**23 June 2020**

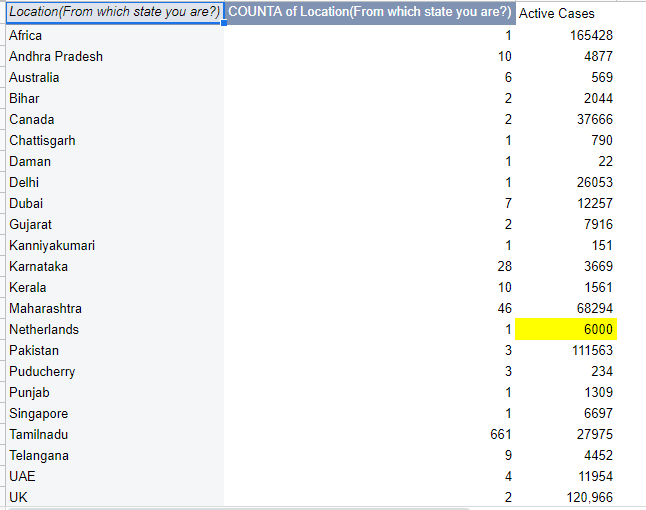
1)Filtered the location column

We made the different spellings of states uniform

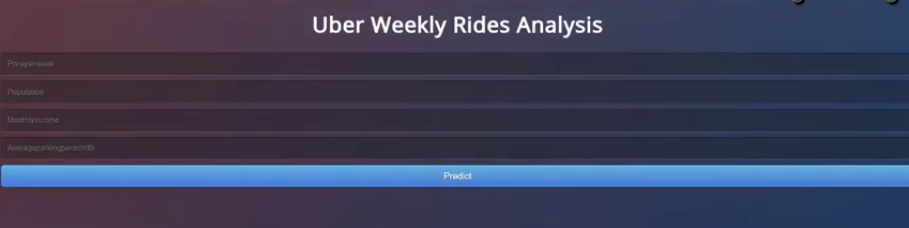
2)Created an anxiety column using GAD-7 Scale

<https://med.dartmouth-hitchcock.org/documents/GAD-7-anxiety-screen.pdf>

3)Found out the active cases in each state



4)Planned to a web app using Flask



Create an UI that asks the user for a set of questions to answer, determine the anxiety of that person and give appropriate suggestions.

5)Changed all columns to numerical data

Key used:

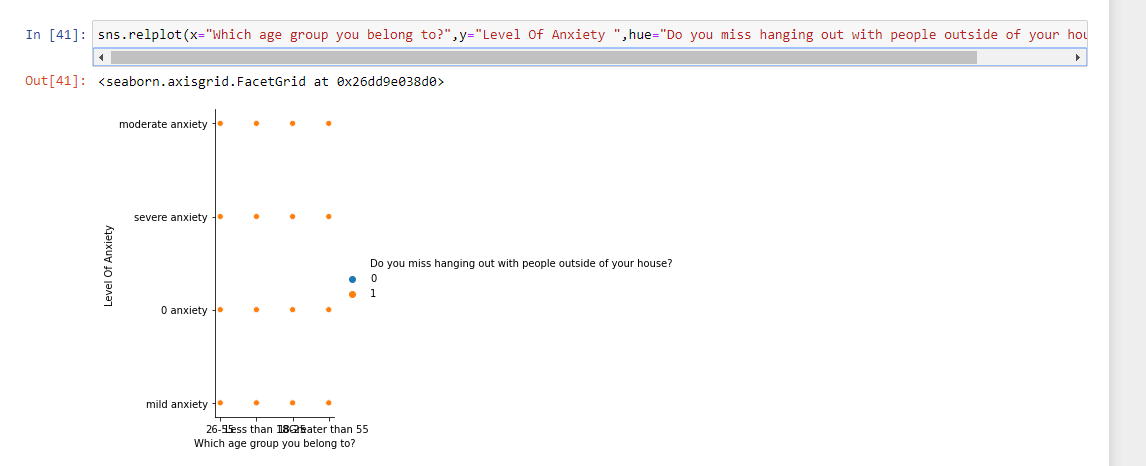
Frequently, Yes,Female- 1

Rarely,No ,Male- 0

**Exploratory Data Analysis(EDA):**

A lot of things to understand the data in a better way

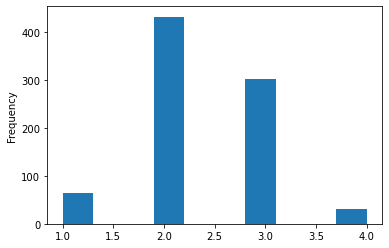
* data.shape=Rows,columns
* data.describe()=Shows mean,count,min,max,sd for each column
* data["Active Cases in State"].unique()=Shows all the unique values and the number of unique values
* data.isnull().sum()=Sum of null values in each column
* Pairplot-Shows the spread of data
* Rel plot=Attached pic



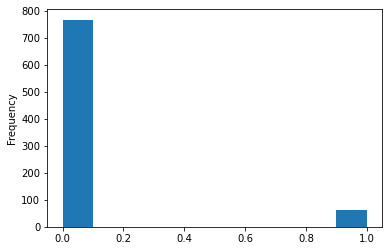
**DATA VISUALIZATION**

* The Dataframe’s histogram plot function was used to count the frequency for each feature.

