**Mental Health Prediction with Machine Learning**

Ravindra Neralla

DSC 680

Milestone 3 – Project 2

# Abstract

COVID-19 pandemic impacted lives of people around the world. [1] Many of us are facing challenges that can be stressful, overwhelming and cause strong emotions in adults and children. Public health actions, such as social distancing, are necessary to reduce the spread of COVID-19, but they can make us feel isolated and lonely and can increase stress and anxiety.  Increased stress causes worsening mental health conditions.

[2] As the COVID-19 pandemic enters its second year, new fast-spreading variants have caused a surge in infections in many countries, and renewed lockdowns. The devastation of the pandemic — millions of deaths, economic strife, unprecedented curbs on social interaction — has already had a marked effect on people’s mental health. More than 42% of people surveyed by the US Census Bureau in December reported symptoms of anxiety or depression in December, an increase from 11% the previous year.

**Problem Statement/Hypothesis:**

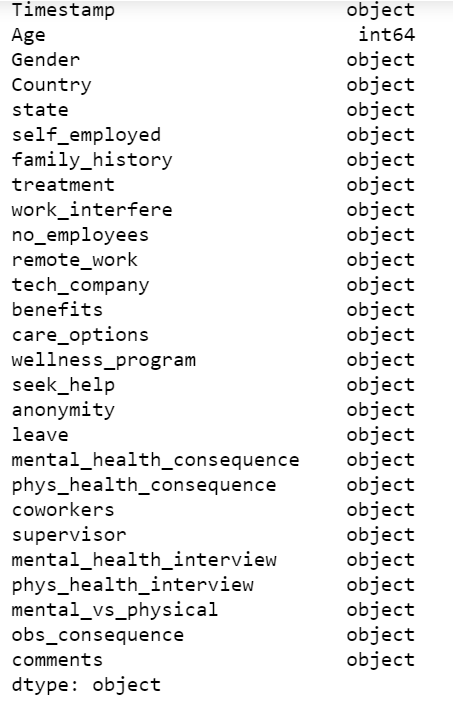
Young adults have experienced a number of pandemic-related consequences, such as closures of universities and loss of income, that may contribute to poor mental health. During the pandemic, a larger than average share of young adults (ages 18-24) report symptoms of anxiety and/or depressive disorder (56%). Compared to all adults, young adults are more likely to report substance use (25% vs. 13%) and suicidal thoughts (26% vs. 11%). Prior to the pandemic, young adults were already at high risk of poor mental health and substance use disorder, though many did not receive treatment [5].

The main research objective in the dataset is identifying the key attributes useful in predicting the mental health status of a person. It will be helpful for health care providers to assess and diagnose the condition appropriately.

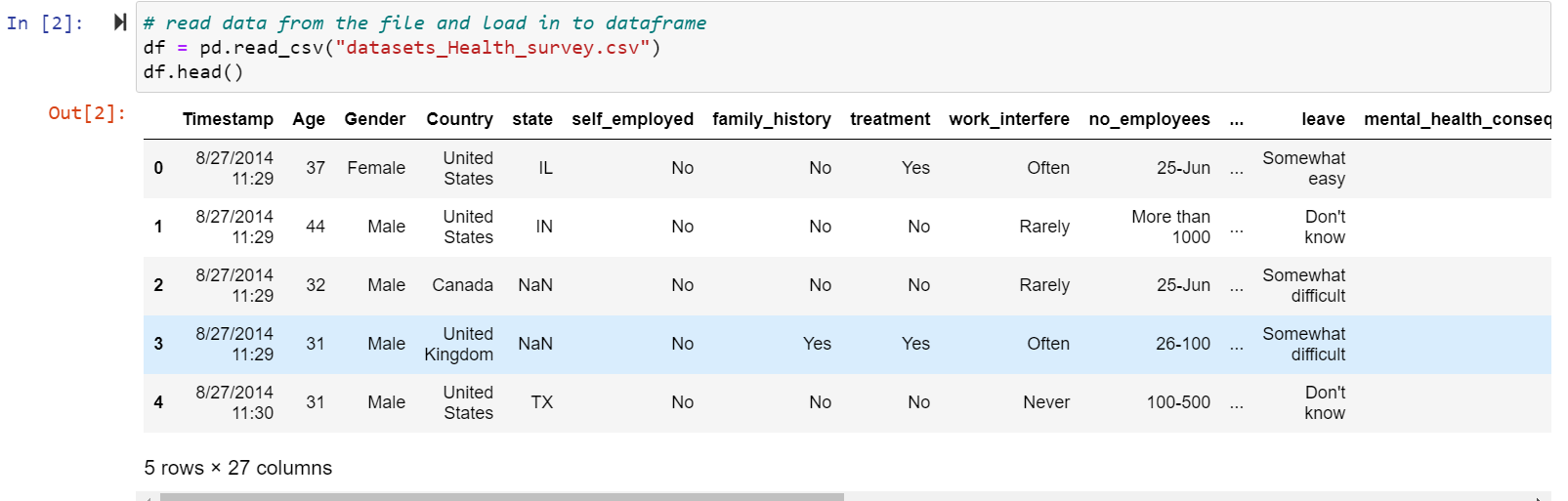
This project is a classification problem in predicting mental health on different individuals with unique attributes. The machine learning algorithms that are going to be compared in performance of accuracy in this project will be Logistic Regression and AdaBoost Classifier.

