**Georgia State University**

### J. Mack Robinson College of Business

CIS 8005 Group Project Team 26

Well Being Prediction Analysis

**Satish Tirumalapudi:** [**stirumalapudi1@student.gsu.edu**](mailto:stirumalapudi1@student.gsu.edu)

**Salma Chanbasha Nandikotkur:** [**schanbashanandikotk1@student.gsu.edu**](mailto:stirumalapudi1@student.gsu.edu)

**Project Summary**

# Project Description:

Health is considered to be “The Most Important Asset” in everyone’s life. A healthy life is the foundation for an happy life. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make healthy choices. Mental health is important at every stage of life, from childhood and adolescence through adulthood.

Because of this busy schedule, people always tend to neglect their mental health rather than getting it treated. This is as important as Physical Health. This disorder can sometimes take up their lives. When we become ill, we tend to spend a lot of money, resulting in a lot of medical expenses.

So, we came up with an idea on designing a model to track the stability of human’s behavior based on multiple factors, so that then can monitor and get proper treatment when needed.

# Problem Statement:

This project is based on designing a Machine Learning model on how to predict whether a person needs an treatment for his well-being and what are the factors causing them to such condition which internally helps HR to identify the employee and provides necessary measures to improvise their condition.

# Data Source:

* The source of the dataset is from OSMH Open Sourcing Mental Health.
* This data file has all the necessary details to find out more on includes all needed information to find out more on age, gender, leave and necessary metrics to make predictions and draw conclusions out of it.

#### Link to the dataset: https://osmhhelp.org/

* This data set includes 1,260 responders in the survey conducted by OSMH. This includes the information for the year 2014.

The following are the independent features:

* **self\_employed:** Are you self-employed?
* **family\_history:** Do you have a family history of mental illness?
* **treatment:** Have you sought treatment for a mental health condition?
* **work\_interfere:** If you have a mental health condition, do you feel that it interferes with your work?
* **no\_employees:** How many employees does your company or organization have?
* **remote\_work:** Do you work remotely (outside of an office) at least 50% of the time?
* **tech\_company:** Is your employer primarily a tech company/organization?
* **benefits:** Does your employer provide mental health benefits?
* **anonymity:** Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
* **leave:** How easy is it for you to take medical leave for a mental health condition?
* **comments:** Any additional notes or comments

**Target Variable:**

* **Treatment**: This column is the target column, and the rest others are independent columns. Independent columns are those which will predict the outcome.

# Steps:

## Data Pre-Processing

* + Encoding
  + Scaling
* Exploratory Data Analysis
* Machine Learning Models
* Evaluation of Models
* Evaluation
* Recommendations
* Future Work
* Conclusion

This is the sample of how a data looks:

#### 

#### Platform used: Google Collab

**Data Pre-Processing:**

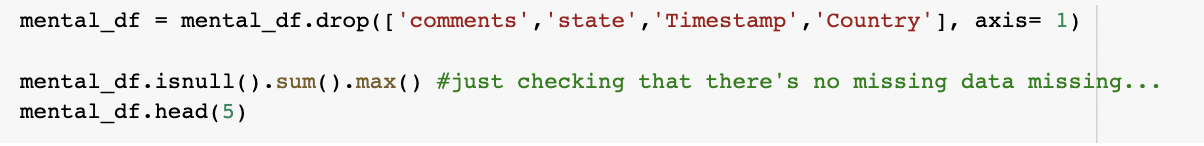
The following are the steps involved in the process of Data Cleaning:

* Checking for the Missing Values:

Table

Description automatically generated

* Now drop the Missing Values Columns and also the other unrelated features.



