



**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Ανάλυση Βιο-δεδομένων

Ακαδημαϊκό έτος 2023-2024, Εαρινό Εξάμηνο

***Ανίχνευση Νόσου Alzheimer μέσω Δεδομένων Γραφής***

**Ονοματεπώνυμο**

**ΑΜ**

Ιωάννης Βόγκας

03400206

Ευγενία Παπούλια

03400228


Κωνσταντίνος Πριμέτης

03400231

Διονύσιος Χοντζάκης

03400238

Ιούλιος 2024

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p> <hr/> <p style="text-align: center;">Σελίδα 1 / 22</p>
----------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------

## ΠΕΡΙΕΧΟΜΕΝΑ


<b>1</b>	<b>ΕΙΣΑΓΩΓΗ .....</b>	<b>3</b>
<b>2</b>	<b>ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ DARWIN .....</b>	<b>4</b>
<b>3</b>	<b>ΝΟΣΟΣ ALZHEIMER ΚΑΙ ΚΛΙΝΙΚΟΣ ΑΝΤΙΚΤΥΠΟΣ.....</b>	<b>8</b>
<b>4</b>	<b>ΜΕΘΟΔΟΛΟΓΙΑ ΠΡΟΣΕΓΓΙΣΗΣ .....</b>	<b>9</b>
4.1	ΑΝΑΛΥΣΗ ΟΛΟΚΛΗΡΟΥ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ.....	9
4.2	ΑΝΑΛΥΣΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΑΝΑ ΔΟΚΙΜΑΣΙΑ.....	14
4.3	ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΩΝ ΑΠΟ ΤΙΣ ΠΙΟ ΧΡΗΣΙΜΕΣ ΚΑΙ ΤΙΣ ΛΙΓΟΤΕΡΟ ΧΡΗΣΙΜΕΣ ΔΟΚΙΜΑΣΙΕΣ .....	15
<b>5</b>	<b>ΑΠΟΤΕΛΕΣΜΑΤΑ.....</b>	<b>16</b>
5.1	ΜΟΝΤΕΛΑ ΑΠΟ ΟΛΟΚΛΗΡΟ ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ .....	16
5.2	ΜΟΝΤΕΛΑ ΑΠΟ ΤΟΝ ΔΙΑΧΩΡΙΣΜΟ ΤΩΝ ΕΓΓΡΑΦΩΝ ΑΝΑ ΔΟΚΙΜΑΣΙΑ .....	18
5.3	ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ ΜΕΤΑΞΥ ΧΡΗΣΙΜΟΤΕΡΩΝ ΚΑΙ ΜΗ ΔΟΚΙΜΑΣΙΩΝ .....	19
<b>6</b>	<b>ΣΥΜΠΕΡΑΣΜΑΤΑ .....</b>	<b>21</b>

## ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ


Σχήμα 1. Εξηγημένη διακύμανση – Διαστάσεις (elbow στο 0.97).....	13
------------------------------------------------------------------	----

## ΛΙΣΤΑ ΠΙΝΑΚΩΝ

Πίνακας 1. Δοκιμασίες που εκτελέστηκαν στο πλαίσιο απόκτησης δεδομένων.....	5
Πίνακας 2. Τα 18 χαρακτηριστικά που εξήχθησαν για κάθε δοκιμασία (Cilia, N. et. al, 2022). .....	6
Πίνακας 3. Υπερπαράμετροι προς βελτιστοποίηση που αναζητήθηκαν κατά την ανάπτυξη του μοντέλου. ....	12
Πίνακας 4. Επίδοση (accuracy) μοντέλων με απομείωση – μετασχηματισμό διαστάσεων.....	16
Πίνακας 5. Επίδοση (accuracy) μοντέλων με επαύξηση δεδομένων και PCA στο 97% της συνολικής διακύμανσης. ....	17

	<p><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΓΡΑΦΗΣ</b></p> <p>ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p>ΙΟΥΛΙΟΣ 2024</p>
		<p>Σελίδα 2 / 22</p>

Πίνακας 6. Αποτελέσματα ακρίβειας (accuracy) – Μοντέλα εκπαιδευμένα στο σύνολο των διαθέσιμων δεδομένων. ....	17
Πίνακας 7. Σειρά κατάταξης μοντέλων, εκπαιδευμένων ανά δοκιμασία. ....	18
Πίνακας 8. Αξιολόγηση (accuracy) μοντέλων βασιζόμενων σε 12 από τις 25 δοκιμασίες. ....	20


	<p><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΓΡΑΦΗΣ</b></p> <p>ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p>ΙΟΥΛΙΟΣ 2024</p>
		<p>Σελίδα 3 / 22</p>

## 1 ΕΙΣΑΓΩΓΗ

Η παρούσα εργασία αποτελεί την εξαμηνιαία εργασία που εκπονήθηκε στο πλαίσιο του Μαθήματος «Ανάλυση Βιο-δεδομένων» του ΔΠΜΣ «Επιστήμη Δεδομένων και Μηχανική Μάθηση».

Στο πλαίσιο της εργασίας, στόχο αποτέλεσε η αναζήτηση και εύρεση ενός συνόλου Βιο-δεδομένων, πάνω στο οποίο επρόκειτο να εφαρμοστούν μέθοδοι και τεχνικές Μηχανικής Μάθησης. Το σύνολο δεδομένων που επιλέχθηκε ονομάζεται “DARWIN” και αφορά σε δεδομένα γραφής που λήφθηκαν από 174 ανθρώπους, κάποιοι από τους οποίους πάσχουν από τη νόσο Alzheimer. Τα δεδομένα αξιοποιήθηκαν με στόχο την ανάπτυξη μοντέλου μηχανικής μάθησης, το οποίο θα ταξινομεί ενδεχόμενους ασθενείς στις κατηγορίες πασχόντων ή μη από τη νόσο με βάση τα δεδομένα από τη γραφή τους.

Η παρούσα έκθεση αποτελεί αναπόσπαστο τμήμα της παρουσίασης με τίτλο «Ανίχνευση της νόσου Alzheimer μέσω της γραφής», ο συνδυασμός των οποίων αποτελεί το παραδοτέο της εργασίας. Ακολουθώς, διακριτοποιείται και παρουσιάζεται σε έξι (6) ενότητες, αποσκοπώντας στη βέλτιστη παρουσίαση και επεξήγηση του συνόλου δεδομένων, του κλινικού αντικτύπου που μπορεί να έχει το αποτέλεσμα, της μεθοδολογίας προσέγγισης του προβλήματος και τελικά των αποτελεσμάτων που προέκυψαν και των συμπερασμάτων που εξήχθησαν.

