# **Project**

Assigned: 25/11/2019

Due: 14/1/2019

## Dataset description

You are given two datasets with measurements from older adults collected within the [FrailSafe](http://frailsafe-project.eu/) project. The first dataset (beacons\_dataset.csv) contains information which is recorded daily with the use of smart beacon devices and concern the older people’s movement in their home setting. Each record of the dataset has the following fields:

* **part\_id**: The user ID, which should be a 4-digit number.
* **ts\_date:** The recording date, which follows the “YYYYMMDD” format (e.g., 14 September 2017, is formatted as 20170914).
* **ts\_time:** The recording time, which follows the “hh:mm:ss” format.
* **room:** The room which the person entered on the specific date and time. We assume that the person remained in the room till the next recording of the same day.

We show some entries in the following table as an example:

|  |  |  |  |
| --- | --- | --- | --- |
| **part\_id** | **ts\_date** | **ts\_time** | **room** |
| 2113 | 20170920 | 9:25:48 | Kitchen |
| 2113 | 20170920 | 9:26:03 | Entrance |
| 2113 | 20170920 | 9:29:25 | Entrance |
| 2113 | 20170920 | 9:53:09 | Bathroom |
| 2113 | 20170920 | 9:57:43 | Bedroom |
| 2113 | 20170920 | 9:58:51 | Entrance |
| 2113 | 20170920 | 9:59:03 | Kitchen |

The second dataset (clinical\_data.csv) contains information which was collected during the clinical evaluation of the older people from medical experts. This information represents the clinical status of the older person across different domains (physical, psychological, cognitive, etc). A list of the recorded clinical parameters and their description is shown in the table below:

|  |  |
| --- | --- |
| **parameter** | **Description** |
| fried | Categorization by Fried |
| hospitalization\_one\_year | Number of hospitalizations in the last year |
| hospitalization\_three\_years | Number of hospitalizations in the last three years |
| ortho\_hypotension | Orthostatic hypotension detection |
| vision | Vision |
| audition | Audition |
| weight\_loss | Unintentional weight loss |
| exhaustion\_score | Self-reported exhaustion |
| raise\_chair\_time | Lower limb strength |
| balance\_single | Single foot station (Balance) |
| gait\_get\_up | Timed Get Up And Go Test |
| gait\_speed\_4m | Speed for 4 meters’ straight walk |
| gait\_optional\_binary | Gait optional evaluation |
| gait\_speed\_slower | Slowed walking speed |
| grip\_strength\_abnormal | Grip strength outside the norms |
| low\_physical\_activity | Low physical activity |
| falls\_one\_year | Number of falls in the last year |
| fractures\_three\_years | Number of fractures during the last 3 years |
| fried\_clinician | Fried’s categorization according to clinician’s estimation |
| bmi\_score | Body Mass Index |
| bmi\_body\_fat | Body Fat (%) |
| waist | Waist circumference |
| lean\_body\_mass | Lean Body Mass |
| screening\_score | Mini Nutritional Assessment (MNA) screening score |
| cognitive\_total\_score | Montreal Cognitive Assessment (MoCA) test score |
| memory\_complain | Memory complain |
| mmse\_total\_score | Folstein Mini-Mental State Exam score |
| sleep | Reported sleeping problems |
| depression\_total\_score | 15-item Geriatric Depression Scale (GDS-15) |
| anxiety\_perception | Anxiety auto-evaluation |
| living\_alone | Living Conditions |
| leisure\_out | Leisure activities |
| leisure\_club | Membership of a club |
| social\_visits | Number of visits and social interactions per week |
| social\_calls | Number of telephone calls exchanged per week |
| social\_phone | Approximate time spent on phone per week |
| social\_skype | Approximate time spent on videoconference per week |
| social\_text | Number of written messages sent by the participant per week |
| house\_suitable\_participant | Subjective suitability of the housing environment according to participant’s evaluation |
| house\_suitable\_professional | Subjective suitability of the housing environment according to investigator’s evaluation |
| stairs\_number | Number of steps to access house |
| life\_quality | Quality of life self-rating |
| health\_rate | Self-rated health status |
| health\_rate\_comparison | Self-assessed change since last year |
| pain\_perception | Self-rated pain |
| activity\_regular | Regular physical activity |
| smoking | Smoking |
| alcohol\_units | Alcohol Use |
| katz\_index | Katz Index of ADL |
| iadl\_grade | Instrumental Activities of Daily Living |
| comorbidities\_count | Number of comorbidities |
| comorbidities\_significant\_count | Number of comorbidities which affect significantly the person’s functional status |
| medication\_count | Number of medication |

