Analysis of survey on Mental health conducted by OSMI

Abstract

Mental health is a widely discussed topic in every industry. Here we are specifically concerned about the tech industry. The respondents of the survey are working professionals, both from technical and non-technical field, majorly from California, US. We were interested to find the major reasons that affect the mental health of these individuals. Reasons such as working region, role type, family history, gender, etc and their interactions with Mental Health of the individuals is studied in this project.

Various tools used for the study and analysis are visualizations for interacting variables, Chi square test for testing independence of the attributes, fitting of multiple classification models.

Through classification models we are able to predict the mental health scenario of a specific individuals having specific attributes (Age, gender, Region, etc).

KEYWORDS

Statistical learning, Logistic Regression, Chi square test, Testing significance of the model, Decision Tree

Index

Acknowledgement

A project usually falls short of its expectations unless guided by the right person at the right time. Success of a project is an outcome of sincere efforts, channelled in the right direction, efficient supervision and the most valuable professional guidance. This project would not have been completed without the direct and indirect help and guidance of such luminaries. They provided us with the necessary resources and atmosphere conductive for healthy learning and training.

We would like to thank Savitribai Phule Pune University for giving us an opportunity to perform the project because of which could apply the theoretical knowledge in Statistics at an undergraduate level we express our gratitude to Principal,

Dr. Rajendra S. Zunjarrao, MODERN COLLEGE OF ARTS, SCIENCE AND COMMERCE, PUNE for allowing us to present this project. At the outset, we would like to take this opportunity to gratefully acknowledge the very kind and patient guidance that we have received from Prof. Sagar Khandagale. We would also like to thank, HOD Dr. Manisha Sane, Dr. Kamble as well as the teaching staff Dr. M. S. Prasad, Mrs. Dhanashree Raskar, Mrs. Trupti Chaudhari and Non-teaching staff Mr.Patil.

Without their critical evaluation and suggestion at every stage of the project, this project could not have reached its present form. Faculty has critically evaluated our each step in developing this project.

We would like to extend the special thanks to our respondents who gave us fruitful information to analyse the data in the survey.

And finally the students of our college, friends and family for their support without which the project could not have been a successful one. Heartfelt gratitude to all of you.

Introduction

There are many horrible ways of people getting depression and anxiety in the world which is highly increasing as the days go on. One of these ways of producing these two elements is mental health. Mental health is not just a bad thing but is also good as you can create resilience, coping strategies and avenues to seek help with anxiety and depression. Psychology workers are one of the many jobs of people helping to create resilience, coping strategies and avenues for people to help them get through anxiety and depression through mental health. Resilience is what the psychology workers aim for first as it has the capacity to help people recover quickly from difficulties.

A way to find out how to use these things towards you as an advantage is by talking to a professional or someone that has used them in the past like a psychologist or a doctor. This is clearly affecting the world that we live in today by a long shot more than any other symptoms out there. For example, from the age 18 years and older 18% of the U.S populations have anxiety disorders, that's 40 million people that are affected by this disorder known as anxiety. Only one-third of the of people anxiety disorders around the world are being treated yet this disease is highly treatable. This leads onto the fact that over $42 billion is spent on anxiety disorders out of the total $148 billion total that is spent on the mental health bill, that's roughly one-third of the total payment. Therefore we can clearly tell by the statistics that over 40million people in the U.S have the anxiety disorder, only one-third of the people are being treated yet it is highly treatable and the fact that it has used over $42 billion in the U.S.

Therefore it can be concluded that the current position that the population of people with anxiety disorders and people that are being treated with and completely different numbers and requires and seeks global attention as it is a big issue. There is a lot we can do on this matter, to show how we as a Statistician, play a big role.

Background

This project contains the analysis of Mental Health Survey data collected from OSMI (Open Sourcing Mental Illness). The project takes into consideration of various factors that affect the mental health of the individual.

Through this analysis we are building a model for prediction of mental health of an individual of specific age, gender, working region and other interacting factors.

Firstly we start with cleaning the data and doing all the pre-processing needed. Then moving forward with the doing some EDA (Exploratory analysis), observing the interaction of the variables with the target variable and finally, Fitting classification models for classifying a specific individual in one of two category that is, whether he is Diagnosed with Mental health issue or not.

1. Methodology

1) The Dataset

- Source : Our dataset is produced by Open Sourcing Mental Illness (OSMI).
- (http link)
- OSMI is a non-profit corporation dedicated to "raising awareness, educating, and providing resources to support mental wellness in the tech and open source communities." What they do in support of this goal includes providing e-books on mental wellness in the workplace, hosting a forum on conversations on mental health, and holding talks at developer conferences about mental health in the community.
- In one of their efforts, OSMI provides a survey on mental health in tech industry. This survey contains a variety of questions pertaining to the mental health of the respondents, the demographics of the respondents, and how employer views on mental health in the workplace. This survey was conducted in 2014 and 2016. For today, we will be using the later year's dataset.
- OSMI representatives listed some of their questions on the survey in the data.world forums. Most notably, OSMI is interested in how certain demographic and work-life components of respondents impact the rate of mental health conditions in the industry.

