Practical for IfI Master Students

Plan, conduct, and document an analysis of wearable sensor data

Topic of the Practical

Goal

Goal of this homework practical is to plan, conduct, and document an analysis of wearable sensor data. Following in data scientists' footsteps, it is your task to make interesting findings (find patterns) in such a dataset, validate these findings through means of creating evidence (hypothesis generation and validation), and write up your gained knowledge in a report.

Dataset Background

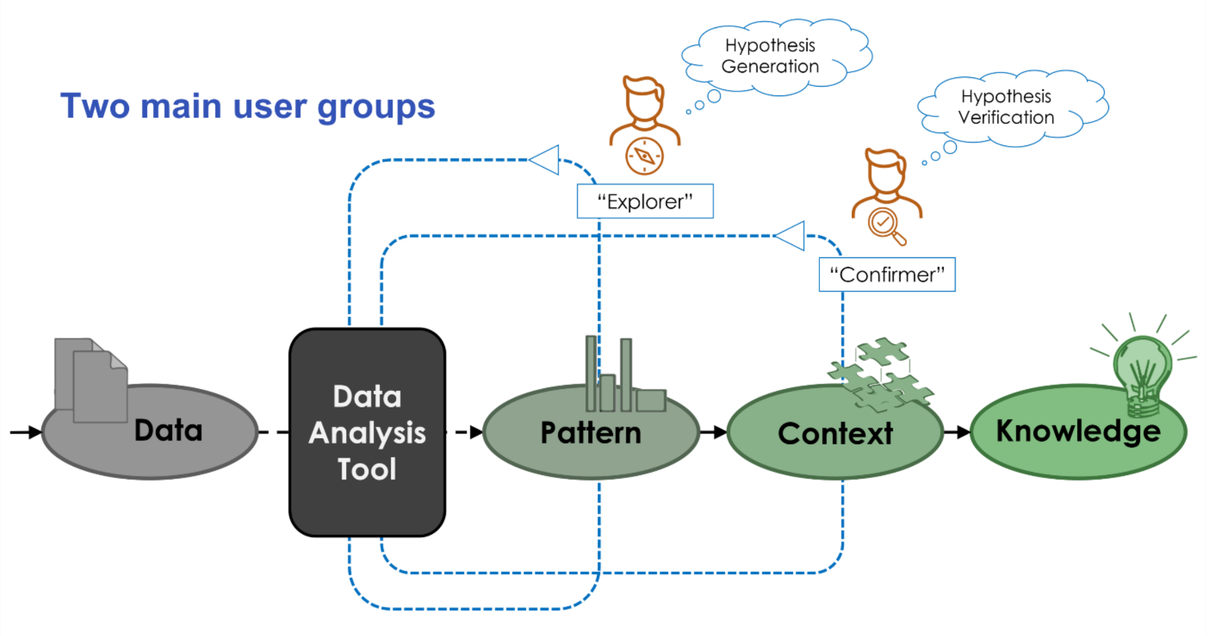
Students receive two files from a measurement study conducted in persons with an unspecified chronic disease. Symptoms commonly include severe tiredness (fatigue) and deteriorating physical fitness. The disease usually starts in adulthood and is not curable. This disease has different forms. Including a "past progressing" form. Most people with the disease progress from early to late-stage disease.

The data were collected using commercial wearable sensors in an intervention that included a strong emphasis on improving physical fitness, with a later follow-up phase. For all participants, the intervention phase varied between 1-3 weeks; the follow-up phase was 4 weeks. Some participants dropped out of the study early.

Expected Tasks

The analysis scenario is in line with the principal IVDA methodology presented in the seminar (see figure below). Based on the data collection, choose and execute data analysis and data visualization methods to identify patterns, making screenshots along the way. Your report on the knowledge you gain will complete the IVDA workflow. You should plan, execute, and document your analysis in a structured process:

1) **Familiarize yourself** with the data files by performing initial data explorations. It is highly recommended to use a statistical programming language such as R or Python, as well as to represent data and data findings visually.
2) **Include additional open-source data** that could provide context on the time series. For example: weather data, calendar of holidays and festivities, environmental data (pollen, pollution), economic data, key indicators of the SARS-CoV-2 epidemic. If the data are bound to a location, assume that all data were collected in the city of Zürich. Selecting and merging **at least one external dataset** is a mandatory step.
3) **Choose a research task**. You can choose between A) data exploration, or B) confirmatory analysis of a pre-specified hypothesis. Specific requirements for each task are described below. Note: you are asked to do **either** A) or B).



