# καταγραφή του περιβάλλοντος υλοποίησης

Περιβάλλον Υλοποίησης:

|  |  |
| --- | --- |
| CPU model | Intel(R) Core(TM) i7-1065G7 |
| CPU clock speed | 1.30GHz |
| Physical CPU cores | 4 |
| Logical CPU cores | 8 |
| RAM | 8 |
| Secondary Storage Type | SSD |

Γενικές βιβλιοθήκες:

Pandas, numpy, os, matplotlib, seaborn, yellowbrick

Βιβλιοθήκη για classification και clustering:

Sklearn

Οι παραπάνω βιβλιοθήκες εγκαθίστανται με την εντολή **pip install library\_name**

# Ερώτημα 1

Αφού εισάγουμε στο πρόγραμμα τα αρχεία .csv με τα δεδομένα δημιουργώ ένα dataframe που περιέχει τις μετρήσεις από όλους τους συμμετέχοντες. Επίσης για διαχωρισμό της κάθε εγγραφής που υπάρχει στο dataframe εισάγω μια επιπλέον στήλη “userID” ώστε να γνωρίζω σε ποιόν ανήκει η μέτρηση.

Το dataframe που δημιουργείται έχει τις στήλες:

timestamp, back\_x, back\_y, back\_z, thigh\_x, thigh\_y, thigh\_z, label

Για αυτό το dataframe παίρνω τα αποτελέσματα:

Πληροφορίες dataframe: df.info()

A screenshot of a computer

Description automatically generated

Έλεγχος για μη ύπαρξη τιμών στα δεδομένα:

A black screen with white text

Description automatically generated

Στατιστικά δεδομένα: df.describe()

A screen shot of numbers

Description automatically generated

Ιστόγραμμα κάθε στήλης:

A graph of a graph

Description automatically generated with medium confidence

Σχολιασμός: οι γραφικές παραστάσεις των στηλών back\_x, back\_y, back\_z, thigh\_x, thigh\_y, thigh\_z φαίνεται να ακολουθούν κανονική κατανομή. Βρίσκονται γύρω από το 0 και παρουσιάζουν κορύφωση στο 0. Βέβαια παρατηρούνται διαφοροποιήσεις στις γραφικές παραστάσεις των back\_x και thigh\_x που δεν έχουν κέντρο στο 0 και στο διάγραμμα thigh\_z κυρίως όπου υπάρχουν τιμές που αποκλίνουν από τις υπόλοιπες (outliers).

Ιστόγραμμα ετικετών (πιο αναλυτικό):

A graph of a bar graph

Description automatically generated

Διάγραμμα Συσχέτισης τιμών dataframe:

A screenshot of a computer screen

Description automatically generated

Σχολιασμός: συσχέτιση παρουσιάζεται μεταξύ των τιμών thigh\_x, back\_x και thigh\_x, thigh\_z. Οι τιμές αυτών των στηλών μοιάζουν μεταξύ τους. Για τις υπόλοιπες τιμές η συσχετίσεις είναι χαμηλές.

Διάγραμμα συσχέτισης στατιστικών δεδομένων:

A screenshot of a graph

Description automatically generated

Σχολιασμός: Από το διάγραμμα συμπεραίνεται ότι υπάρχουν μεγάλες συσχετίσεις μεταξύ των στατιστικών μεγεθών, εκτός από τις ελάχιστες τιμές που δεν έχουν συσχέτιση. Επομένως στα δεδομένα οι ελάχιστες τιμές έχουν μεγάλη απόσταση από τις υπόλοιπες τιμές.

Ερώτημα 2

Για το classifiacation χρησιμοποιήθηκαν Bayesian Networks, Neural Networks, Random Forests.

A black background with multicolored text

Description automatically generated

Αρχικά παράγω από το dataframe το σύνολο δεδομένων εισόδου και το σύνολο δεδομένων με τις ετικέτες. Για το σύνολο δεδομένων εισόδου διαγράφονται οι στήλες του dataframe ('label', 'timestamp', 'userID') και για τις ετικέτες κρατάω μόνο τη στήλη 'label'.

Έπειτα με τη συνάρτηση select\_classifier(option=None) επιλέγεται ένας από τους 3 ταξινομητές και με τη συνάρτηση run\_classification(X,Y,classifier = 1) γίνεται ο διαχωρισμός των δεδομένων σε σύνολο training και testing με αναλογία 0.3 και μετά πραγματοποιείται το training και το testing. Αφού εκτελεστούν και οι 3 ταξινομητές παράγεται διάγραμμα με τα αποτελέσματα ακρίβειας σε εκπαίδευση και δοκιμή.

