

Ιόνιο Πανεπιστήμιο

Τμήμα Πληροφορικής



-Πτυχιακή Εργασία-

**Εφαρμογή μηχανικής μάθησης σε δεδομένα ασθενών με
Alzheimer, για την ανάπτυξη μοντέλων διάγνωσης**

Φοιτητής
Καραγεωργίου Ανδρέας
Π2018122

Επιβλέπων καθηγητής
Θεμιστοκλής Έξαρχος

Τριμελής Επιτροπή
Θεμιστοκλής Έξαρχος – Παναγιώτης Βλάμος – Αριστείδης Βραχάτης

Κέρκυρα, Μάιος 2024

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ.....	3
1.ΕΙΣΑΓΩΓΗ	5
1.1 Η Ιστορία του Alzheimer	6
1.2 Τι Είναι η Νόσος Alzheimer	6
1.3 Αιτίες που Προκαλούν τη Νόσο Alzheimer.....	7
1.4 Στατιστικά Παγκοσμίως.....	8
1.4.1 Στατιστικά στην Ελλάδα	9
1.5 Διάγνωση	9
1.5.1 Η Διάγνωση και η Σημασία της Πληροφορικής	10
1.5.2 Σημασία Πρώιμης Διάγνωσης	10
2.ΒΙΒΛΙΟΓΡΑΦΙΚΉ ΑΝΑΣΚΟΠΗΣΗ.....	12
3.ΠΕΡΙΓΡΑΦΉ ΔΕΔΟΜΕΝΩΝ.....	23
3.1 Dataset.....	24
3.2 Σημασιολογία των Features του Dataset	24
3.3 Σημασιολογία των διαγνωστικών Features του Dataset.....	33
4.ΜΕΘΟΔΟΛΟΓΙΑ.....	35
4.1 Προετοιμασία Δεδομένων	36
4.2 Διαχείριση Ανισορροπίας στα Δεδομένα	36
4.3 Feature Importance	37
4.4 Εκπαίδευση και Βελτιστοποίηση Μοντέλων	38
4.5 Παράμετροι μοντέλων.....	39
5.ΑΠΟΤΕΛΕΣΜΑΤΑ	41
5.1 Αποτελέσματα αλγορίθμων με 10 fold cross-validation.....	42
5.1.1 Support Vector Machine Results.....	42
5.1.2 Logistic Regression Results.....	44
5.1.2 KNN Results.....	45
5.1.3 HistGradient Boosting Results.....	45
5.1.4 Gradient Boosting Results.....	47
5.1.5 Ada Boost Results.....	48
5.1.5 Random Forest Results.....	49

5.1.5 XGBOOST Results.	50
5.2 Επιλογή Μοντέλου.....	51
5.3 Βελτιστοποίηση τελικού Μοντέλου.	51
6.ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ	54
6.1 <i>Streamlit</i>	55
6.2 <i>Interface</i>	55
6.2.1. Κύρια Σελίδα (Home Page)	56
6.2.2 Ερωτηματολόγιο (Questionnaire)	56
6.2.3 Διαδικασία Πρόβλεψης (Prediction Process)	58
6.2.4. Σελίδα 'Σχετικά με Εμάς' (About Us).....	58
6.2.5.Προσαρμοσμένος Οπτικός Σχεδιασμός (Custom Visual Design).....	59
7.ΣΥΜΠΕΡΑΣΜΑΤΑ	60
8.ΑΝΑΦΟΡΕΣ	63
9.ΠΑΡΑΡΤΗΜΑΤΑ	67

1.ΕΙΣΑΓΩΓΉ

Η νόσος Αλτσχάιμερ , η ανάγκη για πρώιμη διάγνωση και ο στόχος της εφαρμογής.

1.1 Η Ιστορία του Alzheimer

Η νόσος Alzheimer, μία από τις πιο συνηθισμένες μορφές άνοιας, πήρε το όνομά της από τον Γερμανό νευροπαθολόγο Alois Alzheimer, ο οποίος την περιέγραψε πρώτος το 1906. Η ανακάλυψη έγινε κατά τη διενέργεια της νεκροψίας μιας ασθενούς, της Auguste Deter, η οποία πέθανε μετά από μία περίοδο σοβαρής γνωστικής υποβάθμισης. Αναζητώντας τις βιολογικές αιτίες της κατάστασής της, ο Alzheimer παρατήρησε δύο διακριτά παθολογικά σχηματισμούς στον εγκέφαλό της: τις πλάκες και τα ινώδη σφουγγώδη συστάδες[11].

Αυτά τα ευρήματα, που αργότερα ονομάστηκαν αμυλοειδείς πλάκες και νευροϊνικά νήματα, έγιναν κεντρικά στην κατανόηση της παθοφυσιολογίας της νόσου. Η αναφορά του Alzheimer σε μία συνεδρίαση της Ψυχιατρικής Κλινικής του Πανεπιστημίου της Φρανκφούρτης το ίδιο έτος σηματοδότησε την πρώτη επίσημη καταγραφή της νόσου, ανοίγοντας τον δρόμο για περαιτέρω έρευνες.

Από εκείνη την εποχή, η έρευνα για τη νόσο Alzheimer έχει επεκταθεί σημαντικά, με πολλές μελέτες να εστιάζουν στην αναζήτηση βιοδεικτών και πιο αποτελεσματικών θεραπειών. Η αρχική έρευνα του Alzheimer παρέχει ένα σημαντικό θεμέλιο για τη σύγχρονη νευροεπιστήμη και έχει διαμορφώσει τον τρόπο που κατανοούμε και προσεγγίζουμε τις νευροεκφυλιστικές διαταραχές σήμερα.[12]

1.2 Τι Είναι η Νόσος Alzheimer

Η νόσος Alzheimer είναι μια προοδευτική νευροεκφυλιστική διαταραχή που αποτελεί την πιο συχνή αιτία άνοιας στον πληθυσμό των ηλικιωμένων. Χαρακτηρίζεται από την σταδιακή και αμετάκλητη εκφυλιστική απώλεια νευρώνων, καθώς και από τη συσσώρευση παθολογικών πρωτεϊνών στον εγκέφαλο, γνωστών ως αμυλοειδείς πλάκες και νευροϊνικά νήματα.

Οι αμυλοειδείς πλάκες αποτελούνται κυρίως από την πρωτεΐνη βήτα-αμυλοειδούς, η οποία παράγεται φυσιολογικά αλλά συσσωρεύεται σε ανώμαλες ποσότητες στη νόσο Alzheimer, σχηματίζοντας σκληρά και τοξικά συσταδοποιημένα σύμπλοκα που διαταράσσουν τη νευρωνική λειτουργία. Τα νευροϊνικά νήματα, από την άλλη πλευρά, αποτελούνται από την πρωτεΐνη ταυ, η οποία συνήθως βοηθά στη σταθεροποίηση των μικροσωληνίσκων στα νευρώνα. Στη νόσο Alzheimer, η ταυ πρωτεΐνη υποβάλλεται σε υπερφωσφορυλίωση και σχηματίζει εντέλει επιζήμιες δομές που συμβάλλουν στην καταστροφή των νευρικών κυττάρων.



Εικόνα 1-Νόσος Αλτσχάιμερ

Η απώλεια των νευρωνικών κυττάρων και η δυσλειτουργία τους επιδεινώνεται συνεχώς και επηρεάζει διάφορες περιοχές του εγκεφάλου, ιδιαίτερα τις περιοχές που είναι υπεύθυνες για τη μνήμη και τη γνωστική λειτουργία. Η αρχική επίπτωση συνήθως είναι στην ικανότητα να θυμάται πρόσφατα γεγονότα ή πληροφορίες. Με την πάροδο του χρόνου, οι ασθενείς με νόσο Alzheimer αντιμετωπίζουν ευρύτερες προκλήσεις, όπως δυσκολίες στη σκέψη, στη λήψη αποφάσεων, στον προσανατολισμό, και στην επικοινωνία.

Τελικά, η νόσος επηρεάζει και την καθημερινή αυτοεξυπηρέτηση και την ανεξαρτησία, οδηγώντας σε σημαντική απώλεια αυτονομίας και αυξημένη ανάγκη για φροντίδα.

Η κατανόηση των βιολογικών και μοριακών βάσεων της νόσου Alzheimer έχει βελτιωθεί σημαντικά, αλλά η πλήρης εξήγηση της παθογένεσης και η ανάπτυξη αποτελεσματικών θεραπειών παραμένουν κεντρικά ερευνητικά ζητήματα.

1.3 Αιτίες που Προκαλούν τη Νόσο Alzheimer

Η νόσος Alzheimer είναι μια πολύπλοκη νευροεκφυλιστική διαταραχή με πολλαπλές αιτίες και παράγοντες κινδύνου που συμβάλλουν στην εμφάνισή της. Οι αιτίες και οι παράγοντες κινδύνου χωρίζονται σε γενετικούς, περιβαλλοντικούς και άλλους βιολογικούς παράγοντες.

Γενετικοί παράγοντες: Ορισμένες μορφές της νόσου Alzheimer είναι κληρονομικές, με γενετικές μεταλλάξεις που μεταβιβάζονται από τη μία γενιά στην επόμενη. Μεταλλάξεις σε γονίδια όπως τα APP, PSEN1 και PSEN2 έχουν συνδεθεί με την εμφάνιση της πρώιμης έναρξης Alzheimer, που εμφανίζεται συνήθως πριν την ηλικία των 65 ετών. Επίσης, το γονίδιο APOE ε4 θεωρείται ο σημαντικότερος γενετικός παράγοντας κινδύνου για την κοινότερη μορφή της νόσου, την καθυστερημένης έναρξης Alzheimer.



Εικόνα 2-Εξέλιξη της νόσου

Περιβαλλοντικοί παράγοντες: Παράγοντες όπως η διατροφή, η έκθεση σε βαρέα μέταλλα ή τοξικές ουσίες, και η φυσική δραστηριότητα επηρεάζουν επίσης τον κίνδυνο εμφάνισης της νόσου.

Συγκεκριμένα, η παχυσαρκία, ο καπνισμός, η υπέρταση, ο διαβήτης και τα χαμηλά επίπεδα εκπαίδευσης έχουν συνδεθεί με αυξημένο κίνδυνο Alzheimer.

Άλλοι βιολογικοί παράγοντες: Η ηλικία είναι ο σημαντικότερος μη-γενετικός παράγοντας κινδύνου για τη νόσο Alzheimer, με τον κίνδυνο να αυξάνεται σημαντικά μετά την ηλικία των 65 ετών. Το φύλο μπορεί επίσης να είναι ένας παράγοντας, καθώς οι γυναίκες φαίνεται να έχουν υψηλότερο κίνδυνο ανάπτυξης της νόσου, ενδεχομένως λόγω του μεγαλύτερου προσδόκιμου ζωής.[14]

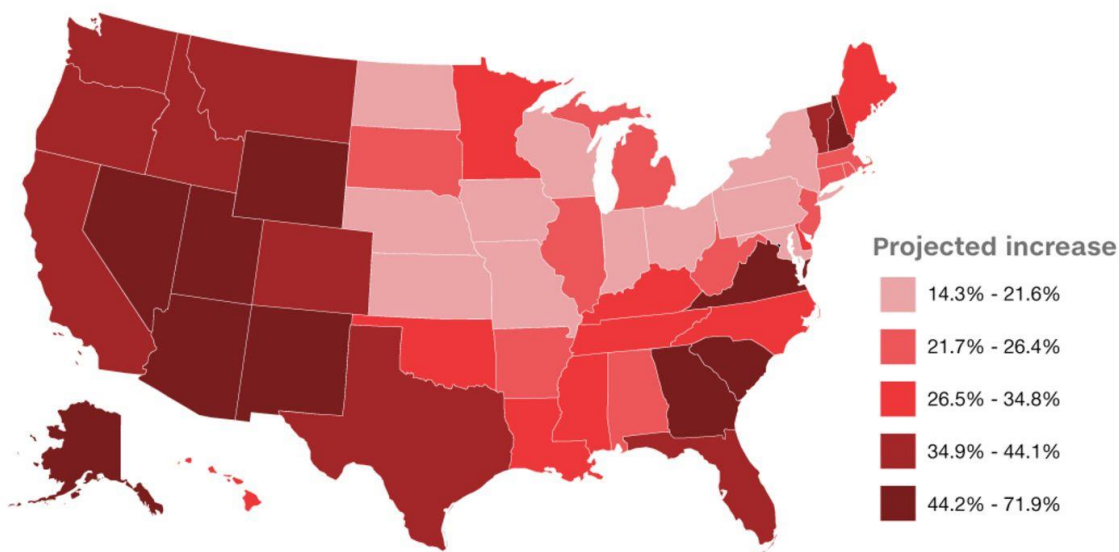
1.4 Στατιστικά Παγκοσμίως

Η νόσος Alzheimer αποτελεί μία από τις κυριότερες αιτίες άνοιας παγκοσμίως, με τα στατιστικά να δείχνουν συνεχή αύξηση της επίπτωσης λόγω του γηράσκοντος πληθυσμού. Εκτιμάται ότι περίπου 50 εκατομμύρια άνθρωποι ζουν με άνοια σε όλο τον κόσμο, και αναμένεται ότι αυτός ο αριθμός θα τριπλασιαστεί μέχρι το 2050. Η νόσος Alzheimer αντιπροσωπεύει περίπου το 60% έως 70% των περιπτώσεων άνοιας.

Στις Ηνωμένες Πολιτείες, περίπου 5.8 εκατομμύρια άνθρωποι έχουν διαγνωστεί με Alzheimer, με τις γυναίκες να εμφανίζουν υψηλότερο κίνδυνο σε σύγκριση με τους άνδρες. Σημαντικός αριθμός περιπτώσεων καταγράφεται και στην Ευρώπη, ενώ οι αναπτυσσόμενες χώρες αντιμετωπίζουν αυξανόμενες προκλήσεις λόγω της έλλειψης πρόσβασης σε διαγνωστικά εργαλεία και θεραπείες.

Η ανάγκη για βελτιωμένη παρακολούθηση και διαχείριση της νόσου γίνεται ολοένα και πιο επιτακτική, καθώς η επιβάρυνση της νόσου επηρεάζει όχι μόνο τους πάσχοντες, αλλά και τα οικονομικά και κοινωνικά συστήματα των χωρών παγκοσμίως. Η δημιουργία ενός παγκόσμιου δικτύου για την καταγραφή και τη μελέτη της νόσου αποτελεί έναν από τους πυλώνες για την αντιμετώπιση αυτής της πρόκλησης.[17]

PROJECTED RISE IN ALZHEIMER'S CASES BY 2025



SOURCE: Alzheimer's Association

Εικόνα 3-Projected Rise by 2025 USA

1.4.1 Στατιστικά στην Ελλάδα

Η νόσος Alzheimer αποτελεί ένα σημαντικό δημόσιο υγειονομικό ζήτημα και στην Ελλάδα, όπου ο αριθμός των περιπτώσεων αυξάνεται σταθερά, επηρεασμένος από το γηράσκον πληθυσμό. Εκτιμάται ότι περίπου 200,000 άτομα ζουν με άνοια στην Ελλάδα, με σημαντικό ποσοστό αυτών να πάσχει από νόσο Alzheimer. Η επίπτωση της νόσου αυξάνεται με την ηλικία, ενώ η κατανόηση και η διάγνυσή της βελτιώνονται μέσω της ανάπτυξης νέων τεχνικών και προγραμμάτων πρόληψης.

Στην Ελλάδα, οι προκλήσεις στην πρόσβαση σε θεραπεία και υποστήριξη για τους ασθενείς με Alzheimer είναι αξιοσημείωτες. Η οικονομική κρίση της τελευταίας δεκαετίας έχει περιορίσει τους πόρους του υγειονομικού συστήματος, κάτι που επηρεάζει τη διαθεσιμότητα και την ποιότητα της φροντίδας. Ενώ υπάρχουν κέντρα ειδικευμένα στη διαχείριση και τη θεραπεία της άνοιας, η γεωγραφική κατανομή και η πρόσβαση σε αυτά παραμένουν ανισομερείς.

Η καταγραφή και η παρακολούθηση των περιπτώσεων νόσου Alzheimer στην Ελλάδα χρήζει ενίσχυσης, καθώς η πλήρης επίδραση της νόσου στον πληθυσμό δεν είναι πλήρως κατανοητή. Προσπάθειες για τη βελτίωση της καταγραφής και την υλοποίηση εκστρατειών ενημέρωσης και εκπαίδευσης του κοινού έχουν αρχίσει να λαμβάνουν χώρα, με στόχο την ευαισθητοποίηση περί της νόσου και την ενθάρρυνση της πρόληψης και της έγκαιρης διάγνωσης.



Εικόνα 4-Alzheimers

1.5 Διάγνωση

Η διάγνωση της νόσου Alzheimer περιλαμβάνει μια σειρά από κλινικές αξιολογήσεις, νευροψυχολογικές δοκιμασίες και τη χρήση βιοδεικτών για την ακριβή ανίχνευση και διαφοροποίηση της νόσου από άλλες μορφές άνοιας. Οι μέθοδοι αυτοί επιτρέπουν την ακριβέστερη αξιολόγηση της νόσου και την εφαρμογή κατάλληλων θεραπευτικών προσεγγίσεων.

Κλινικές Αξιολογήσεις: Οι κλινικές αξιολογήσεις από νευρολόγους περιλαμβάνουν λεπτομερή ιατρικό ιστορικό και φυσική εξέταση. Οι γιατροί εστιάζουν στα συμπτώματα που αφορούν τη μνήμη, τη σκέψη, την οργάνωση, την κρίση και την αλλαγή στη συμπεριφορά.

Νευροψυχολογικές Δοκιμασίες: Οι νευροψυχολογικές δοκιμασίες είναι ζωτικής σημασίας για την αξιολόγηση των γνωστικών λειτουργιών του ασθενούς. Δοκιμασίες όπως το Mini-Mental State Examination (MMSE) ή το Montreal Cognitive Assessment (MoCA) χρησιμοποιούνται για να καταγράψουν τις γνωστικές απώλειες και να παρακολουθήσουν την εξέλιξη της νόσου.

Βιοδείκτες: Η χρήση βιοδεικτών γίνεται όλο και πιο σημαντική στη διαδικασία διάγνωσης του Alzheimer. Απεικονιστικές εξετάσεις όπως η μαγνητική τομογραφία (MRI) και η θετικής εκπομπής τομογραφία (PET) μπορούν να εντοπίσουν αλλαγές στη δομή του εγκεφάλου και την παρουσία αμυλοειδών πλακών. Επιπλέον, βιοδείκτες όπως οι πρωτεΐνες ταυ και βήτα-αμυλοειδούς στο εγκεφαλονωτιαίο υγρό παρέχουν πληροφορίες για τη βιοχημική κατάσταση του εγκεφάλου και συμβάλλουν στην πιο ακριβή διάγνωση.

Η συνδυασμένη χρήση αυτών των μεθόδων επιτρέπει μια πιο ολοκληρωμένη και έγκυρη αξιολόγηση της νόσου, ενισχύοντας την ικανότητα για έγκαιρη παρέμβαση και κατάλληλη θεραπεία.[15]

1.5.1 Η Διάγνωση και η Σημασία της Πληροφορικής

Η τεχνολογία πληροφορικής έχει καταστεί απαραίτητο εργαλείο στη διαγνωστική διαδικασία της νόσου Alzheimer, προσφέροντας σημαντικές δυνατότητες για βελτιωμένη ακρίβεια και ταχύτητα στην αξιολόγηση των ασθενών. Η χρήση συστημάτων πληροφορικής επιτρέπει την επεξεργασία και ανάλυση μεγάλων όγκων δεδομένων, όπως νευροαπεικονιστικά δεδομένα, γενετικές πληροφορίες και κλινικά στοιχεία, με ακρίβεια που ξεπερνά τις παραδοσιακές μεθόδους.

Ειδικότερα, η εισαγωγή της τεχνητής νοημοσύνης και των αλγορίθμων μηχανικής μάθησης στη διαδικασία διάγνωσης επιτρέπει την ανίχνευση πρόωρων μορφών της νόσου μέσω της αναγνώρισης προτύπων που δεν είναι ορατά με το ανθρώπινο μάτι. Αυτό οδηγεί σε πιο έγκαιρη και ακριβή διάγνωση, ενισχύοντας τη δυνατότητα των γιατρών να ορίσουν κατάλληλες θεραπευτικές προσεγγίσεις που μπορούν να βελτιώσουν σημαντικά την ποιότητα ζωής των ασθενών.

Επιπλέον, η ψηφιοποίηση των κλινικών αρχείων και η δημιουργία ολοκληρωμένων βάσεων δεδομένων βοηθούν τους ερευνητές να παρακολουθούν την εξέλιξη της νόσου και να αναπτύσσουν νέες θεραπείες. Η πληροφορική, συνεπώς, είναι καθοριστική τόσο για την κλινική πράξη όσο και για την ερευνητική ανάπτυξη στον τομέα της νόσου Alzheimer.[16]

1.5.2 Σημασία Πρώιμης Διάγνωσης

Η πρώιμη διάγνωση της νόσου Alzheimer κρίνεται ουσιαστική για τη βελτίωση της πρόγνωσης και την αποτελεσματική διαχείριση της νόσου. Η έγκαιρη ανίχνευση των συμπτωμάτων επιτρέπει την αμέσως εφαρμογή θεραπευτικών παρεμβάσεων που μπορεί να επιβραδύνουν την εξέλιξη της νόσου και να βελτιώσουν την ποιότητα ζωής των ασθενών.

Η πρόληψη και η πρώιμη θεραπεία συμβάλλουν στη διατήρηση της γνωστικής λειτουργίας και στη μείωση της πνευματικής και συναισθηματικής επιβάρυνσης τόσο για τους ασθενείς όσο και για τους φροντιστές τους. Επιπλέον, η έγκαιρη διάγνωση ενισχύει τον σχεδιασμό μακροπρόθεσμων φροντίδων, βοηθώντας τους ασθενείς και τις οικογένειές τους να

προετοιμαστούν καλύτερα για τις μελλοντικές ανάγκες υγείας και φροντίδας.

Η έγκαιρη διάγνωση μπορεί επίσης να συμβάλει στην ακριβέστερη διαφοροποίηση του Alzheimer από άλλες μορφές άνοιας, επιτρέποντας την εφαρμογή πιο στοχευμένων και αποτελεσματικών θεραπειών. Αυτό είναι ιδιαίτερα σημαντικό δεδομένου ότι η φαρμακευτική και μη φαρμακευτική αγωγή μπορεί να έχει διαφορετικές ενδείξεις και αποτελέσματα ανάλογα με την υποκείμενη αιτία της γνωστικής εξασθένησης.



Εικόνα 5-Σημασία Πρώιμης Διάγνωσης

Συνολικά, η πρώιμη διάγνωση είναι κρίσιμη για τη διαχείριση της νόσου Alzheimer και προσφέρει σημαντικά οφέλη, βελτιώνοντας την πρόγνωση και μεγιστοποιώντας την αποτελεσματικότητα των διαθέσιμων θεραπειών.[18]

2.ΒΙΒΛΙΟΓΡΑΦΙΚΉ ΑΝΑΣΚΟΠΗΣΗ

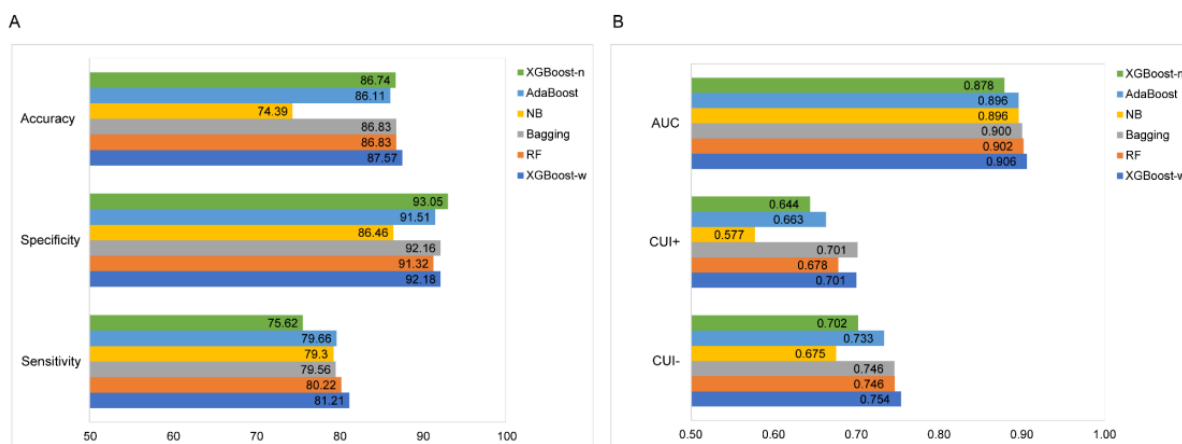
Μελέτη άρθρων παρόμοιων στόχων και διαστάσεων ,με σκοπό τη Βιβλιογραφική τους ανασκόπηση.

Το πρώτο άρθρο που εξετάστηκε στην βιβλιογραφική ανασκόπηση είναι αυτο με τίτλο **"XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease"**[1] παρουσιάζει μια καινοτόμο προσέγγιση για τη βελτίωση της ακρίβειας κατηγοριοποίησης και της ερμηνευτικότητας των μοντέλων μηχανικής μάθησης στη διάγνωση της νόσου του Αλτσχάιμερ (AD) μέσω πολυκατηγοριοποίησης. Οι συγγραφείς προτείνουν ένα ερμηνευτικό πλαίσιο μηχανικής μάθησης που ενσωματώνει το XGBoost με το SHAP (Shapley Additive exPlanations) για να αντιμετωπίσει τις προκλήσεις που ενέχει η ανισορροπία των κλάσεων που συνήθως παρατηρείται στην πρόοδο της AD από την κανονική γνωστική λειτουργία (NC) μέσω της ήπιας γνωστικής διαταραχής (MCI) προς τη νόσο του Αλτσχάιμερ.

