Physical Activity Prediction Using Machine Learning on Wearable and Mobile Devices Data (WeMoD)

Πρόβλεψη Φυσικής Δραστηριότητας με Χρήση Μηχανικής Μάθησης σε Δεδομένα Φορητών και Φορετών Συσκευών (WeMoD)

by

Dimitrios Vasdekis

A thesis submitted in partial fulfillment for the MSc Degree in "Data & Web Science"

in the
Faculty of Sciences
School of Informatics

Supervising Professor: Dr. Athena Vakali

16 Apr 2021
Declaration of Authorship

I, Dimitrios Vasdekis, declare that this thesis titled, "Physical Activity Prediction Using Machine Learning on Wearable and Mobile Devices Data (WeMoD)" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a MSc degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____________________________________________

Date: ____________________________________________________________________
When it comes to health and well-being, regular exercise is about as close to a magic potion as you can get.

Tich Nhat Hanh
A hot topic which nowadays has aroused the interest of many research projects, is the effect of a person’s physical activity (PA) levels on his personal health and his overall quality of life. It has been proven in many works, as we discuss later, that PA increment can bring multiple benefits not only individually but also collectively, as it is an issue with wider social and economic implications. A lifestyle with an emphasis on physical exercise leads to reduced pollutants, a cleaner atmosphere and generally better living conditions. Furthermore, healthy people tend to be more efficient in their work and more productive, which in turn creates multiple economic benefits for the society as a whole. Despite the significant benefits described above, relevant researches indicate that only a small percentage of the population meets the recommended guidelines regarding the levels of physical exercise in their daily lives.

Since the importance of PA increment has been established, an issue that arises is the definition of an appropriate metric that expresses and evaluates physical exercise satisfactorily. The number of steps that are taken by one person during a day, is such a metric, as it achieves the quantification of an aerobic activity of moderate intensity, walking. Monitoring walking levels as a PA is an approach that has been widely adopted due to its simplicity, objectivity and universality.

Once the metric on which a person’s level of PA can be measured has been determined, so that appropriate actions can be taken to increase it if necessary, the next question is how to ensure a clear and objective picture of the number of steps performed during a day by an individual. Modern technology and engineering come to answer this problem. The technological advance of our time, have made possible the creation of portable and wearable devices which through integrated sensors can be used as PA trackers. Such devices, mobile phones, smart watches, belts, etc. have been widely used in the relevant literature to extract daily step-count measurements.
A successful approach to improving PA levels, has been found to be the setting of daily step-count goals, which are set through portable and wearable devices. The activity tracking devices, set a number of steps as a daily goal based on an algorithmic functionality implemented in their software. Next, they monitor the steps that a user takes during the day and inform them of the progress towards achieving their goal, often intervening to provide encouragement or adjusting the original goal if necessary. One of the most important aspects in such approaches is the algorithm used in the goal setting process as it has been found that appropriate goal setting methodology leads in fact to the increment of PA levels.

At the core of this research, are step-count data collected through portable and wearable devices and goal setting methodologies like those described above. The purpose of this work considers two pillars. On the one hand, we seek to create a dataset that can be used by machine learning (ML) algorithms, in order to predict a user’s future daily step-count. On the other hand, this research aims to develop such a prediction model that satisfies the above description. Our goal for this model is to be a stand-alone, end-to-end solution, appropriate for further integration into applications aimed at improving users’ PA levels, either through the establishment of appropriate step-count goals, or through positive intervention techniques.

More specifically, the contribution of this work can be analyzed in 4 key points. We design and implement an integration framework for combining activity data from activity tracking devices created by different manufacturers. An important tool in order to enhance and diversify the pool of available data. In addition, we experiment with the inclusion of a novel combination of activity-related, personal and contextual features in the created dataset and the evaluation of their impact on the predictive ability of the ML models. Moreover, in our research we utilize “in-the-wild” data. Deferentially to the majority of relevant researches, our data reflect the participant’s daily routines in an undoubtedly accurate manner. Finally, we develop a ML model which can efficiently forecast future daily step-count activity.

The results of our experimental process indicate that our contributions have been implemented effectively. The dataset we created has been used efficiently in the model development process and the novel feature sets tested, have been found to improve the predictive ability of the ML models. The different approaches adopted in our methodology, such as outlier handling, dimensionality reduction, algorithm optimization and more, have led to the development of an efficient ML prediction model which has achieved to predict daily step-counts of different users with an average deviation of 1930 steps, by using 5 days of previous activity as input.
Περίληψη

Σχολή Θετικών Επιστημών
Τμήμα Πληροφορικής

Μεταπτυχιακό Δίπλωμα Ειδίκευσης

Πρόβλεψη Φυσικής Δραστηριότητας με Χρήση Μηχανικής Μάθησης σε Δεδομένα Φορητών και Φορτιστών Συσκευών (WeMoD)

του Δημητρίου Βασδέκη

Ένα πολύ σοβαρό ζήτημα το οποίο στις μέρες μας έχει προκαλέσει το ενδιαφέρον πολλών σχετικών έρευνων, είναι η επίδραση των επιπέδων φυσικής δραστηριότητας ενός ατόμου στην προσωπική του υγεία αλλά και σε άλλους παράγοντες της καθημερινότητάς του. Έχει πλέον τεκμηριωθεί ότι η αύξηση της φυσικής δραστηριότητας, μπορεί να επιφέρει πολλά οφέλη, όχι μόνο σε ατομικό επίπεδο αλλά και σε συλλογικό, καθώς αποτελεί ένα ζήτημα με ευρύτερες και οικονομικές προοπτικές. Οι άνθρωποι οι οποίοι ασκούνται επαρκώς είναι γενικά υγιείς και επιμορίζονται λιγότερο στο σύστημα υγείας τους. Επιπλέον, ένα γενικότερο ζήτημα σε σχέση με την φυσική άσκηση, οδηγεί σε μειωμένους και καταράτηση γυμνόφρακτη και εν τέλει σε συνθήκες ποιοτικής διαβίωσης. Επίσης, οι άτομοι οι οποίοι ασκούνται επαρκώς είναι γενικά πιο αποδοτικοί στην εργασία τους και περισσότερο παραγωγικοί, παρά τους επίπεδος ποσοστό του πληθυσμού ανταποκρίνεται και ασκείται στα επίπεδα που συνιστώνται από τους ειδικούς. Αυτό είναι ένα ζήτημα το οποίο χρήζει αντιμετώπισης και κατά συνέπεια αποτελεί έναν στόχο της σύγχρονης ερευνής.

Ένα σημαντικό ζήτημα στη σχέση με την θέσπιση ενός στόχου αύξησης της φυσικής δραστηριότητας, είτε σε ατομικό είτε σε συλλογικό επίπεδο, είναι ο ορισμός ενός κατάλληλου μέτρου που να την ευρύξει και να την αξιολογεί υιοθετηθεί λόγω της απλότητας, της αντικειμενικότητας και της καθοδοκιμάτωσης της.
Εφόσον προσδιορίστηκε η μετρική βάση της οποίας μπορεί να μετρηθεί το επίπεδο φυσικής δραστηριότητας ενός ατόμου, ώστε να ληφθούν κατάλληλα μέτρα για την αύξησή του αν αυτό χρειάζεται αναγκαίο, το επόμενο ζήτημα είναι ο τρόπος με τον οποίο μπορεί να εξασφαλιστεί μια σαφής και αντιμετωπική ευκολία για τον αριθμό των βημάτων που εκτελούνται κατά τη διάρκεια μία ημέρας. Σε αυτό το πρόβλημα έρχεται να δώσει απάντηση η σύγχρονη τεχνολογία και μηχανική. Τα τεχνολογικά άμαξα της εποχής μας, με την δραματική αύξηση της υπολογιστικής υποτατικής των διαφόρων συσκευών, σε συνδυασμό με την εξίσου δραματική μείωση του μεγέθους των κάθε λογικής εξαρτημάτων, έχουν καταστήσει εφικτή την δημιουργία φορητών και φορητών συσκευών οι οποίες μέσω ενσωματωμένων ασθενήρων μπορούν να χρησιμοποιούνται ευρέως, ως καταγραφείς φυσικής δραστηριότητας. Τέτοιους είδους συσκευές, κινητά τηλέφωνα, έχουν ρόλο για την επίτευξη του στόχου του σετάντησης ενός χρήστη σε μια απλή και αναγνώριστη για την άντληση δεδομένων ημερήσιου αριθμού βημάτων.

Μια επιτυχημένη προσέγγιση όσον αφορά τη βελτίωση των επιπέδων φυσικής δραστηριότητας, έχει βρεθεί ότι είναι ο καθορισμός καθημερινών στόχων βημάτων, οι οποίοι τίθενται μέσω φορητών και φορητών συσκευών. Οι συσκευές παρακολούθησης δραστηριότητας, θέτουν έναν συγκεκριμένο αριθμό βημάτων ως καθημερινό στόχο με βάση κάποιον αλγόριθμο που υλοποιείται στο λογισμικό τους. Στη συνέχεια, παρακολουθούν τα βήματα που πραγματοποιεί ένας χρήστης κατά τη διάρκεια της ημέρας και τον ενημερώνουν για την πρόοδο προς την επίτευξη του στόχου του, παρεμβαίνοντας συχνά προειμένουν να παρέχουν ενθάρρυνση να προσαρμόσουν τον αρχικό στόχο εφόσον κάτι τέτοιο χρειάζεται απαραίτητο. Μία από τις πιο σημαντικές πτυχές σε τέτοιες προσεγγίσεις είναι ο αλγόριθμος που χρησιμοποιείται στη διαδικασία καθορισμού στόχων, καθώς έχει βρεθεί ότι αν αυλομοίρασε η κατάλληλη μεθοδολογία ο χρήστης όντως οδηγείται σε αύξηση των επιπέδων φυσικής δραστηριότητας.

Η παρούσα έρευνα, έχει στον πυρήνα της τέτοιας φύσεως δεδομένα φυσικής δραστηριότητας ημερήσιου αριθμού βημάτων, όπως αυτά έχουν συλλεχθεί μέσω φορητών και φορητών συσκευών καθώς και μεθοδολογίες δέσποινας ημερήσιων στόχων. Ο σκοπός της συγκεκριμένης εργασίας εδράζεται σε δύο πυλώνες. Αφ’ ενός, επιδιώκουμε την κατασκευή ενός συνόλου δεδομένων (dataset) το οποίο θα μπορεί να χρησιμοποιηθεί από αλγόριθμους μηχανικής μάθησης, με σκοπό την πρόβλεψη του ημερήσιου αριθμού βημάτων ενός χρήστη σε μεταγενέστερο χρόνο. Αφ’ ετέρου, η έρευνα αποσκοπεί στην ανάπτυξη και αξιολόγηση ενός μοντέλου με τις προδιαγραφές που περιγράφονται παραπάνω. Στόχος μας για αυτό το μοντέλο είναι να αποτελέσει μια ολοκληρωμένη λύση για την περαιτέρω ενσωμάτωσή του σε εφαρμογές που αποσκοπούν στην βελτίωση των επιπέδων φυσικής δραστηριότητας των χρηστών. Πιο συγκεκριμένα, η συνεισφορά αυτής της εργασίας εντοπίζεται σε 4 βασικά σημεία, τα οποία αναφέρουμε επιγραμματικά εδώ, αλλά αναλύουμε σε βάδας στα επόμενα κεφάλαια.
Αρχικά, ένα πρώτο μέρος της συνεισφοράς μας έγινεται στον σχεδιασμό και την ανάπτυξη ενός μοντέλου εργασίας που αποσκοπεί στην ομογενοποίηση δεδομένων που προέρχονται από διαφορετικές συσκευές καταγραφής δραστηριότητας (integration framework). Οι διάφορες συσκευές, αποθηκεύουν τα δεδομένα που συλλέγουν με αρκετά ανοικοδομητική τρόπο και κατά συνέπεια προκειμένου αυτά να αξιοποιηθούν σε ένα ενιαίο dataset, πρέπει πρώτα να υποστούν προετοιμασία. Το integration framework που αναπτύχθηκε, επιτυγχάνει να ομογενοποιεί δεδομένα των δύο πιο δημοφιλών κατασκευαστών φορέτων συσκευών καταγραφής δραστηριότητας, της Apple και της Xiaomi, συζητάται έστω τις δυνατότητες των ερευνητών σε αυτήν αλλά και σε μελλοντικές εργασίες όσον αφορά την επιλογή πηγών δεδομένων δραστηριότητας.

Στις περισσότερες σχετικές έρευνες, όπως θα αναλύσουμε στο κεφάλαιο που αφορά την ανασκόπηση της βιβλιογραφίας, προκειμένου να δημουργηθούν τα διάφορα μοντέλα, χρησιμοποιούνται αμιγώς δεδομένα δραστηριότητας, δηλαδή ημερήσιου αριθμού βημάτων. Στην παρούσα έρευνα αποφασίστηκε να αναλυθούν στο dataset και να αξιολογηθεί η επίδραση τους στην προβλεπτική ικανότητα των μοντέλων, εκτός από δεδομένα δραστηριότητας, δεδομένα που σχετίζονται με την προσωπικότητα ενός χρήστη αλλά και δεδομένα περιβάλλοντος που πιθανόν να επηρεάζουν την φυσική δραστηριότητα ενός ατόμου. Τα περιμετρικά αποτελέσματα δυνατούν αυτή μας την επιλογή δείχνοντας ότι ο συνυπολογισμός τέτοιων δεδομένων βελτιώνει την απόδοση των μοντέλων. Ένα επιπλέον τμήμα της συνεισφοράς μας αποτελείται από όλες αυτές τις ενέργειες ένταξης καινοτόμων δεδομένων στο dataset και της αξιολόγησης της επίδρασής τους στα μοντέλα μηχανικής μάθησης.

Επίσης, ένας συχνός τρόπος με τον οποίο αυτή η εργασία συνεισφέρει στο σχετικό ερευνητικό πεδίο, είναι η ένταξη στο dataset και η μετέπειτα χρήση σε μοντέλα μηχανικής μάθησης, δεδομένων που προέρχονται από την φυσική ροή της καθημερινότητας των συμμετέχοντων της έρευνας. Στις περισσότερες σχετικές εργασίες, προκειμένου να συλλέξουν τα απαραίτητα δεδομένα, ζητείται από τους συμμετέχοντες να εντάξουν στην καθημερινότητά τους μια συσκευή καταγραφής δραστηριότητας η οποία θα καταγράφει τα βήματα που εκτελούν για μία συγκεκριμένη χρονική περίοδο κάστων εβδομάδων ή μηνών. Η προσέγγιση αυτή, όπως θα δούμε στη συνέχεια, έχει βρεθεί ότι μπορεί να οδηγηθεί σε αλλοιωμένα αποτελέσματα σε σχέση με την πραγματική ευκόνα των επιπέδων άσκησης των συμμετέχοντων. Ο αρχικός ενδονοσιασμός τους, σε συνδυασμό με την επίγνωσή τους ότι συμμετέχουν σε μία εν εξέλιξις έρευνα μπορεί σε κάποιες περιπτώσεις να αποτελέσει κίνητρο για συνεχή δραστηριότητα σε σχέση με τα πραγματικά επίπεδα. Στην παρούσα εργασία τα δεδομένα συλλέχθηκαν από άτομα που χρησιμοποιούν συσκευές παρακολούθησης δραστηριότητας στην καθημερινότητα τους και τα δεδομένα αυτά έχουν καταγραφεί σε φυσικές συνθήκες. Το γεγονός αυτό εγγυάται την ακρίβεια των δεδομένων που χρησιμοποιήθηκαν, ως προς την αποτύπωση της πραγματικότητας.
Τέλος, ένα ακόμη μέρος της συνεισφοράς μας, αποτελείται από την ανάπτυξη ενός μοντέλου μηχανικής μάθησης με υπανότητα πρόβλεψης του ημερήσιου αριθμού βημάτων ενός χρήστη, με βάση τα δεδομένα μιας ακολουθίας προηγούμενων ημερών. Για την επίτευξη αυτού του σκοπού, εκτελέστηκε ένα πλήθος πειραμάτων, σχετικών με προσεγγίσεις που αφορούν διάφορα κομμάτια μεθοδολογίας για την ανάπτυξη ενός τέτοιου μοντέλου, με γνώμονα να βρεθεί η καλύτερη δυνατή σύνθεση της οποία θα οδηγήσει στο βέλτιστο αποτέλεσμα. Στο πλαίσιο αυτό εξετάστηκε σε ένα αρχικό επίπεδο η συμπεριφορά ενός προσωποποιημένου μοντέλου, εκπαιδευμένου σε δεδομένα ενός μόνο χρήστη, έναντι της απόδοσης ενός γενικού μοντέλου εκπαιδευμένου στα δεδομένα του συνόλου των συμμετεχόντων της έρευνας.

Τα αποτελέσματα της πειραματικής διαδικασίας που ακολουθήθηκε, δείχνουν ότι η υλοποίηση της συνεισφοράς μας όπως παρουσιάστηκε παραπάνω, ήταν επιτυχής. Το dataset που δημιουργήθηκε, χρησιμοποιήθηκε αποτελεσματικά στη διαδικασία ανάπτυξης των μοντέλων πρόβλεψης και οι καινοτόμες κατηγορίες δεδομένων δραστηριότητας, προσωπικότητας και περιβάλλοντος που δοιμάστηκαν, βρέθηκαν να βελτιώσουν την υπανότητα πρόβλεψης των μοντέλων μηχανικής μάθησης. Οι διαφορετικές προσεγγίσεις που υιοθετήθηκαν στη μεθοδολογία μας, όπως ο χειρισμός των outlier, η μείωση της διαστασιμότητας του dataset, η βελτιστοποίηση των αλγορίθμων που χρησιμοποιήθηκαν κ.α., οδήγησαν στην ανάπτυξη ενός αποτελεσματικού μοντέλου μηχανικής μάθησης που επιτυγχάνει την πρόβλεψη ημερήσιου αριθμού βημάτων διαφορετικών χρηστών, με μέση απόκλιση 1930 βημάτων, χρησιμοποιώντας 5 ημέρες προηγούμενης δραστηριότητας ως δεδομένα εισόδου.
Acknowledgements

Before presenting this research, I feel obliged to thank the people who have made a decisive contribution to its implementation, with the support and guidance they have given me generously. First of all, I would like to thank all the faculty members of the M.Sc. "Data & Web Science" for the professionalism with which they treated me and the knowledge they offered me, without which it would have been impossible for me to complete this work. In particular, I would like to thank my supervising professor, Pr. Athena Vakali. Always approachable and positive, she enabled me to overcome whatever difficulties arose during my research. In addition, I would like to thank the Ph.D. candidate Sofia Yfantidou, without whose guidance and help, the implementation of this work would have been impossible. Finally, I would like to thank my family for their support and patience throughout my journey towards completing this research.
## Contents

Declaration of Authorship ................................................................. i

Abstract ............................................................................................ iii

Περίληψη ............................................................................................. v

Acknowledgements ............................................................................. ix

List of Figures ..................................................................................... xii

List of Tables ....................................................................................... xiv

Abbreviations ...................................................................................... xv

1 Introduction ....................................................................................... 1

2 Fundamentals ................................................................................... 8

2.1 Wearable Devices and Adaptive Goal Setting .................................... 8

2.1.1 Wearable Devices ........................................................................ 8

2.1.2 Goal Setting ................................................................................. 10

2.2 ML and Time-Series Analysis ........................................................... 12

2.2.1 Basic Concept of Machine Learning .............................................. 12

2.2.2 Fundamental Knowledge on Time-Series .................................... 14

3 Literature Review ............................................................................ 16

3.1 Goal Setting Approaches & PA ......................................................... 16

3.2 Context-Aware Coaching & PA ......................................................... 17

3.3 ML Prediction models & PA ............................................................... 18

3.4 Intervention Through Wearable Devices ........................................... 20

3.5 Related Work Limitations ............................................................... 20

4 Methodology ..................................................................................... 23

4.1 Data Collection ............................................................................... 25

4.1.1 Wearable Device Data ................................................................. 26

4.1.2 Questionnaire Data ...................................................................... 27
4.2 Data Preprocessing .................................................. 28
  4.2.1 Data Integration .................................................. 29
  4.2.2 Data Cleaning ................................................... 31
  4.2.3 Feature Engineering ............................................ 33
    4.2.3.1 Date & Activity Features .................................... 33
    4.2.3.2 Features Related to Covid-19 .............................. 34
    4.2.3.3 Personality & Identity Features ........................... 35
  4.2.4 Data Transformation .......................................... 41
  4.3 Initial Feature Selection ........................................ 44
  4.4 ML Algorithm Selection & Optimization ......................... 46
    4.4.1 Algorithm Selection .......................................... 47
    4.4.2 Hyper-parameter Tuning ...................................... 47
  4.5 Dimensionality Reduction ....................................... 48
    4.5.1 Feature Selection ............................................ 49
    4.5.2 Dimensionality Reduction through Dataset Projection ..... 50
  4.6 Obtaining Optimal Window Size ................................ 51