**Data Source:**

The dataset I have used in the analysis is taken from Kaggle, [Data mining of mental health](https://www.kaggle.com/diegocalvo/data-mining-of-mental-health/data). Data set has 27 variables with 1259 observations. Dataset contains survey questions related to mental health of an individual person.



Sample Data:

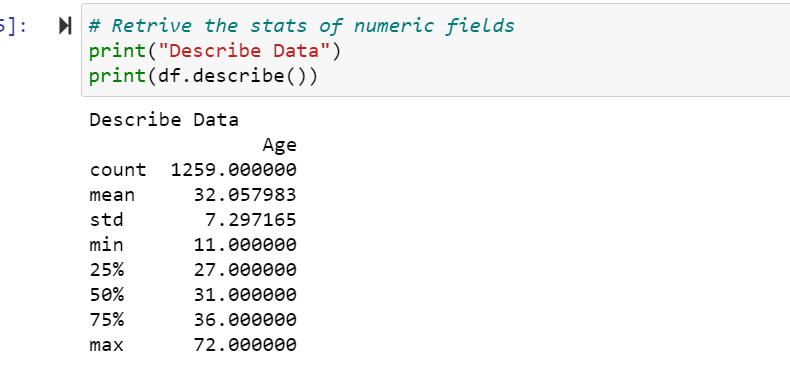


#### **Technical Approach:**

Data Analysis:

I have loaded the data into a python data frame and looked at sample data to understand the variables. After looking at the stats in the data, I observed that data looks clean only one duplicate entry in the dataset. After doing null value analysis, observed that work\_interfere and state variables have null values. To understand more about data, completed the analysis on categorical and numerical variables.

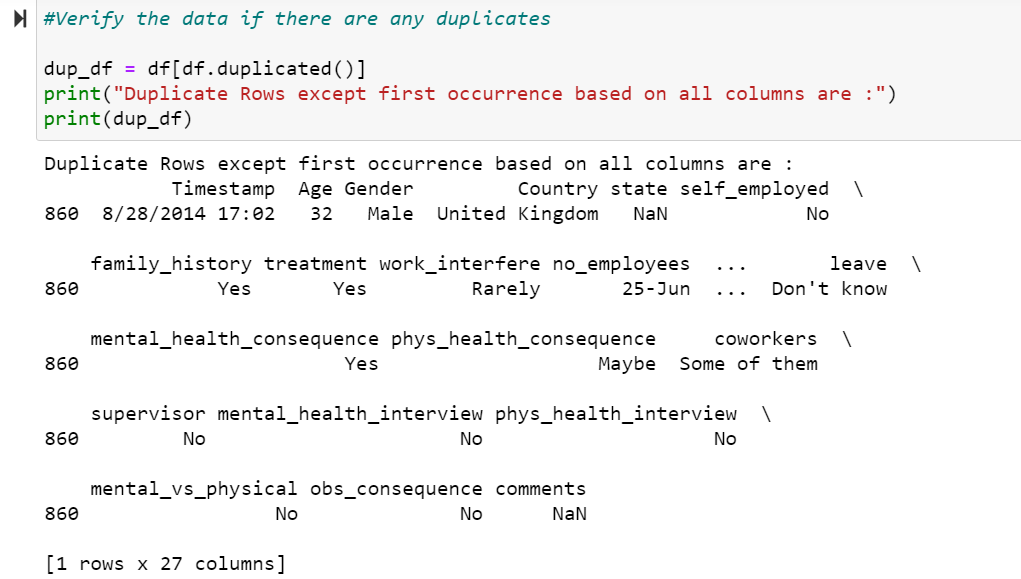
* 1. Get the stats on numerical variables:



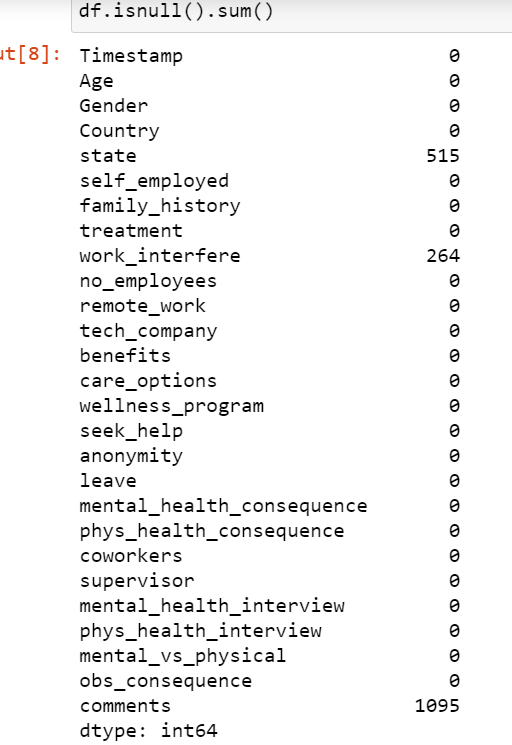
* 1. Get the stats on categorical variables:



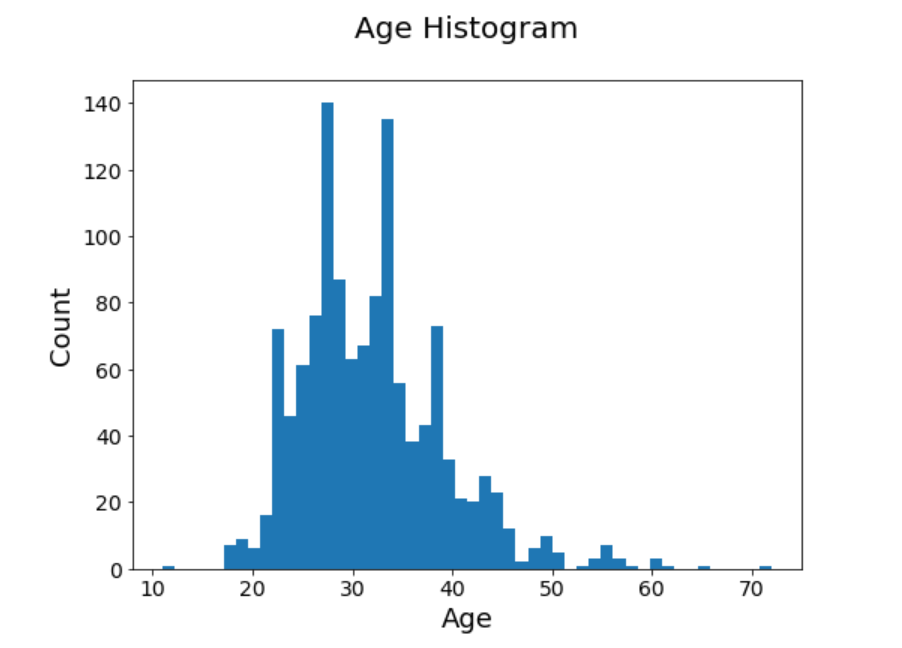
* 1. Verify for duplicate values in dataset:



* 1. Verify for null values in dataset:

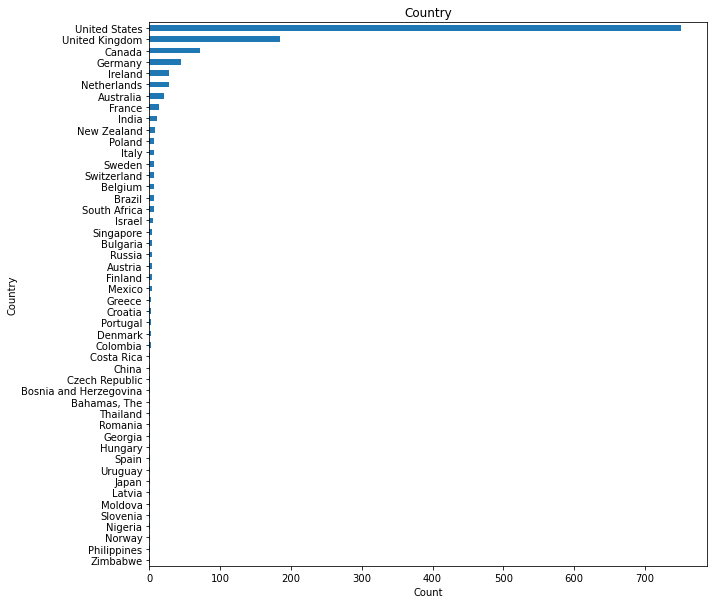


* 1. Generate bar charts:

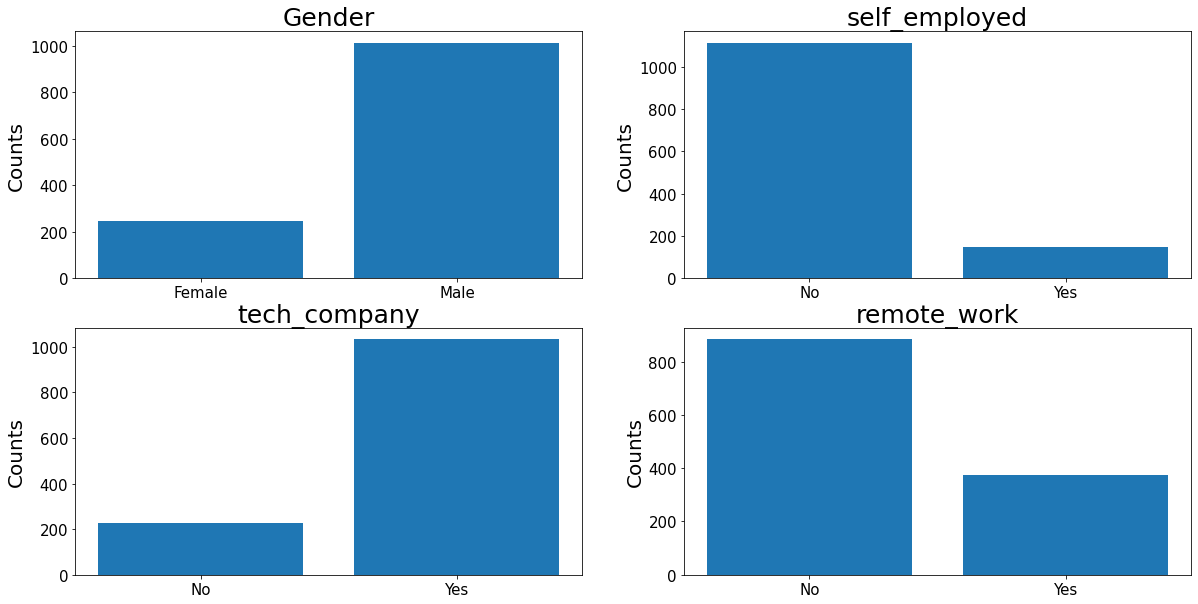


**Observation:**

Most of the participants age is between 20 to 50 years

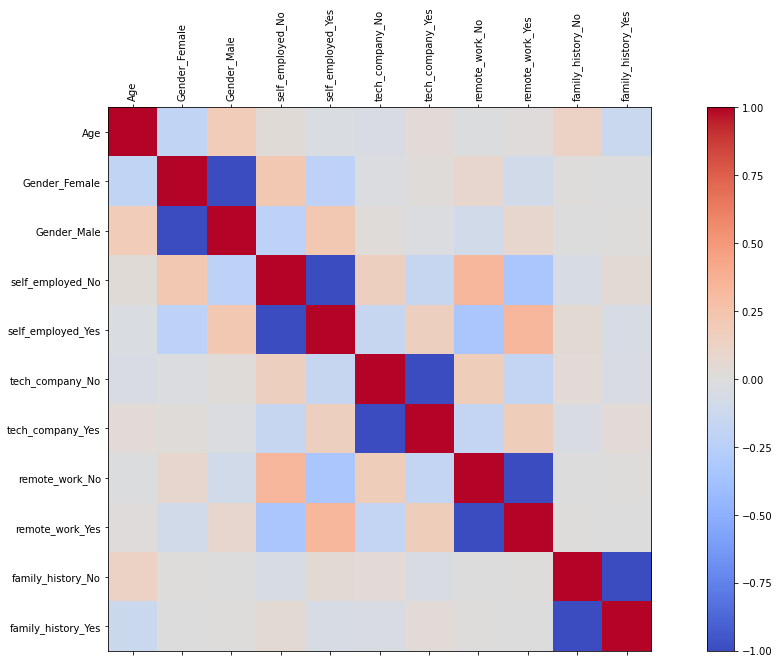


* Looking at the above graph, we can see that maximum number of observations are from Country = United States so filtered the data based on the ‘Country’ field.
* Number of Elements Before filtering on ‘Country’ is: (1259, 27)
* Number of Elements After filtering on ‘Country’ is: (751, 27)
* With this we have reduced the number of rows to 751.
  1. Categorical data analysis:

****

Observations:

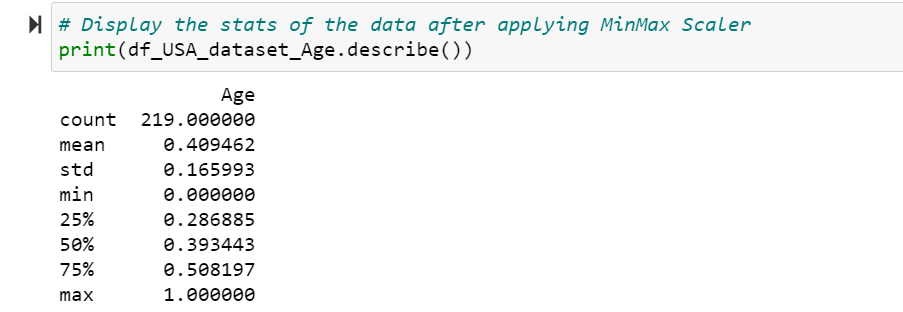
* Most of the participants in the survey are Male.
* Most of the participants are not self-employed.
* Majority of the participants are working for tech companies.
  1. Heat Map analysis:

****

* + 1. Min Max Scalar:

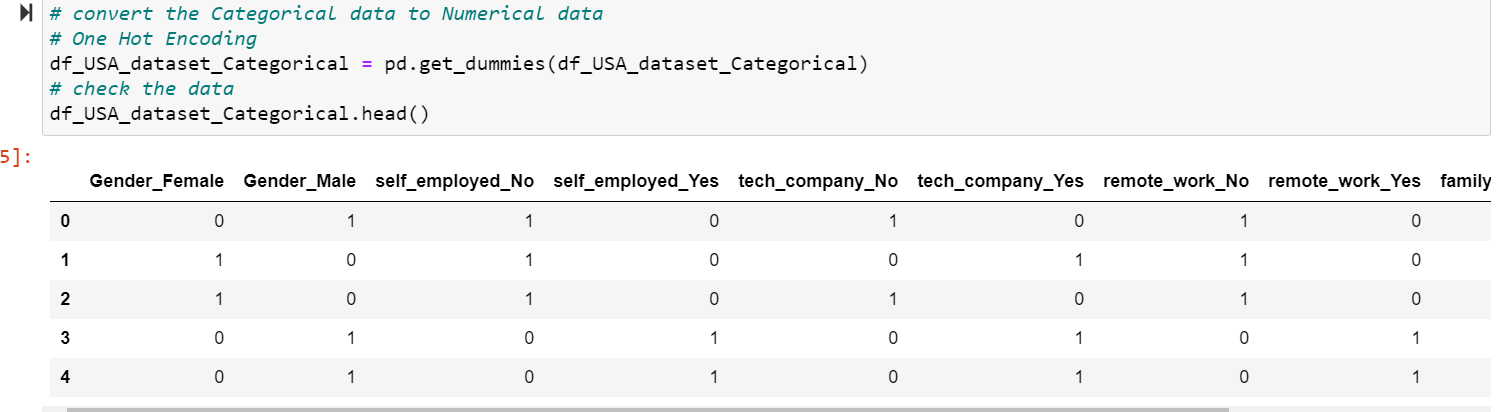
Age variables has values ranging from 11 to 70, due to wide range of the values inconsistent with other values magnitude, applied min max scalar to unify the values to same range.

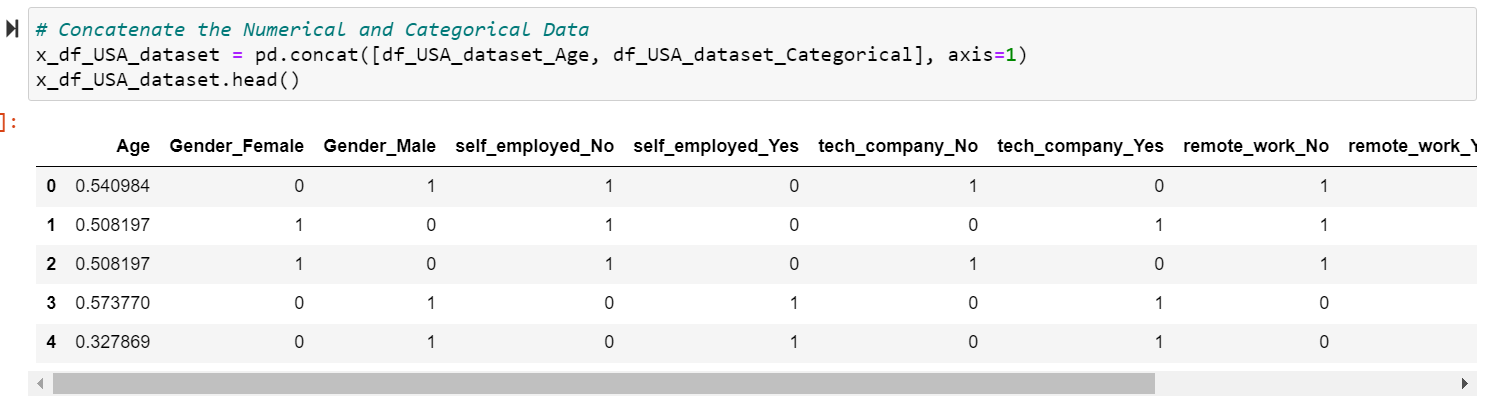




Convert Categorical to Numerical:

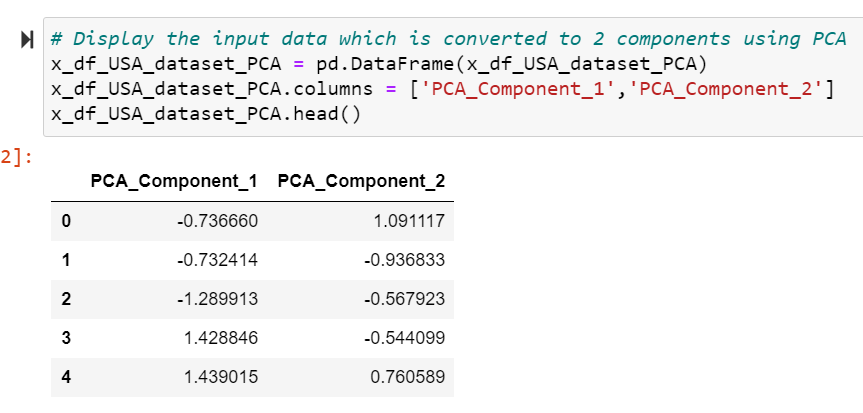
As the model accepts numerical values only, I have converted categorical variables to Numerical.

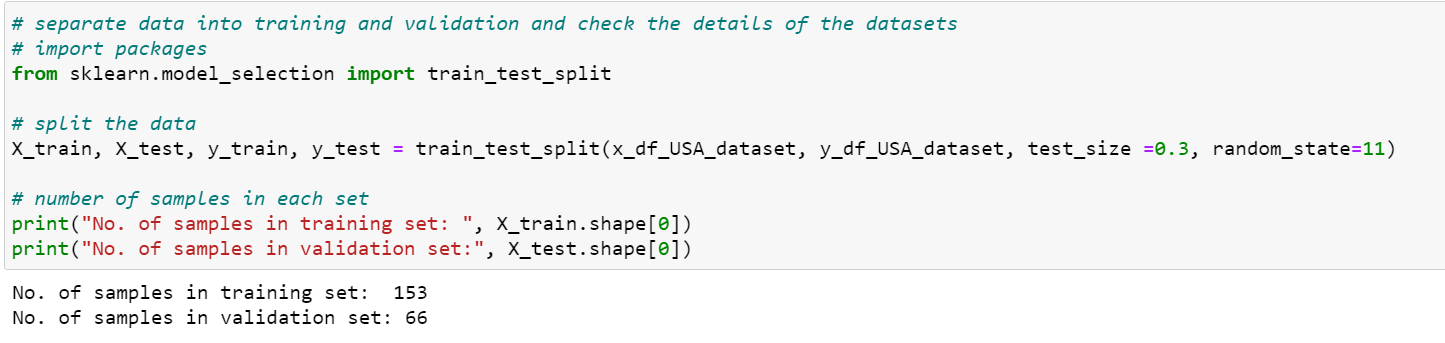




Dimensionality Reduction:

After converting categorical variables to numeric values, the number of input variables have increased. So, dimensionality reduction was performed by applying Principal Component Analysis, reducing the number of variables to 2.





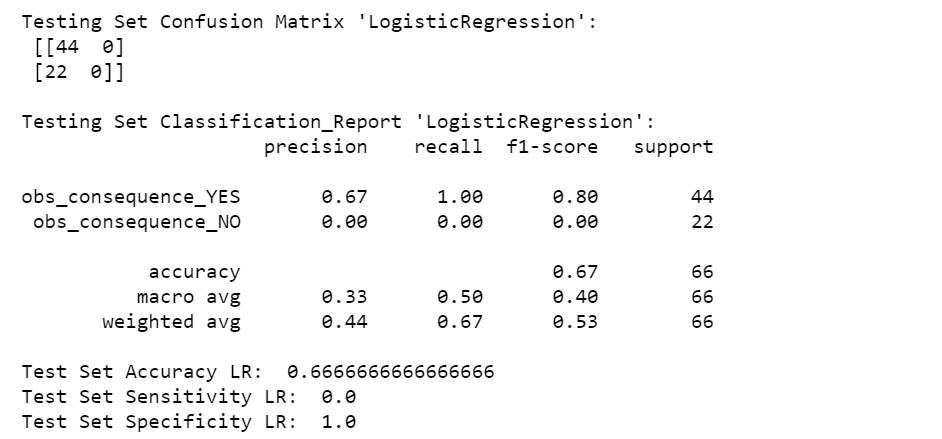
**Modeling:**

1. **Logistic Regression:**

I have selected ‘LogisticRegression’ algorithm for the prediction algorithm. When trained the model using training data and then validated using testing dataset.