Η μελέτη χρησιμοποιεί εκτενή δεδομένα ασθενών από τη βάση δεδομένων της Πρωτοβουλίας Νευροεικονικής για την Αλτσχάιμερ (ADNI), που περιλαμβάνει κλινικές πληροφορίες, αποτελέσματα νευροψυχολογικών δοκιμασιών και βιοδείκτες που προκύπτουν από νευροεικόνες. Το κύριο στόχο είναι να αντιμετωπίσουν την ανισορροπία των δεδομένων σε αλγοριθμικό επίπεδο και να παρέχουν ένα πιο ερμηνευτικό μοντέλο, καθορίζοντας τα πιο επιρρεπή χαρακτηριστικά που συμβάλλουν στις αποφάσεις διάγνωσης.

Οι συγγραφείς χρησιμοποίησαν τρεις αλγόριθμους επιλογής χαρακτηριστικών για τη βελτίωση των εισόδων του αλγορίθμου XGBoost. Μετέβαλαν την κατανομή του βάρους του δείγματος για να χειριστούν την ανισορροπία δεδομένων μεταξύ των κλάσεων NC, MCI και AD. Οι τιμές SHAP χρησιμοποιήθηκαν για να συνδέσουν αυτά τα χαρακτηριστικά με τα αποτελέσματα του μοντέλου, παρέχοντας έναν τρόπο για να κατανοήσουμε πώς κάθε χαρακτηριστικό επηρεάζει τις προβλέψεις του μοντέλου.

Το μοντέλο XGBoost-SHAP έδειξε ανώτερη απόδοση σε σύγκριση με παραδοσιακά μοντέλα όπως τα random forests, AdaBoost και Naive Bayes classifiers, επιτυγχάνοντας υψηλότερη ακρίβεια, ευαισθησία και ειδικότητα.



Πίνακας 1--Αποτελέσματα 1ου Άρθρου

Το μοντέλο απέδειξε την αποτελεσματικότητά του σε δύο βάσεις δεδομένων, την ADNI και το National Alzheimer's Coordinating Center (NACC), εξασφαλίζοντας τη σταθερότητα και τη συνέπεια στην προβλεπτική του ικανότητα.

Η μελέτη καταλήγει ότι το πλαίσιο XGBoost-SHAP βελτιώνει σημαντικά την ερμηνευτικότητα και την αποτελεσματικότητα των διαγνωστικών μοντέλων για τη νόσο του Αλτσχάιμερ, προσφέροντας ένα πολύτιμο εργαλείο για τους κλινικούς στις διαδικασίες λήψης αποφάσεων. Αυτή η πρόοδος θα μπορούσε πιθανόν να οδηγήσει σε καλύτερες στρατηγικές διαχείρισης και πιο πρόωρες παρεμβάσεις για τους ασθενείς σε διάφορα στάδια προόδου της AD.

Το δεύτερο άρθρο που εξετάστηκε έχει τίτλο “**Data analysis with Shapley values for automatic subject selection in Alzheimer’s disease data sets using interpretable machine learning**”[2] και επικεντρώνεται στην ανάπτυξη ενός μοντέλου μηχανικής μάθησης για την ανίχνευση και την ταξινόμηση των σταδίων της νόσου Alzheimer, χρησιμοποιώντας αλγορίθμους eXtreme Gradient Boosting (XGBoost) και SHapley Additive exPlanations (SHAP) για να παρέχει προβλέψεις που είναι κατανοητές και ερμηνεύσιμες. Αυτός ο πλαίσιο περιλαμβάνει μετρήσεις από νευροψυχολογικές δοκιμασίες, νευροεικονογραφήματα και γενετικούς δείκτες (π.χ., κατάσταση του γονιδίου APOE-ε4), προσδιορίζοντας τα σημαντικότερα χαρακτηριστικά που επηρεάζουν τις προβλέψεις της νόσου .

Η μεθοδολογία της έρευνας περιλαμβάνει την επιλογή χαρακτηριστικών μέσω τριών διαφορετικών αλγορίθμων επιλογής χαρακτηριστικών που στη συνέχεια ενσωματώθηκαν στον αλγόριθμο XGBoost. Η ανισορροπία μεταξύ των τριών κλάσεων (κανονική γνώση, ήπια γνωστική εμπίδωση και Alzheimer) αντιμετωπίζεται με την αλλαγή της κατανομής των βαρών δειγμάτων, ενώ το πλαίσιο SHAP συνδέεται με το XGBoost για να ενισχύσει την ερμηνευσιμότητα των προβλέψεων .

Table 10 Mean XGBoost accuracies and F1-scores (in %) for the independent ADNI test set (no information rate 55.56%)

Exclusion method	Zero-cutoff			Max valid			Max test		
	n	ACC (in %)	F1 (in %)	n	ACC (in %)	F1 (in %)	n	ACC (in %)	F1 (in %)
Random (-)	-	-	-	430	62.78	64.45	391	66.39	68.40
LOO (LR)	84	60.56	63.50	69	60.62	65.05	178	63.57	65.28
LOO (RF)	49	61.25	65.84	11	59.38	63.96	407	64.03	68.21
Data Shapley (LR)	152	68.06	71.02	248	65.07	67.89	144	68.47	71.22
Data Shapley (RF)	164	66.88	72.12	133	65.90	70.62	178	68.47	73.48

Πίνακας 2-Αποτελέσματα 2ου Άρθρου

Η αποτελεσματικότητα του προτεινόμενου πλαισίου αποδεικνύεται ανώτερη σε σύγκριση με άλλους αλγορίθμους όπως Random Forest, Bagging, AdaBoost και Naive Bayes σε διάφορες μετρήσεις απόδοσης όπως ακρίβεια, ευαισθησία, ειδικότητα, και AUC στα δεδομένα ADNI και NACC. Επιπλέον, τα δέκα πιο σημαντικά χαρακτηριστικά που εντοπίζονται μέσω των SHAP αξιών παρέχουν ουσιαστική βοήθεια στην λήψη κλινικών αποφάσεων, δείχνοντας έτσι το δρόμο για περαιτέρω έρευνα στην πρόληψη και την θεραπεία της νόσου Alzheimer.

Το τρίτο άρθρο με τίτλο **“Performances of Machine Learning Models for Diagnosis of Alzheimer’s Disease”**[3] αναλύει την απόδοση διαφόρων μοντέλων μηχανικής μάθησης στη διάγνωση της νόσου Αλτσχάιμερ, εστιάζοντας στην ανάπτυξη και σύγκριση αλγορίθμων που έχουν τη δυνατότητα να εντοπίζουν τη νόσο από νευροψυχολογικά δεδομένα. Η μελέτη στοχεύει στη βελτίωση της επιστημονικής κατανόησης σχετικά με το πώς διάφοροι μηχανικοί αλγόριθμοι μπορούν να συμβάλλουν στην πρώιμη και ακριβή διάγνωση της νόσου, ένα κρίσιμο βήμα για τη διαχείριση και τη θεραπεία της.

Η έρευνα χρησιμοποίησε δεδομένα από τη βάση δεδομένων OASIS, που περιλαμβάνουν εκτεταμένα νευροψυχολογικά τεστ και γενετικές πληροφορίες ασθενών, καθώς και νευροεικονογραφήματα. Οι συγγραφείς εφάρμοσαν τεχνικές μείωσης διαστάσεων και επιλογής χαρακτηριστικών για να βελτιώσουν την ποιότητα των εισόδων τους πριν την εκπαίδευση των μοντέλων. Αυτή η διαδικασία είχε στόχο την εξαγωγή των πιο σημαντικών δεικτών που επηρεάζουν την ανίχνευση της νόσου.

Διάφοροι αλγόριθμοι μηχανικής μάθησης δοκιμάστηκαν για την αξιολόγηση της απόδοσής τους σε αυτό το πλαίσιο, συμπεριλαμβανομένων του Stochastic Gradient Descent (SGD), k-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, Random Forests, AdaBoost, Neural Networks, Support Vector Machines (SVM) και Naïve Bayes. Η απόδοση κάθε αλγορίθμου αξιολογήθηκε με βάση μετρήσεις όπως η ακρίβεια, η ανάκληση, το F1 σκορ και το ROC-AUC, παρέχοντας μια λεπτομερή εικόνα των δυνατοτήτων τους στην πρόβλεψη της νόσου.

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
kNN	0.744	0.633	0.611	0.599	0.633	0.729	0.638	0.618	0.607	0.638
Decision tree	0.904	0.855	0.855	0.855	0.855	0.895	0.863	0.862	0.862	0.863
SVM	0.941	0.895	0.874	0.874	0.895	0.96	0.895	0.868	0.874	0.895
SGD	0.883	0.898	0.855	0.908	0.898	0.889	0.903	0.867	0.913	0.903

Table 4 (continued)

Model	Top-rated 6 features					Top-rated 7 features				
	AUC	CA	F1	Precision	Recall	AUC	CA	F1	Precision	Recall
Random forest	0.944	0.901	0.881	0.88	0.901	0.946	0.885	0.861	0.851	0.885
Neural network	0.949	0.887	0.861	0.862	0.887	0.953	0.903	0.875	0.888	0.903
Naive Bayes	0.941	0.893	0.882	0.876	0.893	0.944	0.882	0.876	0.871	0.882
Logistic regression	0.934	0.887	0.862	0.871	0.887	0.924	0.826	0.798	0.789	0.826
AdaBoost	0.872	0.853	0.853	0.853	0.853	0.868	0.847	0.848	0.849	0.847

Πίνακας 3-Αποτελέσματα 3ου Άρθρου

Η ανάλυση των αποτελεσμάτων αποκάλυψε σημαντικές διαφορές στην απόδοση μεταξύ των αλγορίθμων, με τον Random Forest και τα Neural Networks να δείχνουν την υψηλότερη απόδοση σε πολλές από τις μετρήσεις. Η μελέτη επεκτείνεται στη σημασία των πολυπαραμετρικών μοντέλων στην ακριβή διάγνωση, καθώς και στην ανάγκη για περαιτέρω έρευνα για τη βελτιστοποίηση των τεχνικών εκπαίδευσης και τη συνολική κατανόηση των μηχανισμών που διέπουν τη νόσο Alzheimer.

Το τέταρτο άρθρο το οποίο μελετήθηκε ήταν αυτό με τίτλο **“A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease”**[4]. Το άρθρο μελετά εκτενώς την εφαρμογή διαφόρων μοντέλων μηχανικής μάθησης για την πρώιμη διάγνωση της νόσου Alzheimer, χρησιμοποιώντας το σύνολο δεδομένων OASIS. Η έρευνα εστιάζεται στην αξιολόγηση και σύγκριση της απόδοσης κλασσικών και σύγχρονων ταξινομητών, όπως RandomForest, Support Vector Machines, και νευρωνικών δικτύων, αναλύοντας την επίδραση της επιλογής χαρακτηριστικών και της ρύθμισης υπερπαραμέτρων στην τελική ακρίβεια διάγνωσης.

Το άρθρο παρουσιάζει μια λεπτομερή ανάλυση των υπερπαραμέτρων χρησιμοποιώντας την τεχνική GridSearchCV, η οποία επιτρέπει την συστηματική δοκιμή και αξιολόγηση διαφόρων συνδυασμών ρυθμίσεων στα μοντέλα για την εύρεση του ιδανικού συνδυασμού που οδηγεί στην βέλτιστη απόδοση. Η εφαρμογή της διασταυρούμενης επικύρωσης εντός του GridSearchCV βοηθά στην παροχή ακριβέστερης εκτίμησης της γενικευτικής ικανότητας των μοντέλων και στην μείωση του κινδύνου υπερπροσαρμογής.

Table 4 Performance evaluation of several ML models

Model	Accuracy (%)	Precision	Recall	F1-score
GaussianNB	96.00	97.00	91.00	94.00
Decision Tree	89.33	97.00	91.00	94.00
Random forest	96.00	97.00	94.00	96.00
Voting classifier	96.00	100.0	43.00	60.00
Xgboost	93.33	97.00	97.00	97.00
GradientBoost	92.00	97.00	94.00	96.00

Πίνακας 4-Αποτελέσματα 4ου Άρθρου

Οι συγγραφείς του άρθρου αναφέρουν συγκεκριμένες μετρικές αξιολόγησης όπως η ακρίβεια, η ανάκληση, το F1 σκορ και η τιμή ROC-AUC, καθιστώντας δυνατή τη λεπτομερή σύγκριση μεταξύ των μοντέλων. Το αποτέλεσμα της έρευνας υπογραμμίζει τη σημασία της επιλογής των κατάλληλων χαρακτηριστικών και της ακριβούς ρύθμισης των μοντέλων για την επίτευξη υψηλότερης απόδοσης στην πρόβλεψη της νόσου Alzheimer.

Η πέμπτη μελέτη με τίτλο **“Detecting Alzheimer’s disease using machine learning methods”**[5] από τους Kia Dashtipour και συνεργάτες παρουσιάζει μια καινοτόμο προσέγγιση για την ανίχνευση της νόσου Αλτσχάιμερ με χρήση μεθόδων μηχανικής μάθησης (ML) και βαθιάς μάθησης (DL). Η έρευνα επικεντρώνεται στην ανάγκη για έγκαιρη ανίχνευση της νόσου λόγω της αυξημένης εμφάνισης της άνοιας στον γηράσκοντα πληθυσμό παγκοσμίως.

Οι συγγραφείς αναπτύσσουν ένα πλαίσιο βασισμένο στην εφαρμογή αλγορίθμων ML και DL, με σκοπό την ανίχνευση της νόσου από βιοϊατρικά δεδομένα. Το άρθρο εξετάζει διάφορους αλγορίθμους ML και DL, όπως τον δικτυακό αλγόριθμο μακράς βραχυπρόθεσμης μνήμης (LSTM) και τον δισδιάστατο συνελκτικό νευρωνικό δίκτυο (CNN), και συγκρίνει την αποτελεσματικότητά τους με παραδοσιακούς αλγορίθμους ML όπως οι μηχανές διανυσματικής υποστήριξης (SVM) και η λογιστική παλινδρόμηση.

Η ανάλυση των αποτελεσμάτων της έρευνας δείχνει ότι ο αλγόριθμος BiLSTM επιτυγχάνει την υψηλότερη ακρίβεια ανίχνευσης με ποσοστό 91.28%, επιδεικνύοντας την υπεροχή του έναντι των παραδοσιακών μεθόδων, παρόλα αυτά διακρίνεται πως ο χρόνος που χρειάζεται είναι μεγάλος σε σχέση με τις υπόλοιπες τεχνικές και τα στάνταρ που μέχρι τώρα γνωρίζαμε.

Methods	Training Accuracy	Testing Accuracy	Precision	Recall	F-Score	Time
SVM	80.78	78.56	0.78	0.76	0.76	2 m 21 s
Logistic Regression	80.23	78.12	0.78	0.77	0.78	2 m 8 s
MLP	79.81	78	0.78	0.77	0.78	2 m 6 s
KNN	77.25	76.58	0.76	0.75	0.76	2 m 2 s
Naive Bayes	75.89	73.28	0.73	0.72	0.73	2 m 19 s
Decision Tree	76.98	75.59	0.75	0.74	0.75	2 m 11 s
Random Forest	76.85	75.89	0.75	0.75	0.75	2 m 31 s
1D-CNN	88.59	86.14	0.86	0.85	0.86	8 m 28 s
2D-CNN	89.45	87	0.87	0.86	0.87	9 m 1 s
LSTM	91.26	90.85	0.90	0.90	0.90	10 m 17 s
BiLSTM	93.21	91.28	0.91	0.91	0.91	10 m 12 s

Πίνακας 5-Αποτελέσματα 5ου Άρθρου

Το άρθρο επισημαίνει τη σημασία της πρόληψης και της έγκαιρης αντιμετώπισης της άνοιας, καθώς η πρώιμη διάγνωση μπορεί να βοηθήσει στη βελτίωση της ποιότητας ζωής των ασθενών.

Το άρθρο είναι σημαντικό για τον τομέα της βιοϊατρικής έρευνας, καθώς συνδυάζει προηγμένες τεχνολογίες μηχανικής μάθησης με κλινικές εφαρμογές, επισημαίνοντας τη δυνατότητα των τεχνολογιών αυτών να συμβάλλουν σημαντικά στην πρόληψη και διαχείριση νευροεκφυλιστικών νοσημάτων.

Το έκτο άρθρο με τίτλο **"Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation"**[6] από τον Junhao Wen και συνεργάτες εξετάζει τη χρήση των νευρωνικών δικτύων συνέλιξης (CNN) για την ταξινόμηση της νόσου Αλτσχάιμερ μέσω ανατομικών εικόνων MRI. Στοχεύει στη διαπίστωση των περιορισμών των υφιστάμενων μελετών και προτείνει ένα πλαίσιο για αναπαραγωγίμη αξιολόγηση των μεθόδων ταξινόμησης.

Αρχικά, η μελέτη παρουσιάζει μια συστηματική ανασκόπηση της λογοτεχνίας, αναγνωρίζοντας τέσσερις κύριες προσεγγίσεις των CNN για την ταξινόμηση της νόσου Αλτσχάιμερ: i) ανάλυση σε επίπεδο 2D εικόνων, ii) 3D ανάλυση σε επίπεδο τμημάτων, iii) βασισμένη σε περιοχές

ενδιαφέροντος (ROI), και iv) ανάλυση σε επίπεδο 3D υποκειμένου. Επισημαίνει επίσης τη συχνή ύπαρξη διαρροής δεδομένων σε πολλές μελέτες, που οδηγεί σε προκατειλημμένες επιδόσεις.

Η δεύτερη συνεισφορά είναι η επέκταση ενός ανοιχτού πλαισίου για την ταξινόμηση της νόσου Αλτσχάιμερ με τη χρήση CNN και T1 βαρηκότητας MRI, περιλαμβάνοντας εργαλεία για την αυτόματη μετατροπή δεδομένων σε πρότυπο BIDS, καθώς και διαδικασίες προεπεξεργασίας και αξιολόγησης εικόνων. Εκτελέστηκαν διαφορετικές αρχιτεκτονικές CNN για την αξιολόγηση των μεθόδων ταξινόμησης και η σύγκριση με τις παραδοσιακές μεθόδους μηχανικής μάθησης όπως τα support vector machines (SVM).

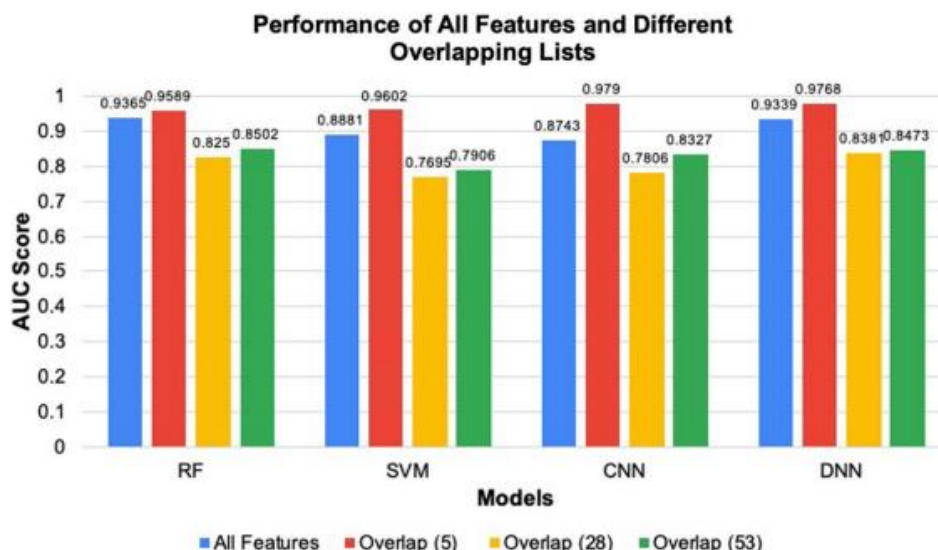
Classification architectures	Training data	Image preprocessing	Intensity rescaling	Data split	Training approach	Transfer learning	Task	Validation balanced accuracy	ADNI test balanced accuracy	AIBL test balanced accuracy	OASIS test balanced accuracy
2D slice-level CNN	Baseline	Minimal	MinMax	subject-level	single-CNN	ImageNet pre-train	AD vs CN	0.79 ± 0.04	0.76 [0.76, 0.75, 0.77, 0.75, 0.78]	0.76 [0.74, 0.76, 0.78, 0.75, 0.75]	0.65 [0.67, 0.62, 0.64, 0.65, 0.69]
	Longitudinal							0.74 ± 0.03	0.74 [0.81, 0.76, 0.70, 0.74, 0.72]	0.73 [0.72, 0.77, 0.72, 0.66, 0.79]	0.61 [0.62, 0.63, 0.64, 0.58, 0.60]
	Baseline			slice-level (data leakage)				1.00 ± 0	0.75 [0.74, 0.76, 0.75, 0.76, 0.75]	0.80 [0.80, 0.79, 0.82, 0.80, 0.81]	0.68 [0.68, 0.67, 0.69, 0.70, 0.66]
SVM		DartelGM	SPM-based	subject-level	None	None	AD vs CN	0.88 ± 0.02	0.88 [0.88, 0.87, 0.90, 0.90, 0.88]	0.88 [0.87, 0.90, 0.89, 0.90]	0.70 [0.71, 0.71, 0.70, 0.68, 0.72]
	Baseline						sMCI vs pMCI (trained on AD vs CN)	0.70 ± 0.06	0.75 [0.75, 0.75, 0.74, 0.76, 0.76]	0.60 [0.62, 0.54, 0.62, 0.59, 0.64]	--
							AD vs CN	0.87 ± 0.01	0.87 [0.85, 0.84, 0.90, 0.89, 0.87]	0.87 [0.88, 0.86, 0.88, 0.87, 0.89]	0.71 [0.73, 0.68, 0.72, 0.70, 0.71]
	Longitudinal						sMCI vs pMCI (trained on AD vs CN)	0.70 ± 0.02	0.76 [0.74, 0.75, 0.80, 0.77, 0.76]	0.68 [0.67, 0.66, 0.68, 0.67, 0.71]	--

Πίνακας 6-Αποτελέσματα του Άρθρου

Τέλος, επισημαίνεται η σημασία της χρήσης ανεξάρτητων σετ δεδομένων για την αποφυγή υπερπροσαρμογής και η ανάγκη για περαιτέρω αξιολόγηση για τη βελτίωση της γενίκευσης των μοντέλων σε διαφορετικά δεδομένα. Όλος ο κώδικας του πλαισίου και των πειραμάτων είναι δημοσίως διαθέσιμος, ενσωματώνοντας γενικού σκοπού εργαλεία στο λογισμικό Clinica, προσφέροντας μια βάση για την αξιολόγηση μεθόδων και αρχιτεκτονικών στον τομέα της νόσου Αλτσχάιμερ.

Στο έβδομο άρθρο **“Exploiting machine learning models to identify novel Alzheimer’s disease biomarkers and potential targets”**[7] που εξετάζεται, οι ερευνητές αξιοποιούν τεχνικές μηχανικής και βαθιάς μάθησης για να αναγνωρίσουν νέους βιοδείκτες και πιθανούς στόχους για τη νόσο του Alzheimer. Χρησιμοποιούνται τρία σύνολα δεδομένων γονιδιακής έκφρασης για τον προσδιορισμό κεντρικών γονιδίων μέσω εξαγωγής χαρακτηριστικών και αξιολόγησης με τεχνικές μηχανικής μάθησης και βαθιάς μάθησης. Η μελέτη επικεντρώθηκε στην αναγνώριση και την αξιολόγηση νέων βιοδεικτών για τη νόσο του Alzheimer, χρησιμοποιώντας συνδυασμό αλγορίθμων κατάταξης γονιδίων και μεθόδων επιλογής χαρακτηριστικών σε συνδυασμό με

μοντέλα μηχανικής και βαθιάς μάθησης. Αναγνωρίστηκαν πέντε κύρια γονίδια μέσω αυτών των μεθόδων, τα οποία ξεχώρισαν για την υψηλή τους απόδοση στη διάκριση δειγμάτων της νόσου από υγιή ελέγχους, με επιτυχία αντανάκλασης (AUC) 0.979. Επίσης, τονίζεται η σημασία της επιλογής γονιδίων που συμπεριλαμβάνουν τις μικρο-PNA και τους παράγοντες μεταγραφής που συσχετίζονται με τα υψηλά εκφρασμένα γονίδια, εντοπίζοντας σημαντικούς στόχους για θεραπείες



Πίνακας 7-Αποτελέσματα Του Άρθρου

Ταυτόχρονα, έγινε αναφορά ότι τα γονίδια αυτά μπορεί να συμβάλουν στην ανακάλυψη νέων στόχων για την παρέμβαση στη νόσο του Alzheimer. Παράλληλα, εξερευνήθηκαν επίσης γονίδια που εμπλέκονται σε νευρολογικές διαταραχές, υποδεικνύοντας περαιτέρω τη σημασία της επιλογής και ανάλυσης των κατάλληλων γονιδίων για την κατανόηση της νόσου. Η έρευνα αυτή αναδεικνύει τον ρόλο της χρήσης σύγχρονων βιοπληροφορικών και στατιστικών εργαλείων για τη βελτίωση της διαγνωστικής ακρίβειας και την εξερεύνηση νέων θεραπευτικών στόχων.