5 Experimentation & Results ........................................... 54
  5.1 Evaluation ......................................................... 55
    5.1.1 Metrics ....................................................... 55
    5.1.2 Evaluation on Dataset ........................................ 56
  5.2 Selection of Important Feature-groups ............................. 56
  5.3 Outlier Handling ................................................. 59
    5.3.1 Manual Outlier Detection .................................... 60
    5.3.2 Outlier Detection with Isolation Forest .................... 61
  5.4 Dimensionality Reduction ....................................... 63
    5.4.1 Recursive Feature Elimination (RFE) ......................... 63
    5.4.2 Principal Component Analysis (PCA) ........................ 64
  5.5 Obtaining Optimal Window Size .................................. 66
  5.6 Generalized and Personalized Model’s Performance ............... 68

6 Conclusions & Future Work ........................................... 74

A Questionnaires ....................................................... 77
  A.1 Demographics Questionnaire ...................................... 77
  A.2 Personality Questionnaire IPIP ................................... 78
  A.3 Questionnaire of Transtheoretical Model of behavior change .. 82

B COVID-19 Features .................................................... 87
# List of Figures

1.1 Applications of step count tracking in various devices .......................... 2
1.2 Top 5 Wearable Device Companies 2020 (shipments in millions) [19] .......... 5
1.3 Global smartphone market share: 2018 Q1 - 2020 Q4 [58] ........................ 5

2.1 Fitness tracker market by device type [54] ............................................. 9
2.2 Various activity trackers ........................................................................ 10
2.3 U.S Wearable technology market size [51] ............................................. 10
2.4 ML categories and applications [26] ..................................................... 13

4.1 Schematic representation of the adopted methodology ............................ 23
4.2 Device type and age group distributions in dataset ................................. 26
4.3 Operation of a cryptographic hash function [31] .................................... 26
4.4 Schematic representation of the adopted data preprocessing methodology .. 28
4.5 Apple & Xiaomi raw activity data ......................................................... 29
4.6 Apple data after XML to CSV conversion, parsed as Pandas DataFrame .... 30
4.7 Apple data after time-related preprocessing ......................................... 30
4.8 Histogram of daily step-counts in dataset ............................................ 32
4.9 Schematic representation of the different feature categories .................... 45
4.11 Number of dimensions and available data in dataset, in relation with the window size $n$ ................................................................. 52
4.12 Schematic representation of the model development methodology amongst with the optimal configuration ......................................................... 53

5.1 CV MAE & MAPE for the best performing dataset .................................. 58
5.2 Predictions of the models on 20 days of an unknown dataset .................... 59
5.3 Outlier and non-outlier distribution in dataset for different threshold values .. 60
5.4 MAE and MAPE values for different thresholds in the training and the test sets. ........................................................................................................ 60
5.5 Outlier and non-outlier distribution in dataset for different threshold values (ISO) .. 61
5.6 MAE and MAPE values for different contamination values in the training and the test sets. ........................................................................................................ 62
5.7 CV MAE for various numbers of features chosen ...................................... 63
5.8 MAE and MAPE values for different number of components used by PCA .... 65
5.9 Dataset’s dimensions, MAE & MAPE for each dimensionality reduction approach ........................................................................................................ 65
5.10 GBR model’s performance on the different phases of the adopted methodology .... 66
5.11 MAE and MAPE in relation to the number of days used as input .......... 67
5.12 Outline of the personalized vs generalized approach ............................. 69
5.13 MAE and MAPE of the generalized and the personalized models on the unknown test set ................................................................. 69
5.14 Personalized and generalized models’ prediction VS the actual step-count for
20 days of an unknown test set. ........................................... 70
5.15 MAE and MAPE of the generalized and the personalized models on the un-
known test set of a completely unknown to the generalized model user. . . . . 71
5.16 Personalized and generalized models’ prediction VS the actual step-count for
20 days of the unknown test set of a completely unknown to the generalized
model user. ................................................................. 72
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Outline of the different goal setting approaches</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>Example of how MAE is not indicative of a regression’s algorithm efficiency</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison between WeMoD and other relevant works</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>Activity data format after integration process</td>
<td>31</td>
</tr>
<tr>
<td>4.2</td>
<td>Activity and date features</td>
<td>34</td>
</tr>
<tr>
<td>4.3</td>
<td>Demographics features</td>
<td>36</td>
</tr>
<tr>
<td>4.4</td>
<td>Big-Five features</td>
<td>37</td>
</tr>
<tr>
<td>4.5</td>
<td>Processes of change and corresponding questionnaire’s questions numbers</td>
<td>40</td>
</tr>
<tr>
<td>4.6</td>
<td>Participant’s dataset format after feature engineering process</td>
<td>41</td>
</tr>
<tr>
<td>4.7</td>
<td>General format of input-output data after transformation in time-series format</td>
<td>42</td>
</tr>
<tr>
<td>4.8</td>
<td>Blueprint of datasets after transformation in time-series format</td>
<td>42</td>
</tr>
<tr>
<td>5.1</td>
<td>MAPE versus MAE in step-count prediction</td>
<td>55</td>
</tr>
<tr>
<td>5.2</td>
<td>MAE for datasets with different feature-groups (training set)</td>
<td>57</td>
</tr>
<tr>
<td>5.3</td>
<td>MAPE for datasets with different feature-groups (training set)</td>
<td>57</td>
</tr>
<tr>
<td>5.4</td>
<td>MAE for datasets with different feature-groups (test set)</td>
<td>57</td>
</tr>
<tr>
<td>5.5</td>
<td>Performance for different outlier threshold values</td>
<td>61</td>
</tr>
<tr>
<td>5.6</td>
<td>Performance for different thresholds (ISO)</td>
<td>62</td>
</tr>
<tr>
<td>5.7</td>
<td>MAE &amp; MAPE values for the dataset obtained with the RFE approach</td>
<td>64</td>
</tr>
<tr>
<td>5.8</td>
<td>MAE &amp; MAPE values of the RFE and the PCA approaches</td>
<td>65</td>
</tr>
<tr>
<td>5.9</td>
<td>Dimensionality and available data of the dataset in relation with the window size value.</td>
<td>67</td>
</tr>
<tr>
<td>5.10</td>
<td>MAE and MAPE values for the personalized and the generalized models on an unknown test set.</td>
<td>70</td>
</tr>
<tr>
<td>6.1</td>
<td>Research’s contributions summarization</td>
<td>75</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>Physical Activity</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>AUTh</td>
<td>Aristotle University of Thessaloniki</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>SDG</td>
<td>Sustainable Development Goals</td>
</tr>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>TTM</td>
<td>The Transtheoretical Model</td>
</tr>
<tr>
<td>RFE</td>
<td>Recursive Feature Elimination</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percentage Error</td>
</tr>
<tr>
<td>CV</td>
<td>Cross Validation</td>
</tr>
<tr>
<td>GBR</td>
<td>Gradient Boosting Regressor</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Physical Activity (PA) is any movement that sets our bodies into motion. According to the U.S. Department of Health and Human Services, when adults do the equivalent of at least 150 minutes of moderate-intensity aerobic activity each week, the benefits are substantial [42]. These benefits include lower risk of all-cause mortality, coronary heart disease, stroke, hypertension, type 2 diabetes, certain types of cancer, anxiety, depression, and Alzheimer’s disease and other dementias. Physically active adults also sleep better, have improved cognition, and have better quality of life. As a person moves from 150 minutes a week toward 300 minutes a week, the health benefits become more extensive. For example, a person who does 300 minutes a week has an even lower risk of heart disease or type 2 diabetes than a person who does 150 minutes a week. Regarding PA levels, the principal “some is better than none and more is better than some” strongly applies. In addition to the health-related benefits of PA stated above, there are also several socioeconomic benefits for societies that integrate and encourage PA in people’s daily lives. The cultivation of an enhanced PA culture can lead to reduction in usage of fossil fuels, safer roads, less air pollution and generally higher quality of life [43]. Moreover, as stated by the World Health Organization (WHO), investment in policy actions to increase PA will have a significant contribution in achieving many of the sustainable development goals identified in Bangkok declaration in 2016 [22], including but not limited to, SDG3 (good health and well-being), SDG2.2 (ending all forms of malnutrition), SDG5.1 (gender equality) and more [43]. Despite the significant benefits described above, only 26% of men, 19% of women, and 20% of adolescents report sufficient activity to meet the relevant aerobic and muscle-strengthening guidelines [42]. This lack of PA is linked to approximately $117 billion in annual health care costs and about 10% of premature mortality in America [42]. According to WHO’s report [43], globally, physical inactivity’s cost is estimated in $54 billion in direct health care 57% of which is covered by the public sector and an additional $14 billion from productivity loss.
As we discuss in detail in chapter 3, where we present literature relevant to this research, it has been found in a number of works, including but not limited to [28, 29, 65], that appropriate goal setting and activity tracking of an individual, may lead to improved lifestyle and an increment in the levels of this individual’s PA. One of the most common ways to track PA, is through step-count data. This is due to the relative easiness of obtaining such measurements, through the usage of wearable and mobile activity tracking devices. Most step-based PA programs that aim to encourage the user into reaching the recommended PA guidelines described above, set a daily **step goal of 10,000 steps**. This is due to the fact that for most people 10,000 steps a day can be translated into about 5 miles or 30 minutes of moderate-intensity aerobic activity (walking), which is in accordance with the global recommendations for PA mentioned above.

A major challenge towards increasing the percentage of people that follow the above-described guidelines regarding PA levels, is to encourage individuals who are not physically active in their daily lives, to include PA to a greater extent in their daily routines. Wearable and mobile activity tracking devices can be used in order to increase an individual’s levels of PA [28, 29, 65]. With the appropriate configuration, wearable devices can provide a personalized coaching experience similar to the effective approach of in-person coaching sessions without their high cost.

The most common way in which wearable devices are used in order to increase PA is by tracking the daily step-count of a user and setting a daily number of steps as a goal for the user to achieve. As discussed above, the step-count can be a useful indicator of PA as it is easy to track, is common across the whole population either engaging in PA or not and finally, the recommended 150 minutes of moderate aerobic exercise can be easily translated into approximately 10,000 daily steps. A few examples of such goal setting applications in smartwatches and smartphones can be seen in Figure 1.1 below.

![Figure 1.1: Applications of step count tracking in various devices](image)

The most crucial aspect of this approach is the algorithm by which the daily step count goal is selected by the wearable and mobile devices. As discussed later in section 2.1, there are more than one categories of goal setting methods regarding step count. The simplest and easier to
implement is a "Fixed Goal Approach" in which the device sets a predefined fixed goal each day for the user. It may be the recommended 10,000 steps or a goal that the user has selected for himself. The category which has been found to perform the best in increasing adherence and PA levels [29, 65], is the "Personalized & Adaptive Goal Setting Approach", in which the device personalizes the daily goal set, by taking into account various aspects of the user’s behaviour, while also adapts overtime to incorporate possible changes of different user-specific or contextual factors.

There have been many methodologies to tackle the task of adaptive goal setting utilizing statistics, formulas based on experimental results that are applied to the users historical activity data and more (relevant literature is presented in detail in chapter 3). However, there is a lot of space for further research on this particular field of adaptive goal setting due to the fact that an optimal model, able to achieve the best results regarding enhancing PA and user engagement, is yet to be found. In this work we concentrate on a personalized & adaptive goal setting methodology, since it is more sophisticated to apply and has been found to have better results than the fixed goal approach.

Up to this point, we have already discussed the importance of increasing PA levels of the general population while also explaining how the number of steps an individual takes during the day relates to this task. We have established how the usage of the relatively new technology of wearable and portable activity tracking devices can contribute to this goal and the importance of the goal setting methodology in the efficiency of the adopted approaches. Although as already stated above, there is a lot of research on the topic, an optimal approach in enhancing PA by efficiently incorporating wearable and portable devices’ capabilities has not been defined yet and that is due to the fact that the task on hand is quite challenging. In our work, we have identified a number of open research issues that we attempt to tackle through our contributions. These issues are as follows:

■ [I1] Multitude of different wearable devices and activity data formats. As we discuss in detail in 2.1.1, there is is the huge variety of wearable and portable activity tracking devices leading to great inhomogeneity in the way activity data is recorded and stored. The fact above, has a significant impact in terms of their use in ML algorithms that aim to increase PA. As a result, relevant research using activity data from wearable and portable devices is either limited to the usage of a single type of activity tracker in order to ensure data homogeneity, or need to design and apply an appropriate integration framework that transforms data from different sources to one single usable format. If the integration framework approach is not followed, conducting the research becomes significantly more difficult, as it limits the amount of available data that must necessarily come from one device.
[I2] Daily step-count depends on a number of factors not yet sufficiently identified. Although numerous factors have been considered in predicting a user’s daily step-count (see Chapter 3), the broad nature of the task always leaves room for further research. So far, in most relevant researches the features used are directly related to activity data (daily step-count, steps taken in specific time granularity, minutes of daily activity etc.) and the researches that do utilize personal or contextual features are doing it in a different manner than the approach used in our work, as we discuss in Chapter 3. In addition to the above, to the best of our knowledge, features related to COVID-19 have never been used in relevant literature, even though instinctively we can assume that PA is strongly affected by the restrictive policies applied by governments worldwide.

[I3] Real-life activity data are rare and difficult to come across. Data scarcity is a common problem in research. Especially when the data considered are related to a user’s personality, identity and daily routine. The fact that in most relevant works the data are collected by individuals participating in an ongoing research, may lead to a distorted reflection of the actual levels of their daily activity, as we discuss in Chapter 3. The importance of data “in-the-wild” lies in the objectivity of capturing daily PA levels during a participant’s daily routine.

[I4] No optimal model in handling the task of daily step-count prediction has been developed. Even though there is a number of relevant works, there is always room for improvement and experimentation on this task. In Chapter 3, we elaborate on the fact that in most related researches, details about the performance of the prediction algorithms used are not provided. In addition to that, even when performance is reported, it is not reported in a manner that allows a safe conclusion to be drawn regarding model efficiency. Finally, as it has not been proven that there is a specific limit to the accuracy of daily step-count forecasting models, the issue of developing such a prediction model remains open.

This thesis major motivation is the development of a ML model, trained on real-life users’ data collected “in-the-wild”(WeMoD). The WeMoD approach is important since it has been clearly indicated from earlier work that participants may genuinely change their behaviour mostly during their everyday lives and not through participating in closed preset experiment [67]. Our approach will consider a variety of features including both historical PA and contextual data, such as data related to the COVID-19 global pandemic or the participant’s personality and social identity. Finally, in order to take advantage of different devices, we also design and develop an integration framework for data coming from heterogeneous (wearable) data sources. Details on our integration framework can be found in section 4.2.1. In summary, our contributions are as follows:
[C1] **Creation of a homogeneous wearable data integration framework** which will merge multi-device data (from Xiaomi and Apple devices), thus tackling the open issue I1. These devices might be wearable activity tracking devices such as wrist-watches, or mobile applications dedicated to recording an individual’s PA. This framework is largely needed since it will enable researchers and experts in this field to expand their sample population and to diversify their available sources of activity data. Also, the inclusion of both mobile applications (Apple Health, Mi Fit) and wearable activity trackers, further enhances the importance of our framework, since it removes the limitation for the research’s participants to own a wearable device. In a realistic scenario, we would like to develop an adaptive goal setting approach that would benefit both people with wearable devices and mobile apps for activity tracking, since it cannot be assumed that everyone will own a wearable device. The Apple and Xiaomi brands were chosen in our research due to fact that they are the leading companies in the wearable device market, while also occupying the first and third place respectively in the global smartphone market as seen below in Figures 1.2 and 1.3.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>151.4</td>
<td>34.1%</td>
<td>111.5</td>
<td>32.2%</td>
<td>35.9%</td>
</tr>
<tr>
<td>Xiaomi</td>
<td>50.7</td>
<td>11.4%</td>
<td>41.7</td>
<td>12.0%</td>
<td>21.7%</td>
</tr>
<tr>
<td>Huawei</td>
<td>43.5</td>
<td>9.8%</td>
<td>28.9</td>
<td>8.3%</td>
<td>50.7%</td>
</tr>
<tr>
<td>Samsung</td>
<td>40.0</td>
<td>9.0%</td>
<td>31.4</td>
<td>9.1%</td>
<td>27.3%</td>
</tr>
<tr>
<td>Fitbit</td>
<td>12.9</td>
<td>2.9%</td>
<td>15.9</td>
<td>4.6%</td>
<td>-18.8%</td>
</tr>
<tr>
<td>Others</td>
<td>146.1</td>
<td>32.9%</td>
<td>117.1</td>
<td>33.8%</td>
<td>24.7%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>444.7</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>346.4</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>28.4%</strong></td>
</tr>
</tbody>
</table>

**Figure 1.2:** Top 5 Wearable Device Companies 2020 (shipments in millions) [19]

**Figure 1.3:** Global smartphone market share: 2018 Q1 - 2020 Q4 [58]
- **[C2]** Combination of activity, personality and contextual features to forecast a user’s upcoming daily step count. We incrementally utilize various feature sets for the prediction task and establish the importance of the different feature sets to the overall model’s predictive ability. A fact worth mentioning about this contribution is the exploration of COVID-19 impact on individual’s PA levels, a factor that has not been tested so far to the best of our knowledge. We take advantage of the publicly available datasets related to the COVID-19 pandemic [12], in order to establish the relation between the COVID-19 movement restrictive policies and the predictive ability of ML models that are used to forecast a user’s daily step count. This contribution aim to address the I2 open issue regarding the limited feature space used in relevant literature. The results presented in 5.2 indicate that the usage of personal and contextual features reduced the MAE of the prediction model from 2277 to 2138 steps without any other modifications on the model’s configuration.

- **[C3]** The experimentation with a real-world dataset containing activity, personality and contextual features. Our data are gathered “in-the-wild” as participants go about their daily routine, using mobile devices, wearable sensors, surveys, and other third-party services’ data, and hence are relevant to use in a real-world PA prediction system. It has been found in relevant works [67], that data gathered during an ongoing research may not reflect the natural step-count activity of the participants and thus, affect the results of a research and the predictive ability of the ML models developed based on those data. In the dataset we created, there is no doubt that the data accurately describe the participants’ PA levels, due to the fact that the activity they represent has not been conducted during an ongoing experiment. In this manner the C3 contribution of this research addressed the open issue I3 discussed above.

- **[C4]** Design and development of an end-to-end, stand-alone model, the WeMoD approach, to forecast future PA from a set of rich features. In this process, we also experiment with the development and comparison of two different ML model designs and a variety of ML algorithms trained on the same features. The two model designs correspond to one personalized ML model, trained on the activity data of one specific user and one generalized, trained on the data of all the research’s participants. We perform this experiment, in order to explore the pros and cons of a personalized versus a generalized approach. Through an extended experimental methodology we managed to obtain the optimal configuration of the prediction model, reducing the MAE from the initial 2138 to 1930 steps assessed through CV. This optimal configuration, corresponds to the generalized approach, using only 5 days of activity as input. With the WeMoD model, this research contributes towards the I4 open issue regarding step-count prediction models. The WeMoD approach, incorporates all of the above mentioned contribution points as
it is designed based on a novel dataset which includes a combination of activity, personal and contextual features (C2), with "in-the-wild" data (C3) transformed through an appropriate integration framework (C1).

In this chapter we established the importance of PA for both individuals and the society as a whole. It has been discussed in depth, the fact that an enhancement in the PA levels will lead to an improved quality of life and thus, it is important to aim researches towards this end. The open issues of relevant research amongst with the manner in which this work contributes to the cause of enhancing PA has also been presented.

In Chapter 2, we will discuss some basic concepts and terminology, necessary for the comprehension of this writing. In chapter 3, the literature relevant to our research will be presented, while in chapter 4 we will describe in depth the methodology adopted in our research. In chapter 5, we will present the results of the various techniques and approaches used in the development of our ML models and finally, in chapter 6, we will discuss the conclusions drawn from this research amongst with ideas for future work that may further contribute to our project.
Chapter 2

Fundamentals

In this chapter we briefly present a few fundamental factors that lie at the core of this research and are crucial for the reader that is interested in understanding the basic aspects of our work. We begin by describing the nature and the importance of wearable devices used as activity trackers, amongst with some basic principles of goal setting methodology. Next, we attempt to give a basic definition of ML and time-series analysis, as they are fundamental in our process, in order to enable users non-familiar to this fields to comprehend the concept of our research.

2.1 Wearable Devices and Adaptive Goal Setting

In this section we focus on the concepts of wearable devices, their potential as activity trackers and their social and economic significance.

2.1.1 Wearable Devices

Wearable devices or gadgets are electronic devices worn by consumers ubiquitously and continually to capture or track biometric information related to health or fitness [18]. They may be present as a piece of software that adds extra functionality in accessories that people use in their everyday lives, such as hand watches or mobile phones or they can be found as dedicated devices designed specifically for the task of activity tracking. Wearable technology can be divided into two categories: fitness activity trackers and healthcare activity trackers. The main difference between the above-mentioned categories is that fitness tracking devices monitor a variety of metrics, such as step-count, pulse, distance walked, sleep quality and more, while healthcare tracking devices are worn out of necessity and are usually oriented in monitoring a specific metric such as glucose levels or blood pressure. Due to their aimed purpose, healthcare activity trackers are significantly more accurate in their tracking.
Within the broad category of fitness activity trackers, there are different types of devices such as smartwatches, fitness-bands, clip-ons, fitness straps and more. In our work we focus on fitness activity trackers and in particular watch-like devices due to the fact that fitness trackers are more widespread in the general population than healthcare devices designed specifically for health-monitoring purposes. The fact that we chose devices of the watch genre is due to their dominance in the market of fitness activity trackers as seen in Figure 2.1 below.