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: center;">Σελίδα 4 / 22</p>


## 2 ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ DARWIN

Για τη λήψη των δεδομένων που αποτελούν το σύνολο δεδομένων DARWIN (Cilia, N. et. al, 2022) πραγματοποιήθηκαν είκοσι πέντε (25) δοκιμασίες γραφής που εντάσσονται στις ακόλουθες κατηγορίες:

- **Γραφικές δοκιμασίες:** Ελέγχθηκε η ικανότητα των συμμετεχόντων στη γραφή στοιχειωδών συμβόλων (π.χ. ένωση σημείων και σχεδιασμός γεωμετρικών σχημάτων).
- **Δοκιμασίες αντιγραφής:** Αξιολογήθηκαν οι ικανότητες των συμμετεχόντων στην επανάληψη σύνθετων γραφικών συμβόλων, τα οποία έχουν εννοιολογική σημασία, όπως γράμματα, λέξεις και αριθμοί.
- **Δοκιμασίες μνήμης:** Ελέγχθηκε η αναπαραγωγή συμβόλων που είχαν προηγουμένως απομνημονευτεί ή συνδέονταν με αντικείμενα που απεικονίζονταν σε μία εικόνα.
- **Δοκιμασίες υπαγόρευσης:** Ερευνήθηκε η μεταβολή του τρόπου γραφής στην περίπτωση που χρησιμοποιείται η μνήμη εργασίας.

Το σύνολο δεδομένων περιλαμβάνει 174 συμμετέχοντες, εκ των οποίων οι 89 πάσχουν από τη νόσο Alzheimer. Οι συμμετέχοντες επιλέχθηκαν βάσει τυπικών κλινικών δοκιμασιών, οι οποίες κάνουν χρήση ερωτηματολογίων μέσω των οποίων αξιολογούνται γνωστικές ικανότητες που καλύπτουν ποικίλους τομείς, από χρονικό και χωρικό προσανατολισμό έως ανάκληση καταγραφών. Επιπλέον, αποκλείστηκαν από αυτό άνθρωποι που λαμβάνουν (ψυχοτρόπα) φάρμακα που επηρεάζουν τις γνωστικές ικανότητες αλλά και όσοι παρουσιάζουν μειωμένες γνωστικές ικανότητες, σύμφωνα με την αξιολόγηση ειδικών ιατρών. Ακόμη, προς αποφυγή μεροληψίας, τα γκρουπ υγιών και πασχόντων επιλέχθηκε να είναι ισορροπημένα ως προς την ηλικία, το επίπεδο εκπαίδευσης, τον τύπο εργασίας τους (χειρωνακτική ή πνευματική) και το φύλο των συμμετεχόντων.


Κατά την εκτέλεση των δοκιμασιών, χρησιμοποιήθηκε τάμπλετ εξοπλισμένο με ειδική γραφίδα που επέτρεπε στους συμμετέχοντες να γράφουν σε φύλλα A4 τοποθετημένα πάνω σε αυτό. Το τάμπλετ κατέγραφε τις συντεταγμένες της κίνησης της μύτης της γραφίδας σε συχνότητα 200Hz. Οι συντεταγμένες κατηγοριοποιούνται ανάλογα με το εάν η γραφίδα βρίσκεται σε ε-

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: right;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: right;">Σελίδα 5 / 22</p>

παφή με το χαρτί ή εάν απέχει από αυτό έως και 3cm. Στην πρώτη περίπτωση μάλιστα, καταγράφεται και πίεση που ασκεί η γραφίδα στο χαρτί. Η εκάστοτε δοκιμασία προβάλλονταν στους συμμετέχοντες και το αποτέλεσμα της εκτέλεσής τους απεικονίζονταν σε πραγματικό χρόνο στον υπολογιστή με τον οποίο ήταν συνδεδεμένο το τάμπλετ. Μετά την ολοκλήρωση και των 25 δοκιμασιών (Πίνακας 1) από τον κάθε συμμετέχοντα, τα πρωτογενή δεδομένα καταγράφονταν στο σύνολο δεδομένων.

**Πίνακας 1. Δοκιμασίες που εκτελέστηκαν στο πλαίσιο απόκτησης δεδομένων.**

Αριθμός δοκιμασίας	Περιγραφή δοκιμασίας	Κατηγορία δοκιμασίας
1	Απεικόνιση υπογραφής	Μνήμη και Υπαγόρευση
2	Ένωση δύο σημείων με μια οριζόντια γραμμή, συνεχόμενα για 4 φορές	Γραφική
3	Ένωση δύο σημείων με μια οριζόντια κάθετα, συνεχόμενα για 4 φορές	Γραφική
4	Σχεδιασμός κύκλου (με διάμετρο 6 εκ.) συνεχόμενα για 4 φορές	Γραφική
5	Σχεδιασμός κύκλου (με διάμετρο 3 εκ.) συνεχόμενα για 4 φορές	Γραφική
6	Αντιγραφή των γραμμάτων “l”, “m” και “p”	Αντιγραφή
7	Αντιγραφή γραμμάτων σε γειτονικές γραμμές	Αντιγραφή
8	Σχηματισμός μίας ακολουθίας τεσσάρων γραμμάτων “l” με μία συνεχόμενη κίνηση	Αντιγραφή
9	Σχηματισμός μίας ακολουθίας τεσσάρων διγραμμάτων “le” με μία συνεχόμενη κίνηση	Αντιγραφή
10	Αντιγραφή της λέξης “foglio”	Αντιγραφή
11	Αντιγραφή της λέξης “foglio” πάνω σε μία γραμμή	Αντιγραφή
12	Αντιγραφή της λέξης “mamma”	Αντιγραφή
13	Αντιγραφή της λέξης “mamma ” πάνω σε μία γραμμή	Αντιγραφή
14	Απομνημόνευση των λέξεων “telefono”, “cane” και “negozio” και συγγραφή τους	Μνήμη και Υπαγόρευση
15	Αντιγραφή της λέξης “bottiglia” αντίστροφα	Αντιγραφή
16	Αντιγραφή της λέξης “casa” αντίστροφα	Αντιγραφή
17	Αντιγραφή έξι λέξεων σε αντίστοιχα κουτιά	Αντιγραφή
18	Συγγραφή του ονόματος ενός αντικειμένου που φαίνεται σε μία εικόνα (μία καρτέλα)	Μνήμη και Υπαγόρευση
19	Αντιγραφή των πεδίων μίας ταχυδρομικής επιταγής	Αντιγραφή


	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	ΙΟΥΛΙΟΣ 2024
		Σελίδα 6 / 22

Αριθμός δοκιμασίας	Περιγραφή δοκιμασίας	Κατηγορία δοκιμασίας
20	Συγγραφή μίας πρότασης μέσω υπαγόρευσης	Μνήμη και Υπαγόρευση
21	Σχεδιασμός ενός περίπλοκου σχήματος	Γραφική
22	Αντιγραφή ενός τηλεφωνικού αριθμού	Αντιγραφή
23	Συγγραφή ενός τηλεφωνικού αριθμού μέσω υπαγόρευσης	Μνήμη και Υπαγόρευση
24	Σχεδίαση ενός ρολογιού, με τους δείκτες να δείχνουν στις 11:05	Γραφική
25	Αντιγραφή μίας παραγράφου	Αντιγραφή

Για κάθε δοκιμασία, εξήχθησαν δεκαοχτώ (18) χαρακτηριστικά βάσει των πρωτογενών δεδομένων (συντεταγμένες, πίεση, χρόνος). Τα χαρακτηριστικά παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 2):