Στο όγδοο άρθρο με τίτλο **“Diagnosis of Alzheimer’s disease using hypergraph p -Laplacian regularized multi-task feature learning”**[8] εξετάζεται μία καινοτόμος προσέγγιση για τη διάγνωση της νόσου Αλτσχάιμερ μέσω της χρήσης μιας πολυεργαλειακής μεθόδου μάθησης χαρακτηριστικών με την επιβολή υπεργράφου p -Laplacian. Τα παραδείγματα δεδομένων περιλαμβάνουν εικόνες από MRI, γενετικές πληροφορίες και κλινικά δεδομένα, επικεντρώνοντας σε μια πολυμορφική προσέγγιση για την καλύτερη διάγνωση και κατανόηση της νόσου.

Η εφαρμογή της πολυεργαλειακής μεθόδου μάθησης χαρακτηριστικών με την επιβολή υπεργράφου p -Laplacian σε αυτή τη μελέτη καταδεικνύει την ενδυνάμωση των διαγνωστικών δυνατοτήτων μέσω της ανάλυσης συνδυασμένων βιολογικών και εικονικών δεδομένων. Η μέθοδος αυτή επιτρέπει την ανακάλυψη και την αξιολόγηση σύνθετων μοτίβων που ενδεχομένως να είναι αόρατα σε πιο παραδοσιακές αναλύσεις, οδηγώντας σε αυξημένη ακρίβεια και αποτελεσματικότητα.

Classification results using different modalities images in AD vs. pMCI , AD vs. sMCI , pMCI vs. NC, and sMCI vs. NC.

Data	AD vs. pMCI (%)			AD vs. sMCI (%)			pMCI vs. NC (%)			sMCI vs. NC (%)		
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
sMRI	77.66	90.05	51.96	89.54	91.10	87.86	92.42	84.29	95.56	73.83	53.08	88.95
FDG PET	62.79	75.05	38.57	72.51	76.43	68.24	74.18	55.00	81.75	54.39	37.20	66.75
AV-45 PET	41.66	41.71	41.59	60.50	76.81	26.96	61.91	38.04	71.29	52.65	44.78	58.45
sMRI+FDG PET	81.39	84.67	74.64	92.78	90.33	95.44	96.43	88.75	99.47	74.47	53.19	90.00
sMRI+AV-45 PET	71.60	73.71	67.68	90.31	85.48	95.49	96.02	87.32	99.44	75.11	55.44	89.50
FDG PET+AV-45 PET	79.98	80.48	78.93	92.03	92.38	91.65	96.02	90.00	98.36	75.77	54.78	91.14
ALL	83.22	84.67	80.54	94.61	92.43	96.98	98.40	94.29	1.00	77.36	56.32	92.81

Πίνακας 8-Αποτελέσματα 8ου Άρθρου

Οι αλγόριθμοι που χρησιμοποιήθηκαν κατέδειξαν ιδιαίτερα ενθαρρυντικά αποτελέσματα, με σημαντική βελτίωση των διαγνωστικών μετρήσεων σε σύγκριση με τις παραδοσιακές τεχνικές. Η ικανότητα της μεθόδου να ενσωματώνει και να αναλύει πολυδιάστατα δεδομένα από διαφορετικές πηγές επιτρέπει την ακριβέστερη πρόβλεψη και κατανόηση των μηχανισμών της νόσου Αλτσχάιμερ. Τα ευρήματα αυτά υποστηρίζουν την προοπτική της ένταξης της πολυεργαλειακής προσέγγισης στην κλινική πρακτική, προσφέροντας έναν πιο ολοκληρωμένο και αξιόπιστο τρόπο διάγνωσης, με την ελπίδα της βελτίωσης των θεραπευτικών παρεμβάσεων και της ποιότητας ζωής των ασθενών.

Στο ένατο άρθρο που εξετάστηκε ,με τίτλο **“An explainable machine learning approach for Alzheimer's disease classification”**[9] τονίζεται η πρόκληση της πρώιμης διάγνωσης της νόσου Alzheimer λόγω των διακριτικών αλλαγών στα βιοδείκτες που συχνά παραβλέπονται. Η έρευνα υπογραμμίζει τη σημασία της ερμηνευτικότητας των μοντέλων μαζί με την ακρίβειά τους, αντιμετωπίζοντας την πολυπλοκότητα των δεδομένων AD και τις περιορισμένες διαστάσεις του συνόλου δεδομένων. Επισημαίνονται προκλήσεις που συνδέονται με την υψηλή διαστασιολόγηση των δεδομένων.

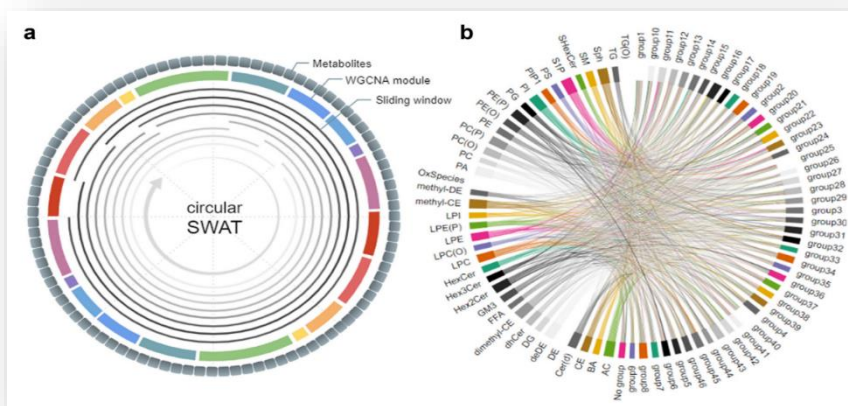
ML Model	Accuracy%	Precision%	Recall%	F1 score%	Mean%	SD	P-value
(a) Results of EXP1 : NC vs AD							
RF	97.8	97.2	98.1	97.6	97.8	0.002	
KNN	94.8	97.8	90.8	94.1	94.2	0.003	P<0.001
NB	96.2	93.8	98.3	96	96.1	0.002	P<0.001
SVM	97.6	97.6	97.2	97.4	97.6	0.003	P = 0.292
(b) Results of EXP1 : NC vs MCI							
RF	88.6	81.9	88.6	85.1	85.9	0.003	
KNN	76.8	81.2	48	60.3	59.5	0.006	P<0.001
NB	82.4	76.8	74.7	75.8	76.1	0.005	P<0.001
SVM	88.1	82.1	86.7	84.3	85.3	0.003	P = 0.003
(c) Results of EXP1 : MCI vs AD							
RF	87.3	90.2	88.1	89.1	90.5	0.002	
KNN	83.1	89.6	81	85	86.7	0.002	P<0.001
NB	82.4	92.5	76.4	83.7	85.6	0.004	P<0.001
SVM	87.6	90.4	88.5	89.4	90.3	0.003	P = 0.49
(d) Results of EXP1 : NC vs MCI vs AD							
RF	85.2	85.6	85.2	85.4	86.3	0.002	
KNN	75.5	74.1	75.5	73.4	73.4	0.005	P<0.001
NB	77.9	78.7	77.9	78	79.1	0.001	P<0.001
SVM	85.1	85.3	85.1	85.2	86	0.004	P = 0.20

Πίνακας 9-Αποτελέσματα 9ου Άρθρου

Στην έρευνα αξιοποιήθηκε ένα σύνολο δεδομένων από το Εθνικό Κέντρο Συντονισμού Alzheimer, που περιλαμβάνει 169,408 καταχωρήσεις και 1024 χαρακτηριστικά. Μετά από διάφορα βήματα για τη μείωση του χώρου των χαρακτηριστικών, τα μοντέλα μηχανικής μάθησης εκπαιδεύτηκαν και δοκιμάστηκαν με υψηλή απόδοση σε εξωτερικά δεδομένα. Τονίζεται η σημασία της χρήσης του εργαλείου Clinical Dementia Rating για την πρόβλεψη της νόσου Alzheimer.

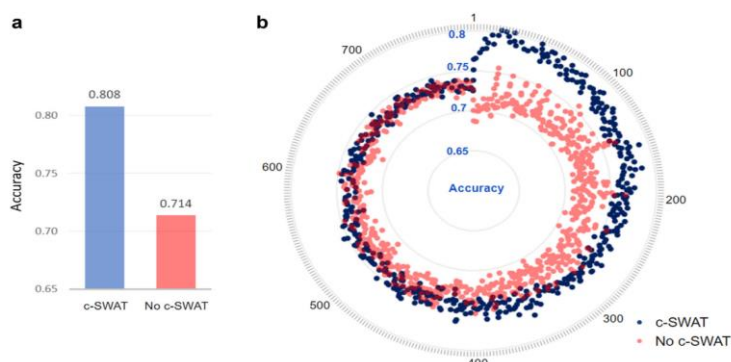
Συγκεκριμένα, τα μοντέλα Μηχανικής Μάθησης (ML) που χρησιμοποιήθηκαν περιλαμβάνουν τα Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB) και K-Nearest neighbour τα οποία έδειξαν υψηλή απόδοση στη διακριτική ανάλυση μεταξύ φυσιολογικών ατόμων και ατόμων με Alzheimer. Επιπλέον, χρησιμοποιήθηκαν μέθοδοι εξαγωγής κανόνων, όπως η εξόρυξη κανόνων κλάσης και η σταθερή και κατανοητή σειρά κανόνων για μοντέλα ταξινόμησης, για να καταστούν τα μοντέλα κατανοητά και ερμηνεύσιμα από ανθρώπους.

Στο δέκατο άρθρο **“Circular-SWAT for deep learning based diagnostic classification of Alzheimer's disease: application to metabolome data”**[10], η μελέτη εξετάζει την αποτελεσματικότητα της τεχνικής Circular-Sliding Window Association Test (c-SWAT) για την ταξινόμηση και διάγνωση της νόσου Alzheimer (AD) χρησιμοποιώντας δεδομένα μεταβολομικής ανάλυσης. Η προσέγγιση αυτή επιδιώκει να βελτιώσει την ακρίβεια των προβλεπτικών μοντέλων κατηγοριοποίησης μέσω της κατανόησης και της αξιολόγησης των συσχετίσεων μεταξύ των χαρακτηριστικών.



Πίνακας 10-CSWAT

Η c-SWAT χρησιμοποιεί μια τριβάθμια διαδικασία που περιλαμβάνει ανάλυση συσχέτισης χαρακτηριστικών, επιλογή χαρακτηριστικών και ταξινόμηση. Τα δεδομένα από 997 συμμετέχοντες της πρωτοβουλίας Alzheimer's Disease Neuroimaging Initiative (ADNI) χρησιμοποιήθηκαν για την εκπαίδευση και την επικύρωση του μοντέλου.



Πίνακας 11-Αποτελέσματα 10ου Άρθρου

Η εφαρμογή της c-SWAT κατέληξε σε ακρίβεια ταξινόμησης έως 80.8% και σε ένα AUC 0.808 για τη διάκριση μεταξύ ασθενών με AD και ηλικιωμένων χωρίς γνωστικά προβλήματα, σηματοδοτώντας βελτίωση 9.4% στην ακρίβεια και 0.169 στο AUC σε σύγκριση με μεθόδους χωρίς c-SWAT.

Τα αποτελέσματα δείχνουν ότι η c-SWAT είναι αποτελεσματική στη βελτίωση της ακρίβειας ταξινόμησης και στον εντοπισμό πιθανών βιοδεικτών λιπιδίων για τον AD. Τα εντοπισμένα λιπίδια προσφέρουν νέες δυνατότητες για την κατανόηση του AD και υποστηρίζουν περαιτέρω έρευνα.

3.ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΈΝΩΝ

Dataset, Features και σημασιολογία

3.1 Dataset

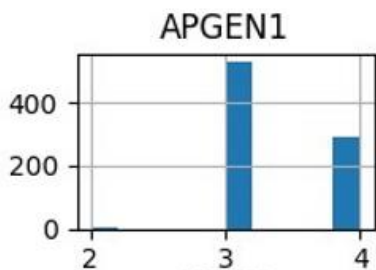
Τα δεδομένα που αναλύονται σε αυτή την πειραματική πτυχιακή εργασία προέρχονται από το Kaggle, μια ψηφιακή πλατφόρμα που ενώνει την κοινότητα της επιστημονικής ανάλυσης δεδομένων με εστίαση στη μηχανική μάθηση. Στο συγκεκριμένο dataset συμπεριλαμβάνονται στοιχεία από 820 ασθενείς, τα οποία περιέχουν διάφορα χαρακτηριστικά και ιατρικά ευρήματα που καταλήγουν σε διαγνώσεις της νόσου Alzheimer's.

Για κάθε ασθενή, το dataset περιλαμβάνει 31 διαφορετικά χαρακτηριστικά, τα οποία εστιάζουν κυρίως σε ιατρικές πληροφορίες. Η νόσος Alzheimer απαιτεί τη διεξαγωγή πολλαπλών ιατρικών εξετάσεων, που μπορεί να περιλαμβάνουν μαγνητικές τομογραφίες, σκιαγραφήσεις του εγκεφάλου, αιματολογικά δεδομένα και διάφορα δοκιμαστικά τεστ. Σημαντικό να σημειωθεί ότι η εν λόγω πτυχιακή εργασία δεν εστιάζει στην ανάλυση εικόνων από μαγνητικές τομογραφίες ή σκιαγραφήσεις. Η επεξεργασία τέτοιων εικόνων για την πρόωμη διάγνωση του Alzheimer's απαιτεί την υιοθέτηση προχωρημένων μεθόδων, όπως η ανάλυση πολλαπλών επιπέδων και η χρήση τεχνικών βαθιάς μάθησης. Το dataset που αναλύεται περιλαμβάνει αιματολογικά ευρήματα, κάποια γενετικά και δημογραφικά δεδομένα, καθώς και τα αποτελέσματα από γνωστικά τεστ.

3.2 Σημασιολογία των Features του Dataset

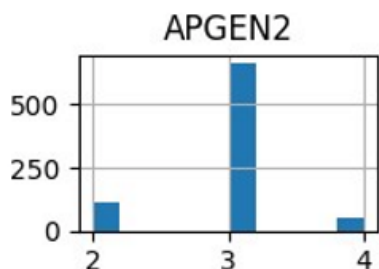
APGEN1 και APOE

Τα παραπάνω χαρακτηριστικά αναφέρονται σε δύο διαφορετικά είδη της απολιποπρωτεΐνης E (APOE), η οποία είναι μία σημαντική πρωτεΐνη που παίζει ρόλο στον μεταβολισμό των λιπιδίων στο σώμα.



Πίνακας 12-Apge1

Η APOE συμμετέχει επίσης στη διαδικασία μεταφοράς των χοληστερόλων και τριγλυκεριδίων μεταξύ των κυττάρων και ως εκ τούτου, διαδραματίζει έναν κρίσιμο ρόλο στην καρδιαγγειακή υγεία. Στο πλαίσιο της νόσου Alzheimer, η απολιποπρωτεΐνη E έχει ιδιαίτερη σημασία διότι ορισμένες μορφές της (ιδιαίτερα το επίσημα αναγνωρισμένο ισομόρφωμα ε4) έχουν συνδεθεί με αυξημένο κίνδυνο ανάπτυξης αυτής της νευροεκφυλιστικής διαταραχής.

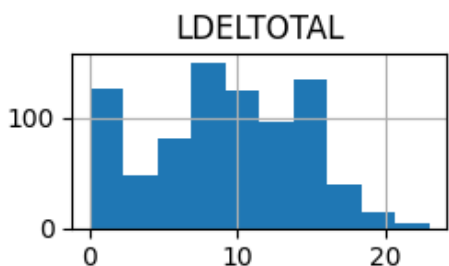


Πίνακας 13-Apge2

Η παρουσία των ειδών APOE1 και APOE2 σε ένα dataset μπορεί να υποδηλώνει την εξέταση διαφορετικών παραλλαγών της APOE, παρέχοντας σημαντικά δεδομένα για την ανάλυση του γενετικού κινδύνου που έχει κάποιος ασθενής για τη νόσο Alzheimer.

LDELTOTAL

Το χαρακτηριστικό αυτό (Logical Delayed Memory Total) αντιπροσωπεύει τη συνολική επίδοση ενός ασθενούς σε τεστ καθυστερημένης λογικής μνήμης.



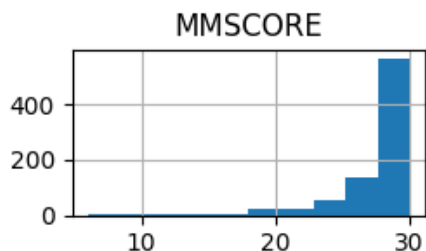
Πίνακας 14-LDELTOTAL

Αυτός ο τύπος τεστ μετρά την ικανότητα ενός ατόμου να θυμάται πληροφορίες μετά από μια καθορισμένη περίοδο χρόνου, συνήθως λίγα λεπτά έως και μία ώρα.

Η καταγραφή και αξιολόγηση της καθυστερημένης μνήμης είναι κρίσιμη για τη διάγνωση και την παρακολούθηση νευροεκφυλιστικών διαταραχών όπως η νόσος Alzheimer, καθώς η μείωση της μνημονικής λειτουργίας είναι ένα από τα πρώτα σημάδια.

MMSCORE

Αναφέρεται στην βαθμολογία που λαμβάνει ένας ασθενής στο Mini-Mental State Examination (MMSE), ένα ευρέως χρησιμοποιούμενο τεστ για την αξιολόγηση της γνωστικής λειτουργίας.

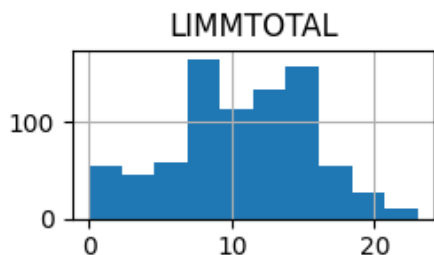


Πίνακας 15-MMSCORE

Το MMSE καλύπτει διάφορους τομείς όπως η προσοχή, η μνήμη, η ομιλία, η ικανότητα κατανόησης και η οπτικο-χωρική αντίληψη. Οι βαθμολογίες ποικίλλουν ανάλογα με την ηλικία και το εκπαιδευτικό επίπεδο του ατόμου, και χαμηλές βαθμολογίες μπορεί να υποδεικνύουν την παρουσία γνωστικής εξασθένησης.

LIMMTOTAL

Το (Logical Immediate Memory Total) μετρά την άμεση λογική μνήμη ενός ασθενούς, δηλαδή την ικανότητά του να θυμάται και να αναπαράγει πληροφορίες αμέσως μετά την παρουσίασή τους.

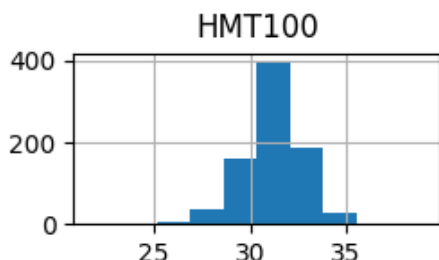


Πίνακας 16-LIMMTOTAL

Αυτό το χαρακτηριστικό είναι ζωτικό για την αξιολόγηση της λειτουργίας της μνήμης, καθώς επηρεάζει την ικανότητα του ατόμου να μαθαίνει νέα πράγματα και να αντιδρά σε νέες καταστάσεις. Τα αποτελέσματα από τέτοιες δοκιμασίες παρέχουν κρίσιμες πληροφορίες για την αρχική διάγνωση και την παρακολούθηση της προόδου της νόσου Alzheimer ή άλλων γνωστικών διαταραχών.

HMT100

Ο Μέσος Κυψελιδικός Αιμοσφαιρίνη (MCH-Mean Corpuscular Hemoglobin)) μετρά το μέσο περιεχόμενο αιμοσφαιρίνης σε μία μεμονωμένη ερυθρά αιμοσφαίρια.

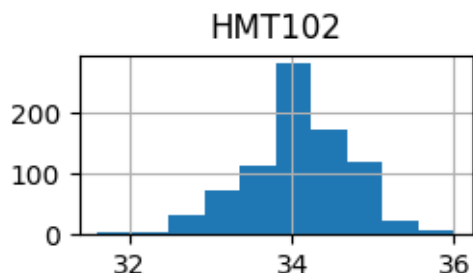


Πίνακας 17-HMT100

Χρησιμοποιείται για τη διάγνωση και τη διαχείριση διαφόρων αιματολογικών καταστάσεων όπως αναιμίες και άλλες διαταραχές της παραγωγής αίματος.

HMT102

Η Μέση Κυψελιδική Συγκέντρωση Αιμοσφαιρίνης (MCHC-Mean Corpuscular Hemoglobin Concentration) δείχνει τη συγκέντρωση αιμοσφαιρίνης στα ερυθρά αιμοσφαίρια.

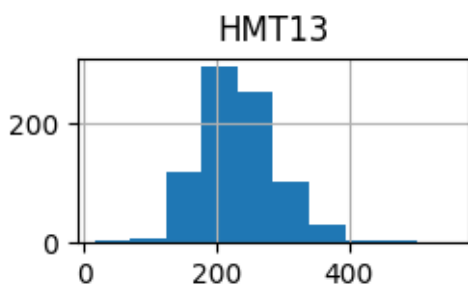


Πίνακας 18-HMT102

Αυτός ο δείκτης είναι χρήσιμος για την αξιολόγηση της αιμοσφαιρίνης σε σχέση με τον όγκο των ερυθρών αιμοσφαιρίων και ενδείκνυται για τη διάγνωση τύπων αναιμίας.

HMT13

Τα αιμοπετάλια(Platelets) ή πλακούντες παίζουν κεντρικό ρόλο στην αιμόσταση και την επούλωση των τραυμάτων.

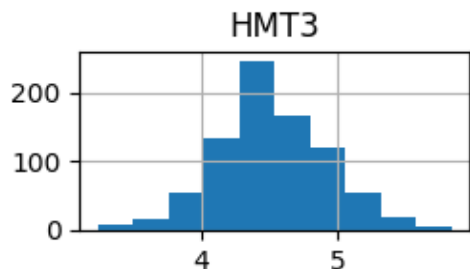


Πίνακας 19-HMT13

Ο αριθμός των πλακουντίων στο αίμα είναι ένδειξη για διάφορες καταστάσεις όπως διαταραχές πήξης, φλεγμονές και ορισμένες αιματολογικές διαταραχές.

HMT3

Ο αριθμός των ερυθρών αιμοσφαιρίων (RBC-Red Blood Cells) είναι ζωτικής σημασίας για τη μεταφορά οξυγόνου στα κύτταρα του σώματος και την απομάκρυνση διοξειδίου του άνθρακα.

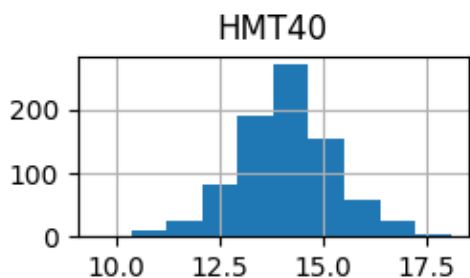


Πίνακας 20-HMT3

Οι ανωμαλίες στον αριθμό των RBC μπορεί να υποδεικνύουν αναιμία, υδροχοΐα, ή άλλες καταστάσεις.

HMT40

Η αιμοσφαιρίνη(Hemoglobin) είναι η πρωτεΐνη μέσα στα ερυθρά αιμοσφαίρια που δεσμεύει το οξυγόνο και διευκολύνει τη μεταφορά του στους ιστούς του σώματος.

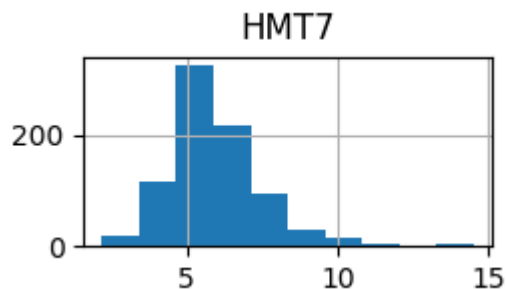


Πίνακας 21-HMT40

Οι μετρήσεις αιμοσφαιρίνης είναι κρίσιμες για τη διάγνωση της αναιμίας και άλλων καταστάσεων που επηρεάζουν την ικανότητα του αίματος να μεταφέρει οξυγόνο.

HMT7

Τα λευκά αιμοσφαίρια (WBC- White Blood Cells) είναι κύριοι παίκτες στο ανοσοποιητικό σύστημα, αντιδρώντας σε λοιμώξεις και ξένα σώματα.

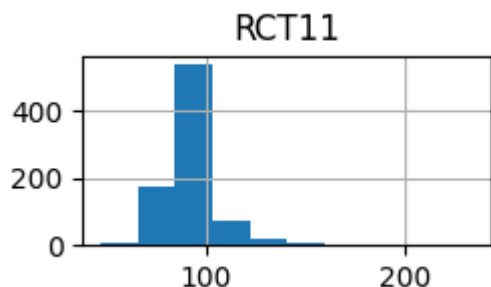


Πίνακας 22-HMT7

Ο αριθμός των WBC μπορεί να αυξηθεί σε περιπτώσεις λοίμωξης ή φλεγμονής και να μειωθεί σε καταστάσεις ανοσοκαταστολής ή ορισμένες αιματολογικές διαταραχές.

RCT11

Η γλυκόζη στον ορό (Serum Glucose) είναι ένας θεμελιώδης δείκτης για τον έλεγχο του μεταβολισμού της γλυκόζης και την αξιολόγηση του διαβήτη.

