Analysis of Possible attributes contributing towards mental disorder in occupational health

Intro

Python (Open Sourcing Mental Illness (OSMI) Survey on Mental health of Workplace

risk factors like :- family history, age, discuss_ coworker discuss_ employer, Sought treatment with ~~with~~ with mental health disorder.

Data cleaning :- Column names were too big & this restricted importing the data into MySQL. * renamed the column names into shorter abbreviations

Data was mostly categorical, converted to numerical for better visualization.

(OSMI) open sourcing mental illness

Total 123 variables → 23 variables aligned with Mental health was selected

Data import to phpmyadmin

Shared data base 'I50\sp2\proj3_db' in (Project)

~~Data tables~~

Data cleaning was done in SQL

Initially the Columns names were too big and of ~~f~~ changed using SQL Queries

ALTER Tables MH Data change.

After removing changing the Column names

Deleted : null values

Delete From MH_Data

WHERE COALESCE (all the Variables) is' NULL

Then

UPDATE the Gender using SQL Query

(SET)

like cis female, cisgender, female, female cis, female ish, trans woman, transfeminine, woman,
my sex is female to female using update

UPDATE MHDATA SET GENDER = 'male', where Gender =
cis heteromale, Cismale, dude, M, male, male cis,
male-ish, androgynous, man,

After Data cleaning

MySQL → python (using mysql Connector)
↓
Correlation matrix (positive correlation between
MHD - currently & family history
Negative correlation (discuss coworker discuss employer,
Sought treatment)
↓ Heatmap

Exploratory Data analysis

Normally distributed, outliers etc

feature analysis

(4) By taking all the columns as features performing
4 ~~randomly~~ features → discuss employer, discuss coworkers, Sought treatment, family history

Steps to identify the most significant features in the dataset

Step-1
* independent features → as 'X'
Y → MHD - currently (predicted) attribute to be

*
Y Step-2
Logistic regression (Categorical outcomes variables)
To perform logistic regression → feature Selection is key

Apply logistic regression on all features showed accuracy of 67.7%

SCFit-learn → Feature importance

Step-1:

Independent features as 'X' and MHDcm as 'y' (attribute to be predicted) then performed logistic regression using all the features of independent variables, then the result showed accuracy of 67.7% (ROC Curve (Sensitivity vs Specificity TPR (false positive

Step-2 :- Then plotted feature importance to analyse most Correlated & significant features in the dataset in python (the '0' value shows no effect where - or +ve high valued variable shows significant feature Importance). Looking at the values features like Discuss employes, discuss-coworks, Sought-treatment & family History has more factors

Step3:- Then performed logistic regression taking top 4 features as independent 'X' variable & MHDcurrently as 'y'. I used SCFit-learn package) &, to import machine learning models. SKlearn linear model imported logistic regression

Step 4:- Assign the input variable (X) with 4 independent features & y → MHD-currently.

Training the model using fit() (train-test-) 25% of data as test set and 75% train set

logistic regression model showed I- 68.7%

& Simultaneously plotted ROC curve

ROC (Receiver operator characteristics) (How model thinks works)

Family-history → Higher feature importance

Confusion matrix

to analyze the performance of the Ada-Boost model

precision = Total positive / Total ~~negatives~~ true + False pos

Data visualization

Scatter plot - plotly

(TPR-FPR)

average good line data conforms ideal ROC curve

(Goodness

flipping coin

dep → with ind → Sought treatment

Middle aged people seek treatment seek treatment

multiple Bar Chart

↳ pivot table

dep - MHD currently, discuss_Co.workers Spenders

not willing to discuss & with their Co workers

AUC

ideal AUC = 1

lower limit = 0.5

Count plot using Seaborn library

Acc to dataset more people were suffering from mental health.

Tableau

Chloropleth maps (Colour mapping technique)

x → MHD currently, areas in the world suffering from MHD.

because dependent variable & categorical in nature

mann whitney utest (nonparametric)

for Hypothesis testing

Pvalue < 0.05

why training and testing for machine learning

used to estimate the performance of machine learning algorithms for making predictions (unbiased performance) and evaluation of algorithms how model performs with new data