**Πίνακας 2. Τα 18 χαρακτηριστικά που εξήχθησαν για κάθε δοκιμασία (Cilia, N. et al, 2022).**


Α/Α	Χαρακτηριστικό
1	Συνολικός χρόνος εκτέλεσης της εργασίας
2	Χρόνος εκτέλεσης κινήσεων στον αέρα
3	Χρόνος εκτέλεσης κινήσεων στο χαρτί
4	Μέση ταχύτητα κινήσεων στο χαρτί
5	Μέση ταχύτητα κινήσεων στον αέρα
6	Μέση επιτάχυνση κινήσεων στο χαρτί
7	Μέση επιτάχυνση κινήσεων στον αέρα
8	Μέσος ρυθμός μεταβολής της επιτάχυνσης κινήσεων στο χαρτί
9	Μέσος ρυθμός μεταβολής της επιτάχυνσης κινήσεων στον αέρα
10	Μέση πίεση που ασκείται από τη γραφίδα
11	Διακύμανση πίεσης που ασκείται από τη γραφίδα
12	Μέσο σχετικό τρέμουλο κινήσεων στο χαρτί

	<p><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΓΡΑΦΗΣ</b></p> <p>ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p>ΙΟΥΛΙΟΣ 2024</p>
		<p>Σελίδα 7 / 22</p>

Α/Α	Χαρακτηριστικό
13	Μέσο σχετικό τρέμουλο κινήσεων στον αέρα
14	Μέσο τρέμουλο (στο χαρτί και στον αέρα)
15	Αριθμός που διακόπηκε η κίνηση της γραφής κατά την εκτέλεση της εργασίας
16	Μέγιστη έκταση γραφθέντων στον άξονα Χ
17	Μέγιστη έκταση γραφθέντων στον άξονα Υ
18	Δείκτης βαθμού κάλυψης του χαρτιού

Τέλος, το σύνολο δεδομένων αποτελείται από 451 χαρακτηριστικά (18 \* 25 για τις δοκιμασίες και τη στήλη ID) και από μία στήλη που περιέχει τον χαρακτηρισμό του κάθε συμμετέχοντα σχετικά με το εάν πάσχει (P) ή όχι (H) από Alzheimer, ενώ δεν παρουσιάζει απουσιάζουσες τιμές.



	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: center;">Σελίδα 8 / 22</p>


### 3 ΝΟΣΟΣ ALZHEIMER ΚΑΙ ΚΛΙΝΙΚΟΣ ΑΝΤΙΚΤΥΠΟΣ

Η νόσος Alzheimer προκαλεί αργή και προοδευτική βλάβη σε νοητικές λειτουργίες, όπως η μνήμη, η σκέψη, η κριτική ικανότητα και άλλες μαθησιακές ικανότητες. Στα πρώιμα στάδια της νόσου, το κυρίαρχο σύμπτωμα είναι η επεισοδική διαταραχή της μνήμης, η οποία αποτελεί ένδειξη δυσλειτουργίας του κροταφικού λοβού (Cilia, N. et. al, 2022). Συνήθως, το στάδιο αυτό ακολουθείται από προοδευτική αμνησία και επιδείνωση σε περαιτέρω γνωστικούς τομείς, φανερώνοντας την παθολογική εμπλοκή ευρύτερων νευρικών συστημάτων.

Η νόσος Alzheimer δεν έχει θεραπεία και η επιβράδυνση της εξέλιξής της θεωρείται ο καλύτερος δυνατός τρόπος αντιμετώπισής της. Με την παγκόσμια αύξηση του προσδόκιμου ζωής, αναμένεται να αυξηθεί σημαντικά η εμφάνισή της σε απόλυτους αριθμούς μέσα στις επόμενες δεκαετίες. Η εκτίμηση αυτή οδηγεί σε επιτακτική ανάγκη για τη βελτίωση των σημερινών προσεγγίσεων για την έγκαιρη διάγνωση της νόσου.

Λόγω του γεγονότος ότι τόσο οι γνωστικές όσο και οι κινητικές λειτουργίες εμπλέκονται στον σχεδιασμό και την εκτέλεση των κινήσεων, καθώς και επειδή η γραφή απαιτεί ακριβή και απόλυτα συντονισμένο έλεγχο του σώματος, η ανάλυση του γραφικού χαρακτήρα μπορεί να παρέχει μια φθηνή και μη επεμβατική μέθοδο αξιολόγησης της εξέλιξης της νόσου.

Υπό αυτό το πρίσμα, η εφαρμογή της Μηχανικής Μάθησης μπορεί να συμβάλει σημαντικά στη μείωση του απαιτούμενου χρόνου κλινικής διάγνωσης της νόσου και κατά συνέπεια στην έναρξη της αντιμετώπισης από τα πολύ πρώιμα στάδια.

	<p><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΓΡΑΦΗΣ</b></p> <p>ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	ΙΟΥΛΙΟΣ 2024
		Σελίδα 9 / 22

## 4 ΜΕΘΟΔΟΛΟΓΙΑ ΠΡΟΣΕΓΓΙΣΗΣ

Στόχο της παρούσας εργασίας αποτέλεσε τόσο η ανάπτυξη του βέλτιστου δυνατού μοντέλου μηχανικής μάθησης που θα προβλέπει τη νόσηση ανθρώπων από Alzheimer μέσω της γραφής τους, όσο και η κατανόηση της σπουδαιότητας καθεμιάς από τις δοκιμασίες στην ανάπτυξη του παραπάνω μοντέλου.


Για αυτό τον σκοπό, η μεθοδολογία που ακολουθήθηκε κινήθηκε σε τρεις βασικούς άξονες. Το πρώτο από αυτά περιλαμβάνει την ανάλυση ολόκληρου του συνόλου δεδομένων, ενώ το δεύτερο επικεντρώνεται ξεχωριστά στην ανάλυση καθεμιάς από τις 25 δοκιμασίες. Ο τρίτος άξονας λαμβάνει υπόψη του τις πιο χρήσιμες για την περίπτωση δοκιμασίες, όπως αυτές προέκυψαν από τα αποτελέσματα του δεύτερου άξονα και δημιουργεί ένα μοντέλο από αυτές, τα αποτελέσματα του οποίου θα συγκριθούν με αυτά του αντίστοιχου μοντέλου που θα προκύψει από τις λιγότερο χρήσιμες δοκιμασίες.

Σε όλες τις περιπτώσεις, τα μοντέλα αξιολογήθηκαν με βάση τη μετρική της ακρίβειας (accuracy), ενώ καταγράφηκε ενισχυτικά και αυτή της ευαισθησίας (sensitivity) που είναι κρίσιμη για την περίπτωση ιατρικών δεδομένων.

### 4.1 ΑΝΑΛΥΣΗ ΟΛΟΚΛΗΡΟΥ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Για την ανάδειξη της δομής του συνόλου δεδομένων DARWIN, δημιουργείται ένα DataFrame από τα διαθέσιμα δεδομένα και στη συνέχεια εκτυπώνονται οι διαστάσεις του. Αυτό επιτρέπει την ύπαρξη σαφούς εικόνας για το μέγεθος και τη μορφή των δεδομένων που θα χρησιμοποιηθούν στην ανάλυσή μας, εξασφαλίζοντας έτσι μια πιο οργανωμένη και αποδοτική προσέγγιση στην έρευνα.

