on the current fold
    loss, acc = model.evaluate(X_test, y_test, verbose=0)
    acc_scores.append(acc)
    loss_scores.append(loss)

# Print the cross-validation results
print("Cross-
validation accuracy: {:.2f}% (+/- {:.2f}%)".format(np.mean(acc_scores)*100,
 np.std(acc_scores)*100))
print("Cross-
validation loss: {:.2f} (+/- {:.2f})".format(np.mean(loss_scores), np.std(l
oss_scores)))
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ξένη

[1] Davie CA. (2008). A review of Parkinson's disease, *British Medical Bulletin*, **86**, 109-27.

[2] Bloem BR, Okun MS, Klein C. (2021). Parkinson's disease, *The Lancet*, **397**, 2284-2303.

[3] Tolosa E, Garrido A, Scholz SW, Poewe W. (2021). Challenges in the diagnosis of Parkinson's disease, *The Lancet Neurology*, **20**, 385-397.

[4] Jagadeesan AJ, Murugesan R, Vimala Devi S, Meera M, Madhumala G, Vishwanathan Padmaja M, Ramesh A, Banerjee A, Sushmitha S, Khokhlov AN, Marotta F, Pathak S. (2017). Current trends in etiology, prognosis and therapeutic aspects of Parkinson's disease: a review, *Acta Biomed*, **23**, 249-262.

[5] E. Ray Dorsey et al. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016, *The Lancet Neurology*, **17**, 939-953.

[6] Goetz CG. (2011). The history of Parkinson's disease: early clinical descriptions and neurological therapies, *Cold Spring Harbor Perspectives Medicine*, **1**.

[7] Jankovic J, Tan EK. (2020). Parkinson's disease: etiopathogenesis and treatment, *Journal of Neurology, Neurosurgery, and Psychiatry*, **91**, 795-808.

[8] Blauwendraat, C., Nalls, M. A., & Singleton, A. B. (2019). The genetic architecture of Parkinson's disease, *The Lancet Neurology*, **19**, 170-178.

[9] Rodríguez-Martín D, Samà A, Pérez-López C, Català A, Moreno Arostegui JM, Cabestany J, et al. (2017). Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer, *PLoS ONE*, **2**.

[10] Camps, J., Samà, A., Martín, M., Rodríguez-Martín, D., Pérez-López, C., Alcaine, S., et al. (2017). Deep Learning for Detecting Freezing of Gait Episodes in Parkinson's Disease Based on Accelerometers, *Lecture Notes in Computer Science*, **10306**, 344–355.

[11] Samà, A., Pérez-López, C., Rodríguez-Martín, D., Català, A., Moreno-Aróstegui, J. M., Cabestany, et al. (2017). Estimating bradykinesia severity in Parkinson's disease by analysing gait through a waist-worn sensor, *Computers in Biology and Medicine*, **84**, 114–123.

[12] Rodríguez-Martín, D., Samà, A., Pérez-López, C. et al. (2020). Posture transition analysis with barometers: contribution to accelerometer-based algorithms, *Neural Comput & Applic*, **32**, 335–349.

[13] Lu M, Wei X, Che Y, Wang J, Loparo KA. (2020), Application of Reinforcement Learning to Deep Brain Stimulation in a Computational Model of Parkinson's Disease, *IEEE Trans Neural Syst Rehabil Eng*, **1**, 339-349.

[14] Tatsuoka, M. M., & Tiedeman, D. V. (1954). Chapter IV: Discriminant Analysis, *Review of Educational Research*, **5**, 402–420.

[15] Nagel, S. (2021). Chapter 2 Supervised Learning: Machine Learning in Asset Pricing, *Princeton: Princeton University Press*, 11-30.

[16] Andrew, A. (1999). REINFORCEMENT LEARNING: AN INTRODUCTION by Richard S. Sutton and Andrew G. Barto, Adaptive Computation and Machine Learning series, MIT Press (Bradford Book), Cambridge, Mass., **2**, 229-235.

[17] Trevor Hastie, Robert Tibshirani and Jerome Friedman (2008). The Elements of Statistical Learning.

[18] Cosma Rohilla Shalizi (2021). Advanced Data Analysis from an Elementary Point of View.

[19] Vishal Maini, Samer Sabri (2017). Machine Learning for Humans.

[20] https://jakevdp.github.io/PythonDataScienceHandbook/

Διαδίκτυο

[1] STAT-ON Holter, https://www.statonholter.com/ (Τελευταία πρόσβαση: 03/2023)

[2] Sense4Care, https://www.sense4care.com/ (Τελευταία πρόσβαση: 03/2023)

[3] Parkinson's Foundation, https://www.parkinson.org/understanding-parkinsons/statistics (Τελευταία πρόσβαση: 03/2023)