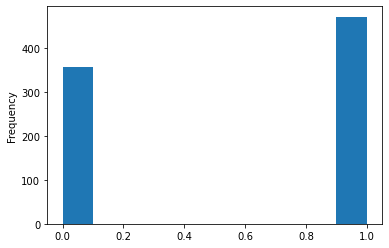
1. Age-Group



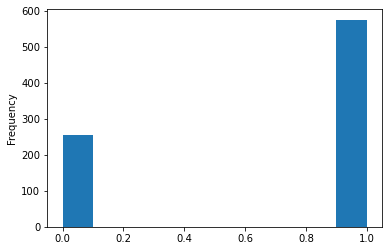
1. Is it difficult to get essentials(food/medicine)?



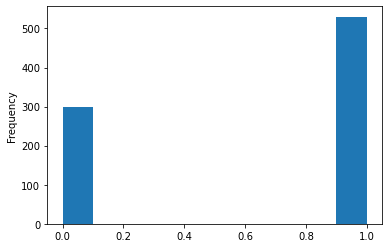
1. Gender



1. Increase in household work

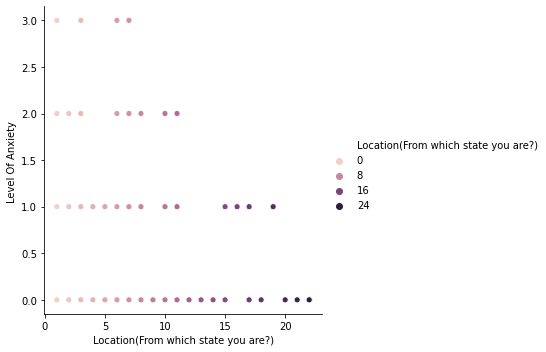


1. Usage of masks/hand-sanitizers



* Used Seaborn python module to visualize the data

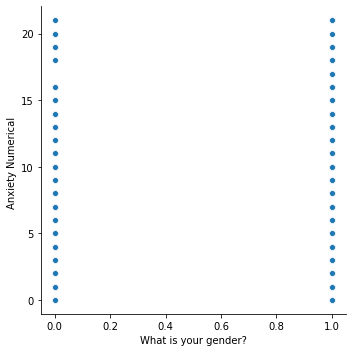
1. Location vs Level Of Anxiety



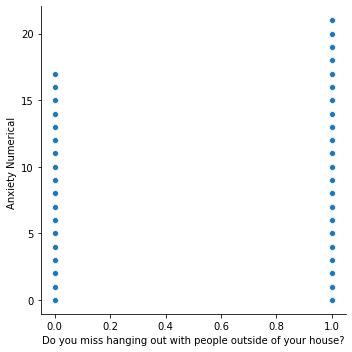
The encoding of the location is given below



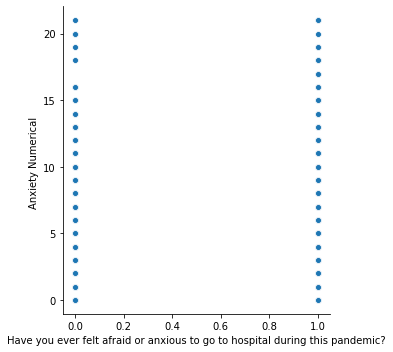
1. Gender vs Anxiety Numerical



1. Do you miss hanging out vs Anxiety Numerical

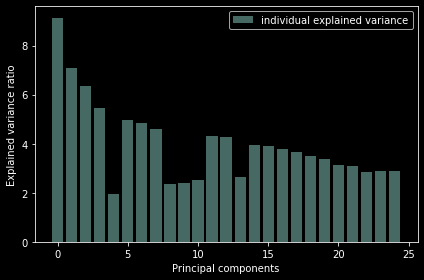


1. Anxious to go to hospital vs Anxiety Numerical



**PCA**:

Applied PCA



Here are the corresponding feature names ranging from 0 to 24 in the above graph

**'Which age group you belong to?'**

**'Location(From which state you are?)',**

**'Is it difficult to get essentials(food/medicine)?',**

**'What is your gender?',**

'How you ever felt symptoms of COVID? ',

**'Do you have any of the following conditions?'**,

**'During the lockdown have you been doing more household work?'**,

**'Have you or anyone you know have corona virus/recovered from COVID/been tested for COVID? '**,

'How often do you use masks/hand sanitizers(per day)?',

'Do you feel like you have wasted your time during the lock down or have you ever felt pressurized by your peer’s accomplishments during the lock down?',

'How often do you eat outside food? (per week) ',

**'Are you in a containment zone? ',**

**'Are you stuck somewhere alone without support from friends or family?** ',

'Do you have children less than 3 yrs, or elders 65+ at your house?',

**'Have you ever felt afraid or anxious to go to hospital during this pandemic?'**,

**'Has your sleep cycle changed drastically?',**

**'Have you felt frustrated by having your activities, major life events or opportunities affected by the virus? ',**

‘Is someone from your family in the medical field everyday?',

'Do you have online classes/assignments/assessments/tests? ',

'Did you face a pay cut/job loss/take a0ther job to cover expenses?',

'How often do you leave your house?(per week) ',

'How often do you interact with people outside your house(friends,neighbors etc.,)?',

'Have you taken any medication to prevent corona(Chloroquine or kabasura kudineer) or any medication to increase your immunity?',