![Fitness tracker market by device type](image)

**Figure 2.1:** Fitness tracker market by device type [54]

By focusing on the most popular activity tracking devices, we aim at implementing a more generic and efficient integration framework thus addressing I1. Such a framework will be able to utilize data originated to the largest possible part of the population, while also maintaining low levels of complexity and ensuring the feasibility of its incorporation into different applications that use step-count data from wearable and mobile devices.

Wearable devices became possible due to the advancements in various technological fields. This kind of devices utilize a number of appropriate sensors (3-axis accelerometer, gyroscope, altimeter, etc.) to perform their tracking. The fact that the size of these sensors reduced to such a degree that allowed their integration in devices like mobile phones or watches was vital for the development of wearable technology. Another crucial factor was the development of appropriate platforms that made the gathered data easily accessible to the user. The transformation of the Internet from slow dial-up connections to fast broadband has made it possible to embed data into user-friendly mobile applications with which the user can interact and gain useful insights.
Wearable technology is a field worth studying and working upon developing, since it has a constantly growing share of the global market ($32.63 billion in 2019 at a compound annual growth rate (CAGR) of 15.9% from 2020 to 2027) and as experts suggest, it is expected to expand further with the growing popularity of internet of things (IoT) [51]. The above statement is depicted in figure 2.3 below.

**Figure 2.3: U.S Wearable technology market size [51]**

### 2.1.2 Goal Setting

Most wearable activity trackers set by default a daily step-goal for the user to reach. The purpose of this goal is to encourage the users to track and increase their PA in order to improve their overall health and the quality of their lives as described in chapter 1. The goal setting methodology can be roughly organized in 4 categories that will be discussed in the following subsections.
**Fixed Goal Setting**

Fixed goal setting is the simplest method by which a goal can be defined. A predefined goal value is set as the desirable outcome and the user has to adapt their behaviour in order to succeed. In the fixed goal setting methodology, the most common daily step goal used, is the 10,000 steps (due to the PA guidelines [42] mentioned in chapter 1) which is considered as the equivalent of 30 minutes of moderate intensity exercise. As we discuss later in Chapter 3, there are many studies that indicate that fixed goal setting is the least efficient methodology in increasing PA levels. This approach may discourage adherence as the 10,000 steps per day may be extremely challenging for not so active users or unambitious for users that exercise regularly and are used to being active during their day.

**Personalized Goal Setting**

Personalized goal setting is the methodology of assigning goals, specifically tailored for a specific individual, considering various factors about their identity or activity patterns. Although a personalized but not adaptive approach is rarely found in literature, it can be used as a goal setting methodology in order to set person-specific daily step count goals. Such a research that uses personalized but not adaptive goal setting can be found in [48] where PA goals were assigned to participants based on a clustering process regarding certain aspects of their identities. This research indicated that the personalized approach outperforms fixed goal setting in enhancing an individual’s PA levels. In the personalized goal setting methodology, the goal setting algorithm takes into account various features describing a specific individual and assign appropriate PA goals, but fails to adapt overtime as the user’s behaviour and other factors may change.

**Adaptive Goal Setting**

Goal setting in an adaptive manner means that goals are being adapted through time depending on various factors. Most of the times, adaptive goal setting is also personalized since the factors considered by the goal setting algorithm are most likely to include user-specific features. Having said that, it is not necessary for an adaptive goal setting approach to be personalized since the features used could be only contextual or environmental. To our knowledge there is not such work utilizing a solely adaptive but not personalized approach and the best performing models presented in chapter 3 make use of approaches combining the personalized and adaptive goal setting methodologies.

**Personalized & Adaptive Goal Setting**

Finally, there is the combination of the two previous approaches were the goals are tailored to a specific user and also adapt through time depending on certain factors. In this approach, the daily number of steps required from the user, constantly changes as it is adapted based on
previous user’s behaviour or other features considered relevant with an individual’s daily PA levels. The step count goal is progressively increasing in order to achieve the desired increment of PA. There have been many studies during the recent years which indicate that the adaptive & personalized goal setting approach, is the most efficient methodology regarding enhancement of PA levels [55, 67] and leads to greater user engagement. The relevant literature which utilizes adaptive goal setting is presented in more detail in Chapter 3.

In this section we presented basic information regarding wearable devices and their usage in modern applications as activity trackers, amongst with some fundamental terminology of goal setting approaches used in this research. In Table 2.1 bellow, we provide a summarization of the characteristics of the different goal setting methodologies.

### Table 2.1: Outline of the different goal setting approaches

<table>
<thead>
<tr>
<th>APPROACH</th>
<th>DESCRIPTION</th>
<th>COMPLEXITY</th>
<th>EFFICIENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>Assigns a fixed goal regardless of the user</td>
<td>Simplest approach</td>
<td>Least efficient approach</td>
</tr>
<tr>
<td>Personalized</td>
<td>Goal depends on user’s personal characteristics</td>
<td>More sophisticated than fixed</td>
<td>Better than fixed but not the most efficient</td>
</tr>
<tr>
<td>Adaptive</td>
<td>Goal adapts over-time, based on user’s behaviour and other factors</td>
<td>Need for an appropriate adaptation algorithm</td>
<td>Has not been seen in literature</td>
</tr>
<tr>
<td>Personalized &amp; Adaptive</td>
<td>Goal is personalized for each user and also adapts overtime</td>
<td>Needs algorithms for personalization and adaptation. The most complex approach</td>
<td>The most efficient approach, with a number of works proving it</td>
</tr>
</tbody>
</table>

The question that occurs naturally, is how can we set the appropriate challenging yet achievable personalized and adaptive goals for a user. The best way in achieving this, would be to have the ability to predict a user’s future activity and set the appropriate goals based on the given predictions. Fortunately, ML and its applications can help us in such endeavor. In the following section we will discuss the basic concepts of ML and time-series analysis, concepts on which this research is built upon.

### 2.2 ML and Time-Series Analysis

#### 2.2.1 Basic Concept of Machine Learning

ML is a field of computer science that utilizes algorithms able to learn from given data [8]. The process of learning is not a goal by itself but a means towards accomplishing a certain
task. In other words, ML can be used in order to complete a task that is too complicated for a human-designed program to accomplish.

The algorithms used in ML can be divided into 3 categories (Figure 2.4) regarding the data used as input and the kind of expected outcome [53]. The first category is supervised ML in which the task is to learn a function that maps an input to an output based on example input-output pairs [52]. It infers such a function from labeled training data consisting of a set of training examples [40]. The second category is unsupervised ML where the task is to uncover previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision [15]. There is also an intermediate ML setting, reinforcement ML, in which an AI agent must learn how to efficiently operate in a given environment in order to accomplish a specific goal. In reinforcement learning, there is no fixed training dataset, rather a goal or set of goals that an agent is required to achieve, actions they may perform, and feedback about performance toward the goal [23].

![Figure 2.4: ML categories and applications [26]](image)

In this work we will focus on supervised ML techniques in order to perform a throughout analysis of time-series PA data and design a model capable of predicting the daily PA of a user in the form of the number of steps taken during a day. For this goal to be achieved, a number
of preprocessing steps and transformations of the raw data is necessary in order to obtain a format appropriate for usage in supervised ML algorithms that will enable us to handle data presented in a time-series format. Our methodology towards this end is discussed extensively in section 4.2.

### 2.2.2 Fundamental Knowledge on Time-Series

Time-series is a set of observations each being recorded at an equally spaced time interval (annual GDP of a country, monthly sales figure, daily step-count, etc.) [14]. The objectives of analyzing time-series mainly fall into three major categories [14]. Firstly, forecasting: predicting the future has always been a task that humans strive to accomplish throughout our history. Secondly, control: time-series analysis is used in production processes to assess if everything is going according to plan or whether the system faces any anomalies. Finally, analyzing and identifying features of the data including seasonality, cycle, trend, and its nature. In many cases the above objectives may overlap, as for example understanding trends and seasonality of the time-series data may be useful in forecasting and control applications. In our research, the task at hand falls in the first category, forecasting, as we aim to predict the number of steps a user will take on a specific day in the future, based on observations of certain features in the past and present. As we have mentioned, predicting the future is a skill that has been of great importance to humanity and therefore, the existence of different methods that aim in this direction is expected. Time-series analysis is a field that has been approached by a number of different techniques, many of them based on mathematics and statistics, existing much earlier than ML approaches. As found in [33], traditional methods are by no means obsolete, since in some simpler problems of univariate time-series analysis they may perform better than complex ML models.

In this research, and given the complexity of the problem which includes a multitude of variables for each time point in the time-series (multivariate time-series), it was decided that it would be most efficient to focus on approaches that utilize ML. An additional factor considered in the decision of utilizing ML instead of statistical approaches is the fact that mathematical techniques take into account one time-series at a time, which means that we would be limited in using data from a single participant. On the contrary, ML approaches enable us to combine data from all users, a possibility we wanted to explore for our generalized model.

Based on the above and since the traditional supervised ML algorithms have been found to perform satisfactorily on corresponding prediction problems and even outperform classical approaches [24], we decided to focus our research on supervised ML approaches. Thus, time-series analysis is used in our work in order to identify meaningful patterns in the user’s PA through data gathered from wearable activity tracking devices. By analyzing historical activity
data regarding the user’s step count, combined with contextual and personal features we build ML models which through supervised ML approaches, are able to predict the daily step count of the next day.

This chapter’s aim was to introduce the basic concepts of wearable activity trackers, goal setting methodologies, ML and time-series analysis to a non-familiar to this terms user, in order for them to be able to follow our research as reflected in this writing. There is a vast body of work about various aspects of the above concepts and a throughout analysis is out of the scope of this research. In the following chapter we will present our methodology regarding the various phases of our process, all the way from data collection to the evaluation of the performance of the step-count predicting ML models developed.
Chapter 3

Literature Review

As stated in Chapter 1, increasing the PA levels of the general population is a goal of high importance which, if achieved up to an adequate degree, would result in great individual and social benefits. Thus, it is only natural that in the recent years there has been a lot of research towards achieving this goal. A significant body of relevant work, explores the manner in which wearable and mobile devices could be used towards achieving the improvement of PA levels or positive behavioural change in general. In this chapter we will present briefly, part of this work and further clarify our contribution, stated in Chapter 1.

3.1 Goal Setting Approaches & PA

There have been several works studying the impact that appropriate goal setting has on increasing an individual’s PA levels, measured in the form of daily step-count data.

In [67] the authors designed a mobile app that delivered daily step-count goals to an intervention and an active control group. The intervention group received personalized goals adapted to their previous activity data whereas the active control group received fixed goals of 10,000 steps. This research led to a statistically significant difference in favour of the intervention group indicating that adaptive goal setting is indeed more effective in increasing PA than fixed goal setting. Both groups had reduced daily step-counts compared to their run-in activity and this is most likely due to the fact that participants were receiving step goals and step-count monitoring from the beginning of the experiment and thus the run-in step-counts were higher than their natural baselines. In our work, it is granted that the data collected are representing the participants’ natural behaviour and daily routines as they have been recorded throughout the usage of their activity tracking devices, not limited to a certain research period. Our data collection process is described in detail in 4.1.
Similar work to [67] has been conducted in [63] and [68]. In [63], the impact of context-aware personalized goal setting was evaluated through testing on a context-aware and an active control group. Once again the results indicate that personalized goal setting is more efficient towards increasing PA levels as expressed through the daily step-count. In [68] the researchers explored the utilization of reinforcement learning algorithms in personalized and adaptive goal setting, on 13 college students. The participants receiving personalized goals increased their PA levels, in contrast with a control group that received fixed daily goals of 10.000 steps.

In [39] the researchers developed an Artificial Intelligence (AI) coach who set PA goals after consideration of various factors. Through a 6-week observational study, this approach has proven to successfully lead to an increment of the exercise volume performed during the time period of one week.

Finally, in [48], an idiographic (person-specific) approach has been tested in order to establish the efficiency of dynamic models of PA in the context of goal setting and positive reinforcement interventions. Similarly to the previous works, the research concluded that the personalized goal setting approach had a positive impact on increasing PA levels.

It is worth to be mentioned that for all the researches presented above, the performance of the prediction algorithm that is used in order to define the step-count goal is not reported. Thus, we have no way of telling if the algorithm behind the personalization was actually successful in predicting a user’s PA. In our work we will develop a ML approach based on an "in-the-wild" dataset containing activity, contextual and personality features able to predict a user’s daily step-count for a day in the future. A model like this could be used in personalized goal setting approaches, which amongst with appliance of appropriate intervention methods could lead to a complete application that can effectively contribute to the increment of PA levels.

3.2 Context-Aware Coaching & PA

Another relevant research topic is the impact of personalized, context-aware coaching towards increasing PA.

In [63], the authors developed a mobile application which sets a daily step goal and delivers personalized text notifications that take into account a variety of contextual features, such as the location of the user or the local weather data. The participants were divided into two groups. The first group received only a daily step-count goal where the second group received a daily step-count as well as personalized text notifications aimed to increase the individual’s daily PA. The results of the research indicate that although the differences between the two groups are not statistically significant the context-aware coaching is effective.
In [48], personalized goals (but not adaptive) were delivered, based on a participant’s profile developed through system identification (system ID) practices and the baseline step-count obtained in a 2-week period in the beginning of the research. The conclusion of this approach was that different person-specific factors (stress levels, typicality of day, weekend/weekday and more) can affect the intervention’s outcome and thus personalized approaches are more appropriate in order to enhance PA.

In another study [29] the researchers used once again a mobile application to deliver personalized notifications (e.g. reminders, encouraging text messages etc) based on users past PA. Their model estimated hourly the probability of user’s adherence to his daily step goal and delivered appropriate coaching messages or even suggestions of new adapted activity plans.

Artificial intelligence agents have also been used in order to deliver context-aware coaching. In [39] the researchers embodied an AI coaching agent in a smartphone application which through interaction with the user could provide encouragement or suggestions about activity plans. A limitation of this research is that the step-count data collected solely through self-reports and no integrated sensors were used for objective measurements.

Our work is relevant to the above researches as it utilizes a number of contextual and personal features in order to effectively predict a user’s daily step-count. A model of such kind can be incorporated in personalized and adaptive goal setting approaches, a vital attribute of an efficient intervention application that aims to increase a user’s PA levels. The combination of social identity, activity, date and COVID-19 features used in our models, has never been used before to the best of our knowledge. Moreover, we make use of integrated functionality of mobile and wearable devices in order to obtain objective step-count measurements instead of relying to participant’s self-reporting records, an approach that could easily lead to inconsistent data and reduced user adherence.

### 3.3 ML Prediction models & PA

Since our research in its core explores the contribution of ML techniques towards obtaining accurate predictions of an individual’s daily step-count, it is considered appropriate to include in our relevant literature review, other approaches that utilize ML in handling the broader task of increasing PA through the usage of wearable and mobile activity tracking devices. In a similar manner to the one of section 3.1 most of the works presented here do not report the performance of the step-count prediction models used. This matter is discussed in greater extent in 3.5.
In [67], the researchers developed a ML model that used the user’s historical data regarding daily step-count and goal achievement rate and generated challenging yet realistic goals that would maximize the future PA.

Another research [5], uses ML in order to develop a model capable of predicting if a user is going to achieve his daily step-count goal or not, based on variables such as the hour of the workday, the number of steps for that hour and the cumulative number of steps up to the moment of the prediction. A similar approach is presented in [29] where the researchers used ML in order to develop a model that calculates hour by hour the probability of a user to achieve his daily step goal. The model takes into account past activity patterns as well as the current activity target in order to deliver the desired prediction. Based on the predicted probability a number of intervention methods (notifications, reminders etc) or even a dynamic update of the daily activity plan could be applied.

A possible limitation of the above works is that they do not take into account contextual and personal features that could probably enhance further the quality of the suggested goals, an approach that we explore in our work.

In [38] the authors developed a neural network model that considered a number of contextual features derived from questionnaires (personal, social and environmental features) as well as the daily step-counts over the period of a week (7 days) and predicted the average weekly step-count of an individual. This weekly average step-count could later be distributed among the days of the week in order for an appropriate personalized goal to be set. In our work we also use contextual and personal features but deferentially from the research described above, we train the model on a period of 5 days, in order to obtain a prediction about the estimated step-count of the next day (day number 6). We chose not to adopt the weekly average approach since it does not consider the different characteristics of the days in a week (weekends/weekdays, holidays, day-specific activity patterns etc).

Another approach that makes use of ML with data gathered from wearable devices, is presented in [66]. The researchers used logistic regression and support vector machine methods to predict if a user is likely to adhere to his PA plan. The models used the individual’s past data (e.g. activity duration, activity intensity and goal achievement rate) to estimate the probability of discontinuing in the forthcoming week. A participant having an exercise relapse is defined as a participant who had a lower average weekly step-count than his baseline step-count obtained through a 3-week run in period. Such an approach can be used in targeted interventions or personal coaching sessions with individuals considered more likely to relapse. This work, also does not consider the importance of features other than activity-related ones, a factor that is explored in depth in our work. In addition to that, as we discuss in 3.5, the transformation of a regression problem such as step-count prediction to a binary classification problem is an oversimplification that ignores vital aspects of the problem.
3.4 Intervention Through Wearable Devices

A vast number of studies have also focused on intervention methods which utilize activity data obtained through wearable and mobile devices in order to increase an individual’s PA levels.

In some works [29, 63], the users were receiving personalized text notifications in order to inform them about their progress or encourage them to continue working towards their assigned goal.

Another intervention approach is proposed in [48, 67], where the participants of the research were given financial motives in the form of Amazon giftcards whenever a certain degree of goal accomplishment was reached. This approach has the limitation of the fact that since the financial motivation can only take place inside the closed research environment, upon terminating the experiment and the financial motives cease to exist the participants are more likely to relapse.

In [65] the researchers experimented with a different intervention approach. Throughout the experiment they asked the participants to submit daily forms about their stress levels and self determination on this particular day. Later they combined this data and examined the correlation with the PA performed and informed the participants about factors that affect their activity. The result of this research was that a one-time brief message describing personalized exercise predictors can have a positive impact on increasing PA levels.

In [28] the researchers studied the possibility of predicting the effect of an intervention on a specific individual before the intervention is given. They clustered the participants considering appropriate features and experimented with two different intervention methods in order to examine if one method is more effective than the other for similar users. In this way a new user could be assigned to a specific cluster and the optimal intervention method can be applied automatically to him.

In all of the above mentioned works, it has been established that appropriate intervention is an effective approach towards increasing PA levels, when combined with an efficient goal setting methodology. Our work, aims to fill this gap by providing a tool for personalized goal setting through the development of a ML model, capable of predicting a user’s future PA through their daily step-count by utilizing a number of novel features towards this end.

3.5 Related Work Limitations

In regards to the relevant works that have been presented above, we can point out certain limitations that we have identified. Firstly, most of them, even though they report positive
results for the significance of personalized goal setting, they do not offer evaluation details and do not mention the performance of the algorithm used in setting the daily goals. Our work differs, since it provides proof (through evaluation) about the efficiency of the recommended algorithm.

Secondly, even in [38] which is the only work presented in this chapter that reports the goal setting algorithm’s performance, the performance is reported as a MAE of 1672 steps and an adjusted- $r^2$ of 0.85. Those metrics, as we discuss further in 5.1.1, only partially reflect the efficiency of a prediction algorithm, since the MAE depends on the real step-count values included in the dataset. For this to be understood, consider the example of Table 3.1.

<table>
<thead>
<tr>
<th>REAL</th>
<th>PREDICTED</th>
<th>MAE</th>
<th>ERROR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.000</td>
<td>1.500</td>
<td>1.500</td>
<td>50%</td>
</tr>
<tr>
<td>10.000</td>
<td>8.500</td>
<td>1.500</td>
<td>15%</td>
</tr>
</tbody>
</table>

We can see that the MAE is inadequate to indicate the performance of a regression algorithm when there is no information of the range of the values included in the dataset, information that is not provided in [38]. In our work, besides MAE we also report the MAPE metric which gives a better insight of our model’s performance and predictive efficiency.

Finally, in a number of the above researches [5, 29, 66], PA prediction is essentially considered as a binary problem with 2 possible outcomes, success and failure, while in truth it’s a much more complex regression problem. As stated in Chapter 1, the guidelines for PA levels recommend a daily step-count of 10,000 steps. By treating this goal (or any other step-count goal) as a binary task which is either achieved or not, one fails to identify all the underlying intermediate stages. In such an approach, let the given goal be 10,000 steps, a step-count of 9,000 is treated in the same manner as a step-count of 4,000, while in fact there is significant difference between those two performances, a difference that should be taken into account by an efficient step-count increasing application.

