



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ,
ΤΟΜΕΑΣ ΘΕΩΡΗΤΙΚΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΘΕΩΡΙΑΣ ΣΥΣΤΗΜΑΤΩΝ ΚΑΙ ΕΛΕΓΧΟΥ –
ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ

ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ

ΟΝΟΜΑΤΕΠΩΝΥΜΑ ΦΟΙΤΗΤΩΝ:

ΔΑΜΙΑΝΙΔΟΥ ΜΑΡΙΑ ΙΩΑΝΝΑ (ΑΕΜ: 852)

ΜΑΡΚΟΥ ΠΑΝΤΕΛΕΗΜΩΝ (ΑΕΜ: 829)

ΧΟΛΕΒΑΣ ΜΗΝΑΣ (ΑΕΜ: 837)

ΘΕΣΣΑΛΟΝΙΚΗ, ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2022

ΜΑΘΗΜΑ: ΕΙΔΙΚΑ ΘΕΜΑΤΑ Ι: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ,
ΔΙΔΑΣΚΩΝ: ΧΑΡΙΣΗΣ Β.

Πίνακας περιεχομένων

Περίληψη 4

Εισαγωγή 5

1. Προ – επεξεργασία Δεδομένων 6

2. Εξαγωγή χαρακτηριστικών 7

2.1 Θηκογράμματα 8

3. Εκπαίδευση μοντέλων Ταξινόμησης 11

3.1 Μηχανές διανυσμάτων υποστήριξης (SVM με πυρήνα RBF) 12

3.2 Δένδρα αποφάσεων (Decision Trees) 23

3.3 Νευρωνικά δίκτυα MLP με χρήση ορμής (momentum) 37

4. Σύγκριση μοντέλων Ταξινόμησης 49

Περίληψη

Στην παρούσα εργασία έγινε μια προσπάθεια εφαρμογής μεθόδων μηχανικής μάθησης για την αναγνώριση ορισμένων ανθρώπινων δραστηριοτήτων. Το πρόβλημα αυτό ανήκει στο ευρύτερο πεδίο της αναγνώρισης μοτίβων και πιο συγκεκριμένα αποτελεί ένα πρόβλημα κατηγοριοποίησης πολλαπλών κλάσεων. Το σύνολο των δεδομένων που χρησιμοποιήθηκε, περιέχει καταγραφές από τον αισθητήρα του επιταχυνσιόμετρου ενός κινητού τηλεφώνου.

Στα πλαίσια της εργασίας εφαρμόστηκαν οι εξής τρεις ταξινομητές:

1. **Μηχανές διανυσμάτων υποστήριξης με πυρήνα Radial Basis Function** (SVM με πυρήνα RBF)
2. **Δένδρα αποφάσεων** (Decision Trees)
3. **Νευρωνικά δίκτυα MLP με 1 – 2 κρυφά στρώματα και χρήση ορμής** (momentum)

Με την ολοκλήρωση της εφαρμογής των ταξινομητών και την εξαγωγή των αποτελεσμάτων μπορέσαμε στη συνέχεια να συγκρίνουμε τις μεθόδους και να αυξήσουμε την ακρίβεια των ταξινομητών μέσω συνένωσης συγκεκριμένων δραστηριοτήτων.

Η υλοποίηση πραγματοποιήθηκε σε γλώσσα προγραμματισμού *Python* με τη βοήθεια της βιβλιοθήκης μηχανικής μάθησης *sklearn*.

Εισαγωγή

Στόχος της παρούσας εργασίας είναι η κατάστρωση ενός συστήματος αναγνώρισης δραστηριότητας και συγκεκριμένα η εφαρμογή μεθόδων μηχανικής μάθησης (*Machine Learning*) για την αυτοματοποιημένη αναγνώριση και κατηγοριοποίηση δραστηριοτήτων από καταγραφές αισθητήρων κινητού τηλεφώνου. Η αναγνώριση δραστηριότητας είναι ένας τομέας που έχει γνωρίσει ανάπτυξη, τα τελευταία κυρίως χρόνια, κάτι που οφείλεται στις πρόσφατες τεχνολογικές εξελίξεις γύρω απ' τη τεχνητή νοημοσύνη καθώς και την διάδοση των έξυπνων φορητών συσκευών.

Η αναγνώριση ανθρώπινης δραστηριότητας, ως πεδίο, αφορά στην ανάπτυξη μοντέλων και αλγορίθμων ικανών να ερμηνεύουν και να ξεχωρίζουν κινήσεις του ανθρώπινου σώματος με σκοπό το προσδιορισμό ενεργειών ή δραστηριοτήτων. Συστήματα αναγνώρισης ανθρώπινης δραστηριότητας έχουν αναπτυχθεί και βρει εφαρμογή, τα τελευταία χρόνια, σε ποικίλους και σημαντικούς τομείς:

- Ασφάλεια και παρακολούθηση
- Υγειονομική περίθαλψη και μέριμνα
- Αλληλεπίδραση ανθρώπου – υπολογιστή

Το πρόβλημα της αναγνώρισης ανθρώπινης δραστηριότητας αποτελεί ένα πρόβλημα κατηγοριοποίησης πολλαπλών κλάσεων. Η κατηγοριοποίηση πολλαπλών κλάσεων (*multiclass classification*) στα πλαίσια της μηχανικής μάθησης αφορά ένα είδος εποπτευόμενης μάθησης (*supervised learning*), όπου ένας αλγόριθμος κατηγοριοποίησης καλείται να κατατάξει ένα δείγμα σε μια και μοναδική κλάση από ένα σύνολο τριών ή παραπάνω κλάσεων.

Η εποπτευόμενη μάθηση είναι μια διαδικασία μάθησης όπου, μέσω ενός μαρκαρισμένου (*labelled*) συνόλου δεδομένων εκπαίδευσης, ένας αλγόριθμος σχηματίζει μια συνάρτηση f , η οποία αντιστοιχίζει κάθε είσοδο με κάποια έξοδο. Ο στόχος είναι, η συνάρτηση που σχηματίζει ο αλγόριθμος, να μπορεί να συσχετίζει δείγματα καινούρια, τα οποία δεν έχει ξαναδεί, με τις σωστές κλάσεις. Θέλουμε, δηλαδή, ο αλγόριθμος να μπορεί να γενικεύει πέραν απ' το σύνολο εκπαίδευσης, να αποδίδει ικανοποιητικά όταν του παρουσιάζονται νέα δεδομένα.

Στα πλαίσια της εργασίας καταστρώσαμε ένα σύστημα που μπορεί να ταξινομήσει με ακρίβεια τις κινήσεις που εκτελεί ένα άτομο σε μία από τις εξής δραστηριότητες: **περπάτημα (walking), όρθιος (standing), τρέξιμο (jogging), καθιστός (sitting), ποδηλασία (biking), ανέβασμα σκαλοπατιών (upstairs), κατέβασμα σκαλοπατιών (downstairs)** βασιζόμενο αποκλειστικά στο σήμα του επιταχυνσιόμετρου ενός «έξυπνου» κινητού. Το επιταχυνσιόμετρο αποτελεί έναν αισθητήρα, ο οποίος μετρά και καταγράφει κατά μήκος τριών αξόνων x , y , και z την επιτάχυνση ενός σώματος ως προς την επιτάχυνση της βαρύτητας, που προκύπτει ως αποτέλεσμα όλων των δυνάμεων που ασκούνται επί αυτού. Τα δεδομένα μετά από κατάλληλο μετασχηματισμό και εξαγωγή χαρακτηριστικών τροφοδοτήθηκαν στις παρακάτω κύριες μεθόδους μηχανικής μάθησης, οι οποίες εφαρμόζονται στην εργασία:

- 1) SVM με πυρήνα Radial Basis Function (RBF)
- 2) Δένδρα απόφασης
- 3) Νευρωνικά δίκτυα ΜΚΡ με 1-2 κρυφά στρώματα και χρήση ορμής (momentum).

1. Προεπεξεργασία Δεδομένων

Το σύνολο δεδομένων που έχουμε στη διάθεσή μας προέρχεται από την εργασία των Shoaib et al. [1] και περιλαμβάνει καταγραφές από 10 συμμετέχοντες (διαθέτουμε ένα csv αρχείο για τον καθένα) που χρησιμοποιούν κινητά τηλέφωνα σε διαφορετικές θέσεις του σώματος ενώ εκτελούν μία δραστηριότητα εκ των biking, downstairs, jogging, sing, standing, upstairs, walking.

Η κάθε καταγραφή δίνει τις τιμές των επιταχυνσιόμετρων από κινητά σε διαφορετικές θέσεις. Η τελευταία στήλη αντιστοιχεί στη δραστηριότητα που εκτελείται. Για την προεπεξεργασία των σημάτων επιλέξαμε μια θέση του κινητού, συγκεκριμένα την “right pocket” (εκ των “right pocket” και “left pocket”).

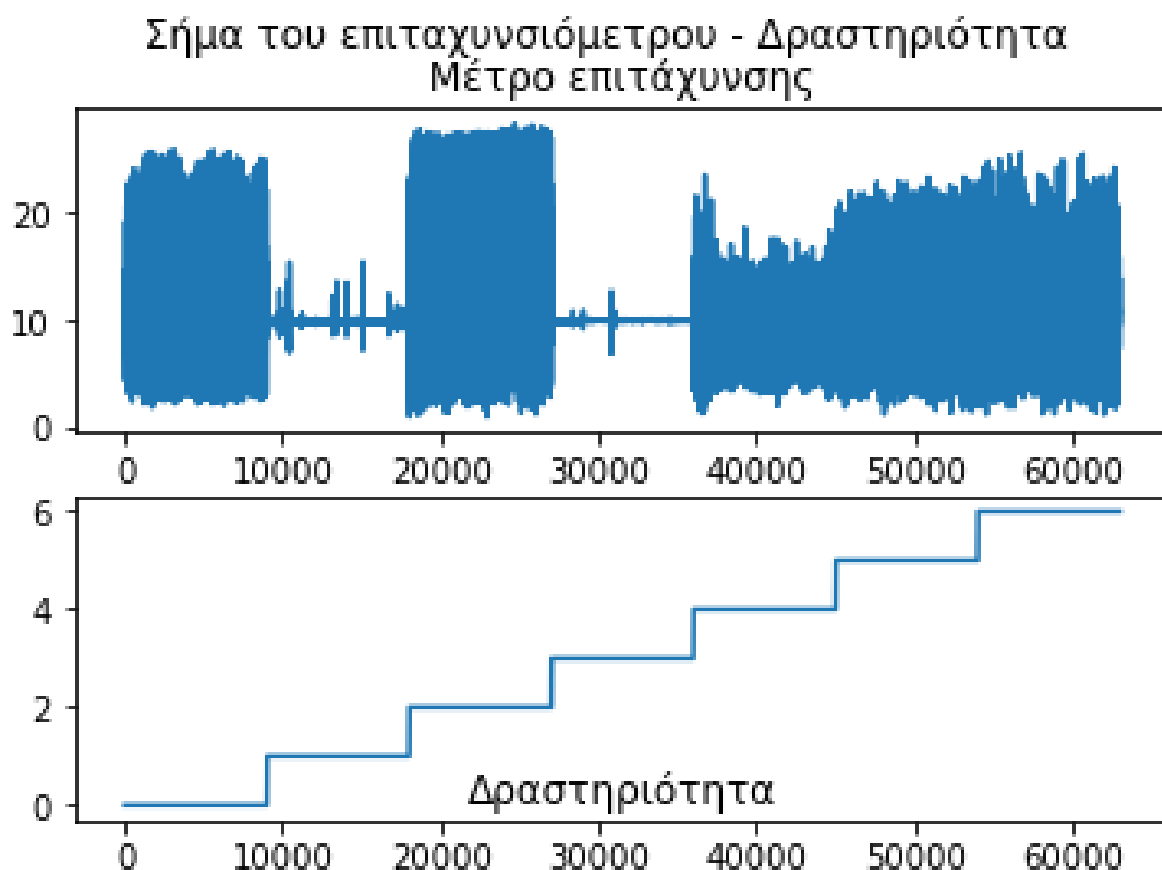
Μετατρέψαμε το τρισδιάστατο σήμα του επιταχυνσιόμετρου σε μονοδιάστατο, παίρνοντας το μέτρο του διανύσματος επιτάχυνσης, δημιουργώντας τη συνάρτηση $\text{norm}()$, η οποία υπολογίζει την τετραγωνική ρίζα των τετραγώνων των συνιστωσών. Ελέγξαμε εάν υπήρχαν τιμές μέτρου που είναι πάνω από 1000, πράγμα που αντιστοιχεί σε σφάλμα, (δεν υπήρχαν τέτοιες τιμές) και προσθέσαμε στον κώδικα μια εντολή “if” η οποία ελέγχει εάν το μέτρο είναι πάνω από 1000 και, σε περίπτωση που είναι, η καταγραφή αυτή αντικαθίσταται από την προηγούμενη.

Στην συνέχεια έγινε έλεγχος του αριθμού καταγραφών ανά δραστηριότητα:



και διαπιστώθηκε ότι **το πλήθος των καταγραφών είναι ίσο ανά κάθε δραστηριότητα.**

2. Εξαγωγή χαρακτηριστικών



Εικόνα 1: σήμα του μέτρου της επιτάχυνσης κατά τη διάρκεια του χρόνου και ανά δραστηριότητα

Πίνακας 1: υπόμνημα εικόνας 1

0 – 8.999	'walking'
9.000 – 17.999	'standing'
18.000 – 26.999	'jogging'
27.000 – 35.999	'sitting'
36.000 – 44.999	'biking'
45.000 – 55.999	'upstairs'
54.000 – 62.999	'downstairs'

Με τη βοήθεια μιας δομής επανάληψης “for” επιλέξαμε παράθυρα 20 δευτερολέπτων του σήματος, με βήμα 1 δευτερόλεπτο. Επειδή η συχνότητα δειγματοληψίας ήταν 50 Hertz, κάθε παράθυρο περιείχε 1000 τιμές (50x20) ενώ το βήμα ήταν 50 τιμές (50x1), και με αυτόν τον τρόπο θέσαμε τα όρια της δομής επανάληψης.

Σε κάθε παράθυρο αντιστοιχίσαμε μία δραστηριότητα, αυτήν που εμφανιζόταν τις πιο πολλές φορές στα δείγματα του κάθε παραθύρου (πλειοψηφία).

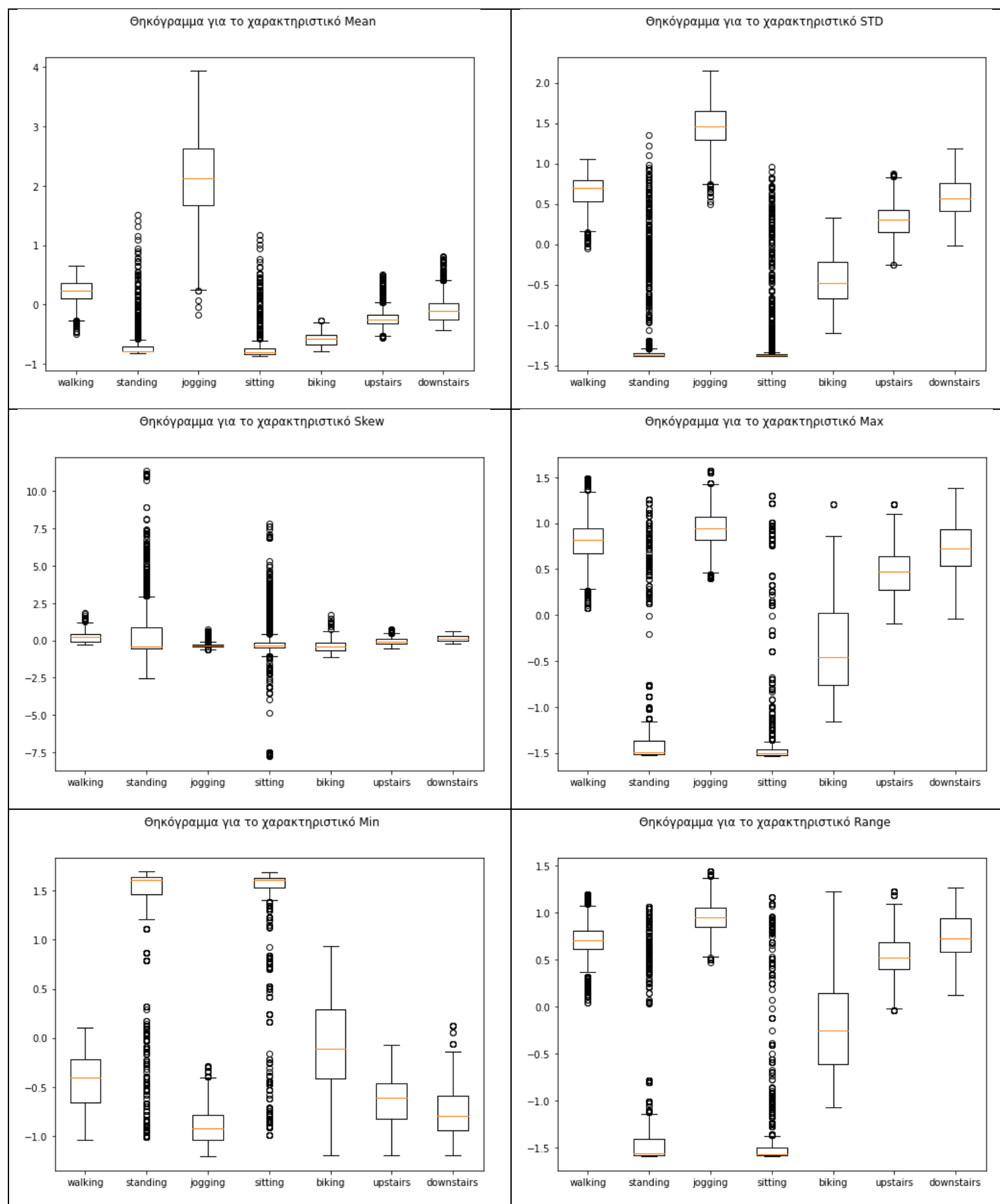
Από κάθε παράθυρο υπολογίσαμε τα χαρακτηριστικά που ζητούνταν: μέση τιμή, τυπική απόκλιση, ασυμμετρία κατανομής, μέγιστη και ελάχιστη τιμή, εύρος, εκτίμηση ισχύος φάσματος με τη μέθοδο του Welch. Το τελευταίο χαρακτηριστικό χρειαζόταν περαιτέρω επεξεργασία καθώς επέστρεφε παραπάνω από μία εξόδους (outrput), συγκεκριμένα δύο πίνακες: έναν με τις τιμές συχνοτήτων των δειγμάτων και έναν με το φάσμα ισχύος της τιμής εισόδου (input). Αυτό που κάναμε ήταν να επιλέξουμε 8 τιμές, αυτές που αντιστοιχούσαν στις μέγιστες τιμές φάσματος ισχύος, για τα 8 ίσα διαστήματα στα οποία χωρίσαμε τις αντίστοιχες τιμές (από τον δεύτερο πίνακα). Με αυτό τον τρόπο παρατηρήσαμε πως λαμβάναμε υψηλότερη ακρίβεια σε σχέση με το να επιλέγαμε μία μόνο μέγιστη τιμή φάσματος.

Χαρακτηριστικά	Περιγραφή
Mean	Μέσος όρος
Std	Τυπική απόκλιση
Skew	Ασυμμετρία κατανομής
Max	Μέγιστη τιμή
Min	Ελάχιστη τιμή
Range	Διαφορά μεταξύ μέγιστης και ελάχιστης τιμής
Welch1, ... , Welch8	Μέγιστες τιμές της ισχύος φάσματος με τη μέθοδο Welch ανά 8 ίσα διαστήματα

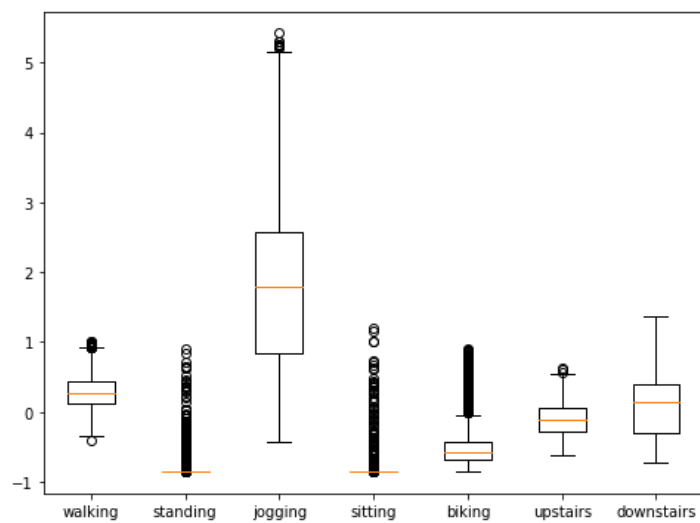
Τέλος, κανονικοποιήσαμε το κάθε χαρακτηριστικό από τα παραπάνω αφαιρώντας τη μέση τιμή και διαιρώντας με την τυπική απόκλιση.