* Assign default values for the variables based on the data types.

int =0, string=”NaN”,float=0.0

A picture containing text

Description automatically generated

* We have got multiple responses for the “Gender” feature, Hence categorized that into 3 sub types, namely Male, Female, Trans

A picture containing chart

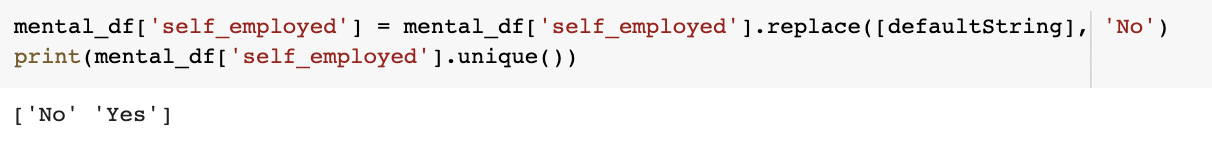
Description automatically generated

* Replaced the missing values with mean and the extreme values i.e., outliers with median this is especially in the “Age” feature.

Graphical user interface, text, application

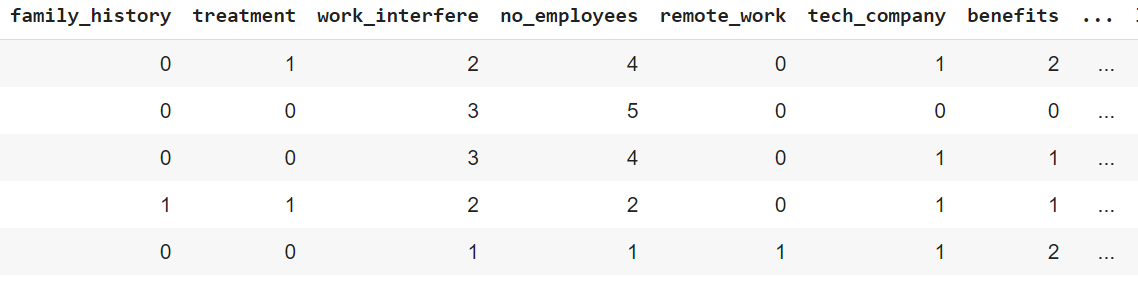
Description automatically generated

* In the “self\_employed” column we have replaced the smaller default values with the less frequency valued variable.



**Encoding:**

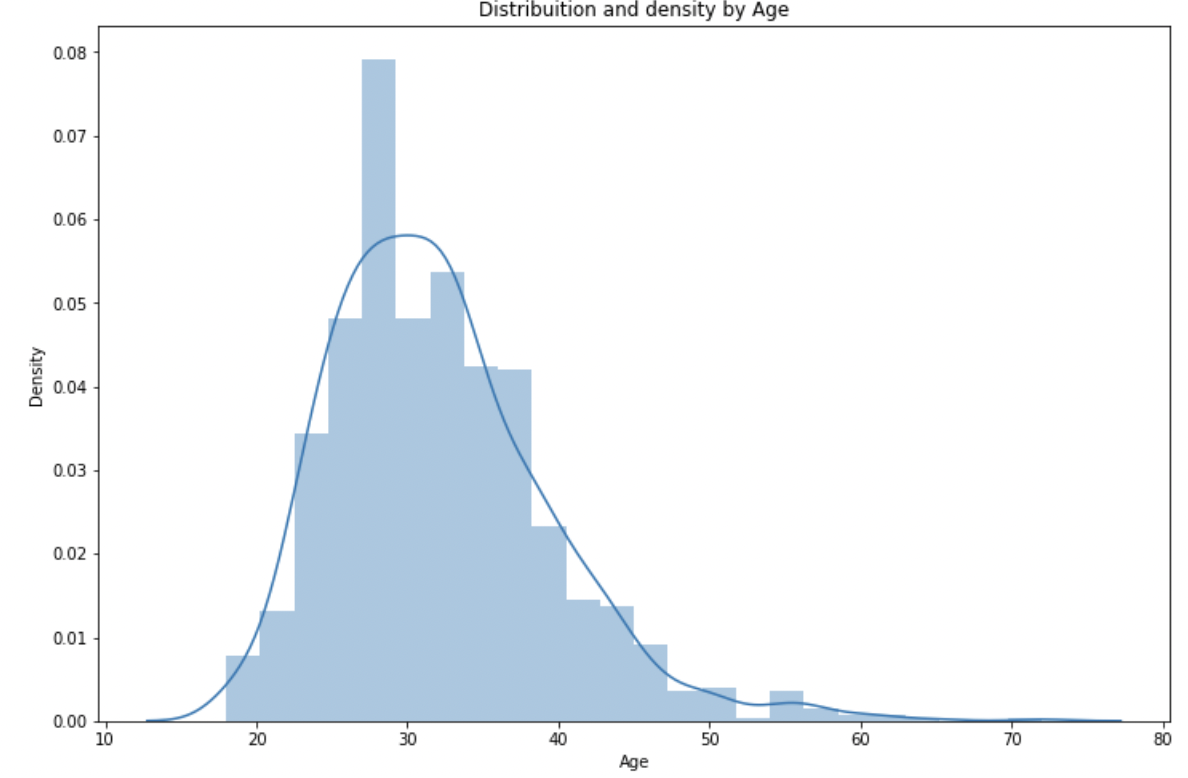
* Encoding is a part of data preprocessing. It refers to converting the labels into numeric form, to convert them into machine readable form.
* Most machine learning models only accept numerical variables, preprocessing the categorical variables becomes a necessary step.
* Here we did Label Encoding, each label is converted into an integer value.