Η ύπαρξη τόσο πολλών χαρακτηριστικών (452 συνολικά) αποτελεί μια πρόκληση από πολλές απόψεις. Αρχικά, η διαχείριση και ανάλυσή τους αυξάνει την ανάγκη για υπολογιστικούς πόρους και μνήμη. Η ανάλυση δεδομένων υψηλής διάστασης μπορεί να είναι πολύπλοκη και χρονοβόρα, καθώς αυξάνει τον αριθμό των πιθανών συνδυασμών και αλληλεπιδράσεων μεταξύ των χαρακτηριστικών.

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: right;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: right;">Σελίδα 10 / 22</p>

Έπειτα, η παρουσία μεγάλου αριθμού χαρακτηριστικών αυξάνει τον κίνδυνο της υπερπροσαρμογής (overfitting) στα μοντέλα μηχανικής μάθησης. Για την αντιμετώπιση του προβλήματος, συχνά απαιτούνται τεχνικές όπως η επιλογή χαρακτηριστικών (feature selection) ή η μείωση διαστάσεων (dimensionality reduction), όπως η ανάλυση κύριων συνιστωσών (PCA), προκειμένου να περιοριστεί ο αριθμός των χαρακτηριστικών σε ένα πιο διαχειρίσιμο σύνολο.

Επιπλέον, η ύπαρξη τόσων πολλών στηλών μπορεί να περιπλέξει την ερμηνεία των αποτελεσμάτων. Η κατανόηση του τρόπου επιρροής κάθε χαρακτηριστικού στο αποτέλεσμα ή την πρόβλεψη μπορεί να είναι δύσκολη και η διαδικασία εξαγωγής συμπερασμάτων απαιτεί προσεκτική ανάλυση και συχνά εξειδικευμένες τεχνικές οπτικοποίησης.


Το σύνολο δεδομένων περιλαμβάνει τρεις (3) τύπους δεδομένων:

- 300 στήλες με δεδομένα τύπου float, που αντιπροσωπεύουν μετρήσεις γραφής.
- 100 στήλες με δεδομένα τύπου object, που περιλαμβάνουν κατηγοριοποιήσεις ή περιγραφικές πληροφορίες.
- 52 στήλες με δεδομένα τύπου int, που αναφέρονται σε αριθμητικούς δείκτες ή μετρήσεις.

Το σύνολο δεδομένων δεν περιέχει ελλιπή δεδομένα (missing values). Αυτό είναι εξαιρετικά ευεργετικό, καθώς απλοποιεί τη διαδικασία της ανάλυσης και της προετοιμασίας των δεδομένων. Επιτρέπει την απ' ευθείας εστίαση στην ανάπτυξη και δοκιμή των μοντέλων μηχανικής μάθησης χωρίς να χρειάζεται να αντιμετωπιστούν προβλήματα που σχετίζονται με την απουσία δεδομένων.

Πρόκειται για πρόβλημα δυαδικής ταξινόμησης, με στόχο τη διάκριση μεταξύ υγιών και πασχόντων από τη νόσο Alzheimer. Το σύνολο δεδομένων είναι καλά ισορροπημένο, διευκολύνοντας την εκπαίδευση των μοντέλων και βελτιώνοντας την ακρίβεια των αποτελεσμάτων.

Κατά την ανάλυση του συνόλου δεδομένων, δημιουργήθηκε ένας πίνακας συσχέτισης για την εξέταση της σχέσης των χαρακτηριστικών με την κλάση "class". Διαπιστώθηκε ότι 43 στήλες έχουν συσχέτιση μικρότερη ή ίση με 0.05 με αυτή, υποδεικνύοντας ότι αυτές οι στήλες είναι

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: center;">Σελίδα 11 / 22</p>

πιθανόν λιγότερο σημαντικές για την ανάπτυξη αποτελεσματικών μοντέλων. Επιπλέον, βρέθηκαν 34 ζεύγη χαρακτηριστικών με πολύ υψηλή συσχέτιση μεταξύ τους, μεγαλύτερη από 0.99. Αυτή η ανάλυση βοηθά στον καθορισμό των πιο σχετικών χαρακτηριστικών για τη βελτίωση της ακρίβειας των προβλέψεων.


Αποπειράθηκε διαχωρισμός των δεδομένων με δύο τρόπους: αρχικά, με **τυχαίο διαχωρισμό (Random Split)**, όπου το 70% χρησιμοποιήθηκε για εκπαίδευση και το 30% για επαλήθευση, διατηρώντας την τυχαιότητα μέσω καθορισμένου seed. Στη συνέχεια, εφαρμόστηκε η μέθοδος **Stratified Shuffle Split**, η οποία τελικά χρησιμοποιήθηκε για την πραγματοποίηση κάθε εκτέλεσης, διασφαλίζοντας ότι η αναλογία των κλάσεων στο σύνολο εκπαίδευσης και στο σύνολο επαλήθευσης παραμένει αντιπροσωπευτική του αρχικού συνόλου. Το 30% των δεδομένων διατέθηκε για επαλήθευση, με χρήση τυχαίου seed για την αναπαραγωγή των αποτελεσμάτων και τη διατήρηση της τυχαιότητας.

Για την κανονικοποίηση των χαρακτηριστικών, επιλέχθηκε ο **Min-Max Scaler**. Αυτή η μέθοδος μετατρέπει τις τιμές των χαρακτηριστικών στο εύρος [0, 1], διευκολύνοντας τη σύγκριση των δεδομένων και την εκπαίδευση των μοντέλων. Η επιλογή του Min-Max Scaler έγινε λόγω της μη ύπαρξης απουσιαζουσών τιμών, γεγονός που επιτρέπει σταθερότητα και αξιοπιστία στη διαδικασία κανονικοποίησης.

Ένας από τους στόχους αυτής της εργασίας είναι η εύρεση του βέλτιστου μοντέλου πρόβλεψης με τις κατάλληλες υπερπαραμέτρους. Για την επίτευξή του, εξετάστηκαν διαφορετικά μοντέλα και ρυθμίσεις, στοχεύοντας στη βελτιστοποίηση της ακρίβειας και της αξιοπιστίας των προβλέψεων.

Ο τρόπος για την εύρεση των βέλτιστων υπερπαραμέτρων και, εκ των υστέρων, του βέλτιστου μοντέλου σε κάθε προσέγγιση πραγματοποιήθηκε μέσω GridSearchCV. Το GridSearchCV εφαρμόστηκε με διάφορες σταυρωτές επικυρώσεις (όπως Leave-One-Out ή 5-fold CV), και το καλύτερο μοντέλο επιλέχθηκε με βάση τη μετρική της ακρίβειας (accuracy).

Οι υπερπαραμέτροι που εξετάστηκαν στο παρόν στάδιο, παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 3). Η διαδικασία αυτή παρείχε μια ολοκληρωμένη σύγκριση και βελτιστοποίηση για κάθε ταξινομητή, επιτρέποντας την επιλογή του καλύτερου συνδυασμού υπερπαραμέτρων με βάση τη μετρική της ακρίβειας (accuracy).