'Have you been affected by not being able to visit religious places?',

'Do you miss hanging out with people outside of your house?',

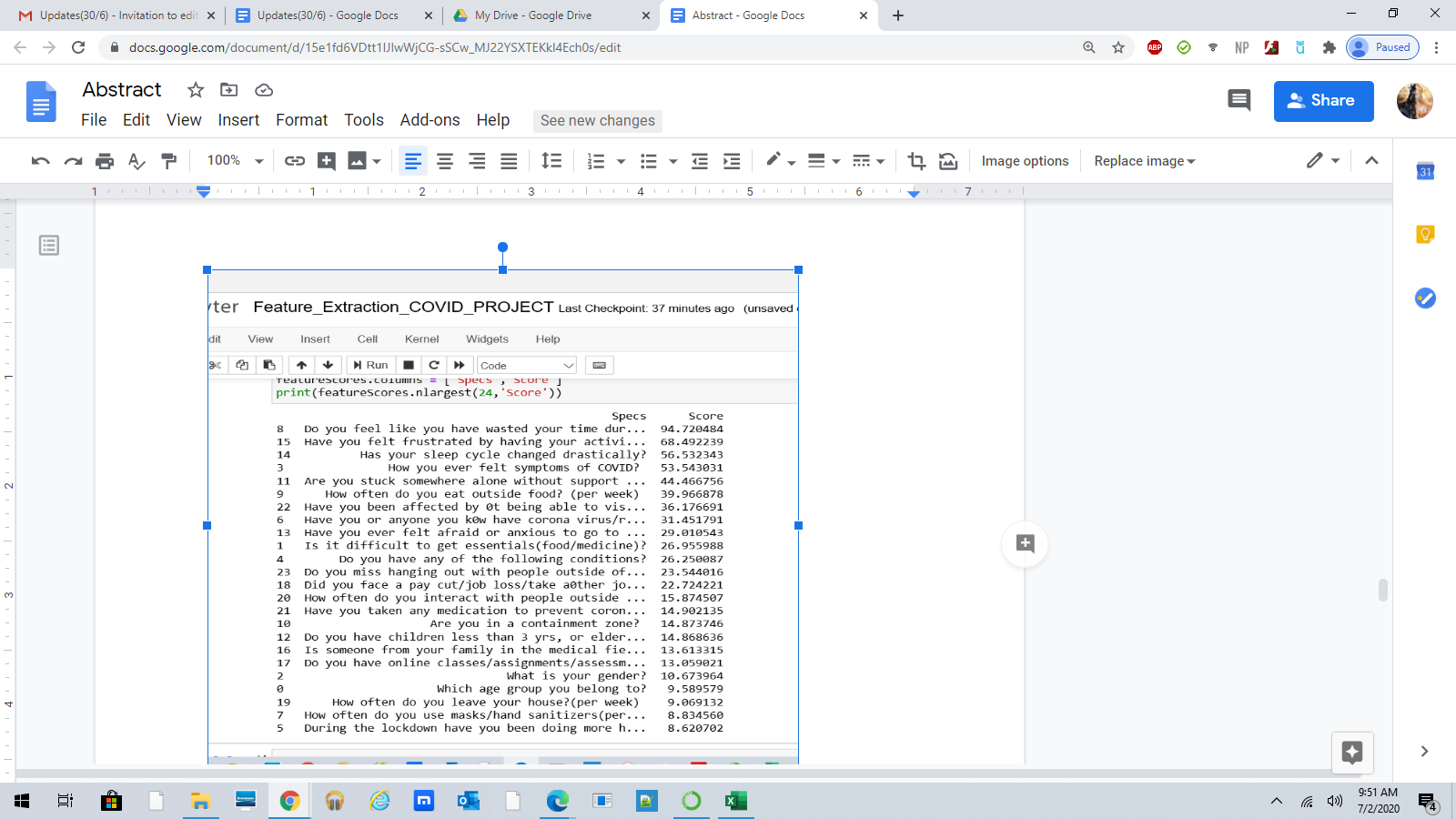
The Highlighted features are top 12 principal components suggested by pca

