Name: Bhargava Gaggainpali

Course: DSC550 – Data Mining

Week11.2\_Final\_Project\_Original\_Case\_Study\_Final\_Documentation

**Title: Mental Health Prediction**

I have selected the data from the Kaggle on the topic ‘Mental Health Survey’ for the term project. When looking for a dataset for the project, this caught my interest. As recently I have received a form from my employer to complete the survey on similar Mental health conditions. Because of the Corona Virus situation, everyone is working remotely and restricting themselves not to go out other than to get groceries. Spending most of the time at home, now a days leading to depression and other mental health issue.

I would like to explore to see what factors are influencing the mental health conditions by relating different factors variables and find which categories of people are more effected.

Using the dataset, I am planning to Identify the variables which can determine the influencing factors of the mental health issue. Build and run the prediction models to find the similar category people who could possibly be facing issues and test the program with different age, gender, country, state, and remote working conditions.

Data Source: Kaggle

<https://www.kaggle.com/diegocalvo/data-mining-of-mental-health/data>

Variables: 26

Observations: 1260

Variables:

Timestamp - Survey Date and Time

Age - Age of the Participant

Gender - Sex of the Participant

Country - Country of the Participant

State - State of the Participant

self\_employed - Self-employment status

family\_history - Family History of Mental Health Issues

treatment - Previous Treatment for Mental Health Issues

work\_interfere - Mental Health issue because of Work interference

no\_employees - Number of co-Workers

remote\_work - Remote work job

tech\_company - Type of company

benefits - Company providing medical benefits

care\_options - Additional Care options

wellness\_program - Enrolled in any Wellness program

seek\_help - Currently seeking help

anonymity - Anonymity issues

leave - Currently on leave

mental\_health\_consequence - Facing Mental health issues

phys\_health\_consequence - Facing Physical Health issues

coworkers - Coworker facing any similar mental health issues

supervisor - Supervisor facing any similar mental health issues

mental\_health\_interview - Did go through any mental health interview

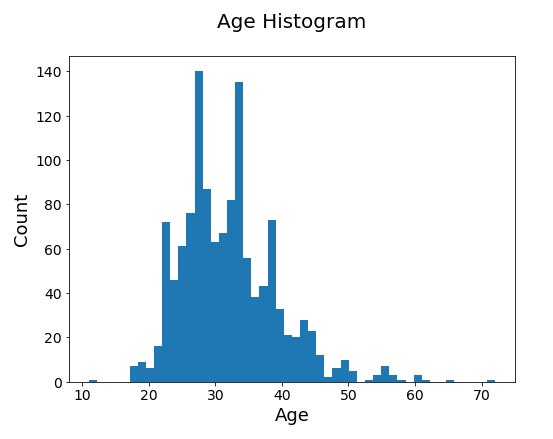
phys\_health\_interview - Did go through any physical health interview

mental\_vs\_physical - Did go through any mental vs physical health interview

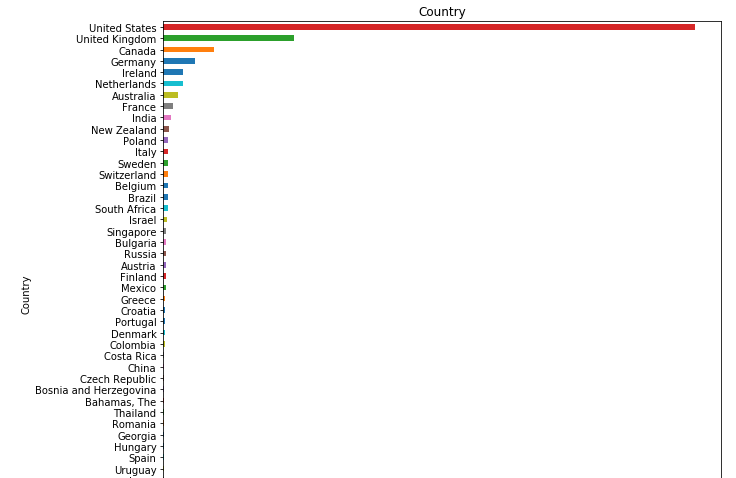
obs\_consequence - Facing any other consequence

