# Identification of Profiles Based on Physical Behaviour Patterns

Arwin Sleutjes
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
a.r.sleutjes@student.utwente.nl

## ABSTRACT

Continuous monitoring of physical activity using wearable technology is becoming ubiquitous and allows for real-time interventions through digital means aiming at the promotion of an active lifestyle. Many people use fitness apps and wearable technology to support it [20, 14]. However, digital health interventions still pose some problems in their effectiveness and usability. Bad tailoring of the interventions has been recognised as one of the main problems [11]. As a result, there is a sudden drop in adherence to digital health interventions after a novelty period. A one-size-fits-all solution is likely to be ineffective, thus there is a need for personally customised interventions [26]. This project applies knowledge discovery techniques to physical activity data and some contextual variables to find patterns which can support personally tailored and comprehensible interventions. It also takes into account the additional dimension of the experienced pleasure during activity.

## Keywords

data mining, subgroup discovery, clustering, k means, dbscan, physical behaviour, physical activity, pleasure, older adults

## 1. INTRODUCTION

Nowadays technology is applied in many fields to assist people and to enhance their quality of life. On the domain of health, e-health makes use of electronic systems and communication to improve healthcare. One important feature of e-health is that it allows real-time intervention in daily life. Of course, these interventions have to be based on data. Moreover, health data can be tracked real-time with the use of wearable technology. The market size of fitness trackers was valued at $17,907 million in 2016 and is expected to value more than triple that by 2023 [14]. Besides that, many people make use of fitness tracking or e-health apps. One of the more popular fitness applications, Runkeeper, has around 45 million users globally [20].

One of the more recurrent applications of e-health is in the promotion of physical activity through digital health interventions [24, 6]. These interventions have been proven to be very effective for young people [15].The large amount of use of tracking technology and fitness applications results in a growing amount of data from a wide range of sources. Combined with the various techniques for knowledge discovery in databases [4], that data can support these digital health interventions.

There already exist applications of knowledge discovery techniques to support digital health interventions to promote an active lifestyle [15]. But not many consider the experienced pleasure during activities. Besides that, digital health interventions still pose some problems in their effectiveness and usability. Bad tailoring of the interventions has been recognised as one of the main problems [11]. As a result, there is a sudden drop in adherence to digital health interventions after a novelty period. A one-size-fits-all solution is likely to be ineffective, thus there is a need for personally customised interventions [26].

This project focuses on the discovery of patterns that can support the personalisation of digital health interventions. It approaches the problem of personalisation in a novel way by applying knowledge discovery techniques, that provide comprehensible results, to behavioural data. The project also takes into account the additional dimension of the experienced pleasure during activity.

## 2. RESEARCH AIMS

To aid in solving the problem of personalisation of the digital interventions in physical activity, this project answers the following research questions:

RQ1 How can knowledge discovery be applied to data on the physical activity and pleasure of older adults and some contextual variables?

RQ2 What patterns can be found in the data on the physical activity of older adults based on the contextual variables?

RQ3 What patterns can be found in the data on the pleasure of older adults based on the contextual variables?

RQ4 How are the physical activity and the pleasure of older adults related?

## 3. CONTEXT

### 3.1 Knowledge Discovery in Databases

Knowledge discovery in databases (KDD) is the process of extracting knowledge from data. It is a field of research in data science. In the literature, KDD is defined as: ”the non-trivial process of identifying valid, novel, potentially

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

$32^{nd}$ *Twente Student Conference on IT* Jan. $31^{st}$, 2020, Enschede, The Netherlands.

Copyright 2020, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

useful, and ultimately understandable patterns in data" [3]. The fact that KDD is defined as a process means that it consists of multiple steps. These steps are data preparation, search for patterns, knowledge evaluation, and refinement [2]. These steps are often repeated for multiple iterations of the process.

### 3.1.1 Data Preparation

A big part of the manual work of KDD is often found in the data preparation step. It consists of selecting the right data, preprocessing it and transforming it. Preprocessing means removing any noise in the data, dealing with any missing values and integrating any other knowledge that could affect the data. In the transformation step, the variables in the data are being analysed. Variables that do not add a significant dimension to the data can be removed from consideration in the search for patterns. Also, the data will be converted to a format that is suitable for the techniques used in the search for patterns.

### 3.1.2 Search for Patterns

The next important part of knowledge discovery is the search for patterns in the prepared data. This part is popularly called data mining. It is concerned with applying analysis methods to data to find patterns and models [4]. There are various methods to perform data mining, but most of these methods can broadly be classified into one of the following 2 categories.

- Predictive methods aim to discover knowledge that is useful for the prediction of a variable of future objects
- Descriptive methods aim to discover knowledge that is useful for the description of the current objects

Principles of data mining techniques are often similar to or based on methods from the fields of machine learning, pattern recognition and statistics. Examples of techniques are classification, clustering and regression [4].

This project mostly deals with descriptive methods to be able to draw comprehensible conclusions that can support the design of personally customised digital health interventions.

# 4. SELECTED DATA MINING METHODS

For this project, a selection of data mining techniques has been made. The application of these methods will later be described. First, the focus lies on the theory and functioning of the selected methods.