# Exploratory Data Analysis:

### Data Visualization:

### The graph here shows the density of people by age, which helps us in understanding the age-range within the dataset.

****

#### Visualizing data using Age

* By Sub classifying this we got 65% females seeking for treatment 35% don’t and 55% males doesn’t need treatment and 45% do need and trans is into negligible fractions.
* The below graph depicts the percentages of survey responders based on Gender. Out of this 79% are male, 20% are female,1% is trans

Chart, waterfall chart

Description automatically generated

b. Visualizing data using Gender

* For the question Do you have a family history of mental illness? For this around 40% people have responded saying that they have a family history with the mental illness and people with family history of mental illness requires treatment compared to the no family history of mental illness. Family history is a significant factor in menta health disorders.

Chart

Description automatically generated

c. Visualizing data using family\_history

* For the question If you have a mental health condition, do you feel it interferes with your work? The results are around 37% people have responded that work interference that it sometimes where people doesn’t want to share about it or shy about sharing it. And they still require treatment for the mental health and the similar pattern is observed with Often.

**Chart, bar chart

Description automatically generated**

d. Visualizing data using work\_interfere column

* For the question does your employer provide mental health benefits? To this We see 38% and around 32% doesn’t know whether they provide benefits or not. Around 63 % people who know that their company provides mental benefits were seeking medical help and treatment.

**Chart, bar chart

Description automatically generated**

e. Visualizing data using benefits column

* Correlation matrix is built here using the corr() function between all the variables within the dataset.
* We observe some strong correlations here between the work\_interfere, anomity ,benefits, leave, gender, family\_history.
* It helps us to understand the direction, the form of the relationship and the degree(strength) of the relationship.

Chart

Description automatically generated

f. Correlation Matrix on encoded data

**Machine Learning Models:**

The below are the Machine Learning Models used for prediction analysis.

* Logistic Regression Model
* K-Nearest Neighbors Model
* Decision Tree
* Random Forest Model

## **Logistic Regression:**

* Logistic regression is a statistical method used to predict the outcome of a dependent variable based on previous observations.
* It's a type of regression analysis and is a commonly used algorithm for solving binary classification problems.
* As our target variable is continuous dependent variable, Logistic Regression is one of the best choice for predicting whether a person requires treatment or not ,with an accuracy of 80.15

Graphical user interface, text, application

Description automatically generated

Linear Regression Model

* The below is the confusion matrix with:
* True positive rate(sensitivity)=0.86
* False Positive rate=0.25
* Precision=0.68

