-Confidential-

May 2022

Example Project – Genetika+

We have access to data sets from patients who have completed a clinical trial during which they were administered a variety of antidepressant medications. The data includes their patient history prior to treatment, as well as parameters measured during and after drug administration.

The provided data of this exercise is a mockup for the real-world data being processed at Genetika+.

Data Description:

* The data set consists of 1953 patients.
* The analysis consists of > 40 potential predictor fields (fewer can be used if required).
  + File: 1\_demographics\_GenetikaExercise\_Confidential.xlsx
  + File: 2\_symptoms\_GenetikaExercise\_Confidential.xlsx
  + File Format:
    - Sheet 1 – data
    - Sheet features: feature types description
    - Sheet code\_dictionary: Encoded values of the features.
* Definition of response: reduction of DEPRESSION SCORE PERCENT IMPROVEMENT scoring > 50% at **any given week**. (see file 3\_dependant\_GenetikaExcercise\_Confidential.xlsx)

Task:

Q1. Provide predictive model of identifying responders vs non-responders, using the dependent features, or the best possible subset of them. Your report should describe the analysis process, feature selection approach, the model you chose to use and the evaluation methods you used.

Q2. Please suggest any additional tools or approaches you would take if you had additional time or could access additional data about these patients?

Deliverables:

* Please provide a word/pptx document to summarize your conclusions.
* The file containing the code used to generate these conclusions, and any other input/processed data you used.

Sincerely,

Sasha Kugel

Genetika+