Introduction

* Pandemics affect billions of people in turn around the world in their daily lives, places of work, or communities, thus creating a serious public health emergency together with a threat to mental health. Therefore, critically examining the complex relationship between such crises and mental health becomes a matter of importance as the world is faced with the devastating effects of the COVID-19 pandemic.
* Our objective is to shed light on the mental health issues that many individuals faced during the COVID-19 pandemic using data on anxiety and depression. To provide advice on how to prevent encountering such disruptions in the future, our aim is to recognize the trends and origins of these feelings. Through our research, we are developing a model to predict and mitigate similar crises, consequences on mental health. Our ultimate goal is to provide individuals with the knowledge and tools necessary to maintain mental resilience in any challenging circumstance, not just pandemics.

Dataset

* This dataset was provided by data.cdc.gov and is last updated on 21 March 2024. The Household Pulse Survey was established by the US Census Bureau in collaboration with five other government agencies to collect data on the socioeconomic effects of the COVID-19 epidemic on American homes. This creative survey sought to determine how the pandemic affected areas including employment status, spending patterns, food security, housing circumstances, interruptions to schooling, and general physical and mental health.
* The data consists of 15157 rows and 14 columns, and the data is collected in between 04/23/2020 and 03/04/2024.

Dataset Variables

* The most important feature of our analysis are variables as these are the main characteristics that help us to achieve or predict the insights that we want to capture or uncover. In this dataset, there are 14 variables with Indicator, Group, State, and Subgroup being categorical variables, Phase, time period, value, Low CI, High CI, confidence interval and quartile range being numerical, and Time period start date, Time period end date being date -time categorical variable.
* The indicator variables contain the information about the symptoms of depressive order, anxiety order, and both by each group such as age, disability status, education, gender identity, race/Hispanic ethnicity, sex, sexual orientation, state, and national estimate. Below is the image that indicates that almost 60% of the observations that are present in the dataset are mainly based on or grouped on state.

Data Reduction

* we can get rid of unnecessary variables and reduce the dimensionality of our dataset which makes our data readable in a faster way. When removing the variables, it’s important to be careful enough that we do not exempt the variables that are crucial for our modeling or analysis.
* As per the research questions that we are working on and the dataset, Confidence interval and Quartile range are the two variables that are unnecessary for the research among the 14 variables. Also, the quartile range variable has the greatest number of blanks or missing values. So, removing these two variables enhances the dataset better and helps us in improving efficiency of the analysis.
* Leveraging Microsoft Excel tool, we have reduced the data by removing the two unnecessary columns such as Confidence interval and Quartile range. Below is the screenshot of the dataset after performing data reduction.

Data Cleaning

* Getting rid of observations that has missing values or blank values, variables that have values with inappropriate format, removing duplicate observations, and identifying is there any variables with inconsistencies and correcting them.
* Handling Inconsistencies: using Microsoft Excel replace option, we have replaced the date value with empty value. From the figure, we can see that inconsistency error in the Phase variable has been handled perfectly without removing or deleting any observations.
* Handling Missing Values: filling the missing value with a default value or with an average value or most repeated value may make our dataset more biased or imbalance. Considering this point, we would like to remove the entire observation that has a blank or missing value and then perform analysis on the cleaned dataset.

Feature engineering

* Transforming our data is a critical step of the data preparation process. In this step, we perform normalization of the data, scaling of the data, feature engineering or creating new feature/variables from the existing variables, encoding the variables, identifying the outliers and many more depending on the research questions and the data that we have.

Outliner Detection

* Detecting outlier is also a crucial part of the data analysis project. Outliers indicate data points that are present very far from the other data points. So, having an outlier in our data impacts our analysis and modeling. Sometimes outliers are good and help us in enhancing the results and providing better insights and uncover or understand the patterns.
* The box plot identifies that there are outliers present in the variables such as Value, Low CI, and High CI. After carefully observing the values in the dataset, we had identified that the values in these 3 variables do not have any outliers and it is just data that has been starting from least value to high value. So, it is not necessary to remove these outliers as removing these outliers may remove the whole data, information and insights related to the Value, High and Low confidence intervals for depression disorder and anxiety disorder.

Correlation Analysis

* Correlation Analysis is very important to know the correlation or linear relationship between the variables. As we have 24 variables after performing data transformation, it is very important to know the relationship between these variables as we are performing analysis on these variables. Correlation can be performed on both numerical and categorical variables, but usually it is mainly performed to find the strength and relationship between the numerical variables. We can also perform correlation analysis on categorical variables but using Chi-Square test.
* The correlation plot suggests that Phase and Time period, Low CI and High CI, and Time Period and Year have strong positive correlation between each other.
* The table indicates that the ‘By Age’ and ‘By State’, ‘By Race/Hispanic ethnicity’ and ‘By State’, ‘By Sex’ and ‘By State’ are strongly correlated with each other.