**Chart

Description automatically generated**

Logistic Regression Confusion Matrix

**Decision Tree Model:**

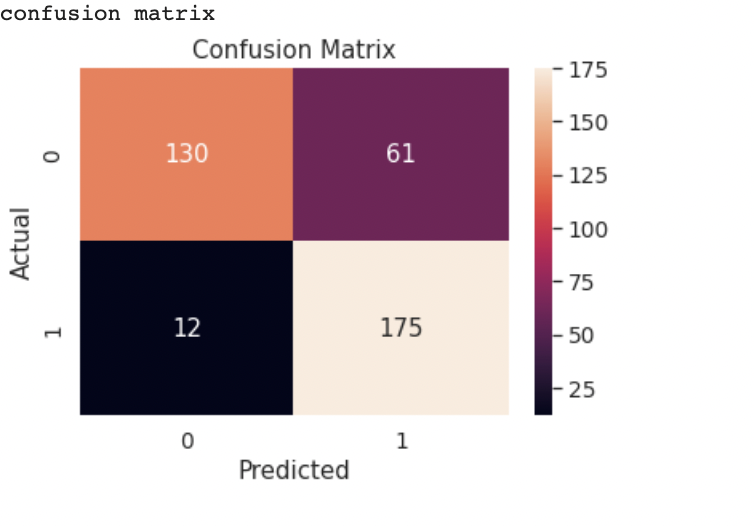
* Decision trees are a way of modeling decisions and outcomes, mapping decisions in a branching structure.
* Decision trees are an approach used in supervised machine learning, a technique which uses labelled input and output datasets to train models, with an accuracy of 80.68

**Graphical user interface, text, application

Description automatically generated**

Decision Tree Model

* The below is the confusion matrix with:
* True positive rate(sensitivity)=0.93
* False Positive rate=0.31
* Precision=0.72



Confusion Matrix for Decision Tree

### Random Forest Model:

* Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.
* **Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.**
* Here also we predicted whether treatment is required using Random Forest with an accuracy of 81.21

**Graphical user interface, text, application, email

Description automatically generated** Random Forest Model

* The below is the confusion matrix with:
* True positive rate(sensitivity)=0.92
* False Positive rate=0.29
* Precision=0.73

**Chart

Description automatically generated**

Confusion Matrix for Random Forest

### K-Nearest Neighbors Model:

* K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems.
* K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set and accuracy we obtained is 83.59

**Graphical user interface, text, application, email

Description automatically generated**

K-Nearest Neighbor Model

The below is the confusion matrix with:

* True positive rate(sensitivity)=0.90
* False Positive rate=0.23
* Precision=0.71

**Chart

Description automatically generated**

Confusion Matrix for K-Nearest Neighbor Model

**Comparison of Models:**

* We have used 4 models for our prediction. They are: Logistic Regression, Random Forest, Decision Tree and K-Nearest Neighbors.
* As per the accuracy, the K-Nearest Neighborsis the Best Choicewhereas the Logistic Regression Model comparatively gave us the least accuracy.
* We have successfully built our predictive model and compared these predictive models based on their accuracies.

Graphical user interface, text, application, email

Description automatically generated

* Below is the bar plot to show the accuracies of the four used models.

Chart, bar chart

Description automatically generated

**Evaluation:**

* We came to know that the most important factors to predict whether the person requires treatment for the mental health condition are age, family\_history, anonymity ,leave and work\_interfere.
* Also, we did observe a strong relationship with the Family History as the people with mental health history requires more attention towards the treatment than the other.
* Another insight we observed is as companies tend to provide more health benefits the people requiring treatment is less. So, the companies should concentrate more on the benefits to provide for the wellbeing of a person.

**Recommendations:**

* There are some features which are not utilized in the data set. With usage of the other valid features, there can be better predictions on the person’s well-being.
* We can try some more predictive models on this data set to get more accuracy.

**Future Work:**

* We intend to predict and detect What is the exact treatment they require and reasons for the abnormal behaviour by further introducing new features into our data set like Treatment\_Type, Work\_Department and many more.
* This analysis will help people understand their stability so that they can get treated and lead a better life.

# Conclusion:

# We have built four models, among them the K-Nearest Neighbor Model shows the best result through which we can say 83.6% variability of treatment can well be explained by predictor variables and which yields on predicting whether a person requires treatment or not is more reliable through this model. This will not vary much from the actual outcome.