our dataset:

<https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016>

OPEN SOURCING MENTAL ILLNESS, LTD

OSMI Mental Health in Tech Survey 2016

Currently over 1400 responses, the ongoing 2016 survey aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers.

How Will This Data Be Used?

We are interested in gauging how mental health is viewed within the tech/IT workplace, and the prevalence of certain mental health disorders within the tech industry. The Open Sourcing Mental Illness team of volunteers will use this data to drive our work in raising awareness and improving conditions for those with mental health disorders in the IT workplace.

file name: machine\_learning\_model

pre\_encoded\_survey:

32 columns indexed 0 to 31

column 8 and 26 contains int data

mental\_cat: categorical data from mental\_df

encode\_df: after using one hot encoder on mental\_cat

after merging encode\_df and dropping mental\_cat from mental\_df

mental\_df contains 122 columns

index dropped from encoded\_survey because sqlite adds index automatically

<next steps are model fitting, train test split etc>

file name: cleaning\_data

reads original csv file mental-heath-in-tech-2016\_20161114

1433,63 shape

drop low\_response\_columns columns with <70% responses

mental\_df becomes 1433,48 shape

many\_distinct\_columns: columns with more than 10 uniques

remove age, country of work, country of residence and age column from many\_distinct\_columns

drop many\_distinct\_columns from mental\_df

gender\_counts: count of gender types

for gender types with count less than 20, replace gender by ‘other’. gender counts now become

Male 1047

Female 286

Other 97

same for country of residence. replace countries with less than 20 count by other

same for country of work

chi-square test:

for reference -

<https://www.khanacademy.org/math/ap-statistics/chi-square-tests/chi-square-goodness-fit/v/goodness-of-fit-example>

<https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test>

used in feature selection. Previously we were using random forest for feature selection

Chi-square is most commonly used by researchers who are studying survey response data because it applies to categorical variables.

<we need to find a research paper related to chi square test. especially the one that says that a chi square test is good for data obtained from surveys>

chi\_square : returns p value, user defined

functions used(in built):

crosstab: used to compute a simple cross-tabulation of two (or more) factors. computes a frequency table of the factors by default( default case used in our code)

stats.chi2\_contingency:

This function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table observed.

if p value is >0.5 the feature becomes insignificant

<how?, I’m confused>

insignificant columns dropped from mental\_df

shape becomes (1433, 32)

rows with nan values dropped using dropna()

shape becomes (861, 32)

replace(",", ""): removes ‘,’

data frame saved to a csv file cleaned\_data.csv

using to\_csv