	<b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b> <b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b> <b>ΓΡΑΦΗΣ</b>  ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 12 / 22

**Πίνακας 3. Υπερπαράμετροι προς βελτιστοποίηση που αναζητήθηκαν κατά την ανάπτυξη του μοντέλου.**

Ταξινομητής	Υπερπαράμετρος
KNN	'n_neighbors', 'weights', 'algorithm', 'metric'
Random Forest	'n_estimators', 'max_depth', 'min_samples_split', 'min_samples_leaf', 'class_weight'
Logistic Regression	'Penalty', 'C', 'solver', 'class_weight'
Decision Tree	'max_depth', 'min_samples_split', 'min_samples_leaf', 'class_weight'
Naive Bayes	-
MLP	'hidden_layer_sizes', 'alpha', 'solver', 'max_iter'
SVM	'Kernel', 'gamma', 'C', 'class_weight'

Πέρα από τη στατιστική ανάλυση για τη μείωση των χαρακτηριστικών, πραγματοποιήθηκε και ανάλυση κύριων συνιστωσών (PCA). Η PCA είναι μια τεχνική μείωσης διαστάσεων (μέσω μετασχηματισμού τους), που χρησιμοποιείται για να μειώσει τον αριθμό των χαρακτηριστικών, διατηρώντας όσο το δυνατόν περισσότερη από την αρχική πληροφορία. Η PCA λειτουργεί με τον προσδιορισμό των κατευθύνσεων (συνιστωσών) που μεγιστοποιούν τη διακύμανση στα δεδομένα και στη συνέχεια με την προβολή των δεδομένων πάνω σε αυτές τις κατευθύνσεις.

Για τον προσδιορισμό του ιδανικού αριθμού κύριων συνιστωσών, χρησιμοποιήθηκε η μέθοδος του αγκώνα (elbow). Η μέθοδος του αγκώνα είναι μια οπτική τεχνική που βοηθά στον εντοπισμό του σημείου, έπειτα από το οποίο η προσθήκη επιπλέον συνιστωσών δεν προσφέρει σημαντική αύξηση της εξηγούμενης διακύμανσης. Έτσι, δημιουργήθηκε ένα γράφημα που απεικονίζει την εξηγούμενη διακύμανση σε σχέση με τον αριθμό των κύριων συνιστωσών και αναζητήθηκε το σημείο όπου η καμπύλη αρχίζει να γίνεται πιο επίπεδη, σχηματίζοντας έναν "αγκώνα". Μετά από αυτή την ανάλυση, προέκυψε ότι ο ιδανικός αριθμός κύριων συνιστωσών είναι αυτός που διατηρεί το 97% της συνολικής διακύμανσης. Για να διατηρηθεί το 97% , χρειάζονται 87 συνιστώσες. Αυτή η προσέγγιση επέτρεψε την περαιτέρω μείωση της πολυπλοκότητας των δεδομένων, εξασφαλίζοντας ταυτόχρονα τη διατήρηση της μέγιστης δυνατής πληροφορίας από το αρχικό σύνολο δεδομένων.

Επιπλέον, εφαρμόστηκε η ίδια διαδικασία με στόχο τη διατήρηση του 96% της συνολικής διακύμανσης. Σε αυτό το πλαίσιο, υπολογίστηκε ότι απαιτούνται 82 συνιστώσες.

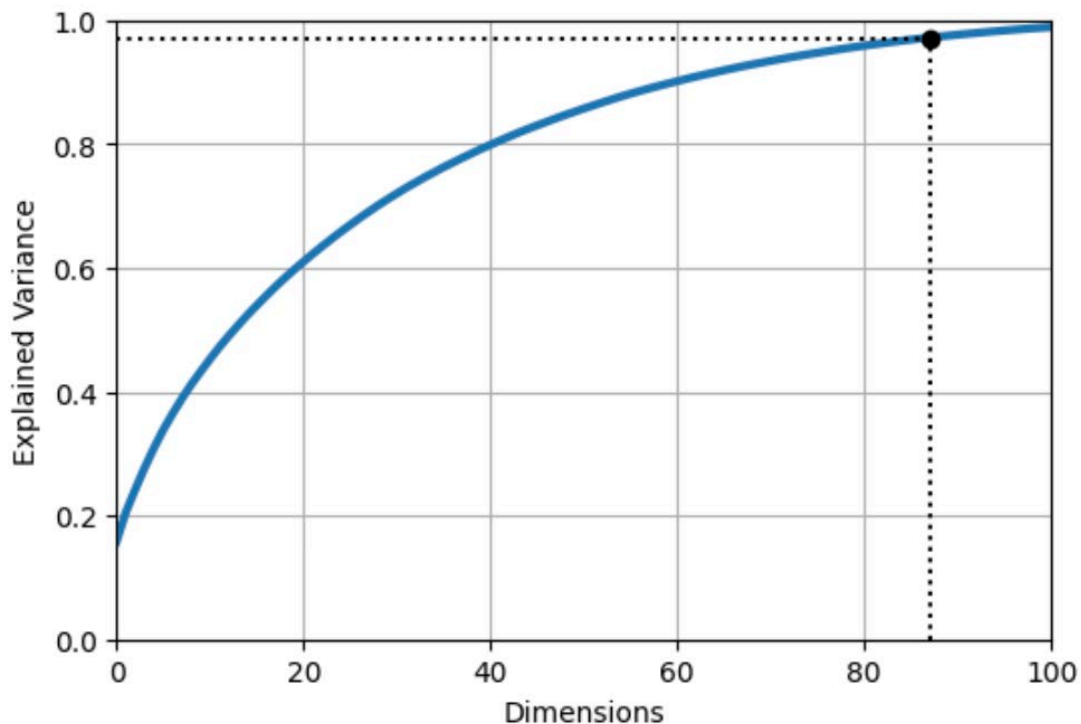


ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ  
ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ  
ΓΡΑΦΗΣ

ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ

ΙΟΥΛΙΟΣ 2024

Σελίδα 13 / 22




Σχήμα 1. Εξηγημένη διακύμανση – Διαστάσεις (elbow στο 0.97).

Σε μια προσπάθεια ανάπτυξης βελτιωμένων μοντέλων εφαρμόστηκε η τεχνική Mixup για την επαύξηση των δειγματικών στοιχείων του συνόλου εκπαίδευσης. Πρόκειται για μια μέθοδο εκπαίδευσης στην οποία εισάγονται νέα δείγματα δεδομένων από τον συνδυασμό δύο ή περισσότερων δειγμάτων εκπαίδευσης, συνήθως με τρόπο γραμμικό.

Δοκιμάσαμε δύο προσεγγίσεις με τη μέθοδο mixup στο σύνολο εκπαίδευσης. Στην πρώτη προσέγγιση, χρησιμοποιήθηκε συντελεστής  $\alpha=0,2$ , ο οποίος ελέγχει την κατανομή της παραμέτρου μίξης και καθορίζει τον βαθμό συνδυασμού των δειγμάτων δεδομένων. Πραγματοποιήθηκε εκπαίδευση μοντέλων με αυτή την προσέγγιση, εφαρμόζοντας mixup και PCA για διατήρηση του 97% της διακύμανσης.

Στη δεύτερη προσέγγιση, χρησιμοποιήθηκε συντελεστής  $\alpha=1$ , που αντιστοιχεί στην απλή αντιγραφή των εγγραφών του συνόλου εκπαίδευσης. Με τον ίδιο τρόπο όπως και παραπάνω

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: right;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: right;">Σελίδα 14 / 22</p>

πραγματοποιήθηκε εκπαίδευση μοντέλων με αυτή την προσέγγιση, εφαρμόζοντας mixup και PCA για διατήρηση του 97% της διακύμανσης.