2) Cleaning the data set

- The data set has about 31 columns/variables. Some are numerical and most the variables are categorical. There are some questions that are not answered by the respondent hence some variables have no values in them but the good part is that those variables are not providing value to the analysis and as a result are not considered as important.
- For cleaning, encoding and analysis purposes we have used python programming language and python libraries.

3) Selection of Variables :

- Variables(Input variable) are shortlisted on the basis of their interaction with Mental health diagnosis (Target variable).

4. Usage of Statistical tools:

i. Data Visualizations.

ii. Chi-square tests .

iii. Contingency tables.

iv. Classification models.

Statistical Tools

Data Visualizations

Graphical Representation is a visual display of data and statistical results. It is often more effective than presenting the data in tabular form. There are many different types of graphical representations which is used depending upon the nature of data and type of the statistical results. It is very effective way to serve the purpose of comparison at a glance and revealing the patters in the data. Graphs and diagrams are easy to understand and create an effect. Graphs and charts are often used to easy understanding of large quantities of data and relationships between parts of the data. Graphs can usually read more quickly than the raw data that they are produced form. They are used in wide variety of fields and can be created by hands often on graphs papers or by computer using a chart application. Therefore, Graphs and Charts believed to be powerful tools to convey information.

▪ Bar Diagram

Bar graph is used frequently in practice for the comparative study of two or more items or values of single variable or a single classification or category of data. Bar diagrams are one of the easiest and the most commonly used devices of presenting most of the business and economic data. These are especially satisfactory for categorical data or series.

▪ Multiple Bar Diagram

A multiple bar diagram is used for two or three-dimensional comparison. For comparison of magnitudes of one variable in two or three aspects or comparison of magnitudes of two or three variables, rectangles in a group are placed side by side. The R-commands are similar to those used for subdivided bar plot. Only we change the default value of argument beside to true. We illustrate the construction of multiple bar plot using the following data.

CHI-SQUARE (X_2) Test for independence of attributes:

•

Suppose that the given data are classified into r levels of attribute A denoted as $A_1,\ldots..A_r$ and s levels of attribute B represented by $B_1,\ldots..B_s$.

Then different class frequencies can be represented in the following tabular form:

A \ B	B_1	B_2	....	B_j	....	B_s	TOTAL
A_1	O_{11}	O_{12}	....	O_{1j}	....	O_{1s}	(A_1)
A_2	O_{21}	O_{22}	....	O_{2j}	....	O_{2S}	(A_2)
....	....	....	....	....	....	....	....
A_i	O_{j1}	O_{j2}	....	O_{ij}	....	O_{is}	(A_i)
....	....	....	....	....	....	....	....
A_r	O_{r1}	O_{r2}	....	O_{rj}	....	O_{rs}	(A_r)
TOTAL	(B_1)	(B_2)	....	(B_j)	....	(B_s)	N

This table is as (r x s) contingency table.

$N=\sum\sum O_{ij}$=Total observed frequency

$(A_i)=\sum O_{ij}$=Total of observed frequencies in i^{th} row; i=1,2,....,r.

$(B_j)=\sum O_{ij}$=Total of observed frequencies in j^{th} column; j=1,2,...,s.

Here, Hypothesis under consideration is,

H_0: Two attributes A and B are independent.

v/s

H_1: Two attributes A and B are not independent.

$e_{ij} = (Ai)(Bj)/N$; i=1,2,....,r; j=1,2,....,s. The test statistic under H_0 is,

$$\chi^2=\sum\sum (O_{ij}-e_{ij})/e_{ij} =\sum\sum(O_{ij}^2/e_{ij})-N$$

Criteria: We reject H_0 at α% l.o.s. if $\chi^2_{r-s-1} \geq \chi^2_{(r-s-1),\alpha}$, Otherwise accept it.

Decision Trees

Decision tree analysis is a supervised machine learning method that are able to perform classification or regression analysis (Table 1). At their basic level, decision trees are easily understood through their graphical representation and offer highly interpretable results.

A tree begins with a root node which is split into two branches; each subsequent split occurs at an intermediary node, also sometimes called a decision node. The tree ends with the terminal or leaf nodes and any subset of connected nodes is referred to as a sub-tree (Fig 1).

