ΕΡΓΑΣΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ (2026-2027)

Σενάριο και στόχος

Μια εταιρεία wearable sensors αναπτύσσει ένα σύστημα αναγνώρισης καθημερινών και αθλητικών δραστηριοτήτων. Ως ομάδα ανάλυσης δεδομένων, καλείστε να αξιοποιήσετε πραγματικές μετρήσεις αισθητήρων για να ανιχνεύετε τη δραστηριότητα του χρήστη και να ανακαλύπτετε ομάδες παρατηρήσεων με παρόμοια χαρακτηριστικά. Στόχος είναι να κατανοήσετε τα δεδομένα, να διερευνήσετε την ποιότητά τους και να συγκρίνετε εναλλακτικές επιλογές με βάση μετρήσιμα αποτελέσματα.

Δεδομένα και πρόσβαση

Για την εργασία θα χρησιμοποιήσετε το σύνολο δεδομένων PAMAP2 το οποίο περιλαμβάνει 3.850.505 καταγραφές από 9 συμμετέχοντες και 18 δραστηριότητες συνολικά. Οι μετρήσεις προέρχονται από τρεις αισθητήρες αδράνειας και έναν μετρητή καρδιακού ρυθμού. Το αρχείο περιλαμβάνει πραγματικές τιμές.

Σελίδα δεδομένων, λήψη και τεκμηρίωση — UCI Machine Learning Repository

Χρησιμοποιήστε τα αρχεία του φακέλου Protocol, με τις 12 προβλεπόμενες δραστηριότητες.

Παρότι οι καταγραφές αποτελούν χρονοσειρές, τα αρχεία έχουν ήδη μορφή πίνακα. Για τη βασική προσέγγιση μπορείτε να χρησιμοποιήστε κάθε γραμμή ως ξεχωριστή παρατήρηση, αξιοποιώντας τις τιμές αισθητήρων της συγκεκριμένης χρονικής στιγμής.

Κάθε γραμμή έχει 54 πεδία: χρονική σήμανση, ετικέτα δραστηριότητας και 52 μεταβλητές που αφορούν τους αισθητήρες. Οι 40 αφορούν τιμές μέτρησης και οι 12 αφορούν προσανατολισμό και μπορούν να αγνοηθούν. Οι εγγραφές με activityID = 0 αφορούν μεταβατικές δραστηριότητες και εξαιρούνται από τη μελέτη.

Μπορείτε ως επέκταση να μετατρέψετε το αρχείο σε χρονικά παράθυρα για τα οποία να εξάγετε χαρακτηριστικά, όπως μέση τιμή ή τυπική απόκλιση.

Μελετήστε τα συνοδευτικά αρχεία τεκμηρίωσης πριν από την επεξεργασία.

ΕΡΓΑΣΙΕΣ

<u>Α) Προετοιμασία και αρχική ανάλυση</u>

Α1. Φόρτωση και έλεγχος ποιότητας

- Δημιουργήστε μια διαδικασία φόρτωσης των αρχείων. Καταγράψτε τα ονόματα και τους τύπους των μεταβλητών που θα χρησιμοποιήσετε τελικά, καθώς και το πλήθος των παρατηρήσεων πριν και μετά από βήμα προεπεξεργασίας.

- Παρουσιάστε κάποια στατιστικά δεδομένα όπως: τις συχνότητες των δραστηριοτήτων, την κάλυψη ανά συμμετέχοντα και την κατανομή των ελλειπουσών τιμών ανά μεταβλητή.
- Ελέγξτε διπλότυπες εγγραφές, σταθερές μεταβλητές και ύποπτες τιμές.
- Διερευνήστε την υπάρξη θορύβου και ακραίων τιμών. Τεκμηριώστε πότε μια τιμή είναι πιθανό σφάλμα μέτρησης και πότε μπορεί να εκφράζει μια έγκυρη, έντονη δραστηριότητα.

Α2. Διερευνητική ανάλυση

- Εξετάστε τις κατανομές βασικών χαρακτηριστικών συνολικά και ανά δραστηριότητα. Χρησιμοποιήστε στοχευμένα διαγράμματα με σαφείς άξονες, μονάδες και σύντομη ερμηνεία.
- Εντοπίστε ισχυρά συσχετισμένα ή πλεονάζοντα χαρακτηριστικά. Δείξτε έναν ευανάγνωστο πίνακα ή χάρτη συσχετίσεων και αναλύστε επιλεγμένα ζεύγη.
- Προτείνετε τα σημαντικότερα χαρακτηριστικά για την αναγνώριση δραστηριοτήτων. Τεκμηριώστε την κατάταξη με κατάλληλη μέθοδο και σχολιάστε πώς επηρεάζεται η ερμηνεία από τη συσχέτιση μεταξύ χαρακτηριστικών.