## 4.1 Subgroup Discovery

### 4.1.1 Definition

Subgroup discovery is used to discover statistically interesting subgroups with respect to a target variable. It can be defined as follows [25]:

In subgroup discovery, there is a population of objects (time-series samples in this project). All these objects have a value for a certain interesting variable, the target variable (physical activity or pleasure in this project). The task of subgroup discovery is to discover subgroups of the population (subsets of samples) that are statistically interesting. Subgroups are statistically more interesting if they are as large as possible and have the most unusual statistical (distributional) characteristics regarding the target variable. Subgroups are induced by conditions based on other variables besides the target variable.

### 4.1.2 Position in Data Mining

Subgroup discovery lies between the extraction of classical association rules and classification rules [7]. It is different from classical association rules in that subgroup discovery is targeted at one variable where the extraction of association rules does not have one specific target variable. Subgroup discovery is also different from classification rules as it does not look for an exhaustive model for the object space. In this way, it is also different from most predictive data mining methods. However, if need be, it can be used to make predictions on the entire object space. This can be done by using knowledge from interesting subgroups to predict the value of the target variable. If no subgroups have been found, then the average or the most common value should be predicted.

### 4.1.3 Output

The output of subgroup discovery is shaped as rules with conditions on the left-hand side and a description of the value(s) for the target variable on the right-hand side. This is often represented as shown in Equation 1.

$$R : Cond \implies Target_{value} \tag{1}$$

Where rule $R$ states that, if the conditions in $Cond$ have been met, then the (statistically interesting) (distribution of the) $_{value}$ for the $Target$ (variable) is found. The part on the right-hand side of the rule can be a single value, where the probability of finding that value is interesting. It can also be the distribution of all found values that is interesting. This depends on the type of the target variable.

### 4.1.4 Important Elements

As said, the type of the target variable is one of the important elements to be considered when applying subgroup discovery. Others are the description language, quality measures and search strategy [7].

**The types of the target variables** that this project evaluates are all numeric. These variables represent physical activity and pleasure. In subgroup discovery, a numeric variable can be studied by dividing the variable into two ranges with respect to the average, discretizing the target variable in a determined number of intervals [16], or searching for significant deviations of the mean of the distribution of the values among others [7].

**The description language** of the output rules in subgroup discovery must be simple. The conditions in Equation 1 represented as $Cond$ are described in terms of pairs of variables and values. These values can have the form of a simple boolean, some fuzzy logic or they can be used in an expression of (in)equality [7].

There is also a need for **quality measures** to evaluate the interestingness of the rules resulting from subgroup discovery. There exist many methods to evaluate a rule [7], but since the project considers only numerically typed target variables the focus lies on quality measures for those. One such quality measure of a subgroup is the **Z-score**. This is defined as the distance between the mean of the subgroup and that of the entire population in terms of the standard deviation of the population. However, besides the Z-score, more factors play a role in the interestingness of a subgroup rule. One other factor is the size of the subgroup. The standard deviation of the subgroup itself is another. If a nominally typed variable is targeted, then a relevant factor is the accuracy with which a subgroup rule correctly describes the value of the target variable.

A quality measure that covers most important factors for nominally typed target variables is the **WRAcc** as defined in Equation 2 [13].

$$WRAcc = p(Cond) * (p(Target_{value}|Cond) - p(Target_{value})) \quad (2)$$

- $p(Cond)$ means the probability that $Cond$ holds.
- $p(Target_{value}|Cond)$ is the probability that the target variable has the specified value given that the conditions of the subgroup rule hold.
- $p(Target_{value}|Cond) - p(Target_{value})$ is the accuracy gain relative to the entire population.

The WRAcc is not useful for numerically typed target variables, but the Z-score, combined with the size and the standard deviation of the subgroup, functions similarly.

The **search strategy** in subgroup discovery is important for finding interesting subgroups. The default search strategy is called beam search. It is based on best-first search which again is based on breadth-first search. Best-first search applies breadth-first but visits child nodes of a node in the order of some quality measure. Beam search does the same but also has a threshold for the number of nodes it will consider visiting at each level of the search tree and is thus a greedy algorithm since it might miss interesting results. The search tree that is traversed and the nodes therein are constructed from conditions that form a (partial) subgroup rule.

#### *4.1.5 Software*

Subgroup discovery was performed using Cortana Open Source Subgroup Discovery [17]. This is an open-source tool with a graphical user interface developed at the Leiden Institute of Advanced Computer Science. It has a few useful features that can be used in this project. Firstly, it supports binary, nominal and numerical data. Secondly, it has many quality measures including the Z-score for numerically valued data. Lastly, Cortana has integrated visualisations for the data and any subgroups. It is also possible to filter subgroups based on statistical significance to reduce the number of results before human analysis. The Z-score quality measure in Cortana also takes into account the size of the subgroup and its standard deviation.

### 4.2 Clustering

Data clustering is another method of data mining. It is described as the process of identifying natural groupings or clusters within multidimensional data based on some similarity measure [18]. It is different from subgroup discovery in that subgroup discovery traverses possible conditions on the data and looks for interesting distributions of values for the target variable based on those conditions. On the other hand, clustering traverses all seemingly interesting distributions of the target variable. There exist many algorithms to perform clustering [12]. This project examines and applies two of them.