2.1 Θηκογράμματα (boxplots)

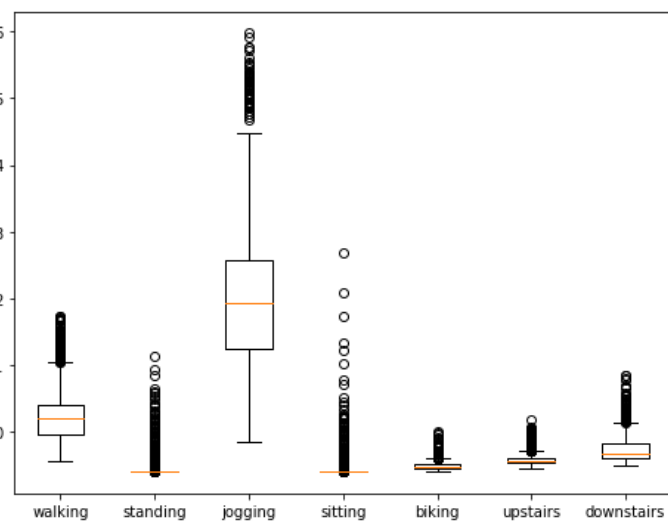
Στην συνέχεια θα ήταν χρήσιμο να εξετάσουμε το βαθμό διαφοροποίησης μεταξύ των δραστηριοτήτων ανά χαρακτηριστικό. Για αυτό τον σκοπό επιλέγουμε την απεικόνιση θηκογραμμάτων (boxplot) για κάθε χαρακτηριστικό ανά δραστηριότητα.



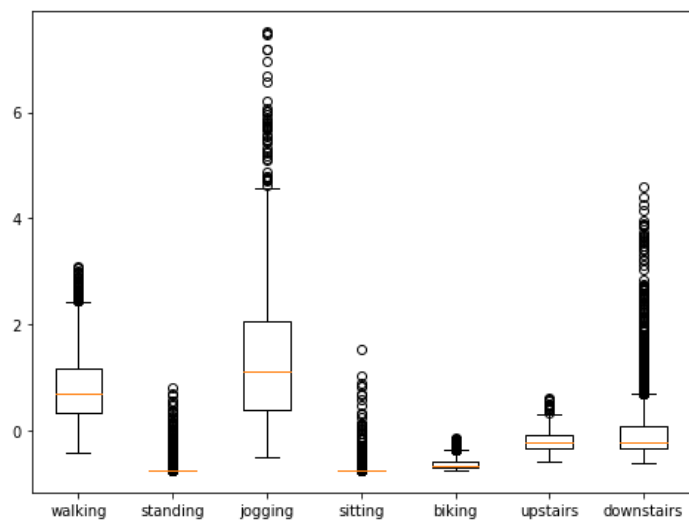
Θηκόγραμμα για το χαρακτηριστικό Welch1



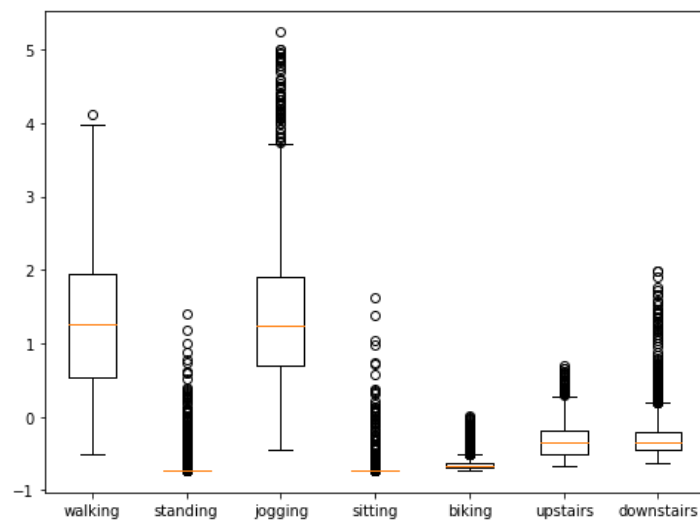
Θηκόγραμμα για το χαρακτηριστικό Welch2



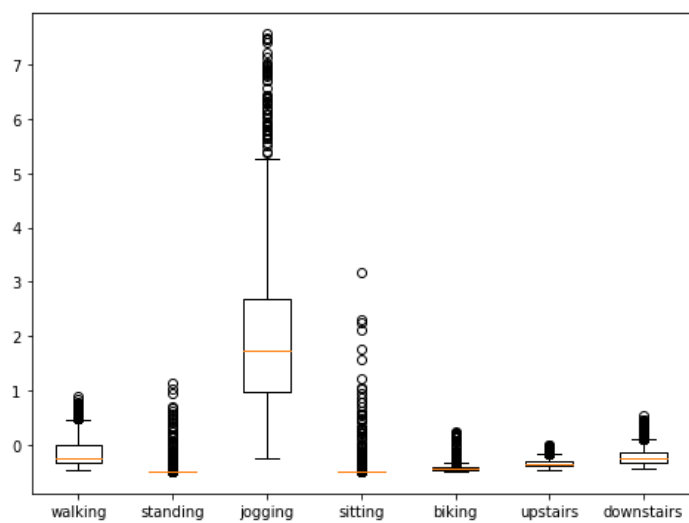
Θηκόγραμμα για το χαρακτηριστικό Welch3



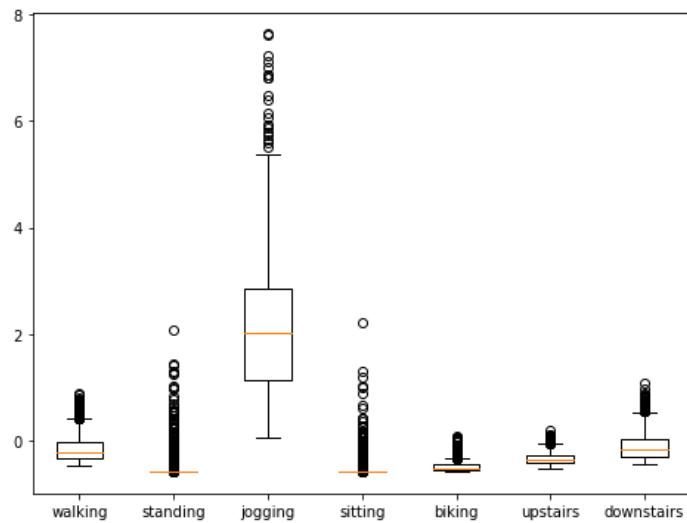
Θηκόγραμμα για το χαρακτηριστικό Welch4

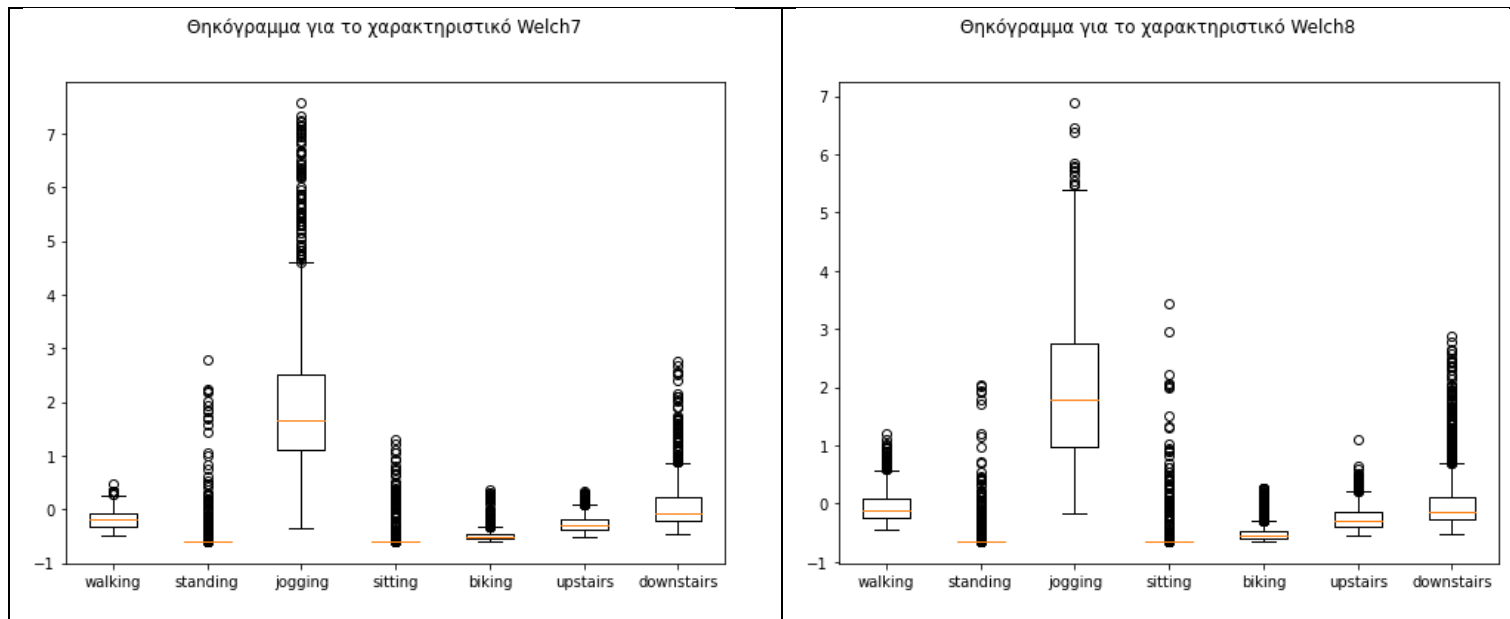


Θηκόγραμμα για το χαρακτηριστικό Welch5



Θηκόγραμμα για το χαρακτηριστικό Welch6





Μια άμεση παρατήρηση είναι ότι οι δραστηριότητες *'standing'* και *'sitting'* δεν δείχνουν να διαφέρουν σημαντικά σε κανένα χαρακτηριστικό, ενώ λιγότερο αλλά εξίσου σημαντικά δείχνουν να μοιάζουν και τα θηκογράμματα των *'upstairs'* και *'downstairs'*. Αντιθέτως, οι υπόλοιπες δραστηριότητες δείχνουν να διαφέρουν στα περισσότερα χαρακτηριστικά που επιλέξαμε. Τα αποτελέσματα αυτά θα επιβεβαιώσουμε – αναλύσουμε παρακάτω.

3. Εκπαίδευση μοντέλων ταξινόμησης

Ακολουθούμε τη μέθοδο αξιολόγησης Leave-One-Subject-Out. Εκπαιδεύουμε κάθε φορά με τα δεδομένα από τους συμμετέχοντες $1, \dots, i-1, i+1, \dots, 10$ και αξιολογούμε στα δεδομένα του συμμετέχοντα i , για i από 1 έως 10. Τελικά προκύπτει έτσι ένας πίνακας σύγχυσης (confusion matrix) για κάθε συμμετέχοντα. Αθροίζοντας τους πίνακες σύγχυσης προκύπτει ένας τελικός πίνακας από τον οποίο μπορούμε να υπολογίσουμε και τις μετρικές αξιολόγησης.

Για την εκπαίδευση χρησιμοποιούμε:

- 1) SVM με πυρήνα Radial Basis Function (RBF). (Διερεύνηση τιμών παραμέτρων C και γ)
- 2) Δένδρα απόφασης. (Διερεύνηση τιμών παραμέτρων)
- 3) Νευρωνικό δίκτυο MLP με 1 – 2 κρυφά στρώματα και χρήση ορμής (momentum). (Διερεύνηση τιμών παραμέτρων)

3.1 Μηχανές διανυσμάτων υποστήριξης (SVM με πυρήνα RBF)

1. Για την εκπαίδευση του μοντέλου, αρχικά χρησιμοποιήθηκαν οι μηχανές διανυσμάτων υποστήριξης SVM με χρήση συναρτήσεων πυρήνα Gauss (*Radial Basis Function* – RBF). Οι παράμετροι του μοντέλου που διερευνήθηκαν είναι η παράμετρος C (εκφράζει το βάρος του κόστους των λάθος ταξινομήσεων) και η θετική παράμετρος γ (εκφράζει το περιθώριο ταξινόμησης).

Αρχικά, πραγματοποιήθηκε δοκιμή με *Randomized Search* και στη συνέχεια έγινε πιο στοχευμένη αναζήτηση σε μικρότερο εύρος τιμών μέσω *Grid Search*. Τα αποτελέσματα από το *Grid Search* αποτυπώνονται στο παρακάτω γράφημα:

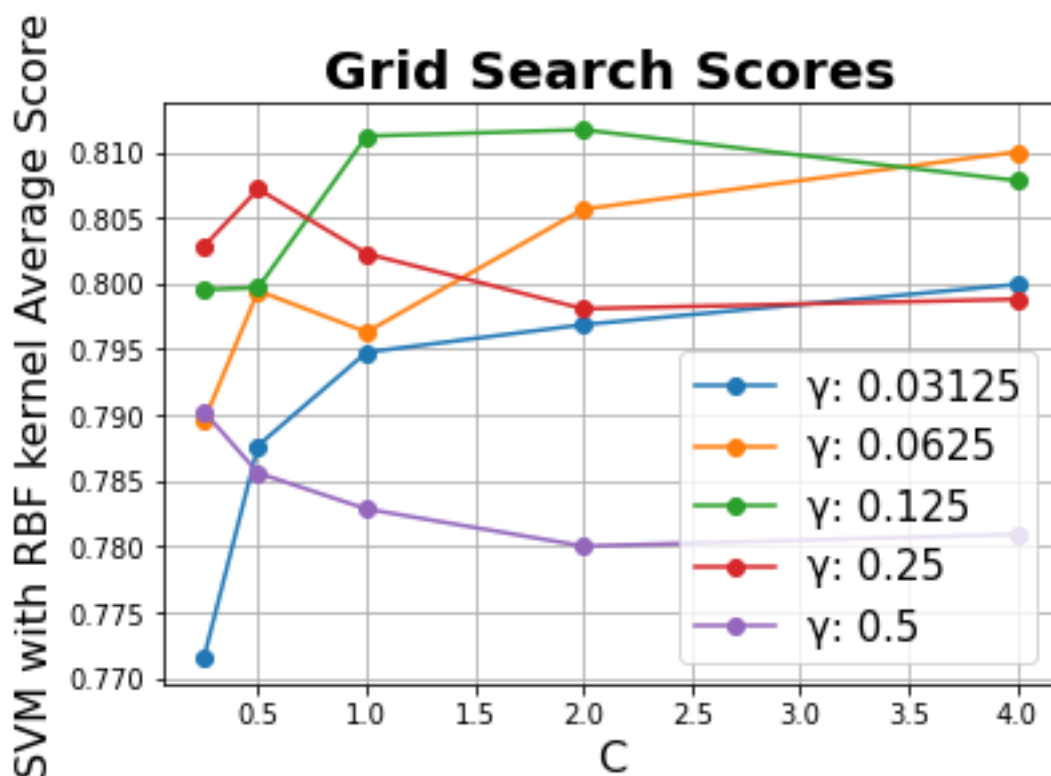


Figure 1: Ακρίβεια του μοντέλου για $C=\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ και $\gamma=\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}\}$

Παρατηρούμε ότι παίρνουμε τη μέγιστη ακρίβεια (**0.81**) για τις τιμές των παραμέτρων **$C=2$** , **$\gamma=0.125$** .

Η επίδραση κάθε παραμέτρου ξεχωριστά στην ακρίβεια του μοντέλου φαίνεται στα παρακάτω γραφήματα:

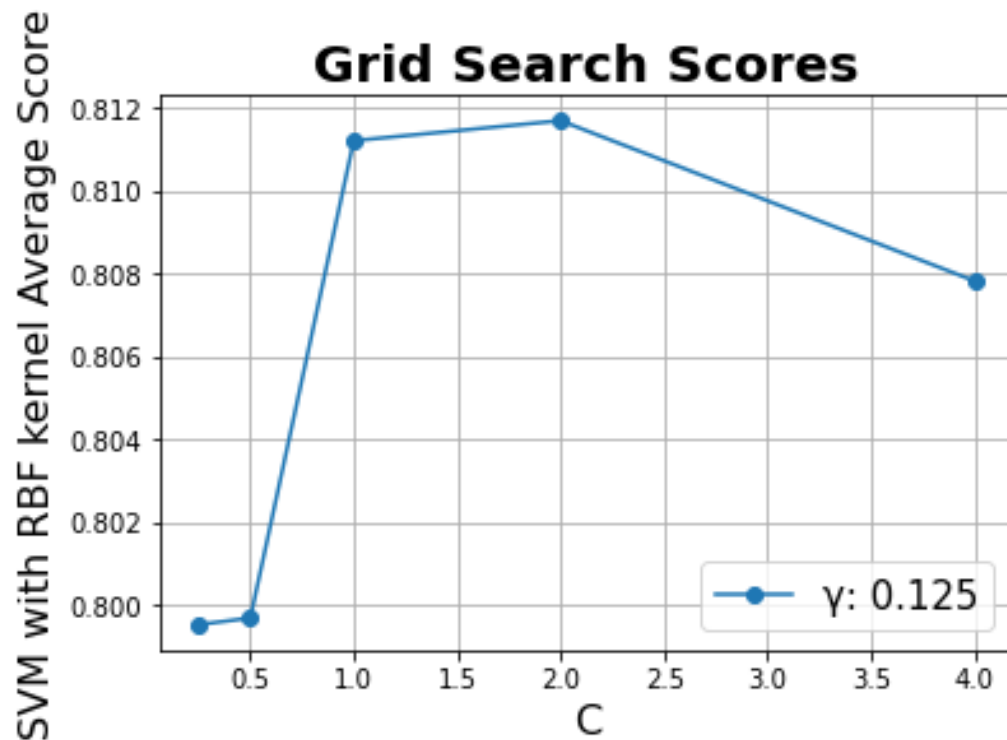


Figure 2: Επίδραση στην ακρίβεια του μοντέλου της παραμέτρου C

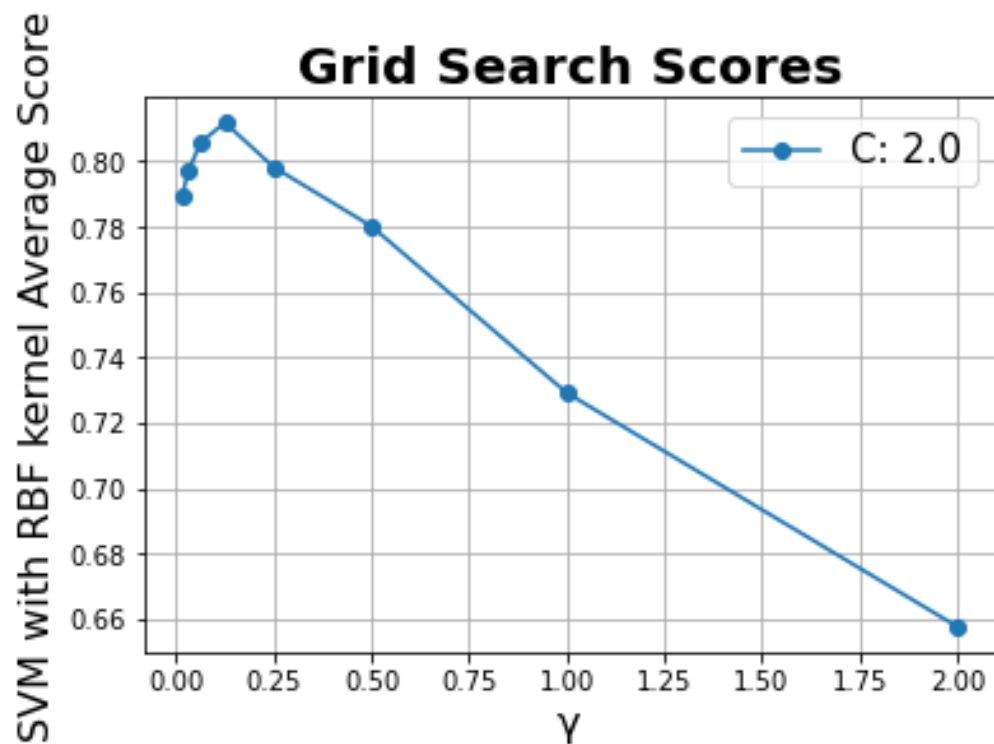


Figure 3: Επίδραση στην ακρίβεια του μοντέλου της παραμέτρου γ

2. Οι τιμές των παραμέτρων $C=2$, $\gamma=0,125$ δίνουν την υψηλότερη ακρίβεια και όσο το δυνατόν λιγότερη υπερπροσαρμογή (*overfitting*) στα δεδομένα.

