Homework 10: First steps of the project

KAGGLE - STRESS DETECTION: Predicting stress levels based on personality traits, sleep quality, movement data and mobile phone use.

Project repository: https://github.com/liisa-j/IDS-project

TEAM: Liisa Jullinen (Group 7)

Business understanding

Business goals

According to WHO, mental disorders accounted for 13% of the total global burden of disease in 2004 (with depression alone accounting for 4.3% and being among the largest single causes of disability worldwide - 11% of all years lived with disability globally), particularly for women[1]. The economic consequences of these health losses are large: estimated global impact amounting to 16.3 trillion USD between 2011 and 2030[2]. Due to the COVID pandemic, both anxiety disorder and depression prevalence increased even more (26% and 28% respectively), leaving about 246 million people suffering from major depressive disorder and 374 million from anxiety disorders[3]. These numbers are also reflected locally in Estonia[4,5].

Among other factors that increase the likelihood of obtaining a mental disorder, stress can lead to many physical and mental disorders, especially anxiety disorders and depression[6]. According to many

[1] Comprehensive mental health action plan 2013–2030. Geneva: World Health Organization; 2021 (https://iris.who.int/bitstream/handle/10665/345301/9789240031029-eng.pdf?sequence=1)

[2] World Economic Forum, Harvard School of Public Health. The global economic burden of non-communicable diseases. Geneva: World Economic Forum; 2011 (https://www.weforum.org/publications/global-economic-burden-non-communicable-diseases/)

[3] World mental health report: transforming mental health for all. Geneva: World Health Organization; 2022 (https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf?sequence=1)

[4] Eesti rahvastiku vaimse tervise uuringu konsortsium (2022). Eesti rahvastiku vaimse tervise uuringu lõpparuanne. Tallinn, Tartu: Tervise Arengu Instituut, Tartu Ülikool: 12 (https://tai.ee/sites/default/files/2022-06/Eesti%20rahvastiku%20vaimse%20tervise%20uuring.pdf)

[5] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

[6] Eesti rahvastiku vaimse tervise uuringu konsortsium (2022). Eesti rahvastiku vaimse tervise uuringu lõpparuanne. Tallinn, Tartu: Tervise Arengu Instituut, Tartu Ülikool: 50. (https://tai.ee/sites/default/files/2022-06/Eesti%20rahvastiku%20vaimse%20tervise%20uuring.pdf)

studies, smartphone (especially problematic) usage increases stress level of the user[7,8]. It has also been established that Big Five personality traits (extraversion, agreeableness, openness, conscientiousness, neuroticism) have an effect on how they perceive stressors (people higher on neuroticism perceive events as highly stressful and uncontrollable, whereas those higher on conscientiousness, extraversion, and openness tend to perceive such events as within their control)[9], thus mediating their stress response. It is also known that poor sleep quality[10] and lack of physical activity are highly correlated with higher stress levels, whereas personality traits act as mediating factors[11].

The goal of this analysis is to combine several of the factors mediating stress levels into one model and see how personality traits, movement data, mobile phone usage data and data on sleep patterns predict stress levels. The possible results should imply which personality types are most vulnerable to different factors causing stress and which of the factors have the biggest effect on stress levels. These insights gained could be used as a basis for preventive campaigns for mental health (promoting positive mental health hygiene) and possibly doing targeted campaigns.

Assessing the situation

The data used for this project originates from Kaggle and is publicly available. The dataset has been compiled recording various behavioural and psychological features of 100 participants (during 30 days) and consist of 3000 rows and 20 features. The features include scores of the Big Five personality traits, sleep patterns (including the Pittsburgh Sleep Quality Index (PSQI) score), screen time and phone use statistics, movement data (mobility radius, accelerometer data, distance) and Perceived Stress Scale score and skin conductance as stress level measures.

Data will be analysed with relevant Python libraries and results with relevant code will be presented within a Jupyter notebook.

The deadline for finishing the analyses is 9.12. No legal and security requirements are relevant.

Most important risks are associated with data sufficiency (among possible measures are either generating data synthetically or getting more data, which could be complicated taking into account the goals and the amount of the features necessary to achieve the goals).

[7] Longitudinal Effects of Excessive Smartphone Use on Stress and Loneliness: The Moderating Role of Self-Disclosure. Karsay, K., Schmuck, D., Matthes, J., Stevic, A. Cyberpsychology, Behavior, and Social Networking 2019 22:11, 706-713. (https://doi.org/10.1089/cyber.2019.0255)

[8] The association between smartphone use, stress, and anxiety: A meta-analytic review. Vahedi, Z., Saiphoo, A. Stress and Health. 2018: 1–12. (https://doi.org/10.1002/smi.2805)

[9] Characterizing stress processes by linking big five personality states, traits, and day-to-day stressors. Ringwald, W., R., Nielsen, S.R., Mostajabi, J., Vize, C.E, Berg, T., Manuck, S.B., Marsland, A.L., Wright, A.G.C. Journal of Research in Personality. 2024: 110. (https://www.sciencedirect.com/science/article/abs/pii/S0092656624000357?via%3Dihub)

[10] Alotaibi, A. D., Alosaimi, F. M., Alajlan, A. A., & Bin Abdulrahman, K. A. (2020). The relationship between sleep quality, stress, and academic performance among medical students. Journal of family & community medicine, 27(1), 23–28. (https://doi.org/10.4103/jfcm.JFCM_132_19)

[11] Hegberg, N.J., Tone, E.B. 2015. Physical activity and stress resilience: Considering those at-risk for developing mental health problems. Mental Health and Physical Activity. 8, 1-7. (https://doi.org/10.1016/j.mhpa.2014.10.001)

Terminology: Big Five personality traits (extraversion, agreeableness, openness, conscientiousness, neuroticism).