Τέλος, καθώς παρά το γεγονός ότι ο αριθμός των χαρακτηριστικών είναι σημαντικά μεγάλος, επιλέχθηκε λόγω του σχετικά μικρού αριθμού εγγραφών, να αναπτυχθεί και ένα μοντέλο από ολόκληρο το σύνολο δεδομένων, χωρίς να εφαρμοστεί αφαίρεση χαρακτηριστικών ή κάποια μέθοδος μείωσης των διαστάσεων του (PCA). Η προσέγγιση που ακολουθήθηκε σε αυτή την περίπτωση (διαχωρισμός συνόλων εκπαίδευσης και ελέγχου, κανονικοποίηση, υπερπαραμέτροι που αναζητήθηκαν κλπ.) είναι αντίστοιχη με όσα περιγράφονται παραπάνω.

## 4.2 ΑΝΑΛΥΣΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΑΝΑ ΔΟΚΙΜΑΣΙΑ

Κατά τον δεύτερο άξονα της εργασίας, διερευνήθηκε κάθε δοκιμασία ξεχωριστά με στόχο την ανίχνευση ενός συνόλου από αυτά που θα προσεγγίζουν τη διάγνωση του εξεταζόμενου κατά το δυνατό καλύτερα. Το μικρότερο σύνολο δοκιμασιών θα δώσει τη δυνατότητα να σχεδιαστεί μια νέα διαδικασία εξέτασης που θα απαιτεί σημαντικά λιγότερο χρόνο για την ολοκλήρωσή της. Είναι ιδιαίτερα σημαντικό, η προσέγγιση των εξεταζόμενων να γίνεται με τέτοιο τρόπο ώστε να διατηρείται ακέραια η θέλησή τους για συμμετοχή. Μόνο σε αυτή τη περίπτωση μπορεί να διασφαλιστεί πως η επίδοσή τους θα είναι αδιάβλητη.

Η διαδικασία που ακολουθήθηκε είναι η εξής: Διαχωρίστηκε το πλήρες DataFrame σε 25 επιμέρους DataFrames, καθένα από τα οποία διατηρεί τις 18 στήλες που αφορούν σε μία δοκιμασία. Στη συνέχεια, διέτρεξε μια επαναληπτική διαδικασία για την αναζήτηση του βέλτιστου μοντέλου σε κάθε δοκιμασία. Συγκεκριμένα, για κάθε δοκιμασία και κάθε επανάληψη, διαχωρίστηκαν τα δεδομένα σε σύνολα ελέγχου και επικύρωσης με αντίστοιχες ποσοστώσεις 75% και 25%. Επιλέχθηκε αυτός ο διαχωρισμός να εφαρμόζεται σε κάθε επανάληψη, ώστε τα αποτελέσματα που θα παρουσιαστούν στη συνέχεια να είναι εύρωστα. Επιπλέον, διασφαλίστηκε ότι η αναλογία των κλάσεων στο σύνολο εκπαίδευσης και στο σύνολο επικύρωσης θα παραμείνει αντιπροσωπευτική του αρχικού συνόλου μέσω **Stratified Shuffle Split**. Στη συνέχεια, μετασχηματίστηκαν τα δεδομένα χρησιμοποιώντας τον **Min-Max Scaler**. Μέσω χρήσης **GridSearchCV** στο σύνολο δεδομένων εκπαίδευσης, αναζητήθηκαν οι βέλτιστες υπερπαραμέτροι για τα μοντέλα: **SVM**, **Random Forest**, **Logistic Regression**, **MLP** και **XGB**. Το τελευταίο βήμα είναι η αξιολόγηση κάθε μοντέλου μέσα από το σύνολο επικύρωσης. Σημειώνεται



	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: center;">Σελίδα 15 / 22</p>

πως το πλήθος των επαναλήψεων που επιλέχθηκαν για αυτή τη διαδικασία ήταν πέντε (5), καθώς η επιλογή μικρότερου πλήθους δεν θα έδινε αρκετή πληροφορία για την απόδοση των μοντέλων, ενώ η επιλογή μεγαλύτερου πλήθους θα καθιστούσε την υλοποίηση της διαδικασίας εξαιρετικά χρονοβόρα.

#### **4.3 ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΩΝ ΑΠΟ ΤΙΣ ΠΙΟ ΧΡΗΣΙΜΕΣ ΚΑΙ ΤΙΣ ΛΙΓΟΤΕΡΟ ΧΡΗΣΙΜΕΣ ΔΟΚΙΜΑΣΙΕΣ**

Σύμφωνα με τα αποτελέσματα που προέκυψαν από τα μοντέλα των επί μέρους δοκιμασιών, αυτές κατατάσσονται ως προς τη συμβολή τους στην επίτευξη του καλύτερου μοντέλου βάσει της μετρικής της ακρίβειας. Έτσι, στην προσπάθεια επιβεβαίωσης του παραπάνω αποτελέσματος αλλά και επίτευξης ακόμη μεγαλύτερης ακρίβειας, αποφασίστηκε η ανάπτυξη ενός μοντέλου βάσει των δεδομένων που προήλθαν από τις δώδεκα πιο «καθοριστικές» δοκιμασίες. Τα αποτελέσματά του θα συγκριθούν με αυτά που θα προκύψουν από το αντίστοιχο μοντέλο των δώδεκα «λιγότερο καθοριστικών» δοκιμασιών. Καθώς στα μοντέλα που αναπτύχθηκαν στο πλαίσιο της παραγράφου 4.2, ο καλύτερος ανά περίπτωση ταξινομητής είναι (αναμενόμενα) διαφορετικός, εδώ επιλέγεται να χρησιμοποιηθεί ο SVC και στις δύο περιπτώσεις, ο οποίος είναι ο πιο αποδοτικός για την πλειοψηφία των προαναφερθέντων μοντέλων. Συμπληρωματικά, επιλέγεται και ο RandomForest καθώς είναι αυτός που έδωσε τα καλύτερα αποτελέσματα για τα μοντέλα της παραγράφου 4.1. Η προετοιμασία και εκπαίδευση των μοντέλων κατά τα λοιπά γίνεται σε αντιστοιχία με τα μοντέλα της παραγράφου 4.1. Δεν εφαρμόστηκε η τεχνική PCA ούτε επαύξηση δεδομένων σε αυτή την περίπτωση.



5 ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 ΜΟΝΤΕΛΑ ΑΠΟ ΟΛΟΚΛΗΡΟ ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

Οι επιδόσεις των μοντέλων που περιγράφονται στην παράγραφο 4.1 παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 4).

Πίνακας 4. Επίδοση (accuracy) μοντέλων με απομείωση – μετασχηματισμό διαστάσεων.