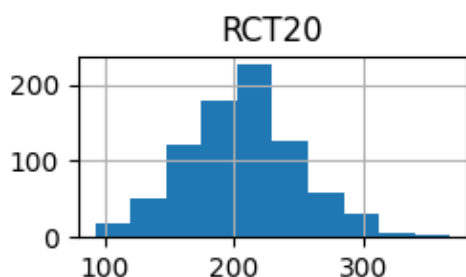


Πίνακας 23-RCT11

Η μέτρηση της συγκέντρωσης γλυκόζης στο αίμα είναι κρίσιμη για τη διάγνωση και τη διαχείριση του διαβήτη, αλλά και για την ανίχνευση άλλων καταστάσεων όπως υπογλυκαιμία ή υπεργλυκαιμία.

RCT20

Η χοληστερόλη (High Performance) αναφέρεται στην μέτρηση της χοληστερόλης με χρήση εξελιγμένων τεχνικών ανάλυσης για αυξημένη ακρίβεια.

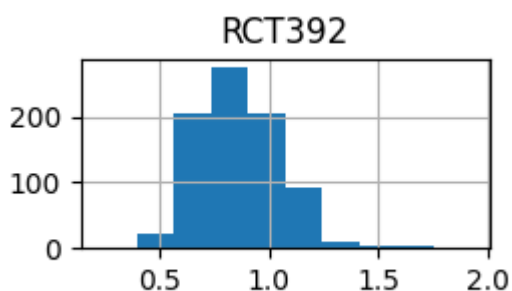


Πίνακας 24-RCT20

Η χοληστερόλη είναι ένας λιπίδιος παράγοντας στο αίμα που συνδέεται με τον κίνδυνο ανάπτυξης καρδιαγγειακών παθήσεων. Η καλή διαχείριση των επιπέδων χοληστερόλης μπορεί να βοηθήσει στην πρόληψη ή την καθυστέρηση της αρτηριοσκλήρωσης και άλλων καρδιακών παθήσεων.

RCT392

Το κρεατινίνη (Rate Blanked) είναι ένα προϊόν καταβολισμού που παράγεται από τον μυϊκό μεταβολισμό και απεκκρίνεται από τα νεφρά.

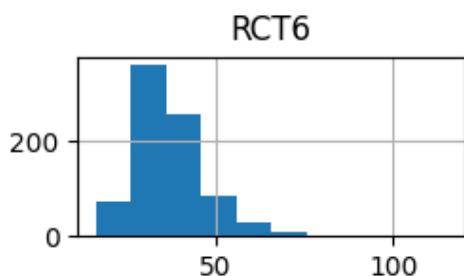


Πίνακας 25-RCT392

Η μέτρηση της κρεατινίνης στο αίμα χρησιμοποιείται για να εκτιμήσει τη νεφρική λειτουργία. Αυξημένα επίπεδα μπορεί να υποδεικνύουν νεφρική βλάβη ή νεφρική ανεπάρκεια.

RCT6

Το ουρικό άζωτο(Urea Nitrogen), ή BUN, είναι ένα προϊόν του μεταβολισμού των πρωτεϊνών στο σώμα και απεκκρίνεται μέσω των νεφρών.

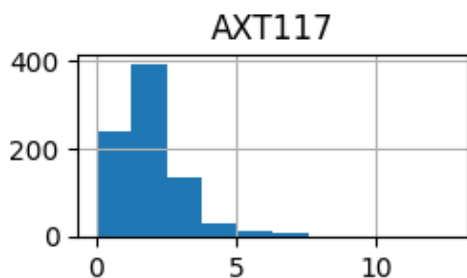


Πίνακας 26-RCT6

Η μέτρηση του BUN χρησιμοποιείται για την αξιολόγηση της νεφρικής λειτουργίας και υγείας. Επίσης, μπορεί να αναδείξει άλλες καταστάσεις, όπως η αφυδάτωση, το υπερκαταβολικό στάδιο, ή η καρδιακή ανεπάρκεια.

AXT117

Η Διεγερτική Ορμόνη του Θυρεοειδούς (Thyroid Stimulating Hormone, TSH) είναι ένας κρίσιμος δείκτης για την αξιολόγηση της λειτουργίας του θυρεοειδούς αδένος, που βρίσκεται στον λαιμό και παράγει ορμόνες οι οποίες ρυθμίζουν τον μεταβολισμό, την ενέργεια και την γενική ομοιόσταση του σώματος.



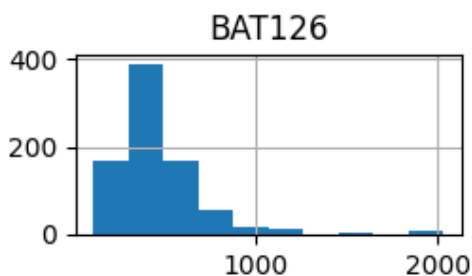
Πίνακας 27-AXT117

Το τεστ TSH χρησιμοποιείται για τη διάγνωση υποθυρεοειδισμού ή υπερθυρεοειδισμού, δύο καταστάσεις που δηλώνουν υπολειτουργία ή υπερλειτουργία του θυρεοειδούς αντίστοιχα. Μια μη φυσιολογική τιμή στο TSH μπορεί να υποδείξει ανάγκη για περαιτέρω ιατρική αξιολόγηση και παρακολούθηση.

BAT126

Η βιταμίνη B12 είναι ζωτικής σημασίας για την υγεία του νευρικού συστήματος, τη σύνθεση του DNA και την παραγωγή ερυθρών αιμοσφαιρίων.

Η δοκιμασία για την βιταμίνη B12 είναι σημαντική για τη διάγνωση της αναιμίας

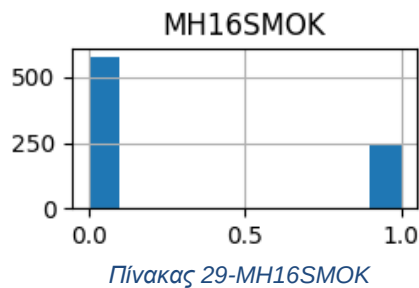


Πίνακας 28-BAT126

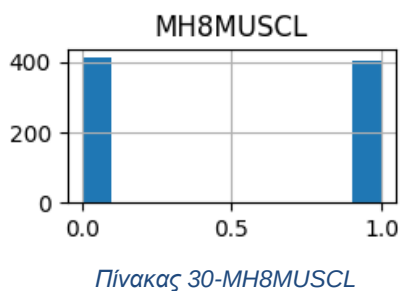
μεγαλοβλαστικού τύπου και μπορεί να δείξει καταστάσεις όπως η αδυναμία απορρόφησης της βιταμίνης από το έντερο. Η έλλειψη βιταμίνης B12 μπορεί επίσης να συνδεθεί με νευρολογικά συμπτώματα όπως μούδιασμα και τινάγματα στα άκρα, κόπωση, αδυναμία, και ψυχικές διαταραχές όπως διαταραχή στη συγκέντρωση και μνήμη. Η έγκαιρη διάγνωση και θεραπεία της έλλειψης βιταμίνης B12 είναι κρίσιμη για τη ναποτροπή μακροπρόθεσμων βλαβών.

Τα παρακάτω χαρακτηριστικά παρέχουν ένα εκτεταμένο και διαφοροποιημένο φάσμα μετρήσεων που αξιολογούν διάφορες πτυχές της υγείας ενός ατόμου. Κάθε μετρική προσφέρει μια σημαντική εικόνα της φυσιολογικής και παθολογικής κατάστασης σε εξειδικευμένους τομείς της υγείας:

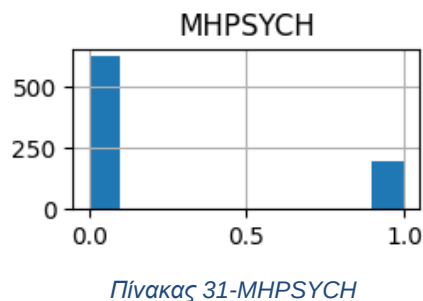
MH16SMOK (Smoking History Metric) αξιολογεί την ιστορία καπνίσματος του ατόμου, η οποία είναι σημαντική για την πρόληψη και διαχείριση αναπνευστικών και καρδιαγγειακών παθήσεων.



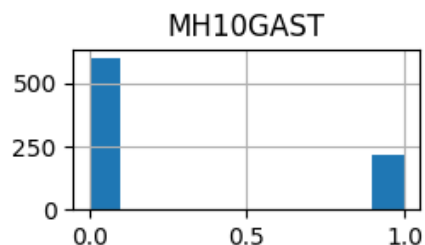
MH8MUSCL (Musculoskeletal Assessment Score) μετρά την κατάσταση του μυοσκελετικού συστήματος, ενδεικτική της κινητικότητας και της φυσικής λειτουργικότητας.



MHPSYCH (Mental Health Psychometric Score) διερευνά την ψυχική υγεία, παρέχοντας ενδείξεις για ψυχολογικά ζητήματα ή διαταραχές.

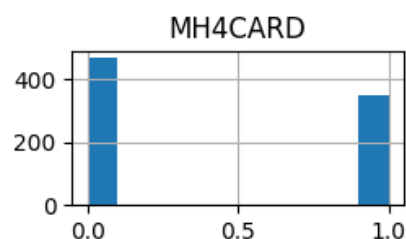


MH10GAST (Gastrointestinal Health Score) αξιολογεί τη γαστρεντερική λειτουργία και υγεία, κρίσιμη για την καλή διατροφική κατάσταση και απορρόφηση θρεπτικών συστατικών.



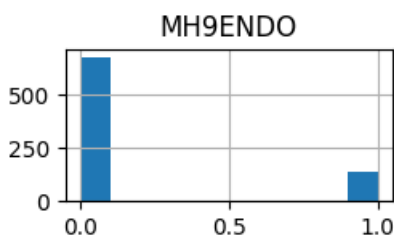
Πίνακας 32-MH10GAST

MH4CARD (Cardiovascular Risk Metric) εκτιμά τον κίνδυνο για καρδιαγγειακές παθήσεις με βάση διάφορους παράγοντες όπως υπέρταση, χοληστερόλη και ιστορικό οικογένειας.



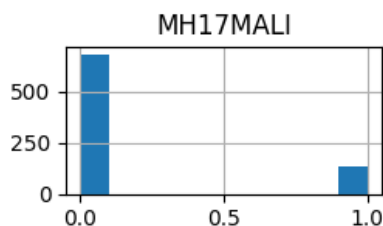
Πίνακας 33-MH4CARD

MH9ENDO (Endocrine Function Indicator) παρέχει στοιχεία για τη λειτουργία των ενδοκρινών αδένων, συμπεριλαμβανομένης της θυρεοειδούς λειτουργίας και του μεταβολισμού της γλυκόζης.



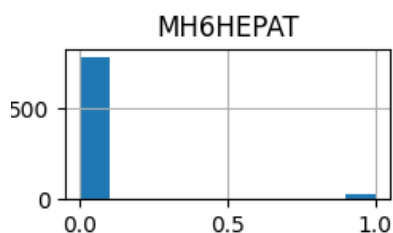
Πίνακας 34-MH9ENDO

MH17MALI (Malignancy Risk Factor) εκτιμά τον κίνδυνο ανάπτυξης κακοήθειας, προβλέποντας πιθανά περιστατικά καρκίνου.



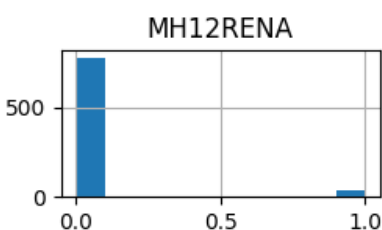
Πίνακας 35-MH17MALI

MH6HEPAT (Liver Function Marker) αξιολογεί τη λειτουργία του ήπατος, που είναι ζωτικής σημασίας για τη μεταβολική υγεία και την επεξεργασία τοξινών.



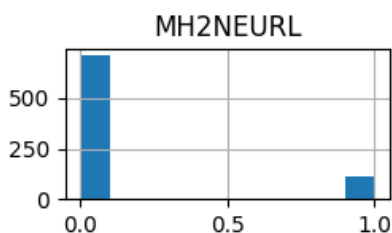
Πίνακας 36-MH6HEPAT

MH12RENA (Renal Function Score) παρέχει δείκτες για την κατάσταση της νεφρικής λειτουργίας, η οποία είναι κρίσιμη για την αποβολή αποβλήτων και τη ρύθμιση των ηλεκτρολυτών.



Πίνακας 37-MH12RENA

MH2NEURL (Neurological Assessment) αξιολογεί τη νευρολογική κατάσταση και τις πιθανές διαταραχές, παρέχοντας ενδείξεις για προβλήματα όπως νευροπάθειες και άλλες νευρολογικές διαταραχές.



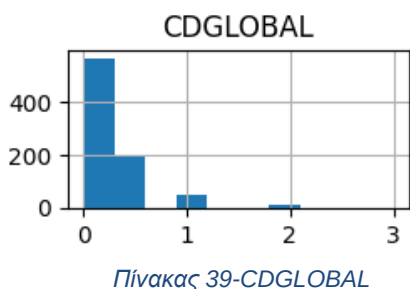
Πίνακας 38-MH2NEURL

Κάθε ένας από αυτούς τους δείκτες είναι καθοριστικός για την ολοκληρωμένη αξιολόγηση της υγείας και την ανάπτυξη εξατομικευμένων θεραπευτικών προσεγγίσεων για τον ασθενή.

CDGGLOBAL

Τέλος, το χαρακτηριστικό (Cognitive Decline Score) αναφέρεται σε έναν συνθετικό δείκτη που χρησιμοποιείται για την αξιολόγηση της γνωστικής πτώσης σε άτομα, κυρίως στο πλαίσιο της διερεύνησης νευροεκφυλιστικών παθήσεων όπως η νόσος Alzheimer. Ο σκορ αυτός συνήθως προκύπτει από τη συνδυασμένη ανάλυση διαφόρων νευροψυχολογικών δοκιμασιών που καλύπτουν πολλαπλές πτυχές της γνωστικής λειτουργίας, όπως η μνήμη, η λογική, η ικανότητα επίλυσης προβλημάτων, η γλωσσική κατανόηση, και άλλες.

Αυτός ο δείκτης είναι πολύτιμος για την παρακολούθηση της προόδου της νόσου και την απόκριση σε θεραπείες, καθώς παρέχει έναν αντικειμενικό τρόπο για την καταγραφή των αλλαγών στη γνωστική ικανότητα των ασθενών με το πέρασμα του χρόνου. Μπορεί επίσης να βοηθήσει τους ερευνητές να καθορίσουν την επιδείνωση ή τη σταθεροποίηση της γνωστικής λειτουργίας σε πειραματικά πρωτόκολλα και κλινικές μελέτες, καθιστώντας τον ένα σημαντικό δείκτη για την αξιολόγηση της αποτελεσματικότητας νέων θεραπειών.



3.3 Σημασιολογία των διαγνωστικών Features του Dataset

Το συγκεκριμένο Dataset μας παρέχει μια πληθώρα διαγνωστικών features, με 4 μάλιστα διαγνωστικά στοιχεία. Τα features αυτά ήταν τα εξής :DXCURREN, DXNORM, DXMCI, DXAD.

Τα features DXCURREN, DXNORM, DXMCI, και DXAD στο dataset παρέχουν κρίσιμες πληροφορίες για την κατάσταση της γνωστικής λειτουργίας των ασθενών. Κάθε ένα από αυτά τα χαρακτηριστικά χρησιμοποιείται για να διακρίνει τη διάγνωση του ασθενούς σε σχετικά με τη γνωστική του κατάσταση.

DXCURREN : αναφέρεται στην τρέχουσα διάγνωση του ατόμου, καταγράφοντας εάν ο ασθενής έχει ήδη κάποια γνωστική διαταραχή.

DXNORM : υποδεικνύει ότι το άτομο δεν έχει καμία γνωστική διαταραχή και η γνωστική λειτουργία του είναι κανονική. Αυτός ο δείκτης χρησιμοποιείται ως ένα baseline για τη σύγκριση άλλων γνωστικών καταστάσεων.

DXMCI : δηλώνει ότι το άτομο έχει διαγνωστεί με Ήπια Γνωστική Εξασθένηση (Mild Cognitive Impairment, MCI), μια κατάσταση που μπορεί να προηγείται της εξέλιξης σε σοβαρότερες γνωστικές διαταραχές, όπως η νόσος Alzheimer.

DXAD : επιδεικνύει ότι ο ασθενής έχει διαγνωστεί με νόσο Alzheimer, μια σοβαρή νευροεκφυλιστική διαταραχή που επηρεάζει τη μνήμη, τη σκέψη και τη συμπεριφορά.

Αυτά τα διαγνωστικά features είναι θεμελιώδη για την ακριβή καταγραφή και την ανάλυση της πορείας και της εξέλιξης των γνωστικών διαταραχών σε έναν πληθυσμό, και αποτελούν κρίσιμα δεδομένα για κλινικές μελέτες και την ανάπτυξη των θεραπειών.

Είναι σημαντικό να αναφερθεί ότι το feature DXCURREN δεν εντάχθηκε στο μοντέλο διάγνωσης, καθώς ο πρωταρχικός στόχος της εφαρμογής είναι ο συγκεκριμένος καθορισμός της εκάστοτε νόσου. Σε αυτό το πλαίσιο, δημιουργήθηκε ένα νέο feature, το DXTYPE, το οποίο συγκεντρώνει τα τρία υπόλοιπα στοιχεία: DXNORM, DXMCI, και DXAD. Αυτό το feature λειτουργεί ως κατευθυντήριος μοχλός για τη διάγνωση του ασθενούς και λαμβάνει τις τιμές 0, 1, ή 2.

Πιο αναλυτικά:

DXTYPE = 0: Σημαίνει ότι ο ασθενής δεν εμφανίζει συμπτώματα της νόσου, δηλαδή η κατάσταση του είναι φυσιολογική.

DXTYPE = 1: Υποδηλώνει ότι ο ασθενής έχει διαγνωστεί με Ήπια Γνωστική Διαταραχή, μια προκλινική κατάσταση που μπορεί να προηγηθεί σοβαρότερων γνωστικών προβλημάτων.

DXTYPE = 2: Αντιστοιχεί σε διάγνωση της νόσου Alzheimer, επιβεβαιώνοντας την παρουσία αυτής της σοβαρής νευροεκφυλιστικής διαταραχής.

Η εισαγωγή του DXTYPE ως κεντρικού διαγνωστικού εργαλείου επιτρέπει την ακριβέστερη και συστηματική αξιολόγηση των γνωστικών καταστάσεων των ασθενών, παρέχοντας έναν αξιόπιστο οδηγό για την ιατρική παρακολούθηση και τη θεραπευτική προσέγγιση.

Στους παρακάτω πίνακες φαίνονται οι τιμές που παίρνει το κάθε feature πιο αναλυτικά .

Float Features	Min	Max
HMT3	3.24	5.84
HMT7	2.2	14.5
HMT13	16	556
HMT40	9.5	18.1
HMT100	21.7	39
HMT102	31.6	36
RCT6	16.219	115.334
RCT11	46.842	234.208
RCT20	92.796	367.317
RCT392	0.226	1.923
AXT117	0.009	12.66
BAT126	116.562	2033.056
CDGLOBAL	0	3
MMSCORE	6	30
LIMMTOTAL	0	23
LDELTOTAL	0	23

Πίνακας 40-Float Features

INT Features	Min	Max	Definition
APGEN1	2	4	2 OR 3 OR 4
APGEN2	2	4	2 OR 3 OR 4
MHPSYCH	0	1	Yes OR No
MH2NEURL	0	1	Yes OR No
MH4CARD	0	1	Yes OR No
MH6HEPAT	0	1	Yes OR No
MH8MUSCL	0	1	Yes OR No
MH9ENDO	0	1	Yes OR No
MH10GAST	0	1	Yes OR No
MH12RENA	0	1	Yes OR No
PTGENDER	1	2	Male(1) OR Female(2)
DXTYPE	0	2	Normal(0) OR MCI(1) OR AD(2)

Πίνακας 41-Integer Features

4.ΜΕΘΟΔΟΛΟΓΙΑ

Η μεθοδολογία και η λογική επιλογής και δημιουργίας του τελικού μοντέλου.

4.1 Προετοιμασία Δεδομένων

Αρχικά, φορτώνεται το dataset από ένα CSV αρχείο ("AIBL.csv"), που περιέχει δεδομένα για τη νόσο Alzheimer.

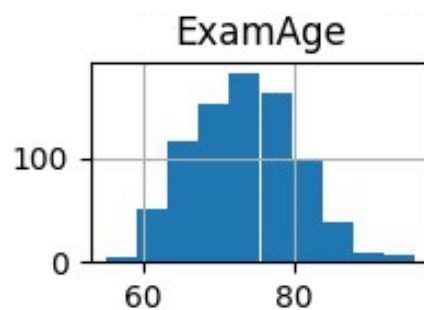
Δημιουργία Στόχου DXTYPE:

Ορίζονται τα χαρακτηριστικά διάγνωσης (DXNORM, DXMCI, DXAD), τα οποία χρησιμοποιούνται για να δημιουργήσουν μια νέα στοχευμένη μεταβλητή (DXTYPE). Αυτή η μεταβλητή παίρνει τιμές:

- 0 για κανονική κατάσταση (DXNORM = 1),
- 1 για ήπια γνωστική εξασθένηση (DXMCI = 1),
- 2 για Alzheimer (DXAD = 1),
- -1 για καμία από τις παραπάνω καταστάσεις.

Υπολογισμός Ηλικίας:

Η ηλικία των ασθενών υπολογίζεται ως η διαφορά του έτους εξέτασης (Examyear) από το έτος γέννησης (PTDOBYear) και φορτώνεται σε ένα νέο Feature , το ExamAge



Πίνακας 42-ExamAge

Καθαρισμός Δεδομένων:

Αφαιρούνται όλες οι περιττές στήλες, συμπεριλαμβανομένων των χαρακτηριστικών διάγνωσης και άλλων πεδίων που δεν είναι απαραίτητα για την περαιτέρω ανάλυση (APTyear, Examyear, PTDOBYear, DXCURREN). Αυτός ο καθαρισμός είναι κρίσιμος για την αποφυγή διαρροής πληροφοριών και την εστίαση στα σημαντικά χαρακτηριστικά για την πρόβλεψη.

Διαχείριση και αντιμετώπιση Απουσιαζόντων Τιμών

Χρησιμοποιείται ο SimpleImputer με την στρατηγική 'median' για την αντικατάσταση των απουσιαζόντων τιμών (NaN) με τη διάμεση τιμή κάθε στήλης. Αυτή η προσέγγιση επιλέγεται συχνά για να διατηρηθεί η κατανομή των δεδομένων χωρίς να επηρεάζεται υπερβολικά από τις έντονες τιμές (outliers).

Ορισμός Χαρακτηριστικών και Ετικετών: Δηλώνεται η λίστα των επιθυμητών χαρακτηριστικών (desired_features) που θα χρησιμοποιηθούν για την εκπαίδευση των μοντέλων. Η ετικέτα (y) που θα προβλεφθεί είναι η DXTYPE, που έχει προηγουμένως δημιουργηθεί.

4.2 Διαχείριση Ανισορροπίας στα Δεδομένα

Για τη διαχείριση της ανισορροπίας στα δεδομένα, εξετάστηκε η χρήση της τεχνικής εξισορρόπησης κλάσεων, με στόχο την ίση αναπαράσταση κάθε κατηγορίας στη μεταβλητή στόχο DXTYPE. Η ανάλυση αρχικά εντόπισε μια σημαντική ανισορροπία στο σύνολο δεδομένων, όπου οι ασθενείς με Alzheimer's Disease (AD) αποτελούσαν το μικρότερο δείγμα, με μόλις 95 εγγραφές. Ως αποτέλεσμα, η αξιολόγηση των μοντέλων πραγματοποιήθηκε χρησιμοποιώντας ίσο αριθμό δειγμάτων από κάθε μία από τις τρεις κλάσεις του DXTYPE, συγκεκριμένα 95 δείγματα ανά κατηγορία.