comments - Additional comments

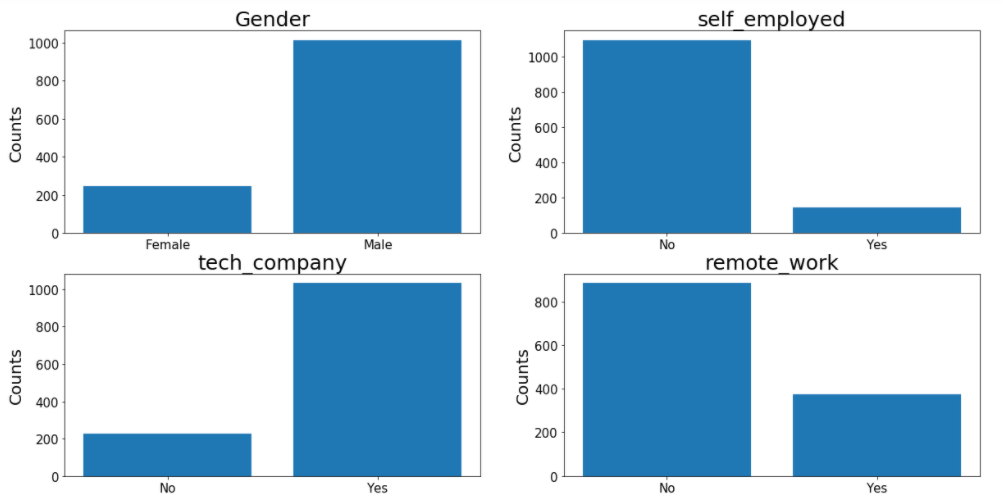
Representing the Histogram of the “Age” values in the dataset.



Representing the Bar Graph of the “Country” values in the dataset.



Representing the Bar Graph of the “Country, Self\_employed, tech\_company, remote\_work” values in the dataset.



Looking at all the variables in the dataset, I tried to find the data types of the variables to see how many are Numerical and Categorical for further analysis.

Variables: 26

Observations: 1260

Variables:

Timestamp - Survey Date and Time

Age - Age of the Participant

Gender - Sex of the Participant

Country - Country of the Participant

State - State of the Participant

self\_employed - Self-employment status

family\_history - Family History of Mental Health Issues

treatment - Previous Treatment for Mental Health Issues

work\_interfere - Mental Health issue because of Work interference

no\_employees - Number of co-Workers

remote\_work - Remote work job

tech\_company - Type of company

benefits - Company providing medical benefits

care\_options - Additional Care options

wellness\_program - Enrolled in any Wellness program

seek\_help - Currently seeking help

anonymity - Anonymity issues

leave - Currently on leave

mental\_health\_consequence - Facing Mental health issues

phys\_health\_consequence - Facing Physical Health issues

coworkers - Coworker facing any similar mental health issues

supervisor - Supervisor facing any similar mental health issues

mental\_health\_interview - Did go through any mental health interview

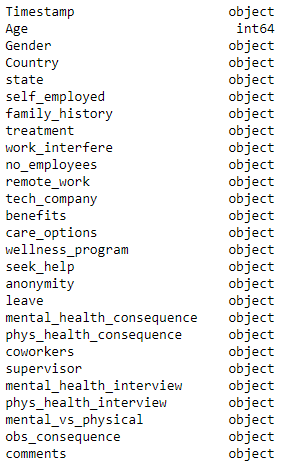
phys\_health\_interview - Did go through any physical health interview

mental\_vs\_physical - Did go through any mental vs physical health interview

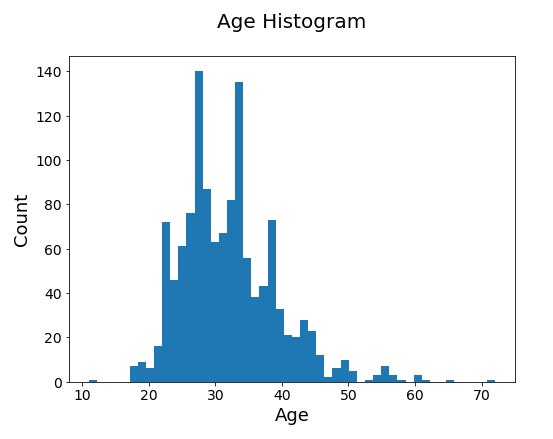
obs\_consequence - Facing any other consequence