All of the above presented works, contribute to some extent in addressing some of the open issues (I1 - I4) identified in Chapter 1. In Table 3.2 bellow, we present a comparison of the WeMoD approach and relevant researches regarding the open issues they are addressing. Through this comparison we hope to highlight the contribution of our work in this field of research.
### Table 3.2: Comparison between WeMoD and other relevant works

<table>
<thead>
<tr>
<th></th>
<th>[67]</th>
<th>[63]</th>
<th>[68]</th>
<th>[39]</th>
<th>[48]</th>
<th>[29]</th>
<th>[5]</th>
<th>[38]</th>
<th>[66]</th>
<th>WeMoD</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1: Usage of Integration Framework</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>N/A</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>I2: Usage of Personal &amp; Contextual Features</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>I3: Usage of &quot;In-the-wild&quot; Data</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>I4: Creation/Incorporation of a step-count Prediction Model</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Regarding Table 3.2, we should clarify 2 points. Firstly, while in [38] 3 different activity tracking devices were used (FitBit Charge, FitBit One and FitBit Zip), they are all different models of the same brand and no integration framework is mentioned in the research. Thus, [38] does is not considered to address I1. Secondly, [29] addresses all of the I2, I3, and I4 open issues and it is not mentioned by the authors (N/A) if it also addresses I1. Despite that, [29] is one of the researches mentioned above, that consider the PA prediction problem as a binary classification task and not as a complex regression problem. Thus, this research is experimenting with the task of PA prediction and improvement from a totally different point of view than the one we use in our work.

In this chapter we presented a review of literature relevant to our work and discussed how our research differs, amongst with our contribution. It is obvious from the work that has been done in the short period that wearable and mobile devices are used for activity tracking, that the utilization of such data in order to increase PA levels is an important topic that has attracted a lot of interest. In our work, we build upon previous knowledge and propose the WeMoD solution, that utilizes not only historical activity data but also contextual and personal features of the participants, acquired through appropriate questionnaires as presented in depth in 4.1 and 4.2.3. We apply time-series analysis in order to predict the number of steps a user will take in the next day, a factor that can be used in future works in order to assign a doable yet challenging daily step goal. We elaborate more on our methodology in Chapter 4 and the results of our research in Chapter 5.
Chapter 4

Methodology

The purpose of this chapter is to present the methodology adopted in our research. Our process aims to implement our contributions as described in Chapter 1. All the individual approaches that have been tested and are discussed in this chapter, can be divided into 2 broad categories, depending on the ultimate goal they are aimed to achieve as seen in Figure 4.1.

Figure 4.1: Schematic representation of the adopted methodology
On the one hand we have the approaches that aim at creating an innovative dataset thus satisfying the contributions C1, C2 and C3 (sections 4.1 to 4.3), while on the other hand there are the techniques and practices that aim at developing the best possible ML model (WeMoD) that will be based on the above mentioned dataset, thus implementing contribution C4 (sections 4.4 to 4.6).

From the above image, one may notice that the "dataset creation" process results into 2 different datasets, one generalized and one personalized while the same also applies to the "model development" process which results in 2 different prediction models. This is due to the fact that in our experimentation we also compare the performances of a model based on a single individual’s data and a model which is trained on the whole dataset, including data from all the participants of the research. We adopt this approach, in order to conduct an experiment on the performance, pros and cons of a generalized model versus a personalized one. Through this experimental process, we draw some useful conclusions and lay the foundations for further research on the issue.

Of course, each one of the individual actions presented in Figure 4.1, such as "data preprocessing" or "dimensionality reduction" is in truth, a complex set of different techniques and approaches which are analyzed and discussed in detail in the following sections of this chapter.

Firstly, we focus on the part of our methodology which is aimed at the creation of an appropriate dataset and we explain in what manner we implement the contributions C1, C2 and C3. In 4.1 we focus on the data collection process applied in order to create a novel dataset appropriate for usage by ML algorithms (C3). The methodology by which the participants and the wearable activity tracking devices were chosen, as well as information regarding the various aspects of the data gathering process are discussed in depth. In the next section, 4.2, we describe the necessary preprocessing steps taken, in order to convert the raw information collected in the "data collection" phase into meaningful features that will contribute to the predictive ability of the ML models (C2). In the section following data preprocessing, section 4.3, we present the manner in which the initial set of features that should be included in the final dataset were selected based on their contribution on the performance of the ML models.

Secondly, in the sections of this chapter that follow the ones mentioned above, we discuss the next part of our methodology, the part that includes the approaches that are aimed in the development of an efficient ML prediction model. In 4.4 we describe the manner in which we selected the best performing algorithm for our model, amongst with our approach in obtaining its optimal configuration. Next, in 4.5, we explore the necessity of reducing the number of features of the original dataset. Such an approach may reduce the time required for ML models to be trained and even lead to improved quality in predictions [25]. In this section, we present the two different approaches tested in reducing the dimensionality of the original
dataset: RFE and PCA. Finally, in section 4.6, we describe the process adopted in selecting the optimal window size, meaning the number of days that are given as input to the ML model in order to predict the step-count of a day in the future.

4.1 Data Collection

One of the most important factors in tackling a problem utilizing ML methods is the dataset that will be used to train and evaluate the ML models. In our case, the problem we are tackling using ML is the problem of predicting the number of steps a person will take during a day. As stated in Chapter 1, there are 2 open issues regarding relevant data. The first is that they are mostly focused on features related to activity while a multitude of other factors (personality, context) that may affect PA have been inadequately tested (I2). The second issue, is the lack of relevant “in-the-wild” data. Data are in most cases (see Table 3.2) gathered from users during a period in which they were actively participating in an ongoing research (I3). This fact has been found that may lead to a distorted reflection of an individual’s actual PA levels [67].

In this work, we aimed to create a dataset that is based on features relevant to the PA levels of previous days, the personality of this particular individual, time-relevant features deriving from the specific date that a number of steps occurred and features related to COVID-19 movement-restrictive government policies (C2). In addition to that, we make use of “in-the-wild” data (C3), submitted in a voluntary basis by the participants of this research. So, the first step in our project was the collection and processing of the necessary data.

For the purpose of this research, 21 adults, users of activity trackers of either Xiaomi or Apple brand, enrolled on a voluntary basis from the research team. The participants were of various ages with the majority of them, 8, belonging in the 25-34 age group, 7 in the 35-44 age group, 5 in the 18-24 age group and 1 in the 45-54 age group. Finally, 10 of the participants were male while 11 were female. The distribution of Xiaomi and Apple devices amongst with the different age groups of the participants can be seen in Figure 4.2 bellow.

In order to create the dataset required in our work we had to collect various information from our sample population that volunteered to participate in this experiment. This information can be roughly divided in 3 broad categories: (1) activity and time related data, extracted from wearable devices, (2) personal data obtained through questionnaires that the volunteers were required to submit and (3) data related to the COVID-19 global pandemic [44].

Data collection and processing were conducted under the guidelines of EU General Data Protection Regulation (GDPR) (2016/679) [46] with pseudonymization of the data collected, using a one-way cryptographic hash function [34] (SHA256), in order to make it impossible to match a
piece of information to the specific participant that submitted it. All participants provided written, informed consent and the study was approved by AUTh Ethics Committee (254324/2020).

Figure 4.3 presents an example of the SHA hash function’s operation.

![Diagram](image)

**Figure 4.2: Device type and age group distributions in dataset**

**Figure 4.3: Operation of a cryptographic hash function [31]**

### 4.1.1 Wearable Device Data

As stated above, all of the participants were users of either an Apple or a Xiaomi activity tracking device. The choice of these specific brands was made due to their popularity as discussed in Chapter 1 (Figures 1.2, 1.3). The lack of works that incorporate more than one type of activity tracking device (Table 3.2) is an open issue (I1) of the relevant literature as discussed in Chapter 1. In our work, we incorporate an integration framework of the 2 most popular brands
of mobile and wearable activity tracking devices in order to expand the pool of available data sources (C1).

All of the participants were asked to extract their recorded activity data, available through the respective mobile applications or the web platforms of the activity trackers they use and submit this data to our research team. Upon submission, the data were pseudonymized and stored for further processing. From the activity data collected, besides the number of steps taken on each day, we were able to extract time-related features derived from the metadata of the recorded dates of activity. The feature engineering process is described in detail in 4.2.3.

4.1.2 Questionnaire Data

In order to obtain and include contextual features to our research we asked the participants to fill and submit three questionnaires. Due to the fact that all of the participants’ mother tongue is Greek, the language used in the questionnaires was also Greek. For two of the questionnaires that were originally published in English, a formal published Greek translation was used.

The restrictive measures that the Greek government has decided to apply due to the COVID-19 global pandemic were active throughout the whole time period this research was being conducted and thus, in the process of questionnaire submission, the Google Forms [10] platform was used. The above means, that the data gathered were temporarily stored on a European server which as stated by Google is GDPR-compliant [9].

In detail, the questionnaires used were the following:

- Demographic questionnaire regarding general information about each participant.
- The Greek version [62] of the Big-Five 50-item personality questionnaire [7].
- The Greek version [1] of the questionnaire of transtheoretical model of behavior change TTM [16, 17, 49].

All of the above questionnaires can be found in Appendix A. The process of feature extraction from the raw information provided by the questionnaires is presented in 4.2.3.

In this section we discussed the process by which the data used in our research were collected and in what manner we implemented the contributions C2 and C3. We also described the sample population that was considered in the data collection process amongst, with the nature of the data gathered. In the following section, we will present the data preprocessing methodology, a vital step in the process of creating our dataset.
4.2 Data Preprocessing

A simple definition of data preprocessing could be that it is the data mining process by which raw data gathered from various sources is transformed into meaningful information, appropriate for statistical analysis and usage in ML algorithms [27]. Data preprocessing includes cleaning, normalization, transformation, feature extraction and more [27]. The most common problems that can be found in raw data can be assigned into 3 categories [6]:

- **Inconsistent data**: data often comes from different sources and may be stored in various formats and files. Possible duplicates across the different files, misspelling errors and errors in used notation may lead to low quality data that would have a negative impact in the data analysis process. In our work we tackled this problem in the integration process that is presented in 4.2.1.

- **Noisy data**: human mistakes, malfunctioning sensors and rare exceptions may provide data that offer meaningless information. Before any analysis, in order to ensure the robustness of the drawn conclusions, it is necessary to handle the existence of such data in the dataset used. The process towards this end, that was adopted in our research is described in depth in subsection 4.2.2

- **Missing data**: it is very common for missing data to create gaps that can affect the final analysis. In this project, missing values were a significant factor to consider, since as stated in section 4.1 which described the data collection process, the activity data gathered were not recorded in a fixed time period and thus, there were a lot of dates with no recorded activity for every participant. Furthermore, even the data originated to each individual, contained dates with no recorded activity. In subsections 4.2.2 and 4.2.4, the manner in which the above issues were tackled is presented in detail.

Based on the above, it is obvious that data preprocessing is a step of great importance in the creation of a dataset appropriate for usage by ML algorithms. In our methodology, and since the creation of such a dataset is part of our contribution, we had to experiment in depth with different techniques and approaches in order to ensure a robust and viable outcome. A schematic representation of our data preprocessing methodology is presented in Figure 4.4 bellow.

![Schematic representation of the adopted data preprocessing methodology](image-url)
In the following subsections we further elaborate on the preprocessing stages shown in the above figure and discuss the methods utilized in our research in order to implement them in the most efficient manner possible.

4.2.1 Data Integration

Due to the fact that our data originated from wearable devices and mobile applications of two different brands, Apple and Xiaomi, there were many differences in the format of the activity data that the participants submitted to the research team. Apple users are given the ability by Apple to download their activity data in the form of an "XML" file while for Xiaomi the same data can be exported to a "CSV" file. In image 4.5 below, an example of the initial format of the data upon submission is presented.

![Image of Apple and Xiaomi raw activity data](image)

**Figure 4.5: Apple & Xiaomi raw activity data**

It was decided that the most efficient manner in handling the different data formats would be to extract the relevant information of Apple’s data and present it in the same format as Xiaomi. As XML is a markup language used to present information in a structured manner [3], the conversion from XML to CSV is a relatively simple process. There is a number of such converters available online and for the sake of this project we used the converter that can be found in [20]. After the initial conversion of the XML files to CSV, Apple data were transformed to a relatively easier for humans to comprehend format with rows and columns. A snapshot of the new format, parsed as a Pandas DataFrame object [45], is given in Figure 4.6 bellow.

At this stage there are two important issues regarding Apple’s data. Firstly, it is obvious from Figure 4.6, that step activity is recorded with corresponding timestamps for the beginning and ending of the activity. Moreover, the timestamps are in UTC time and have not been adjusted to the local timezone something necessary in order to extract the daily step-counts we find in Xiaomi data. Secondly, an issue not so obvious at first glance, is that if a participant was using
both an iPhone and an Apple Watch there would be overlapping or duplicated values in the data due to the different data sources.

<table>
<thead>
<tr>
<th>type</th>
<th>unit</th>
<th>creationDate</th>
<th>startDate</th>
<th>endDate</th>
<th>value</th>
</tr>
</thead>
</table>

Figure 4.6: Apple data after XML to CSV conversion, parsed as Pandas DataFrame

In order to adjust the time-related information, some data preprocessing was needed. Various time references deriving from each date such as year, month, day of month, hour and day of week were extracted using the Python timezone package [2]. The format of the Apple data after the time-related preprocessing was as in Figure 4.7 bellow.

<table>
<thead>
<tr>
<th>type</th>
<th>unit</th>
<th>creationDate</th>
<th>startDate</th>
<th>endDate</th>
<th>value</th>
<th>year</th>
<th>month</th>
<th>date</th>
<th>day</th>
<th>hour</th>
<th>dow</th>
</tr>
</thead>
<tbody>
<tr>
<td>StepCount</td>
<td>count</td>
<td>2015-03-19</td>
<td>01:34:57 +0200</td>
<td>00:52:57 +0200</td>
<td>18</td>
<td>2015-03-19</td>
<td>2015</td>
<td>2015</td>
<td>19</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 4.7: Apple data after time-related preprocessing

To tackle the second issue, the possible duplicated values, we decided that for overlapping or duplicated time periods we should discard the data originated to the Apple Watch as both Apple Watch and iPhone are quite similar regarding accuracy in step-count recording [37].

After the process described above, it was quite simple to transform the Apple users’ data to the same format as Xiaomi though an aggregation of the various step-count measurements by date. Due to the fact that Apple’s data was containing only step-count information, the additional columns (distance, calories etc.) of activity data shown in Figure 4.5, originated to Xiaomi devices were discarded. The final format of both Apple and Xiaomi activity data, after
applying the integration framework described in this subsection was as shown in Table 4.1 below.

Table 4.1: Activity data format after integration process

<table>
<thead>
<tr>
<th>DATE</th>
<th>STEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020-01-20</td>
<td>8808</td>
</tr>
<tr>
<td>2020-01-21</td>
<td>3870</td>
</tr>
<tr>
<td>2020-01-22</td>
<td>8822</td>
</tr>
<tr>
<td>2020-01-23</td>
<td>4352</td>
</tr>
<tr>
<td>2020-01-24</td>
<td>12410</td>
</tr>
</tbody>
</table>

In this subsection we described our approach in designing and implementing an integration framework for data originated to Apple and Xiaomi activity tracking devices. As already stated, this integration framework between the 2 most popular brands of the wearable activity trackers’ market, is considered as part of the contribution of this research (C1) since it enables future researchers to expand the pool of data available to them.

After designing and applying the integration framework to all of the activity data submitted by the research’s participants, the next step was data-cleaning in order to ensure that the dataset to be used in ML algorithms in later phases does not contain noisy data and meaningless information that could have a negative impact to the data analysis process. The data-cleaning methodology is described in detail in subsection 4.2.2 that follows.

4.2.2 Data Cleaning

This subsection describes the methodology followed in order to handle noisy data that would have a negative impact on the efficiency of the ML prediction models developed in the following phases of this research.

Handling of No-Wear Days

After the data integration process, the first issue that had to be addressed was the handling of no-wear days. These are the days whose daily step-count indicate that the user was not wearing or carrying the activity tracking device during that date. For activity data recorded by iPhone, there were no days with zero number of steps. If during a date there were no steps recorded, this date was excluded from the final dataset. Xiaomi devices on the other hand, assign a zero number of steps to days with no recorded activity and do not discard them from the dataset exported by the user. One more factor that should be considered in handling no-wear days is that even if the user carries the activity tracker for a limited amount of time, for example to move it from a desk to a drawer, the activity tracker would record a small number
of steps. A day like this is not representative of the user’s daily PA and should not be taken into consideration in the data analysis process as it would distort the drawn conclusions and results. Based on the above, it was decided that days with less than 500 recorded steps should be discarded from the dataset as no-wear days. The limit of 500 steps has been proven to effectively distinguish no-wear days in a number of relevant works [30, 35, 36, 48, 56].

**Outlier Handling**

After no-wear days were removed from the dataset, another factor to be considered regarding the data cleaning process was the existence of days in the dataset that had an unnaturally large number of daily steps recorded. Such measurements, as mentioned at the beginning of section 4.2 may occur due to malfunctioning sensors, human error during data collection, very rare occasions that are not representative of a participant’s daily routine and more. Thus, they should not be included in the dataset, as they may distort the predictive ability of the ML models or lead to overfitting to the training data. A histogram of the numbers of daily steps in our dataset as seen in Figure 4.8 indicates that such values are included (e.g. there is a value of over 70,000 steps in a day) and thus, our models will most likely improve their performance through appropriate outlier handling.

![Figure 4.8: Histogram of daily step-counts in dataset](image)

In this work, 2 approaches were tested in order to experiment with the impact of outlier identification and removal, the manual approach and the usage of an Isolation Forest [32].

For the manual approach, the recordings of the dataset were sorted by their daily step-count numbers in a descending order and the top recordings up to a threshold (e.g. 1%, 2% of the dataset etc.) were removed. We tested various threshold values and evaluated the affect on the
models’ performance with both CV and the usage of a test set as we describe in 5.1. The results of this method for the different threshold values are presented in 5.3.1.

In our second approach, we utilized the Isolation Forest algorithm as implemented in [47]. In contrast with our manual approach, the Isolation Forest takes into account the input features of a data point in order to decide if it is an outlier or not. Thus, the step-count number of a day, that in our research is the target variable, is not considered in the outlier identification process. Isolation Forest, also needs the user to define a contamination value, which indicates the percentage of outliers to be detected in the dataset, in a similar manner to the threshold value of our manual approach. A number of different contamination values were also tested in applying the Isolation Forest approach and the given results of our evaluation process can be found in 5.3.2.

4.2.3 Feature Engineering

This subsection is about the methodology adopted in order to extract meaningful features, from the raw information presented in the format that has occurred after the data integration and data cleaning processes, described in subsections 4.2.1 and 4.2.2 respectively. By the approach discussed here, we describe the manner in which we implement the C2 contribution of this research, meaning the creation of a dataset that includes features derived not only from PA patterns but also from personal and contextual factors.

4.2.3.1 Date & Activity Features

After applying the integration process described in 4.2.1, the format of the activity data recorded from the different activity tracking devices was as presented in Table 4.1. The daily step-count numbers were already in a format appropriate to be utilized in ML algorithms and thus no further preprocessing was needed. Furthermore, it was the most important piece of information, as besides a feature for previous days in a timeseries sequence, it is also the variable that we are trying to predict for a day in the future. Deferentially, the date alone does not provide much information in order to be considered as a meaningful feature in its raw format, but the meta-information derived from a date may be used effectively towards this end. By utilizing the implemented functionality in Python’s datetime module [50] and holidays package [41] the following features have been extracted for each date in the dataset:

- **IS_HOLIDAY**: A feature that had the value "TRUE" if a date is an official holiday in Greece and "FALSE" otherwise.
- **IS_WEEKEND**: Whether a day is either Saturday or Sunday or neither of them.
- **DAY_OF_WEEK**: The day of the week presented as a number with 0 being Monday and 6 being Sunday.

- **DAY_OFMONTH**: A number ranging from 1 to 31 representing the place of a date in a month.

- **MONTH**: The month of a date as a number ranging from 1 to 12.

After applying the feature engineering process described above to the data gathered from the participants’ activity tracking devices, the dataset had the format presented in Table 4.2 below:

<table>
<thead>
<tr>
<th>DATE</th>
<th>STEPS</th>
<th>IS_HOLIDAY</th>
<th>DAY_OF_WEEK</th>
<th>IS_WEEKEND</th>
<th>MONTH</th>
<th>DAY_OF_MONTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018-08-12</td>
<td>840</td>
<td>False</td>
<td>6</td>
<td>True</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>2018-08-13</td>
<td>1105</td>
<td>False</td>
<td>0</td>
<td>False</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>2018-08-14</td>
<td>1766</td>
<td>False</td>
<td>1</td>
<td>False</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>2018-08-15</td>
<td>867</td>
<td>True</td>
<td>2</td>
<td>False</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>2018-08-16</td>
<td>2206</td>
<td>False</td>
<td>3</td>
<td>False</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

### 4.2.3.2 Features Related to Covid-19

Due to the fact that this research is being conducted during the global pandemic of COVID-19 that has been first identified in December of 2019 in China [44] and it is still ongoing at the time of this writing, we decided that relevant features should be added in the dataset as they might bear meaningful information about the PA patterns of the participants on specific days of recorded activity.