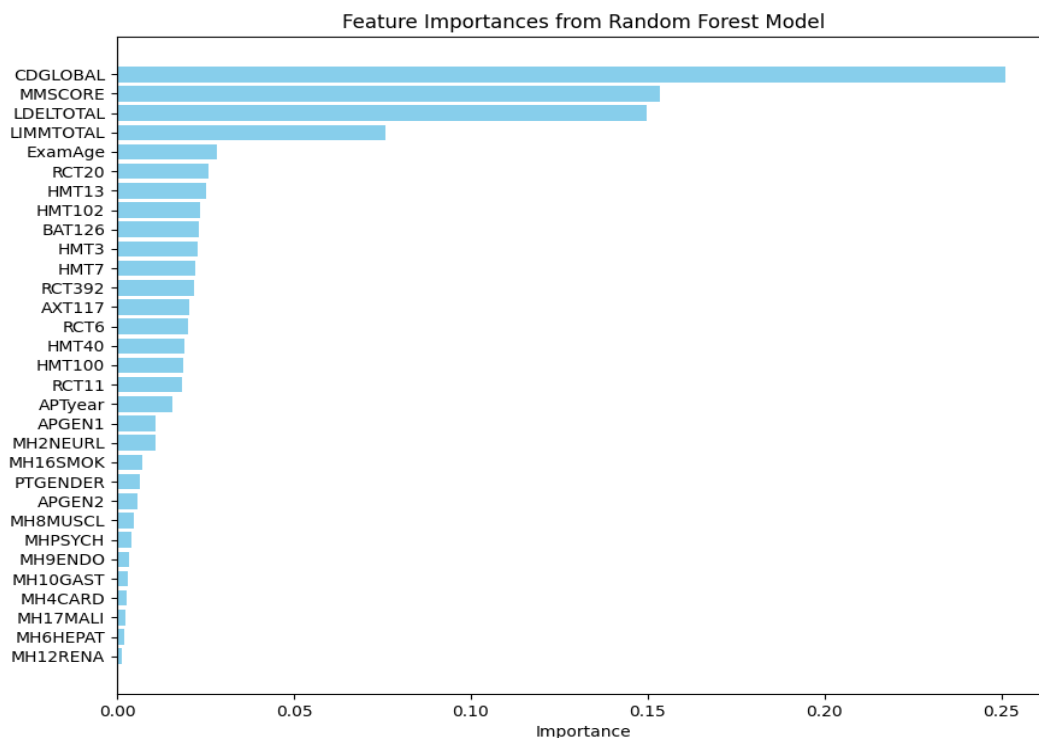
Παρ' όλα αυτά, η εφαρμογή της εξισορρόπησης κλάσεων δεν οδήγησε σε σημαντική βελτίωση των μετρικών απόδοσης των μοντέλων. Αυτό το εύρημα υποδηλώνει ότι, παρά την αρχική προσπάθεια για θεωρητικά πιο ισορροπημένη ανάλυση, η απλή εξισορρόπηση των δειγμάτων δεν είναι απαραίτητα επαρκής για την επίτευξη βελτιωμένων διαγνωστικών προβλέψεων. Έτσι, αποφασίστηκε να προχωρήσει η ανάλυση χωρίς την εφαρμογή της τεχνικής εξισορρόπησης στις τελικές διαδικασίες, αναζητώντας άλλες μεθόδους και προσεγγίσεις που θα μπορούσαν να ενισχύσουν την ακρίβεια και την αξιοπιστία των προβλέψεων.

4.3 Feature Importance

Στην παρούσα έρευνα, εξετάστηκε η επίδραση των διάφορων χαρακτηριστικών στην απόδοση ενός μοντέλου μηχανικής μάθησης για την πρώιμη διάγνωση της νόσου Alzheimer, χρησιμοποιώντας το μοντέλο ταξινόμησης Random Forest. Το σύνολο δεδομένων AIBL υποβλήθηκε σε μια διαδικασία εκκαθάρισης και μετασχηματισμού, με την αφαίρεση απαρχαιωμένων και λιγότερο σημαντικών χαρακτηριστικών. Αφού καθορίστηκαν τα χαρακτηριστικά εισόδου και η μεταβλητή στόχος, δηλαδή το 'DXTYPE', προχωρήσαμε στην εκπαίδευση του μοντέλου.

Η σημασία των χαρακτηριστικών καταγράφηκε μέσω της μεθόδου των feature importances που προσφέρει το Random Forest, η οποία αποκάλυψε τη συνεισφορά κάθε χαρακτηριστικού στην απόδοση του μοντέλου.

Παρ' όλα αυτά, η ανάλυση των αποτελεσμάτων έδειξε ότι η συμβολή της συγκεκριμένης προσέγγισης στη βελτίωση της ακρίβειας του τελικού μοντέλου δεν ήταν σημαντική. Αυτό οδήγησε



Πίνακας 43-Feature Importance Results

στην απόφαση να μην υιοθετηθεί αυτή η τεχνική στην τελική φάση της εφαρμογής μας, καθώς η επίπτωση της στη βελτίωση της απόδοσης ήταν περιορισμένη.

Αξίζει να σημειωθεί όμως ότι η διαδικασία έδωσε σημαντικές διορατικότητες για την επιρροή των επιμέρους χαρακτηριστικών στη διαγνωστική διαδικασία.

4.4 Εκπαίδευση και Βελτιστοποίηση Μοντέλων

Πριν την επιλογή του τελικού μοντέλου, χρησιμοποιήθηκαν πολλαπλά μοντέλα μηχανικής μάθησης. Πιο συγκεκριμένα χρησιμοποιήθηκαν τα εξής μοντέλα:

1. RandomForestClassifier

Περιγραφή: Ο RandomForest είναι ένας αλγόριθμος ensemble που λειτουργεί κατασκευάζοντας πολλά δέντρα αποφάσεων κατά την εκπαίδευση και εξάγοντας τη μέση πρόβλεψη από τα διάφορα δέντρα για να βελτιώσει την προβλεπτική ακρίβεια και να ελέγξει την υπερεκπαίδευση.

Εφαρμογή: Ιδιαίτερα χρήσιμος σε περιπτώσεις με πολλαπλά χαρακτηριστικά, όπως στην περίπτωση διάγνωσης Alzheimer, όπου η ακρίβεια και η ικανότητα ανάλυσης σύνθετων σχέσεων είναι κρίσιμη.

2. GradientBoostingClassifier

Περιγραφή: Ο Gradient Boosting δουλεύει κατασκευάζοντας δέντρα σε σειρά, όπου κάθε νέο δέντρο βοηθά στην διόρθωση των λαθών που έγιναν από τα προηγούμενα στο σύνολο δεδομένων.

Εφαρμογή: Εξαιρετικά αποτελεσματικός σε δεδομένα όπου οι σχέσεις είναι περίπλοκες και η συμβολή του κάθε χαρακτηριστικού είναι διαφορετική, κάτι που είναι συχνό σε ιατρικά δεδομένα.

3. AdaBoostClassifier

Περιγραφή: Ο AdaBoost (Adaptive Boosting) είναι μια άλλη τεχνική ensemble που στοχεύει να βελτιώσει την απόδοση των αλγορίθμων μηχανικής μάθησης, ενισχύοντας τα 'αδύναμα' μοντέλα για να δημιουργήσει ένα ισχυρότερο τελικό μοντέλο.

Εφαρμογή: Χρησιμοποιείται για την βελτίωση της ακρίβειας στη διάγνωση, ακόμη και όταν τα αρχικά μοντέλα δεν είναι τόσο αποτελεσματικά.

4. HistGradientBoostingClassifier

Περιγραφή: Αυτός ο αλγόριθμος είναι μια βελτίωση του τυπικού Gradient Boosting και είναι κατασκευασμένος για να είναι πιο αποτελεσματικός σε μεγάλα δεδομένα, με τη χρήση ιστογραμμάτων για την αποδοτικότερη βελτιστοποίηση.

Εφαρμογή: Ιδανικός για την επεξεργασία των πολύπλοκων και μεγάλων σετ δεδομένων στην ιατρική έρευνα.

5. KNeighborsClassifier (KNN)

Περιγραφή: Ο KNN είναι ένας αλγόριθμος βασισμένος στην απόσταση μεταξύ των δειγμάτων και προβλέπει την κατηγορία ενός δείγματος με βάση τις πιο κοντινές του γειτονιές.

Εφαρμογή: Ευρέως χρησιμοποιείται στην ιατρική διάγνωση ως μέθοδος πρόβλεψης λόγω της απλότητας και αποτελεσματικότητας του.

6. Logistic Regression

Περιγραφή: Αποτελεί μια βασική στατιστική τεχνική που προβλέπει την πιθανότητα εμφάνισης ενός γεγονότος με βάση δεδομένα εισόδου.

Εφαρμογή: Χρησιμοποιείται για την αξιολόγηση της πιθανότητας ένας ασθενής να έχει Alzheimer με βάση διάφορα χαρακτηριστικά.

7. Support Vector Machine (SVM)

Περιγραφή: Ο SVM είναι ένας δυναμικός αλγόριθμος ταξινόμησης που χρησιμοποιεί την έννοια των υπερεπιπέδων για να καθορίσει την απόσταση μεταξύ των κατηγοριών δεδομένων.

Εφαρμογή: Εξαιρετικά αποτελεσματικός για δεδομένα με υψηλή διάσταση και περίπλοκες διακρίσεις μεταξύ των κατηγοριών, όπως είναι τα ιατρικά δεδομένα.

8. XGBClassifier

Περιγραφή: Πρόκειται για μια βελτιστοποιημένη διανομή του Gradient Boosting Framework που είναι ιδιαίτερα αποδοτική, ευέλικτη και φιλική προς τον προγραμματιστή.

Εφαρμογή: Χρησιμοποιείται ευρέως σε προβληματικές προβλέψεις στον τομέα της υγείας, προσφέροντας ταχύτητα και ακρίβεια.

Η χρήση αυτών των αλγορίθμων και η αντίστοιχη βελτιστοποίηση των παραμέτρων τους επιτρέπει τη δημιουργία αξιόπιστων και αποτελεσματικών μοντέλων που μπορούν να διαχειριστούν την πολυπλοκότητα και τις προκλήσεις της διάγνωσης της νόσου Alzheimer.

Για κάθε μοντέλο, πραγματοποιήθηκαν τέσσερις δοκιμές:

1. με πλήρη δεδομένα
2. με πλήρη δεδομένα και παραμετροποίηση (hypertuning)
3. με ανάλυση σημαντικότητας χαρακτηριστικών(Feature Importance)
4. και μια συνδυαστική των δύο τελευταίων, δηλαδή με Feature Importance και hypertuning

Η βελτιστοποίηση των παραμέτρων (hypertuning) επιτεύχθηκε μέσω του StratifiedKFold και cross_val_score, ενώ η αξιολόγηση των μοντέλων βασίστηκε σε μετρήσεις όπως το precision, recall, το accuracy_score και το f1_score.

4.5 Παράμετροι μοντέλων

Hyperparameter Tuning

Ο όρος "hyperparameter tuning" αναφέρεται στη διαδικασία επιλογής των ιδανικών παραμέτρων για ένα μοντέλο μηχανικής μάθησης. Αντίθετα με τις παραμέτρους του μοντέλου, που μαθαίνονται αυτόματα κατά τη διάρκεια της εκπαίδευσης, οι υπερπαραμέτροι πρέπει να οριστούν πριν από την εκπαίδευση του μοντέλου.

Οι υπερπαραμέτροι επηρεάζουν άμεσα τη συμπεριφορά των αλγορίθμων κατά την εκπαίδευση και μπορούν να έχουν σημαντικό αντίκτυπο στην απόδοση του τελικού μοντέλου. Η

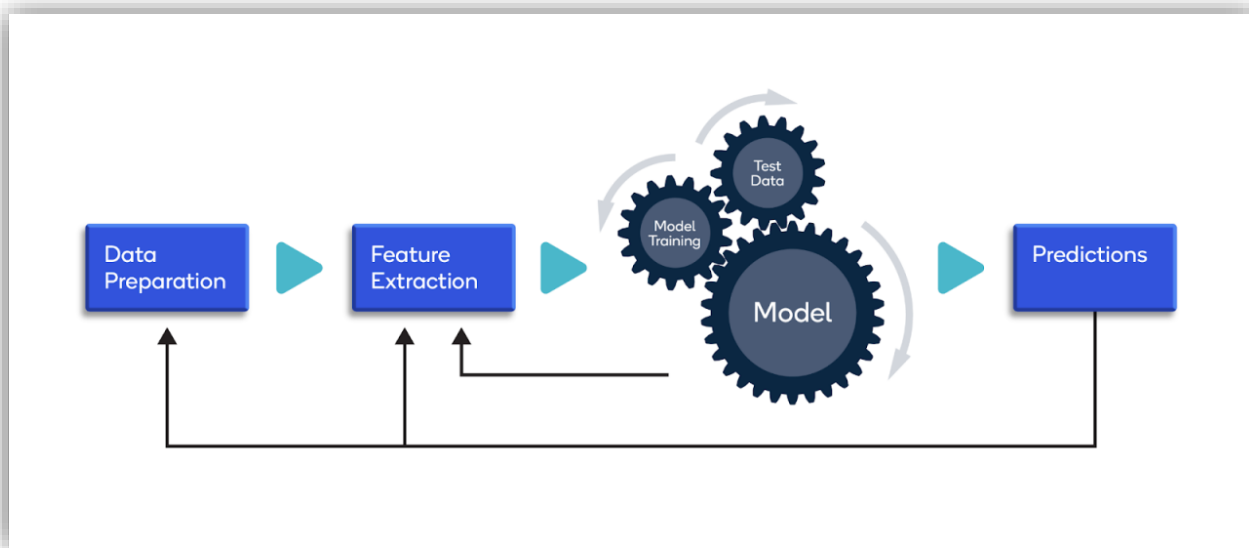
βελτιστοποίηση αυτών των υπερπαραμέτρων μπορεί να οδηγήσει σε πιο αποτελεσματικά και ακριβή μοντέλα.

GridSearchCV

Το GridSearchCV είναι μια τεχνική που χρησιμοποιείται για τη συστηματική εξερεύνηση πολλαπλών συνδυασμών υπερπαραμέτρων, αυτοματοποιώντας τη διαδικασία της βελτιστοποίησης. Ουσιαστικά, ορίζετε ένα "πλέγμα" από διαφορετικές υπερπαραμέτρους και το GridSearchCV δοκιμάζει όλους τους δυνατούς συνδυασμούς σε αυτό το πλέγμα.

Για κάθε συνδυασμό, το GridSearchCV εκπαιδεύει ένα μοντέλο στο σετ εκπαίδευσης και το αξιολογεί χρησιμοποιώντας διασταυρούμενη επικύρωση (cross-validation). Η απόδοση κάθε μοντέλου αξιολογείται με βάση μια συγκεκριμένη μετρική, συνήθως την ακρίβεια ή άλλες σχετικές μετρικές. Τελικά, το GridSearchCV επιλέγει τις υπερπαραμέτρους που οδήγησαν στην καλύτερη απόδοση κατά τη διαδικασία της διασταυρούμενης επικύρωσης.

Στην περίπτωσης σας, χρησιμοποιήθηκε το GridSearchCV για τη βελτιστοποίηση των αλγορίθμων μηχανικής μάθησης όπως RandomForest, Gradient Boosting, AdaBoost, και άλλοι, με διάφορες υπερπαραμέτρους όπως `n_estimators`, `max_depth`, `learning_rate`, κ.ά. Αυτό βοήθησε στην εύρεση του ιδανικού μοντέλου για την πρόβλεψη του DXTYPE από το σύνολο δεδομένων AIBL.



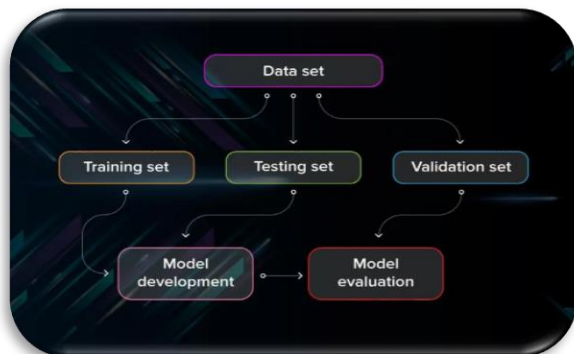
Εικόνα 6-Διαδικασία Υλοποίησης

5.ΑΠΟΤΕΛΕΣΜΑΤΑ

Αποτελέσματα αλγορίθμων , επιλογή τελικού μοντέλου και ενέργειες βελτιστοποίησης .

5.1 Αποτελέσματα αλγορίθμων με 10 fold cross-validation.

Πριν προχωρήσουμε στη στατιστική ανάλυση των αποτελεσμάτων των αλγορίθμων που χρησιμοποιήθηκαν στην παρούσα πτυχιακή εργασία, θα πρέπει να επισημάνουμε τη σημασία της αξιολόγησης των μοντέλων μηχανικής μάθησης στη διαδικασία της πρώιμης ανίχνευσης νόσων. Η αξιοπιστία και η αποδοτικότητα των προγνωστικών μοντέλων εξαρτώνται άμεσα από την ορθότητα των αποτελεσμάτων τους, καθιστώντας έτσι την στατιστική ανάλυση κεντρικό στοιχείο της ερευνητικής διαδικασίας.



Εικόνα 7-Διαδικασία Εύρεσης Βέλτιστου Μοντέλου

Στο πλαίσιο αυτό, η στατιστική ανάλυση που θα ακολουθήσει αποσκοπεί στο να αξιολογήσει την αποτελεσματικότητα κάθε αλγορίθμου, να εντοπίσει τα δυνατά και τα αδύναμα σημεία τους και να παρέχει μια ολοκληρωμένη εικόνα για τις διαφορές στην απόδοσή τους. Η ανάλυση αυτή θα βασιστεί σε διάφορες στατιστικές μετρήσεις όπως η μέση ακρίβεια, η τυπική απόκλιση, ο χρόνος διασταυρούμενης επικύρωσης, καθώς και μετρικές σχετικές με την ακρίβεια, την ανάκληση και το σκορ F1 για κάθε κατηγορία της πρόβλεψης.

Με τη συστηματική αυτή προσέγγιση, ελπίζουμε να κατανοήσουμε καλύτερα τις δυνατότητες και τις περιορισμούς των μοντέλων που έχουν αναπτυχθεί, να προσδιορίσουμε ποιες παράμετροι επηρεάζουν περισσότερο την απόδοσή τους, και να προτείνουμε βελτιώσεις για μελλοντικές εφαρμογές στον τομέα της πρώιμης διάγνωσης της νόσου Alzheimer.

Όπως προαναφέρθηκε στο προηγούμενο κεφάλαιο, ο κάθε αλγόριθμος εκπαιδεύτηκε και δοκιμάστηκε σε 4 διαφορετικές παραλλαγές του Dataset.

1. Simple Dataset→ απλό data clearance
2. Feature Imp(15) → data clearance + feature Importance με τα 15 πιο σημαντικά features
3. Hyperparameter Tuned→ data clearance + hyperparameter Tuning
4. Hyper Tuned with Feature Imp(15)→ data clearance + hyperparameter Tuning+ feature Importance

Παρακάτω θα γίνει στατιστική απεικόνιση των αποτελεσμάτων για κάθε πείραμα που έγινε πριν την επιλογή του τελικού μοντέλου.

5.1.1 Support Vector Machine Results.

Βάσει των δεδομένων του πίνακα, η καλύτερη απόδοση του αλγορίθμου Support Vector Machine παρατηρείται στη συνθήκη με την εφαρμογή της υπερπαραμετροποίησης και τη σημαντικότητα

των 15 χαρακτηριστικών (Hyper Tuned with Feature Imp(15)). Συγκεκριμένα, σε αυτή την ρύθμιση εμφανίζεται υψηλή μέση ακρίβεια 0.912, σταθερή τυπική απόκλιση 0.024, και άριστες επιδόσεις σε όλες τις μετρήσεις αξιολόγησης (Precision, Recall, F1 Score) με τιμές πάνω από 0.84. Αυτό υποδεικνύει ότι η συνδυασμένη ρύθμιση προσφέρει τα βέλτιστα αποτελέσματα για το συγκεκριμένο μοντέλο.

Support Vector Machine	Simple Dataset	Feature Imp(15)	Hyperparameter Tuned	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.887	0.912	0.884	0.912
Standard Deviation	0.010	0.024	0.028	0.024
C-V Time (seconds)	0.2290	0.1860	0.2500	0.1810
Precision	0.800	0.839	0.831	0.839
Recall	0.817	0.848	0.835	0.848
F1 Score	0.805	0.842	0.833	0.842
Classification Report precision- recall-f1- score	0 0.88 0.96 0.92 1 0.58 0.44 0.50 2 0.71 0.62 0.67	0 0.90 0.96 0.93 1 0.69 0.59 0.63 2 0.71 0.62 0.67	0 0.91 0.93 0.92 1 0.62 0.59 0.61 2 0.73 0.69 0.71	0 0.90 0.96 0.93 1 0.69 0.59 0.63 2 0.71 0.62 0.67
Class Accuracy	0.820	0.850	0.840	0.850
Macro Avg Precision- Recall-F1	(0) 0.72 (1) 0.67 (2) 0.69	(0) 0.77 (1) 0.72 (2) 0.74	(0) 0.75 (1) 0.74 (2) 0.74	(0) 0.77 (1) 0.72 (2) 0.74
Weighted Avg Precision-Recall-F1	(0) 0.80 (1) 0.82 (2) 0.81	(0) 0.84 (1) 0.85 (2) 0.84	(0) 0.83 (1) 0.84 (2) 0.83	(0) 0.84 (1) 0.85 (2) 0.84

Πίνακας 44-SVM Results

5.1.2 Logistic Regression Results.

Η ανάλυση των αποτελεσμάτων για τον αλγόριθμο Logistic Regression δείχνει ότι οι καλύτερες επιδόσεις επιτυγχάνονται στα σετ δεδομένων με υπερπαραμετροποίηση και όταν συνδυάζεται με τη σημαντικότητα των 15 χαρακτηριστικών. Συγκεκριμένα, η μέση ακρίβεια, η ακρίβεια, το recall και το F1 Score είναι υψηλότερα σε αυτές τις ρυθμίσεις, υποδεικνύοντας ότι αυτές οι παράμετροι βελτιστοποιούν την απόδοση του μοντέλου.

Logistic Regression	Simple Dataset	Feature Imp(15)	Hyperparameter Tuned	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.902	0.916	0.899	0.916
Standard Deviation	0.036	0.035	0.032	0.035
C-V Time (seconds)	0.1710	0.1100	0.1070	0.1080
Precision	0.904	0.879	0.853	0.879
Recall	0.902	0.878	0.860	0.878
F1 Score	0.902	0.878	0.855	0.878
Classification Report precision- recall-f1- score	0 0.95 0.96 0.95 1 0.75 0.79 0.77 2 0.92 0.75 0.83	0 0.95 0.95 0.95 1 0.69 0.74 0.71 2 0.79 0.69 0.73	0 0.92 0.96 0.94 1 0.69 0.59 0.63 2 0.75 0.75 0.75	0 0.95 0.95 0.95 1 0.69 0.74 0.71 2 0.79 0.69 0.73
Class Accuracy	0.900	0.880	0.860	0.880
Macro Avg Precision- Recall-F1	(0) 0.87 (1) 0.83 (2) 0.85	(0) 0.81 (1) 0.79 (2) 0.80	(0) 0.79 (1) 0.76 (2) 0.77	(0) 0.81 (1) 0.79 (2) 0.80
Weighted Avg Precision-Recall-F1	(0) 0.90 (1) 0.90 (2) 0.90	(0) 0.88 (1) 0.88 (2) 0.88	(0) 0.85 (1) 0.86 (2) 0.86	(0) 0.88 (1) 0.88 (2) 0.88

Πίνακας 45-Logistic Regression Results

5.1.2 KNN Results.

Στα δεδομένα που ακολουθούν παρουσιάζονται λεπτομερώς τα αποτελέσματα του αλγορίθμου KNN, τα οποία δείχνουν σαφώς κατώτερη απόδοση σε σχέση με τους υπόλοιπους αλγορίθμους, ακόμα και μετά την τροποποίηση της παραμέτρου -k-.

KNN	Simple Dataset	Feature Imp(15)	Hyperparameter Tuned	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.802	0.864	0.805	0.864
Standard Deviation	0.020	0.036	0.022	0.028
C-V Time (seconds)	0.2900	0.1030	0.3230	0.1090
Precision	0.768	0.798	0.720	0.739
Recall	0.793	0.805	0.768	0.762
F1 Score	0.762	0.787	0.719	0.740
Classification Report precision- recall-f1- score	0 0.82 0.98 0.89 1 0.59 0.29 0.39 2 0.80 0.50 0.62	0 0.83 0.96 0.89 1 0.58 0.41 0.48 2 1.00 0.50 0.67	0 0.79 0.98 0.88 1 0.45 0.15 0.22 2 0.75 0.56 0.64	0 0.82 0.94 0.87 1 0.46 0.32 0.38 2 0.78 0.44 0.56
Class Accuracy	0.790	0.800	0.770	0.760
Macro Avg Precision- Recall-F1	(0) 0.74 (1) 0.59 (2) 0.63	(0) 0.81 (1) 0.63 (2) 0.68	(0) 0.67 (1) 0.56 (2) 0.58	(0) 0.68 (1) 0.57 (2) 0.60
Weighted Avg Precision-Recall-F1	(0) 0.77 (1) 0.79 (2) 0.76	(0) 0.80 (1) 0.80 (2) 0.79	(0) 0.72 (1) 0.77 (2) 0.72	(0) 0.74 (1) 0.76 (2) 0.74

Πίνακας 46-KNN Results

5.1.3 HistGradient Boosting Results.

Από τον παραπάνω πίνακα παρατηρείται ότι ο αλγόριθμος HistGradient Boosting διατηρεί συνεπή αποτελέσματα μέσης ακρίβειας και F1 σκορ σε όλες τις διαμορφώσεις του, με ελάχιστες διακυμάνσεις. Η σταθερότητα των επιδόσεων του αποτυπώνεται στην ομοιομορφία των μέσων

ακριβειών και των F1 σκορ, ενώ το χρονικό πλαίσιο επικύρωσης διακυμαίνεται, δείχνοντας ότι οι πιο απαιτητικές ρυθμίσεις έχουν επιβραδύνει το μοντέλο. Η μικρή μεταβολή στις επιδόσεις από την προσθήκη της σημασίας χαρακτηριστικών και τη ρύθμιση των υπερπαραμέτρων υποδεικνύει ότι ο βασικός αλγόριθμος παρέχει ήδη ισχυρή βάση για ανίχνευση και διάκριση στην πρόβλεψη.

