Ανίχνευση της νόσου Alzheimer μέσω της γραφής

Ανάλυση Βιο-Δεδομένων 2023-2024

Βόγκας Ιωάννης (03400206)
Παπούλια Ευγενία (03400228)
Πριμέτης Κωνσταντίνος (03400231)
Χοντζάκης Διονύσιος (03400238)

Σκοπός Εργασίας

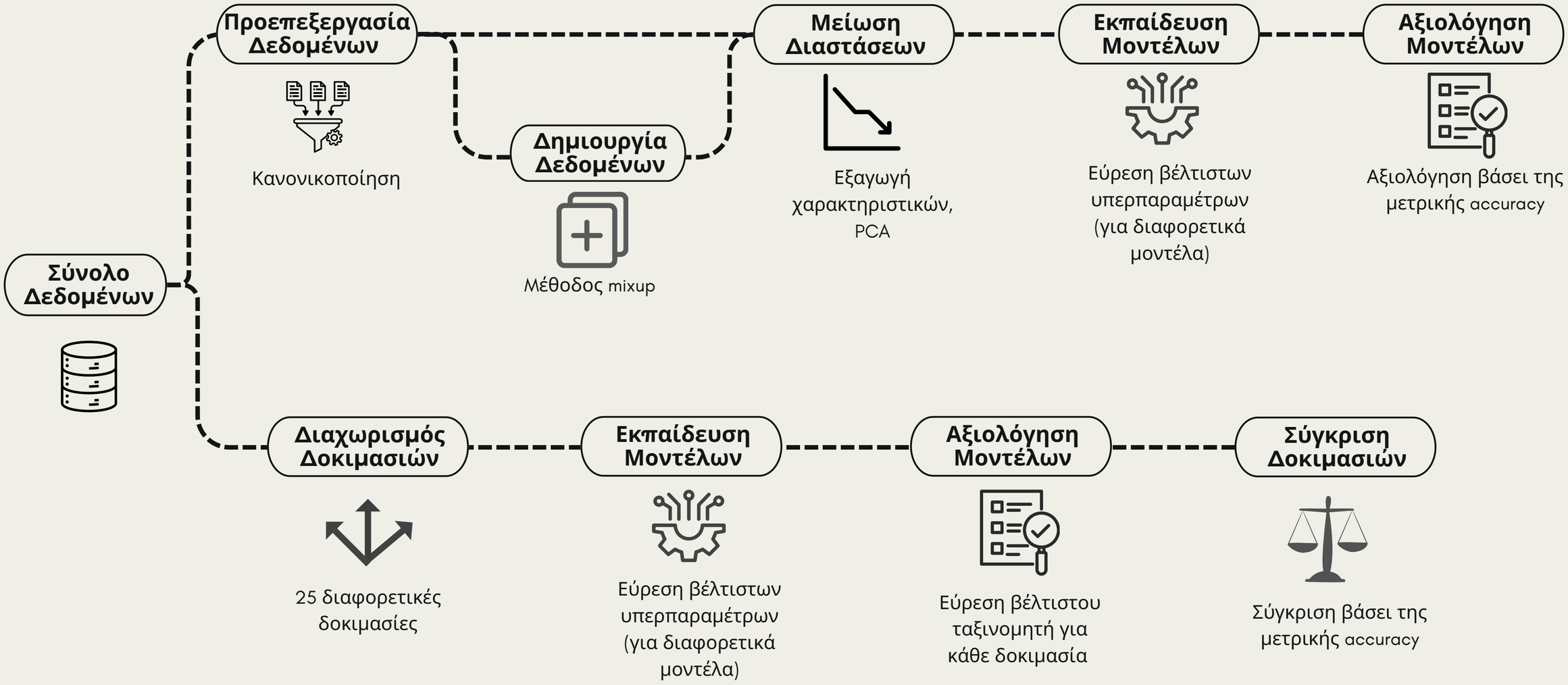
Η εργασία αποσκοπεί στη σύγκριση διαφόρων μοντέλων μηχανικής μάθησης ως προς την αποτελεσματικότητά τους στη διάγνωση της νόσου Alzheimer μέσω ανάλυσης της γραφής. Δηλαδή, αξιοποιώντας δεδομένα που περιγράφουν τον τρόπο γραφής τόσο ασθενών όσο και υγιών ατόμων, στόχος είναι να βρεθεί το καταλληλότερο μοντέλο για την ανίχνευση της συγκεκριμένης νόσου. Η προσπάθεια αυτή αποσκοπεί στη βελτίωση της έγκαιρης και ακριβούς διάγνωσης της νόσου Alzheimer, με ανώδυνο τρόπο, προωθώντας έτσι την καλύτερη κατανόηση και διαχείριση της ασθένειας.