In response to the COVID-19 outbreak, governments around the world have been taking a wide range of measures. Many of these measures such as restrictions of traveling, closing of certain types of businesses, restrictions around movement after a certain hour and more, are restrictive regarding the PA of citizens that are affected by them. In our research all of the participants are in Greece whose government has decided in applying various such measures for an extended time period.

The recorded time periods for the different participants in our research are ranging from the first months of 2015 up to the last months of 2020. The majority of them had to alter their typical behaviour regarding their daily step-counts due to the facts presented above about the restrictive measures of Greek government’s response to the COVID-19 global pandemic. In order to take that fact into account, we utilized the dataset in [12] that includes information...
on several different common policy responses that governments of more than 180 different countries have taken to respond to the pandemic from 01-01-2020 to 17-02-2021.

The above mentioned dataset contains 19 different indicators regarding policy responses which can be organized into 4 different groups:

- Containment and closure policies
- Economic policies
- Health system policies
- Miscellaneous policies

The groups and the indicators of the dataset are described in detail in its documentation page that can be found in [13]. From the above groups we were only interested in the group of containment and closure policies as they have a greater direct impact on the PA of an individual. We also kept only the records about Greece and discarded features considered irrelevant, such as country code and region name that either did not apply to the Greek state or were the same for every record. The containment and closure policies group contains 8 indicators and 7 flags that give further information about each of the indicators. The full set of COVID-19 related features is given in Appendix B.

### 4.2.3.3 Personality & Identity Features

Lastly, besides activity, date and COVID-19 features presented in 4.2.3.1 and 4.2.3.2 respectively, we included personality and identity features in our dataset, derived from appropriate questionnaires which the research’s participants were required to submit. The questionnaires used can be found in Appendix A and the process of collecting these information is described in 4.1.2.

As stated in 4.1.2, 3 different questionnaires were used in order to gather information about multiple aspects of a participant’s personality and social identity. Each one of these questionnaires after applying appropriate preprocessing steps contributed to the final dataset with a particular set of features. In the following paragraphs the preprocessing process and the extracted features for each questionnaire will be presented in detail.

**Demographics**

By using the demographic questionnaire of Appendix A.1, we extracted 5 features describing the social identity of the research’s participants. These features were:
- Gender
- Age
- Family status
- Educational level
- Career status

All of the above features with the exception of age were categorical, meaning that they could only have a value belonging to a predefined set of values. In order to handle the exception of the age feature we organized the different ages submitted into 4 age groups: 18-24, 25-34, 35-44 and 45-54. Next, depending on their age, we assigned each participant to the corresponding age group. After the above process we applied a label encoding process for each of these features in order to replace the string values with numerical that can be interpreted better by ML algorithms. A snippet of the demographic features and their values for five random participants is presented in Table 4.3 below.

<table>
<thead>
<tr>
<th>GENDER</th>
<th>AGE</th>
<th>FAMILY STATUS</th>
<th>EDUCATION LEVEL</th>
<th>CAREER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

**Big-Five**

The concept of Big-Five questionnaire is a quite established approach in the research field of human personality. The main idea as has been defined firstly in [60] and later in [61], is that there are 5 relatively strong and recurrent personality factors that are quite common amongst people, and can be used for a rough categorization of human personalities. The Big-Five questionnaire [7] and its Greek version [62] used in our work, uses 50 questions towards this end. These questions are designed in such a manner that each of them aims to identify certain personality traits that correspond to one of the five personality factors. The recurrent personality factors and their corresponding personality traits that define them are as follows [61]:

- **Extroversion**: Silent vs. Talkative; Secretive vs. Frank; Cautious vs. Adventurous; Submissive vs. Assertive; Languid, Slow vs. Energetic and Self Contained vs. Sociable.
- **Agreeableness**: Spiteful vs. Good-Natured; Obstructive vs. Cooperative; Suspicious vs. Trustful; Rigid vs. Adaptable; Cool, Aloof vs. Attentive to People; Jealous vs. Not So; Demanding vs. Emotionally Mature; Self-Willed vs. Mild; and Hard, Stern vs. Kindly.

- **Conscientiousness**: Frivolous vs. Responsible; Unscrupulous vs. Conscientious; Relaxed, Indolent vs. Insistently Orderly; Quitting vs. Persevering; and Unconventional vs. Conventional.

- **Emotional stability**: Worrying, Anxious, vs. Placid; Easily Upset vs. Poised, Tough; Changeable vs. Emotionally Stable; Neurotic vs. Not So; Hypochondriacal vs. Not So; and Emotional vs. Calm.

- **Intellect/Imagination**: Boorish vs. Intellectual, Cultured; Clumsy, Awkward vs. Polished; Immature vs. Independent-Minded; Lacking Artistic Feelings vs. Esthetically Fastidious; and, Practical, Logical vs. Imaginative.

Each question is answered on a scale of 1 to 5 depending on how representative it is to the user, with 1 being "very inaccurate" and 5 being "very accurate". The score of each question is either added or subtracted from the total score of one of the 5 recurrent personality factors, depending on how the corresponding personality traits are being handled in the question. With the submission of all 50 questions the final score for each factor for a participant is calculated [21]. These scores of the 5 personality factors of the Big-Five questionnaire were used as features in this work. In Table 4.4 below, a snippet of the features extracted from the Big-Five questionnaire for 5 random participants is presented.

<table>
<thead>
<tr>
<th>EXTROVERSION</th>
<th>AGREEABLENESS</th>
<th>CONSCIENTIOUSNESS</th>
<th>EMOTIONAL STABILITY</th>
<th>INTELLECT/IMAGINATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>-12</td>
<td>19</td>
<td>15</td>
<td>-11</td>
<td>8</td>
</tr>
<tr>
<td>-2</td>
<td>24</td>
<td>1</td>
<td>-20</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>16</td>
<td>-27</td>
<td>25</td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>7</td>
<td>-12</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>23</td>
<td>13</td>
<td>-27</td>
<td>27</td>
</tr>
</tbody>
</table>

**Transtheoretical Model of Behavior Change**

The transtheoretical model of behavior change is a model that attempts to conceptualize the process of change in an individual’s behavior. It includes key elements from other related theories (hence, the name "transtheoretical") in order to be able to be applied to a variety of populations and settings [49].
The main idea behind the TTM model is that people move through a series of stages when modifying their behavior. The time spent in each stage may vary across individuals but the stages remain the same. Moreover, there are certain processes that need to be implemented in order to progress through the stages of change. These 10 processes can be organized in 2 groups as follows [49]:

■ **Cognitive and Affective Experiential Processes**

1. **Consciousness Raising**: Includes increased awareness of the causes, consequences, and treatments for a particular problematic behavior. Awareness-raising interventions include feedback, education, debate, interpretation, book therapy, and media campaigns.

2. **Dramatic Relief**: Initially produces increased emotional experiences followed by reduced affect, if appropriate action can be taken. Psychodrama, role playing, mourning, personal testimonies and media campaigns are examples of techniques that can move people emotionally.

3. **Environmental Reevaluation**: Combines both emotional and cognitive assessments of how the presence or absence of a personal habit affects a person’s social environment, such as the effect of smoking on others. It can also involve realizing that one can serve as a positive or negative role model for others. Empathy training, documentaries and family interventions can lead to such reassessments.

4. **Self-Reevaluation**: Combines both cognitive and emotional assessments of a person’s image with and without a particular unhealthy habit, such as the image of someone as an inactive or active person. Clarifying values, healthy role models and images are techniques that can move people effectively.

5. **Social Liberation**: Requires an increase in social opportunities or alternatives, especially for those who are relatively deprived or oppressed. Advocacy, empowerment processes and appropriate policies can produce increased opportunities for promoting minority health, promoting gay health and promoting health for poor people. These same procedures can also be used to help all people change, such as smoke-free zones, salads at school meals, and easy access to condoms and other contraceptives.

■ **Behavioral Processes**

1. **Self-Liberation**: Is both the belief that one can change and the commitment to act on that belief. New Year’s resolutions, public testimonies and multiple rather than simple choices can enhance self-liberation or what is commonly called willpower. Motivation research shows that people with two choices have more commitment
than people with one choice. Those with three available options have even greater commitment. Having four options though, does not further enhance the individual’s willpower. So with smokers, for example, three options that could be given are cold turkey, nicotine fading and nicotine replacement.

2. **Counter Conditioning**: Requires learning healthier behaviours that can replace problematic ones. Relaxation can compensate for stress. Having the courage to express and stand for your opinion can offset peer pressure. Nicotine replacement can replace cigarettes and low fat foods can be chosen instead of an unhealthy diet.

3. **Helping Relationships**: Combines caring, trust, ease of expression and acceptance, as well as support for healthy behavior change. Building relationships, having allies and reaching for appropriate counseling can be sources of social support.

4. **Reinforcement Management**: Provides the consequences of taking steps towards a certain direction. While the use of punishments may be included, it has been found that people who successfully change their behavior rely on rewards far more than punishments. So, the positive reinforcements are emphasized, as the philosophy of this model is to work in accordance to the way people change naturally.

5. **Stimulus Control**: Removes stimuli that lead to unhealthy habits and adds prompts for healthier alternatives. Avoiding unwanted habits, redesigning the environment and self-help groups can provide stimuli that support the desired change and reduce the risk of relapse. The design of parking lots that are a 2-minute walk from a company’s establishments and the placement of artworks in stairwells are examples of environmental redesign that aims to encourage exercise.

In our research, we used a Greek questionnaire [1] that applies TTM to exercise-related behavior change, based on the work in [16, 17]. The above mentioned questionnaire can be found in A.3. It contains 1 multiple-choice question regarding the attitude one has towards exercise and 39 questions scaled from 1 to 5 in which the participant has to state the degree each statement is representative of their behavior, with 1 being "very inaccurate" and 5 being "very accurate".

After label-encoding of the 6 available answers, the multiple-choice question was ready to be used as feature. Regarding the 39 scaled questions further preprocessing was needed in order to transform the raw information into features appropriate for ML. The questionnaire we used is designed in such a way that each of the 39 questions corresponds to 1 of the 10 processes of change presented above. The processes of change and their corresponding questions are presented in Table 4.5 below:
Each of these questions has a value ranging from 1 to 5, depending on the degree of relevance to the participant’s behavior. The mean value of each group of questions was assigned as the value of the corresponding process of change for each individual. Amongst with the standard deviation of each process’s value we extracted 20 features to be used in our ML models, thus having in total 21 features from TTM when we add the feature created from the multiple-choice question.

In this subsection we have described in detail our feature engineering process and the various features we have extracted from raw information. The goal of this process was the creation of a dataset that includes a combination of activity, personal and contextual features. The lack of availability of similar datasets has been identified as an open issue (I2) of the relevant researches as stated in Chapter 1.

Some of the features like those extracted from date metadata describe a specific day of recorded activity where others, such as demographics and personality features are participant-oriented. Step-count on the other hand, which is also the target variable that we aim to predict for future days, is a feature whose value depends on both the date’s and the participant’s features. After the completion of the feature engineering process, each day of recorded activity of each participant was being described by 53 features.

In the following subsection, we describe the process by which these features, separated for each participant at this phase, will be transformed in a format appropriate for time-series analysis and usage in ML prediction models.
4.2.4 Data Transformation

After the feature engineering process described above in subsection 4.2.3, for each research’s participant we have obtained a dataset of the following format:

<table>
<thead>
<tr>
<th>DATE</th>
<th>FEATURE_1</th>
<th>FEATURE_2</th>
<th>…</th>
<th>FEATURE_52</th>
<th>FEATURE_53</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date_1</td>
<td>value_1_1</td>
<td>value_2_1</td>
<td>…</td>
<td>value_52_1</td>
<td>value_53_1</td>
</tr>
<tr>
<td>Date_2</td>
<td>value_1_2</td>
<td>value_2_2</td>
<td>…</td>
<td>value_52_2</td>
<td>value_53_2</td>
</tr>
<tr>
<td>Date_3</td>
<td>value_1_3</td>
<td>value_2_3</td>
<td>…</td>
<td>value_52_3</td>
<td>value_53_3</td>
</tr>
<tr>
<td>Date_4</td>
<td>value_1_4</td>
<td>value_2_4</td>
<td>…</td>
<td>value_52_4</td>
<td>value_53_4</td>
</tr>
<tr>
<td>Date_5</td>
<td>value_1_5</td>
<td>value_2_5</td>
<td>…</td>
<td>value_52_5</td>
<td>value_53_5</td>
</tr>
</tbody>
</table>

Although we have obtained features appropriate for ML, the datasets in their current format are not usable by prediction models due to the fact that the data are not presented in a time-series format in order to be used by ML algorithms for forecasting. In addition to that, at this stage there are 21 datasets one for each research’s participant, an issue that needs to be addressed for the development of the generalized prediction model (Figure 4.1). In this subsection, the process by which the different individual datasets of the 21 research’s participants will be transformed, in order to be combined in one aggregated dataset appropriate for usage in ML algorithms and time-series prediction will be described step by step.

Non-consecutive Day Handling

As stated in subsection 4.2.2 in which we described the data cleaning process, we decided to remove from the dataset, days with a recorded daily step-count less than 500 steps, as no-wear days. Moreover, as seen on the same subsection, Apple devices do not include at all days with zero number of steps. These two factors have resulted in the creation of gaps in the recorded time periods.

In order to apply time-series analysis by using supervised ML algorithms we had to obtain a dataset in the format of (input variables, target variable) pairs. The input variables in our case will be the 53 features of each of the \( n \) days in the past, amongst with all the features of day \( n+1 \) except the step-count of that day. The target variable would be the step-count of day \( n+1 \).

It is obvious from the above that in order to ensure the robustness of our results, the sequence of \( n \) days in the past used as input, should be consecutive in order for patterns in activity to be identified effectively. In order for this to be feasible, gaps in the recorded time periods should be handled appropriately. At this point it should be clarified that we are referring to the number of days used as input as \( n \), due to the fact that different values of \( n \) were tested in order to find the optimal value that leads to the most accurate predictions (see model development in Figure...
4.1). The methodology adopted in finding the optimal window size is discussed in detail in 4.6 and the relevant results are presented in 5.5.

In order to handle the issue of gaps in the sequence of recorded days of activity, each participant’s individual dataset was parsed separately day by day and was divided into sub-datasets which included only consecutive days. The number of days in these sub-datasets was ranging from 1 day to a few months of activity based on the length of the consecutive-day sequences of the original individual datasets. The sub-datasets were written to separate files in order to be parsed later in the next phases of the transformation process. From the 21 original individual datasets of the participants, after this step, 645 new sub-datasets with only consecutive days were created. Sub-datasets with days less or equal to \( n \) could not be used, as it was not possible to form a valid input-output pair which requires at least \( n + 1 \) days of activity. More information regarding this matter is provided in the following paragraph.

**Transformation in Time-Series Format**

As stated above, before ML can be used in our approach, the time-series forecasting problem must be re-framed as supervised learning problem. The information has to be transformed from a sequence of recordings, to pairs of input and output sequences. In order for this to be achieved, each dataset of consecutive days of activity created by the process described in 4.2.4 was parsed separately and was transformed from its original form, to the input-output pairs format that is presented in Table 4.7 below.

**Table 4.7:** General format of input-output data after transformation in time-series format

<table>
<thead>
<tr>
<th>INPUT</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features of ( n ) days of activity + all features of day ( n+1 ) except the step-count</td>
<td>Step-count of day ( n+1 )</td>
</tr>
</tbody>
</table>

The “window” of \( n \) days iterated through all the consecutive days of the each dataset, moving forward by one day in each iteration. In order to clarify this statement, a blueprint of the datasets occurred after this process is presented in Table 4.8.

**Table 4.8:** Blueprint of datasets after transformation in time-series format

<table>
<thead>
<tr>
<th>INPUTS</th>
<th>ALL FEATURES</th>
<th>ALL FEATURES EXCEPT STEP COUNT</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td>( D_2 )</td>
<td>...</td>
<td>( D_n )</td>
</tr>
<tr>
<td>( D_2 )</td>
<td>( D_3 )</td>
<td>...</td>
<td>( D_{n+1} )</td>
</tr>
<tr>
<td>( D_3 )</td>
<td>( D_4 )</td>
<td>...</td>
<td>( D_{n+2} )</td>
</tr>
<tr>
<td>( D_3 )</td>
<td>( D_4 )</td>
<td>...</td>
<td>( D_{n+3} )</td>
</tr>
<tr>
<td>( D_4 )</td>
<td>( D_5 )</td>
<td>...</td>
<td>( D_{n+4} )</td>
</tr>
</tbody>
</table>
Since we have already handled missing data and gaps in the recorded activity periods in 4.2.2 and 4.2.4 respectively, the only missing (NaN) values in the dataset are those that have occurred after the time-series transformation. For example if an individual dataset contained less than \( n \) days (it is reminded that after the non-consecutive handling process described in 4.2.4 even 1-day datasets were created), an input sequence containing NaN values would be created as the remaining days of the sequence up to \( n + 1 \) would be filled with NaN. The same would also happen for the last days of a dataset were there would be less than \( n \) remaining days in order for a sequence to be created. These NaN-containing rows have been discarded from each individual dataset, leaving us with completed time-series sequences of input-output data appropriate for usage in supervised learning techniques of ML. Moreover, due to the methodology adopted, it is certain that each row contains sequences of consecutive days without any gaps or missing data.

**Creating the Final Dataset**

At this stage of our methodology we have already obtained a number of input-output sequences, ready to be used in supervised ML approaches. In order though to create the final format of the personalized and the generalized datasets and implement our C3 contribution, a final step in this process is needed.

For the creation of the generalized dataset, which includes data from all of the 21 research’s participants, we concatenated all of the individual datasets into one. As stated above in 4.2.4, each row of every individual dataset is a sequence of consecutive days with features of both the days of the sequence and the identity of the particular user who provided these recordings. There is no individual dataset that contains information about more than one participants and thus, there is no way that data from more than one participants would be mixed in the same sequence during this concatenation, something that would distort the information reflected in the dataset.

In order to create the personalized dataset that will be used to develop a model trained only on the data of a specific user, we repeated the same concatenation process described above. In this approach, instead of using every individual dataset that has occurred, we used only data originated to a single participant. The selection of the participant that was used for the creation of the personalized dataset was based on the number of available days of recorded activity.

In this section we discussed a very important part of our methodology, the data preprocessing phase. We presented the integration framework we developed in order to use data originated to more than one activity tracking devices, a contribution (C1) to an open issue (I1) considering relevant researches. Next, we described the data cleaning process we followed, in order to ensure the quality of the dataset we create and enhance the efficiency of the ML models that will be developed based on this dataset. Following data cleaning, we discussed the feature
engineering process by which we converted the raw information submitted by the research’s participants into features appropriate for usage in ML algorithms. In this part, we elaborated on the manner in which we combined activity, personal and contextual features, thus implementing our C2 contribution. Finally, in the last subsection, we presented the data transformation process designed and implemented in order to ensure that our dataset contains time-series data in such a format that their usage by supervised ML approaches is feasible.

In the next section, we describe the last step in the "dataset creation" part of our methodology (Figure 4.1), the process in which we select the features that will contribute to the predictive ability of a ML model and thus, should be included in the personalized and the generalized datasets. This process is a first approach in identifying redundant features that do not bear information related to an individual’s daily step-count.

### 4.3 Initial Feature Selection

Through the data collection and data preprocessing steps of our methodology, we have obtained a set of different features regarding personal, contextual and activity factors. These features were inspired by similar works in relevant literature [38] and also by common sense, considering aspects that could possibly affect an individual’s daily step-count. However, it cannot be ruled out that there may be specific features that are redundant and have nothing to contribute in improving the predictive ability of the ML models. Such features, if any, should not be included in the final dataset, as they may even have a negative impact on the performance of the models.

At this stage, the way chosen in order for us to evaluate the contribution of the different sets of features to the development of efficient ML models is through an appropriate experimentation approach. The purpose of this section is to describe this process and give information about the assessment of the importance of the different feature categories.

As discussed in section 4.2.3, the features extracted through the feature engineering process can be organized into 4 categories:

- **Activity-related Features**: Meaning the actual recorded daily step-count.
- **Time-related Features**: Features deriving from a date’s metadata.
- **Personality-related Features**: Features related to the participant’s identity extracted from relevant questionnaires.
- **COVID-19-related Features**: Features that describe the movement-restrictive measures taken by the Greek government in order to deal with the global pandemic of COVID-19.

A schematic representation of the above mentioned categorization can be found is presented in Figure 4.9 below:

![Schematic representation of the different feature categories](image)

**Figure 4.9**: Schematic representation of the different feature categories

Our approach in this initial feature selection process (feature selection takes also place in the "model development" part of our methodology 4.5) was to establish the value of each of the 4 categories by training ML models with different subsets of features. The following 3 different datasets were created, containing:

- Activity and date features.
- Activity, date and personality features.
- Activity, date, personality and COVID-19 features.