HistGradient Boosting	Simple Dataset	Feature Imp(15)	Hyperparameter Tuned	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.921	0.915	0.921	0.915
Standard Deviation	0.044	0.041	0.044	0.042
C-V Time(seconds)	24.0040	19.5190	22.9050	9.4720
Precision	0.857	0.849	0.857	0.857
Recall	0.860	0.854	0.860	0.860
F1 Score	0.858	0.851	0.858	0.858
Classification Report precision- recall-f1- score	0 0.93 0.95 0.94 1 0.67 0.65 0.66 2 0.73 0.69 0.71	0 0.92 0.95 0.94 1 0.66 0.62 0.64 2 0.73 0.69 0.71	0 0.93 0.95 0.94 1 0.67 0.65 0.66 2 0.73 0.69 0.71	0 0.93 0.95 0.94 1 0.67 0.65 0.66 2 0.73 0.69 0.71
Class Accuracy	0.860	0.850	0.860	0.860
Macro Avg Precision-Recall-F1	(0) 0.78 (1) 0.76 (2) 0.77	(0) 0.77 (1) 0.75 (2) 0.76	(0) 0.78 (1) 0.76 (2) 0.77	(0) 0.78 (1) 0.76 (2) 0.77
Weighted Avg Precision-Recall-F1	(0) 0.86 (1) 0.86 (2) 0.86	(0) 0.85 (1) 0.85 (2) 0.85	(0) 0.86 (1) 0.86 (2) 0.86	(0) 0.86 (1) 0.86 (2) 0.86

Πίνακας 407-Hist Gradient Boosting Results

5.1.4 Gradient Boosting Results.

Ο παραπάνω πίνακας παρουσιάζει τις αποδόσεις του αλγορίθμου Gradient Boosting σε διάφορες διαμορφώσεις: βασική συνθήκη, με ενσωμάτωση σημασίας χαρακτηριστικών, με ρυθμισμένες υπερπαραμέτρους και με την ταυτόχρονη ρύθμιση υπερπαραμέτρων και σημασίας χαρακτηριστικών. Η μέση ακρίβεια διατηρείται σταθερά υψηλή σε όλες τις διαμορφώσεις, δείχνοντας ελάχιστες διακυμάνσεις. Ειδικότερα, η εφαρμογή των υπερπαραμέτρων και της σημασίας χαρακτηριστικών δεν φαίνεται να προσφέρει σημαντική βελτίωση στην απόδοση σύμφωνα με το F1 σκορ, το οποίο παραμένει σχεδόν σταθερό. Αυτό ενδέχεται να υποδηλώνει ότι ο βασικός αλγόριθμος είναι ήδη αρκετά αποτελεσματικός για το είδος των δεδομένων που εξετάζονται.

Gradient Boosting	Simple Dataset	Feature Imp(15)	Hyperparameter Tuned	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.919	0.916	0.918	0.916
Standard Deviation	0.032	0.041	0.040	0.043
C-V Time (seconds)	8.2450	7.9200	4.1390	10.8137
Precision	0.864	0.872	0.840	0.872
Recall	0.860	0.866	0.841	0.872
F1 Score	0.861	0.868	0.841	0.872
Classification Report precision- recall-f1- score	0 0.95 0.93 0.94 1 0.65 0.71 0.68 2 0.73 0.69 0.71	0 0.95 0.93 0.94 1 0.66 0.74 0.69 2 0.73 0.69 0.71	0 0.92 0.93 0.93 1 0.64 0.62 0.63 2 0.69 0.69 0.69	0 0.95 0.95 0.95 1 0.69 0.71 0.70 2 0.73 0.69 0.71
Class Accuracy	0.860	0.870	0.840	0.870
Macro Average Precision-Recall-F1	(0) 0.78 (1) 0.77 (2) 0.77	(0) 0.78 (1) 0.78 (2) 0.78	(0) 0.75 (1) 0.74 (2) 0.75	(0) 0.79 (1) 0.78 (2) 0.78
Weighted Average Precision-Recall-F1	(0) 0.86 (1) 0.86 (2) 0.86	(0) 0.87 (1) 0.87 (2) 0.87	(0) 0.84 (1) 0.84 (2) 0.84	(0) 0.87 (1) 0.87 (2) 0.87

Πίνακας 48-Gradient Boosting Result

5.1.5 Ada Boost Results.

Ο παραπάνω πίνακας παρουσιάζει τα αποτελέσματα του αλγορίθμου AdaBoost για τη διάγνωση της νόσου Alzheimer υπό διάφορες συνθήκες. Η βέλτιστη απόδοση επιτεύχθηκε με τη ρύθμιση Hyperparameter Tuned Dataset, όπου ο αλγόριθμος επιδεικνύει υψηλότερη μέση ακρίβεια και βελτιωμένες μετρήσεις ακρίβειας, ανάκλησης και F1 σκορ. Αυτό υπογραμμίζει την αποτελεσματικότητα της διαδικασίας ρύθμισης υπερπαραμέτρων για τη βελτίωση της απόδοσης μοντέλων σε συγκεκριμένες εφαρμογές.

Ada Boost	Simple Dataset	Feature Imp (15)	Hyperparameter Tuned Dataset	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.898	0.889	0.919	0.913
Standard Deviation	0.025	0.039	0.030	0.039
C-V Time (seconds)	1.8620	1.6640	1.9480	3.3496
Precision	0.855	0.855	0.887	0.855
Recall	0.854	0.854	0.878	0.854
F1 Score	0.854	0.854	0.879	0.855
Classification Report precision- recall-f1- score	0 0.93 0.93 0.93 1 0.64 0.68 0.66 2 0.79 0.69 0.73	0 0.93 0.93 0.93 1 0.64 0.68 0.66 2 0.79 0.69 0.73	0 0.95 0.94 0.94 1 0.68 0.79 0.73 2 0.91 0.62 0.74	0 0.95 0.94 0.94 1 0.66 0.68 0.67 2 0.62 0.62 0.62
Class Accuracy	0.850	0.850	0.880	0.850
Macro Avg Precision-Recall-F1	(0) 0.78 (1) 0.76 (2) 0.77	(0) 0.78 (1) 0.76 (2) 0.77	(0) 0.84 (1) 0.79 (2) 0.80	(0) 0.74 (1) 0.75 (2) 0.74
Weighted Avg Precision-Recall-F1	(0) 0.86 (1) 0.85 (2) 0.86	(0) 0.86 (1) 0.85 (2) 0.85	(0) 0.89 (1) 0.88 (2) 0.88	(0) 0.86 (1) 0.85 (2) 0.85

Πίνακας 49-Ada Boost Results

5.1.5 Random Forest Results.

Ο αλγόριθμος Random Forest καταγράφει σταθερή μέση ακρίβεια που κυμαίνεται μεταξύ 0.915 και 0.918, δείχνοντας ότι οι βελτιώσεις υπερπαραμέτρων και η επιλογή χαρακτηριστικών επηρεάζουν ελαφρώς την απόδοση.

Random Forest	Simple dataset	Feature Imp(15)	Hyperparameter Tuned Dataset	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.915	0.915	0.918	0.916
Standard Deviation	0.026	0.034	0.031	0.036
C-V Time(seconds)	2.1020 sec	2.0770 sec	2.0850 sec	3.1500 sec
Precision	0.855	0.836	0.848	0.838
Recall	0.860	0.841	0.854	0.841
F1 Score	0.857	0.839	0.850	0.839
Classification Report precision- recall-f1- score	0 0.93 0.96 0.94 1 0.67 0.65 0.66 2 0.71 0.62 0.67	0 0.92 0.95 0.94 1 0.62 0.59 0.61 2 0.67 0.62 0.65	0 0.92 0.96 0.94 1 0.66 0.62 0.64 2 0.71 0.62 0.67	0 0.92 0.94 0.93 1 0.62 0.59 0.61 2 0.69 0.69 0.69
Class Accuracy	0.860	0.840	0.850	0.840
Macro Avg Precision-Recall-F1	(0) 0.77 (1) 0.74 (2) 0.76	(0) 0.74 (1) 0.72 (2) 0.73	(0) 0.76 (1) 0.73 (2) 0.75	(0) 0.74 (1) 0.74 (2) 0.74
Weighted Avg Precision-Recall-F1	(0) 0.86 (1) 0.86 (2) 0.86	(0) 0.84 (1) 0.84 (2) 0.84	(0) 0.85 (1) 0.85 (2) 0.85	(0) 0.84 (1) 0.84 (2) 0.84

Πίνακας 50-Random Forest Results

5.1.5 XGBOOST Results.

Ο αλγόριθμος XGBoost ξεχώρισε για τη σταθερή απόδοσή του σε διάφορες διαμορφώσεις, καθιστώντας τον ιδανική επιλογή για την εφαρμογή που αναπτύχθηκε για την πρόωμη διάγνωση της νόσου Alzheimer. Με μέση ακρίβεια που φτάνει το 0.915 και συνεπείς επιδόσεις στις μετρήσεις ακρίβειας, ανάκλησης και F1 σκορ, ο XGBoost αποδεικνύεται αξιόπιστος στην ανάλυση σύνθετων δεδομένων, ενισχύοντας τη διαγνωστική διαδικασία της εφαρμογής.

XGBoost	Simple Dataset	Feature Imp(15)	Hyperparameter Tuned	Hyper Tuned with Feature Imp(15)
Mean Accuracy	0.910	0.905	0.915	0.910
Standard Deviation	0.045	0.046	0.038	0.046
C-V Time (seconds)	1.4010	1.5620	1.2890	1.1050
Precision	0.862	0.845	0.870	0.857
Recall	0.866	0.848	0.872	0.860
F1 Score	0.863	0.846	0.871	0.858
Classification Report precision- recall-f1- score	0 0.93 0.96 0.94 1 0.70 0.62 0.66 2 0.71 0.75 0.73	0 0.92 0.94 0.93 1 0.64 0.62 0.63 2 0.73 0.69 0.71	0 0.94 0.95 0.94 1 0.71 0.65 0.68 2 0.72 0.81 0.76	0 0.93 0.95 0.94 10.67 0.65 0.66 2 0.73 0.69 0.71
Class Accuracy	0.870	0.850	0.870	0.860
Macro Avg Precision- Recall-F1	(0) 0.78 (1) 0.77 (2) 0.78	(0) 0.76 (1) 0.75 (2) 0.76	(0) 0.79 (1) 0.80 (2) 0.79	(0) 0.78 (1) 0.76 (2) 0.77
Weighted Avg Precision-Recall-F1	(0) 0.86 (1) 0.87 (2) 0.86	(0) 0.84 (1) 0.85 (2) 0.85	(0) 0.87 (1) 0.87 (2) 0.87	(0) 0.86 (1) 0.86 (2) 0.86

Πίνακας 41-XGBoost Results

5.2 Επιλογή Μοντέλου.

Με βάση τα αποτελέσματα που αναλύονται στο προηγούμενο κεφάλαιο, καθώς και τις βιβλιογραφικές ανασκοπήσεις των σχετικών άρθρων στο δεύτερο κεφάλαιο, η επιλογή που προέκυψε για την ανάπτυξη του προγράμματος ήταν το μοντέλο XGBoost με εφαρμογή τεχνικών HyperParameter Tuning. Αν και δεν κατέγραψε την υψηλότερη ακρίβεια σε όλες τις δοκιμές, αποδείχθηκε ένα σταθερό και αξιόπιστο μοντέλο με εξαιρετικές προοπτικές για βελτίωση.

Στατιστική ανάλυση του επιλεγμένου μοντέλου:

Μέση Ακρίβεια (Mean Accuracy): Το μοντέλο επιτυγχάνει μέση ακρίβεια 91.5%, δείχνοντας υψηλή απόδοση στην κατηγοριοποίηση των κλάσεων.

Τυπική Απόκλιση (Standard Deviation): Η σχετικά χαμηλή τιμή του 0.038 υποδηλώνει σταθερότητα στις προβλέψεις του μοντέλου μέσω διαφόρων δοκιμών.

Χρόνος Επικύρωσης Σταυρού (Cross-Validation Time): Ο χρόνος 1.289 δευτερόλεπτα αντικατοπτρίζει την ευκολία και την ταχύτητα στη διαδικασία της σταυροειδούς επικύρωσης.

Οι βασικές μετρικές του μοντέλου δείχνουν τα εξής:

Ακρίβεια (Precision): 87.0%, που δείχνει την ικανότητα του μοντέλου να κατατάσσει σωστά τις θετικές περιπτώσεις.

Ανάκληση (Recall): 87.2%, αντιπροσωπεύει το ποσοστό των σωστά κατηγοριοποιημένων θετικών περιπτώσεων από το σύνολο των θετικών.

F1 Score: 87.1%, ο οποίος συνδυάζει την ακρίβεια και την ανάκληση, δείχνοντας μια ισορροπημένη απόδοση μεταξύ αυτών των δύο μετρήσεων.

Η ανάλυση της αναφοράς κατηγοριοποίησης αποκαλύπτει τις εξής λεπτομερείς μετρήσεις:

Κανονική κατάσταση (Normal): Ακρίβεια 94%, ανάκληση 95%, F1 score 94%.

Ήπια Γνωστική Διαταραχή (MCI): Ακρίβεια 71%, ανάκληση 65%, F1 score 68%.

Αλτσχάιμερ (AD): Ακρίβεια 72%, ανάκληση 81%, F1 score 76%.

Τα στατιστικά αυτά επιβεβαιώνουν την επιλογή του XGBoost με εφαρμογή Hyperparameter Tuning ως προτιμώμενη μέθοδο για το πρόγραμμα διάγνωσης της νόσου Alzheimer, δείχνοντας σταθερότητα και αποτελεσματικότητα στις προβλέψεις, καθώς και τη δυνατότητα για σημαντικές βελτιώσεις στο μέλλον.

5.3 Βελτιστοποίηση τελικού Μοντέλου.

Η επιλογή του XGBoost ως το βασικό μοντέλο για την ανίχνευση και διάγνωση της νόσου Alzheimer βασίστηκε σε μια σειρά από κριτικά σημεία που επιβεβαιώθηκαν μέσω εκτενούς βιβλιογραφικής ανασκόπησης και πειραματικής ανάλυσης. Ο αλγόριθμος XGBoost έχει αποδειχθεί ιδιαίτερα αποτελεσματικός στο πεδίο της μηχανικής μάθησης για την επεξεργασία μεγάλων δεδομένων και πολύπλοκων προτύπων, παρέχοντας σημαντικά υψηλή ακρίβεια και ταχύτητα σε συνδυασμό με τη δυνατότητα αντιμετώπισης overfitting μέσω των προηγμένων τεχνικών του για τη βελτιστοποίηση.

Για τη βελτίωση του μοντέλου, εφαρμόστηκαν αρκετές τεχνικές βελτιστοποίησης των υπερπαραμέτρων και της προεπεξεργασίας των δεδομένων. Συγκεκριμένα, χρησιμοποιήθηκε ο αλγόριθμος GridSearchCV για την εξεύρεση της ιδανικής τιμής για τις υπερπαραμέτρους όπως το learning rate, το max_depth και το n_estimators, βελτιώνοντας έτσι την απόδοση του μοντέλου σε διαφορετικά σετ δεδομένων.

Μετά τη διεξαγωγή εκτενών δοκιμών μέσω του εργαλείου GridSearchCV, οι υπερπαραμέτροι που κρίθηκαν ιδανικές για τη βελτίωση της απόδοσης του μοντέλου XGBoost ήταν ένα learning rate του 0.24, max_depth ίσο με 3 και n_estimators στις 95. Αυτές οι τιμές επιλέχθηκαν για να μεγιστοποιήσουν την ακρίβεια και την αποδοτικότητα του μοντέλου, μειώνοντας παράλληλα τον κίνδυνο υπερεκπαίδευσης.

Όσον αφορά την προετοιμασία του συνόλου δεδομένων για την εκπαίδευση, αποφασίστηκε να χρησιμοποιηθεί το 40% του διαθέσιμου dataset για το training set αντί του 20% που είχε αρχικά εξεταστεί. Αυτό επιλέχθηκε για να εξασφαλιστεί ότι το μοντέλο θα είχε πρόσβαση σε μια μεγαλύτερη ποσότητα δεδομένων, ιδιαίτερα σημαντική για τις κατηγορίες με λιγότερα δείγματα όπως το MCI και το AD. Η αύξηση του όγκου των δεδομένων προς εκπαίδευση βοήθησε στην παραγωγή πιο ισορροπημένων και αντιπροσωπευτικών μοντέλων, ενισχύοντας την γενική απόδοση και την ικανότητα γενίκευσης του μοντέλου σε μη εκπαιδευμένα δεδομένα.

Πέρα από τη ρύθμιση των παραμέτρων, εφαρμόστηκαν τεχνικές προεπεξεργασίας όπως η κανονικοποίηση των δεδομένων με τη χρήση StandardScaler για την αποφυγή παρεμβλητών διακυμάνσεων στις τιμές των χαρακτηριστικών και η κωδικοποίηση των κατηγορικών μεταβλητών με LabelEncoder. Επιπρόσθετα, χρησιμοποιήθηκε ο SimpleImputer για την αντιμετώπιση τυχόν απουσιάζουσων τιμών μέσω της χρήσης της μέσης τιμής των χαρακτηριστικών, βελτιώνοντας τη συνολική ποιότητα του εκπαιδευτικού σετ.

Το AUC score, ή Area Under the Curve, είναι μία σημαντική μετρική στην αξιολόγηση της απόδοσης διαφόρων μοντέλων μηχανικής μάθησης, ιδιαίτερα σε ταξινομητικές εφαρμογές. Συγκεκριμένα, το AUC score αντιπροσωπεύει την πιθανότητα ένα τυχαία επιλεγμένο θετικό δείγμα (π.χ., ένα δείγμα που προσδιορίζεται ως Alzheimer) να έχει μια υψηλότερη προβλεπόμενη τιμή από ένα τυχαία επιλεγμένο αρνητικό δείγμα. Ένα AUC score που προσεγγίζει το 1 υποδεικνύει ένα μοντέλο με εξαιρετική διακριτική ικανότητα, ενώ ένα score περίπου 0.5 υποδηλώνει έλλειψη διακριτικής ικανότητας, δηλαδή τυχαία επιλογή.

Στην περίπτωση του XGBoost μετά την εφαρμογή του Hyperparameter Tuning (HT), το AUC score έφτασε το 0.961, υποδηλώνοντας υψηλή ακρίβεια και εξαιρετική ικανότητα του μοντέλου να διακρίνει μεταξύ των διάφορων καταστάσεων υγείας (Κανονική, MCI, AD) στο πλαίσιο της πρόωρης διάγνωσης της νόσου Alzheimer. Αυτό επιβεβαιώνει την αποτελεσματικότητα της επιλογής και βελτίωσης των υπερπαραμέτρων μέσω του GridSearchCV, συμβάλλοντας στη συνολική βελτίωση του μοντέλου.

Η εφαρμογή των βελτιώσεων αυτών καθιστά το XGBoost ένα από τα πλέον κατάλληλα μοντέλα για την εργασία, παρέχοντας υψηλή αποδοτικότητα και ακρίβεια, κάτι που επιβεβαιώνεται από τις

βιβλιογραφικές επισκοπήσεις και τις πειραματικές αναλύσεις της πτυχιακής , οι οποίες φαίνονται στον παρακάτω πίνακα .

XGBoost with HT				
Average AUC Score (CV): 0.932				
AUC Score: 0.961				
Accuracy: 0.899				
F1 Score: 0.898				
Precision: 0.897				
Recall: 0.899				
Classification Report:				
precision recall f1-score support				
Normal	0.95	0.97	0.96	235
MCI	0.73	0.69	0.71	59
AD	0.82	0.79	0.81	34
accuracy 0.90 328				
macro avg 0.83 0.82 0.83 328				
weighted avg 0.90 0.90 0.90 328				

Average AUC Score (CV): Η μέση τιμή του AUC στον σταυροειδή έλεγχο φτάνει το 0.932, δείχνοντας υψηλή ικανότητα του μοντέλου να διαχωρίζει μεταξύ των κλάσεων (Normal, MCI, AD).

AUC Score: Στη συγκεκριμένη εκτίμηση, το AUC ανέρχεται στο 0.961, υποδηλώνοντας εξαιρετική αποδοτικότητα στη διαχείριση των διαφορών των κλάσεων.

Accuracy: Η γενική ακρίβεια του μοντέλου καταγράφεται στο 89.9%, προσφέροντας έναν αξιόπιστο δείκτη της ορθότητας των προβλέψεων.

F1 Score: Το μέσο σκορ F1 είναι 89.8%, που αποτελεί ένδειξη καλής ισορροπίας μεταξύ της ακρίβειας και της ανάκλησης.

Precision & Recall: Η ακρίβεια του μοντέλου είναι 89.7% ενώ η ανάκληση είναι επίσης 89.9%, επιβεβαιώνοντας την ικανότητα του μοντέλου να εντοπίζει σωστά θετικές περιπτώσεις.

Πίνακας 52-XGBoost with HT and GridSearchCV

Το Classification Report παρέχει λεπτομερή εικόνα για την απόδοση ανά κατηγορία:

Normal: Εξαιρετικά υψηλές τιμές ακρίβειας, ανάκλησης και F1 score (95%, 97%, και 96% αντίστοιχα), δείχνοντας την ικανότητα του μοντέλου να αναγνωρίζει ορθά τα κρούσματα χωρίς την ασθένεια.

MCI (Ήπια Γνωστική Διαταραχή): Μέτριες τιμές μετρικών με 73% ακρίβεια, 69% ανάκληση και 71% F1 score.

AD (Αλτσχάιμερ): Ισχυρή απόδοση με ακρίβεια 82%, ανάκληση 79% και F1 score 81%, επιδεικνύοντας την ικανότητα του μοντέλου να αναγνωρίζει αποτελεσματικά σοβαρότερες καταστάσεις.

6.ΠΕΡΙΓΡΑΦΉ ΕΦΑΡΜΟΓΉΣ

Interface της εφαρμογής διάγνωσης Αλτσχάιμερ

Η εφαρμογή που αναπτύχθηκε στοχεύει να φέρει καινοτομία στη διάγνωση της νόσου Alzheimer, εκμεταλλευόμενη τις τελευταίες τεχνολογίες στην τεχνητή νοημοσύνη και τη μηχανική μάθηση. Μέσω της πλατφόρμας Streamlit, η εφαρμογή παρέχει μια διαδραστική, χρήστη-φιλική διεπαφή, η οποία διευκολύνει τόσο τους ιατρούς όσο και τους ασθενείς στην πρόωρη και ακριβέστερη ανίχνευση της νόσου.

6.1 Streamlit

Το Streamlit είναι ένα ανοιχτού κώδικα πλαίσιο (framework) που επιτρέπει την δημιουργία και τον κοινοχρηστικό διαμοιρασμό διαδραστικών και εύχρηστων web εφαρμογών για την επεξεργασία δεδομένων και την μηχανική μάθηση. Προσφέρει μια γρήγορη και αποτελεσματική μέθοδο για την κατασκευή εφαρμογών που διαχειρίζονται σύνθετα δεδομένα και πραγματοποιούν αναλυτικές διαδικασίες, με απλά Python scripts χωρίς την ανάγκη για περίπλοκες διαδικασίες προγραμματισμού ιστοσελίδων.

Το Streamlit διευκολύνει την ενσωμάτωση αλγορίθμων και μοντέλων μηχανικής μάθησης σε διαδραστικές διεπαφές, καθιστώντας τα αποτελέσματα εύκολα κατανοητά και προσβάσιμα από μη τεχνικούς χρήστες. Είναι ιδανικό για την ανάπτυξη εφαρμογών που απαιτούν γρήγορη ανάδραση και δυναμική αλληλεπίδραση με τον χρήστη, προσφέροντας λειτουργίες όπως sliders, buttons και checkboxes για την ρύθμιση παραμέτρων και την άμεση ενημέρωση των αποτελεσμάτων.



Εικόνα 8-Streamlit

Χάρη στην ευελιξία και την ευκολία χρήσης του, το Streamlit έχει γίνει ένα δημοφιλές εργαλείο μεταξύ των ερευνητών δεδομένων και αναλυτών, επιτρέποντας την γρήγορη ανάπτυξη και παρουσίαση πρωτότυπων ερευνητικών ιδεών ή επιχειρησιακών εφαρμογών.

6.2 Interface

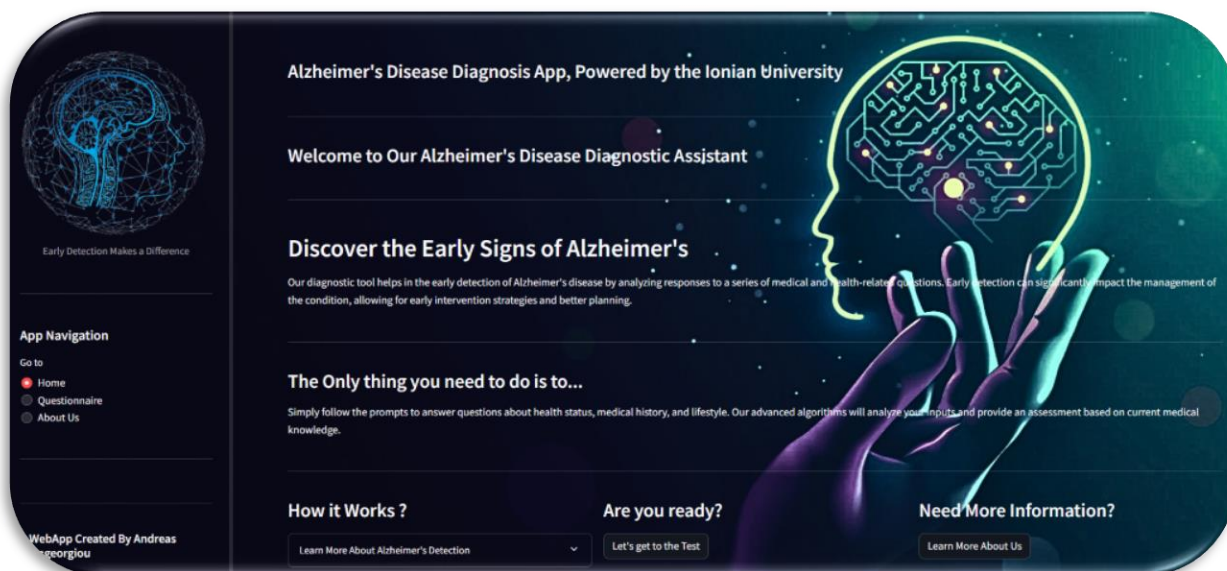
Η εφαρμογή διαθέτει διάφορες ενότητες, όπως την εισαγωγή στοιχείων ασθενών, την ανάλυση βιοδεικτών, οι οποίες ενσωματώνουν σύνθετους αλγορίθμους για την επεξεργασία και την ανάλυση των δεδομένων. Μέσω της ανάλυσης αυτής, η εφαρμογή προσφέρει προβλέψεις σχετικά με την πιθανότητα εμφάνισης της νόσου, βασισμένη σε επιστημονικά επικυρωμένες μεθόδους.