Τέλος με τη συνάρτηση  classification(df, option=None) θα εκτελεστούν και οι 3 ταξινομητές, αλλά μπορεί με την παράμετρο option να εκτελεστεί μόνο 1 από τους 3.

Ερώτημα 3

Για το clustering προηγείται η διαδικασία επεξεργασίας των δεδομένων. Από το dataframe παράγω μια λίστα με 22 dataframes. Το καθένα αντιστοιχεί σε ένα συμμετέχοντα του πειράματος. Οι στήλες αυτών των dataframes είναι οι τιμές mean, std, min, 25%, 50%, 75%, max για κάθε σένσορα που χρησιμοποιείται, δηλαδή 7\*6=42 στήλες. Στο τέλος όταν δημιουργηθούν και τα 22 dataframes τα ενώνω σε ένα και κάθε γραμμή του νέο dataframe αντιστοιχεί σε έναν από τους 22 συμμετέχοντες.

Μετά τη φάση της δημιουργίας του νέου dataframe το χρησιμοποιώ για τη διαδικασία του clustering. Η συσταδοποίηση γίνεται με Kmeans, MeanShift και DBscan. Με τη συνάρτηση select\_clusterer(option) επιλέγεται ένα από τους 3 clusterers και εκτελείται η συσταδοποίηση. Ο κάθε αλγόριθμος συσταδοποίησης που χρησιμοποιείται εκτελείται με κλήση στην κατάλληλη συνάρτηση (kmeans, meanshift, dbscan).

Στη συνάρτηση kmeans() εκτελείται ο αλγόριθμος και μετά καλείται η συνάρτηση KElbowVisualizer ώστε να βρεθεί η βέλτιστη τιμή του αριθμού των συστάδων. Αφού βρεθεί εκτελείται ο αλγόριθμος με τη βέλτιστη τιμή του k.

A graph with a line graph

Description automatically generated with medium confidence

Στη συνάρτηση meanshift() εκτελείται ο αλγόριθμος. Ο αριθμός των συστάδων δεν δίνεται ως όρισμα διότι η συνάρτηση MeanShift() βρίσκει από μόνη της το βέλτιστο αριθμό συστάδων.

Στη συνάρτηση dbscan() πριν εκτελεστεί ο αλγόριθμος DBSCAN χρησιμοποιείται η μέθοδος εύρεσης των πλησιέστερων γειτόνων ώστε να υπολογιστεί η μέγιστη ακτίνα που θα έχει ένα cluster. Από το διάγραμμα εντοπίζεται το σημείο elbow και αυτή θα είναι η παράμετρος eps που θα χρησιμοποιηθεί στον dbscan.

A graph with a line going up

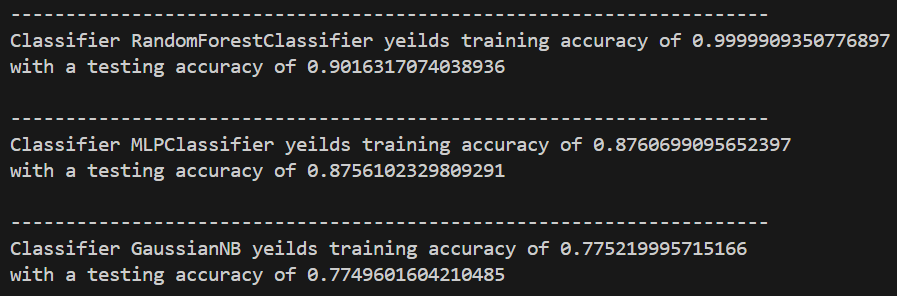
Description automatically generated

Καλώντας τη συνάρτηση  clustering(X, option=None) θα εκτελεστούν και οι 3 αλγόριθμοι συσταδοποίησης, αλλά μπορεί με την παράμετρο option να εκτελεστεί μόνο 1 από τους 3.

# o Σχολιασμό των τελικών αποτελεσμάτων

Για το πρώτο ερώτημα ο σχολιασμός έχει γίνει στην προηγούμενη ενότητα.

Ερώτημα 2



A graph of different colored bars

Description automatically generated with medium confidence

Ερώτημα 3

A computer screen shot of numbers

Description automatically generated