4 ML regression algorithms were fitted to each dataset and the best performing models were compared in order to decide if the contribution of each set of features has a positive impact on the efficiency of the predictive models. Details on the algorithm selection and optimization process are given in section 4.4. The results of this approach indicated that the gradual addition of features kept improving the performance of the models and thus, all of the feature sets used, had provided meaningful insight that improved the predictive ability of the ML algorithms.
This conclusion, justifies our hypothesis that such a combination of features would contribute (C2) to the field of relevant research. Details on the methodology of evaluation of the different models can be found in section 5.1 and the given results are presented and discussed in detail in Chapter 5.2.

In this section we have described the last phase of the part of our methodology that concerns the creation of the dataset. A dataset that meets the specifications we set out in Chapter 1 regarding the contribution that this research seeks to make. We described the approach taken in order to select the feature categories to be included in the final dataset and also presented the experimental process, which proved that our C2 contribution was valid, as it was found that all feature categories contribute positively to the predictive ability of the ML models.

Up to this point, through the previous sections of this chapter, we have discussed the methodology in which an appropriate dataset regarding PA of individuals has been created. We also elaborated on the manner in which our contribution points C1, C2 and C3 as presented in Chapter 1 have been implemented. The design of the integration framework amongst the 2 most popular brands (Apple & Xiaomi) in the market of wearable activity trackers, described in 4.2.1 implemented C1, as it can be used in future works in order to enable researches to utilize a greater number of activity data originated to different sources. C2, was implemented by the incorporation in the dataset of a number of activity-related, personal and contextual features, features that have been proven in this work to contribute positively in the efficiency of ML models. Finally, as discussed in 4.1, we implemented C3 through the usage of ”in-the-wild” data, an approach that is rarely adopted in relevant literature 3.2.

The following sections of this chapter, present the part of our methodology that aims to the development of efficient ML models which are able to predict a user’s future daily step-count, by using the novel dataset created in this research. In the next section, we present the first phase of the model development process, the process of selecting the best performing ML regression algorithm and optimizing its performance through hyper-parameter tuning.

### 4.4 ML Algorithm Selection & Optimization

In this section, the methodology adopted in selecting the ML algorithm that will be used in the development of the prediction models, is presented. In addition to that, we discuss in detail the concept of hyper-parameter tuning, a necessary step in order to obtain the best possible performance of the ML models.
4.4.1 Algorithm Selection

Since the target variable of the prediction models is a continuous integer value, the task at hand is a regression problem and thus, algorithms with the ability to perform this kind of predictions were considered in our research. Furthermore, as seen in 4.2.4, the dataset is structured in a format designed for usage with supervised ML approaches, meaning approaches that require data representations in an input-output pair format. Due to these facts, the algorithms used in our work are ML regression algorithms for supervised ML.

The first experiment of our research that involved evaluation of ML models, was the experiment presented in 4.3 in which the feature categories that would be included in the dataset were selected. In this experiment we tested the performance of the following 4 ML regression algorithms:

- Ridge Regression.
- Decision Tree.
- Random Forest.
- Gradient Boosting Regressor.

Further discussion about the theoretical background of the above algorithms is out of the scope of this writing, however relevant information can be found in [47], which implements their functionality in Python, implementation that was used in our research.

In the above experiment, the GBR has consistently outperformed each of the other algorithms and the relevant results are presented in 5.2. Due to the fact that GBR was the best performing algorithm in all the different feature sets used (see 4.3) regardless of the chosen window size, it was selected as the algorithm to be used in the next phases of our experimentation process.

In the following subsection we present the approach adopted in defining the optimal configuration of the prediction algorithms. This is a vital step in order to obtain the best possible performance of the prediction models. The algorithm selection process described above, has been applied on the best configuration of the different algorithms in order to compare their optimal performances.

4.4.2 Hyper-parameter Tuning

Every ML algorithm has certain hyper-parameters, whose values have to be defined by the user and affect the overall performance and efficiency of the models. Hyper-parameters should not
be confused with parameters which are the internal coefficients set by training the model on a dataset and are learned automatically. Hyper-parameters have no optimal default values appropriate for every case and the values that lead to the best results depend on many factors such as the amount of available data, the specific characteristics of the dataset and more. Although there are general heuristics and rules of thumb, the best way for a researcher to configure the hyper-parameter values to be used on a specific problem is through trial and error, a process particularly expensive in both computational and time assets, especially if there is a number of ML algorithms to be tested on a large dataset.

In this work, as already stated, 4 different ML algorithms were tested throughout the whole model development and experimentation process. These algorithms are Ridge Regression, Regression Decision Tree, Regression Random Forest and Gradient Tree Boosting Regression. In order to conduct the hyper-parameter tuning process, meaning the selection of the hyper-parameter set that leads to the best performance for each algorithm, we used a method implemented in scikit-learn [47], the GridSearchCV. In GridSearchCV, a search space as a grid of hyper-parameter values is defined and each position of the grid is evaluated through CV. This means, that given a list of values for each hyper-parameter, all possible combinations of these values are evaluated through CV and the best performing model is retained. By applying this method on all of the ML algorithms tested, we obtained the best performing hyper-parameter configuration for each model. These configurations were used in order to obtain the experimental results presented in Chapter 5.

In this section, we presented our approach towards selecting the best performing algorithm for the prediction models of this research. We also discussed the manner in which the optimal configurations of the algorithms tested throughout the experimentation process were obtained. This step is necessary in the development process of an efficient prediction model that will implement the contribution C4 of this research. In the next section, the methodology adopted in reducing the dimensionality of the original dataset, created in previous steps of our process, is described in detail.

### 4.5 Dimensionality Reduction

After the completion of the preprocessing phase, each day of recorded activity in the dataset is described by 53 different features. Based on the blueprint of Table 4.8, that means that for \( n = 5 \) (5 days of activity as input in order to predict the daily step-count of the 6th day), the total number of features for each sequence of input data would be 317 features, for \( n = 7 \) the number of features would be 423 and so on. It becomes obvious that the task at hand is characterized by high dimensionality, meaning that the number of features to be used in ML is relatively high. As seen in literature [25], high dimensionality may lead to increased required
training time for the ML models and even reduced accuracy of the obtained predictions. Based on the above, it was decided that it would be appropriate to test if reducing the number of features would lead to the improvement of the performance of the ML models developed.

In order for the dimensions of the dataset to be reduced, 2 different categories of such methodologies were chosen for testing, feature selection and dimensionality reduction through dataset projection into less dimensions. These categories will be further discussed in the following subsections, amongst with the approach adopted in implementing them for the sake of this research.

4.5.1 Feature Selection

The feature selection approach includes the procedures related to the selection of high importance features which contribute to the greatest extent to ML models’ predicting ability. In this manner, features which do not bear much information can be discarded from the dataset and thus, reducing the required training time and even possibly improve the accuracy of a model’s predictions.

Our approach towards the task of feature selection was to apply a methodology called recursive feature elimination (RFE) [11] as implemented in Python by scikit-learn [47]. RFE is a feature selection method that uses a ML algorithm in its core to help in selecting the most important features. It works by starting with all the features of the dataset and gradually removing features until the desired number remains. This is achieved by fitting an appropriate core ML algorithm to the training dataset, ranking the features by their importance, discarding the least important features and re-fitting the model to the new reduced dataset. This process is repeated until the desired number of features remains. Figure 4.10 below describes the above process in a pseudo-code format.

The core algorithm has to be a ML algorithm, capable of assigning an importance score to each feature. Such algorithms are linear algorithms which assign weights to the various features and tree-based models (Decision Tree, Random Forest etc.).

In this work, we used a specialized implementation of RFE, the RFECV of scikit-learn [47]. RFECV, can be used in order to automatically select the best number of features chosen by RFE, instead of requiring from the user to define this number. This is achieved by evaluating the performance of different numbers of features through CV and automatically selecting the number of features that resulted in the best mean score. The user has to define the minimum number of features, the number of least important features to be eliminated in each iteration and the type of CV to be used for the evaluation of the performance. For the sake of this project, the Gradient Boosting Regressor ML algorithm was used as core model in RFECV as it has been
found through experimentation in our research that it has the best performance compared to other models. The relevant results of the RFE approach can be found in Chapter 5.4.1.

### 4.5.2 Dimensionality Reduction through Dataset Projection

Another approach towards reducing the size of the input data is to transform the initial dataset into one that has less dimensions (features), an approach tested in this work through the popular technique of Principal Component Analysis, or PCA [59] as implemented in [47]. Although RFE also reduces the dimensions of the input data, it achieves this by completely eliminating the least important features from the dataset as seen in 4.5.1. PCA on the other hand, accomplishes the task by applying transformations from the field of linear algebra to the data, making the resulting dataset a projection of the original one. This means, that the original features do not exist anymore as they are projected in such a manner that new features are created while maintaining all the original important relationships. When PCA is applied in the development of a model, it means that any data that are to be used (test set, new data etc) have to be also transformed in the same manner to the one of the original dataset in order for the model to be able to process them. The results of the PCA approach adopted in the dimensionality reduction of our dataset can be found in 5.4.2.

In this section of our methodology chapter, we discussed the different approaches tested in reducing the dimensionality of the original dataset used in our experiments. We described 2 different methodologies, RFE which discards the least important features, thus leading to a dataset of fewer dimensions and the PCA which converts the dataset into one with less
dimensions, with features which are completely different from those of the initial dataset. In the following section, the process of defining the optimal number of days that should be given as input to the ML models in order to achieve the best possible performance, is presented in detail.

### 4.6 Obtaining Optimal Window Size

Our experimentation process is structured on a supervised ML methodology with regression algorithms. In supervised approaches, data should be formatted as input-output pairs. In our case, the input is a sequence of $n$ days and the output is the step-count value of day $n + 1$ as seen in Table 4.8.

In a similar manner to the one of the hyper-parameter tuning process described in 4.4.2, there is no pre-defined optimal value of the parameter $n$ that would lead to the best performance of the prediction models and thus, it is defined in our research, through experimentation with different values of $n$.

Towards this end, 4 different window sizes were tested and the performances of the corresponding prediction models were compared. It has to be noted that as the size of the window used gets bigger, so does the dimensionality of the dataset, while the number of available input-output pairs is reduced. This "side-effect" may lead to poor performance of the prediction models and overfitting to the training data and thus, dictates a limit on the size of the window to be used. Considering the above, the different $n$ values used, were 5, 7, 14 and 20. These values, cover an adequate number of experimental window sizes for the sake of this research, while also maintain the feasibility of efficient training of ML models, as the feasibility of the training process is the strongly related to the amount of available data and the dimensionality of the dataset.

The dimensionality increases due to the fact that more and more days (each originally described by 53 features as we discussed in 4.2.3) are added to each input sequence of the dataset. The decrement of the number of available data points in the dataset, is explained by the process of handling non-consecutive days, described in 4.2.4. For example, the requirement of 20 consecutive days of recorded activity is far more strict than the requirement of 5 consecutive days. Thus, as the window size is increasing the number of sequences appropriate for usage is reducing. In Figure 4.11 below we present the dimensions of the dataset and the available data points in relation with the value of the window size $n$.

The results of this experimental process are presented in 5.5 and indicate that no significant difference in the performance of the ML models has occurred through the usage of the different
window sizes. In order to obtain a more robust conclusion on this matter, it would be appropriate in future works to repeat this experiment with a greater number of available data. Since in our research the performances of models developed on different values of $n$ were similar, the window size of 5 input days was used for further experimentation due to the fact that utilized fewer features while leading to a bigger dataset with a greater number of available data points.

In this section we described the methodology of defining the number of days that should be used as input on the development process of ML prediction models. The impact of the different window sizes on the dataset's format was presented and through this we explained the limitations imposed on this issue by the data structure and availability. Finally, we discussed the results given by this experiment and elaborated on the reasons which led to the usage of value 5 for the $n$ hyper-parameter of our models.

In sections 4.4 to 4.6, we have presented the part of our methodology that is aimed in the development of a ML model which efficiently predicts a user’s daily step-count based on activity-related, personal and contextual features. We discussed different approaches regarding each issue that was explored, each of them giving different results. In this manner, we were able to choose the best performing model, thus implementing our contribution C4 described in Chapter 1. In Figure 4.12 bellow, a schematic representation of the methodology adopted in obtaining the model configuration with the optimal performance is presented. The path in blue represents the approaches and decisions that led to the model with the best predictive ability.

In this chapter we presented and discussed in depth, our adopted methodology towards firstly, creating a novel dataset under the specifications set in Chapter 1 where we stated our contributions and secondly developing a ML model capable of efficiently forecasting a user’s future daily step-count. We began by describing our data collection process, the challenges met and our approaches towards overcoming them. In this section we elaborated on the design of our
integration framework, which is also part of this research’s contribution (C1) and the process of gathering “in-the-wild” data from the research’s participants (C3). Next, in section 4.2, we discussed in depth our methodology and experimentation with different approaches in data preprocessing, a phase vital for the creation of a dataset that implements C2 by including a number of novel activity, personal and contextual features. In section 4.3, we presented the last stage of the dataset creation process, in which we proved that all of the above feature categories can contribute positively to the predictive ability of a ML model. Finally, in sections 4.4 to 4.6, we described the process of developing an efficient ML model for step-count forecasting (C4) and discussed the different approaches tested and evaluated towards this end.

In the next chapter, we will present the adopted evaluation approach, amongst with the results of this research and the impact that the different configurations had to the models’ performances. All of the approaches and techniques presented in Chapter 4, are assessed through the evaluation methodology described in 5.1 and the results of this experimental process are presented collectively.
Chapter 5

Experimentation & Results

In Chapter 4, we presented in detail our methodology towards creating a novel dataset for step-count forecasting and the development of an efficient ML prediction model. Throughout this process, multiple regression algorithms were used and a number of different approaches were applied at the different stages of the research. These approaches, as well as the models that emerged through the various combinations of methods, had different performances in terms of their ability to predict a user’s daily number of steps. The purpose of this chapter is to present the results regarding the accuracy of prediction of the above mentioned models and compare the different approaches to each other. Through the results presented here, we provide proof-of-work and justify our claim of implementing our contribution points, in an objective and quantified manner.

The first section of this chapter, 5.1, is dedicated in our evaluation methodology that was used in the different steps of our approach, in order to compare models and techniques, assessing their predicting ability and obtaining the best performing model that implements our C4 contribution in the most efficient manner. Next, in section 5.2, we present the results of the feature categories selection process described in 4.3, which we adopted in order to decide for each group of features (following the categorization given in 4.3) if they had a positive impact on the predictive ability of the regression algorithms and thus, if they should be included on the final dataset to be used by the ML models. In section 5.3, the results of the 2 outlier handling approaches described in 4.2.2 are presented. We discuss the impact of the manual approach, compare it to the one of the Isolation Forest and elaborate on the decisions taken based on this results. Later, in section 5.4, we quote the results of the 2 different approaches implemented for reducing the number of features in the dataset, RFE and PCA. Again, we compare the 2 approaches based on the given results and explain the manner in which they affected the following steps in our methodology. In the next section, section 5.5, the results of the experimental process of identifying the optimal number of days to be used as input, are presented.
and discussed. Finally, in section 5.6, we compare the performance of a model trained in the data of all research participants, as opposed to the performance of a personalized model trained with data from only one specific user.

5.1 Evaluation

For the sake of this project, a number of different ML models have been developed in order to assess the impact of the various approaches adopted. Different algorithms have been tested, each with each own hyper-parameters whose optimal values had to be defined though experimentation. Also, different versions of the dataset were used regarding their dimensionality (original dataset, RFE, PCA) as well as the number of data points included (dataset before and after outlier removal, different window sizes). All of these approaches amongst with their combinations, resulted in different models that had to be evaluated in order to identify the optimal configuration that would provide the most accurate predictions as discussed in the previous section. In the following subsections, we will discuss the metrics used in the evaluation process as well as our methodology regarding assessing the performance of the models on the dataset.

5.1.1 Metrics

When presenting our evaluation methodology, the first factor that should be discussed is the metrics used in assessing the accuracy of the predictions of the different models. Given the fact that the task at hand is a regression problem, we decided to base our evaluation on two different metrics, Mean Absolute Error (MAE) that has been proven to be effective in assessing the average performance of a regression model [64] and Mean Absolute Percentage Error (MAPE), a metric that calculates the average error in predictions as an absolute percentage of the real number of steps. To our best knowledge no such metric has been used in the relevant literature in order to evaluate the performance of daily step-count predicting models. As an example of the importance of MAPE in the evaluation process consider the following:

<table>
<thead>
<tr>
<th>REAL</th>
<th>PREDICTED</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.000</td>
<td>2.500</td>
<td>2.500</td>
<td>50%</td>
</tr>
<tr>
<td>10.000</td>
<td>7.500</td>
<td>2.500</td>
<td>25%</td>
</tr>
</tbody>
</table>

It is obvious that the MAPE provides additional insight about the quality of the predictions that cannot be extracted by the MAE metric alone. Both of the predictions in Table 5.1 have a MAE of 2.500 steps but while in the first example this represents half of the real number of steps
taken during that day, in the second example the error percentage is 25%. Thus, we would say that the prediction of our model is far more accurate for the second example than it is for the first.

As discussed in Chapter 3 considering relevant literature, most works do not report the efficiency of their approaches towards the task of predicting a user’s daily step-count. Even those who do report relevant performance [38], do not report the MAPE or any other similar metric or even the range of the daily step-counts in the dataset in order to obtain a better insight on the predictive ability of their models. Through our evaluation process, we wish to promote an additional approach in assessing and reporting the efficiency of step-count prediction models.

5.1.2 Evaluation on Dataset

In the process of evaluating the performance of the different models, the dataset to be used each time (before or after outlier removal, before or after each method of dimensionality reduction etc.), was separated into a train and a test set. The train set being 90% of the original dataset and the remaining 10% being the test set.

The evaluation process being described was the same in every relevant experimentation process that has been discussed in the above sections. Firstly, the performance of the model was evaluated through a CV process on the training data. The CV process was taking into account both of the 2 metrics discussed in 5.1.1 in order to assess the performance of the algorithm. Following the evaluation on the training data, each model was assessed on the test set in order to test its performance on completely unknown data and ensure the robustness of the evaluation results by excluding the possibility of overfitting. The values of the MAE and MAPE metrics were calculated based on the model’s predictions thus completing the evaluation process.

5.2 Selection of Important Feature-groups

As discussed above, in section 4.3, in order to decide which features should be included in the dataset, we organized them in 4 categories and tested the performance of the chosen regression algorithms in datasets that were containing different subset of these features. Following the evaluation methodology described in 5.1, the results indicated that all feature groups had a positive contribution to the predictive ability of the ML models used and thus, no group as a whole should be excluded from the process. This result supports our C2 contribution point, meaning our decision to include personal and contextual features, in addition to activity-related ones to the created dataset.
In Tables 5.2 and 5.3 below, the results of the 4 ML models for the different datasets are summarized for the MAE and MAPE metrics respectively. These tables, present the models’ performance on training data as assessed by CV. The corresponding performance on the test set is presented later in Table 5.4. Details regarding the evaluation process and the metrics used can be found in 5.1.1.

**Table 5.2: MAE for datasets with different feature-groups (training set)**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Activity &amp; Date</th>
<th>Activity, Date &amp; Personality</th>
<th>Activity, Date &amp; Personality &amp; COVID-19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Regression</td>
<td>2363</td>
<td>2281</td>
<td>2274</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>2412</td>
<td>2301</td>
<td>2290</td>
</tr>
<tr>
<td>Random Forest</td>
<td>2277</td>
<td>2185</td>
<td>2149</td>
</tr>
<tr>
<td>Gradient Boosting Regressor</td>
<td>2277</td>
<td>2174</td>
<td>2138</td>
</tr>
</tbody>
</table>

**Table 5.3: MAPE for datasets with different feature-groups (training set)**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Activity &amp; Date</th>
<th>Activity, Date &amp; Personality</th>
<th>Activity, Date &amp; Personality &amp; COVID-19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Regression</td>
<td>85.05%</td>
<td>80.20%</td>
<td>79.78%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>75.44%</td>
<td>71.24%</td>
<td>70.91%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>73.60%</td>
<td>72.27%</td>
<td>70.62%</td>
</tr>
<tr>
<td>Gradient Boosting Regressor</td>
<td>72.83%</td>
<td>68.16%</td>
<td>66.35%</td>
</tr>
</tbody>
</table>

As stated above, Tables 5.2 and 5.3, present the performance of the ML algorithms as obtained through the 5-Fold CV process we utilized throughout our evaluation process. As described in 5.1, we also tested our algorithms on a separate portion of the dataset used as test-set. The algorithms’ performances on the test-set for the MAE and MAPE metrics is similar to the one of the CV, a fact that indicates that our models have not been overfitted to the training data. The results on the test-set are summarized in Table 5.4 below.