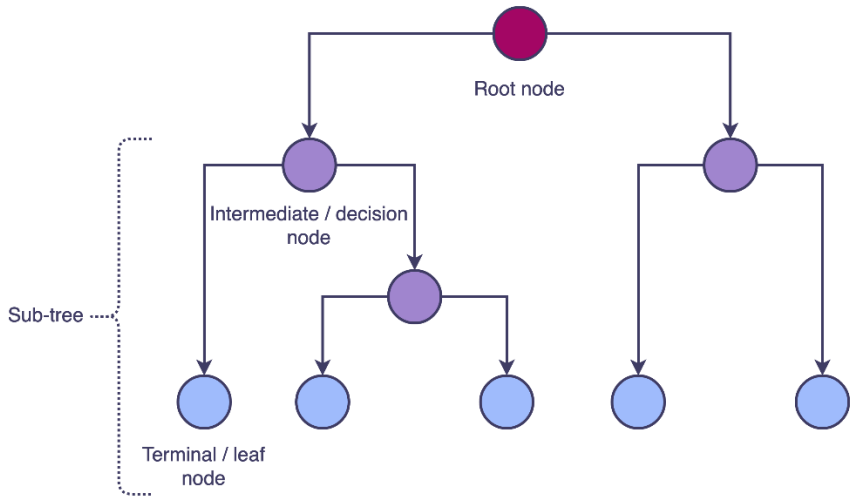


Fig . Decision tree structure where each node split results in two branches. The initial split is made at the root node, subsequent splits are made at intermediate/decision nodes and the tree ends with unsplit terminal/leaf nodes. A subset of connected nodes is a sub-tree.

Decision trees are most commonly used for prediction. By way of following the path determined by the decision split at each node, a new observation can be predicted to belong to one of *k* terminal nodes and from there obtain its predicted value.

In classification, the predicted outcome class will be the class with the highest proportion of training observations in that terminal node.

2. Exploratory Data Analysis

We see that our dataset contains survey responses from 1433 users, and each filled-out survey represents the answers to 63 questions. Because this survey is collected through a *Typeform*, there are missing values in each row that indicate questions that were not answered by the respondent. Thankfully, the variables we are considering are filled out by the majority of users, and so we are not dealing with a severe missing values problem :)

Selection of the feature:

Question/Variable that represent the diagnosis of the mental health of the person.

1) Have you had a mental health disorder in the past ?

2) Do you currently have a mental health disorder ?

3) Have you been diagnosed with a mental health condition by a medical professional?

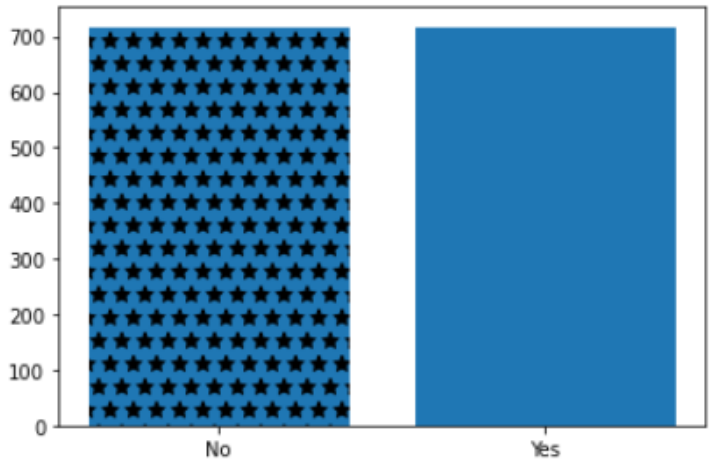
Why this parameter?

Of the three questions above third question represents most reliable answer for diagnosis of mental health in a person. The other two questions seems to have to bias in them since answers are not according to the medical professional. We will definitely go with ones that have less bias. Hence we chose the third question as representative of the diagnosis outcome.

Data Visualizations and Observing interaction effects between features of the data.

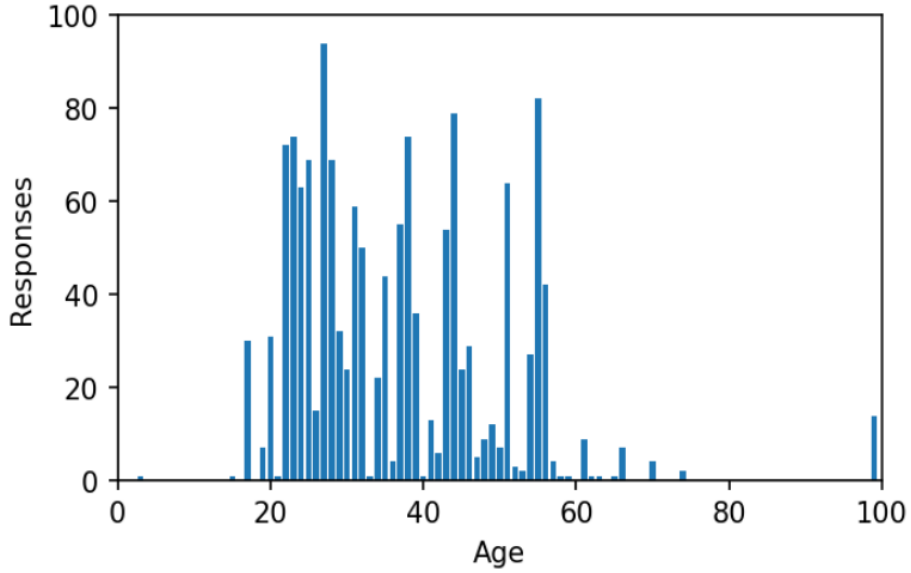
Analysis of number of respondents (according to age, gender, role type,etc)