Καταρχάς, θα εξετάσουμε την υπερπροσαρμογή στα δεδομένα, απεικονίζοντας σε γράφημα την ακρίβεια στο σύνολο εκπαίδευσης (*train test*) και στο σύνολο δοκιμής (*test set*) ανά επανάληψη που έγινε με τη μέθοδο *Leave One Subject Out* (10 συνολικά).

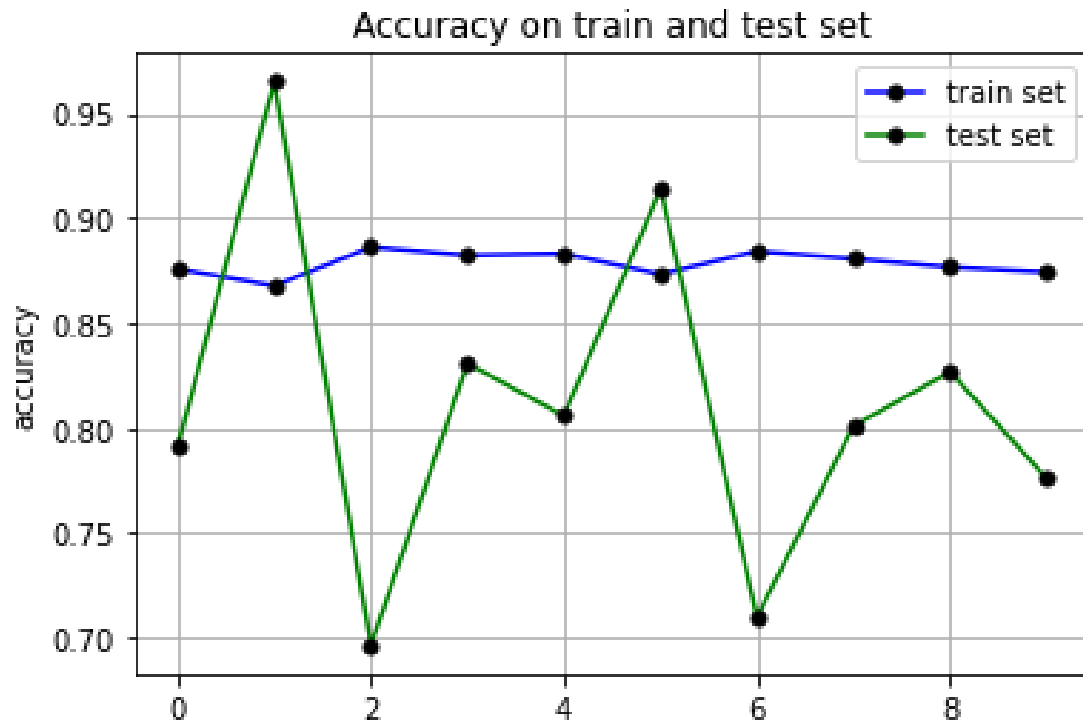


Figure 4: Ορθότητα ταξινόμησης για κάθε επανάληψη της μεθόδου *LOSO*

Παρατηρούμε ότι η ακρίβεια του συνόλου εκπαίδευσης δεν είναι πολύ κοντά στο 1, που θα ήταν μια σοβαρή ένδειξη υπερπροσαρμογής. Αντιθέτως η ακρίβεια του συνόλου εκπαίδευσης παραμένει σταθερή για κάθε επανάληψη, περίπου στο **0,875**.

Σε δύο επαναλήψεις έχουμε υψηλότερη ακρίβεια στο σύνολο δοκιμής (*test set*) σε σχέση με το σύνολο εκπαίδευσης, αλλά είναι μεμονωμένες περιπτώσεις και αποκλείουμε την πιθανότητα να προκλήθηκαν λόγω υπερπροσαρμογής.

Γενικότερα, δεν παρατηρούμε έντονες ενδείξεις υπερπροσαρμογής του μοντέλου στα δεδομένα εκπαίδευσης.

Στην συνέχεια θα γίνει υπολογισμός του πίνακα σύγχυσης.

Λόγω του ότι η εκπαίδευση έγινε με τη μέθοδο *Leave One Subject Out*, (με τη χρήση μιας δομής επανάληψης *for* 10 φορές, μία για κάθε συμμετέχων) σε κάθε επανάληψη λαμβάναμε και έναν πίνακα σύγχυσης (confusion matrix) και στο τέλος αθροίσαμε τους 10 αυτούς πίνακες για να λάβουμε έναν μόνο πίνακα για το μοντέλο που δημιουργήσαμε

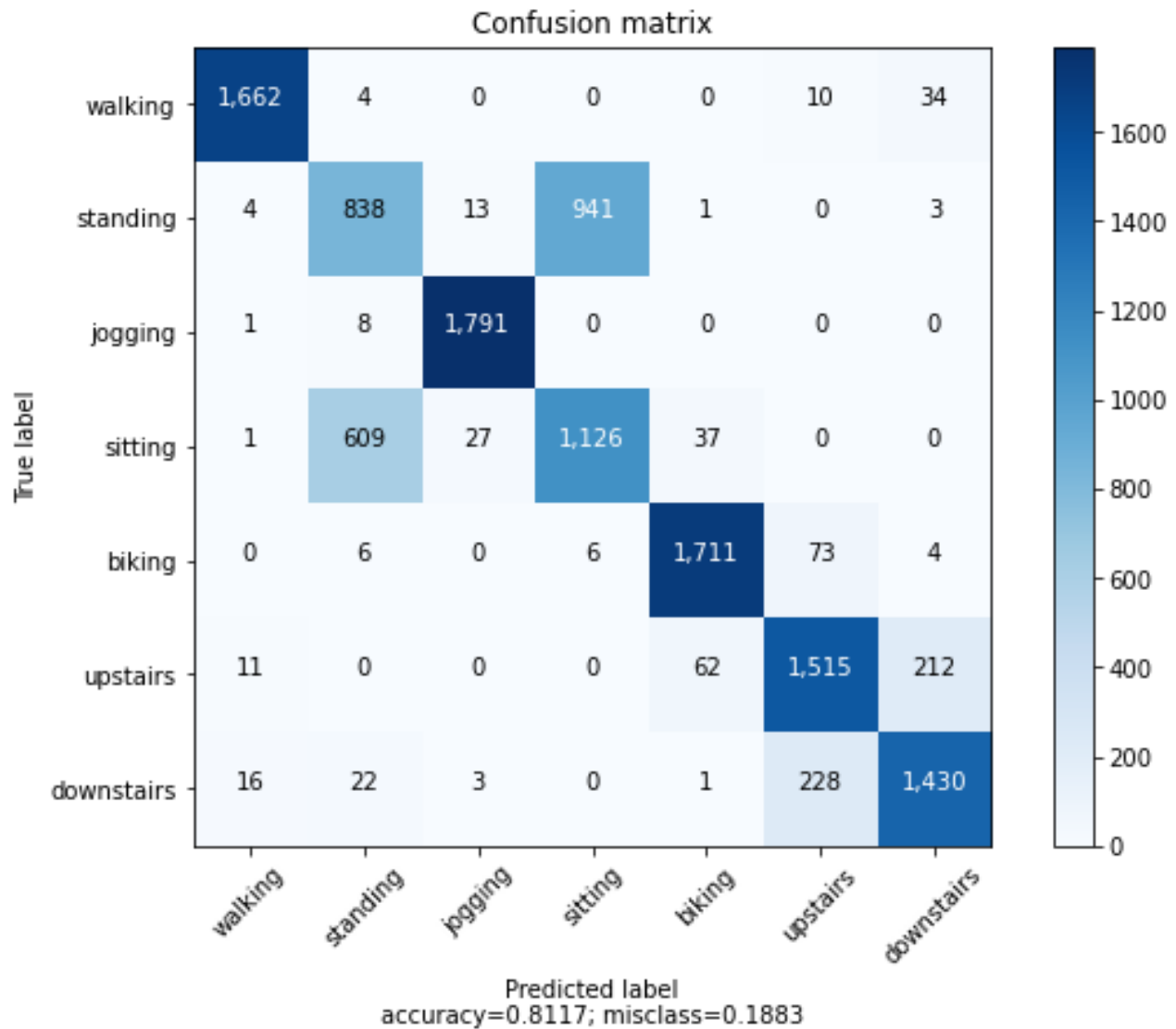


Figure 5: Πίνακας σύγχυσης (Confusion matrix) για $C=2$, $\gamma=0,125$

Υπολογίζουμε την ακρίβεια, από τον λόγο των ορθώς ταξινομημένων προτύπων προς το σύνολο των προτύπων:

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

$$Accuracy = \frac{1662 + 838 + 1791 + 1126 + 1711 + 1515 + 1430}{12410} = 0,8117$$

Κανονικοποιούμε τις τιμές στον πίνακα σύγχυσης και παίρνουμε το κριτήριο *Recall* ανά περίπτωση, έτσι ώστε να κατανοήσουμε ευκολότερα τα αποτελέσματα του πίνακα σύγχυσης:

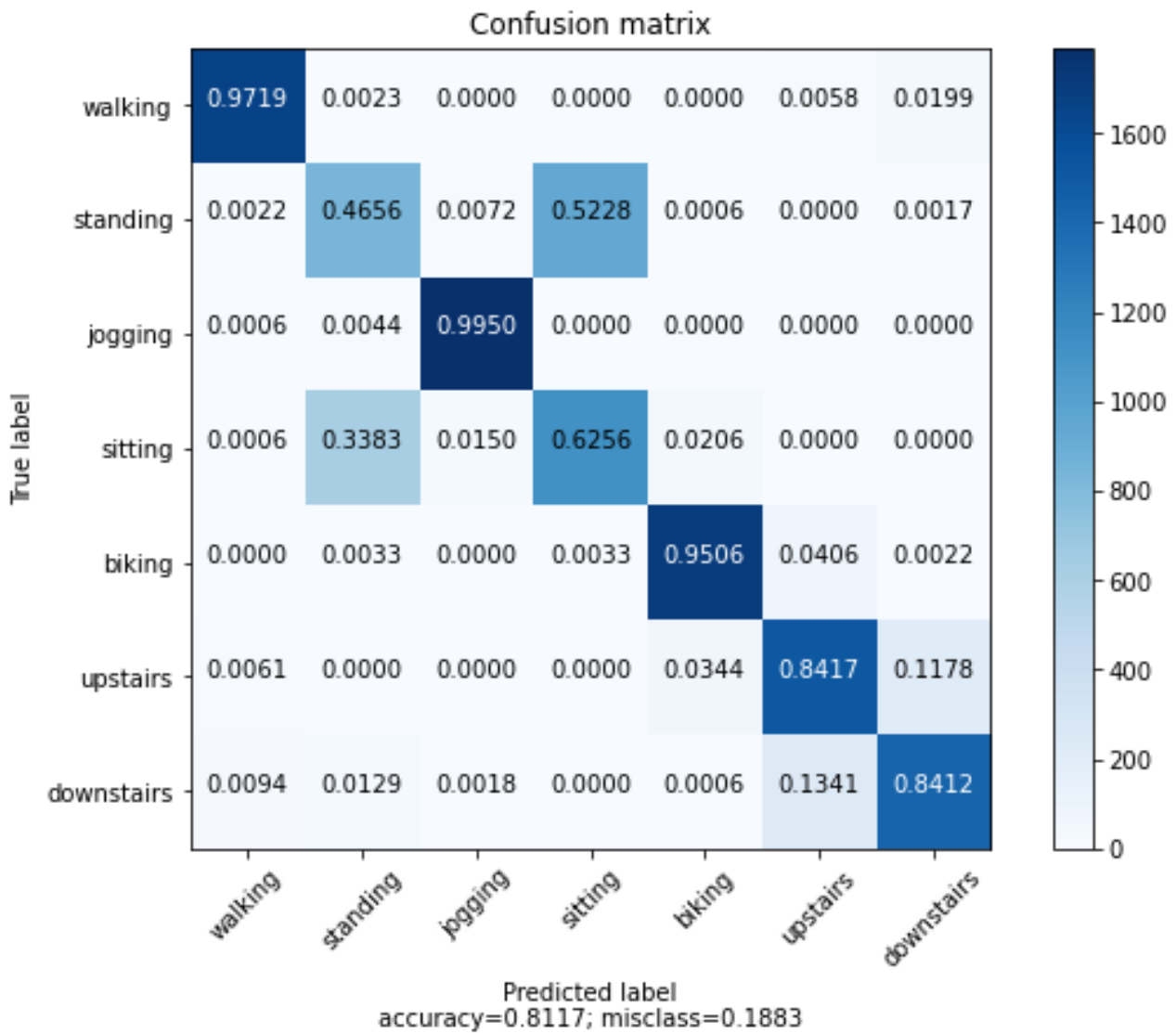


Figure 6: Κανονικοποιημένος πίνακας σύγχυσης (Confusion matrix) για $C=2$, $\gamma=0,125$

Στην συνέχεια παρουσιάζουμε τα αποτελέσματα της συνάρτησης **classification_report** που παρουσιάζουν συνοπτικά όλες τις βασικές μετρικές που θέλουμε , τόσο για κάθε κλάση ξεχωριστά, όσο και *macro-averaged*.

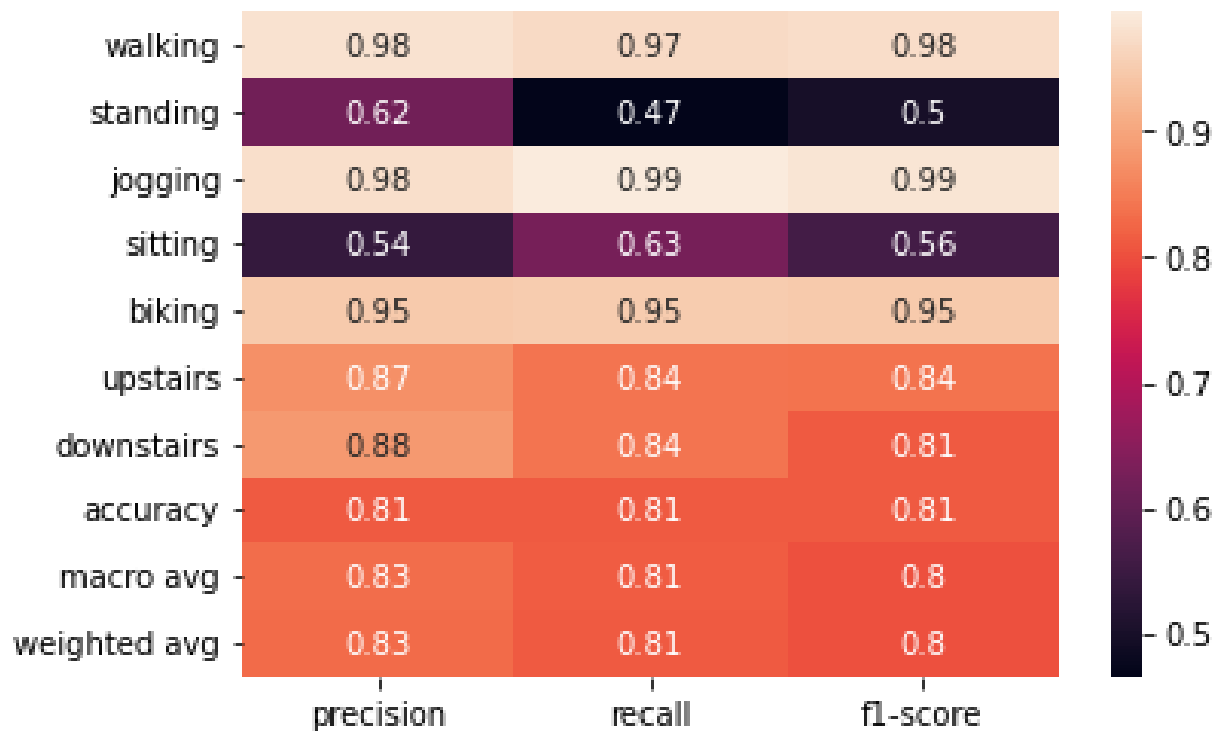


Figure 7: Classification report για $C=2$, $\gamma=0.125$

Στο **classification_report** εμφανίζονται τα κριτήρια *Precision* (Ευστοχία), *Recall* (Ανάκληση) και **F1-score** ανά κάθε κλάση ξεχωριστά:

Precision (Ευστοχία):

$$Precision = \frac{TP}{POSITIVE}$$

Recall (Ανάκληση):

$$Recall = \frac{TP}{Κλάση\ i}$$

F1-score:

$$F - measure = \frac{Precision * Recall}{(Precision + Recall)/2}$$

3. Παρατηρούμε λοιπόν από το **classification_report** ότι οι μέσες τιμές του **f1-score** ανά δραστηριότητα (μέση τιμή από τις 10 τιμές, μία για κάθε επανάληψη) είναι υψηλότερες από 0,80 με εξαίρεση τις δραστηριότητες **standing (0,50)** και **sitting (0,56)**. Ταυτόχρονα με τη βοήθεια του πίνακα σύγχυσης, παρατηρούμε ότι οι ίδιες δραστηριότητες (**standing**,

sitting) συγχέονται μεταξύ τους από τον αλγόριθμο. Το ίδιο συμβαίνει, σε αρκετά μικρότερο βαθμό όμως, μεταξύ των δραστηριοτήτων **upstairs** και **downstairs**.

Στο στάδιο της εξαγωγής χαρακτηριστικών είχε γίνει απεικόνιση των θηκογραμμάτων ανά δραστηριότητα σε κάθε χαρακτηριστικό. Σε αυτές τις απεικονίσεις, οι δραστηριότητες *standing* και *sitting* διαφέραν ελάχιστα μεταξύ τους. Ενδεικτικά, παραθέτουμε την απεικόνιση για το χαρακτηριστικό της Μέσης τιμής παραθύρου -Mean.