**Table 5.4: MAE for datasets with different feature-groups (test set)**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Activity &amp; Date</th>
<th>Activity, Date &amp; Personality</th>
<th>Activity, Date &amp; Personality &amp; COVID-19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Regression</td>
<td>2429</td>
<td>2281</td>
<td>2378</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>2455</td>
<td>2447</td>
<td>2485</td>
</tr>
<tr>
<td>Random Forest</td>
<td>2353</td>
<td>2276</td>
<td>2239</td>
</tr>
<tr>
<td>Gradient Boosting Regressor</td>
<td>2373</td>
<td>2262</td>
<td>2239</td>
</tr>
</tbody>
</table>
For the best performing dataset, the one that includes all of the activity, personal and contextual features, we present the performance of the different algorithms in the format of a bar chart in Figure 5.1.

![CV MAE & MAPE for the best performing dataset](image)

**Figure 5.1: CV MAE & MAPE for the best performing dataset**

From the charts above, we can conclude that the GBR algorithm outperforms the others, considering both the MAE and the MAPE metrics and this difference in performance becomes more apparent for the MAPE. On the other hand, the Ridge Regression algorithm, while seems to have a MAE similar to the other algorithms, even outperforming Decision Tree, produces a significantly higher MAPE in comparison to the other approaches. This fact, is an additional indication of the importance of a more complete performance evaluation and reporting process than the one currently used in relevant literature, as discussed in Chapter 3. It also, supports our decision to include the MAPE metric in our evaluation process as it contributes in obtaining a more accurate reflection of the predictive ability of a ML model. The ensemble approaches, Random Forest and GBR, seem to perform generally better in this particular dataset and this is another useful conclusion that can be drawn through this experiment.

As stated above, after fitting our models to the training data, we evaluated their performance on a test set that contained data previously unseen by them. In Figure 5.2 bellow, we present the step-count values predicted by the best performing models (those trained on the full-featured dataset), amongst with the actual step-counts of these days.

The above plots, even though they provide a partial reflection of the performance of the models by presenting predictions for 20 days, are indicative of their ability to identify patterns in the user’s activity.

In this section, we presented the results of the initial feature selection process described in 4.3. It can be safely concluded, that the best performing dataset is the one that contains all the feature groups and this conclusion backs our contribution C2 which is the inclusion of personal, contextual and activity-related features to the dataset. The best performing algorithm is the
GBR and in general, ensemble approaches are more efficient on this dataset. Considering this fact, as discussed in Chapter 4, for the next phases of our methodology we will use this section’s performance of the GBR as our baseline and we will experiment with the different approaches described in “model-development” part of our process, in order to further improve it. In the following section we will focus on our adopted methodology, aimed to identify outliers and we will test 2 different approaches of outlier removal in order to see how the model’s performance will be affected.

5.3 Outlier Handling

In this section we will present the results of the 2 outlier handling approaches described in 4.2.2. Outlier handling, as discussed in 4.2.2, is a vital step in the creation of a high-quality dataset, a fact that strongly applies in this research where "in-the-wild" data are included, thus implementing our C3 contribution point. Firstly, we will discuss the impact of the manual
oulier identification and removal approach on the GBR model’s performance and then, the corresponding results of the Isolation Forest approach will be presented.

5.3.1 Manual Outlier Detection

As described in 4.2.2, in the manual outlier detection and removal approach, the days with the higher recorded step-count are removed from the dataset. The number of days to be removed is defined as a threshold value indicating the percentage of the total number of recordings in the dataset, that should be considered outliers. The distribution of outliers and non-outliers in the dataset for different thresholds is presented in Figure 5.3.

![Outlier and non-outlier distribution in dataset for different threshold values](image)

Figure 5.3: Outlier and non-outlier distribution in dataset for different threshold values

After experimentation with different thresholds, the graphs shown in Figure 5.4 have emerged. They present the impact of outlier removal on the MAE and MAPE metrics for the training and the test sets.

![MAE and MAPE values for different thresholds in the training and the test sets](image)

Figure 5.4: MAE and MAPE values for different thresholds in the training and the test sets.
Based on the above plots, we can conclude that there is such a threshold for which the performance of the model is improved while overfitting is avoided and this optimal threshold is 2%. As seen in Figure 5.4a, up to the value of 2%, the MAE of both the training and the test sets is reduced, while for thresholds greater that 2% the MAE of the test set is constantly increasing, a clear sign of overfitting. In Table 5.5 bellow, the scores achieved for different threshold values are presented collectively.

**Table 5.5: Performance for different outlier threshold values**

<table>
<thead>
<tr>
<th>Threshold</th>
<th>TRAINING SET</th>
<th>TEST SET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>1%</td>
<td>2021</td>
<td>65.49%</td>
</tr>
<tr>
<td>2%</td>
<td>1940</td>
<td>64.62%</td>
</tr>
<tr>
<td>5%</td>
<td>1788</td>
<td>62.96%</td>
</tr>
<tr>
<td>10%</td>
<td>1617</td>
<td>61.80%</td>
</tr>
</tbody>
</table>

In this subsection we experimented with a manual outlier handling approach which has been proven to have a positive impact on the predictive ability of our ML model. Next, we will test the Isolation Forest approach which automatically identifies outliers in a given dataset.

### 5.3.2 Outlier Detection with Isolation Forest

In the Isolation Forest approach, the methodology is similar to that of the manual outlier handling. A number of values for the contamination parameter (see 4.2.2), were tested in order to establish which of them leads to the optimal performance and if the approach as a whole can be used efficiently in our work. As stated in 4.2.2, Isolation Forest considers the input data and not the values of the target variable in order to identify outliers. This fact leads to the distribution shown of Figure 5.5.

![Distribution of outliers for threshold values 1% and 5%](image)

**Figure 5.5: Outlier and non-outlier distribution in dataset for different threshold values (ISO)**
The corresponding diagrams for the MAE and MAPE metrics in the training and the test sets are shown in Figure 5.6 below.

*Figure 5.6: MAE and MAPE values for different contamination values in the training and the test sets.*

It is obvious that the given results are not nearly as good as those obtained through the manual approach, with significantly worse values for the MAE and MAPE metrics. Moreover, the model is clearly overfitting even for low values of contamination parameter. In Table 5.6 below, the corresponding results for different contamination values are presented.

*Table 5.6: Performance for different thresholds (ISO)*

<table>
<thead>
<tr>
<th>THRESHOLD</th>
<th>TRAINING SET</th>
<th>TEST SET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>1%</td>
<td>2131</td>
<td>66.20%</td>
</tr>
<tr>
<td>2%</td>
<td>2130</td>
<td>66.49%</td>
</tr>
<tr>
<td>5%</td>
<td>2097</td>
<td>66.06%</td>
</tr>
<tr>
<td>10%</td>
<td>2097</td>
<td>66.55%</td>
</tr>
</tbody>
</table>

In this section we presented the results of the experimental process described in 4.2.2, aimed at the detection and removal of outlier values in the dataset. Such values are not representative of real-life activity patterns and lead to less efficient ML prediction models if not handled appropriately. Through the results of this section, it became clear that our models can benefit from outlier identification and removal techniques and that the most efficient approach is the manual handling of outliers with the threshold of 2%. For the following experiments we will remove the outliers completely from the dataset considering the above experimental results. In the next section we will present the results of the dimensionality reduction approaches tested, in order to establish if we can further improve the performance of our model.
5.4 Dimensionality Reduction

In this section we present the results of the RFE and PCA approaches described in 4.5.1 and 4.5.2 respectively, towards the reduction of the dataset’s dimensionality. In the first part, subsection 5.4.1, we discuss the efficiency of the RFE technique and in subsection 5.4.2 we present the performance of PCA. Finally, we conclude if any of these methods should be applied in our research based on the given results.

5.4.1 Recursive Feature Elimination (RFE)

As described in 4.5.1, the RFE is a feature elimination technique which utilizes a core algorithm in order to assign importances to the various features and through this, eliminate the least important ones. Hence, the first step in the evaluation of such a technique was to train the core algorithm and obtain the optimal number of features. Due to the fact that GBR has been found to outperform every other algorithm tested, while also having the ability to identify the importance of the dataset’s features, we chose to use the GBR algorithm as the core algorithm of the RFE technique. In the plot given in Figure 5.7, we present the MAE as obtained through a 5-Fold CV, for different numbers of features taken as input by the GBR algorithm.

![Figure 5.7: CV MAE for various numbers of features chosen](image)

The graph above has occurred for \( n = 5 \), meaning that 5 days were used as input in order for the step-count of the 6th day to be predicted. As shown in Table 4.8, for \( n = 5 \) the total number of features is 317. Based on 5.7 above, the optimal number of features that led to the best MAE was 247 and thus, the dimensionality of our dataset was reduced by 70 features.
The next step of our evaluation of the RFE technique is to establish if the GBR model used for the predictions (not the GBR model in the core of the RFE) will benefit from applying the RFE approach, compared with its performance it had on the full dataset.

After transforming the initial dataset with the RFE model of the 247 features described above, we tested our GBR model on a training and a test set following the evaluation process described in 5.1. The results of this experiment are shown in Table 5.7 below.

Table 5.7: MAE & MAPE values for the dataset obtained with the RFE approach

<table>
<thead>
<tr>
<th>TRAINING SET</th>
<th>TEST SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>1930</td>
<td>65.10%</td>
</tr>
</tbody>
</table>

Recalling the results of the manual approach of outlier handling shown in Table 5.5, we conclude that we practically achieved the same MAE and MAPE but with a dataset reduced by 70 features, a fact that as stated in 4.5 will result in less required training time for our ML models and can be considered as improvement. The corresponding performances on the test sets are not to be compared due to the fact that the outliers were removed from the dataset in which the RFE approach was tested. In the following subsection we will experiment with a different approach in reducing the dimensions of the dataset, the dimensionality reduction with PCA.

5.4.2 Principal Component Analysis (PCA)

As described in 4.5.2, the PCA approach towards reducing the dimensions of a dataset, differentially from RFE, instead of eliminating unimportant features creates a projection of the original dataset to fewer dimensions through mathematical techniques of linear algebra. In its algorithmic implementation, PCA needs from the user to define the number of components, meaning the dimensions, of the dataset to be created. The optimal number of components depends on many factors such as the nature of the dataset and the relations amongst the different features and has to be decided through an appropriate experimental process.

In order explore the effect of PCA on the GBR model’s performance we adopted the following process. We established a set of different number of components that we decided to experiment with, and for each of these numbers we repeated the same process: through the usage of a pipeline, we firstly transformed the original dataset to one with the required number of components and after that, we evaluated the performance of a GBR model trained on the transformed dataset, through the evaluation process described in 5.1. In Figure 5.8 bellow, we present the corresponding MAE and MAPE results.

Based on the graphs above, we can conclude that the optimal number of components which yielded the best results is 57, with a MAE of 1975 steps. Although the best MAPE is 67.62%
with 22 components, it is practically the same with the MAPE of the 57 components which is 67.65%.

In Table 5.8 below, we compare the best performances of the RFE and the PCA approaches towards the dimensionality reduction of the dataset used.

<table>
<thead>
<tr>
<th></th>
<th>RFE</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>1930</td>
<td>1975</td>
</tr>
<tr>
<td>MAPE</td>
<td>65.10%</td>
<td>67.65%</td>
</tr>
</tbody>
</table>

Since the performance of the 2 approaches is quite similar with RFE slightly outperforming PCA, it is useful to consider the factor of the final number of dimensions of the dataset after each of these approaches. In Figure 5.9 bellow, we present a comparison of the different dimensionalities of the dataset, amongst with the performances of the dimensionality reduction approaches.

The 2 dimensionality reduction approaches achieve roughly the same level of accuracy in predictions with that of the original dataset. Having said that, PCA is the method that leads to the...
worst performance, but on the other hand is using much less dimensions than those of both the original and the RFE datasets. Fewer dimensions, as already mentioned in 4.5, lead to less complex models that are easier to handle and train. Thus, the conclusion to be drawn after our experimentation with dimensionality reduction approaches, is that our model certainly benefits from both RFE and PCA, but the optimal approach depends on the researchers preferences regarding the trade-off between model’s complexity and performance. In this research, we decided to adopt the RFE approach (as seen in Figure 4.12) towards dimensionality reduction, since it yields the best results while also leads to less complex models than those of the original approach, the one prior to any dimensionality reduction.

In the previous sections of this chapter, we presented the results of different approaches regarding feature selection, outlier detection and feature elimination, phases of our adopted methodology that all proved to provide a positive contribution to our model’s predictive ability. Figure 5.10 bellow, shows the improvement of the best model’s (GBR) performance through these steps of our process.

![Graphs showing improvement in performance](image)

**Figure 5.10**: GBR model’s performance on the different phases of the adopted methodology

In the following section, we will present the results of experimentation towards a different direction. We do not seek to improve the model’s performance, but rather identify the optimal value of a hyper-parameter relevant to the dataset on a higher level. We are experimenting with the impact that different numbers of input days has on the model’s performance and present the corresponding results.

### 5.5 Obtaining Optimal Window Size

Following the methodology presented in 4.6, we experimented with different values for the number of days to be given as input to the prediction models, in order for them to forecast the daily step-count of a day in the future.
As discussed in 4.6, there are certain constraints regarding the range of the number of input days, constraints set by the fact that as the window size increments, so does the dimensionality of the dataset while the number of available data points for training is reducing. Based on the above, the numbers of input days chosen for testing were 5, 7, 14 and 20. In Table 5.9 below, we present the number of features and the available data points of the dataset, in relation to each of the above mentioned window sizes.

Table 5.9: Dimensionality and available data of the dataset in relation with the window size value.

<table>
<thead>
<tr>
<th>WINDOW SIZE</th>
<th>DIMENSIONS</th>
<th>DATA POINTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>317</td>
<td>9970</td>
</tr>
<tr>
<td>7</td>
<td>423</td>
<td>9322</td>
</tr>
<tr>
<td>14</td>
<td>794</td>
<td>7582</td>
</tr>
<tr>
<td>20</td>
<td>1112</td>
<td>6472</td>
</tr>
</tbody>
</table>

Considering the above Table and the corresponding schematic representation of Figure 4.11, the limitations on the feasible window sizes become apparent. Adding additional data to the dataset would allow more window sizes to be tested and lead to safer conclusions about their contribution to the performance of ML models. Such an approach is therefore an opportunity for future research.

In Figure 5.11, bellow we present the values of the MAE and the MAPE metrics, obtained through our evaluation process in relation to the different window sizes tested.

![Figure 5.11: MAE and MAPE in relation to the number of days used as input](image)

As already discussed in 4.6 and is now obvious from the above graphs, there is no significant impact of the window size on the performance of the prediction models. This is a fact that would be interesting to be examined in future works, utilizing a bigger dataset, with more data available in order for a safer conclusion on this matter to be drawn. For the sake of this research the window size of 5 days was chosen to be used, due to the fact that it results in the
best MAE, provides the most available data point to be included in the dataset and also leads to the least complex models since they need the smallest number of features in training.

In this section we discussed the manner in which the number of days to be used as input for the prediction models was decided. We presented the limitations on the appropriate numbers of input days to be tested and stated the numbers included in our experimental process. The given results indicated that there was no significant difference in the performance of the models in relation with the number of input days and thus, for efficiency and performance purposes explained above, we decided to use the window size of 5 days. Despite that, since this experiment is strongly depended on the amount of data included in the dataset, experiments regarding the optimal window size are proposed as an issue for future research with a dataset which would contain a greater amount of data.

In the next and final section of this chapter, we will present and compare the performance of a generalized versus a personalized model. The generalized model is a ML model trained on a pooled dataset, containing data from all research’s participants while the personalized model is a model specifically trained on data originated to one participant. While the MAE of the given results presented in this chapter may be satisfactory, the MAPE leaves plenty of room for improvement. An approach that has shown promising results in the health and well-being domain is Personalized ML [4, 57]. By personalized ML, we mean the models that are expected to work well for each individual, and not just for the average population, as traditionally approached in ML. However, a throughout experimentation on such an important and broad topic is out of the scope of this research and we include it in our work in order to conduct an initial experiment and set the foundation for future work on this matter.

5.6 Generalized and Personalized Model’s Performance

As stated above, an in-depth comparison of the efficiency of generalized and personalized models is out of the scope of this research. The purpose of this section’s experiment, is to provide an initial insight regarding the generalized versus the personalized approaches and set the foundations for further research on this field.

For this experiment, we utilized two versions of a dataset with identical features. The first one, the personalized dataset, was created on data originated to a single participant of the research. This participant, was the one with the most recorded days of activity, as this factor would enable us to train our models more efficiently. 10% of this user’s data were held out to be used as a test set for evaluation purposes. The second dataset, used for the generalized model, contained data from all the participants, except those of the above mentioned user, that was selected for the personalized approach. In order to create the training set of the generalized
model, the personalized training set was added to the generalized dataset. At this stage and due to the experimental set-up described above, the data included in the test set were unknown to both the personalized and the generalized models. A schematic representation of the above described specifications, is given in Figure 5.12 below:

Through this experimental process, we wish to assess the performance of the generalized and the personalized models in forecasting a user’s future daily step-count. The two models were trained on their respective datasets and next, the MAE and MAPE metrics were calculated for their predictions on the unknown test set. The results of this experiment are presented in Figure 5.13.

![Figure 5.12: Outline of the personalized vs generalized approach.](image)

![Figure 5.13: MAE and MAPE of the generalized and the personalized models on the unknown test set](image)

(a) MAE  
(b) MAPE
The test set in which the two models were evaluated, was a set of 160 days of activity, originated to one specific user. As seen in the above graphs, the two models have a quite similar performance regarding the MAE metric. Although the same applies for the MAPE, considering this metric, the generalized models seems to slightly outperform the personalized one. In Table 5.10 bellow we present the values of the two metrics used in the evaluation process.

**Table 5.10:** MAE and MAPE values for the personalized and the generalized models on an unknown test set.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized</td>
<td>1789</td>
<td>46.92%</td>
</tr>
<tr>
<td>Personalized</td>
<td>1817</td>
<td>50.66%</td>
</tr>
</tbody>
</table>

From the results of Table 5.10, we observe that the two models have a similar performance with the generalized slightly outperforming the personalized one. A possible explanation of this observation is that the incorporation of personal features in the dataset, gave the ability to the generalized model to distinguish between different users to some extent. This fact is yet another indication of the importance of this research’s C2 contribution, since this was the goal we hoped to achieve when including personal features to our dataset. In addition to the above, the generalized model seems not only to adequately predict PA of a single user, but also benefit to some extent from the information originated to different participants of the research since it outperforms the personalized model. In Figure 5.14 bellow, we present a plot with the predictions of the two models amongst with the actual step-count values, for 20 days of the unknown test set.

**Figure 5.14:** Personalized and generalized models’ prediction VS the actual step-count for 20 days of an unknown test set.
Having said the above, as already stated, this experiment is only a small step in experimenting with a topic of great importance and space for research and thus, the true meaning of the given results is an issue that should be explored further in future works in order to be adequately addressed.

As a second experiment, we trained a new generalized model, but this time we did not include any data of the user selected for the personalized approach. That means, that our generalized prediction model is trained on data from the other 20 users and we use it to predict the step-count of user 21, a user whose data are never before seen by the generalized ML model. The MAE and MAPE metrics of this approach are presented in Figure 5.15 below, in comparison with the MAE and MAPE of the personalized model.

![Figure 5.15: MAE and MAPE of the generalized and the personalized models on the unknown test set of a completely unknown to the generalized model user.](image)

In this scenario, it is obvious that the performance of the generalized model has significantly worsen, but on the other hand as seen in Figure 5.16, the model is able to identify some patterns in the user’s step-count, even though it has no previous knowledge of any information related to this specific participant of the research. Most likely, with a greater number of users whose data will be utilized for the generalized model’s training, the efficiency of its predictions, even for completely unknown individuals, would be further improved. This topic, as already stated, requires a more in-depth analysis that is out of the scope of this work, and is left as an inspiration for future research.

At first glance, the given results indicate that the generalized model outperforms the personalized one when both have been trained on data of a specific user and as expected, the contrary applies when there is no previous knowledge of a user for the generalized model. Despite that, the selection of one model over the other and the proclamation of an approach as ”better than the other” is not as simple as running a few experiments. It is a statement that depends on a
number of factors. Consider the following question: are 100 personalized models, each quite efficient in predicting exactly one user’s step-count, better than one model that is able to provide not ideal but adequate predictions for all of them? The answer is probably “it depends”. The application that incorporates these models, the data availability of each user and the available resources for model development and maintenance are only a few, of the multitude of factors to be considered when making such a decision.

In this section, we presented an experiment, conducted towards our goal to implement our C4 contribution point, in which we evaluated the performance of a generalized and a personalized model in predicting a specific user’s future daily step-count. The results to some extent were expected, with the 2 models performing similarly when they both have been trained on data originated to this specific individual and with the personalized model performing better when the generalized model had no previous knowledge of the user in question. In addition to that, we provided some possible explanations regarding the meaning of the given results, while underlying the importance of further research on this topic. Finally, we elaborated on the complexity of the "personalized vs generalized" question and presented our take on this matter.