#### *4.2.1 K Means*

In K Means clustering the algorithm starts with a predetermined number of clusters based on the initialisation parameters. For each cluster, a centre is chosen at the start, based on the initialisation parameters of the algorithm. For all data points, the algorithm determines which of the cluster centres is the closest based on the euclidean distance. The data points are associated with the cluster that has the closest centre. Then the location of the centre of each cluster is shifted to the mean of the data points in it. Again, all data points are associated with the (possibly new) closest cluster centre. This is repeated until no significant changes happen during an iteration or a maximum amount of iterations has been reached.

#### *4.2.2 DBSCAN*

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. In DBSCAN clustering there are 2 important variables. One is the minimum amount of points in a cluster and the other is the maximum distance between two points to be considered nearby. The algorithm starts off by picking one data point at random. It will then look for all data points in the neighbourhood. All points that are closer than the maximum distance are considered to be in the same new group. They are all evaluated in the same manner until no more nearby points are found. Then the algorithm checks if the minimum amount of points has been reached. If this is the case then the new group will be considered as a cluster. If not, then the points will be considered as outliers. If there are any remaining data points not in a cluster or considered as an outlier, then one of them is picked at random and the procedure is repeated.

#### *4.2.3 Software*

The python library scikit-learn has implementations for clustering using K Means and DBSCAN [19]. These implementations were used in this research.

## 5. DATA

Table 1 describes the relevant variables of the data set that was used in this project [1]. The data was collected from 10 older adults of ages 65 to 83. The experiments had a duration ranging from 24 to 38 days with varying starting dates. Throughout the experiment, the subjects were all in a free-living environment. This also resulted in slightly noisy data (subjects have varying waking hours).

Subjects were prompted with a question asking what they were doing, approximately every hour from 08:00 until 20:00. Following that question, subjects were asked a question about their location, the type of participants in the activity (if any) and the experienced pleasure. The combined answers to these questions constitute 1 sample.

Besides that, the physical activity of the subjects has been measured using hip-worn accelerometers. The subjects were instructed to wear the sensor during waking hours. This accelerometer data has been quantified as integral module of acceleration (ima) per minute. The variable ima0505 associated with each sample consists of the sum of raw ima values in a 10-minute window around the time the sample was taken.

Additionally, a data set containing daily weather data from the Royal Dutch Meteorological Institute has been used [8]. The data set contained all available data about the weather during the experiment dates. The relevant variables are presented in Table 2.

### 5.1 Software

All data preparation was done using python [23] in a Jupyter [10] project with the Jupyter Lab [9] user interface. During this process pandas [21], a python library, was used to select, preprocess and tranform the data. The python library seaborn [22] was used for visualisation of the data.

**Table 1. Description of Data Set**

| Variable | Type | Comment |
|---|---|---|
| sub | numerical | id of the subject |
| date | date | date |
| time | time | time |
| exp_day | numerical | relative day in the experiment |
| pleasure | numerical | pleasure graded 1 through 10 |
| activity_id | nominal | sanitised activity_text |
| activity_text | nominal | description of activity |
| location_id | nominal | sanitised location_text |
| location_text | nominal | description of location |
| participant_id | nominal | sanitised participant_text |
| particpant_text | nominal | description of other participants |
| ima0505 | numerical | sum of activity in 10 minute window |
| weekday | numerical | weekday |

**Table 2. Description of KNMI Data Set**

| Variable | Comment |
|---|---|
| date | date |
| TG | daily mean temperature in 0.1 Celsius |
| FG | daily mean wind speed in 0.1 meter/second |

# 6. METHODOLOGY

## 6.1 Data Preparation

During the data preparation phase, the data has been transformed and combined to build a data set that was more suitable for analysis. The following changes were made.

1. The variable date has been split into the year, month and day. It was later only used to match the weather data to the right samples. It would make no sense to use the whole date as a variable in data mining since a date will never repeat. Using the month would make the most sense, but the data set covers each month only barely. Therefore it was also removed from consideration in the search for patterns.

2. The variable time has been split into the hour, minute and second. It only makes sense to use the hour to generally indicate the part of the day. Moreover, samples were taken approximately hourly.

3. The raw ima data (not ima0505) has been aggregated to an hourly mean, standard deviation, minimum and maximum.

4. The reported values for pleasure have been normalised per subject to create the variable pleasure_normalized. This way, any general positivity or negativity from any subject could not influence the results. The values were normalised by subtracting the mean and dividing by the standard deviation.

5. The variables TG and FG were combined to calculate the daily mean apparent temperature with the KNMI JAG/TI method [5].

## 6.2 Cortana Subgroup Discovery Tool

Cortana has 4 sections of settings: Data Set, Target Concept, Search Conditions and Search Strategy.

At the data set part, it is possible to change the interpreted type of a variable and one can enable or disable it for consideration in discovery. For every iteration of analysis, only a certain set of variables was enabled.