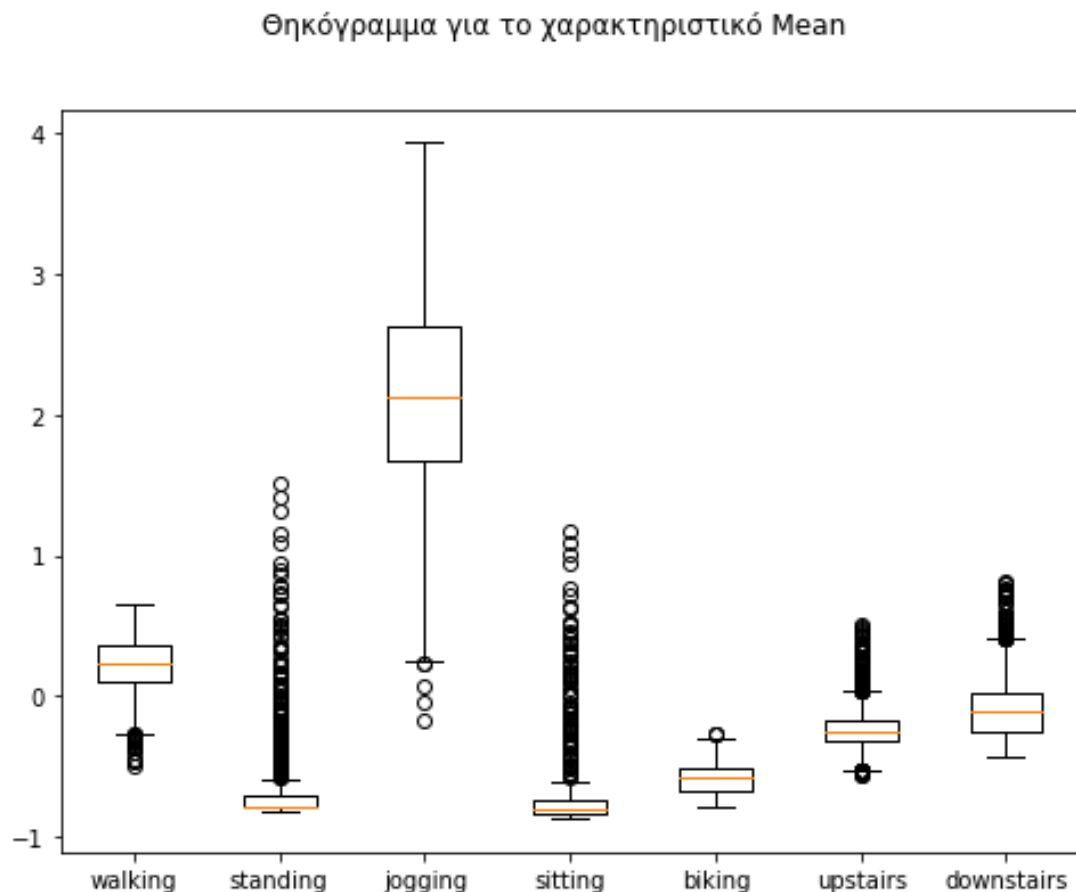
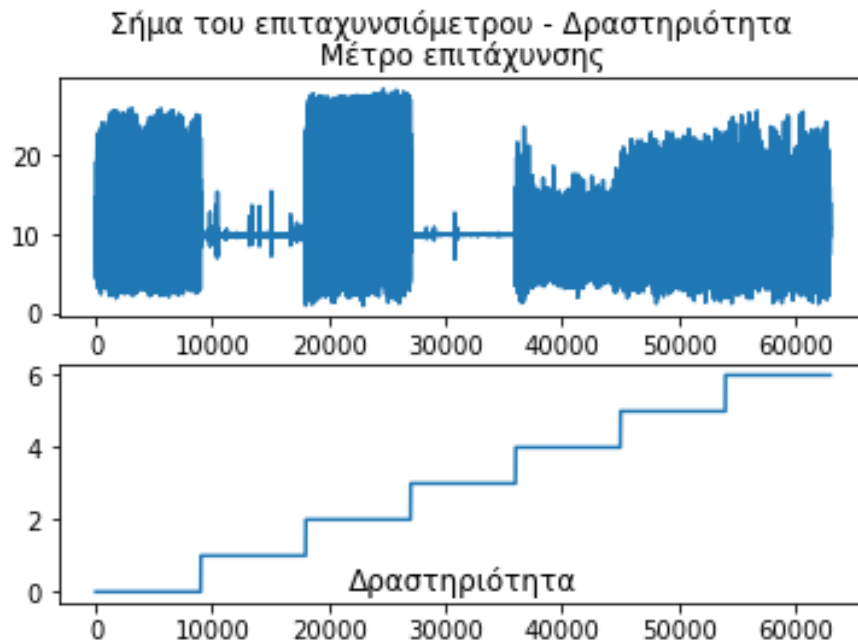


Figure 8: Θηκόγραμμα για το χαρακτηριστικό Mean

Αντίστοιχη εικόνα παρατηρούμε και στα θηκογράμματα όλων των υπόλοιπων χαρακτηριστικών. Συνεπώς, τα θηκογράμματα μας δίνουν μια πρώτη εξήγηση της αιτίας που ο αλγόριθμος συγχέει τις δραστηριότητες *standing* – *sitting*.

Αντιστοίχως, αν εξετάσουμε την απεικόνιση του μέτρου της επιτάχυνσης σε σχέση με το χρόνο (raw data):



Παρατηρούμε και πάλι ότι το σήμα που παίρνουμε και για τις δύο δραστηριότητες χαρακτηρίζεται από μια αντίστοιχη «στασιμότητα», που κάνει δύσκολη την διάκριση τους.

4. Για να ξεπεράσουμε τη δυσκολία του αλγόριθμου να ξεχωρίσει τις δραστηριότητες *standing*, *sitting* και *upstairs*, *downstairs* εφαρμόζουμε ομαδοποίηση τους ως νέες δραστηριότητες ***standing-sitting*** και ***upstairs-downstairs***.

Στην συνέχεια μέσω της συνάρτησης *GridSearch* αναζητούμε τις βέλτιστες τιμές των παραμέτρων C-γ για τα ομαδοποιημένα δεδομένα:

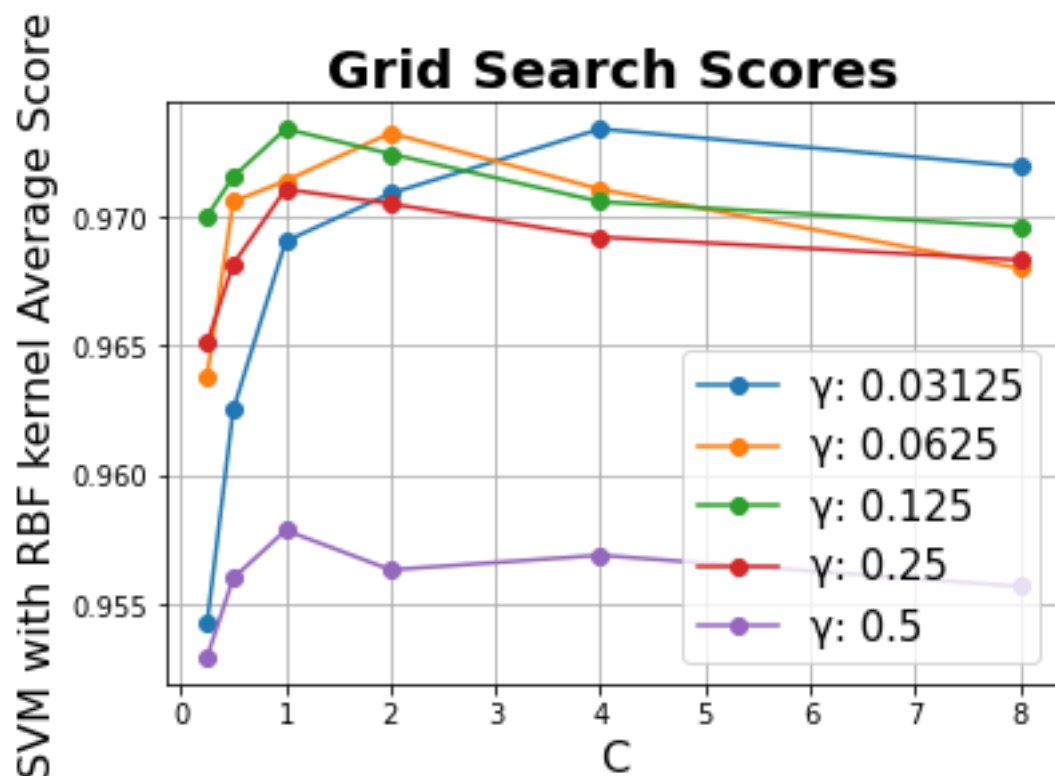


Figure 9: Ακρίβεια του μοντέλου για $C=\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$ και $\gamma=\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}\}$

Παρατηρούμε ότι παίρνουμε τη μέγιστη ακρίβεια (**0.97**) για τις τιμές των παραμέτρων **C=1**, **γ=0.125**, σημαντικά αυξημένη σε σχέση με τα μη ομαδοποιημένα δεδομένα.

➔ Υπολογίζουμε τον καινούργιο πίνακα σύγχυσης (άθροισμα πινάκων).

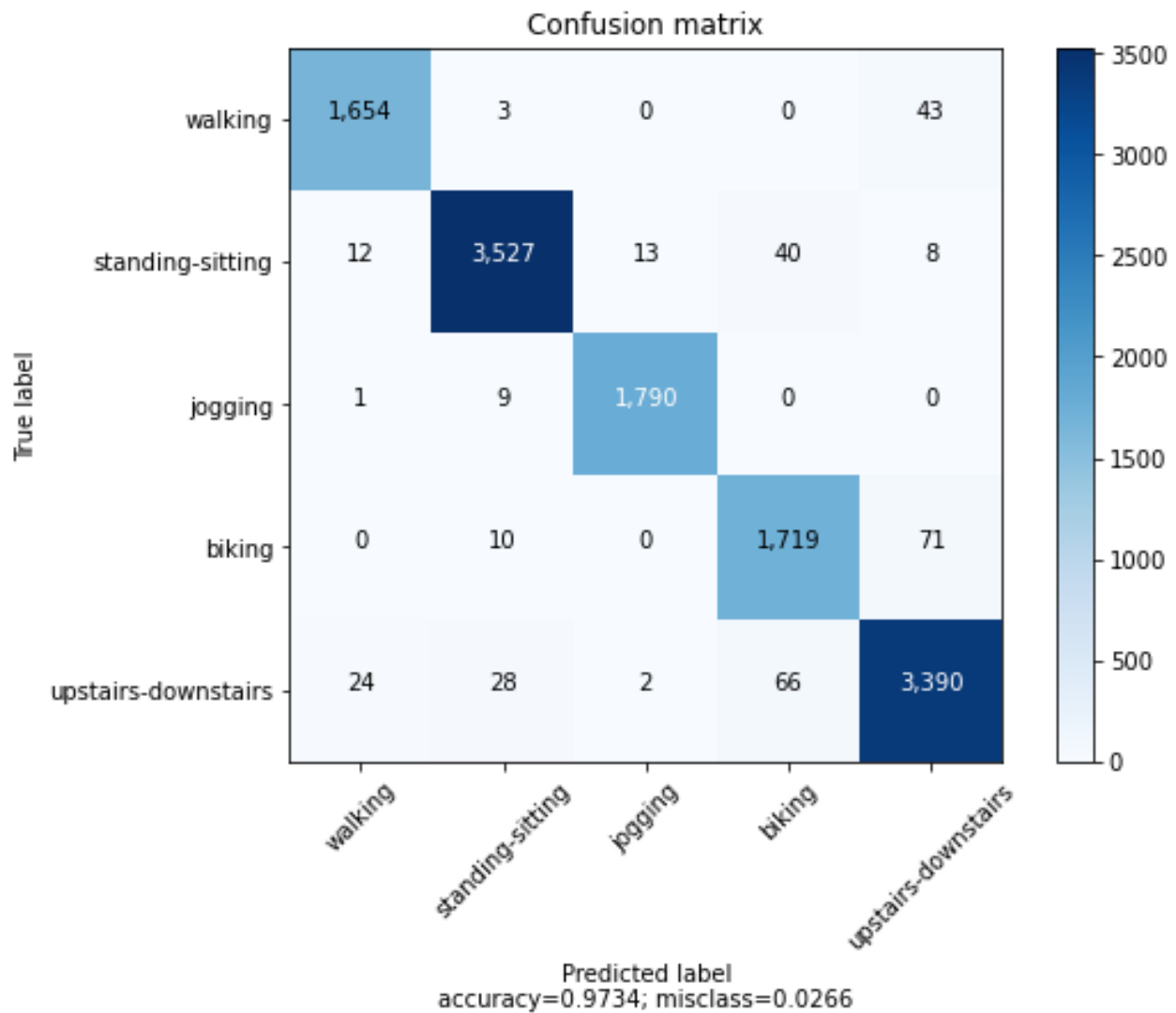


Figure 10: Πίνακας σύγχυσης (Confusion matrix) για C=1, γ=0,125

Υπολογίζουμε την ακρίβεια, από τον λόγο των ορθώς ταξινομημένων προτύπων προς το σύνολο των προτύπων:

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

$$Accuracy = \frac{1654 + 3527 + 1790 + 1719 + 3390}{12410} = 0,9734$$

Κανονικοποιούμε τις τιμές στον πίνακα σύγχυσης και παίρνουμε το κριτήριο Recall ανά περίπτωση, έτσι ώστε να κατανοήσουμε ευκολότερα τα αποτελέσματα του πίνακα σύγχυσης:

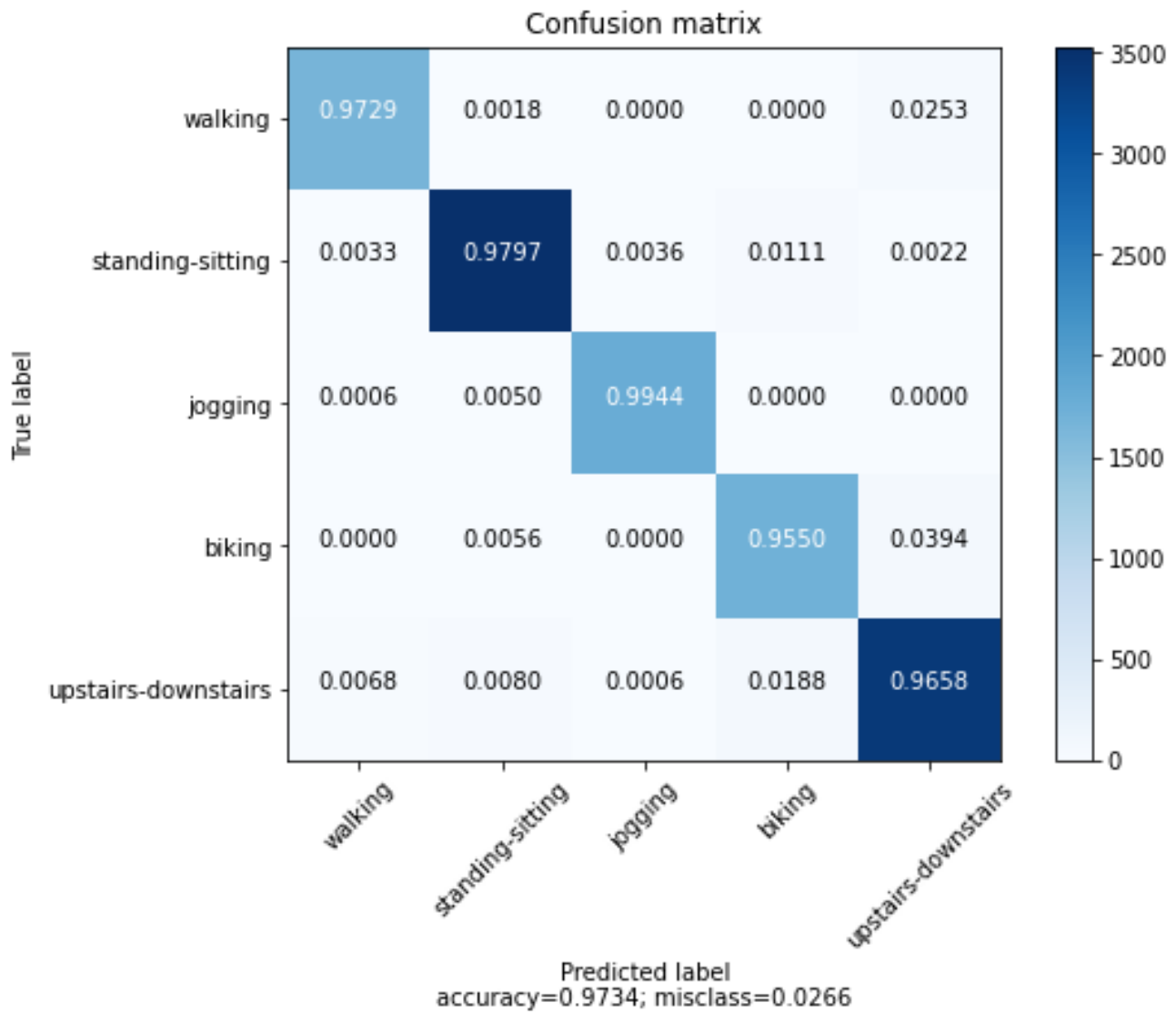
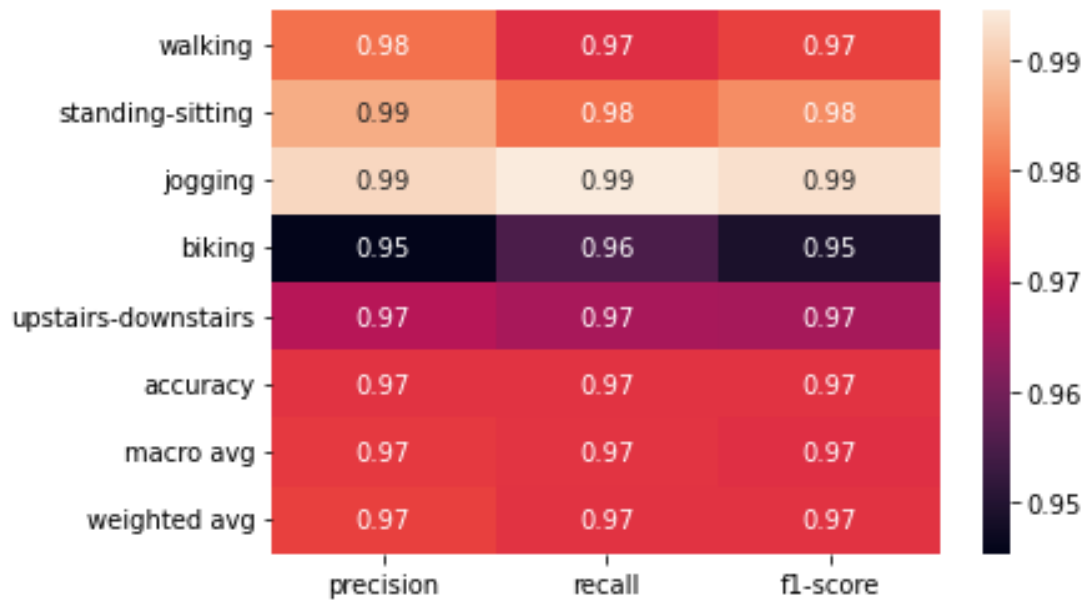


Figure 11: Κανονικοποιημένος πίνακας σύγχυσης (Confusion matrix) για $C=1$, $\gamma=0,125$

Παρατηρούμε ότι μετά την ομαδοποίηση των δραστηριοτήτων, **η ακρίβεια έχει αυξηθεί κατά πολύ.**

Στην συνέχεια παρουσιάζουμε τα αποτελέσματα της συνάρτησης `classification_report`:



Πλέον, οι τιμές του f1-score είναι πάνω από 0,95 για όλες τις δραστηριότητες!

2. Δένδρα αποφάσεων (Decision Trees)

Οι παράμετροι του μοντέλου που διερευνήθηκαν, οι τιμές που διερευνήθηκαν και οι βέλτιστες παράμετροι φαίνονται στον Πίνακα 1 παρακάτω.

Περιγραφή διαδικασίας: Αρχικά, πραγματοποιήθηκαν δοκιμές *Grid Search* και στη συνέχεια έγινε μια πιο στοχευμένη διερεύνηση των παραμέτρων χειροκίνητα ως εξής: όλες οι παράμετροι εκτός από την υπό διερεύνηση παράμετρο παραμένουν σταθερές και η άλλη τρέχει σε ένα εύρος τιμών. Τα αποτελέσματα που λαμβάνουμε, τα αποθηκεύουμε και παρατηρούμε την εξέλιξη της ακρίβειας συναρτήσει της υπό διερεύνησης παραμέτρου (βλέπε οπτικοποιήσεις παρακάτω). Επαναλαμβάνοντας αυτή τη διαδικασία πολλές φορές (*Grid Search* και επιλεγμένη διερεύνηση) πήραμε τελικά τη μέγιστη δυνατή τιμή ακρίβειας (**0.79**).

Οι τιμές αυτές δίνουν μεγαλύτερη τιμή ακρίβειας και λιγότερη κατά το δυνατόν υπερπροσαρμογή στα δεδομένα. Οι τιμές μέσης ακρίβειας *f1 score* (μέση τιμή από τις 10 τιμές, μία για κάθε επανάληψη) κυμαίνονται από 0.7 έως 0.79 περίπου, ενώ η υπερπροσαρμογή υποθέτει απόσταση μικρότερη του 0.05 της ακρίβειας του συνόλου εκπαίδευσης (*train set*) από το σύνολο δοκιμής (*test set*). Να σημειωθεί πως οι τιμές αυτές ακρίβειας που μετρήσαμε, κυμαίνονταν μεταξύ 0.5 και 0.9 (βλ. παρακάτω: *Ορθότητα Ταξινόμησης*)

Λόγω του ότι η εκπαίδευση έγινε με τη μέθοδο *Leave One Subject Out*, (με τη χρήση μιας δομής επανάληψης *for* 10 φορές, μία για κάθε συμμετέχων) σε κάθε επανάληψη λαμβάναμε και έναν πίνακα σύγχυσης (*confusion matrix*) και στο τέλος αθροίσαμε τους 10 αυτούς πίνακες για να λάβουμε έναν μόνο πίνακα για το μοντέλο που δημιουργήσαμε.