Costs and benefits: since it has been established that every $1 invested in scaled-up treatment for depression and anxiety, there is a $4 return in better health and productivity[12], it is obvious that prevention could yield even bigger gains.

Data-mining goals

The goal is to find the best suited machine learning model with the highest predictive values for both variables indicating stress response.

The data will be divided into 70-30 train-test split and the models will be evaluated by using explained variance, mean squared error and R^2.

Data understanding

The data is publicly available on Kaggle.com in a csv format. The dataset consists of 3000 rows and 20 features that correspond to the expectations described above (including features that refer to the Big Five personality traits, sleep patterns, screen time and phone use statistics, movement data and Perceived Stress Scale score and skin conductance as stress level measures).

The dataset consists of the following columns/features: **participant_id** (unique identifier for each participant; Data type: Integer; Range: 1 to 100 (as there are 100 participants)), **day** (The day of observation for each participant; Data type: Integer; Range: 1 to 30 (each participant is observed over 30 days)), **PSS_score** (Perceived Stress Scale score, measuring stress levels; Data type: Integer; Range: 10 to 40), **Openness** (Measure of openness to experience, a personality trait; Data type: Float; Range: 1.0 to 5.0), **Conscientiousness** (Measure of conscientiousness, a personality trait; Data type: Float; Range: 1.0 to 5.0), **Extraversion** (Measure of extraversion, a personality trait; Data type: Float; Range: 1.0 to 5.0), **Agreeableness** (Measure of agreeableness, a personality trait; Data type: Float; Range: 1.0 to 5.0), **Neuroticism** (Measure of neuroticism, a personality trait; Data type: Float; Range: 1.0 to 5.0), **sleep_time** (The time (in hours) the participant went to sleep; Data type: Float; Range: 5.0 to 9.0 hours), **wake_time** (The time (in hours) the participant woke up; Data type: Float; Range: 5.0 to 9.0 hours), **sleep_duration** (The duration (in hours) the participant slept; Data type: Float; Range: 6.0 to 9.0 hours), **PSQI_score** (Pittsburgh Sleep Quality Index (PSQI) score, measuring sleep quality; Data type: Integer; Range: 1 to 5), **call_duration** (Total duration of phone calls for the day (in minutes); Data type: Float; Range: 0 to 60 minutes), **num_calls** (Number of phone calls made during the day; Data type: Integer; Range: 0 to 20 calls), **num_sms** (Number of SMS messages sent during the day; Data type: Integer; Range: 0 to 50 messages), **screen_on_time** (Total screen-on time for the day (in hours); Data type: Float; Range: 1.0 to 12.0 hours), **skin_conductance** (Measure of skin

[12] Mental health matters. 2020. The Lancet Global Health, Volume 8, Issue 11, e1352. (https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(20)30432-0/fulltext#:~:text=The%20economic%20case%20for%20investment,in%20better%20health%20and%20productivity.)

conductance, indicating arousal or stress response; Data type: Float; Range: 0.5 to 5.0 µS (microsiemens)), **accelerometer** (Accelerometer data representing physical movement; Data type: Float; Range: 0.1 to 2.5 g (g-force)), **mobility_radius** (The radius of mobility for the participant (in kilometers); Data type: Float; Range: 0.1 to 1.5 km), **mobility_distance** (Total distance moved during the day (in kilometers); Data type: Float; Range: 0.5 to 5.0 km).

There are no missing values in the dataset (assuming it has been pre-processed to some extent previously). Stress features, personality traits, sleep quality and mobile use statistics seem to have adequate measures and ranges. When looking at the distributions of mobility features (mobility_radius, mobility_distance) they seem intuitively very low (maximum mobility radius being 1.5 km and distance 5 km). This raises a lot of questions about data validity and perhaps it would be better not to include the mobility readings in the analyses (since all of the participants seem to fall into the sedentary category). It is also similarly hard to logically interpret the accelerometer data (g-force).

Thus, most probably only 17 of the original features will be used (excluding movement/mobility data).

The data quality is however sufficient for most goals of this project (excluding mobility data).

Planning your project

1. Data preparation: cleaning, constructing data, integrating and formatting data. In this phase the abovementioned problems will be addressed, the unnecessary features removed, additional synthetic data generated, necessary data transformations conducted. Some of the features will be combined into new features and possibly also clustering of the participants according to their personality traits will be conducted.
2. Considering additional/ alternative data sources. Since this is a circular process and already thus far some limitations have occurred with the data selected, returning to data selection (or additional data) should definitely be considered to fulfil the “research” goals of this project.
3. Splitting data into train and test data.
4. Possible models will be selected and listed. Tests will also be outlined, in order to determine how well the models worked. Models will be built and trained on the data. Models will be evaluated.
5. Evaluating the success of the process, possibly returning to some previous steps.

All of the relevant parts of the whole process will be documented and the workload will be approximately 30 hours.