The target concept (or variable) varied between ima0505 and pleasure_normalized. However, both are numeric, therefore the Z-score was used as a quality measure. As mentioned earlier, Cortana also takes into account the size (described as coverage) of the subgroup in this quality measure. This has a similar result as multiplying the quality measure score with the probability of meeting the conditions; larger subgroups will have a higher score.

The default search strategy (beam search) was used. However, the search conditions were modified for every iteration of analysis.

- The refinement depth indicates the maximum number of variable value pairs in the conditions of a subgroup rule. It varied from 1 to 4. It was set up like this to ensure that the subgroup rules would still be meaningful.

- The minimum coverage indicates the minimum amount of data points that are needed to form a subgroup. This was set to 25 for refinement depth = 1 and it was set to 5 for refinement depth > 1. It was set up like this to prevent a few random outliers from being considered as an interesting subgroup.

Lastly, the subgroups were filtered based on a threshold of the quality measure. This was done by generating 500 random subgroup descriptions (conditions) and analysing the distribution of the quality scores on those. Then a threshold value was chosen based on a level of significance (at least 5%). The level of significance indicates the chance of finding a subgroup with at least the associated quality measure score at random in those 500 subgroups.

After everything was set up, the discovery algorithm would run and a list of subgroups came out. These subgroups were examined manually to see if any interesting insights could be obtained. The results of this are discussed in Section 7.1.

## 6.3 Clustering

Clustering has been applied to the variables ima0505 and pleasure_normalized to look for a relationship between the two. Both variables were previously scaled to the range of 0 to 1. This way a change in either of the variables would be equally meaningful. This ensures that the results from the clustering algorithms would be meaningful.

K Means clustering has been applied with various numbers of clusters to look for any interesting clustering of the data that could indicate a relation between the variables.

DBSCAN clustering was applied with a minimum distance of 0.05 between data points and a minimum of 5 samples to form a cluster.

# 7. RESULTS

## 7.1 Subgroup Discovery

The density distribution plots resulting from subgroup discovery can be found in Appendix A. Subgroup discovery was used to target physical activity and pleasure. Furthermore, the following two sections will discuss the corresponding results in greater detail.

### *7.1.1 Physical Activity*

The physical activity was represented as ima0505. The mean value for the physical activity in all 2288 samples of the data set was 15487 and the standard deviation was 16879. Table 3 describes the characteristics of each of the subgroup targeted at physical activity. The interpretation of the results is described in the following list. Numbers at the end of each item indicate the subgroup(s) on which they are based.

1. While outside or going for a walk as an activity or as transport, subjects were more active than average (1, 2, 3).
2. Subjects go for a walk as an activity or as transport with widely varying amounts of physical activity (2, 3).
3. While watching TV, subjects were less active than average (4).
4. While outdoors, with friends and before 16:00, subjects were more active than average (5).
5. While outdoors, with friends, on Friday, subjects were more active than average (6).

### *7.1.2 Pleasure*

The pleasure was represented as pleasure_normalized. The mean value for pleasure in all 2288 samples of the data set was 0 and the standard deviation was 1. Table 4 describes the characteristics of each of the subgroup targeted at pleasure. The interpretation of the results is described in the following list. Numbers at the end of each item indicate the subgroup(s) on which they are based.

1. While relaxing, with friends or walking as an activity, subjects were more pleased than average (7, 8, 9).
2. While doing the laundry, subjects were less pleased than average (10).
3. While with friends and at least slightly active, subjects were more pleased than average (11).
4. While outdoors, with friends, before 18:00 and at least slightly active, subjects were more pleased than average (12).

### *7.1.3 Combination*

By combining the results from the above two sections and selected other facts, more knowledge can be discovered.

The measurements were taken approximately during waking hours. This is also visible in Figure 1. It means that the condition "hour <= 18" of subgroup 12 implies "during the day". This reveals that being active during the day with friends is more pleasurable than average.

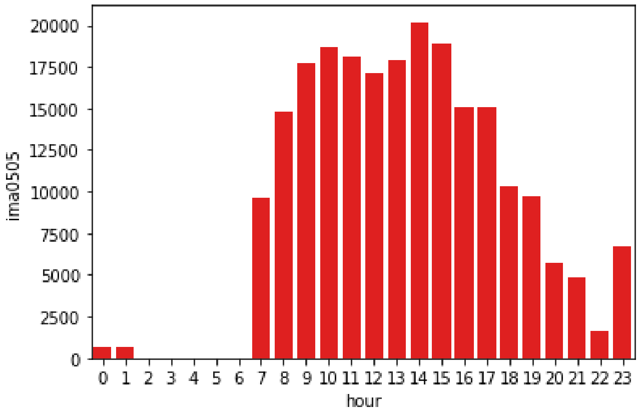

**Figure 1. Mean ima0505 throughout the day**

The average pleasure experienced in the 237 samples where subjects were watching TV was only 0.055559 with a standard deviation of 0.822837. That, combined with subgroup 4 and 9, reveals the following. On average subjects were more pleased while going for a walk or being active with friends than they were when watching TV. In those cases, they were also more active.