Α3. Προεπεξεργασία και μείωση διαστάσεων

- Σχεδιάστε διαδικασία αντιμετώπισης ελλειπουσών τιμών, διαφορετικών κλιμάκων και προβληματικών μετρήσεων. Καταγράψτε τις εναλλακτικές που εξετάσατε και αιτιολογήστε τις τελικές επιλογές.
- Συγκρίνετε την αρχική έγκυρη αναπαράσταση με τουλάχιστον μία αναπαράσταση μειωμένης διάστασης, π.χ. PCA. Αιτιολογήστε τη μέθοδο και τον αριθμό διαστάσεων. Αξιολογήστε τη νέα αναπαράσταση τόσο στην ταξινόμηση όσο και στη συσταδοποίηση.
- Για τις υπολογιστικά απαιτητικές αναλύσεις μπορείτε να χρησιμοποιήσετε υποσύνολα. Περιγράψτε τον τρόπο δειγματοληψίας και εξετάστε πόσο αντιπροσωπευτικά είναι ως προς τις δραστηριότητες και τη σύνθεση των δεδομένων.

<u>Β) Κατηγοριοποίηση/ταξινόμηση</u>

Β1. Πειραματική οργάνωση

Ορίστε και τεκμηριώστε τη στρατηγική διαχωρισμού των δεδομένων, την επιλογή παραμέτρων και τη διαδικασία τελικής αξιολόγησης. Αναφέρετε το μέγεθος και τη σύνθεση των συνόλων που χρησιμοποιείτε. Επιλέξτε μια συνεπή διαδικασία ώστε οι συγκρίσεις των μοντέλων να είναι ουσιαστικές.

Β2. Μοντέλα και μετρικές

- Υλοποιήστε μια γραμμή βάσης που προβλέπει τη συχνότερη δραστηριότητα και έναν απλό ταξινομητή με βασική προεπεξεργασία.
- Συγκρίνετε τουλάχιστον πέντε διαφορετικούς ταξινομητές. Επιλέξτε διαφορετικές οικογένειες και αιτιολογήστε τη χρησιμότητά τους.
- Διερευνήστε την επιλογή υπερπαραμέτρων στους ταξινομητές με λογικό υπολογιστικό όριο. Αναφέρετε τις τιμές που εξετάσατε και τον τρόπο επιλογής της τελικής ρύθμισης.
- Χρησιμοποιήστε ως κύρια μετρική το macro-F1. Παρουσιάστε επιπλέον balanced accuracy, precision και recall ανά δραστηριότητα, καθώς και πίνακα σύγχυσης του τελικού μοντέλου. Συμπεριλάβετε χρόνους εκπαίδευσης και πρόβλεψης.

Β3. Τι βελτίωσε τη λύση;

- Για τουλάχιστον έναν ταξινομητή, συγκρίνετε ως προς την επίδοση, και το χρόνο την αρχική και μειωμένη αναπαράσταση (αναφέρετε το πλήθος γνωρισμάτων).
- Κατασκευάστε την καμπύλη μάθησης με τρία μεγέθη δεδομένων εκπαίδευσης, ενδεικτικά 10%, 30% και 100% του συνόλου εκπαίδευσης που ορίσατε. Περιγράψτε τη δειγματοληψία και συγκρίνετε επίδοση και χρόνο, χρησιμοποιώντας κοινό σύνολο αξιολόγησης.
- Εντοπίστε και αναλύστε τουλάχιστον δύο συχνές συγχύσεις κλάσεων. Προτείνετε πιθανές εξηγήσεις και σχολιάστε την επίδοση στις λιγότερο συχνές κατηγορίες.