Choice A: Data Exploration

You are expected to analyze "the unknown", e.g., by following the information-seeking mantra: "overview first, zoom and filter, details on demand" (see slides). You will certainly identify patterns in the data, like common progressions, outliers, or trends across time. In the knowledge generation process, you succeed when you can represent these patterns visually using the visual mapping principles learned in the seminar. You must present **at least three entirely different findings**, with each finding presented as an individual visualization that visually prioritizes that finding.

Choice B: Confirmatory Analysis

You are expected to form hypotheses about the dataset in advance, and then determine whether these hypotheses can actually be validated with the data at hand, or have to be rejected. Note that a result that rejects a hypothesis can still be a strong analysis result, especially if its rejection is quite surprising. You are required to **present at least three entirely different hypotheses**, each with confirmatory data analysis settings and result proposals:

1) Start by thinking about interesting hypotheses in the data. For each hypothesis, pick a sensor signal and postulate a plausible hypothesis for how the signals may be different according to personal-, disease-, or measurement-context differences. By context, we mean when and where a datapoint was collected, as informed by the external information you will add to the database.
2) Follow the PICO design steps to plan your analysis:
    a. P: Population; which participants will you include? A subset? All participants?
    b. I: Intervention: will you analyze the effect of the intervention as described in the background? If yes, your time-series analysis should distinguish and compare between the intervention- and the follow-up phase. You can also choose to ignore these phases.
    c. C: Comparator: which groups or factors will you compare?
    d. O: Which sensor outcome will you compare? Your outcome may require pre-processing and further aggregation, depending on your hypothesis.
3) Once you have defined your PICO, refine your analysis plan by specifying the statistical methods (e.g. regression analysis, t-tests, Chi^2-tests), as well as how you will deal with missing information, and what "measures" and "significance levels" you will use to reject

or accept your initial hypothesis (e.g. p-values). Think about how you will present the information. On the one hand, this includes describing the participants and sensor measurements in a table. On the other hand, you should use data visualizations for each tested hypothesis, to visually illustrate the evidence you find from your analysis (to reject or accept your hypothesis).

Bonus for Grading

Students who include one of the following four machine learning modeling techniques in their analysis approach receive a 0.25-grade bonus: Clustering, Dimensionality Reduction, Classification, or Regression. The (output of the) algorithmic models must be shown with a visualization, so that the results are visually observable.

File Sensors

```
-----------------------------------------------------------------------------------------------------------
------------------------
time
Date/Time of measurement
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  string (str19)

         unique values:  7,464                    missing "":  0/59,296

              examples:  "2021-04-01 22:00:00"
                         "2021-05-22 08:00:00"
                         "2021-07-07 13:00:00"
                         "2021-09-15 07:00:00"

               warning:  variable has embedded blanks

-----------------------------------------------------------------------------------------------------------
------------------------
steps
Steps
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  numeric (int)

                 range:  [0,6123]                     units:  1
         unique values:  2,857                    missing .:  0/59,296

                  mean:   296.029
              std. dev:   555.523

           percentiles:        10%       25%       50%       75%       90%
                                 0         0        75       360       825

-----------------------------------------------------------------------------------------------------------
------------------------
sleep
Sleep (minutes)
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  numeric (byte)

                 range:  [0,60]                       units:  1
         unique values:  61                       missing .:  0/59,296

                  mean:   17.8363
              std. dev:   26.4667

           percentiles:        10%       25%       50%       75%       90%
                                 0         0         0        60        60

-----------------------------------------------------------------------------------------------------------
------------------------
heartrate                                                                           HeartRate (as beats per
minute, hourly average)
-----------------------------------------------------------------------------------------------------------
------------------------
```

```
                  type:  string (str16)

         unique values:  11,916                   missing "":  0/59,296

              examples:  "62.0535714285714"
                         "70.2413793103448"
                         "77.9833333333333"
                         "87.2"

-------------------------------------------------------------------------------------------------------------
------------------------
id
(unlabeled)
-------------------------------------------------------------------------------------------------------------
------------------------

                  type:  numeric (float)

                 range:  [1120,9926]                  units:  1
         unique values:  44                       missing .:  0/59,296

                  mean:    5615.2
              std. dev:   2534.05

           percentiles:        10%       25%       50%       75%       90%
                              2130      3389      5977      7928      9085
```