Subgroup 10 reveals, that doing the laundry is not very pleasurable. However, often that and similar household tasks just have to be done. Luckily, one does not only sit still when doing any household errands. Thus, if a subject is already quite inactive household errands do provide a way to stay active albeit not very pleasing.

After looking into the data set more, it was found that most of the 16 samples of subgroup 6 came from subject 3. In those samples subject 3 was extremely active. This also corresponds to what is found in our clustering results.

## 7.2 Clustering

As a baseline for the results of the application of clustering, Figure 2 shows a scatter plot of the data that was used, labelled per subject. It already reveals the distribution of the data.

Figure 3 shows the results of DBSCAN clustering and reveals a lot of outliers labelled as -1. There is one big cluster containing most of the data points and one smaller cluster above it. The smaller cluster mostly consists of samples from subject 3, as can be found by visually comparing Figure 2 and Figure 3. This cluster indicates a large amount of activity and seemingly above-average pleasure. These points correlate with our last comment on the results of subgroup discovery. Subject 3 was consistently very pleased and active while outside with friends on Fridays. Furthermore, there are no other insights to be obtained from Figure 3.

K Means clustering and analysis of its results did not provide any meaningful insights yet. Further work will be required here.

**Table 3. Description of Subgroups (Physical Activity)**

| Subgroup | Conditions | Coverage | Mean | Standard Deviation | Figure |
|---|---|---|---|---|---|
| 1 | location = outside | 426 | 23131 | 24235 | 4 |
| 2 | activity = waking as activity | 37 | 62136 | 43610 | 5 |
| 3 | activity = walking as transport | 33 | 45057 | 41952 | 6 |
| 4 | activity = watching TV | 237 | 5656 | 6610 | 7 |
| 5 | location = outside<br>participant = friends<br>hour <= 16 | 49 | 35572 | 36340 | 8 |
| 6 | location = outside<br>participant = friends<br>weekday = Friday | 16 | 49116 | 42861 | 9 |

**Table 4. Description of Subgroups (Pleasure)**

| Subgroup | Conditions | Coverage | Mean | Standard Deviation | Figure |
|---|---|---|---|---|---|
| 7 | activity = relaxing | 126 | 0.587237 | 0.763618 | 10 |
| 8 | participant = friends | 147 | 0.510508 | 0.744027 | 11 |
| 9 | activity = walking as activity | 37 | 0.884981 | 0.859319 | 12 |
| 10 | activity = doing the laundry | 96 | -1.107476 | 1.031741 | 13 |
| 11 | participant = friends<br>physical activity >= 3924.0 | 122 | 0.568315 | 0.733144 | 14 |
| 12 | location = outside<br>participant = friends<br>hour <= 18<br>physical activity >= 3924.0 | 109 | 0.584933 | 0.731156 | 15 |

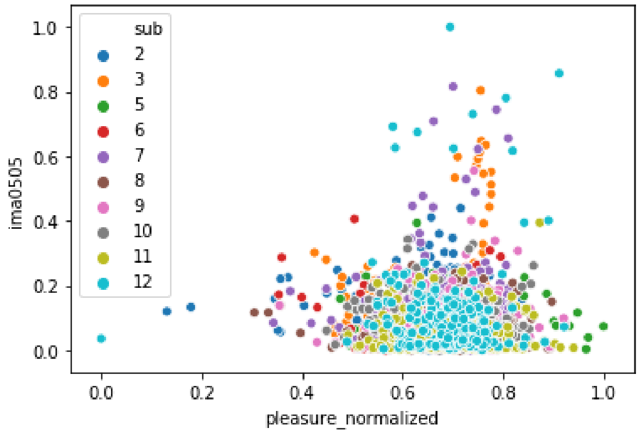

**Figure 2. Results of coloring by subject**

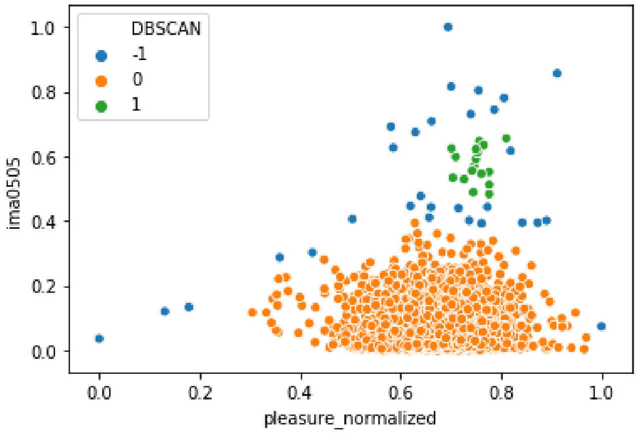

**Figure 3. Results of DBSCAN clustering**

## 8. CONCLUSION

In this project, three data mining methods were applied to find patterns in the physical activity and the pleasure of older adults. The techniques that were applied are subgroup discovery, K Means clustering and DBSCAN clustering. Subgroup discovery proved to be a useful technique to find patterns in the data, which are simple to interpret by non-experts. The subgroup discovery results provide interesting insights into the physical activity and pleasure of older adults and can be used in digital interventions systems to promote physical activity. Two examples of interventions that could be implemented based on the results are:

1. Encourage the user to go for a walk or to become active with a friend instead of watching TV. This will combine an increase in pleasure with an increase in physical.
2. Encourage the user to go for a walk with friends during the day instead of their regular daily activity at that time. This also combines an increase in pleasure with an increase in physical.