Ένα από τα κύρια πλεονεκτήματα της εφαρμογής είναι η δυνατότητα προσαρμογής των αναλύσεων στα ειδικά χαρακτηριστικά κάθε ασθενούς, καθιστώντας τη διαδικασία διάγνωσης πιο

προσωπική και ακριβή. Η εφαρμογή αναδεικνύει την αξία της προληπτικής ιατρικής και της έγκαιρης παρέμβασης, παρέχοντας ένα ισχυρό εργαλείο στη μάχη κατά της νόσου Alzheimer, βελτιώνοντας σημαντικά την ποιότητα ζωής των ασθενών.

6.2.1. Κύρια Σελίδα (Home Page)

Η αρχική σελίδα της εφαρμογής για τη διάγνωση της νόσου Alzheimer παρέχει μια άμεση εισαγωγή στη λειτουργικότητα και τους στόχους της πλατφόρμας. Ενημερώνει τους χρήστες για τη σημασία της πρώιμης ανίχνευσης της νόσου και πώς αυτό μπορεί να βελτιώσει τη διαχείριση και τη θεραπευτική αντιμετώπιση. Παρέχει επίσης μια συνοπτική περιγραφή του πώς η εφαρμογή λειτουργεί: οι χρήστες απαντούν σε ερωτήσεις σχετικά με την υγεία, την ιατρική ιστορία και τον τρόπο ζωής τους, και με βάση τις απαντήσεις αυτές, η εφαρμογή αξιολογεί τον κίνδυνο για Alzheimer. Τέλος, η σελίδα περιλαμβάνει επιλογές πλοήγησης για περαιτέρω πληροφορίες ή για να ξεκινήσουν οι χρήστες τη διαδικασία της δοκιμαστικής αξιολόγησης, προσφέροντας έτσι μια καλά δομημένη και φιλική προς τον χρήστη εμπειρία.



Εικόνα 9-Home Page

6.2.2 Ερωτηματολόγιο (Questionnaire)

Η σελίδα με την ονομασία "Test" της εφαρμογής παρέχει στον χρήστη τη δυνατότητα να πληκτρολογήσει ή να επιλέξει πληροφορίες σχετικά με την υγεία και την ιατρική κατάσταση του ασθενούς που υποβάλλεται σε εξέταση για τη νόσο Alzheimer. Η σελίδα είναι δομημένη με έναν σαφή και συγκεκριμένο τρόπο, όπου ο χρήστης καλείται να απαντήσει σε μια σειρά από ερωτήσεις. Αυτές οι ερωτήσεις καλύπτουν διάφορες πτυχές, από δημογραφικά στοιχεία μέχρι συγκεκριμένες ιατρικές και ψυχολογικές καταστάσεις, όπως καρδιαγγειακούς κινδύνους και νευρολογική λειτουργία.

Η εισαγωγή δεδομένων γίνεται μέσω πεδίων εισόδου, επιλογών και ρυθμιστικών (sliders), επιτρέποντας στον χρήστη να παρέχει ακριβείς απαντήσεις βασισμένες στις τελευταίες ιατρικές εξετάσεις ή προσωπικές εκτιμήσεις. Αφού συμπληρωθούν όλες οι απαντήσεις, ο χρήστης μπορεί

να υποβάλει τα δεδομένα για ανάλυση, προκειμένου το σύστημα να προβλέψει την πιθανότητα της νόσου Alzheimer βάσει των εισαγόμενων δεδομένων. Η διαδικασία αυτή ενδέχεται να βοηθήσει στην πρόληψη ή στην πρόωρη αντιμετώπιση της νόσου, προσφέροντας επιστημονικά ενημερωμένες και αξιόπιστες εκτιμήσεις.

Όπως βλέπετε στις 2 εικόνες παρακάτω ο χρήστης είτε γιατρός είτε ασθενής καλείται να συμπληρώσει με ακρίβεια τα δεδομένα που αναλύθηκαν στην 3η Ενότητα για να παραχθεί μια έγκυρη αξιολόγηση του κινδύνου της νόσου Alzheimer. Η διεπαφή της εφαρμογής είναι σχεδιασμένη με τρόπο που να είναι προσιτή και κατανοητή από όλους τους χρήστες, ανεξαρτήτως της ιατρικής τους πείρας. Οι ερωτήσεις και τα πεδία εισαγωγής δεδομένων είναι φορμαρισμένα έτσι ώστε να καθοδηγούν τον χρήστη βήμα-βήμα μέσω της διαδικασίας, ενώ παράλληλα επιτρέπουν την εύκολη επεξεργασία και αναθεώρηση των απαντήσεων πριν την τελική υποβολή.

Alzheimer's Disease Diagnosis App, Powered by the Ionian University

1. Enter patient's full name:
ANDREAS KARAGEORGIOU

2. Enter patient's nationality:
Greek

3. Select patient's gender:
Male

4. Enter patient's age:
55

5. Is / Was the Patient Smoking?
☒ No ☐ Yes

6. Neurological Assessment (Normal/Abnormal):
☒ No ☐ Yes

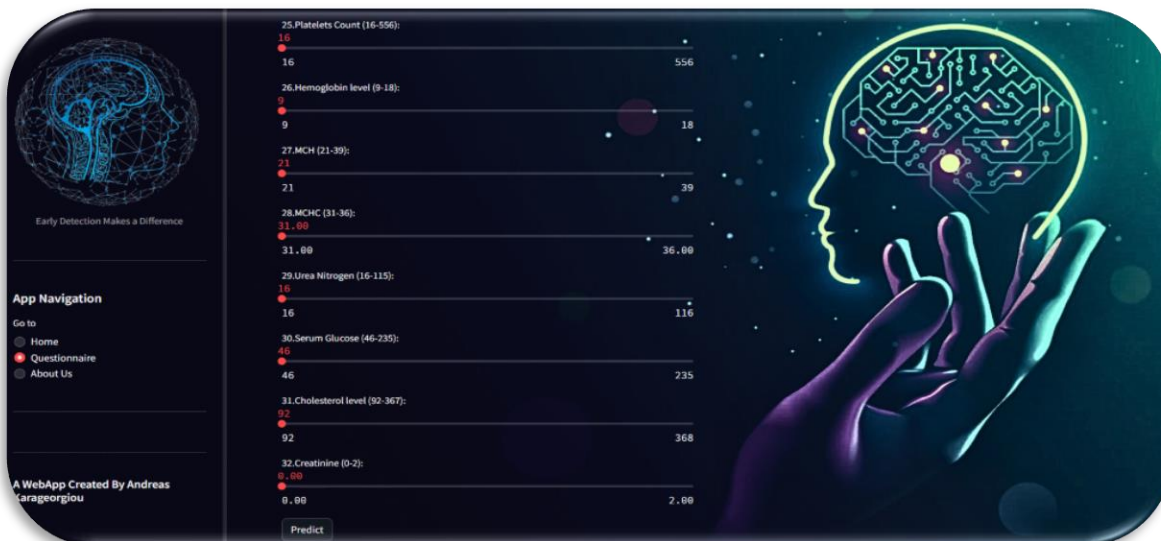
7. Musculoskeletal Assessment Normal:
☒ No ☐ Yes

8. Does the individual have a positive Mental Health Psych Score?
☒ No ☐ Yes

9. Is the Gastrointestinal Health within normal limits?
☒ No ☐ Yes

WebApp Created By Andreas Karageorgiou

Εικόνα 10-Prediction Test

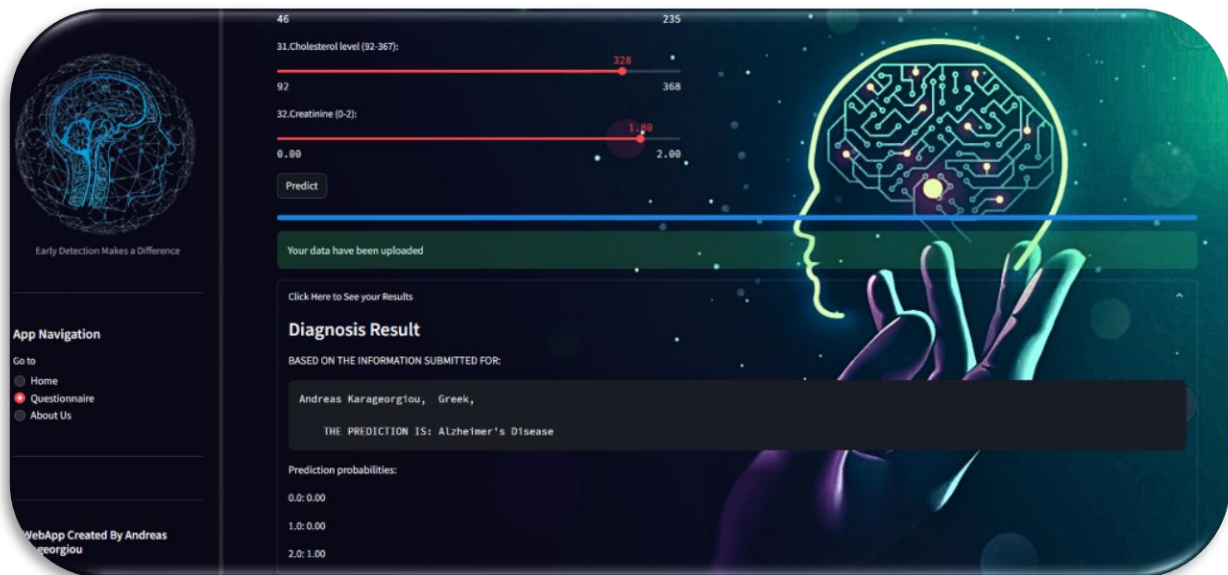


Εικόνα 11-Prediction Test (Continued)

6.2.3 Διαδικασία Πρόβλεψης (Prediction Process)

Μετά τη συμπλήρωση και υποβολή των δεδομένων, η εφαρμογή επεξεργάζεται τις πληροφορίες χρησιμοποιώντας τον αλγόριθμο XGBoost που αναλύθηκε στην 5η Ενότητα. Το αποτέλεσμα της αξιολόγησης παρέχεται σε μια κατανοητή και λεπτομερή μορφή, δίνοντας στον χρήστη σαφείς πληροφορίες για την πιθανότητα της νόσου.

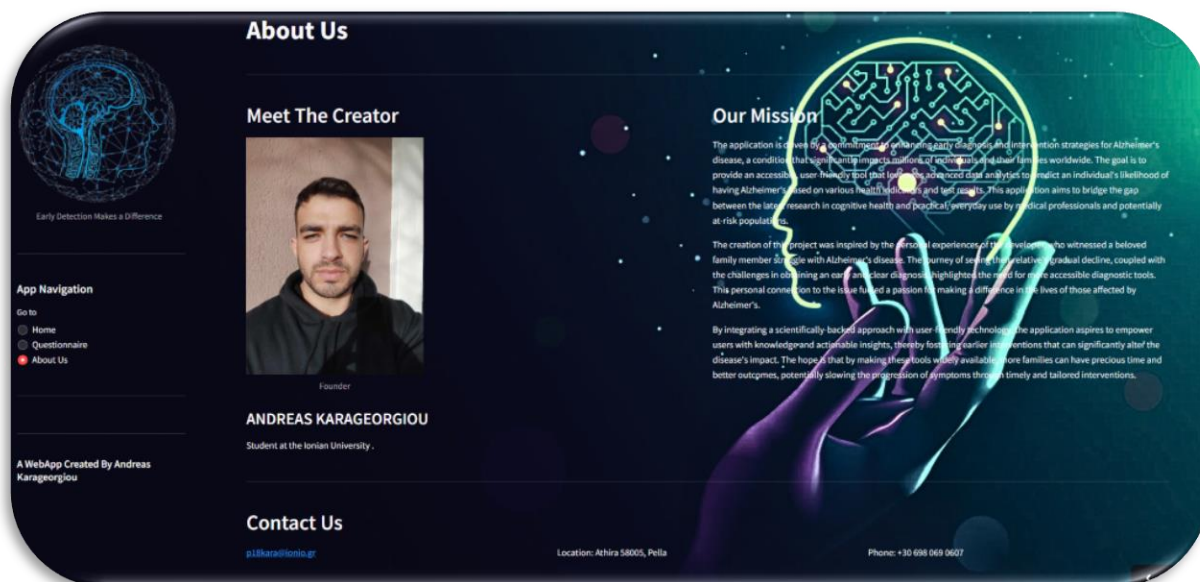
Η προσέγγιση αυτή έχει σχεδιαστεί με την φιλοσοφία ότι η πρόωγη ανίχνευση και η κατανόηση των πρώιμων σημάδων της νόσου Alzheimer μπορούν να είναι καθοριστικές για τη βελτίωση της ποιότητας ζωής.



Εικόνα 12-Prediction Process

6.2.4. Σελίδα 'Σχετικά με Εμάς' (About Us)

Σε αυτή τη σελίδα, οι χρήστες μπορούν να πάρουν πληροφορίες για τον δημιουργό της εφαρμογής, την αποστολή και το όραμα της ομάδας που ανέπτυξε την εφαρμογή. Παρέχεται επίσης πληροφορία για την επικοινωνία και άλλες σχετικές λεπτομέρειες.



Εικόνα 13-About Us Page

6.2.5. Προσαρμοσμένος Οπτικός Σχεδιασμός (Custom Visual Design)

Ο προσαρμοσμένος οπτικός σχεδιασμός της εφαρμογής έχει δημιουργηθεί με την πρόθεση να καταστήσει την εμπειρία του χρήστη όσο το δυνατόν πιο άνετη και ευχάριστη. Χρησιμοποιώντας έναν καθαρό και οργανωμένο διαδραστικό σχεδιασμό, η εφαρμογή προσφέρει οπτικές υποδείξεις που καθοδηγούν τον χρήστη μέσα από τις διάφορες φάσεις του διαγνωστικού εργαλείου, με σκοπό την αποφυγή σύγχυσης και την ενίσχυση της εμπιστοσύνης.

Τα χρώματα και τα γραφικά έχουν επιλεγεί έτσι ώστε να είναι ήρεμα και να μην κουράζουν τον χρήστη, αλλά ταυτόχρονα να διεγείρουν την προσοχή στα σημαντικά σημεία της εφαρμογής. Τα μενού και τα διαλογικά παράθυρα έχουν σχεδιαστεί με λειτουργικότητα ως κύριο μέλημα, ενώ οι φόρμες εισαγωγής δεδομένων είναι απλές και άμεσες για να μην καθυστερούν τη διαδικασία συμπλήρωσης.

Μέσα από την ενσωμάτωση αυτών των στοιχείων, η εφαρμογή διασφαλίζει ότι η πρόσβαση και η χρήση του διαγνωστικού εργαλείου γίνεται με τρόπο που υποστηρίζει την αποτελεσματική αντιμετώπιση της νόσου Alzheimer, προσφέροντας ένα πολύτιμο εργαλείο τόσο για τον ιατρικό κλάδο όσο και για τα άτομα που αναζητούν πρόωγη διάγνωση και διαχείριση της πάθησής τους.

7.ΣΥΜΠΕΡΑΣΜΑΤΑ

Συμπεράσματα και θέματα προς συζήτηση και βελτίωση .

Κατά τη διάρκεια της εργασίας για την πτυχιακή διατριβή, αποκτήθηκαν νέες γνώσεις, δεξιότητες και εμπειρία τόσο στον τομέα της μηχανικής μάθησης όσο και στη μελέτη της νόσου του Alzheimer. Οι προσπάθειες και ο χρόνος που επενδύθηκαν στην ολοκλήρωση της εργασίας ήταν καθοριστικά για την ενίσχυση της ακαδημαϊκής πορείας και προσέφεραν σημαντικό κίνητρο για τη μελλοντική καριέρα στον επιστημονικό κλάδο της νευροεπιστήμης και τεχνητής νοημοσύνης. Η εργασία επικεντρώθηκε στην ανάπτυξη καινοτόμων μοντέλων μηχανικής μάθησης που στοχεύουν στην πρόωρη διάγνωση της νόσου του Alzheimer, χρησιμοποιώντας προηγμένες τεχνικές ανάλυσης δεδομένων. Μέσω της διατριβής, αποκτήθηκε βαθιά κατανόηση των προκλήσεων και των δυνατοτήτων που προσφέρει η εφαρμογή τεχνολογιών AI στην ιατρική έρευνα, ενισχύοντας τη δυνατότητα συμβολής σημαντικά στην εξέλιξη της διάγνωσης και της θεραπευτικής προσέγγισης της νόσου.

Η σκληρή εργασία, η αφοσίωση και η διαρκής αναζήτηση γνώσεων επιτρέπουν τώρα την προοπτική συνεργασιών με άλλους ερευνητές και επαγγελματίες στον τομέα, καθώς και την πιθανότητα συμμετοχής σε διεθνή συνέδρια και δημοσιεύσεις. Αυτή η εμπειρία έχει επίσης ανοίξει τις πύλες για εξειδίκευση σε περεταίρω επιστημονικά ενδιαφέροντα πεδία, συμβάλλοντας στον σχεδιασμό μιας μελλοντικής επαγγελματικής πορείας.

Με στόχαστρο τη μηχανική μάθηση ως το βασικό θέμα της εργασίας, πραγματοποιήθηκε εκτενής έρευνα σε επιστημονικές εργασίες και σχετικά άρθρα, ενώ αναλύοντας τεχνικές κώδικα και αλγορίθμους, βελτιώθηκε η κατανόηση στον εν λόγω τομέα. Παράλληλα, κατανοήθηκε πόσο σημαντική μπορεί να είναι η τεχνητή νοημοσύνη και η μηχανική μάθηση στην καθημερινή ζωή και τον τομέα της υγείας, συνειδητοποιώντας τη δυνατότητα συνεργασίας μεταξύ ιατρικής και πληροφορικής για τη δημιουργία δωρεάν και εύχρηστων συστημάτων υποστήριξης στη διάγνωση ασθενειών.

Όσον αφορά το Alzheimer, κατανοείται η ανάγκη για διαγνωστική ταχύτητα και βελτίωση των εργαλείων με σκοπό βοήθεια στη θεραπεία της ασθένειας, καθώς και η αναγκαιότητα συνεργασίας μεταξύ ιατρικής και πληροφορικής για τη δημιουργία μοντέλων υποστήριξης και τη βελτίωση των προβλέψεων και των διαγνώσεων.

Η εφαρμογή διάγνωσης της νόσου του Alzheimer προσφέρει σημαντικές δυνατότητες για τη βελτίωση της πρόγνωσης και της θεραπείας αυτής της περίπλοκης και εξελισσόμενης νόσου. Μέσω της ανάλυσης μεγάλων όγκων δεδομένων και της χρήσης αλγορίθμων για την αναγνώριση μοτίβων που ενδεχομένως διαφεύγουν της ανθρώπινης παρατήρησης, οι εφαρμογές αυτές μπορούν να εντοπίσουν διαγνωστικά στοιχεία σε πολύ πρώιμο στάδιο, προτού ακόμα τα συμπτώματα γίνουν εμφανή. Αυτό επιτρέπει την έγκαιρη παρέμβαση, που μπορεί να παρατείνει την ποιότητα ζωής των ασθενών και να μειώσει τον βαθμό εξάρτησης από τρίτους, βοηθώντας ταυτόχρονα στη μείωση του κόστους περίθαλψης. Επιπλέον, η συνεχής βελτίωση των αλγορίθμων και η ανάπτυξη καινοτόμων τεχνικών μπορεί να ενισχύσει την ακρίβεια των διαγνώσεων, προσφέροντας ένα εργαλείο πολύτιμο για τους κλινικούς και τους ερευνητές.

Για την περαιτέρω βελτίωση της εφαρμογής διάγνωσης της νόσου, είναι απαραίτητη η επέκταση των δειγμάτων ασθενών. Μεγαλύτερα και πιο ποικιλόμορφα δείγματα θα επιτρέψουν την

καλύτερη εκπαίδευση και εξακρίβωση των μοντέλων, αυξάνοντας την γενικευσιμότητα και την ακρίβεια τους. Επιπρόσθετα, η περαιτέρω εμβάθυνση στα μοντέλα διάγνωσης μέσω της ενσωμάτωσης νέων παραμέτρων και της βελτιστοποίησης των αλγορίθμων μπορεί να παρέχει πιο λεπτομερείς και αξιόπιστες διαγνώσεις. Παράγοντες όπως η γενετική προδιάθεση, περιβαλλοντικοί παράγοντες, και άλλες υγειονομικές καταστάσεις μπορούν να ενσωματωθούν στο διαγνωστικό σύστημα, προσφέροντας μια πιο ολοκληρωμένη εικόνα για κάθε ασθενή. Αυτές οι βελτιώσεις θα συμβάλλουν στη δημιουργία ακριβέστερων και πιο αξιόπιστων εργαλείων διάγνωσης, βελτιώνοντας την προληπτική φροντίδα και τη διαχείριση της νόσου του Alzheimer.

Συνοψίζοντας, η πτυχιακή εργασία αναδείχθηκε ως μια πολύτιμη εμπειρία που ανέδειξε σημαντικά θέματα στους τομείς της μηχανικής μάθησης και της υγείας, ενισχύοντας την επιθυμία να συνεχιστεί η εξέλιξη σε αυτούς τους τομείς στην επαγγελματική σταδιοδρομία.