File Markers

```
-----------------------------------------------------------------------------------------------------------
------------------------
record
Record (not relevant)
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  numeric (byte)

                 range:  [1,45]                       units:  1
         unique values:  44                       missing .:  0/44

                  mean:   22.9773
              std. dev:   13.2849

           percentiles:        10%       25%       50%       75%       90%
                                 5      11.5      22.5      34.5        41

-----------------------------------------------------------------------------------------------------------
------------------------
age
Age
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  numeric (int)

                 range:  [1958,2005]                  units:  1
         unique values:  25                       missing .:  0/44

                  mean:    1977.8
              std. dev:     8.706

           percentiles:        10%       25%       50%       75%       90%
                              1966    1972.5      1978    1982.5      1986

-----------------------------------------------------------------------------------------------------------
------------------------
sex
(unlabeled)
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  string (str6)

         unique values:  2                        missing "":  0/44

            tabulation:  Freq.  Value
                            29  "Female"
                            15  "Male"

-----------------------------------------------------------------------------------------------------------
------------------------
diseasetype
Type/Stage of Disease
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  string (str24)

         unique values:  3                        missing "":  0/44

            tabulation:  Freq.  Value
                            18  "Early Disease Stage"
                             7  "Fast Disease Progression"
                            19  "Late Disease Stage"

               warning:  variable has embedded blanks

-----------------------------------------------------------------------------------------------------------
------------------------
id                                                                                              Participant
Id, Link to Sensor file
-----------------------------------------------------------------------------------------------------------
------------------------

                  type:  numeric (int)

                 range:  [1120,9926]                  units:  1
         unique values:  44                       missing .:  0/44

                  mean:    5782.8
              std. dev:   2469.32

           percentiles:        10%       25%       50%       75%       90%
                              2589      3734      6066      7962      9085
```

Written Report

Result Submission: Report

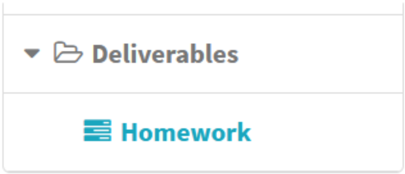
You are expected to write the results of your studies in a report. Once again: you should elaborate on two exploratory data analysis findings, *as well as* two hypothesis-driven confirmatory data analysis scenarios. The report should be compact, but still complete. **Please do not exceed six pages of written text**. Note: **Figures, Tables, and the references (at the end of the report) do not count for page length**, meaning that the size of your report will be longer than the six pages of written text. Create and include as many figures in your report as possible; figures will have a considerable effect on grading, as creating data visualizations is part of the seminar focus. The structure of the report should be as follows:

0) Abstract. Briefly describe your overall approach in a few sentences, including goals/tasks, and findings (structured as background, objective, methods, results, conclusions)
1) Introduction. Describe the context, the motivation/goal of your analysis, and which information needs to be addressed. Tell which methods you applied, and which results you have uncovered. Also, describe shortly how you have validated your findings (e.g., through visualizations, errors of regression models, p-values, etc).
2) Background and Related Work. You may want to tell the reader more about the domain background, other related studies (related work), and data science approaches similar to yours. We expect you to include at least five carefully selected references.
3) Methods. Every scientific approach requires a systematic description of how studies were conducted. Write down HOW you arrived at your results, and which methods/models/techniques/experiments/experiment designs you used. With this information, other researchers will a) understand what you did, b) justify that you did a good job, and c) be able to replicate your study.
4) Data Characterization. Tell the reader more about the dataset used (dataset characterization) and analysis tasks you have executed. Include information about the dataset size, the number of attributes, data quality aspects. Don't forget to relate the dataset to the background you are studying (and that you have described in the Background and Related Work of your report)
5) Implementation and Sofware Use. Tell the reader more about your "technology stack". Describe through which tools/programming languages/architectures your practical was realized. A few sentences will suffice.
6) Results. As you see, it is important to separate between how you did it (the process) and what you found out (results). Use tables, numbers, and visualizations to show your results. Do not forget to show the evidence/significance/strengths/etc. of the results/knowledge you achieved.
7) Discussion. Summarize the most important results and put them into an (international) context. That means: tell readers how your results align with the results of other studies, but also in which ways your results disagree. The discussion section is also the right place to be critical with your own approach.
8) Limitations. Every good study contains a section about the limitations of the study and problems one had to encounter during analyses. Tell, e.g., under which conditions your results may not be valid, or what was impossible to assess due to data problems, etc. Or: which simplifying assumptions needed to be made to conduct your analyses?
9) Conclusions: add advice to subject experts/domain experts based on your finds. Summarize and point out future work.
10) References. This appendix lists the references you have used in a structured way. Make use of your writing environment (e.g., Overleaf + bibtex, or Word). You should not need to hand-craft references in the list at the end of your document.

Homework Submission

Submit your Report to OLAT. You will find a Deliverable called "**Homework**".

Deadline is May 31 2024, 11:59pm/23:59.



Wish you the best of success with the homework!