p8105_fp_pipeitup

pipeitup

11/6/2019

Final Project Proposal: Prevalence trend and Determinants of Mental Illness in Tech Company

Yingxi Ji (yj2579) Yeqing Ji (yj2580) Wurongyan Zhang (wz2507) Yiling Yang (yy3018) Yanni Wang (yw3439)

Mental illnesses are common in the United States. According to the National Institute of Mental Health, in 2017, about one in five U.S. adults live with mental illness. As future technicians, we are wondering the prevalence and determinants of mental illness in the tech companies, which has barely been extensively studied. The anticipated dataset is OSMI Mental Health Survey in Tech Survey, which can be found at https://osmihelp.org/research. This survey is an annual survey regarding mental health of people in tech companies conducted by Open Source Mental Illness (OSMI), a non-profit organization. We will employ the most recent 3 years' data (i.e. 2016-2018). With such dataset, we particularly are interested in: Examining the changes in prevalence of mental illness in recent 3 years / across regions / gender / age / complications between mental illness? Does your employer provide resources to learn more about mental health issues and how to seek help reflects the society's attitude towards mental illness? Building classification model to predict mental illness status. Planned analyses begin with data cleaning and EDA. This step includes visualization, possibly interactive, by using graphs such as bar graphs, correlation matrix and heat map. Proportional hypothesis test and logistic regression are followed. An interactive map might be considered. To build the classification model, decision tree classification tree diagram and random forest are proposed. Expected coding challenges are mostly during the cleaning process since there are many categorical data with NA values or free response questions, which cannot be handled by simple factorization. The intended outputs include: 1) An analysis report, 2) A web page containing report and classification model, and 3) A video demonstration of webpage and report.

We propose a timeline as follows:

Nov 15 Data cleaning and EDA Nov 22 Finish interested questions Dec 1 Finish formal report Dec 4 Finish webpage

The anticipated data sources

OSMI Mental Health Survey in Tech Survey https://osmihelp.org/research The planned analyses / visualizations / coding challenges

Planned analyses:

EDA, proportional hypothesis test, logistic regression, decision tree, random forest

Coding challenges:

Cleaning data will be challenges since the data is categorical outcomes. Some of the data were entered by participants themselves so the data include the problem of misclassification. # Visualizations: using plot.ly to draw Bar graph, heat map, correlation matrix, region map, classification tree