- Παρακάτω υπάρχουν γραφήματα – οπτικοποιήσεις όσων περιγράφηκαν παραπάνω καθώς και μια εφαρμογή του μοντέλου για μια συγκεκριμένη ομαδοποίηση των κλάσεων/ δραστηριοτήτων πρόβλεψης που «συγχέει» ο αλγόριθμος (ερώτηση 3, 4).

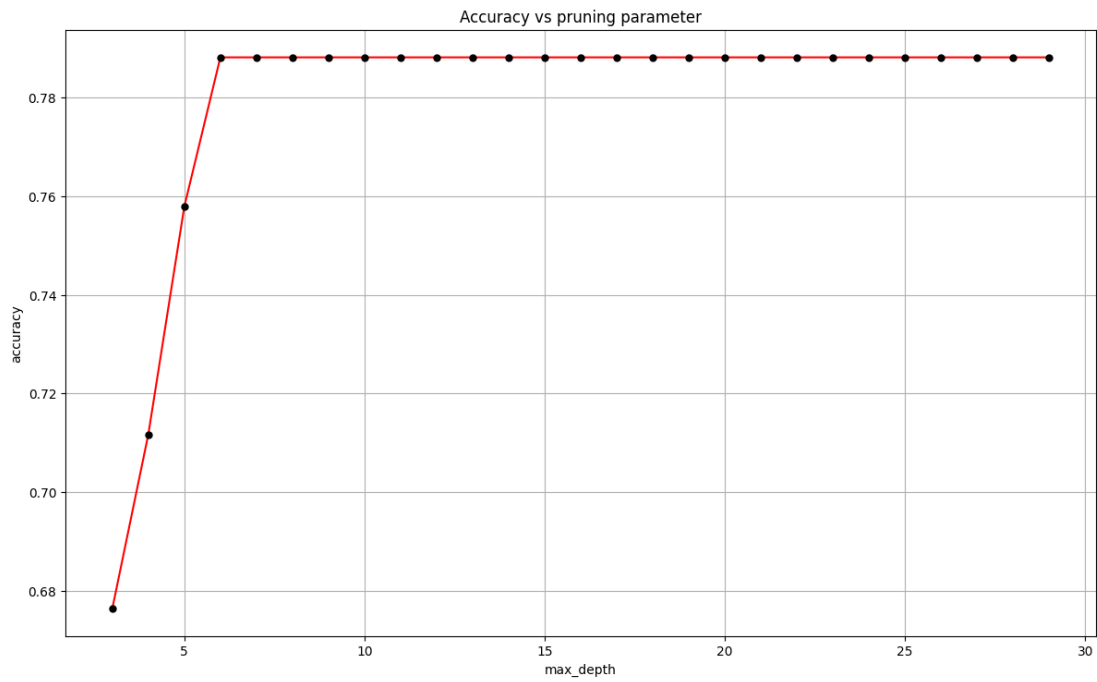
ΠΑΡΑΜΕΤΡΟΣ	ΛΕΙΤΟΥΡΓΙΑ	ΤΙΜΕΣ	ΒΕΛΤΙΣΤΗ ΤΙΜΗ
criterion	επιλογή συνάρτησης απώλειας	"gini", "entropy"	"entropy"
splitter	τρόπος διαχωρισμού κόμβων	"best", "random"	"best"
max_depth	επιλογή μέγιστου βάθους	3, 4, ..., 30	6
min_samples_leaf	ελάχιστος αριθμός προτύπων σε φύλλο	1, ..., 1000	346
min_samples_split	ελάχιστος αριθμός προτύπων για διαχωρισμό κόμβου	2, ..., 500	2
min_weight_fraction_leaf	ελάχιστο βάρος που λαμβάνεται υπόψη για τη δημιουργία φύλλων ¹	1	default
max_features	το πλήθος των χαρακτηριστικών που εξετάζεται στη διερεύνηση του καλύτερου διαχωρισμού ¹	None	default
random_state	έλεγχος της τυχαιότητας του εκτιμητή ¹	None	default
max_leaf_nodes	μέγιστο πλήθος φύλλων	7, ..., 30	15
min_impurity_decrease	έλεγχος «καθαρότητας» (impurity) ¹	0	default
class_weight	βάρη που σχετίζονται με τις κλάσεις ¹	None	default
ccp_alpha	παράμετρος Minimal Cost-Complexity Pruning	0, 0.001, ..., 0.1,	0.014

Πίνακας 2: Παράμετροι Δέντρων Απόφασης

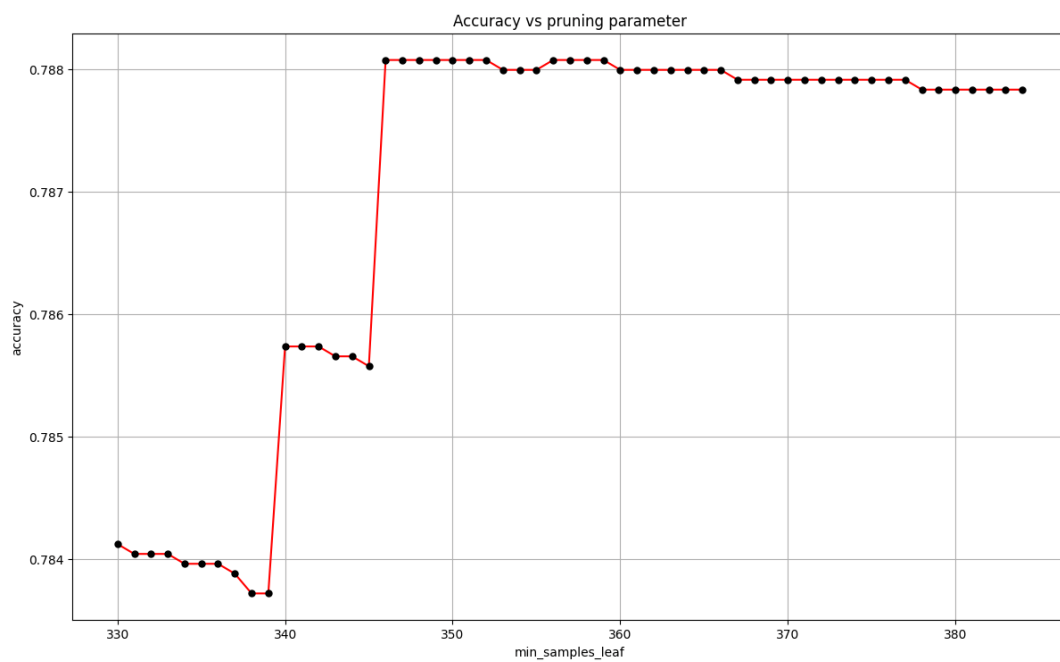
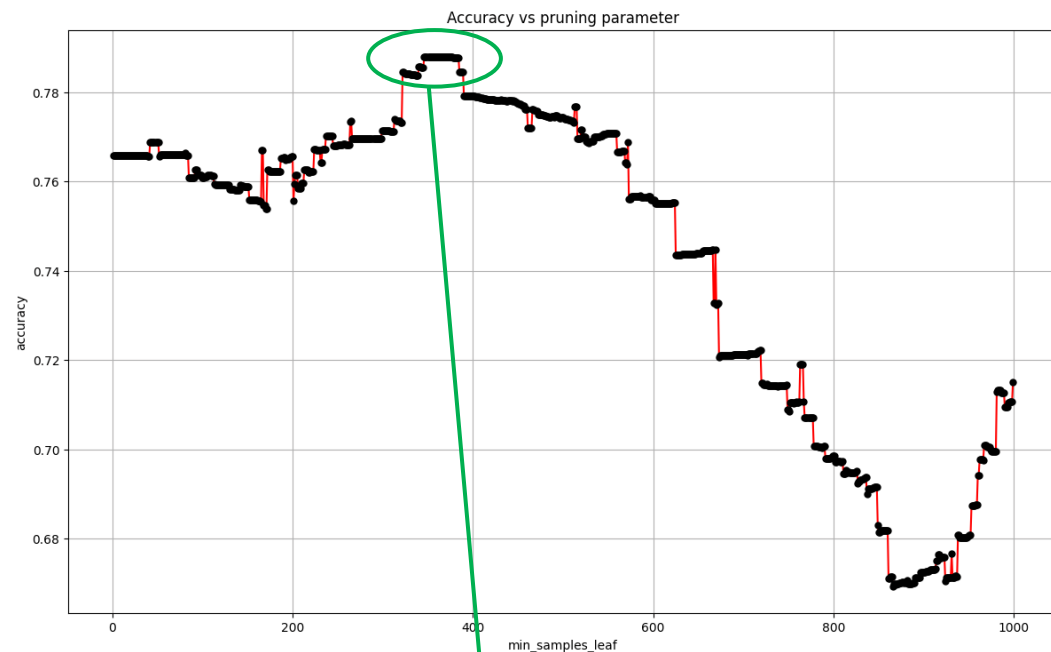
Με τιμή ακρίβειας: **0.79** (0.7880741337630942)

¹ Δεν έχουμε πληροφορία

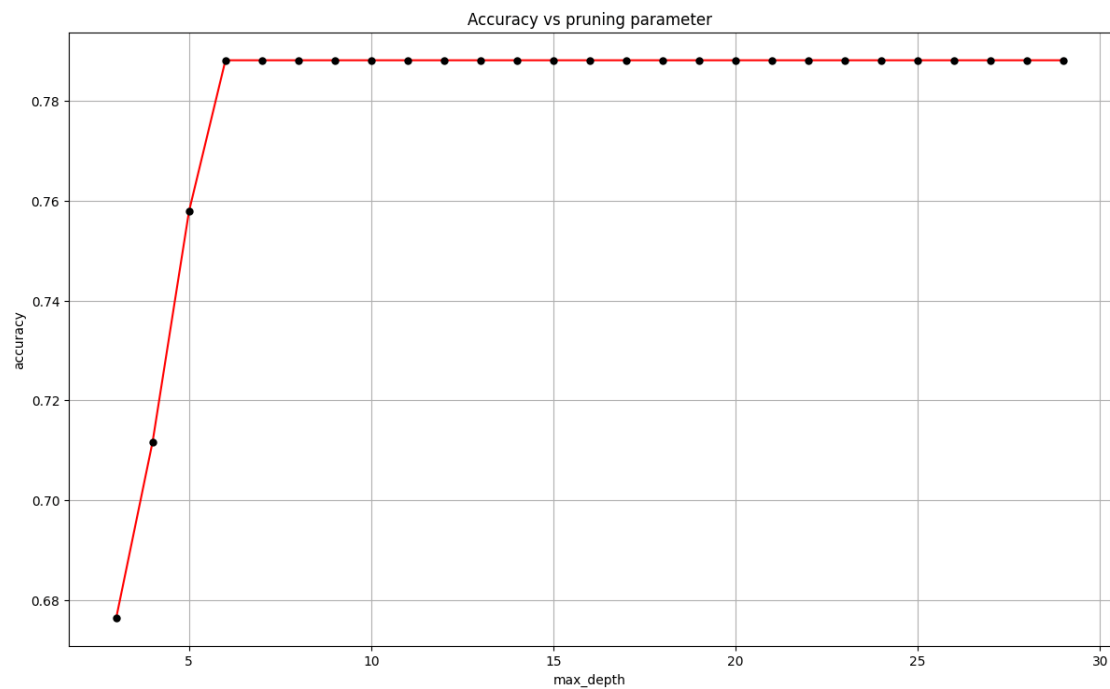
Παράμετρος: μέγιστο βάθος δέντρου



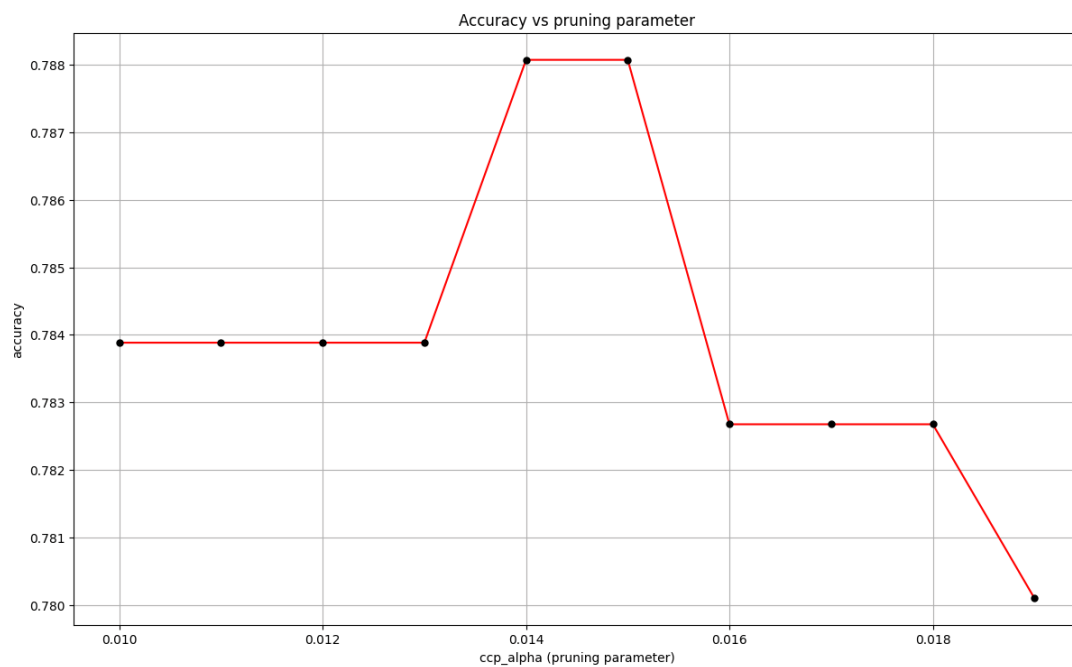
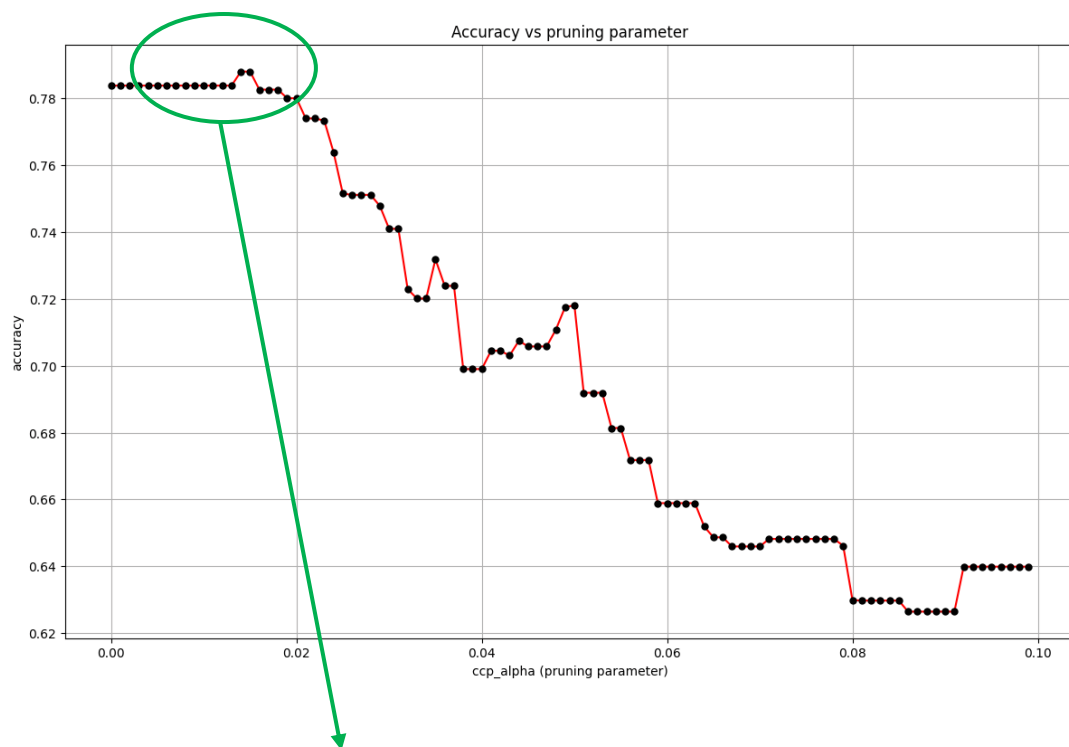
Παράμετρος: ελάχιστος αριθμός
προτύπων σε φύλλο