Σύνολο Δεδομένων

Χαρακτηριστικά συνόλου δεδομένων:

- Πίνακας (174 X 452), το μεγαλύτερο διαθέσιμο σύνολο δεδομένων για το συγκεκριμένο πρόβλημα (Clica et al., 2022).
- Προέρχονται από 89 ασθενή και 85 υγιή άτομα.
- Περιγράφουν 25 διαφορετικές δοκιμασίες (αντιγραφή λέξεων και κειμένου, κατασκευή γραμμών και σχημάτων, απεικόνιση υπογραφής κ.ά.)
- Κάθε δοκιμασία αποτελείται από 18 αριθμητικές μεταβλητές (ταχύτητα γραφής, συνολικός χρόνος, πίεση που ασκείται στο χαρτί κ.ά.)
- Binary labels: “P” ή “H” (πάσχει ή όχι από Alzheimer)

Κατευθύνσεις και Βήματα Ανάλυσης



1ο Μέρος Ανάλυσης

Συγκεντρωτική Ανάλυση Δοκιμασιών

Μεθοδολογία

1. Κανονικοποίηση δεδομένων.
2. Μείωση διαστάσεων (feature selection και PCA).
   2.1 Εκπαίδευση 7 διαφορετικών μοντέλων για την εύρεση των βέλτιστων υπερπαραμέτρων μέσω 5-fold cross-validation και Leave-One-Out cross-validation (Random Forest, SVM, Logistic Regression, MLP, KNN, Naive Bayes, Decision Tree).
3. Διατήρηση όλων των διαστάσεων.
   3.1 Εκπαίδευση 7 διαφορετικών μοντέλων για την εύρεση των βέλτιστων υπερπαραμέτρων μέσω 5-fold cross-validation και Leave-One-Out cross-validation (Random Forest, SVM, Logistic Regression, MLP, KNN, Naive Bayes, Decision Tree).
4. Αξιολόγηση και σύγκριση των αποτελεσμάτων όλων των μοντέλων βάσει της μετρικής accuracy.

Μείωση των Διαστάσεων

ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

1η Περίπτωση: Με τη χρήση πίνακα συσχετίσεων αφαιρέθηκαν 43 χαρακτηριστικά που είχαν αμελητέα συσχέτιση με το label (<= 0.05).

2η Περίπτωση: Με τη χρήση πίνακα συσχετίσεων εντοπίστηκαν 34 ζεύγη χαρακτηριστικών με σημαντικά μεγάλη συσχέτιση (>0.99) και διατηρήθηκε ένα μόνο χαρακτηριστικό από κάθε ζεύγος.

ΑΠΟΤΕΛΕΣΜΑΤΑ

Μέθοδος Επιλογής Χαρακτηριστικών	Αποδοτικότερος Ταξινομητής	Ακρίβεια
1η Περίπτωση	Random Forest	0.887
2η Περίπτωση	Naive Bayes	0.886

Μείωση των Διαστάσεων

ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PCA)

Με τη χρήση ενός γραφήματος της εξηγούμενης διακύμανσης σε σχέση με τον αριθμό των συνιστωσών επιλέχθηκαν:

- 1η Περίπτωση: 87 συνιστώσες, διατηρώντας τη συνολική διακύμανση στο 97%.
- 2η Περίπτωση: 82 συνιστώσες, διατηρώντας τη συνολική διακύμανση στο 96%.