comments - Additional comments

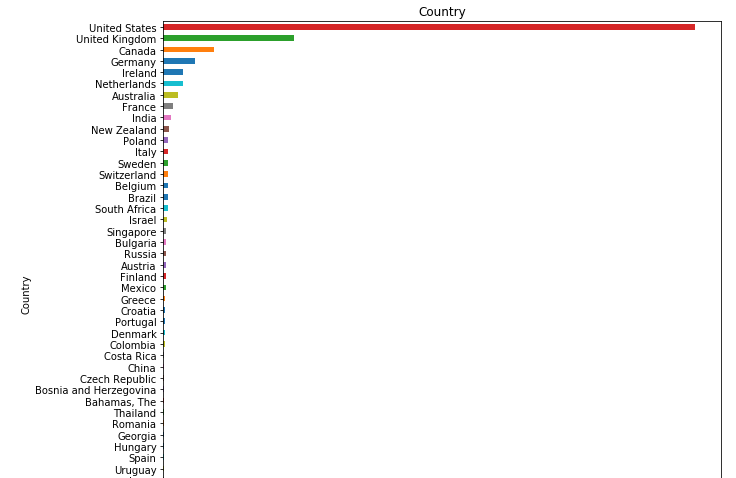
**Only one variable ‘Age’ is Numerical and rest all are is Categorical variables.**



**Representing the Histogram of the “Age” values in the dataset.**



**Representing the Bar Graph of the “Country” values in the dataset.**



**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)

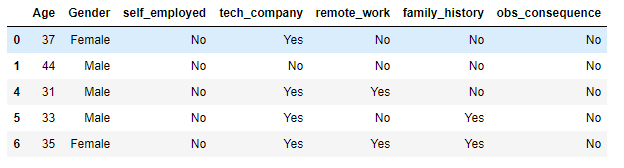
With this we have reduced the number of rows to 751.

-------------------------------------------------------------------------------------------------------------------------

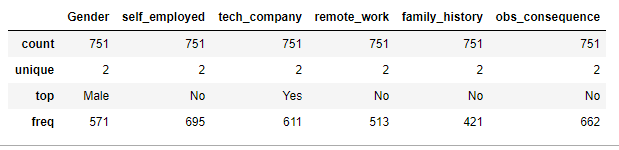
**Out of the List of the variables, we have selected only few variables for further analysis.**

**'Age', 'Gender', 'self\_employed', 'tech\_company', 'remote\_work', 'family\_history', 'obs\_consequence'.**

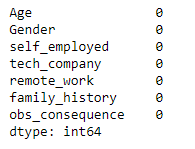
Looking at the sample data:



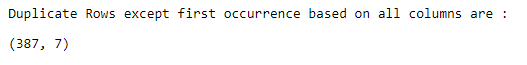
Displaying the stats of the categorical variables.



**Check for the NULL values in the selected variables. Found NO Null values.**



**When checked for duplicate values in the data frame, found 387 rows as duplicates.**



**After removing the Duplicate rows the Final rows remaining are : (219, 7)**

Number of elements in the Filtered dataframe (219, 7)

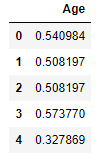
Number of elements in the Input Numerical dataframe (219, 1)

Number of elements in the Input Categorical dataframe (219, 5)

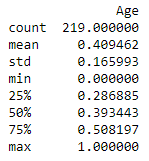
Number of elements in the Output Categorical dataframe (219, 1)

**---------------------------------------------------------------------------------------------------------------------**

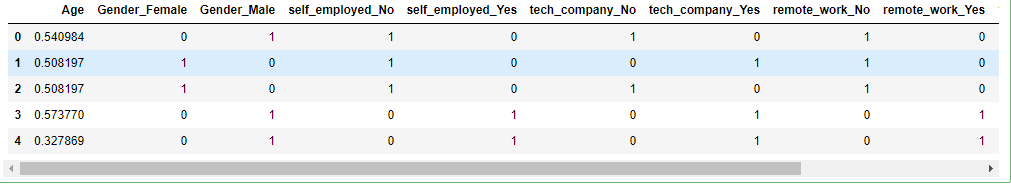
**Applied ‘MinMax Scaler’ to Converted the Numerical Variables ‘Age’ to convert to lower values.**