**Feature selection:**

We tried these feature selection algorithms:-

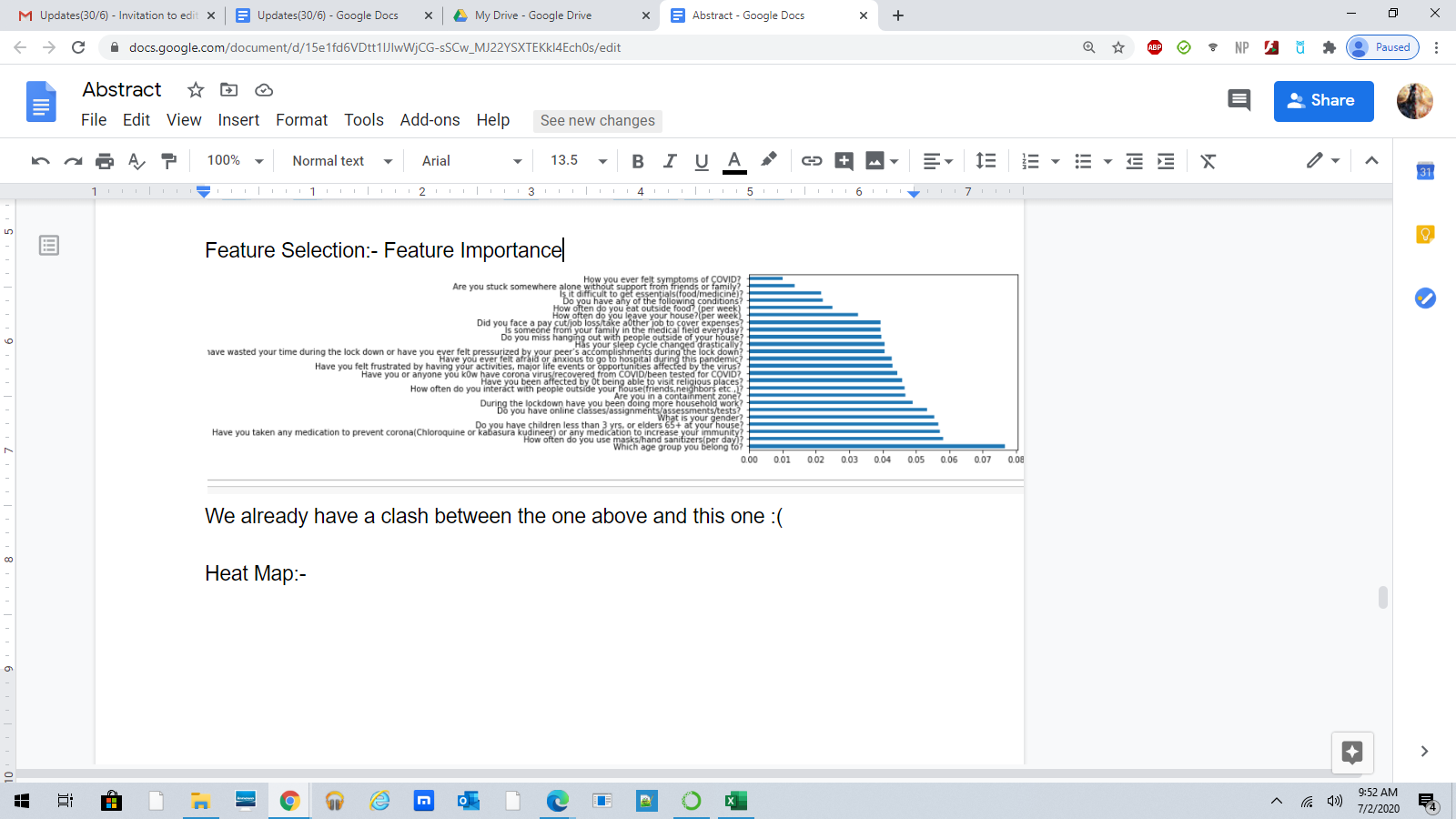
1. Univariate Selection:-

Which yielded:-



1. Feature Importance:-

Which Yielded:-



The issue here is that the most important features of both these algorithms should be the same but they aren’t. For example, the age group has a high importance according to the feature importance output, but it is one of the least important features when you look at the univariate selection. We moved on to the heat map which didn’t yield any conclusive results. Originally we tried to choose all the features that overlapped through various algorithms, but those proved to be too few, so we decided to move on to other techniques.

**Cross Validation:**

Tried Random Forests, SVM,Logistic Regression

Random Forests:

[0.53571429 0.56626506 0.4939759 0.59036145 0.53012048 0.62650602

0.59036145 0.48192771 0.56626506 0.48192771]

**Accuracy of Random Forests is: 55.2352266207688**

SVM:

[0.57142857 0.60240964 0.57831325 0.61445783 0.56626506 0.60240964

0.57831325 0.5060241 0.56626506 0.55421687]

**Accuracy of SVM is: 57.40103270223752**

Log:

[0.58333333 0.59036145 0.4939759 0.60240964 0.59036145 0.56626506

0.59036145 0.44578313 0.5060241 0.48192771]

**0.5490909090909091**

**25 June 2020:**

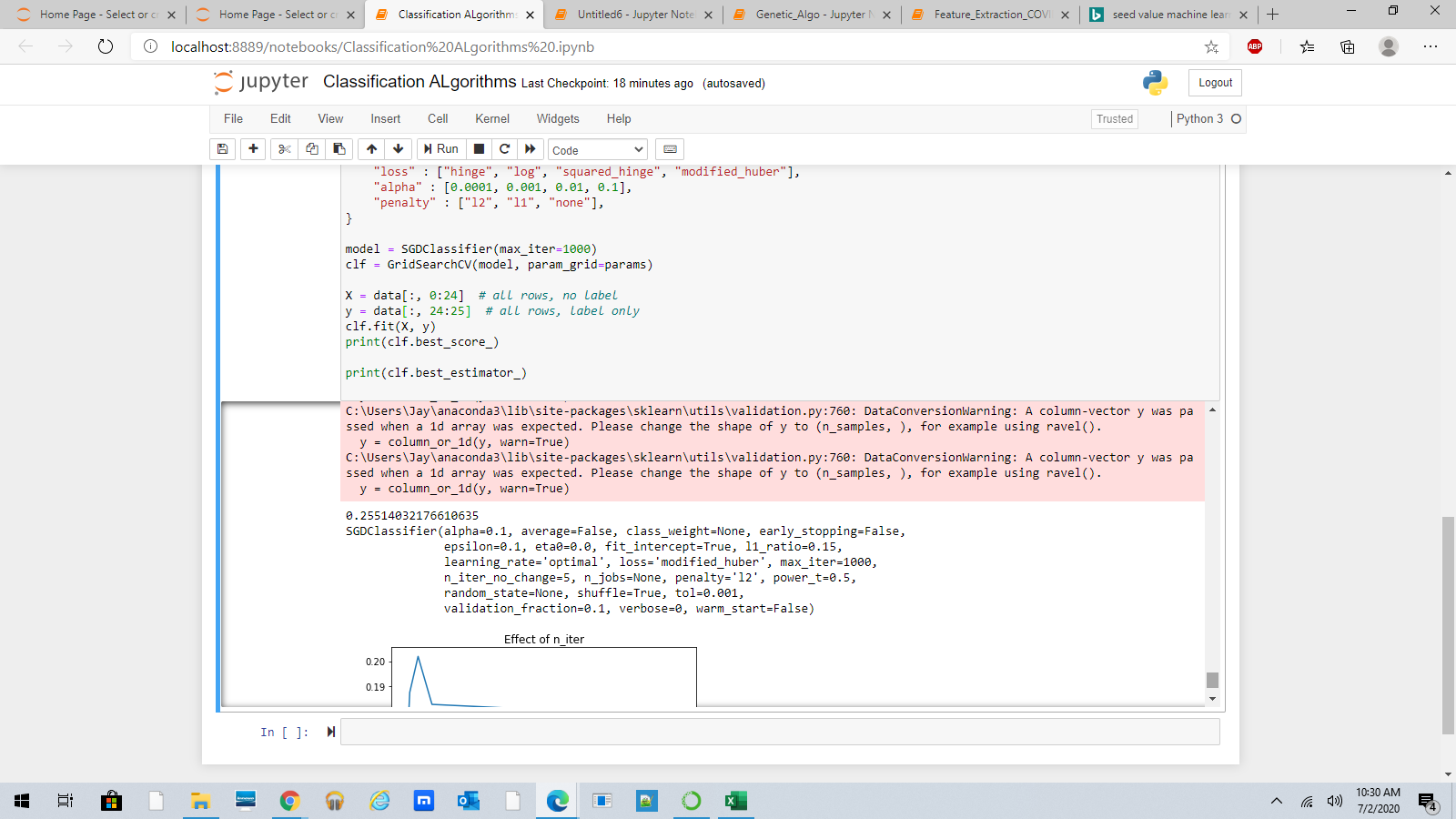
Since our cross validation accuracy was pretty low,we decided to proceed with all the types of algorithms and find out which one works best- Classification,Regression(as we have the anxiety numerical) and Clustering.