1) Total number of diagnosed individuals/respondents.



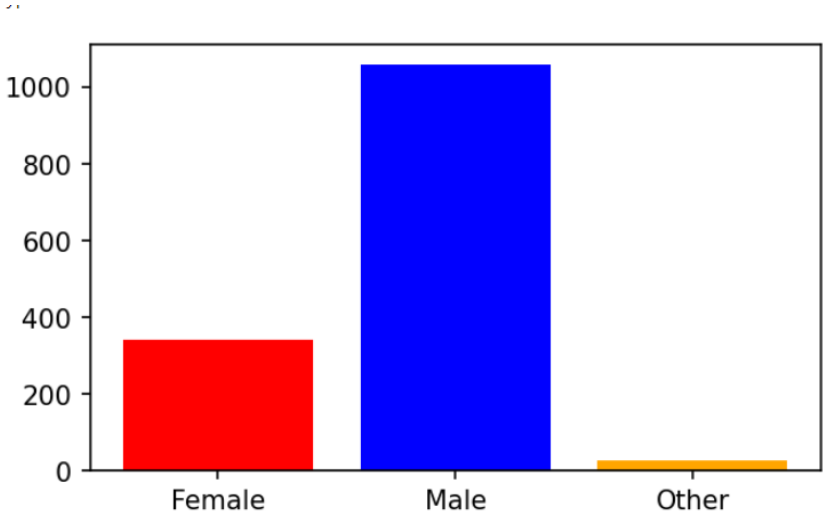
Comment: The total number of respondents that have diagnosis is same as that don't have.

2) Age Wise Respondents



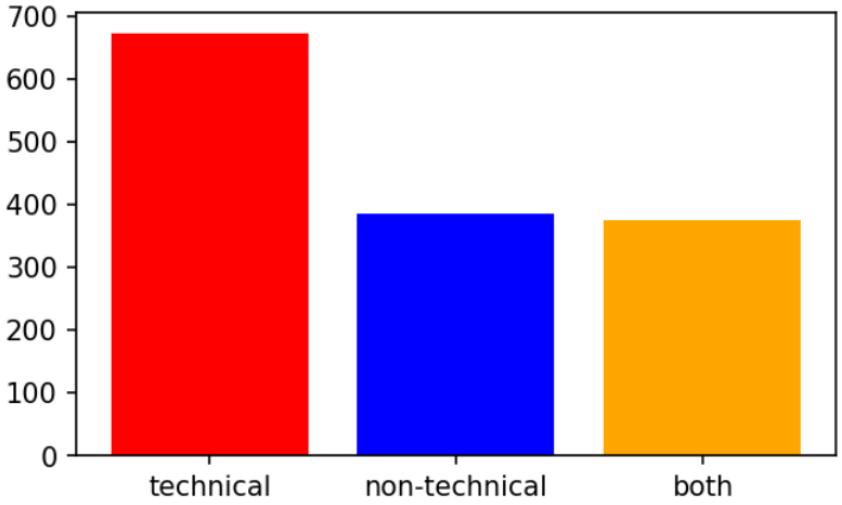
Comment: Maximum number respondents are in the age bracket 20-60, which is relevant because the most the working professional are in this bracket.

5) Genderwise Respondents



Comment: Among the total of number of respondents, Male are high in response.

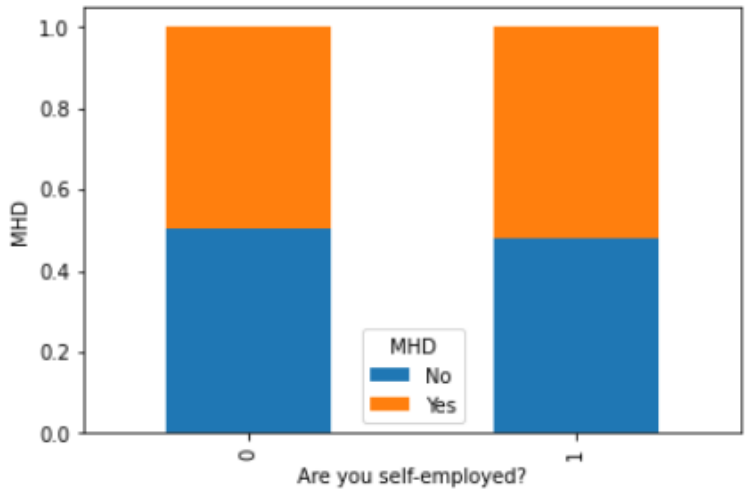
6) Role-Type wise respondents



Comment: Technical working people are high in number as respondents for the survey as compared to non-technical individuals and those who have both roles.