Both of these recommendations can increase the quality of life of older adults.

It was not possible to find great evidence of a relationship between the physical activity and the pleasure with clustering algorithms. However, it seems like being active is generally more pleasurable, unless the activity itself is inherently unpleasurable (think of household errands).

## 9. DISCUSSION AND FUTURE WORK

The results of this project show that a knowledge discovery study of behavioural data can provide insights into patterns in the physical activity and the pleasure of older adults. Due to time limitations, analysis using clustering methods was not as extensive as possible. However subgroup discovery can be of great use in constructing understandable and personalised digital health interventions to promote physical activity and increase the quality of life. Therefore, recommendations for future work are:

- Dedicate more time on the application of clustering techniques on physical activity data and contextual variables.
- Gather more data and contextual information to obtain more insights and draw more generalised conclusions using subgroup discovery.

## 10. ACKNOWLEDGEMENTS

The author of this paper would like to thank the main supervisor Dr. C.F. Pinho Rebelo de Sá and additional supervisor Dr. A.M. Meireles ferreira Cabrita of this project for all the support, feedback and constructive discussions during meetings and through e-mail. Also the author would like to thank family and friends who have proofread this paper.

## 11. REFERENCES

[1] M. Cabrita, R. Lousberg, M. Tabak, H. J. Hermens, and M. M. R. Vollenbroek-Hutten. An exploratory study on the impact of daily activities on the pleasure and physical activity of older adults. *European Review of Aging and Physical Activity*, 14(1), 2017.

[2] U. Fayyad. *Knowledge discovery in databases: An overview*, volume 1297 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Berlin Heidelberg, 1997.

[3] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):40,41,43–45, 1996.

[5] G. Groen. Wind chill equivalente temperatuur (wcet) knmi-implementatie jag/ti-methode voor de gevoelstemperatuur in de winter. Technical Report TR-309, Koninlijk Nederlands Meteorologisch Instituut, 2009.

[6] W. Hardeman, J. Houghton, K. Lane, A. Jones, and F. Naughton. A systematic review of just-in-time adaptive interventions (jitais) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1), 2019.

[7] F. Herrera, C. Carmona, P. González, and M. Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):496–497,499, dec 2011.

[8] K. N. M. Instituut. Weather data set. http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi.

[9] Jupyter. Jupyter lab: The jupyter user interace. https://jupyterlab.readthedocs.io/en/latest/.

[10] Jupyter. Jupyter: The interactive computing project. https://jupyter.org/index.html.

[11] M. Karekla, O. Kasinopoulos, D. Neto, D. Ebert, T. Daele, T. Nordgreen, S. Höfer, S. Oeverland, and K. Jensen. Best practices and recommendations for digital interventions to improve engagement and adherence in chronic illness sufferers. *European Psychologist*, 24(1):49,52–53,59, 2019.

[12] O. Kurasova, V. Marcinkevicius, V. Medvedev, A. Rapecka, and P. Stefanovic. Strategies for big data clustering. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, volume 2014-December, pages 740–747, 2014.

[13] N. Lavrač, P. Flach, and B. Zupan. *Rule evaluation measures: A unifying view*, volume 1634 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Berlin Heidelberg, 1999.

[14] S. Loomba and A. Khairnar. Fitness trackers market overview. https://www.alliedmarketresearch.com/fitness-tracker-market, mar 2018.

[15] J. McIntosh, S. Jay, N. Hadden, and P. Whittaker. Do e-health interventions improve physical activity in young people: a systematic review. *Public Health*, 148:140, 2017.

[16] K. Moreland and K. Truemper. *Discretization of target attributes for subgroup discovery*, volume 5632 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Berlin Heidelberg, 2009.

[17] L. I. of Advanced Computer Science Data Mining Group. Cortana: Open source subgroup discovery. http://datamining.liacs.nl/cortana.html.

[18] M. Omran, A. Engelbrecht, and A. Salman. An overview of clustering methods. *Intelligent Data Analysis*, 11(6):583, 2007.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] S. Perez. Runkeeper's new apple watch app lets you ditch your iphone when tracking your runs. https://techcrunch.com/2015/10/22/runkeepers-new-apple-watch-app-lets-you-ditch-your-iphone-when-tracking-your-runs/, oct 2015. Retrieved at January 22, 2020.

[21] PyData. Pandas: The python data analysis library. https://pandas.pydata.org/.

[22] PyData. seaborn: statistical data visualization. https://seaborn.pydata.org/.

[23] Python. Python: The programming language. https://www.python.org/.

[24] S. Stockwell, P. Schofield, A. Fisher, J. Firth, S. E. Jackson, B. Stubbs, and L. Smith. Digital behavior change interventions to promote physical activity and/or reduce sedentary behavior in older adults: A systematic review and meta-analysis. *Experimental gerontology*, 120:68–87, 2019.