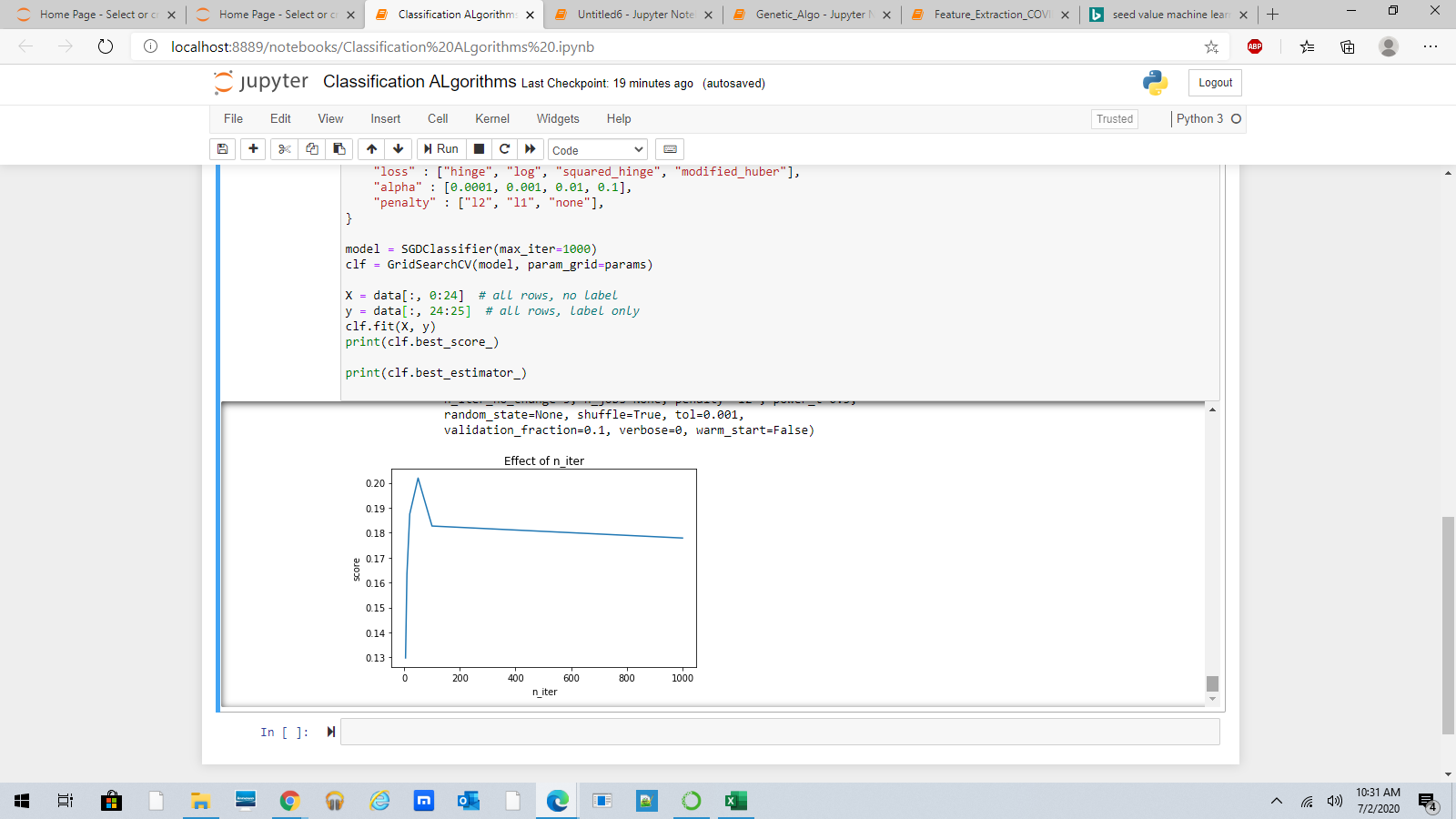
**26 June 2020 to 28 June 2020**

**CLASSIFICATION**

1. SGD Classifer(This was one of the classifiers recommended for data with less than 100K samples)

We found that the score was pretty low, only 0.255 and that the slope couldn’t be reduced as much as we hoped it would be:-