Ταξινομη- τής	Εξαγωγή Χαρακτηριστικών				PCA			
	Ασθενής Συσχέ- τιση με κλάση		Ισχυρή Συσχέτιση χα- κτηριστικών μεταξύ τους		Διατήρηση διακύ- μανσης 97%		Διατήρηση διακύ- μανσης 96%	
	5 folds	One- leave-out	5 folds	One-leave- out	5 folds	One- leave-out	5 folds	One- leave-out
KNN	0.792	0.754	0.698	0.754	0.698	0.698	0.679	0.679
Random Forest	0.887	0.867	0.811	0.792	0.792	0.83	0.754	0.754
Logistic Re- gression	0.773	0.509	0.792	0.509	0.491	0.491	0.491	0.491
Decision Tree	0.735	0.716	0.641	0.698	0.641	0.811	0.773	0.773
Naive Bayes	0.867	0.867	0.886	0.886	0.773	0.641	0.698	0.698
MLP	0.773	0.754	0.773	0.754	0.773	0.773	0.754	0.754
SVM	0.792	0.509	0.754	0.509	0.679	0.679	0.660	0.660

Στη διαδικασία βελτιστοποίησης, αναζητήθηκαν βέλτιστες υπερπαραμέτροι με χρήση GridSearchCV. Οι αλγόριθμοι Random Forest και Naive Bayes ξεχώρισαν για την υψηλή ακρίβεια. Ο Random Forest απέδωσε εξαιρετικά στην ασθενή συσχέτιση με την κλάση, επιτυγχάνοντας ακρίβεια 0.886 σε 5-fold cross-validation και 0.867 με Leave-One-Out. Ο Naive Bayes επίσης έδειξε υψηλές επιδόσεις. Συνολικά, η ανάλυση οδήγησε σε αποτελεσματικά μοντέλα με ακρίβεια και μειωμένη πολυπλοκότητα. Στην περίπτωση της 97% διατήρησης της συνολικής διακύμανσης προέκυψαν καλύτερα αποτελέσματα από την 96% διατήρηση της διακύμανσης.

**Πίνακας 5. Επίδοση (accuracy) μοντέλων με επαύξηση δεδομένων και PCA στο 97% της συνολικής διακύμανσης.**


Ταξινομητής	MixUp με $\alpha=0.2$	MixUp με $\alpha=1$
kNN	0.773	0.698
Random Forest	0.811	0.735
Logistic Regression	0.773	0.509
Decision Tree	0.716	0.811
Naive Bayes	0.868	0.641
MLP	0.773	0.754
SVM	0.781	0.660

Χρησιμοποιώντας PCA για διατήρηση του 97% της διακύμανσης και Grid Search με 5-fold cross-validation, βρέθηκαν οι βέλτιστες υπερπαραμέτροι. Τα αποτελέσματα προέκυψαν από την προσθήκη δειγματικών στοιχείων μέσω mixup. Ο καλύτερος ταξινομητής ήταν ο Naive Bayes, με ακρίβεια 0.868 για συντελεστή mixup  $\alpha=0.2$ . Στη δεύτερη προσέγγιση, με συντελεστή mixup  $\alpha=1$ , δηλαδή αντιγραφή του συνόλου εκπαίδευσης, ο Decision Tree αναδείχθηκε καλύτερος, με ακρίβεια 0.811.

Τα αποτελέσματα της αξιολόγησης των μοντέλων που εκπαιδεύθηκαν στο σύνολο των διαθέσιμων δεδομένων για κάθε ταξινομητή, παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 6).

**Πίνακας 6. Αποτελέσματα ακρίβειας (accuracy) – Μοντέλα εκπαιδευμένα στο σύνολο των διαθέσιμων δεδομένων.**

Ταξινομητής	Επαύξηση δεδομένων	Ακρίβεια (accuracy)
kNN	OXI	0,811
Random Forest		0,868
Logistic Regression		0,830

	<b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b> <b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b> <b>ΓΡΑΦΗΣ</b> ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 18 / 22

Ταξινομητής	Επαύξηση δεδομένων	Ακρίβεια (accuracy)
Decision Tree		0,830
Naive Bayes		0,774
MLP		0,830
SVC		0,774
Random Forest	NAI	0,906


Από τα παραπάνω αποτελέσματα, γίνεται φανερό ότι για το συγκεκριμένο πρόβλημα και το συγκεκριμένο πρόβλημα, η εκπαίδευση στο σύνολο των διαθέσιμων δεδομένων οδηγεί σε καλύτερα αποτελέσματα από τα μοντέλα που εφαρμόζουν τεχνικές μείωσης των διαστάσεων (PCA). Πιο συγκεκριμένα, έπειτα από την επαύξηση δεδομένων επιτυγχάνεται από ακρίβεια 0.906. Αυτή είναι και η υψηλότερη τιμή που επιτυγχάνεται στο πλαίσιο της παρούσας εργασίας.

## 5.2 ΜΟΝΤΕΛΑ ΑΠΟ ΤΟΝ ΔΙΑΧΩΡΙΣΜΟ ΤΩΝ ΕΓΓΡΑΦΩΝ ΑΝΑ ΔΟΚΙΜΑΣΙΑ

Τα αποτελέσματα της διαδικασίας που περιγράφεται στην παράγραφο 4.2 παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 7):

**Πίνακας 7. Σειρά κατάταξης μοντέλων, εκπαιδευμένων ανά δοκιμασία.**

Κατάταξη	Αριθμός δοκιμασίας	Ταξινομητής	Μ.Ο. Ακρίβειας	Εύρος
1	10	SVM	0.89	0.87 - 0.91
2	11	SVM	0.85	0.78 - 0.90
3	9	SVM	0.82	0.74 - 0.90
4	17	XGB	0.81	0.77 - 0.87
5	15	SVM	0.80	0.78 - 0.89
6	13	SVM	0.80	0.73 - 0.91
7	23	XGB	0.77	0.69 - 0.82
8	14	SVM	0.76	0.68 - 0.80
9	16	SVM	0.76	0.68 - 0.86
10	4	SVM	0.76	0.52 - 0.91


	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: center;">Σελίδα 19 / 22</p>

Κατάταξη	Αριθμός δοκιμασίας	Ταξινομητής	Μ.Ο. Ακρίβειας	Εύρος
11	3	SVM	0.76	0.65 - 0.81
12	18	SVM	0.75	0.68 - 0.87
13	21	XGB	0.75	0.70 - 0.83
14	24	Logistic Regression	0.74	0.69 - 0.81
15	2	SVM	0.74	0.68 - 0.89
16	1	SVM	0.74	0.67 - 0.80
17	12	SVM	0.72	0.70 - 0.78
18	5	XGB	0.72	0.65 - 0.78
19	22	Random Forest	0.72	0.68 - 0.76
20	7	Logistic Regression	0.71	0.65 - 0.80
21	6	Random Forest	0.70	0.64 - 0.78
22	20	SVM	0.69	0.57 - 0.73
23	25	MLP	0.69	0.61 - 0.78
24	19	Logistic Regression	0.69	0.65 - 0.70
25	8	MLP	0.66	0.64 - 0.77

Όσον αφορά στο τεχνικό κομμάτι της ανάλυσης, τονίζονται δύο σημεία. Αρχικά, τα αποτελέσματα της ακρίβειας (accuracy) των βελτιστοποιημένων μοντέλων, δηλαδή αυτών με τις υπερ-παραμέτρους που προέκυψαν από το GridSearchCV, στις πρώτες δέκα (10) δοκιμασίες ξεπερνούν το 76% παρέχοντας πολύτιμη πληροφορία για τον γιατρό που θα θελήσει να τα συμβουλευτεί. Επιπλέον, παρατηρείται η επικράτηση του μοντέλου SVM, καθώς σε 14 από τα 25 μοντέλα εμφάνισε μεγαλύτερη προβλεπτική ικανότητα σε σχέση με τους υπόλοιπους ταξινομητές.