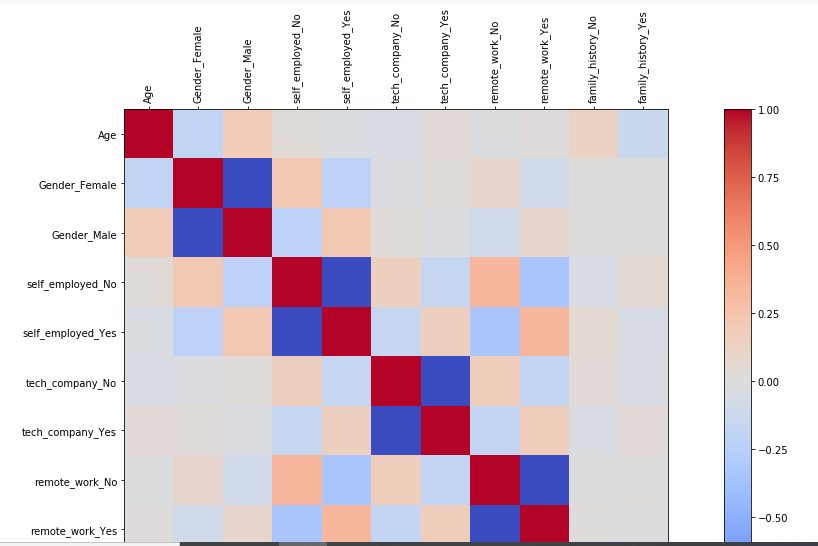
**Stats of the Variables after converting with MinMax Scaler.**



**Using ‘One Hot Encoder’ converted the Categorical values to Numerical values.**

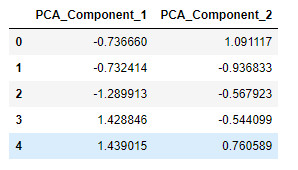


**Using Correlation Map, tried to find the values with much correlation.**

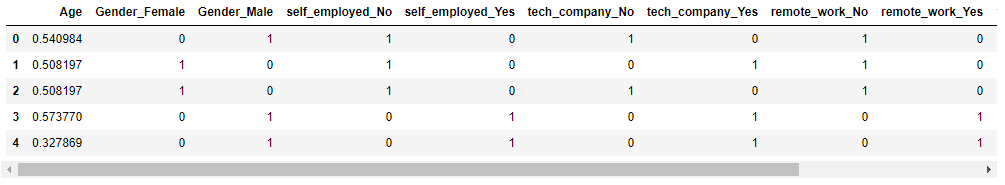


**---------------------------------------------------------------------------------------------------------------------**

**Using the PCA (Principle Component Analysis) converted 11 variables to 2 variables.**

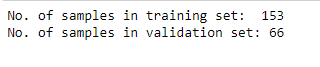


I have selected appropriate variables which has significant impact on the output variable as given.



Out of Total number of rows, I have divided data in to 70% vs 30**%** for the training vs validation datasets.

Given are the count of rows in the training and validation datasets.





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**

Testing Set Confusion Matrix 'LogisticRegression':

[[44 0]

[22 0]]

Testing Set Classification\_Report 'LogisticRegression':

precision recall f1-score support

obs\_consequence\_YES 0.67 1.00 0.80 44

obs\_consequence\_NO 0.00 0.00 0.00 22

accuracy 0.67 66

macro avg 0.33 0.50 0.40 66

weighted avg 0.44 0.67 0.53 66

Test Set Accuracy LR: 0.6666666666666666

Test Set Sensitivity LR: 0.0

Test Set Specificity LR: 1.0

**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.

Training Set Confusion Matrix 'LogisticRegression':

[[119 0]

[ 34 0]]

Training Set Classification\_Report 'LogisticRegression':

precision recall f1-score support

obs\_consequence\_YES 0.78 1.00 0.88 119

obs\_consequence\_NO 0.00 0.00 0.00 34

accuracy 0.78 153

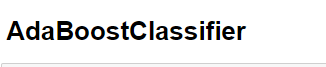
macro avg 0.39 0.50 0.44 153

weighted avg 0.60 0.78 0.68 153

Training Set Accuracy LR: 0.7777777777777778

Training Set Sensitivity LR: 0.0

Training Set Specificity LR: 1.0



I have choose ‘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. Given are the results.