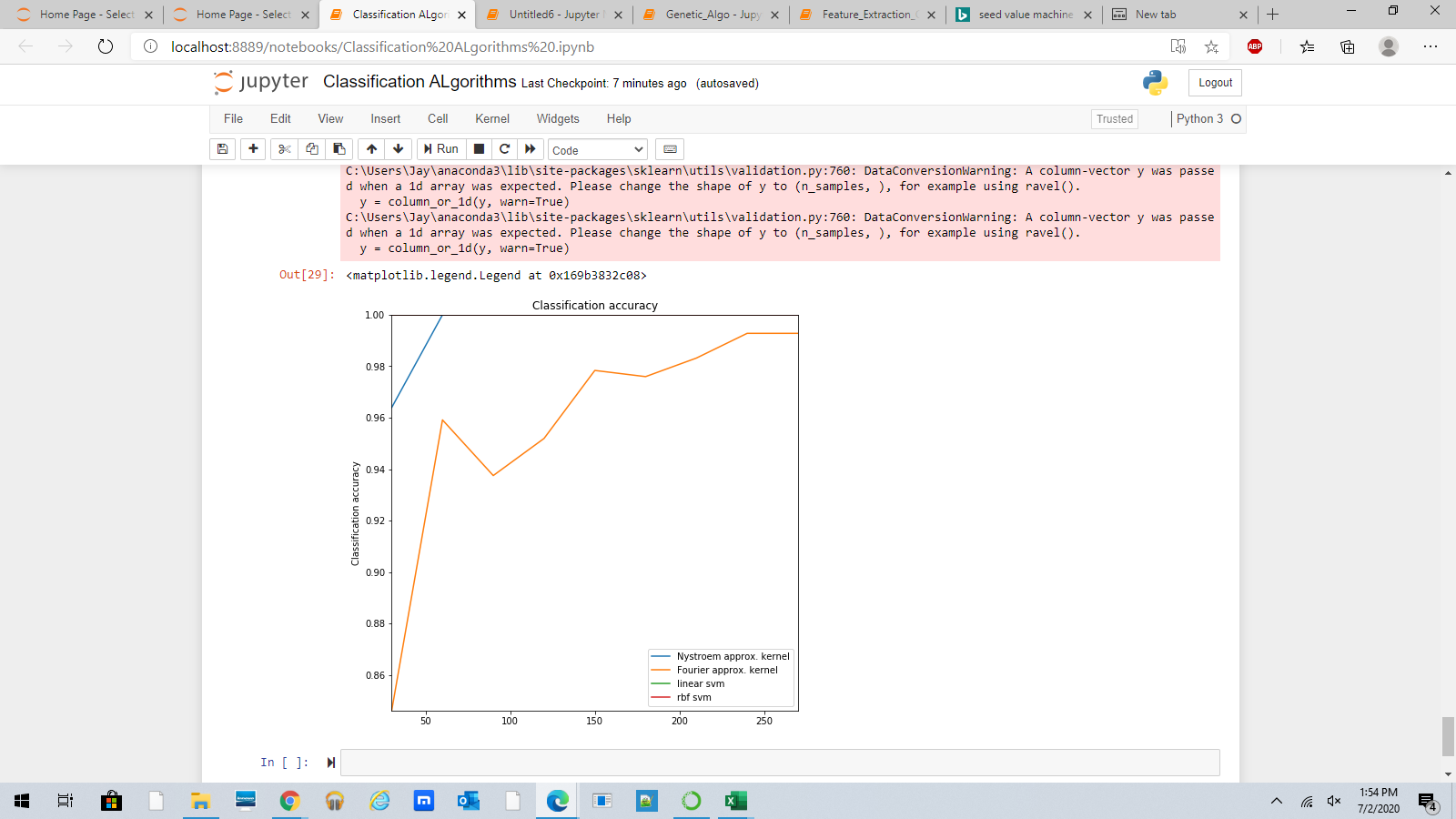




1. Linear SVC:-

Only yielded accuracy of 0.252... so was discarded

1. SVM Kernel Approximation:-



Neither linear svm nor rbf svm show up on the graph, and I couldn’t find the usage of Nystroem and Fourier with respect to classification algos.

**REGRESSION:**

**Feature selection:**

Tried Correlation Feature selection and mutual information feature selection.

**Algorithms**:

In the case of Linear,Ridge,Lasso and Elastic Net, both feature selection methods applied separately give the same rise in accuracy.

* Linear Regression

R squared value - 0.33

After applying feature selection, increases by 0.01( Selecting the top 10 features)

* Ridge Regression

R squared value - 0.35

After applying feature selection, increases by 0.01( Selecting the top 10 features)

* Lasso Regression

R squared value - 0.35

After applying feature selection, increases by 0.01( Selecting the top 10 features)

* Elastic Net Regression

R squared value - 0.35

After applying feature selection, increases by 0.01( Selecting the top 10 features)

* Regression trees

R squared value - 0.98

After applying feature selection,decreases to 0.6

* Gradient Boost regression

R squared value - 0.64

After applying feature selection, increases by 0.01( Selecting the top 10 features)

* Support vector Regression

Explained variance score- 0.27

After applying feature selection drops to 0.26

* Random Forests

Explained variance score-0.0883

After applying feature selection, it remains the same.