### 5.3 ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ ΜΕΤΑΞΥ ΠΕΡΙΣΣΟΤΕΡΟ ΚΑΙ ΛΙΓΟΤΕΡΟ ΧΡΗΣΙΜΩΝ ΔΟΚΙΜΑΣΙΩΝ


Ο Πίνακας 8 που ακολουθεί παρουσιάζει τα αποτελέσματα που προέκυψαν για τα μοντέλα που αναπτύχθηκαν βασιζόμενα στις περισσότερες και λιγότερες χρήσιμες δοκιμασίες, όπως αυτές προέκυψαν από τα αποτελέσματα των μοντέλων της παραγράφου 5.2.

	<b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b> <b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b> <b>ΓΡΑΦΗΣ</b> ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ	ΙΟΥΛΙΟΣ 2024
		Σελίδα 20 / 22

**Πίνακας 8. Αξιολόγηση (accuracy) μοντέλων βασιζόμενων σε 12 από τις 25 δοκιμασίες.**

	<b>Μοντέλο - 12 Πιο Χρήσιμες Δοκιμασίες</b>	<b>Μοντέλο - 12 Λιγότερο χρήσιμες Δοκιμασίες</b>
<b>Random Forest</b>	0,755	0,849
<b>SVC</b>	0,774	0,811

Από τα παραπάνω αποτελέσματα, γίνεται αντιληπτό ότι οι δοκιμασίες που προέκυψαν ως λιγότερο σημαντικές όταν αξιολογήθηκαν ατομικά, οδηγούν σε μοντέλα υψηλότερης ακρίβειας, όταν συνδυαστούν. Ο συνδυασμός των θεωρητικά πιο χρήσιμων δοκιμασιών, μπορεί να περιλαμβάνει συσχετισμένα δεδομένα και έτσι η επίδοση των αντίστοιχων μοντέλων είναι μειωμένη. Σε κάθε περίπτωση και τα δύο μοντέλα οδηγούν σε τιμές της ακρίβειας μικρότερες από τα αντίστοιχα με το σύνολο των δεδομένων, επιβεβαιώνοντας τη γενική αρχή της μηχανικής μάθησης πως εκπαίδευση σε περισσότερα δεδομένα οδηγεί σε καλύτερα αποτελέσματα.

	<p style="text-align: center;"><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p style="text-align: center;"><b>ΓΡΑΦΗΣ</b></p> <p style="text-align: center;">ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p style="text-align: center;">ΙΟΥΛΙΟΣ 2024</p>
		<p style="text-align: center;">Σελίδα 21 / 22</p>

## 6 ΣΥΜΠΕΡΑΣΜΑΤΑ


Ακολούθως συνοψίζονται τα κύρια συμπεράσματα που προκύπτουν έπειτα από την ολοκλήρωση της διερεύνησης του προβλήματος.

Όσον αφορά στον άξονα αξιολόγησης του συνολικού συνόλου δεδομένων, η εφαρμογή μεθόδων μείωσης διαστάσεων, όπως η επιλογή χαρακτηριστικών και η ανάλυση κύριων συνιστωσών (PCA), οδήγησε σε ποικίλα αποτελέσματα. Χρησιμοποιώντας πίνακες συσχέτισεων, αφαιρέθηκαν χαρακτηριστικά με χαμηλή συσχέτιση ( $\leq 0,05$ ) στην πρώτη προσέγγιση, οδηγώντας στην αφαίρεση 43 χαρακτηριστικών. Σε αυτή την περίπτωση, ο Random Forest αποδείχθηκε ο πιο αποδοτικός ταξινομητής με ακρίβεια 0,887. Στη δεύτερη προσέγγιση, εντοπίστηκαν 34 ζεύγη χαρακτηριστικών με σημαντικά μεγάλη συσχέτιση μεταξύ τους ( $> 0,99$ ) και διατηρήθηκε μόνο ένα χαρακτηριστικό από κάθε ζεύγος, με τον Naive Bayes να επιτυγχάνει την υψηλότερη ακρίβεια (0,886).

Στην ανάλυση κύριων συνιστωσών (PCA), διατηρώντας τη συνολική διακύμανση στο 97%, εξάχθηκαν 87 συνιστώσες. Σε αυτή την περίπτωση, ο Random Forest εμφάνισε την υψηλότερη ακρίβεια στο 0,83. Αντίστοιχα, στη δεύτερη περίπτωση ανάλυσης PCA, με συνολική διακύμανση στο 96% και 82 συνιστώσες, οι Random Forest και MLP κατέγραψαν την υψηλότερη ακρίβεια, που ανήλθε σε 0,75.

Η διατήρηση όλων των διαστάσεων, παρά τον μεγάλο αριθμό χαρακτηριστικών, πραγματοποιήθηκε συμπληρωματικά λόγω του σχετικά μικρού αριθμού εγγραφών. Αυτή η προσέγγιση επέτρεψε την ανάπτυξη ενός μοντέλου με ολόκληρο το σύνολο δεδομένων, χωρίς να εφαρμοστεί αφαίρεση χαρακτηριστικών ή μεθόδων μείωσης διαστάσεων όπως το PCA. Στο πλαίσιο αυτό, ο Random Forest αποδείχθηκε ο πιο αποδοτικός ταξινομητής, επιτυγχάνοντας ακρίβεια 0,86, ακολουθούμενος από την Logistic Regression με ακρίβεια 0,85 και τον Decision Tree με ακρίβεια 0,82. Στην περίπτωση που εφαρμόστηκε επαύξηση δεδομένων, η μέγιστη τιμή ήταν ίση με 0,906, επιτεύχθηκε από τον Random Forest και αποτέλεσε την υψηλότερη τιμή της ακρίβειας που λήφθηκε στο πλαίσιο της παρούσας εργασίας.

Τέλος, μελετώντας μοντέλα και την προβλεπτική τους ικανότητα σε κάθε δοκιμασία (task) ξεχωριστά, προέκυψαν αποτελέσματα ελαφρώς χειρότερα από την προαναφερθείσα ανάλυση για τη μέση περίπτωση, ενώ υπήρχαν συγκεκριμένες δοκιμασίες που έδιναν τον χώρο στον

	<p><b>ΑΝΑΛΥΣΗ ΒΙΟΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΑΝΙΧΝΕΥΣΗ ΝΟΣΟΥ ALZHEIMER ΜΕΣΩ ΔΕΔΟΜΕΝΩΝ</b></p> <p><b>ΓΡΑΦΗΣ</b></p> <p>ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ</p>	<p>ΙΟΥΛΙΟΣ 2024</p>
		<p>Σελίδα 22 / 22</p>

ταξινομητή να βελτιώσει τη προβλεπτική του ικανότητα. Τα αποτελέσματα του βέλτιστου ταξινομητή πάνω σε κάθε δοκιμασία κυμαίνονταν μεταξύ 66% - 89%, ενώ οι δοκιμασίες με τις υψηλότερες τιμές ακρίβειας ήταν οι 10, 11, 9, 17 και 15.