<u>Γ) Συσταδοποίηση και ερμηνεία</u>

Γ1. Αναπαραστάσεις και αλγόριθμοι

- Συσταδοποιήστε τις παρατηρήσεις στα χαρακτηριστικά εισόδου χωρίς να λάβετε υπόψη σας την ετικέτα δραστηριότητας. Επιλέξτε κατάλληλη κλίμακα και μέτρο ομοιότητας ή απόστασης και αιτιολογήστε τις επιλογές σας.
- Συγκρίνετε τουλάχιστον τρεις διαφορετικές προσεγγίσεις από διαφορετικές οικογένειες αλγορίθμων (π.χ. βάσει κέντρου, βάσει πυκνότητας, ιεραρχική συσταδοποίηση).
- Χρησιμοποιήστε και την αρχική και τη μειωμένη αναπαράσταση.

Γ2. Επιλογή ρυθμίσεων και αξιολόγηση

- Διερευνήστε διαφορετικά πλήθη συστάδων ή άλλες βασικές παραμέτρους. Δεν απαιτείται το πλήθος των συστάδων να ταυτίζεται με το πλήθος των δραστηριοτήτων.
- Επιλέξτε ρυθμίσεις με βάση εσωτερικές μετρικές (π.χ. silhouette score), και ελέγξτε τη σταθερότητα με τουλάχιστον τρεις εκτελέσεις ή επαναδειγματοληψίες. Δηλώστε τυχόν δειγματοληψία. Εξηγήστε πώς συγκρίνετε τα αποτελέσματα.
- Αφού επιλέξετε τις ρυθμίσεις, συγκρίνετε τις συστάδες με τις γνωστές δραστηριότητες: υπολογίστε Adjusted Rand Index (ARI) και Normalized Mutual Information (NMI) και παρουσιάστε πίνακα συστάδων–δραστηριοτήτων.
- Όταν ένας αλγόριθμος χαρακτηρίζει παρατηρήσεις ως θόρυβο, αναφέρετε το ποσοστό τους και εξηγήστε πώς αντιμετωπίζονται στις μετρικές.

Γ3. Ερμηνεία ευρημάτων

Ποιες δραστηριότητες συνυπάρχουν στις ίδιες συστάδες; Ποιες διασπώνται σε περισσότερες; Περιγράψτε τα προφίλ επιλεγμένων συστάδων μέσα από τα χαρακτηριστικά τους. Εξετάστε αν η δομή σχετίζεται με διαφορές μεταξύ συμμετεχόντων, ένταση κίνησης ή θέση αισθητήρα. Συγκρίνετε τα ευρήματα με τα σφάλματα της ταξινόμησης και εξηγήστε την επίδραση της μείωσης διαστάσεων.

ΠΑΡΑΔΟΣΗ

Η ημερομηνία παράδοσης ορίζεται η 15/1/2027 στο eclass.

Κάθε ομάδα παραδίδει ένα ενιαίο πακέτο που συνδέει τα ερωτήματα, τα πειράματα και τα συμπεράσματα. Το πακέτο πρέπει να καλύπτει τις ενότητες Α, Β, Γ.

<u>Παραδοτέο ενότητας Α</u>

Στην αναφορά: σύνοψη του συνόλου εργασίας, πίνακας ποιότητας δεδομένων, 4–6 στοχευμένα διαγράμματα, τεκμηρίωση συσχετίσεων και σημαντικών χαρακτηριστικών, πίνακας αποφάσεων προεπεξεργασίας και περιγραφή της μειωμένης αναπαράστασης. Στον κώδικα: εκτελέσιμη διαδικασία φόρτωσης και μετασχηματισμών.

Παραδοτέο ενότητας Β

Στην αναφορά: περιγραφή πειραματικού πρωτοκόλλου, ενιαίος πίνακας σύγκρισης ταξινομητών και ρυθμίσεων, καμπύλη μάθησης, πίνακας σύγχυσης και ανάλυση σφαλμάτων. Στον κώδικα: διαδικασία εκπαίδευσης και αξιολόγησης και αρχείο αποτελεσμάτων σε CSV.

Παραδοτέο ενότητας Γ

Στην αναφορά: πίνακας αλγορίθμων, αναπαραστάσεων, παραμέτρων, μετρικών και χρόνων, τεκμηρίωση σταθερότητας, πίνακας συστάδων–δραστηριοτήτων, 1–2 απεικονίσεις και ερμηνεία τουλάχιστον τριών συστάδων. Στον κώδικα: αναπαραγώγιμες εκτελέσεις και εξαγωγή των αντίστοιχων αποτελεσμάτων σε CSV.