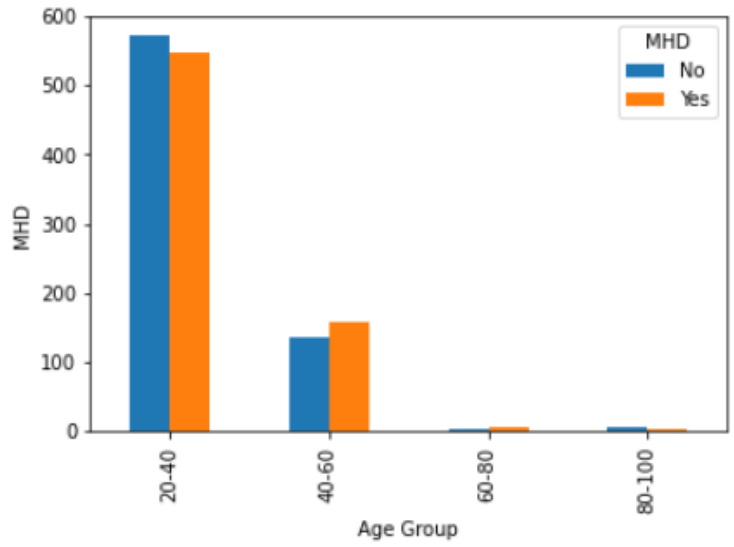
To check for the interaction effects of the Target variable(Mental health status) with other features of the data.

7) Employment type vs Mental Health status.



Comment: There is no significant interaction between the employment type and Mental Health. Those working in a company and those who have their own business / firm have same number of proportion of individuals who are mentally diagnosed and those who are not.

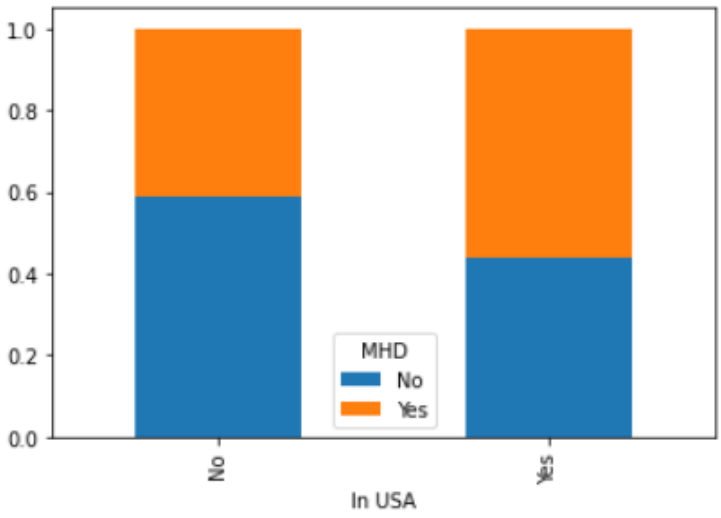
8) Age- Group vs MENTAL HEALTH STATUS



Comment : According to the data, there is slight interaction between the age group variable and Mental health. Individuals falling in the age bracket 20-40 have less mentally diagnosed individuals than those who are not diagnosed. Also it can be

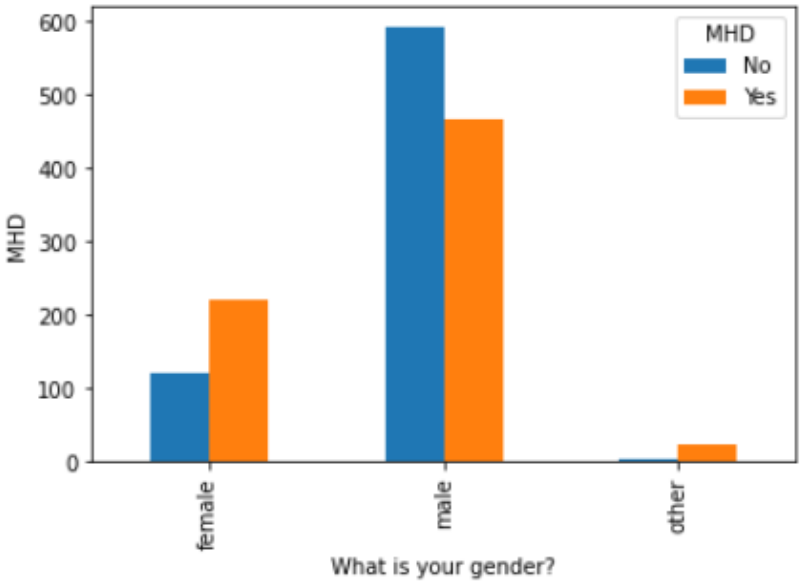
observed that as the age increases the individuals in mentally diagnosed proportion increases.