ΑΠΟΤΕΛΕΣΜΑΤΑ

Μέθοδος Επιλογής Χαρακτηριστικών	Αποδοτικότερος Ταξινομητής	Ακρίβεια
1η Περίπτωση	Random Forest	0.830
2η Περίπτωση	Decision Tree	0.773

Διατήρηση Όλων των Διαστάσεων

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 διαστάσεών του.

	ΑΠΟΔΟΤΙΚΟΤΕΡΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	**ΑΚΡΙΒΕΙΑ**
1	Random Forest	0.86
2	Logistic Regression	0.85
3	Decision Tree	0.82

2ο Μέρος Ανάλυσης

Διαχωρισμός Δοκιμασιών

Μεθοδολογία

1. Διαχωρίστηκαν οι 25 διαφορετικές δοκιμασίες (tasks).
2. Για κάθε δοκιμασία, εκπαιδεύτηκαν 5 διαφορετικά μοντέλα για την εύρεση των βέλτιστων υπερπαραμέτρων μέσω 5-fold cross-validation (Random Forest, SVM, Logistic Regression, MLP, XGB).
3. Επαναλήφθηκε η διαδικασία της εκπαίδευσης 5 φορές (κάθε μία με διαφορετικό split σε train και test set) με σκοπό να εξεταστεί το robustness των μοντέλων.
4. Για κάθε δοκιμασία επιλέχθηκε ο καλύτερος ταξινομητής (εκείνος που προέβλεπε με μεγαλύτερη ακρίβεια τις περισσότερες φορές).
5. Για τον κάθε καλύτερο ταξινομητή υπολογίστηκε ο μέσος όρος των accuracies από τις 5 επαναλήψεις.
6. Ταξινομήθηκαν και συγκρίθηκαν οι δοκιμασίες.

Δοκιμασίες με την καλύτερη προβλεπτική ικανότητα

	ΤΟΡ 5 ΔΟΚΙΜΑΣΙΕΣ	ΤΑΞΙΝΟΜΗΤΗΣ	Μ.Ο. ΑΚΡΙΒΕΙΑΣ	ΔΙΑΣΤΗΜΑ ΑΚΡΙΒΕΙΑΣ ΤΩΝ 5 ΕΠΑΝΑΛΗΨΕΩΝ
1	Αντιγραφή της λέξης “foglio”	SVM	0.89	[0.87, 0.91]
2	Αντιγραφή της λέξης “foglio” πάνω σε μία γραμμή	SVM	0.85	[0.78, 0.90]
3	Σχηματισμός μίας ακολουθίας τεσσάρων διγραμμάτων “le” με μία συνεχόμενη κίνηση	SVM	0.82	[0.74, 0.90]
4	Αντιγραφή έξι λέξεων σε αντίστοιχα κουτιά	XGB	0.81	[0.77, 0.87]
5	Αντιγραφή της λέξης “bottiglia” αντίστροφα	SVM	0.80	[0.78, 0.89]

Δοκιμασίες με τη χειρότερη προβλεπτική ικανότητα

	ΒΟΤΤΟΜ 5 ΔΟΚΙΜΑΣΙΕΣ	ΤΑΞΙΝΟΜΗΤΗΣ	Μ.Ο. ΑΚΡΙΒΕΙΑΣ	ΔΙΑΣΤΗΜΑ ΑΚΡΙΒΕΙΑΣ ΤΩΝ 5 ΕΠΑΝΑΛΗΨΕΩΝ
1	Σχηματισμός μίας ακολουθίας τεσσάρων γραμμάτων “l” με μία συνεχόμενη κίνηση	MLP	0.66	[0.64, 0.77]
2	Αντιγραφή των πεδίων μίας ταχυδρομικής επιταγής	Logistic Regression	0.69	[0.65, 0.70]
3	Αντιγραφή μίας παραγράφου	MLP	0.69	[0.61, 0.78]
4	Συγγραφή μίας πρότασης μέσω υπαγόρευσης	SVM	0.69	[0.57, 0.73]
5	Αντιγραφή των γραμμάτων “l”, “m” και “p”	Random Forest	0.70	[0.64, 0.78]