Παράμετρος: μέγιστο πλήθος φύλλων



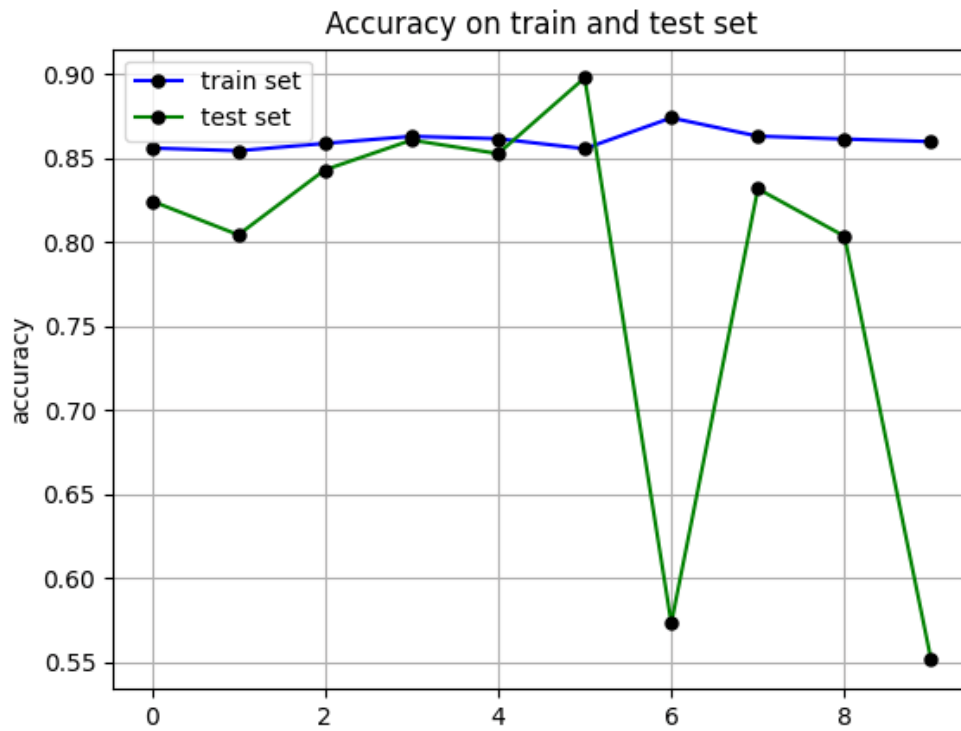
Παράμετρος: ccp_alpha



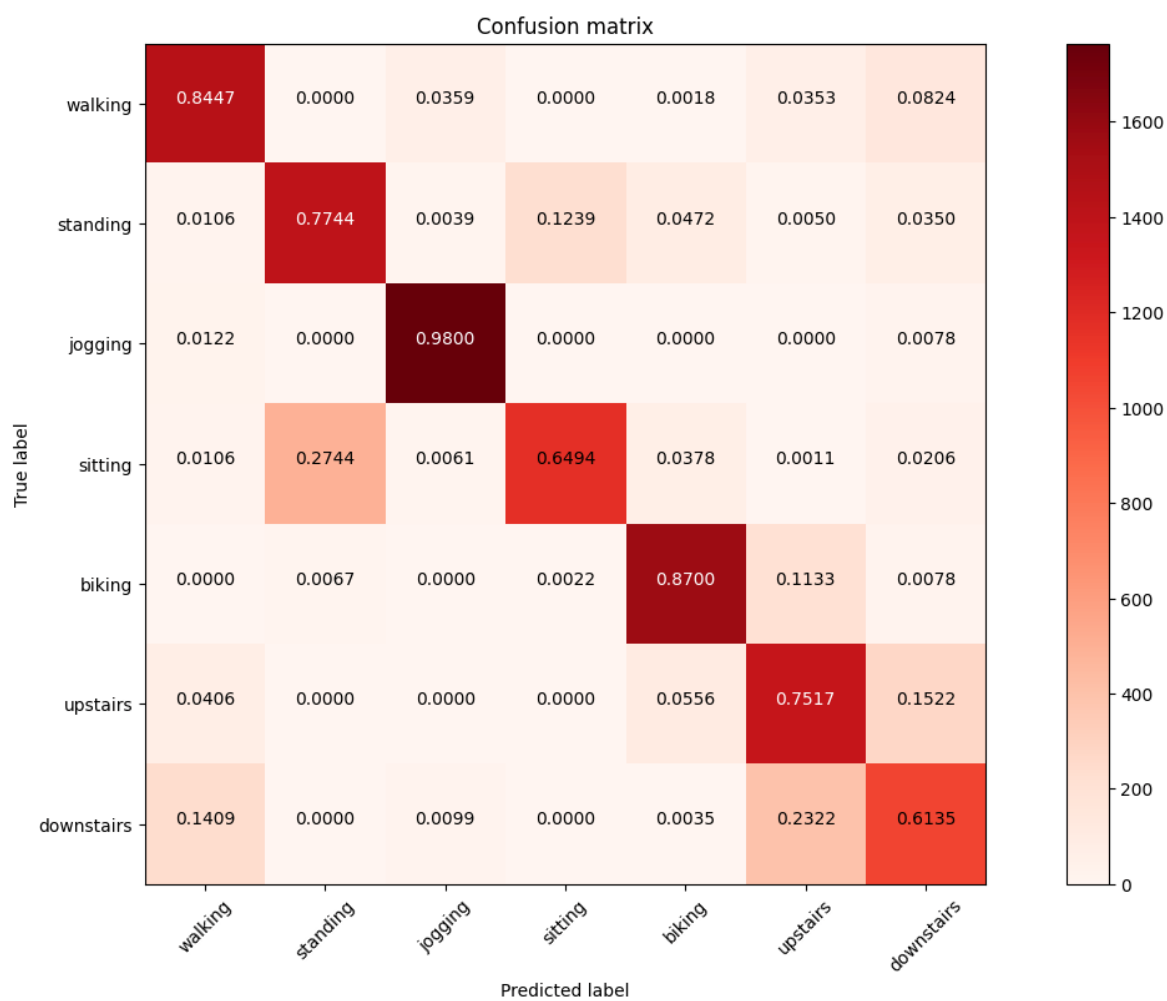
Πίνακας ακρίβειας



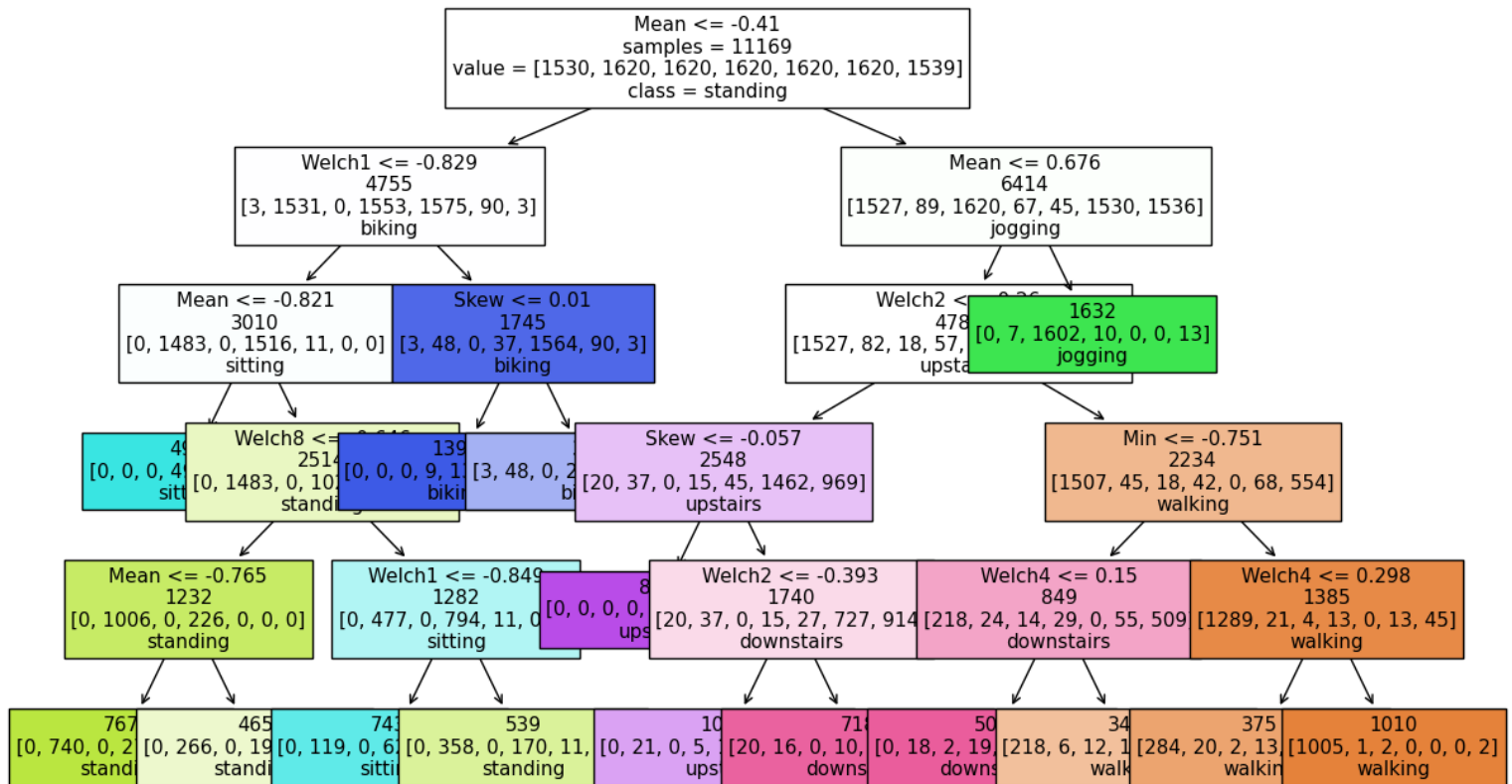
Ορθότητα Ταξινόμησης



Κανονικοποιημένος Confusion Matrix



Οπτικοποίηση του Δέντρου Απόφασης



Επιπλέον δοκιμές

Για τις παρακάτω παραμέτρους είχαμε επίσης καλή ακρίβεια, αλλά περισσότερο υπολογιστικό κόστος, καθώς οι τιμές των παραμέτρων ήταν πιο υψηλές, για αυτό και προτιμήσαμε να παραθέσουμε ως βέλτιστες τιμές αυτές που αναφέρθηκαν παραπάνω. Ωστόσο, οι τιμές αυτές των παραμέτρων μας έδιναν εξίσου καλή ακρίβεια.

ΠΑΡΑΜΕΤΡΟΣ	ΤΙΜΕΣ	ΒΕΛΤΙΣΤΗ ΤΙΜΗ
criterion	"gini", "entropy"	"entropy"
splitter	"best", "random"	"best"
max_depth	3, 4, ..., 30	29
min_samples_leaf	1, ..., 1000	420
min_samples_split	2, ..., 500	600
max_leaf_nodes	7, ..., 30	14
ccp_alpha	0, 0.001, ..., 0.1,	0.005

Με τιμή ακρίβειας: **0.78** (0.7837228041901693)

Σύγχυση Δραστηριοτήτων



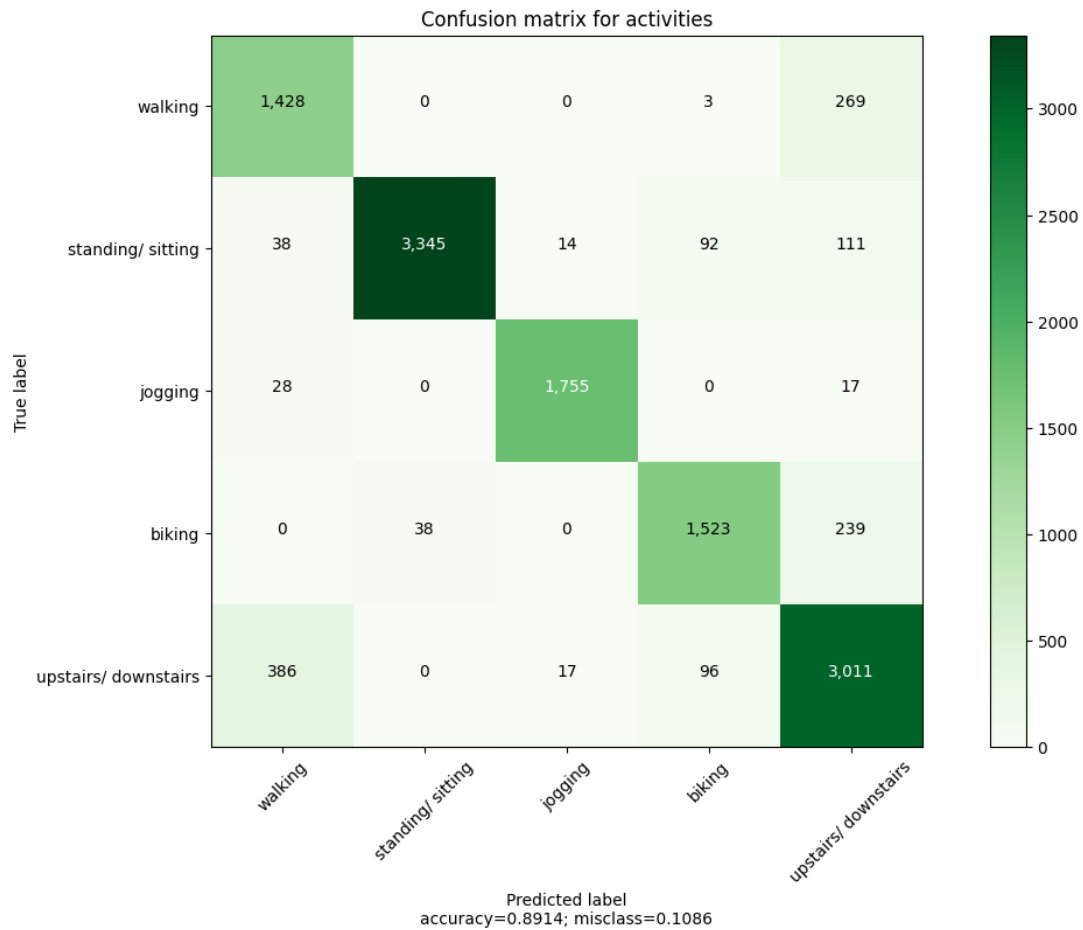
Σύμφωνα με τον Πίνακα Σύγχυσης ο ταξινομητής «μπερδεύει» τις εξής δραστηριότητες:

standing ~ sitting

upstairs ~ downstairs

Μια πιθανή εξήγηση για το ότι το μοντέλο μπερδεύει τα *standing*, *sitting* είναι η στατικότητα των δραστηριοτήτων αυτών: η μόνη διαφορά τους είναι το ύψος του συμμετέχοντα. Όσον αφορά τα *upstairs*, *downstairs* είναι όμοια ως προς την ταχύτητα και φορά, αλλά διαφέρουν ως προς την διεύθυνση κίνησης.

Ο νέος πίνακας σύγχυσης που λαμβάνουμε:



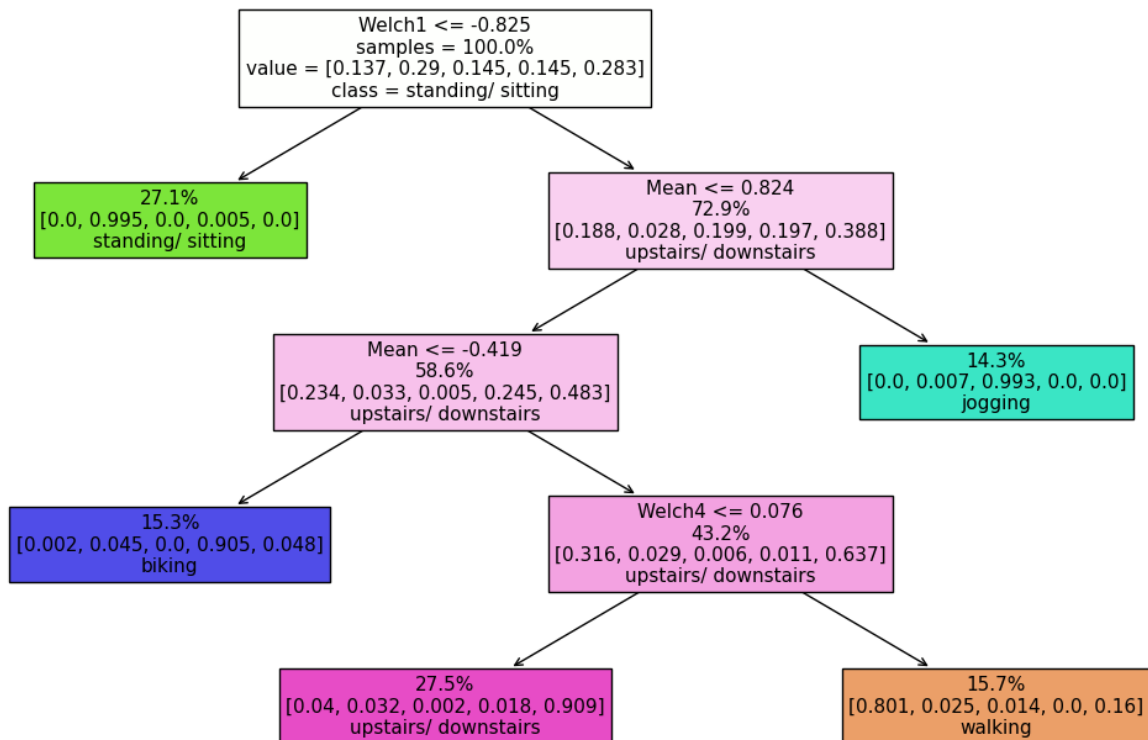
Η *ακρίβεια* σε αυτή την περίπτωση είναι εμφανώς *υψηλότερη*. Το γεγονός ότι ομαδοποιήσαμε τις δραστηριότητες αυτές, τις οποίες ο αλγόριθμος δυσκολευόταν κατά κάποιο τρόπο να ταξινομήσει, είχε θετικό αποτέλεσμα στην ταξινόμηση.

→ Η ακρίβεια αυξήθηκε σχεδόν κατά 10%: από 0.79 ανέβηκε σε 0.89.

Το νέο Δέντρο Απόφασης:

με παράμετρο `csp_alpha` διαφοροποιημένη² [αυξημένη]:

νέα τιμή **0.1**, προηγούμενη τιμή: 0.014

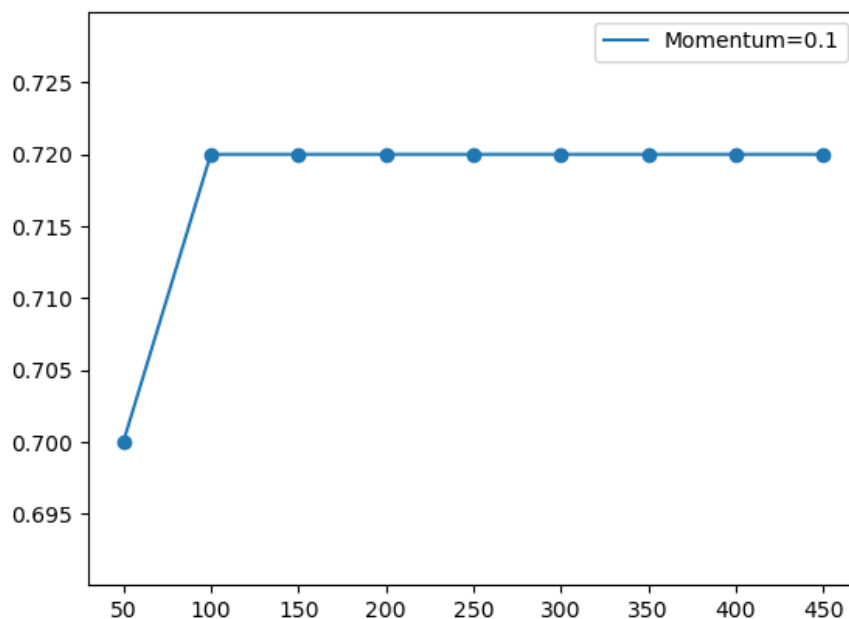


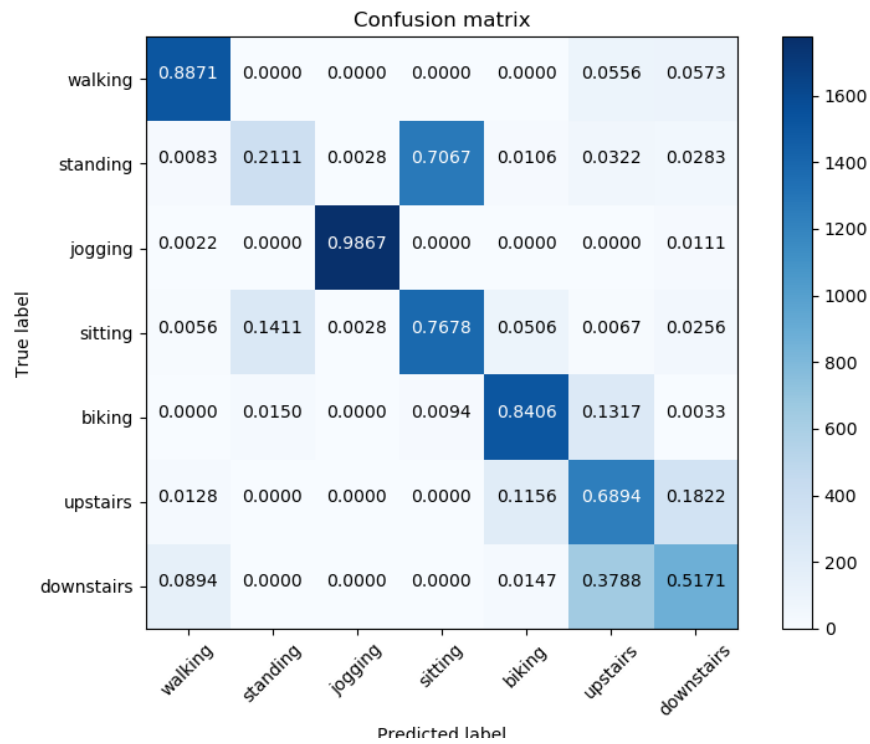
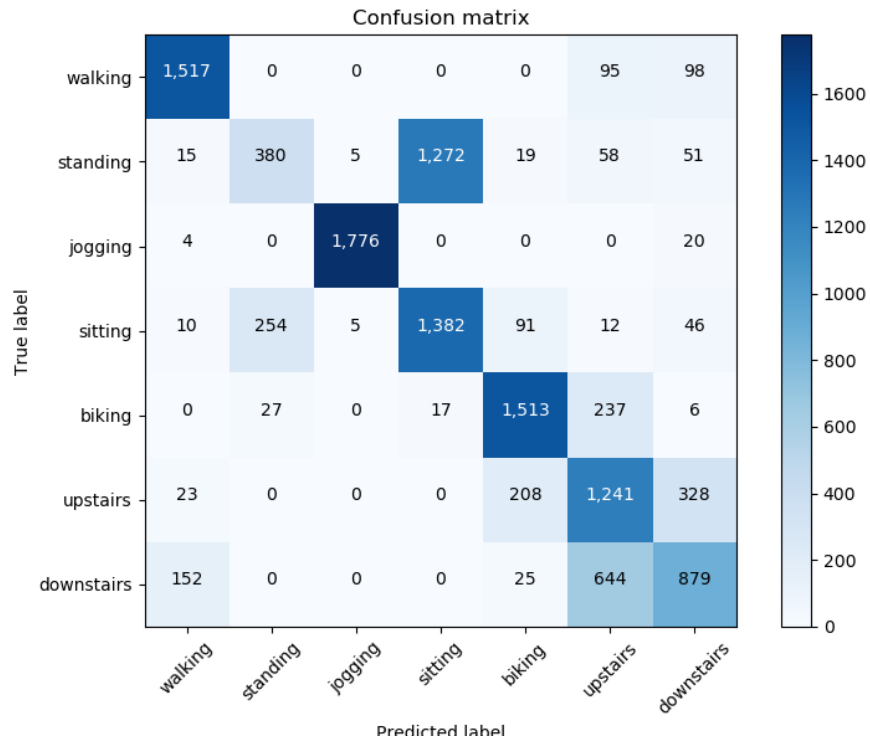
² Κάποιοι κόμβοι διαχωρίζονταν σε φύλλα που είχαν την ίδια κλάση με τον αρχικό κόμβο, οπότε καταλάβαμε πως χρειαζόταν ρύθμιση η παράμετρος `csp_alpha` που αφορά το κλάδεμα του Δέντρου Απόφασης.