[25] S. Wrobel. Inductive logic programming for knowledge discovery in databases. In *Relational Data Mining*, pages 74–101. Springer Berlin Heidelberg, 2001.

[26] Y. Zhou, A. Kankanhalli, and K. Huang. Predicting exercise behavior in fitness applications: A multi-group study. In *Thirty Eighth International Conference on Information Systems Proceedings*, pages 1–3. Association for Information Systems, dec 2017.

# APPENDIX

## A. RESULTS OF SUBGROUP DISCOVERY

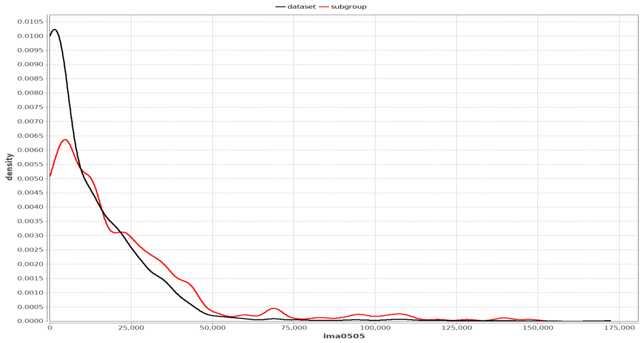

Figure 4. location = outside

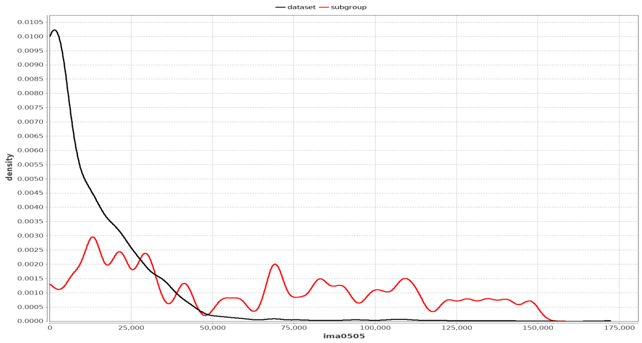

Figure 5. activity = walking as activity

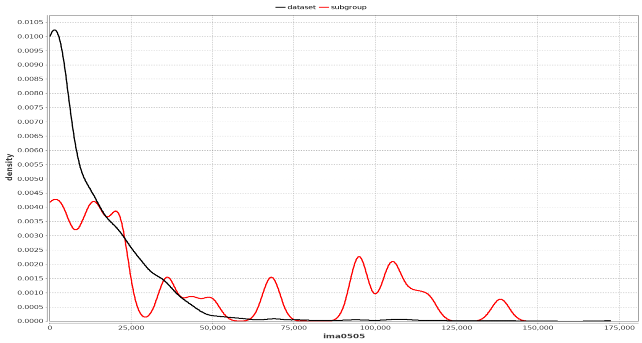

Figure 6. activity = walking as transport

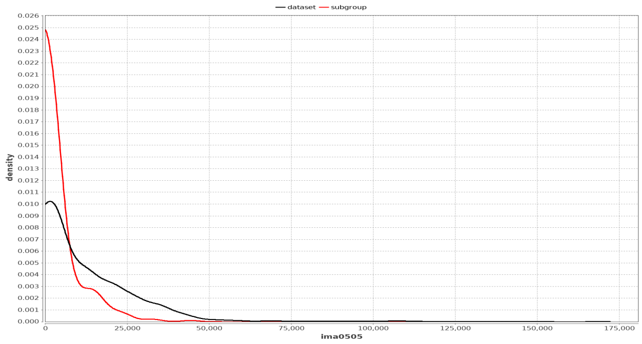

Figure 7. activity = watching TV