**Testing Dataset**

Testing Set Confusion Matrix 'AdaBoostClassifier':

[[42 2]

[19 3]]

Testing Set Classification\_Report 'AdaBoostClassifier':

precision recall f1-score support

obs\_consequence\_YES 0.69 0.95 0.80 44

obs\_consequence\_NO 0.60 0.14 0.22 22

accuracy 0.68 66

macro avg 0.64 0.55 0.51 66

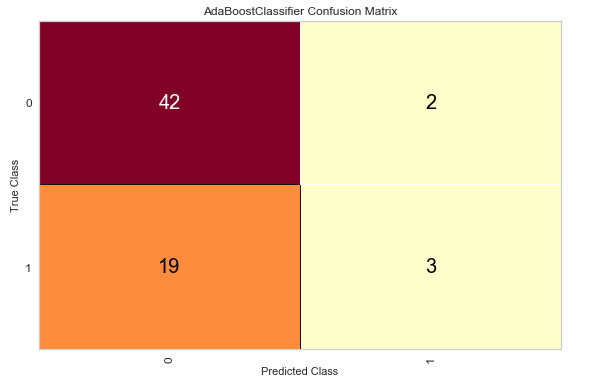
weighted avg 0.66 0.68 0.61 66

Test Set Accuracy AB: 0.6818181818181818

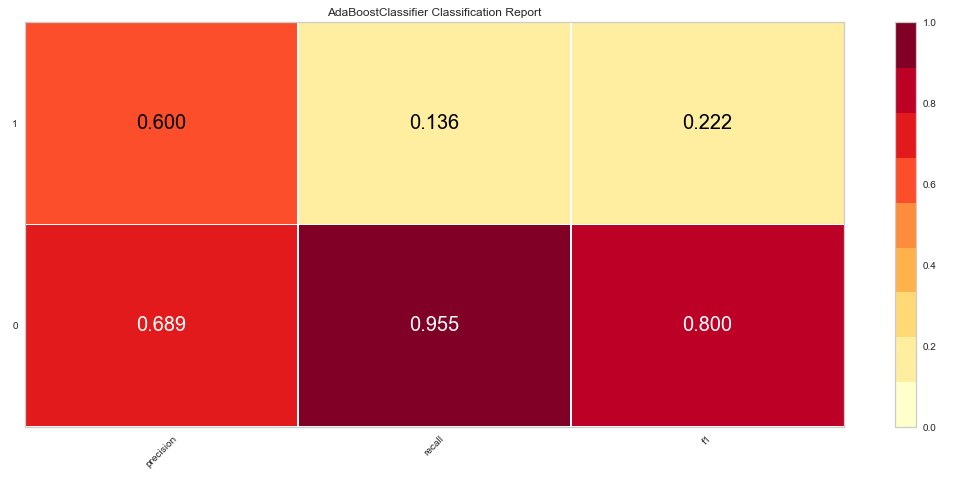
Test Set Sensitivity AB: 0.13636363636363635

Test Set Specificity AB: 0.9545454545454546

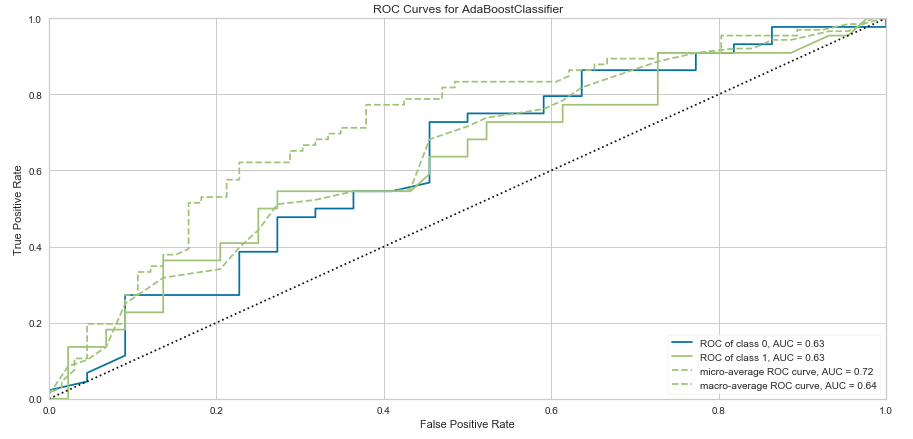
**Plot view of Confusion Matrix**



**Plot view of Classification Report**

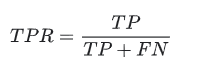


**Plot view of ROC Curves**

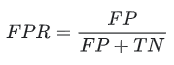


An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds where curve plots two parameters. True Positive Rate and False Positive Rate.

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 have 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 trying couple of models, the Accuracy improved slightly with the ‘AdaBoostClassifier’ algorithm.