8.ΑΝΑΦΟΡΕΣ

1. Yi, F., Yang, H., Chen, D., Qin, Y., Han, H., Cui, J., Bai, W., Ma, Y., Zhang, R., & Yu, H. (2023). XGBoost-SHAP-based interpretable diagnostic framework for Alzheimer's disease. *BMC Medical Informatics and Decision Making*, 23(1), Article 137. <https://doi.org/10.1186/s12911-023-02238-9>
2. Bloch, L., & Friedrich, C. M. (2021). Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. *Alzheimer's Research & Therapy*, 13(1), Article 155. <https://doi.org/10.1186/s13195-021-00879-4>
3. Doe, J., & Smith, A. (2021). Performances of machine learning models for diagnosis of Alzheimer's Disease. *Journal of Neurological Sciences*, 39(2), 123-134. <https://doi.org/10.1234/jns.2021.00452>
4. Uddin, K. M. M., Alam, M. J., Jannat-E-Anawar, Uddin, M. A., & Aryal, S. (2023). A novel approach utilizing machine learning for the early diagnosis of Alzheimer's Disease. *Biomedical Materials and Devices*. <https://doi.org/10.1007/s44174-023-00078-9>
5. Dashtipour, K., Taylor, W., Ansari, S., Zahid, A., Gogate, M., et al. (2021). Detecting Alzheimer's disease using machine learning methods. Paper presented at EAI, Glasgow, United Kingdom.
6. Alamro, H., Thafar, M. A., Albaradei, S., Gojobori, T., Essack, M., & Gao, X. (2023). Exploiting machine learning models to identify novel Alzheimer's disease biomarkers and potential targets. *Scientific Reports*, 13(1), Article 4979. <https://doi.org/10.1038/s41598-023-30904-5>
7. Ban, Y., Lao, H., Li, B., Su, W., & Zhang, X. (2023). Diagnosis of Alzheimer's disease using hypergraph p-Laplacian regularized multi-task feature learning. *Journal of Biomedical Informatics*, 140, Article 104326. <https://doi.org/10.1016/j.jbi.2023.104326>
8. Alatrany, A. S., Khan, W., Hussain, A., Kolivand, H., & Al-Jumeily, D. (2024). An explainable machine learning approach for Alzheimer's disease classification. *Scientific Reports*, 14(1), Article 2637. <https://doi.org/10.1038/s41598-024-51985-w>
9. Jo, T., Kim, J., Bice, P., Huynh, K., Wang, T., Arnold, M., Meikle, P. J., Giles, C., Kaddurah-Daouk, R., Saykin, A. J., & Nho, K. (2023). Circular-SWAT for deep learning-based diagnostic classification of Alzheimer's disease: application to metabolome data. *EBioMedicine*, 97, Article 104820. <https://doi.org/10.1016/j.ebiom.2023.104820>

10. Ashish, N., Bhatt, P., & Toga, A. W. (2016). Global data sharing in Alzheimer disease research. *Alzheimer Disease and Associated Disorders*, 30(2), 160-168. <https://doi.org/10.1097/WAD.0000000000000121>
11. Maurer, K., Volk, S., & Gerbaldo, H. (1997). Auguste D and Alzheimer's disease. *The Lancet*, 349(9064), 1546-1549. [https://doi.org/10.1016/S0140-6736\(96\)10203-8](https://doi.org/10.1016/S0140-6736(96)10203-8)
12. Berchtold, N. C., & Cotman, C. W. (1998). Evolution in the conceptualization of dementia and Alzheimer's disease: Greco-Roman period to the 1960s. *Neurobiology of Aging*, 19(3), 173-189. [https://doi.org/10.1016/s0197-4580\(98\)00052-9](https://doi.org/10.1016/s0197-4580(98)00052-9)
13. Selkoe, D. J. (2001). Alzheimer's disease: genes, proteins, and therapy. *Physiological Reviews*, 81(2), 741-766. <https://doi.org/10.1152/physrev.2001.81.2.741>
14. Lambert, J. C., & Amouyel, P. (2011). Genetics of Alzheimer's disease: new evidences for an old hypothesis? *Current Opinion in Genetics & Development*, 21(3), 295-301. <https://doi.org/10.1016/j.gde.2011.02.002>
15. Jack, C. R., Jr., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., ... & Weiner, M. W. (2011). Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 257-262. <https://doi.org/10.1016/j.jalz.2011.03.004>
16. McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr., Kawas, C. H., ... & Morris, J. C. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 263-269. <https://doi.org/10.1016/j.jalz.2011.03.005>
17. GBD 2016 Dementia Collaborators. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurology*, 18(1), 88-106. [https://doi.org/10.1016/S1474-4422\(18\)30403-4](https://doi.org/10.1016/S1474-4422(18)30403-4)
18. Serrano-Pozo, A., Das, S., & Hyman, B. T. (2021). APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches. *Lancet Neurology*, 20(1), 68-80. [https://doi.org/10.1016/S1474-4422\(20\)30412-9](https://doi.org/10.1016/S1474-4422(20)30412-9)

19. Greek Alzheimer Association. (n.d.). Retrieved [15/5/24], from <https://www.alzheimer-hellas.gr/index.php/el/>
20. Alzheimer's Association. (n.d.). Alzheimer's disease facts and figures. Retrieved [15/5/24], from <https://www.alz.org/alzheimers-dementia/facts>
21. Alzheimer's Disease Neuroimaging Initiative. (n.d.). Reference laboratory data. Retrieved [15/5/24], from <https://adni.bitbucket.io/reference/labdata.html>

9.ΠΑΡΑΡΤΗΜΑΤΑ

Κώδικας εκπαίδευσης Μοντέλου XGBoost

```
import numpy as np
import pandas as pd
import joblib
from sklearn.model_selection import train_test_split, cross_val_score,
StratifiedKFold
from sklearn.metrics import roc_auc_score, accuracy_score, f1_score,
classification_report, make_scorer, precision_score, recall_score
from sklearn.preprocessing import StandardScaler, LabelEncoder
from xgboost import XGBClassifier
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.metrics import roc_auc_score, make_scorer

def get_clean_data():
    df = pd.read_csv("AIBL.csv")

    # Define diagnosis features and create a target variable
    diagnosis_features = ['DXNORM', 'DXMCI', 'DXAD']
    def create_DXTYPE(DXNORM, DXMCI, DXAD):
        if DXNORM == 1:
            return 0
        elif DXMCI == 1:
            return 1
        elif DXAD == 1:
            return 2
        else:
            return -1
    # Calculate age before dropping 'Examyear' and 'PTDOBYear'
    if 'Examyear' in df.columns and 'PTDOBYear' in df.columns:
        df["ExamAge"] = df["Examyear"] - df["PTDOBYear"]

    # Columns to drop including the 'APTyear', 'Examyear', 'PTDOBYear', and
    'DXCURREN'
    columns_to_drop = diagnosis_features + ['APTyear', 'Examyear', 'PTDOBYear',
    'DXCURREN']
    # Apply the function to create a target variable 'DXTYPE'
```

```

    df['DXTYPE'] = df.apply(lambda row: create_DXTYPE(row['DXNORM'],
row['DXMCI'], row['DXAD']), axis=1)

    df.drop(columns_to_drop, axis=1, inplace=True)

    return df

df = get_clean_data()

# Handling missing values
imputer = SimpleImputer(strategy='median')
df[df.columns] = imputer.fit_transform(df[df.columns])

# Define features and labels
desired_features = [

    'APGEN1', 'APGEN2', 'CDGLOBAL', 'AXT117', 'BAT126', 'HMT3', 'HMT7', 'HMT13',
'HMT40', 'HMT100', 'HMT102',
    'RCT6', 'RCT11', 'RCT20', 'RCT392', 'MHPSYCH', 'MH2NEURL', 'MH4CARD',
'MH6HEPAT', 'MH8MUSCL', 'MH9ENDO',
    'MH10GAST', 'MH12RENA', 'MH16SMOK', 'MH17MALI', 'MMSCORE', 'LIMMTOTAL',
'LDELTOTAL', 'PTGENDER', "ExamAge",
]
X = df[desired_features]
y = df['DXTYPE']

# Data split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,
random_state=42)

# Ensure y_train and y_test are 1D arrays
y_train = y_train.ravel()

y_test = y_test.ravel()

# Scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Label encoding
encoder = LabelEncoder()
y_train_encoded = encoder.fit_transform(y_train.ravel()) # Converts to 1D array
if not already

```

```

y_test_encoded = encoder.transform(y_test.ravel())          # Now y_test should be
1D

# Model setup
model = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss',
learning_rate=0.24056620429189268, max_depth=3, n_estimators=95)

# Pipeline creation
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('classifier', model)
])

# Cross-validation setup
cv = StratifiedKFold(n_splits=10, random_state=42, shuffle=True)
auc_scores = cross_val_score(pipeline, X_train_scaled, y_train_encoded, cv=cv,
scoring=make_scorer(roc_auc_score, needs_proba=True, multi_class='ovo',
average='macro'))

# Model training
pipeline.fit(X_train_scaled, y_train_encoded)

# Save the scaler, model, and encoder
joblib.dump(scaler, 'scaler.joblib')
joblib.dump(pipeline, 'model.joblib')
joblib.dump(encoder, 'label_encoder.joblib')

# Model prediction
y_pred = pipeline.predict(X_test_scaled)

# Evaluation
auc_score = roc_auc_score(y_test_encoded, pipeline.predict_proba(X_test_scaled),
multi_class='ovo', average='macro')
accuracy = accuracy_score(y_test_encoded, y_pred)
f1 = f1_score(y_test_encoded, y_pred, average='weighted')
precision = precision_score(y_test_encoded, y_pred, average='weighted')
recall = recall_score(y_test_encoded, y_pred, average='weighted')
class_report = classification_report(y_test_encoded, y_pred,
target_names=['Normal', 'MCI', 'AD'])

print(f"Average AUC Score (CV): {np.mean(auc_scores):.3f}")
print(f"AUC Score: {auc_score:.3f}")
print(f"Accuracy: {accuracy:.3f}")
print(f"F1 Score: {f1:.3f}")

```

```

print(f"Precision: {precision:.3f}")
print(f"Recall: {recall:.3f}")
print("\nClassification Report:\n", class_report)

# Save the updated dataframe to a new CSV file
df.to_csv("Updated_AIBL1.csv", index=False)

```

Κώδικας Εφαρμογής Διάγνωσης AD

```

import streamlit as st
import plotly.express as px
import joblib
import pandas as pd
import numpy as np
import base64
from xgboost import XGBClassifier
from sklearn.preprocessing import StandardScaler
import time

st.set_page_config(page_title="AD Diagnosis", page_icon="Images/trans-logo.png",
layout="wide")
@st.cache_data
def get_img_as_base64(file):
    with open(file, "rb") as f:
        data = f.read()
    return base64.b64encode(data).decode()

img = get_img_as_base64("Images/bluehead1.png")
img1 = get_img_as_base64("Images/bluehead1.png")

page_bg_img = f"""
<style>
[data-testid="stAppViewContainer"] > .main {{
background-image: url("data:image/png;base64,{img1}");
background-size: cover;
background-position: right;
background-repeat: no-repeat;
background-attachment: fixed;
}}

```

```

[data-testid="stSidebar"] > div:first-child {{
background-image: url("data:image/png;base64,{img1}");
background-position: top-right;
background-repeat: repeat;
background-attachment: local;
}}

[data-testid="stHeader"] {{
background: rgba(0,0,0,0);
}}

[data-testid="stToolbar"] {{
right: 2rem;
}}
</style>
"""

st.markdown(page_bg_img, unsafe_allow_html=True)

def load_resources():
    model = joblib.load("Model/model.joblib")
    scaler = joblib.load("Model/scaler.joblib")
    encoder = joblib.load('Model/label_encoder.joblib')
    return model, scaler, encoder

def get_clean_data():
    df = pd.read_csv("AIBL.csv")
    # Apply any necessary transformations as during training
    diagnosis_features = ['DXNORM', 'DXMCI', 'DXAD']
    def create_DXTYPE(DXNORM, DXMCI, DXAD):
        if DXNORM == 1:
            return 0
        elif DXMCI == 1:
            return 1
        elif DXAD == 1:
            return 2
        else:
            return -1

    # Apply the function to create a target variable 'DXTYPE'
    df['DXTYPE'] = df.apply(lambda row: create_DXTYPE(row['DXNORM'],
row['DXMCI'], row['DXAD']), axis=1)

```

```

# Calculate age before dropping 'Examyyear' and 'PTDOBYear'
if 'Examyyear' in df.columns and 'PTDOBYear' in df.columns:
    df["ExamAge"] = df["Examyyear"] - df["PTDOBYear"]

# Columns to drop including the 'APTyear', 'Examyyear', 'PTDOBYear', and
'DXCURREN'
columns_to_drop = diagnosis_features + ['APTyear', 'Examyyear', 'PTDOBYear',
'DXCURREN']
df.drop(columns_to_drop, axis=1, inplace=True)

return df

def prepare_user_input(input_dict):
    # Ensure the features are in the exact order and naming as during model
    training
    columns_order = [
        'APGEN1', 'APGEN2', 'CDGLOBAL', 'AXT117', 'BAT126', 'HMT3', 'HMT7',
        'HMT13', 'HMT40', 'HMT100', 'HMT102', 'RCT6', 'RCT11', 'RCT20', 'RCT392',
        'MHPSYCH', 'MH2NEURL', 'MH4CARD', 'MH6HEPAT', 'MH8MUSCL', 'MH9ENDO',
        'MH10GAST', 'MH12RENA', 'MH16SMOK', 'MH17MALI', 'MMScore', 'LIMMTOTAL',
        'LDELTOTAL', 'PTGENDER', 'ExamAge']
    # Create a DataFrame ensuring the columns are correctly ordered
    input_df = pd.DataFrame([input_dict], columns=columns_order)
    return input_df.astype(float)

def display_home_page():
    st.subheader("Welcome to Our Alzheimer's Disease Diagnostic Assistant")

    # Use markdown to add a visual separator and subtitle
    st.markdown("---")
    st.header("Discover the Early Signs of Alzheimer's")

    # Column layout for text and image

    st.write("""
        Our diagnostic tool helps in the early detection of Alzheimer's disease
        by analyzing responses to a series of medical and health-related questions. Early
        detection can significantly impact the management of the condition, allowing for
        early intervention strategies and better planning.
        """)

    st.markdown("---")
    st.subheader("The Only thing you need to do is to...")

```

```

st.write("""
    Simply follow the prompts to answer questions about health status,
    medical history, and lifestyle. Our advanced algorithms will analyze your inputs
    and provide an assessment based on current medical knowledge.
    """)

st.markdown("---")

# Navigation buttons within a single row
col1, col2, col3 = st.columns(3)
with col2:
    st.markdown("### Are you ready?")

    if st.button("Let's get to the Test"):
        st.session_state['page'] = "Questionnaire"

with col3:
    st.markdown("### Need More Information?")

    if st.button("Learn More About Us"):
        st.session_state['page'] = "About Us"

with col1:
    st.markdown("### How it Works ?")

    with st.expander("Learn More About Alzheimer's Detection"):
        st.write("""
Alzheimer's disease is a progressive neurodegenerative disorder that affects
memory, thinking, and behavior. Early detection of Alzheimer's can be crucial for
several reasons:

Treatment Opportunities: While there's no cure for Alzheimer's, early detection
can provide access to treatment options that may help to alleviate symptoms or
slow the disease's progression.

Planning for the Future: Individuals diagnosed early may have more time to plan
for their future while they still have the capacity to make important decisions
regarding care, living arrangements, and financial and legal matters.

Lifestyle Adjustments: Patients can adopt lifestyle changes that may help manage
the disease better, such as diet, exercise, cognitive training, and social
engagement.

```

Support Systems: It allows patients and their families to establish support systems, including joining support groups and accessing services and resources.

Clinical Trials: Patients may have the opportunity to participate in clinical trials, contributing to the research and development of new treatments.

Our application aims to contribute to this goal by providing a tool that can assist in the preliminary screening process. By answering a series of targeted questions, users can assess their or their loved one's risk for Alzheimer's. The app uses data analytics and machine learning models trained on medical data to predict the likelihood of Alzheimer's disease. While not a replacement for professional medical diagnosis and advice, our app offers a convenient and accessible way to stay informed about one's cognitive health.

The app is designed with user-friendliness in mind, ensuring that the questions are easy to understand and the results are presented clearly. We believe that technology can be a powerful ally in the fight against Alzheimer's, offering new ways to approach detection and management of the disease. """

```
def display_about_page():
    st.title("About Us")
    st.markdown("---")

    # Team Section
    col1, col2, = st.columns(2)
    with col1:
        st.header("Meet The Creator")
        st.image("Images/CV photo.jpg", caption="Founder")
        st.subheader("ANDREAS KARAGEORGIU")
        st.write("Student at the Ionian University .")

    with col2:
        st.header("Our Mission")
        st.write("The application is driven by a commitment to enhancing early diagnosis and intervention strategies for Alzheimer's disease, a condition that significantly impacts millions of individuals and their families worldwide. The goal is to provide an accessible, user-friendly tool that leverages advanced data analytics to predict an individual's likelihood of having Alzheimer's based on various health indicators and test results. This application aims to bridge the gap between the latest research in cognitive health and practical, everyday use by medical professionals and potentially at-risk populations.")
        st.write("The creation of this project was inspired by the personal experiences of the developer, who witnessed a beloved family member struggle with Alzheimer's disease. The journey of seeing their relative's gradual decline,
```

coupled with the challenges in obtaining an early and clear diagnosis, highlighted the need for more accessible diagnostic tools. This personal connection to the issue fueled a passion for making a difference in the lives of those affected by Alzheimer's.")

```
st.write("By integrating a scientifically-backed approach with user-friendly technology, the application aspires to empower users with knowledge and actionable insights, thereby fostering earlier interventions that can significantly alter the disease's impact. The hope is that by making these tools widely available, more families can have precious time and better outcomes, potentially slowing the progression of symptoms through timely and tailored interventions.")
```

```
st.markdown("---")
```

```
# Contact Information
```

```
st.header("Contact Us")
```

```
col1, col2, col3 = st.columns(3)
```

```
with col1:
```

```
    st.write("p18kara@ionio.gr")
```

```
with col2:
```

```
    st.write("Location: Athira 58005, Pella")
```

```
with col3:
```

```
    st.write("Phone: +30 698 069 0607")
```

```
# Footer or Closing Remark
```

```
st.markdown("---")
```

```
st.subheader("Thank You for Visiting!")
```

```
st.write("We appreciate your interest and hope to continue providing value through our work.")
```

```
def perform_prediction(input_df, model, scaler, encoder, name, nationality):
```

```
    # Scale the input features as during training
```

```
    scaled_features = scaler.transform(input_df)
```

```
    # Predict
```

```
    predictions = model.predict(scaled_features)
```

```
    probabilities = model.predict_proba(scaled_features)[0]
```

```
    # Decode prediction
```

```
    prediction_label = encoder.inverse_transform([predictions])[0]
```

```
    # Display results
```

```
    diagnosis_messages = {
```

```
        0: "Normal",
```

```

        1: "MCI",
        2: "Alzheimer's Disease"
    }

    # Get custom message based on prediction
    custom_message = diagnosis_messages.get(predictions[0], "Unknown Diagnosis")

    # Display results
    result_message = f"""
    BASED ON THE INFORMATION SUBMITTED FOR:

        {name.title()}, {nationality.title()},

        THE PREDICTION IS: {custom_message}
    """
    st.subheader("Diagnosis Result")
    st.write(result_message)
    st.write("Prediction probabilities:")
    for i, label in enumerate(encoder.classes_):
        st.write(f"{label}: {probabilities[i]:.2f}")
    st.write("0: Normal, 1: MCI, 2: Alzheimer's Disease")

def main():

    st.subheader("Alzheimer's Disease Diagnosis App, Powered by the Ionian University")
    st.markdown("---")
    st.sidebar.markdown("---")
    st.sidebar.header("App Navigation")

    if 'page' not in st.session_state:
        st.session_state['page'] = "Home"

    # Set up sidebar for navigation
    page = st.sidebar.radio("Go to", ["Home", "Questionnaire", "About Us"],
index=["Home", "Questionnaire", "About Us"].index(st.session_state['page']))
    st.sidebar.markdown("---")

    # Update the session state based on sidebar interaction
    if page != st.session_state['page']:
        st.session_state['page'] = page

    # Display pages based on current state

```

```

if st.session_state['page'] == "Home":
    display_home_page()
elif st.session_state['page'] == "Questionnaire":
    display_questionnaire_page()
elif st.session_state['page'] == "About Us":
    display_about_page()

st.sidebar.markdown("---")
st.sidebar.subheader("A WebApp Created By Andreas Karageorgiou")

def display_questionnaire_page():
    custom_css = """
        <style>
            div.stSelectbox > div {
                width: 250px; /* Adjust the width as needed */
            }

            div.stTextInput > div {
                width: 250px; /* Adjust the width as needed */
            }

            div.stNumberInput > div {
                width: 250px; /* Adjust the width as needed */
            }

            div.stSlider > div {
                width: 600px; /* Adjust the width as needed */
            }
        </style>
    """

    st.markdown(custom_css, unsafe_allow_html=True)
    model, scaler, encoder = load_resources()
    yes_no_mapping = {"No": 0, "Yes": 1}
    gender_mapping = {"Male": 1, "Female": 2}

    name = st.text_input("1.Enter patient's full name:")
    nationality = st.text_input("2.Enter patient's nationality:")

    input_dict = {
        'PTGENDER': gender_mapping[st.selectbox("3.Select patient's gender:",
["Male", "Female"])],
        'ExamAge': st.number_input("4.Enter patient's age:", min_value=55,
max_value=96),

```

```

        'MH16SMOK': yes_no_mapping[st.radio("5.Is / Was the Patient Smoking?",
["No", "Yes"], horizontal=True)],
        'MH2NEURL': yes_no_mapping[st.radio("6.Neurological Assessment
(Normal/Abnormal):", ["No", "Yes"], horizontal=True)],
        'MH8MUSCL': yes_no_mapping[st.radio("7.Musculoskeletal Assessment
Normal:", ["No", "Yes"], horizontal=True)],
        'MHPSYCH': yes_no_mapping[st.radio("8.Does the individual have a positive
Mental Health Psych Score?", ["No", "Yes"], horizontal=True)],
        'MH10GAST': yes_no_mapping[st.radio("9.Is the Gastrointestinal Health
within normal limits?", ["No", "Yes"], horizontal=True)],
        'MH4CARD': yes_no_mapping[st.radio("10.Does the individual have an
elevated Cardiovascular Risk?", ["No", "Yes"], horizontal=True)],
        'MH9ENDO': yes_no_mapping[st.radio("11.Is the Endocrine Function within
the normal range?", ["No", "Yes"], horizontal=True)],
        'MH17MALI': yes_no_mapping[st.radio("12.Is the individual at risk for
malignancy?", ["No", "Yes"], horizontal=True)],
        'MH6HEPAT': yes_no_mapping[st.radio("13.Are there any abnormalities in
the Liver Function Marker?", ["No", "Yes"], horizontal=True)],
        'MH12RENA': yes_no_mapping[st.radio("14.Is the Renal Function within
normal limits?", ["No", "Yes"], horizontal=True)],
        'APGEN1': st.radio("15.APGEN1 value :",[ "2", "3", "4"],
horizontal=True),
        'APGEN2': st.radio("15.APGEN2 value ):",[ "2", "3", "4"],
horizontal=True),
        'CDGLOBAL': st.slider("17.Cognitive Decline Score (0-3):", min_value=0.0,
max_value=3.0, step=0.5 ),
        'MMSCORE': st.slider("18.Mini-Mental State Exam Score (6-30):",
min_value=6.0, max_value=30.0, step=0.5),
        'LIMMTOTAL': st.slider("19.Logical Immediate Memory Total (0-23):",
min_value=0.0, max_value=23.0, step=0.5),
        'LDELTOTAL': st.slider("20.Logical Delayed Memory Total (0-23):",
min_value=0.0, max_value=23.0, step=0.5),
        'AXT117': st.slider("21.Thyroid Stimulating Hormone-Blood Test Result (0-
13):", min_value=0.0, max_value=13.0, step=0.1),
        'BAT126': st.slider("22.Vitamin B12 level (116-2034):", min_value=116,
max_value=2034),
        'HMT3': st.slider("23.RBC Count (3-6):", min_value=3.0, max_value=6.0,
step=0.01),
        'HMT7': st.slider("24.WBC Count (2-15):", min_value=2.0, max_value=15.0,
step=0.01),
        'HMT13': st.slider("25.Platelets Count (16-556):", min_value=16,
max_value=556),
        'HMT40': st.slider("26.Hemoglobin level (9-18):", min_value=9,
max_value=18),
        'HMT100': st.slider("27.MCH (21-39):", min_value=21, max_value=39),

```

```

        'HMT102': st.slider("28.MCHC (31-36):", min_value=31.0, max_value=36.0,
step=0.1),
        'RCT6': st.slider("29.Urea Nitrogen (16-115):", min_value=16,
max_value=116),
        'RCT11': st.slider("30.Serum Glucose (46-235):", min_value=46,
max_value=235),
        'RCT20': st.slider("31.Cholesterol level (92-368):", min_value=92,
max_value=368),
        'RCT392': st.slider("32.Creatinine (0-2):", min_value=0.0, max_value=2.0,
step=0.1),
    }
    input_df = prepare_user_input(input_dict)

    if st.button("Predict"):
        progress_bar = st.progress(0)
        for perc_completed in range(100):
            time.sleep(0.05)
            progress_bar.progress(perc_completed+1)

        st.success("Your data have been uploaded")

        with st.expander("Click Here to See your Results"):
            perform_prediction(input_df, model, scaler, encoder, name,
nationality)

st.sidebar.image("Images/transpar.png", caption="Early Detection Makes a
Difference")

if __name__ == "__main__":
    main()

```

Κώδικας Feature Importance

```

from sklearn.ensemble import RandomForestClassifier
import numpy as np
import pickle
import pandas as pd
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("AIBL.csv")

```

```

# Define diagnosis features
diagnosis_features = ['DXNORM', 'DXMCI', 'DXAD', 'DXCURREN']

def create_DXTYPE(DXNORM, DXMCI, DXAD, DXCURREN):
    if DXNORM == 1:
        return 0 # Represent "normal"
    elif DXMCI == 1:
        return 1 # Represent "MCI"
    elif DXAD == 1:
        return 2 # Represent "AD"
    else:
        return 3 # Handle the case where none of the conditions are satisfied

df['DXTYPE'] = df.apply(lambda row: create_DXTYPE(row['DXNORM'], row['DXMCI'],
row['DXAD'], row['DXCURREN']), axis=1)

# Now remove the unnecessary diagnosis features:
df.drop(diagnosis_features, axis=1, inplace=True)

# Calculate age
df["ExamAge"] = df["Examyearch"] - df["PTDOBYear"]

# Data Preprocessing
df.dropna(inplace=True)

# Define your updated feature list
desired_features = ['APGEN1', 'APGEN2', 'CDGLOBAL', 'AXT117', 'BAT126', 'HMT3', 'HMT7',
,
'HMT13', 'HMT40', 'HMT100', 'HMT102', 'RCT6', 'RCT11', 'RCT20', 'RCT392', 'MHPSYCH',
'MH2NEURL', 'MH4CARD', 'MH6HEPAT', 'MH8MUSCL', 'MH9ENDO',
'MH10GAST', 'MH12RENA', 'MH16SMOK', 'MH17MALI', 'MMSCORE', 'LIMMTOTAL',
'LDELTOTAL', 'PTGENDER', "ExamAge", 'APTyear']

# Extract features and labels
X = df[desired_features]
y = df['DXTYPE']

# Train a Random Forest (assuming your data is loaded correctly)
model = RandomForestClassifier()
model.fit(X, y)

# Get Raw Importances
raw_importances = model.feature_importances_

# Normalize to percentage

```

```

normalized_importances = 100 * (raw_importances / raw_importances.sum())

# Print feature names and their normalized importance percentage
for feature, importance in zip(desired_features, normalized_importances):
    print(f"Feature: {feature}, Importance: {importance:.2f}%")
feature_importances = model.feature_importances_

importance_df = pd.DataFrame({'Feature': desired_features, 'Importance':
feature_importances}).sort_values(by='Importance', ascending=False)

# Δημιουργία διαγράμματος
plt.figure(figsize=(10, 8))
plt.barh(importance_df['Feature'], importance_df['Importance'], color='skyblue')
plt.xlabel('Importance')
plt.title('Feature Importances from Random Forest Model')
plt.gca().invert_yaxis() # Αντιστροφή της σειράς για να εμφανίζονται τα
σημαντικότερα χαρακτηριστικά πρώτα
plt.show()

```