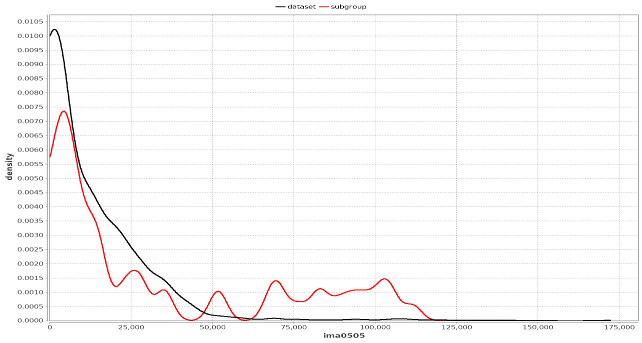

Figure 8. location = outside and participant = friends and hour <= 16

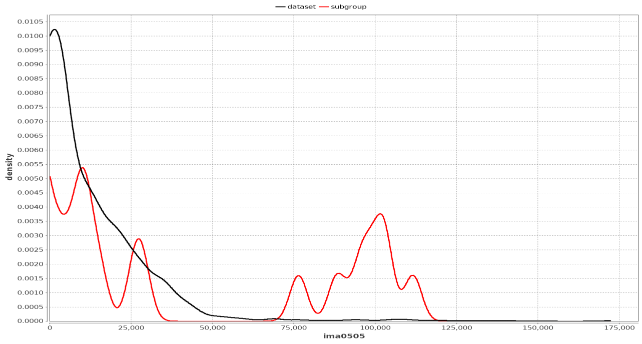

Figure 9. location = outside and participant = friends and weekday = Friday

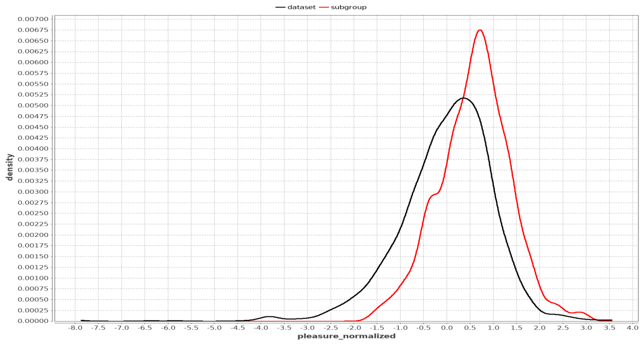

Figure 10. activity = relaxing

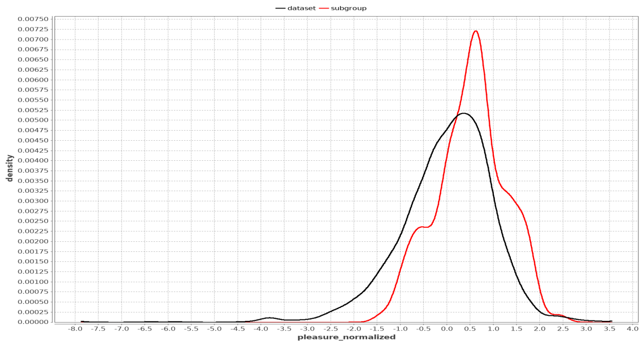

Figure 11. participant = friends

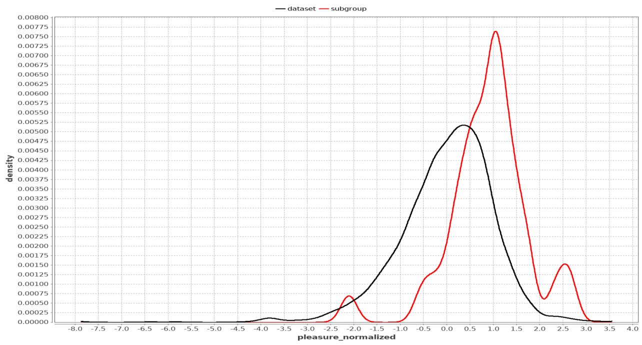

**Figure 12. activity = walking as activity**

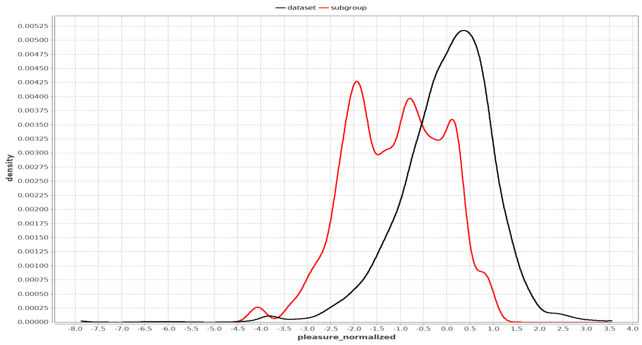

**Figure 13. activity = doing the laundry**

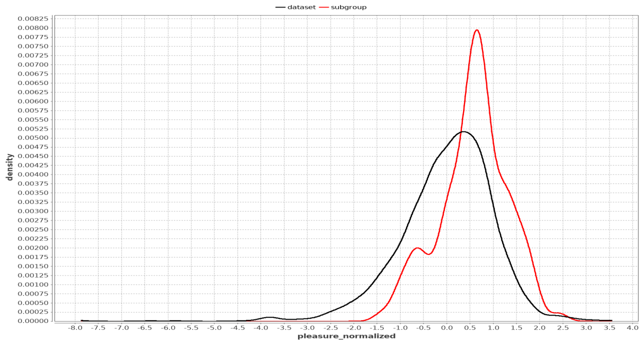

**Figure 14. participant = friends and physical activity >= 3924.0**

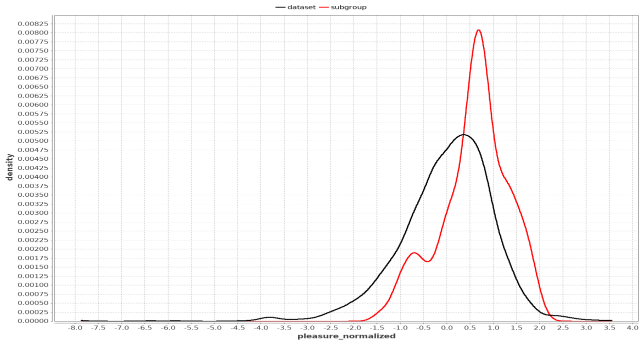

**Figure 15. location = outside and participant = friends and hour <= 18 and physical activity >= 3924.0**