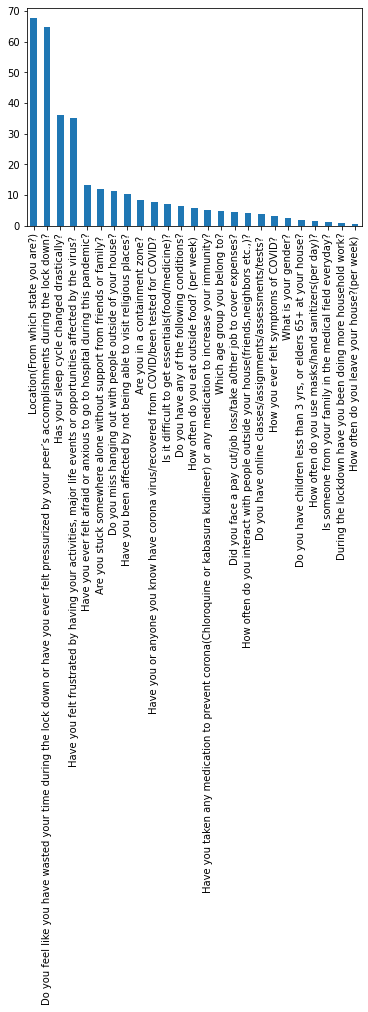
* ExtraTreesClassifier

Explained variance is negative

**CLUSTERING**

For feature selection we used Chi-Square method since the input and output are categorical

This is the output we got for Chi-Square feature Selection

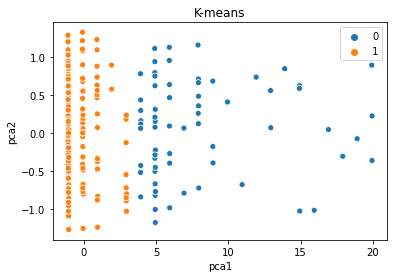


We took the top 12 features and tried applying k-means and Gaussian Mixture Model and tried with 2,3,4 clusters

* K-means

1. Directly used the categorical value and used principal components just for plotting the graph

K=2

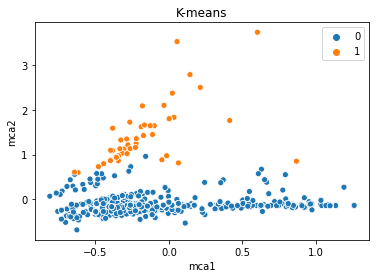


K=3

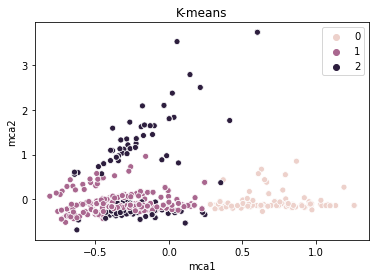
1. Applied MCA(Multiple Correspondence Analysis)

MCA is similar to PCA. It was advised to used MCA for categorical inputs.

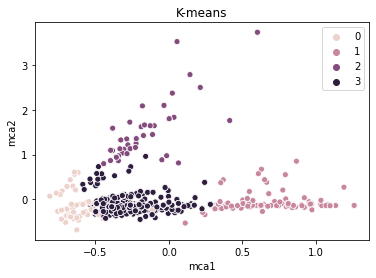
K=2



K=3

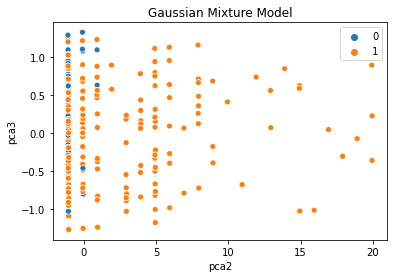


K=4

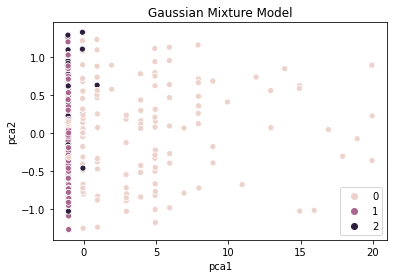


* Gaussian Mixture Model

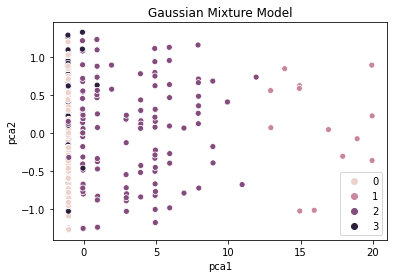
Number of components=2



Number of components=3



Number of components=4



**29th June 2020**

**Regression Trees:**

* We took the Random Forests to select features and recursively eliminated about 12 features.
* And both the R squared and explained variance did not have a drastic drop like in the earlier case (0.98 to 0.6)
* Finally, after dropping 12 features we currently have 13 features and
  + RMSE= 1.7913250837175132
  + Explained Variance =0.8513866491107251
  + R squared coeff value= 0.9043095027187834

**Genetic Algorithm:-**

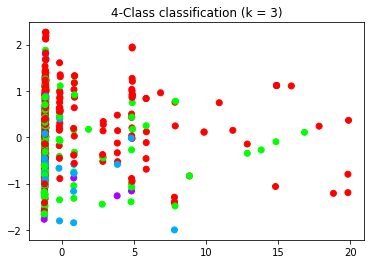
* We realized that our original idea was to implement a neural network using the most important features and giving those features various weightages, so we thought we would apply the genetic algorithm.
* It says online that the genetic algorithm can do one of two things:- 1. Help choose most important variables

2. Assign weightages to most important values

→ The algorithms for selecting most important variables yielded the CV MSE before and after value, which didn’t help with selecting the important factors, so we decided to stick with the most important features selected through random forest.