The Accuracy of the model is 67% and when looking at the confusion matrix the TrueNegative and FalseNegative values are completely zero.

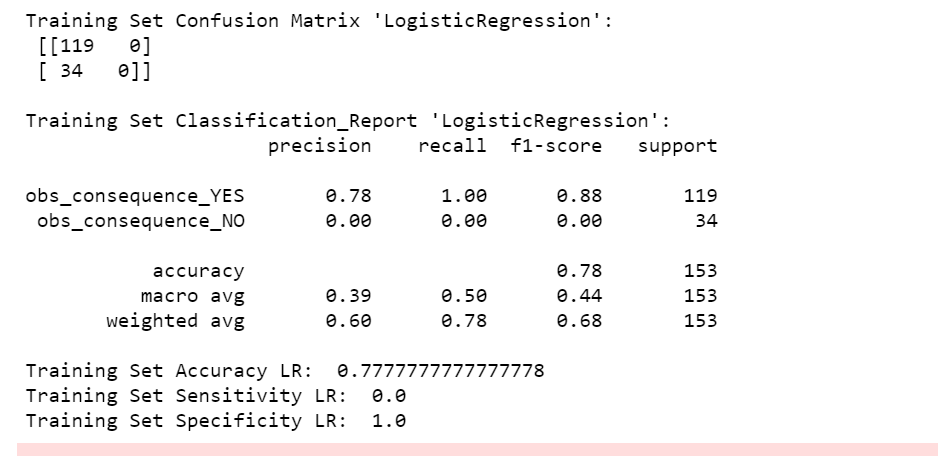
**Validation/Testing Dataset:**



**Training Dataset:**

Since the Accuracy percentage is very low, I tried to run the model with the ‘training’ to see if the model is skewed, and I got the same results with confusion matrix the TrueNegative and FalseNegative values are completely zero.

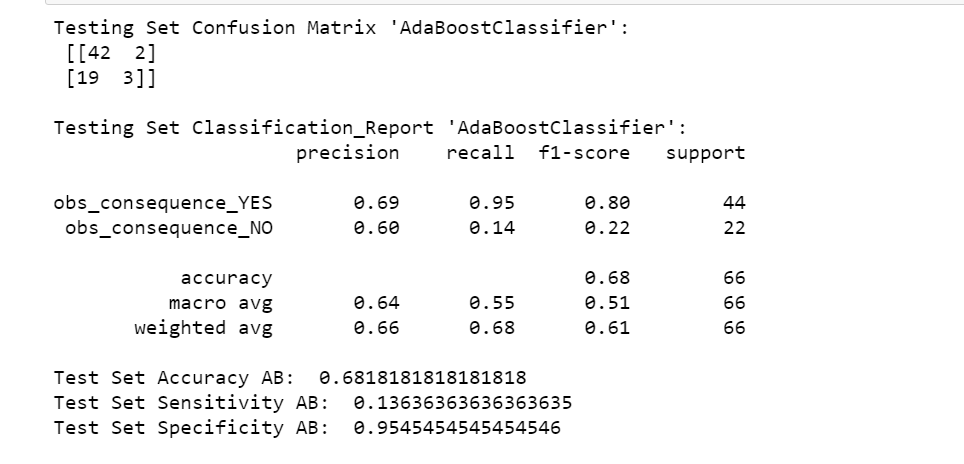
With this I have decided that the ‘LogisticRegression’ model is not appropriate for the given dataset.