Τελικό πακέτο υποβολής

Παραδοτέο	Περιεχόμενο
report.pdf	Τεχνική αναφορά 8–12 σελίδων, χωρίς εξώφυλλο, βιβλιογραφία και προαιρετικά παραρτήματα. Διακριτές ενότητες Α, Β και Γ, συμπεράσματα, περιορισμοί και συνεισφορά μελών.
src/ ή notebooks/	Εκτελέσιμος κώδικας φόρτωσης, προεπεξεργασίας, EDA, ταξινόμησης και συσταδοποίησης. Τα notebooks να περιλαμβάνουν τις βασικές εξόδους.
README.md και περιβάλλον	Οδηγίες λήψης/τοποθέτησης δεδομένων και εκτέλεσης, εκδόσεις βιβλιοθηκών, τυχαίοι σπόροι, απαιτήσεις υλικού και ενδεικτικός χρόνος. requirements.txt ή environment.yml.
results/	Πίνακες πειραμάτων σε CSV, βασικά διαγράμματα και ρυθμίσεις που επιτρέπουν την αντιστοίχιση κάθε αποτελέσματος με την εκτέλεσή του.

Υποβάλετε ένα συμπιεσμένο αρχείο ανά ομάδα. Μην συμπεριλάβετε τα αρχικά δεδομένα, αλλά δώστε σαφείς οδηγίες πρόσβασης.

Τελική παρουσίαση

Προτείνεται παρουσίαση 5–7 διαφανειών με έμφαση στις αποφάσεις και στα ευρήματα. Όλα τα μέλη συμμετέχουν στην προετοιμασία και στην παρουσίαση. Κάθε μέλος πρέπει να μπορεί να εξηγήσει τη συνολική προσέγγιση της ομάδας.

Ενδεικτικό περιεχόμενο

- Πρόβλημα, δεδομένα και ερευνητικά ερωτήματα
- Βασικά ευρήματα EDA και αποφάσεις προεπεξεργασίας
- Ταξινόμηση, συγκρίσεις και ανάλυση σφαλμάτων
- Συσταδοποίηση, αξιολόγηση και ερμηνεία
- Τι βοήθησε, βασικοί περιορισμοί και τελικό συμπέρασμα

Προτιμήστε ευανάγνωστα διαγράμματα και συγκρίσεις αντί για εκτενή αποσπάσματα κώδικα. Να είστε έτοιμοι για σύντομες ερωτήσεις σχετικά με τις επιλογές και τα αποτελέσματά σας.

Αξιολόγηση

Η ποιότητα της τεκμηρίωσης και της ερμηνείας αποτελεί βασικό μέρος της αξιολόγησης. Δεν βαθμολογείται μόνο η υψηλότερη επίδοση. Μια τεκμηριωμένη σύγκριση, μια εύστοχη ανάλυση αποτυχίας ή ένα αρνητικό αποτέλεσμα μπορούν να αποτελούν ουσιαστική συνεισφορά. Δεν απαιτείται κάθε τεχνική να βελτιώνει τις μετρικές· απαιτείται να εξετάζετε και να εξηγείτε την επίδρασή της.

ΚΑΤΑΝΟΜΗ ΒΑΘΜΟΛΟΓΙΑΣ

Κριτήριο	Βαρύτητα	Τι αξιολογείται
Προεπεξεργασία και EDA	20%	Ποιότητα ελέγχων, συσχετίσεις, σημαντικά χαρακτηριστικά, αιτιολόγηση αποφάσεων.
Μείωση διαστάσεων και δειγματοληψία	15%	Ουσιαστικές συγκρίσεις αναπαραστάσεων και μεγεθών δεδομένων, επίδοση και κόστος.
Ταξινόμηση	25%	Πειραματική οργάνωση, baselines, διαφορετικές οικογένειες, μετρικές, και σφάλματα.

Συσταδοποίηση	25%	Σύγκριση μεθόδων, επιλογή ρυθμίσεων, σταθερότητα και ερμηνεία συστάδων.
Αναπαραγωγιμότητα και αναφορά	5%	Εκτελέσιμος κώδικας, πλήρεις οδηγίες, σαφής σύνδεση αποτελεσμάτων και συμπερασμάτων.
Προφορική παρουσίαση	10%	Σαφήνεια, επιλογή ευρημάτων, τήρηση χρόνου, συμμετοχή και κατανόηση της λύσης.

Οι δύο εργασίες με την καλύτερη τεκμηρίωση θα έχουν ένα bonus 10%.