9) Working in USA vs MENTAL HEALTH STATUS



Comment : Whether an individual works in USA or not, does affect his Mental health. Interaction is observed between the variables. Also note that, most of the respondents are from California, USA the outcome is biased for USA.

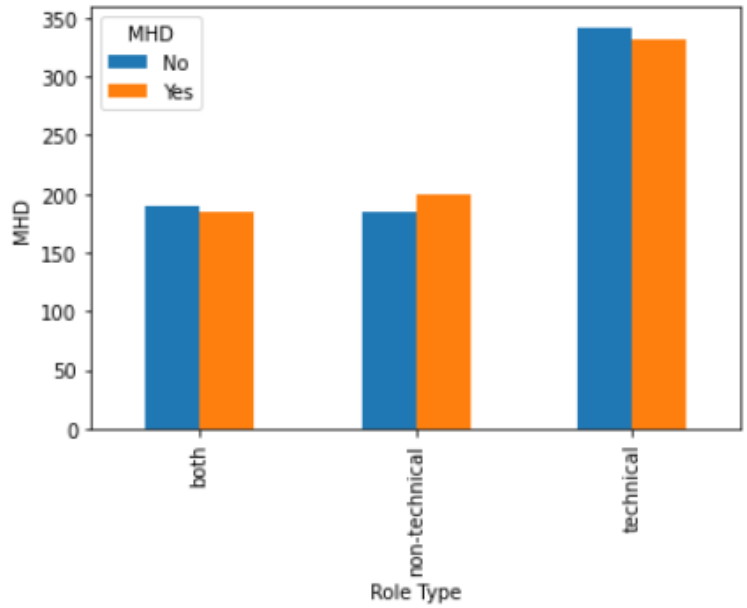
10) Gender vs MENTAL HEALTH STATUS



Comment : We have seen above that male respondents are high in number as compared other gender types. Considering, Female respondents it can be observed that most female respondents are diagnosed with mental health issue and Considering male respondents, most of the respondents are not diagnosed with

mental health issue. Hence it can be concluded that there is interaction effect of age group with the Mental Health Status.

11) Role type vs MENTAL HEALTH STATUS

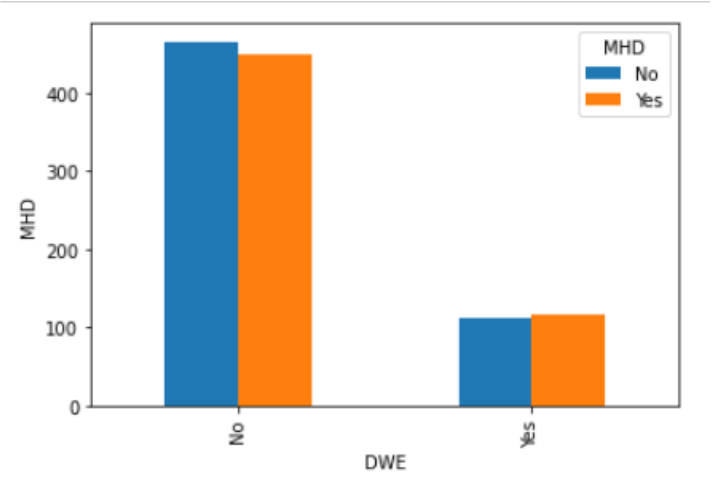


Comment: No big interaction of individuals role type can be seen with the Mental health status.