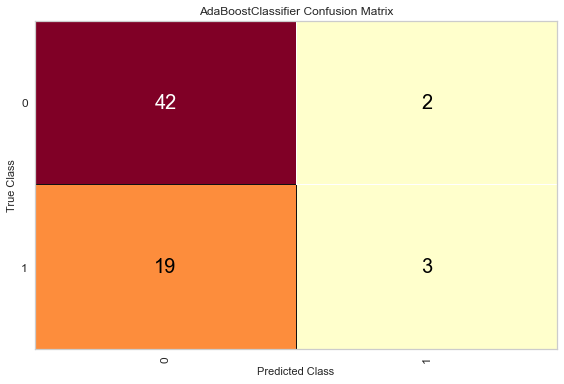


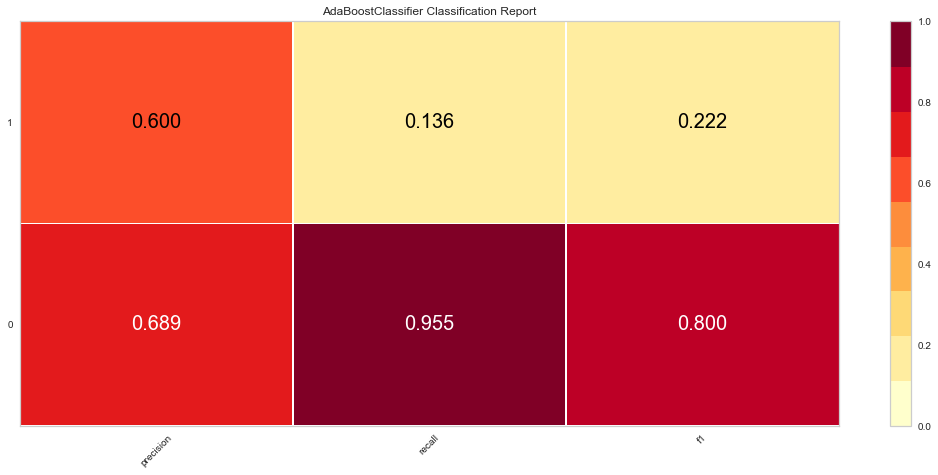
1. **AdaBoost Classifier:**

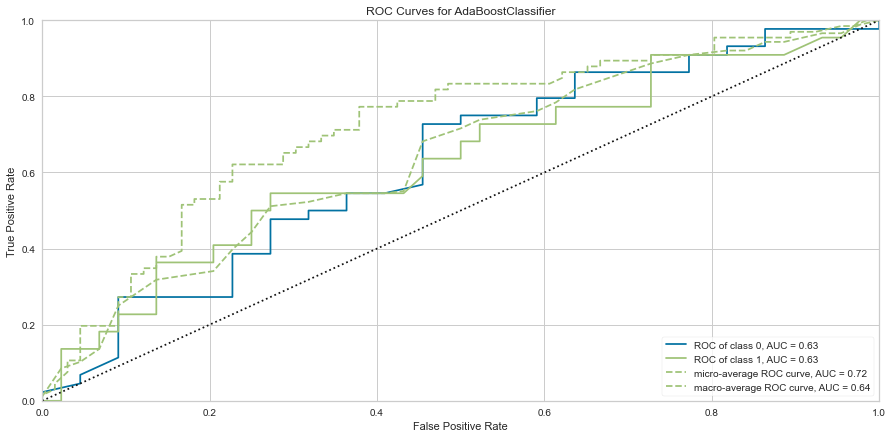
I have chosen ‘AdaBoostClassifier’ as an alternate algorithm to run the model and validate the results. With the ‘AdaBoostClassifier’ the accuracy improved compared to LogisticRegression algorithm.

After running the model with training data, validated with test dataset below are the results.



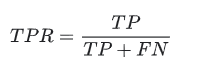
****

****

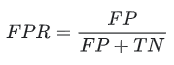
****

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds where curve plots two parameters. True Positive Rate and False Positive Rate.

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:



False Positive Rate (FPR) is defined as follows:



**Training Dataset**

To validate the model, I ran the model with training dataset and got the accuracy percentage of 86%.

Training Set Confusion Matrix 'AdaBoostClassifier':

[[111 8]

[ 18 16]]

Training Set Classification\_Report 'AdaBoostClassifier':

precision recall f1-score support

obs\_consequence\_YES 0.86 0.93 0.90 119

obs\_consequence\_NO 0.67 0.47 0.55 34

accuracy 0.83 153

macro avg 0.76 0.70 0.72 153

weighted avg 0.82 0.83 0.82 153

Training Set Accuracy AB: 0.8300653594771242

Training Set Sensitivity AB: 0.47058823529411764

Training Set Specificity AB: 0.9327731092436975

**Conclusion**:

I have calculated the Accuracy, Sensitivity and Specificity for all the models with training and test datasets to make sure the model is not skewed. After reviewing the models, the Accuracy improved slightly with the ‘AdaBoostClassifier’ algorithm.

**References:**

1. Centers for Disease Control and Prevention -Jan 22nd 2021- Mental Health and Coping During COVID-19. [Coping with Stress](https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/managing-stress-anxiety.html)
2. Alison Abbott - COVID's mental-health toll: how scientists are tracking a surge in depression. [Nature- vol 590-11 Feb 2021](https://www.nature.com/articles/d41586-021-00175-z)
3. Diegocalvo - Data mining of mental health: [Kaggle Dataset](https://www.kaggle.com/diegocalvo/data-mining-of-mental-health/data)

# Joshua Gordon - One Year In: COVID-19 and Mental Health- [Directors Messages](https://www.nimh.nih.gov/about/director/messages/2021/one-year-in-covid-19-and-mental-health.shtml)

## The Implications of COVID-19 for Mental Health and Substance Use- [COVID-19](https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/)