**K-Nearest-Neighbour Algorithm**

* We used the K-Nearest Neighbour algorithm. Splitted the dataset into train and test set using train\_test\_split() function.
* The model predicts the level of Anxiety(i.e., No,Mild,Moderate or severe anxiety)
* The number of neighbours we chose is 3. And tested the model against the test set.

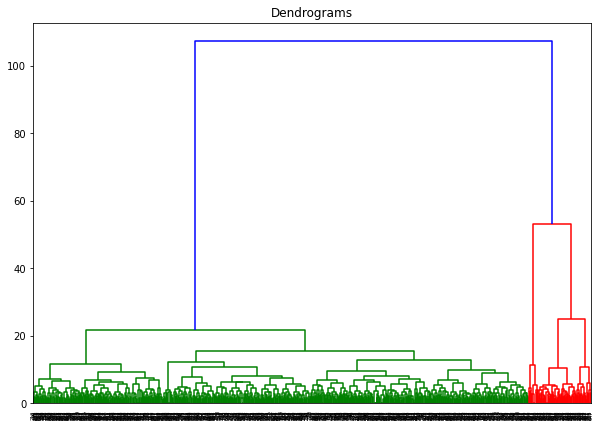


Accuracy obtained=0.59(59%)

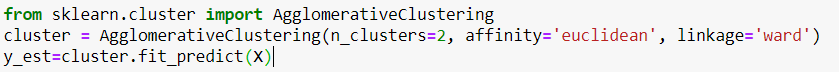
**Hierarchical Clustering**

* We saw that K-means clustering creates clusters of same size and we want to decide the cluster size at the beginning.
* So we used **Agglomerative Hierarchical Clustering** which takes each data point as a separate cluster and after each iteration it merges the nearby datapoints.
* We used Denograms to record the sequence of merge or split

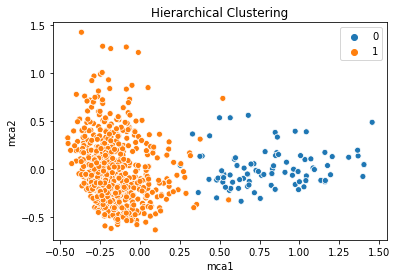
Here is the output that we got from denogram



* Since the denogram splits into two clusters we took the cluster size to be two.



This is the graph obtained.



We split up the data into 4 age groups

I took less than 18 and greater than 55

Preetiha took 26 to 55

Krithika took 18 - 25

Less than 18 and Greater than 55

I did a few analysis techniques in Python (correlation and group by) and a few in excel and dropped some features and then I applied feature importance to drop 2 to 3 features more and I applied regression trees, KNN, Decision tree classifier, K means, Linear regression

The classification yields a maximum accuracy of 73% and in regression I had to fit all the samples to the regressor as there were very few samples in both the age groups.

The R squared value is high.

But the model overfits and did not perform well when I gave entirely new data.

And in classification, since the number of samples corresponding to severe and moderate anxiety are very few the model finds it difficult to predict both these classes.

When I remove these two classes and create a model with only 0 and mild anxiety the accuracy is 87%.

But severe and moderate anxiety can't be generalized with 2 or 3 samples.

K Means didn’t have a great accuracy as well.

Factor analysis: It gives a number of factors and the amount to which it reduces the variance of the op variable. It condenses features and makes factors. But finding which factors corresponds to which features is one hell of a task and nearly impossible so framing questions for the web app becomes difficult.

LDA - dimensionality reduction technique after applying LDA if I apply basic algos like KNN, random forests it works better. It got a 70% accuracy with less than 18. But again, its predicting both moderate and severe wrongly.

I tried the algos that convert multi class (Binary relevance, Chain classifiers) to binary classification and got a max of 55% accuracy

There is a feature selection method called reliefF which is highly recommended for multi class problems, but I couldn't get the selected features finally.

The problem is that when we build a model the maximum, we can get is a 70% accuracy anything more than that is overfitting the data.

We did some visualizations as well. We could not get a lot of intuitions from it as the dataset is too diverse.

We have tried most of the feature selection techniques and most of the algorithms and so far, nothing yields great results.

And in my age groups, there weren't any features that was applicable to a specific gender. Men have been household work and have anxiety. Women too have been taking pay cuts. So, I was not able to narrow down features based on a gender.

We got the idea of grouping the features into 7 questions and map it to 7 GAD questions and use the grouped features against the op anxiety:

The grouping

Essentials and face a pay cut - economic fear

Have you or anyone u know have covid and containment zone - Fear of contracting covid

Masks

Immunity and eat outside - precautious

Wasted time and felt frustrated - Loss of opportunities

Containment zone

Stuck somewhere

How often u go outside

Religious places

- isolation

Children less than 3

Medical field everyday

Felt anxious to go the hosp

- Fear of protecting family

Miss hanging out

How often u interact

Miss social life

Got a max accuracy of 63% when we are using the GAD 7 scale as output

Whereas when we use scales 0 to 4 and then add it up and create our own op, we can get accuracy greater than 80%