Project

Assigned: 2/11/2020

Due: 11/1/2021

Dataset description

You are given two datasets with measurements from older adults collected within the FrailSafe 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

1. 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.

2. 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

1. 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.

2. *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.

3. *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

1. *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 ΒΜΕ 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 and mactom1980@gmail.com.