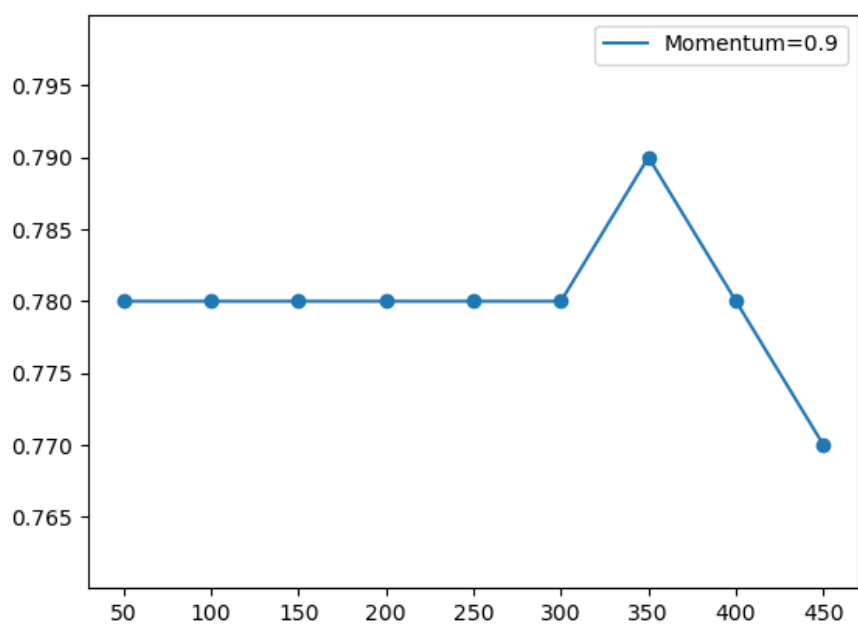
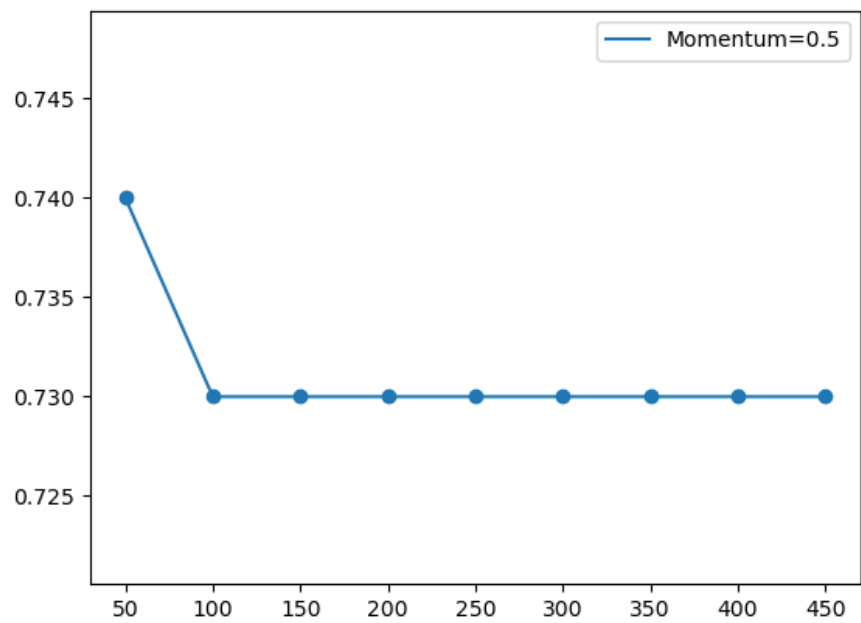
3. Νευρωνικά δίκτυα MLP με 1 – 2 κρυφά στρώματα και χρήση ορμής (momentum)

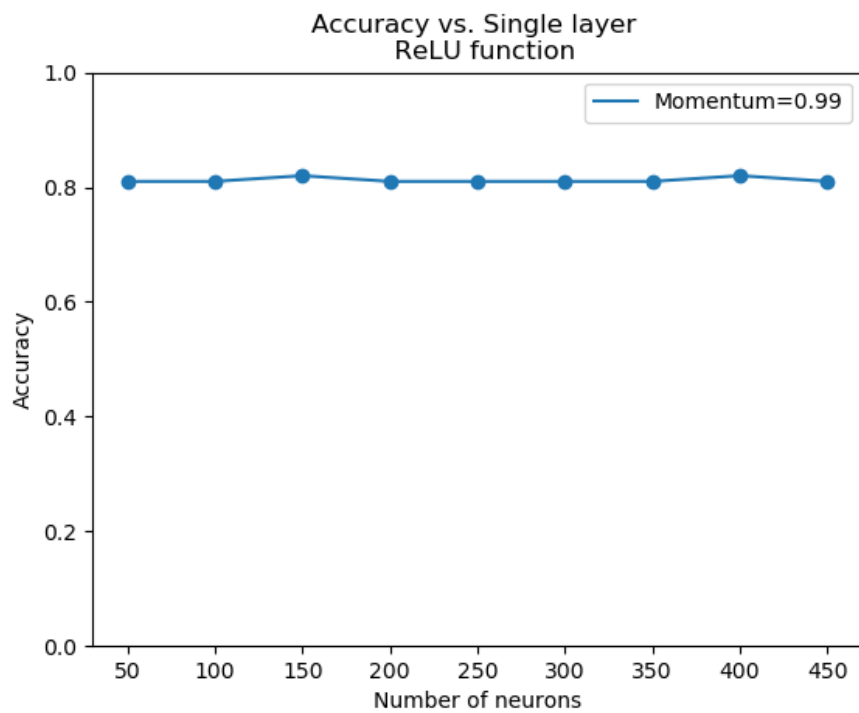
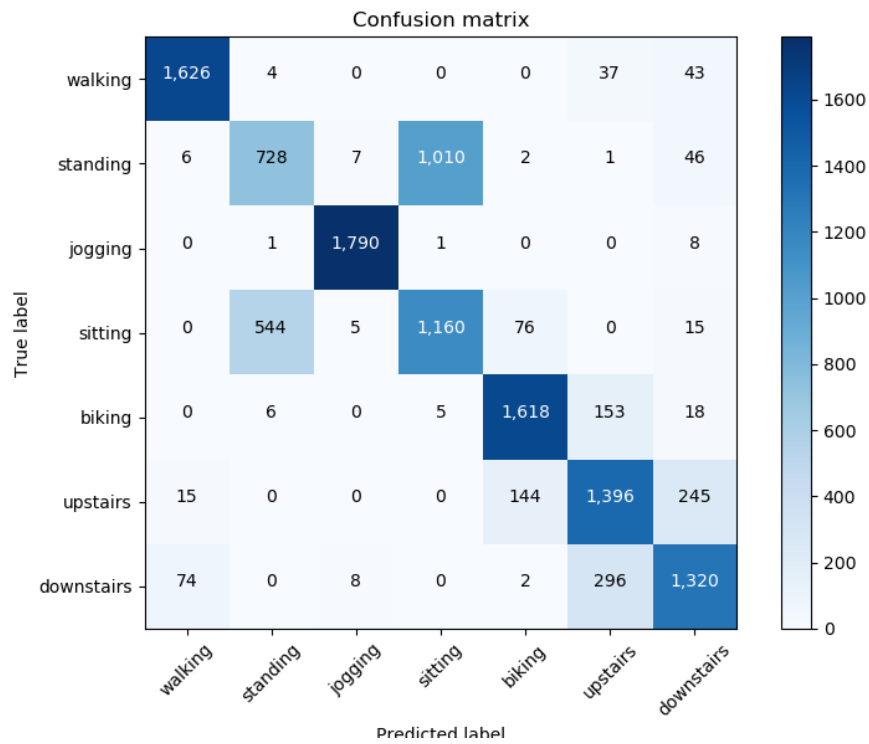
Χρησιμοποιήθηκαν δίκτυα MLP ενός και δύο κρυφών στρωμάτων, μεγέθους από 50 έως 450 νευρώνες και (50,50), (100,100), (150,150), (200,200) και (500,500) νευρώνων αντίστοιχα. Λόγω της χρήσης ορμής διερευνήθηκαν οι τιμές ορμής (momentum) 0.1, 0.5, 0.9, 0.99 (με stochastic gradient descent ως μέθοδο βελτιστοποίησης). Ως συναρτήσεις ενεργοποίησης χρησιμοποιήθηκαν οι: ReLU και logistic. Σε κάθε περίπτωση, όπως φαίνεται από τα παρακάτω διαγράμματα, η ακρίβεια πρόβλεψης φτάνει μέχρι στο 0.82, λόγω των 200 epochs εκπαίδευσης, ενώ αλλάζοντας τις τιμές του momentum, η logistic αλλάζει πιο “απότομα” τιμές ακρίβειας σε σχέση με τη ReLU. Τέλος, οι δραστηριότητες που συγχέει περισσότερο ο αλγόριθμος είναι μεταξύ τους οι: sitting και standing και upstairs και downstairs.

MLP με ένα κρυφό στρώμα και συνάρτηση ενεργοποίησης ReLU

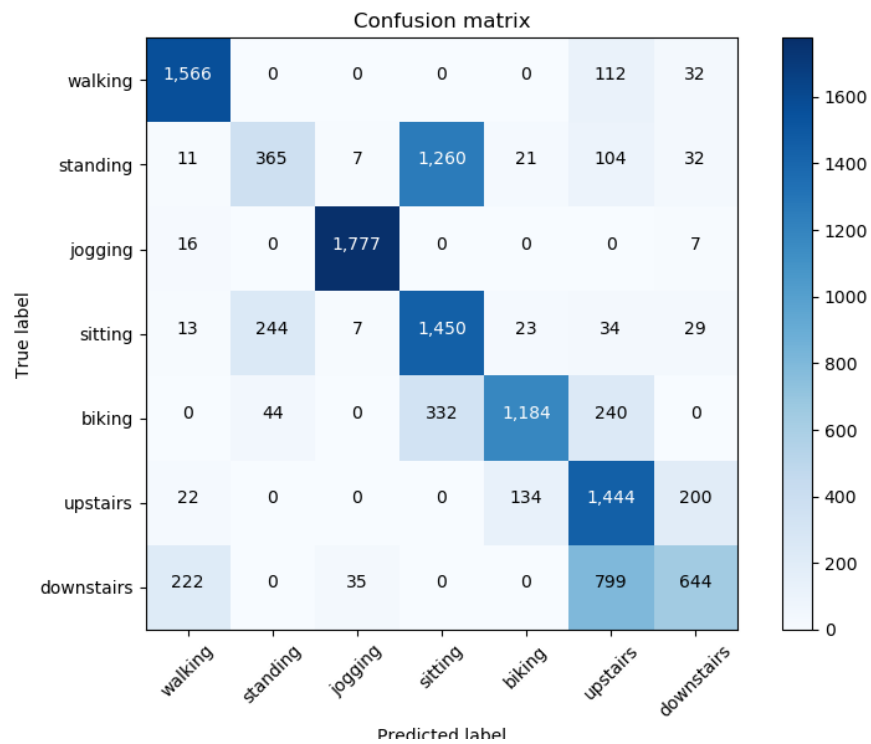
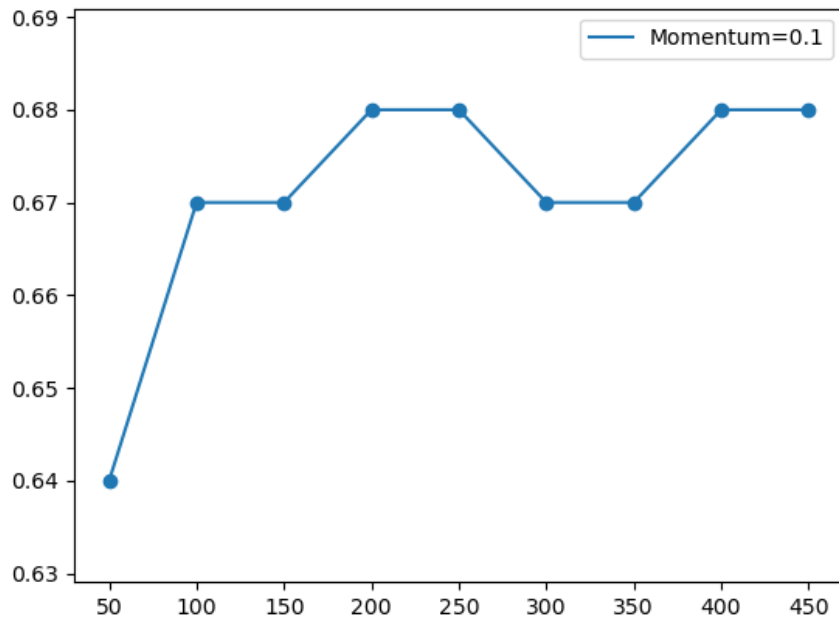


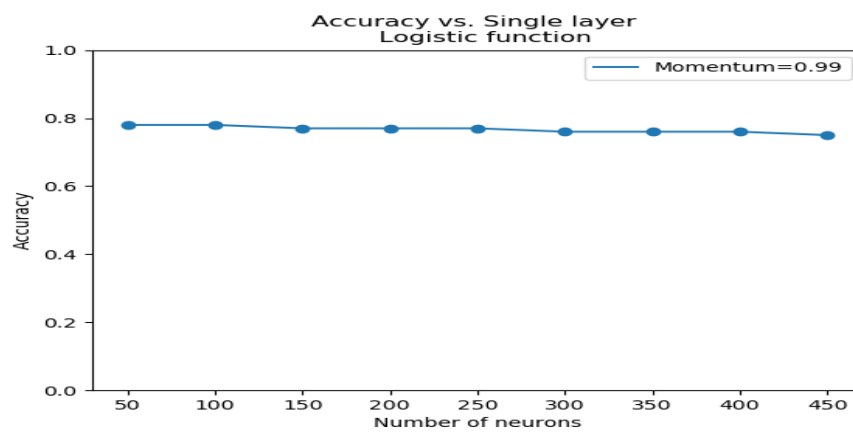
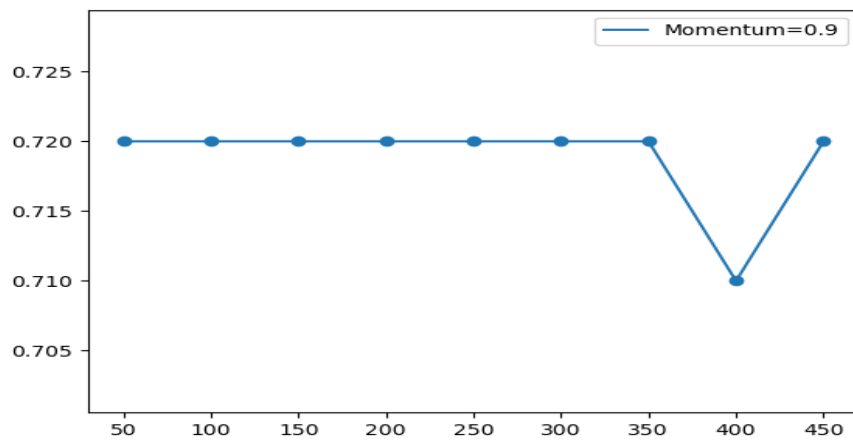
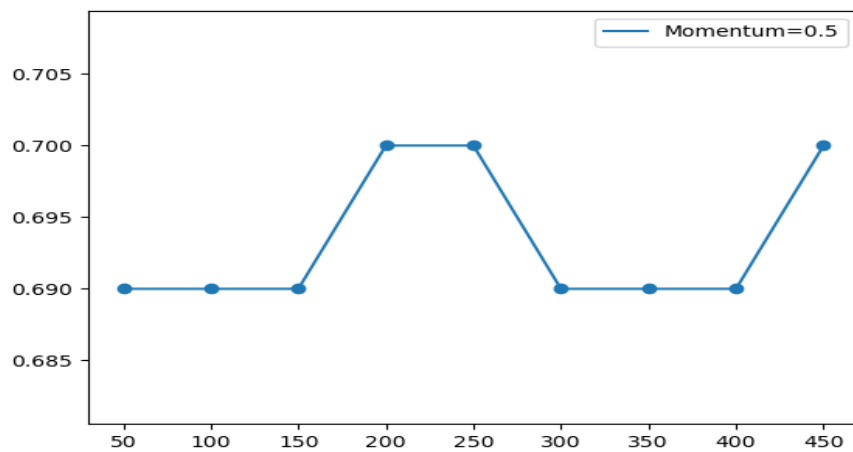


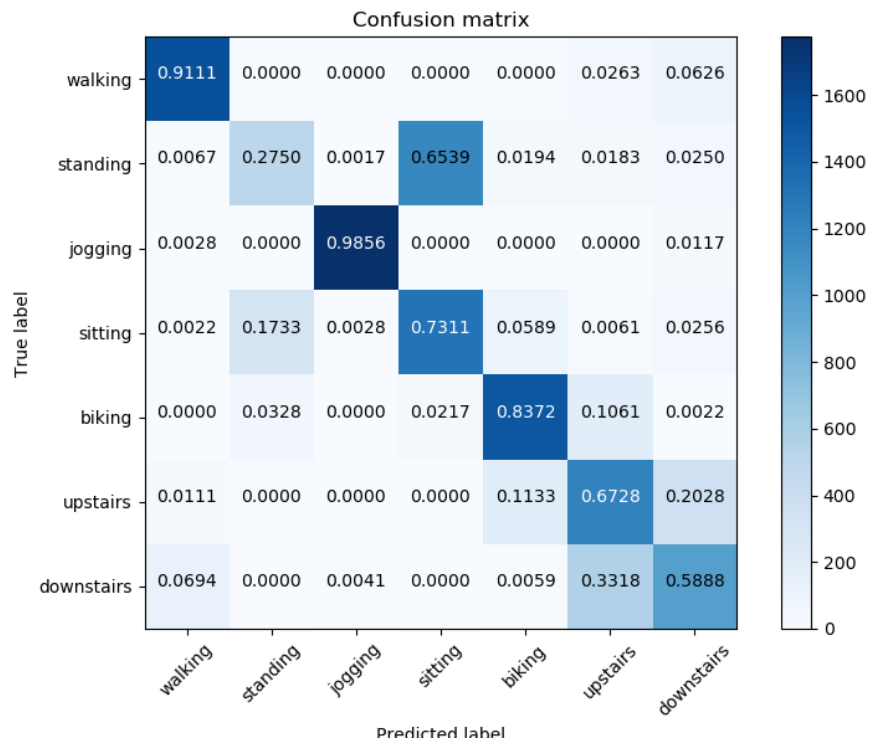




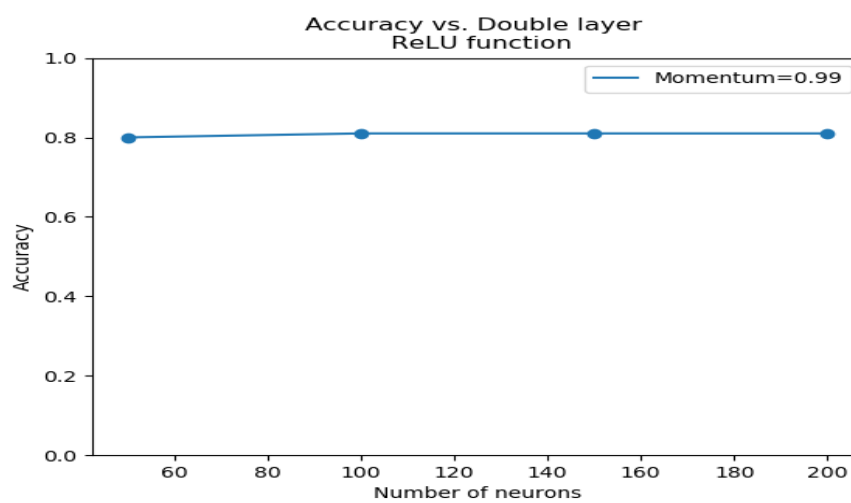
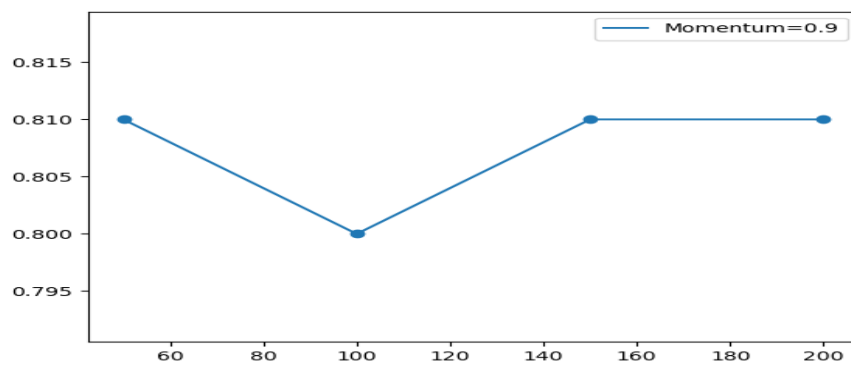
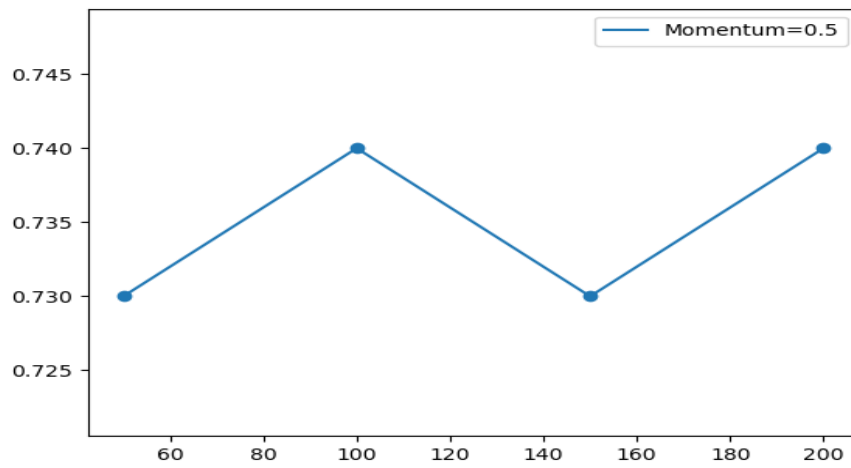
MLP με ένα στρώμα και συνάρτηση ενεργοποίησης logistic