Special attention needs to be given to the “**fried**” parameter, which categorizes the older population into:

* Frail: Older adults which are vulnerable to stressors and have an increased risk of having a major (adverse) life event
* Pre-frail: Older adults which are moving towards frailty
* Non-frail: Healthy older adults

This categorization is generated by 5 of the above measurements, namely the **weight\_loss, exhaustion\_score, gait\_speed\_slower**, **grip\_strength\_abnormal,** and **low\_physical\_activity**.

## **Tasks**

### Part A

#### Preprocessing of the clinical dataset

You need to perform a number of preprocessing steps in the clinical dataset:

* **Convert nominal features to numerical**: As many of the classification and clustering algorithms need datasets with numerical data, you need to convert all nominal features to numerical ones. We give few examples here:
  + Yes/No 🡪 1/0
  + Frail / Pre-frail / Non-frail 🡪 2 / 1 / 0
  + Hears well / moderate / poorly 🡪 2 / 1 / 0
* **Remove erroneous values**: In some entries of the dataset you will find values which are erroneous (e.g., there are «999» and «test non applicable/adequate» values in some of the features). You should identify and remove these values (replace them with empty value) as they will affect the analysis results.
* **Handle missing values**: You need to handle the missing values in your dataset which were created by the previous step or existed from the beginning. You can adopt any strategy you think that fits best to your case such as:
  + Remove entries with missing values in some features
  + Remove features which have many missing values
  + Fill missing values of each feature with the average value of the feature.

#### Classification

Using the above preprocessed dataset, you need to perform classification analysis in order to predict the “fried” parameter. Take care not to include in the analysis the 5 parameters used for generating the fried categorization. You need to use at least one classification algorithm and show your results.

### Part B

#### Preprocessing of the beacons dataset

You need to perform a number of preprocessing steps in the beacons dataset:

* **Correct room labels**: The field “room” of the dataset doesn’t have predefined values and this results into having different strings describing the same room (e.g. you will see «Leavingroom», «Livingroom1», «Leavingroom» and «Sitingroom» values which all refer to the same room). You need to correct the dataset by making the labels homogenous (or as homogeneous as possible).
* **Remove erroneous users**: In the dataset description it is mentioned that the part\_id field is a 4-digit number. You need to remove any entries of the dataset which do not comply with this rule (e.g. «test»).
* **Generate features**: Your task is to generate a new dataset which will have one entry for each user. The entry will contain the percentage of the time the person has spent in the following rooms «Bedroom», «Bathroom», «Livingroom» and «Kitchen». As an example you might find that user 2113 spent 30% of his/her time in «Bedroom», 20% in «Bathroom», 15% in «Livingroom» and 30% in «Kitchen». This new dataset will be used in the next steps.

#### Merging the two preprocessed datasets

* **Merge datasets into one:** As a last preprocessing step, you need to combine the preprocessed clinical and beacons datasets into one. The merged dataset will contain one entry for each person for which there are both clinical and beacons data.

#### Clustering

* You need to apply at least one clustering algorithm on the final preprocessed dataset and evaluate the clustering using internal criteria (e.g., Silhouette index).
* Optionally, you can try some dimensionality reduction methods (e.g., PCA) which might lead to a better clustering (e.g., with higher Silhouette index).

### Part C

#### Data visualization and exploratory analysis

* Present the analysis results using relevant visualization tools.
* Perform an exploratory analysis on the clustering results, by observing the homogeneity of the cluster in terms of clinical parameters (e.g., is cluster 1 consisted mostly of Pre-frail older people?).

This project is modular. Students in the ΠΕΖ program should do only part A. Students in ΣΜΗΝ and BME programs should do parts A and B. Students in ΥΔΑ program should do all parts (A, B and C). You can form groups of 2. You should submit your final report along with any data (after preprocessing) and code you used/developed by email to vasilis@ceid.upatras.gr.