12) Company size vs MENTAL HEALTH STATUS

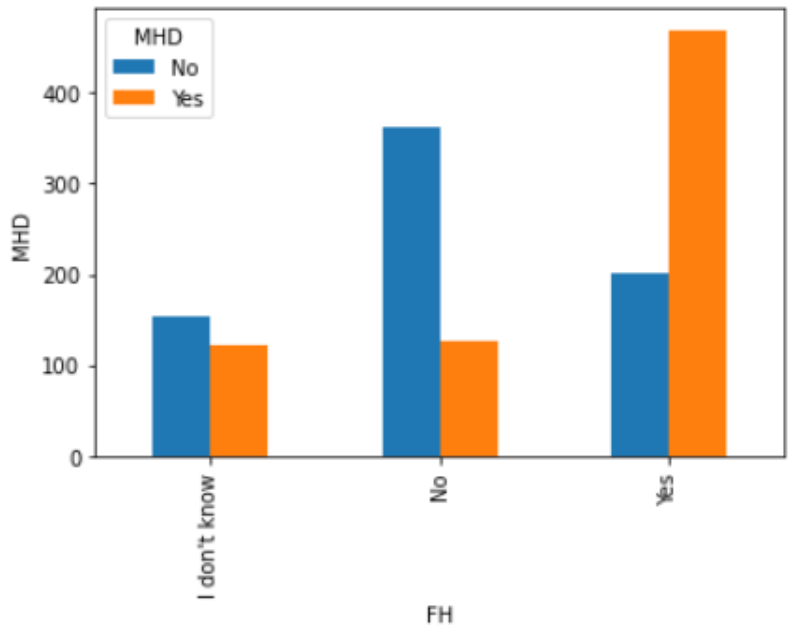
Comment : Interaction is definitely observed. It can be interpreted that with big company size comes huge responsibility and stress could be one of the factors to increase mental health issue.

13) Discussion of Manager with Employees (DWE) vs MENTAL HEALTH STATUS



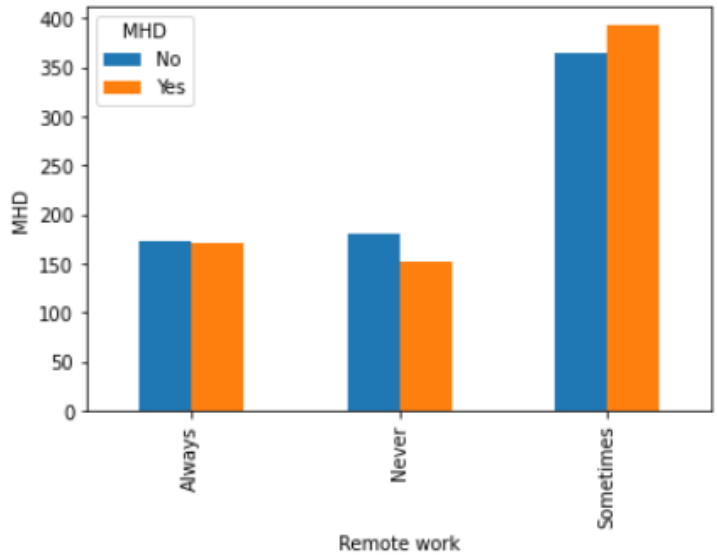
Comment : Whether the team head, manager discuss topic mental health issue with their employees doesn't seem to have to any sort of effect on the Mental health status of an individual.

14) Family History vs MENTAL HEALTH STATUS



Comment: Respondent's family history does have effect on the Mental health of the respondent. Really high interaction effect can be seen. Does it mean ancestoral history in mental health issue have effect in present, for that more analysis might be needed to be done. For right now we will stick with this data and its analysis.

15) Remote work vs MENTAL HEALTH STATUS



Comment: Interaction can be seen between the variables. That is if the individual works remotely or non-remotely or both, does seem to have effect on the mental health status of the respondents.

Using Chi Square test for independence of attributes.

1) Contingency table and Chi-Square test result for features Mental Health status and Role type of respondents.

MENTAL HEALTH STATUS	No	Yes
Role Type		
both	190	185
non-technical	185	200
technical	342	331

Chi Square test results:

observed_values :-
[[190 185]
[185 200]
[342 331]]

p-value is :-

0.6602819108951667

We conclude that ,the categorical variables Remote work and Mental health status are independent and no interaction effect is present. That is we accept the null hypo thesis at 5% l.o.s

2) Contingency table and Chi - square test for features Mental Health status and of respondents.

MENTAL HEALTH STATUS	No	Yes
RW		
Always	172	171
Never	181	152
Sometimes	364	393

Chi square tests result:

Observed values:-
[[172 171]
[181 152]
[364 393]]

p-value is :-
0.16213035241171508

Comment : According to the test Criteria as p-value is greater than 0.05, we accept the null hypothesis and conclude that the attributes are independent that they do not interact with each other.

4) Contingency table and Chi - square test for features Mental Health status and Family history of respondents

MENTAL HEALTH STATUS	No	Yes
Family Hist.		
I don't know	154	121
No	362	126
Yes	201	469

Chi Square test results :

observed_values :-
[[154 121]
[362 126]
[201 469]]

p-value is :-
1.1988824713501208e-49

Comment : According to the test Criteria as p-value is less than 0.05, we reject the null hypothesis and conclude that the attributes are dependent that they do interact with each other.

Shortlisting of the variables that show interaction with the feature – Mental health diagnosed by Medical professional.

The input variables or the parameters of the model will be :

a) Gender
b) Age
c) Work region
d) Company size
e) Family history
f) Remote work
g) Role type

Note : The features i) Role type, ii) Remote work might not show high interaction effect with our target variable, as well as do not pass the criteria of the chisquare test at 5% l.o.s but they do affect the mental stress level of the individual as most of the respondents are working professionals. Hence they have been included in the model.