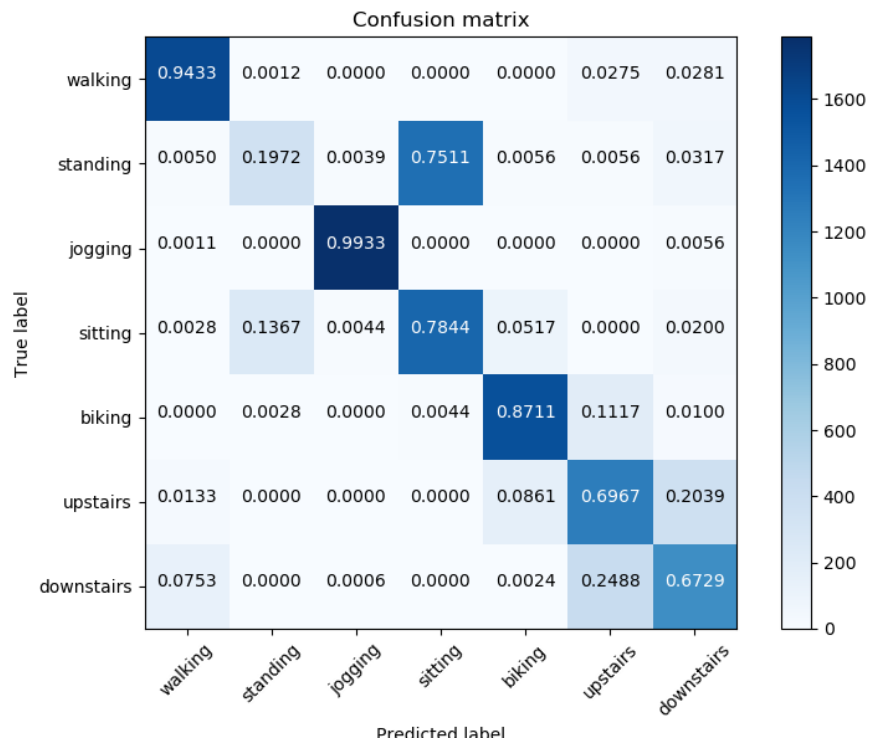
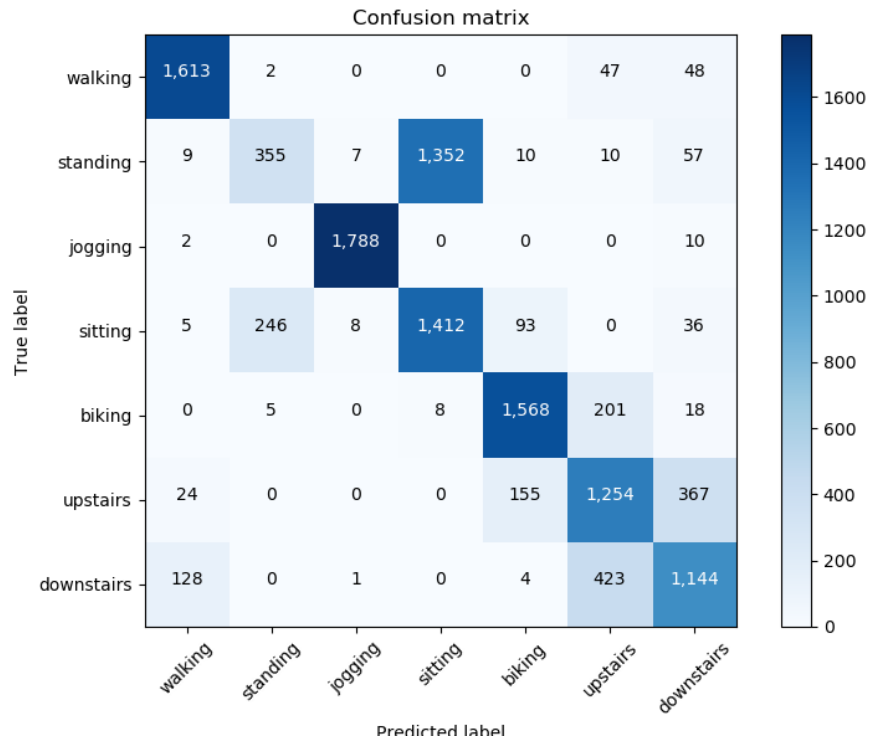




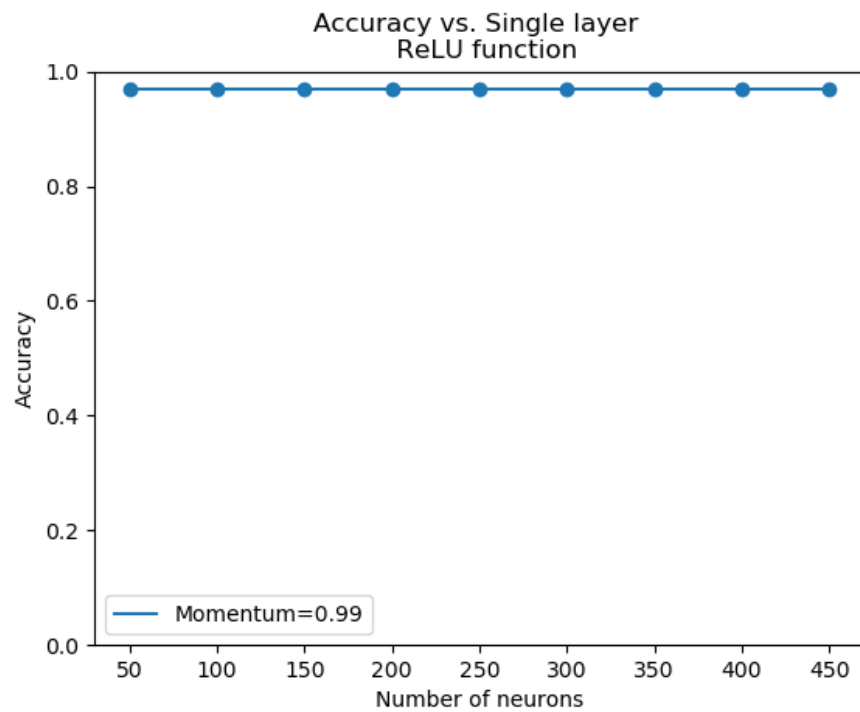
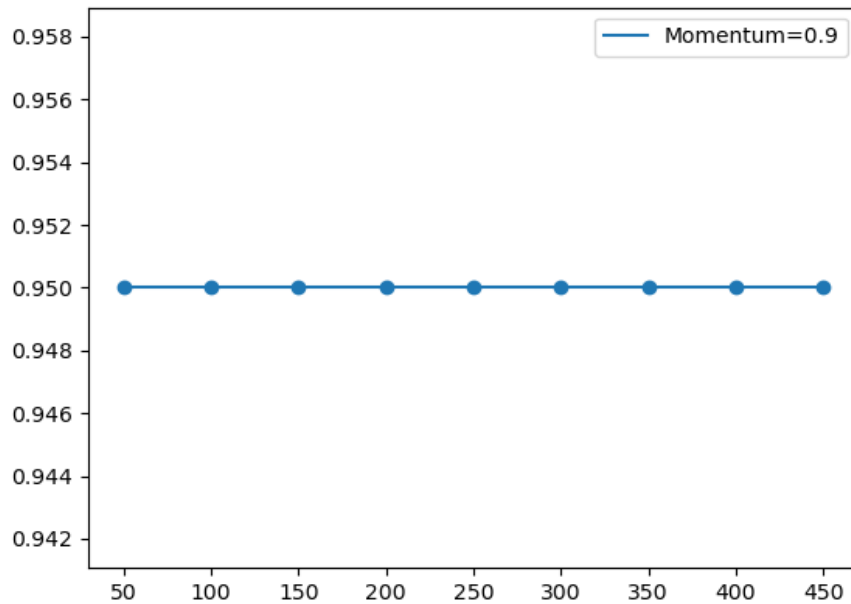


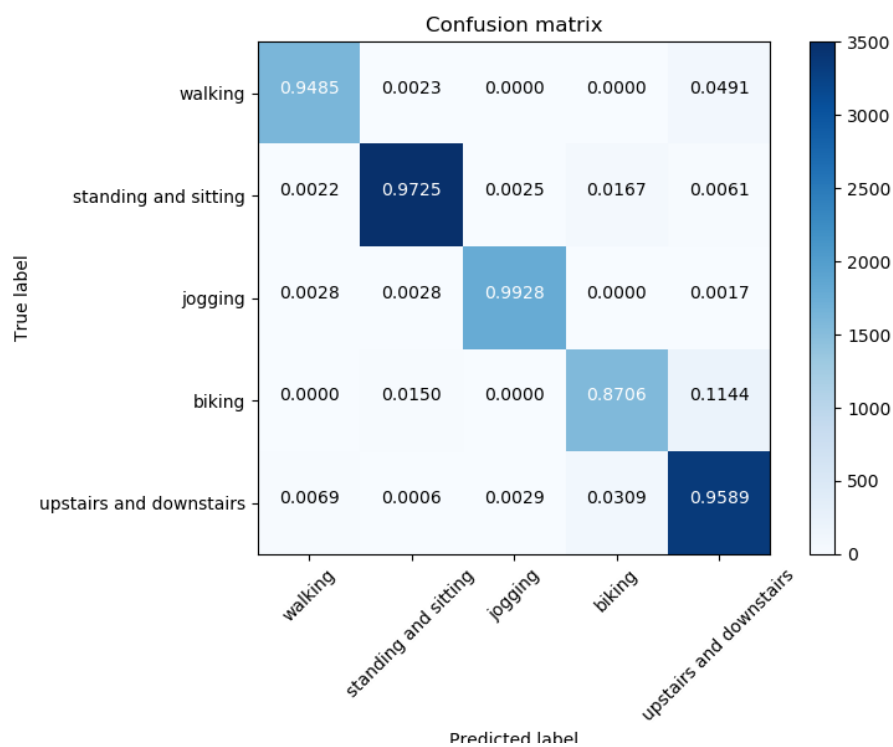
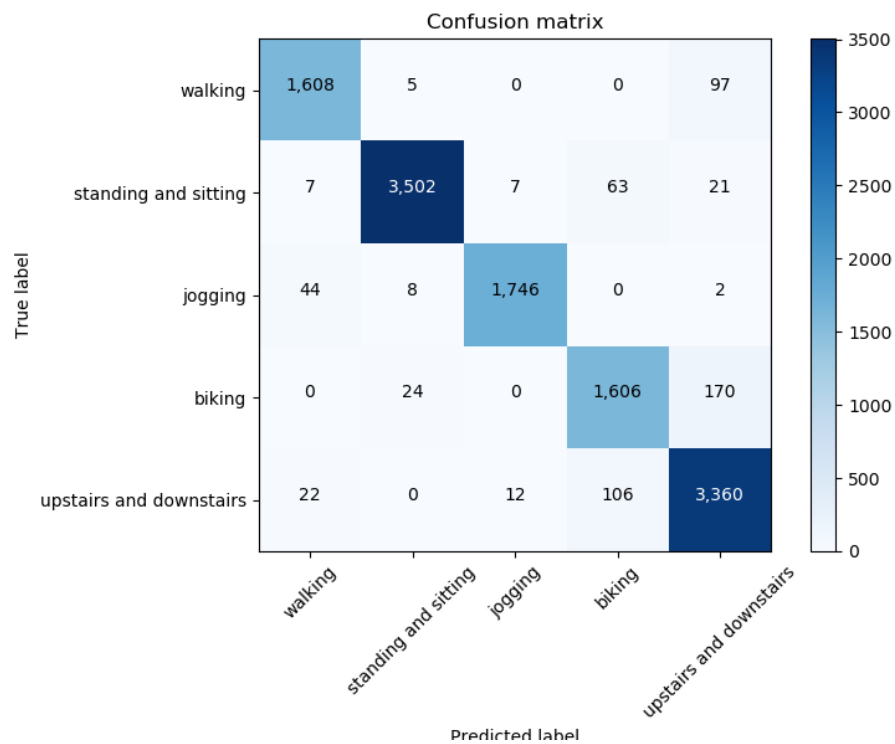
MLP με δύο κρυφά στρώματα και συνάρτηση ενεργοποίησης ReLU

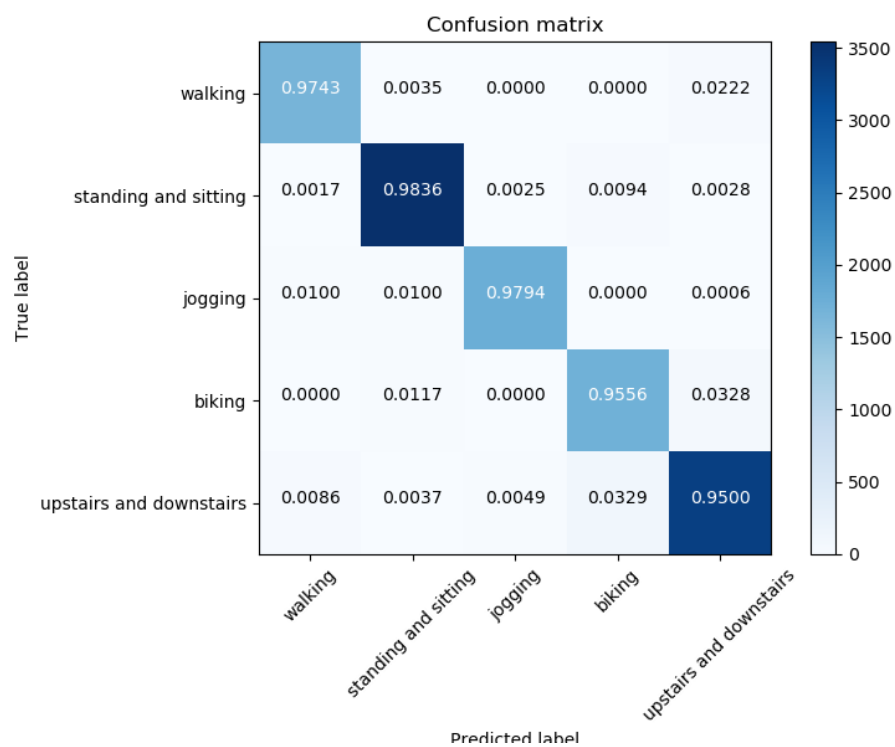
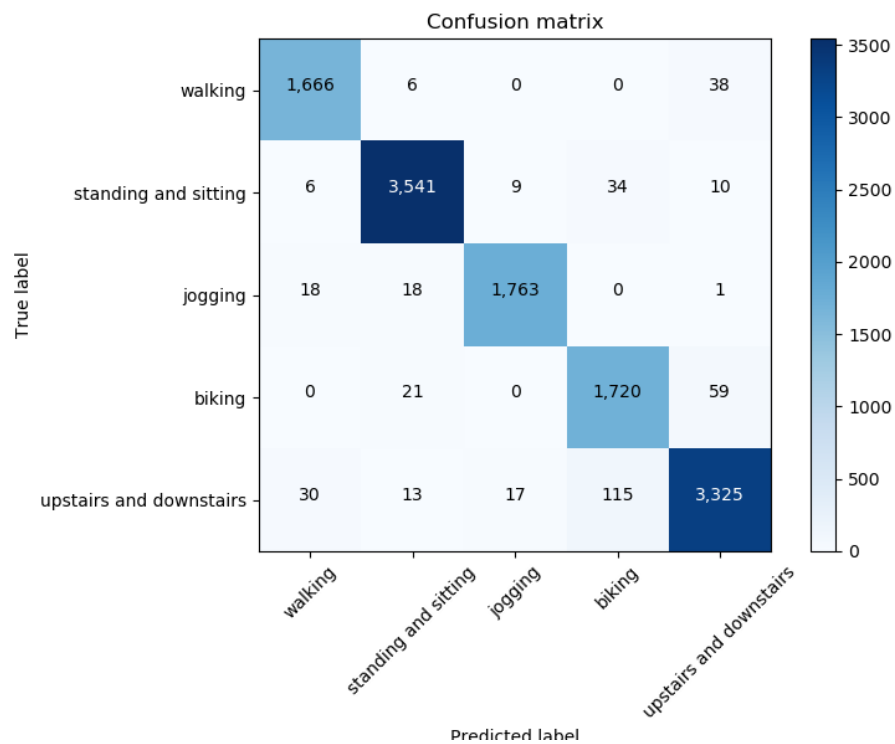




MLP μετά την ομαδοποίηση των δραστηριοτήτων που συγγέει το αρχικό. Χρήση ενός κρυφού στρώματος και συνάρτησης ReLU







4. Σύγκριση μοντέλων Ταξινόμησης

Κάποια χαρακτηριστικά τα οποία μπορούμε να συγκρίνουμε μεταξύ των τριών μοντέλων που εφαρμόσαμε (Μηχανές διανυσμάτων υποστήριξης, Δέντρα Αποφάσεων, Multilayer Perceptron) είναι τα εξής:

- Σύγκριση δραστηριοτήτων
- Ακρίβεια ταξινόμησης πριν και μετά την ομαδοποίηση των δραστηριοτήτων που συγχέονται

χαρακτηριστικό/ μέτρο σύγκρισης	SVM	Decision Tree	MLP (momentum)
σύγκριση δραστηριοτήτων	standing = sitting, upstairs = downstairs	standing = sitting, upstairs = downstairs	standing = sitting, upstairs = downstairs
αρχική ακρίβεια	0.81	0.79	0.82
ακρίβεια μετά την ομαδοποίηση	0.97	0.89	0.99