3ο Μέρος Ανάλυσης

Εμπλουτισμός Συνόλου Δεδομένων

Μεθοδολογία

1. Διαχωρισμός συνόλου δεδομένων σε test και train sets.
2. Προσθήκη 60 νέων δειγμάτων στο σύνολο εκπαίδευσης μέσω της μεθόδου mixup (ανά δύο δείγματα του αρχικού train set δημιουργήθηκε ένα επιπλέον).
3. Μείωση διαστάσεων (PCA) / Διατήρηση όλων των διαστάσεων.
4. Εκπαίδευση 7 διαφορετικών μοντέλων για την εύρεση των βέλτιστων υπερπαραμέτρων μέσω 5-fold cross-validation (Random Forest, SVM, Logistic Regression, MLP, KNN, Naive Bayes, Decision Tree).
5. Αξιολόγηση και σύγκριση των αποτελεσμάτων των μοντέλων βάσει της μετρικής accuracy.

Αποτελέσματα Προσθήκης Δειγμάτων

ΜΕΙΩΣΗ ΔΙΑΣΤΑΣΕΩΝ

Με τη χρήση ενός γραφήματος της εξηγούμενης διακύμανσης σε σχέση με τον αριθμό των συνιστωσών επιλέχθηκαν 87 συνιστώσες.

Πιο αποδοτικός ταξινομητής αποδείχθηκε ο **Naive Bayes** με **ακρίβεια 0.867.**

ΔΙΑΤΗΡΗΣΗ ΟΛΩΝ ΤΩΝ ΔΙΑΣΤΑΣΕΩΝ

Ανάπτυξη μοντέλου με ολόκληρο το σύνολο δεδομένων και την προθήκη των νέων δειγμάτων εκπαιδευσης.

Πιο αποδοτικός ταξινομητής αποδείχθηκε ο **Random Forest** με **ακρίβεια 0.906**.

Σύνοψη Αποτελεσμάτων ανά Περίπτωση

	Περίπτωση	Ταξινομητής	Ακρίβεια
1	Αφαίρεση χαρακτηριστικών λόγω συσχέτισης	Random Forest	0.89
2	Μείωση διαστάσεων μέσω PCA	Random Forest	0.83
3	Διατήρηση όλων των διαστάσεων	Random Forest	0.86
4	Προσθήκη δειγμάτων εκπαίδευσης (μέθοδος mixup)	Random Forest	0.90
5	Διαχωρισμός δοκιμασιών	SVM	0.89

Σχολιασμός Αποτελεσμάτων

- Η μείωση των διαστάσεων αυξάνει την αποδοτικότητα των μοντέλων, ωστόσο πρέπει να διατηρηθεί σημαντικός αριθμός διαστάσεων.
- Ο ταξινομητής Random Forest είναι ο πιο αποδοτικός όσον αφορά την συγκεντρωτική ανάλυση των δοκιμασιών, με μέγιστη ακρίβεια 0,887.
- Απομονώνοντας συγκεκριμένες δοκιμασίες (tasks) επιτυγχάνονται σημαντικά υψηλά αποτελέσματα, κοντά σε αυτά που προκύπτουν από την συγκεντρωτική ανάλυση των δοκιμασιών.
- Ο ταξινομητής SVM ξεχωρίζει για την αποδοτικότητά του στην πρόβλεψη της νόσου μέσω μεμονωμένων δοκιμασιών (tasks).
- Η προσθήκη δειγμάτων εκπαίδευσης μέσω της μεθόδου mixup δύναται να βελτιώσει, σε μικρό βαθμό, την αποδοτικότητα των μοντέλων.

Βιβλιογραφία

Cilia, N. D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A., & Parziale, A, 2022. Diagnosing Alzheimer’s disease from on-line handwriting: A novel dataset and performance benchmarking. Engineering Applications of Artificial Intelligence, 111.

Ευχαριστούμε!

Ερωτήσεις;