Encodes

Gender

Female	0
Male	1
Other	2

#Age

20-40	0
40-60	1
60-80	2
80-100	3

Work in USA or not

Yes	1
No	0

Employment

Employee in company	0
Self employed	1

Role type

Both	0
Non-Technical	1
Technical	2

Remote work

Always	0
Never	1
Sometimes	2

Company Size

1-5	0
100-200	1
26-100	2
500-1000	3
6-25	4
>1000	5

Mental Health Diagnosis

Yes	1
No	0

Fitting of Classification Model

A) Fitting Decision Tree model (Classification model 1)

Here the input variables will be those parameters that have high interaction effect with the Mental health status variable / feature.

Input Variables are :

a) Gender
b) Age
c) Work region
d) Company size
e) Family history
f) Remote work
g) Role type

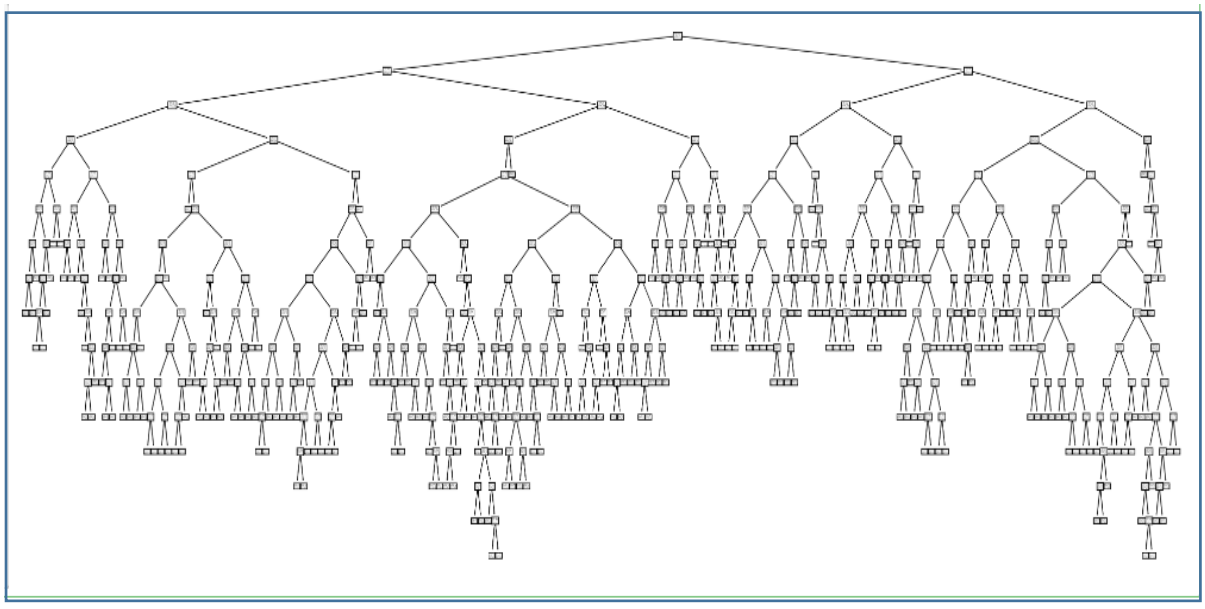
Target variable :

a) Mental health diagnosed or not.

Accuracy of the decision tree model.

The accuracy of the model is
81.85624563852059 %

Decision tree diagram:



Major findings

1) From the analysis we found that the variable 'Family history' displayed high interaction with Mental health status of an individual. So the further question is whether, family history of mental health issue affects the individuals mental health? Separate analysis needs to done to answer this question.

2) Discussion of team leader or the manager of the company with the employees of the company on the topic of mental health issues doesn't affect mental health of the individual.

3) Considering female working professionals, maximum of the females answered yes regarding the mental health diagnosis by medical professional. This might be because most of the females would have to handle their job and household activities. As we don't have any further data regarding this we cannot conclude anything.

Limitations

1) The data source is secondary, hence is biased because the data collection is not done by us. Typically the data is biased in favour of the organization that collected it (In this case it's OSMI)

2) The survey is specifically for tech industry, hence overall mental health status of the whole region or a country can't be concluded.

3) As discussed above, the survey is biased towards California, USA because most of the respondents are from California. It is also worth noting that Silicon valley is present in California.

4) As the survey was an open survey there are few responses from 'female' and 'other' gender category, which again results in bias towards Male since they are high in response number.

Bibliography

Data source :

1) https://osmihelp.org/research

Websites for references :

1) https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
2) https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96
3) https://en.wikipedia.org/wiki/Chi-squared_test