This section was the last section of the chapter in which we presented the results of the experimental process of our research. Firstly, in section 5.1, we presented the evaluation methodology adopted in this research and discussed on the importance of adequately assessing and reporting the performance of similar prediction models. In section 5.2, we presented the results of the process by which the decision about the importance of the different feature categories was
made. Through these results we decided which features should be included on the dataset and proved the value of our C2 contribution, meaning our decision to take into account a combination of activity, personal and contextual features in the creation of the dataset for step-count prediction. Next, in section 5.3, we presented the results of our outlier-handling methodology. We shown the importance of such an approach and discussed the performance of our prediction models before and after outlier removal. In the following section, section 5.4, we shown the results of the 2 approaches tested towards dimensionality reduction of the dataset. We proved that our models can benefit form such approaches, with RFE leading to the best results considering the aims and limitations of this research. In section 5.5, we presented the results given when different numbers of input days were used as input for our models. According to this results, no significant difference in the performance of the prediction models was observed and the window size of 5 days was chosen due to the fact that it led less complex models. Finally, in section 5.6, we discussed the issue of choosing between a personalized and a generalized model and presented the corresponding results of a relevant experimental process in our research.

Through our methodology, discussed in Chapter 4, and the results presented in this chapter we provided proof-of-work for our decisions towards the creation of the dataset that implements our contributions C1, C2 and C3 and the development of an efficient ML model for step-count forecasting that implements our contribution C4. In the next chapter we will present the conclusions of this research amongst with our ideas for future work.
Chapter 6

Conclusions & Future Work

As discussed in detail in Chapter 1, the increment of a person’s PA levels may lead to significant benefits not only on an individual basis but also for the society as a whole. Daily step-count data has been found to be an objective and quantitative measure of PA levels, while it has the additional positive feature of universality, since it can be measured in the vast majority of people, regardless of physical condition, age or other factors that differentiate them. In the recent years, the technological and economical advances have made possible to incorporate activity tracking functionality in devices we use in our everyday lives, such as wrist-watches and mobile phones. This fact has made possible for researchers to explore step-count data, originated to wearable activity tracking devices in order to identify approaches in increasing PA levels.

The purpose of our research is to contribute towards this end in two ways. Firstly, by creating a novel dataset which includes a combination of activity-related, personal and contextual features and secondly, by developing a model which utilizes the above dataset in forecasting a user’s future daily step-count. A model like this could be incorporated in the core of more complex goal setting and intervention approaches aimed into positively altering a user’s behaviour regarding PA.

By studying literature relevant to our work, we have identified a number of issues that in our opinion have not been addressed adequately, issues that in our research are addressed through our contributions described and annotated as C1, C2, C3 and C4 in Chapter 1. By implementing C1, we designed and applied an integration framework that enabled us to pool activity data originated to the 2 most popular brands, Apple and Xiaomi, considering their share of the wearable activity trackers’ market. In C2 we included in our dataset a novel combination of activity, personal and contextual features. Through the assessment of the impact that each of these feature categories had on the predictive ability of the ML models, we proved that our approach was correct and such features can contribute positively to the performance of
prediction models. C3 is related to the data recording and collection process. We noticed that in most relevant works, activity data were gathered from users during the time period in which they were actively participating in an ongoing research. As it has been found that such an approach may lead to inaccurate reflection of an individual’s PA levels [67], in our work, we utilize "in-the-wild" data originated to user activity unrelated to their participation on the ongoing research. Finally, C4 contribution concerns the development of a ML prediction model, capable of efficiently predicting a user’s future step-count, given a set of features for a sequence of previous days. The results given by the different approaches tested towards this end, indicate that we met these requirements through the development of such a model. Table 6.1 bellow, presents the above contributions collectively.

<table>
<thead>
<tr>
<th>CONTRIBUTIONS</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Implementation of an integration framework for wearable devices and mobile applications</td>
</tr>
<tr>
<td>C2</td>
<td>Incorporation of activity, personal and contextual features in the prediction models</td>
</tr>
<tr>
<td>C3</td>
<td>Usage of &quot;in-the-wild&quot; data for more accurate representation of PA levels</td>
</tr>
<tr>
<td>C4</td>
<td>Development of an end-to-end ML step-count prediction model</td>
</tr>
</tbody>
</table>

The importance and significance of our research, lies in all of the contributions described above. The dataset we created can be used in future works that may built on our experimental results. The prediction model we developed, is a stand-alone end-to-end forecasting tool that can be incorporated in any approach that could make use of it, in order to help users in increasing their PA levels and adopt a healthier lifestyle. Based on the above, we believe that the objectives of this research have been achieved, without this of course meaning that there is no room for further improvement through future work.

An important issue that we would like to address in the future is the relatively small number of participant’s in the research. A bigger population sample would provide a greater number of data and that means more days of recorded activity for the ML models to be trained on. In that manner, the prediction models would be more robust regarding their performance and also would be the results of the different experimental approaches applied in our work. In addition to the above, as stated in 5.5, having a greater amount of data available, would allow the experimental process of obtaining the optimal input days number to be further diversified by the addition of more window sizes to experiment with.

In a similar manner to the addition of more data, another future work idea is the incorporation of more features in the dataset created and the assessment of their impact on the predictive ability of the ML models. Such features, might be features related to a user’s location, the
weather conditions on specific days of activity, personality traits different from those used in this work and more. In reality, the available space of features that may be directly or indirectly related to a user’s daily step-count is vast. Thus, in our opinion any future research that attempts to identify the impact of a number of features on the performance of step-count prediction models can be considered as a contribution to this field.

Another idea for future work, is an in depth experimentation with the "personalized versus generalized" approach. As stated in 5.6, a throughout research on this topic is out of the scope of this research. Despite that, the results given through our experimental process, provide interesting insight and indicate that there more than enough space for further work regarding the pros and cons of a personalized over a generalized approach and the contrary.

Finally, as the last future work idea stated here, we propose the incorporation of our model in a complete application, aimed in increasing an individual’s PA levels. The ML prediction model, developed in this research, was designed having this approach in mind and with the hope of contributing to the cause of promoting a healthier lifestyle on people. Personalized and adaptive goal setting approaches and related intervention application could make use of a model like the one described in our work, in the core of their goal setting methodology.
Appendix A

Questionnaires

A.1 Demographics Questionnaire

- Φύλο
  - Άνδρας
  - Γυναίκα
  - Άλλο

- Ηλιοσφαίριση

- Οικογενειακή Κατάσταση
  - Ελεύθερος-η
  - Παντρεμένος-η/Σε ελεύθερη συμβίωση
  - Χήρος-α
  - Χωρισμένος-η/Σε διάσταση
  - Άλλο

- Εκπαίδευση
  - Απόφοιτος-η Δημοτικού
  - Απόφοιτος-η Γυμνασίου/Λυκείου
  - Απόφοιτος-η ΙΕΚ
  - Πτυχιούχος Πανεπιστημίου
Α.2 Personality Questionnaire IPIP

Στις σελίδες που ακολουθούν, υπάρχουν φράσεις που περιγράφουν συμπεριφορές ανθρώπων. Σας παρακαλούμε να χρησιμοποιήσετε τις απαντήσεις που βρίσκονται δίπλα σε κάθε πρόταση για να περιγράψετε πόσο πιστά κάθε πρόταση περιγράφει εσάς. Περιγράψτε τον εαυτό σας όπως είστε στη στιγμή και όχι όπως θα ήθελατε να ήσαστεν στο μέλλον. Περιγράψτε τον εαυτό σας όπως ειλικρινά τον αντιλαμβάνεστε σε σχέση με άλλους ανθρώπους το ίδιο φύλου και περίπου της ίδιας ηλικίας. Σας παρακαλούμε να διαβάσετε κάθε πρόταση προσεκτικά και να κυκλώστε τον αριθμό που αντιστοιχεί στην πρόταση που επιλέγετε.

| Καθόλου Αν- | Λίγο Αν- | Δεν Μπορώ | Κάπως Αν- | Πολύ Αν- |
| τιπροσω- | τιπροσω- | να Απο- | τιπροσω- | τιπροσω- |
| πυτικό | πυτικό | φασίσω | πυτικό | πυτικό |

<p>| Είμαι η ζωή σε ένα πάρτι | 1 | 2 | 3 | 4 | 5 |
| Αισθάνομαι μικρό ενδιαφέρον για τους άλλους | 1 | 2 | 3 | 4 | 5 |
| Είμαι πάντοτε προετοιμασμένος | 1 | 2 | 3 | 4 | 5 |
| Αγχώνομαι εύκολα | 1 | 2 | 3 | 4 | 5 |
| Έχω ένα πλούσιο λεξιλόγιο | 1 | 2 | 3 | 4 | 5 |
| Δεν μιλώ πολύ | 1 | 2 | 3 | 4 | 5 |</p>
<table>
<thead>
<tr>
<th>Ενδιαφέρομαι για τους ανθρώπους</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Αφήνω τα πράγματα μου ολόγυρα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Είμαι χαλαρός/ή τις περισσότερες φορές</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Δυσκολεύομαι να κατανοήσω αφηρημένες ιδέες</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Αισθάνομαι άνετα όταν βρίσκομαι ανάμεσα σε ανθρώπους</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Προσβάλλω τους άλλους</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Δίνω προσοχή στις λεπτομέρειες</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Ανησυχώ για διάφορα πράγματα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Έχω ζωηρή (ζωντανή) φαντασία</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Προτιμώ να μένω στο παρασκήνιο</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Συμπάσχω με τα συναισθήματα των άλλων</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Τα κάνω άνω κάτω</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Σπάνια νοιώθω μελαγχολία</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Δεν ενδιαφέρομαι για αφηρημένες ιδέες</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Αρχίζω συζητήσεις</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Δεν ενδιαφέρομαι για τα προβλήματα των άλλων</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Κάνω τις «αγαρείες» αμέσως</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Ενοχλούμαι εύκολα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Έχω εξαιρετικές ιδέες</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Έχω ελάχιστα πράγματα να πω</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Έχω μαλακή καρδιά</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Συχνά ξεχνώ να βάζω τα πράγματα πίσω στη σωστή τους θέση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Αναστατώνομαι εύκολα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Δεν έχω καλή φαντασία</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Μιλώ με πολλούς διαφορετικούς ανθρώπους στα πάρτι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Δεν ενδιαφέρομαι πραγματικά για τους άλλους ανθρώπους</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Μου αρέσει η τάξη</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Η διάθεσή μου αλλάζει διαρκώς</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Είμαι γρήγορος/η στο να καταλαβαίνω πράγματα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Λέγοντά μου αρέσει να προσελευτώ την προσοχή πάνω μου</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Βρίσκω χρόνο για τους άλλους</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Αποφεύγω αυτά που πρέπει να κάνω (τα καθήκοντά μου)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Έχω συχνές εναλλαγές στη διάθεσή μου</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Χρησιμοποιώ δύσκολες λέξεις</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Λένε με ενοχλεί να είμαι το επίκεντρο της προσοχής</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Αισθάνομαι τα συναισθήματα των άλλων</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Ακολουθώ ένα πρόγραμμα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Εκπαιδεύομαι εύκολα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Αφιερώνω χρόνο για να αξιολογώ τα πράγματα (που κάνω)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Είμαι ήσυχος/η όταν βρίσκομαι ανάμεσα σε ξένους</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Κάνω τους ανθρώπους να αισθάνονται άνετα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Είμαι αφριβής στη δουλεία μου

Συχνά ασθάνωμαι μελαγχολικά

Είμαι γεμάτος/η ιδέες

### Α.3 Questionnaire of Transtheoretical Model of behavior change

- Επιλέξτε μια από τις επόμενες προτάσεις που είναι πιο κοντά στη συμπεριφορά σας ως προς την άσκηση:

  - □ Αυτή την χρονική περίοδο δεν ασκούμαι και δεν σκοπεύω να ξεκινήσω στους επόμενους 6 μήνες
  - □ Αυτή την χρονική περίοδο δεν ασκούμαι, αλλά σκέφτομαι να ξεκινήσω να ασκούμαι στους επόμενους 6 μήνες
  - □ Αυτή την χρονική περίοδο ασκούμαι, αλλά όχι τακτικά
  - □ Ασκούμαι τακτικά τους τελευταίους 6 μήνες
  - □ Ασκούμαι τακτικά για περισσότερο από 6 μήνες
  - □ Έχω ασχημεί τακτικά στο παρελθόν, αλλά δεν ασκούμαι τόσο αυτήν την χρονική περίοδο

<table>
<thead>
<tr>
<th>Καθόλου</th>
<th>Λίγο</th>
<th>Μέτρια</th>
<th>Πολύ</th>
<th>Πάρα/ Πολύ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Θυμάμαι πληροφορίες που κάποιοι μου έχουν δώσει για τα οφέλη της άσκησης</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Χρησιμοποιώ πληροφορίες από άρθρα και διαφημίσεις για το πώς μπορώ να κάνω την άσκηση μέρος της ζωής μου</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Στην προσπάθεια μου να μάθω περισσότερα για την άσκηση, διαβάζω άρθρα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ψάχνω και πληροφορούμαι σχετικά με την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Οι προειδοποιήσεις για τον κίνδυνο της υγείας από την έλλειψη άσκησης με ευαισθητοποιούν</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Η δραματική παρουσίαση των αρνητικών συνεπειών της αδράνειας με ευαισθητοποιούν</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ευαισθητοποιούμε στις προειδοποιήσεις των συνεπειών της μη άσκησης</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Αισθάνομαι πως θα ήμουν πρότυπο για τους άλλους εάν έκανα άσκηση περισσότερο</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Αναφερτώ ότι η έλλειψη άσκησης επηρεάζει τους κοντινούς σε μένα ανθρώπους</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Συνειδητοποίω ότι, εάν ασκούμαι πιο πολύ μπορώ να επηρεάζω τους άλλους στο να ακολουθούν ένα πιο υγιεινό τρόπο ζωής</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Εάν γυμναστώ περισσότερο, κάποιοι από το κοντινό μου περιβάλλον θα ενθαρρυνθούν να κάνουν το ίδιο</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Το σκέφτομαι εάν η συχνή άσκηση θα με έκανε πιο υγιή και πιο ευτυχισμένο</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Σκέφτομαι πώς θα είμαι, εάν εξαιρετικά να ασκούμαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Θυμόμενο με τον εαυτό μου όταν δεν ασκούμαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Με ενδιαφέρει το ότι θα είχα μεγαλύτερη αυτοπεποίθηση εάν έκανα συστηματική άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Εντοπίζω κοινωνικές αλλαγές που βοηθούν τα άτομα να αδληθούν</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Γνωρίζω συνεχώς άτομα που με ενθαρρύνουν να ασκηθώ</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Παρατηρώ πως πολλές επιχειρήσεις ενθαρρύνουν τους υπαλλήλους να γυμναστούν προσφέροντας τους παιδιά προγραμμάτων άσκησης και ελεύθερο χρόνο</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Γνωρίζω πολλά γυμναστήρια που προσφέρουν στα μέλη τους δωρεάν φίλαξη των παιδιών</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Συμμετέχω σε κάποια φυσική δραστηριότητα για να μην μένω αδρανής</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Από το να αντιμετωπίζω την άσκηση ως μια καθημερινή υποχρέωση, προσπαθώ να την βλέπω ως ένα μέσο χαλάρωσης και καθημερινής αναζωογόνησης</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Όταν είμαι κουρασμένος ασκούμαι γιατί γνωρίζω πως μετά θα νιώθω καλύτερα</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Όταν νιώθω πιεσμένος, εκτονώνουμαι με την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Έχω κάποιον στον οποίο μπορώ να βασιστώ όταν αντιμετωπίζω κάποιο πρόβλημα με την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Έχω κάποιον φίλο/ή που γυμνάζεται και με ενθαρρύνει όταν δεν νιώθω καλά</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Έχω κάποιον που μου επισημαίνει τις αρνητικές επιτώσεις όταν δεν γυμνάζομαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Έχω κάποιον που με συμβουλεύει για την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ανταμείβω τον εαυτό μου όταν ασκούμαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Προσπαθώ να βάζω ρεαλιστικούς στόχους όταν ασκούμαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Πείθω τον εαυτό μου, ότι η άσκηση κάνει καλό στο σώμα μου</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Κάνω καλό στον εαυτό μου όταν ασκούμαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Λέω στον εαυτό μου πως είμαι ικανός/η να συνεχίζω να ασκούμαι, εάν το θέλω πραγματικά</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Λέω στον εαυτό μου πως μπορώ να συνεχίσω να ασκούμαι, εάν προσπαθήσω σκληρά</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Δεσμεύομαι να συνεχίσω να ασκούμαι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Θυμίζω στον εαυτό μου ότι είμαι ο μοναδικός υπεύθυνος/η για την υγεία μου και μόνο εγώ μπορώ να αποφασίσω αν θα ασκηθώ ή όχι</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Αφήνω σε εμφανή σημεία στο σπίτι μου τον αθλητικό εξοπλισμό για να μου υπενθυμίζουν την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Αφήνω σε εμφανή σημεία στη δουλεία μου τον αθλητικό εξοπλισμό για να μου υπενθυμίζουν την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Απομακρύνω ότι με αποτρέπει από την άσκηση

<table>
<thead>
<tr>
<th>Επιμέτρηση</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Αποφεύγω να περνώ πολύ χρόνο σε περιβάλλον που δεν ευνοεί την άσκηση</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
## Appendix B

### COVID-19 Features

<table>
<thead>
<tr>
<th>NAME</th>
<th>DESCRIPTION</th>
<th>MEASUREMENT</th>
<th>CODING</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1_School-Closing</td>
<td>Record closings of schools and universities</td>
<td>Ordinal scale</td>
<td>0 - no measures&lt;br&gt;1 - recommend closing or all schools open with alterations resulting in significant differences compared to non-Covid-19 operations&lt;br&gt;2 - require closing (only some levels or categories, eg just high school, or just public schools)&lt;br&gt;3 - require closing all levels&lt;br&gt;Blank - no data</td>
</tr>
<tr>
<td>C1_Flag</td>
<td>Binary flag for geographic scope</td>
<td>Binary flag</td>
<td>0 - targeted&lt;br&gt;1 - general&lt;br&gt;Blank - no data</td>
</tr>
<tr>
<td>C2_Workplace_closing</td>
<td>Record closings of workplaces</td>
<td>Ordinal scale</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 - no measures</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 - recommend closing (or recommend work from home)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 - require closing (or work from home) for some sectors or categories of workers</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 - require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C2_Flag</th>
<th>Binary flag for geographic scope</th>
<th>0 - targeted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 - general</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C3_Cancel_public_events</th>
<th>Record cancelling public events</th>
<th>Ordinal scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 - no measures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 - recommend cancelling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 - require cancelling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C3_Flag</th>
<th>Binary flag for geographic scope</th>
<th>0 - targeted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 - general</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C4_Restrictions_on_gatherings</th>
<th>Record limits on gatherings</th>
<th>Ordinal scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 - no restrictions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 - restrictions on very large gatherings (the limit is above 1000 people)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 - restrictions on gatherings between 101-1000 people</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 - restrictions on gatherings between 11-100 people</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 - restrictions on gatherings of 10 people or less</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C4_Flag</th>
<th>Binary flag for geographic scope</th>
<th>0 - targeted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 - general</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
</tr>
</tbody>
</table>
| C5_Close_public_transport | Record closing of public transport | Ordinal scale | 0 - no measures  
1 - recommend closing (or significantly reduce volume/route/means of transport available)  
2 - require closing (or prohibit most citizens from using it)  
Blank - no data |
|---|---|---|---|
| C5_Flag | Binary flag for geographic scope | 0 - targeted  
1 - general  
Blank - no data |
| C6_Stay_at_home_requirements | Record orders to shelter in place and otherwise confine to the home | Ordinal scale | 0 - no measures  
1 - recommend not leaving house  
2 - require not leaving house with exceptions for daily exercise, grocery shopping, and 'essential' trips  
3 - require not leaving house with minimal exceptions (eg allowed to leave once a week, or only one person can leave at a time, etc)  
Blank - no data |
| C6_Flag | Binary flag for geographic scope | 0 - targeted  
1 - general  
Blank - no data |
| C7_Restrictions_on_internal_movement | Record restrictions on internal movement between cities/regions | Ordinal scale | 0 - no measures  
1 - recommend not to travel between regions/cities  
2 - internal movement restrictions in place  
Blank - no data |
| C7_Flag | Binary flag for geographic scope | 0 - targeted  
1 - general  
Blank - no data |
<table>
<thead>
<tr>
<th>C8_International-travel_controls</th>
<th>Record restrictions on international travel</th>
<th>Ordinal scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note: this records policy for foreign travellers, not citizens</td>
<td></td>
<td>0 - no restrictions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 - screening arrivals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 - quarantine arrivals from some or all regions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 - ban arrivals from some regions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 - ban on all regions or total border closure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blank - no data</td>
</tr>
</tbody>
</table>
Bibliography


[54] Loomba Smirti and Khairnar Aniket. 2018. Fitness Trackers Market by Device Type (Fitness Bands, Smartwatch, and Others), Display Type (Monochrome and Colored), Sales Channel (Online and Offline), and Compatibility (iOS, Android, Windows, Tizen, and Others) - Global Opportunity Analysis and Industry Forecast, 2017-2023. https://www.alliedmarketresearch.com/fitness-tracker-market