In which Age

* This bar chart indicates that the people between the ages 18-29 years are the most common ones to have anxiety and depressive disorders and people who are 80 years of age and above are least likely to have anxiety and depressive disorders. Even the visualization indicates that the lesser the age group, the more percentage of people likely to have anxiety and depressive disorder and the more the age groups the lesser they have these disorders and are mentally healthy enough.

How does pandemic affect

* This line chart indicates that during the start of the pandemic the anxiety and depression disorders are around 20% to 30%, whereas during the peak stage of the pandemic the percentage is in between 25% to 40%, in current year, the percentage of anxiety and depression disorders had a substantial dip which is between 10% and 20%. In addition to that, the line chart also indicates that the symptoms of anxiety disorder are more commonly seen in the people rather than symptoms of depressive disorder over the time.

Do any states or areas

* This bar chart indicates that the national average percentage of these disorders is 28.62% and there are 23 states from 50 states in the United States that have greater than or equal to average prevalence of the anxiety and depression disorders. To be precise, among all these 23 states, Louisiana is the top state that has people who have the symptoms of anxiety and depression disorders the most with 34%, Mississippi being the second top with 33% and Nevada in the top three with 32% approximately.

Predicting the Strongest

* The accuracy of the random forest model for depressive disorder and anxiety disorder are 67.69% and 66.10% respectively which is a good indicator that our model is performing well. But the overall combined accuracy of random forest model is 33.79% which is very low, and it indicates that the complexity of the model in predicting both the target variables as each label is low.
* The accuracy of the decision tree for anxiety disorder is 66% which is a good indicator that our model is performing well. For the Anxiety order class, we got a precision value of 0.66 which suggests that the model is correctly predicting the positive class. The recall value of 1.00 suggests that the recall value is very high and indicates that the random forest is identifying all the actual instances of the depressive order successfully.

Predicting the livelihood

* The accuracy of the random forest model for depressive disorder and anxiety disorder are 67.69% and 66.10% respectively which is a good indicator that our model is performing well. But the overall combined accuracy of random forest model is 33.79% which is very low, and it indicates that the complexity of the model in predicting both the target variables as each label is low.
* The accuracy of the decision tree for anxiety disorder is 66% which is a good indicator that our model is performing well. For the Anxiety order class, we got a precision value of 0.66 which suggests that the model is correctly predicting the positive class. The recall value of 1.00 suggests that the recall value is very high and indicates that the random forest is identifying all the actual instances of the depressive order successfully.

Predicting the future

* Once the model got trained and fit with 20 epochs, the actual and forecasted values of ‘Value’ variable of 211 days from the test data have been plotted as below figure. The orange/predicted value line and blue/actual value line in line graph are like each other which indicates that the LSTM model has correctly predicted the actual values. Thus, indicating that our LSTM model has perfectly predicted the Values.
* The actual values of ‘Value’ variable for the next 15 days has been predicted and forecasted as seen below. The plot indicates or suggests that the value of depression disorder and anxiety disorder indicators will have rise in the next 15 days as the line was slowly increasing with the highest predicted value on 2024/02/15 with a value of 16.5 approximately. As the minimum value of the Value variable is 13.0 and maximum value being 16.5, there can be many people that are dealing with anxiety and depression disorder symptoms.

Interpretations

* From the results we can see that most of the cases are arising mostly because of the anxiety disorder, and it is playing crucial role for the addition of other disorders. Anxiety disorder is very high at the young age groups of 18 – 29 which is suggesting that the upbringing of the young adults, lifestyle choices and demographics plays a vital role for affecting these changes that cannot be cured but controlled. The data from the visualizations and characteristics from the predictions suggest that the awareness about the mental health has to be spread enough so that it can be controlled at the right stages and can be of better help for those who are facing these consequences from the disorders.

* The results are also suggesting that the states which are having huge party cultures and lavish lifestyle is also an indicator which should be looked after. From the visualizations, Louisiana and Nevada being on top suggesting that the usage of substance might also lead to these adverse effects resulting in disorders, But on flip side Colorado has the lowest prevalence suggests that it is not the substance usage primarily but the culture and environment which is affecting young adults to these factors, The laws should be made stringent and awareness from the right age can help people in thinking better which results in making correct lifestyle choices.

Summary

* The pandemic has seemingly profound impact on everybody physically and mentally, From the data we have gathered we can see that the results are pointing towards pandemic where all of us haven’t expected this type of situation to be happening all over the world and it made an impact to the young adults, As they used to live independently and also with lockdown with no options of social life, Already the pandemic has been affecting physically it has led to mental health which has become a piping reason. The sudden change of pandemic has brought everyone to be always cautious and attentive now more than ever to take care of our well-being and this lateral shift of pandemic with remote work culture, no in-person activities, only living at home has become very hard at first but it seemed like a better option later. But this shift has unknowingly created awareness on mental health and lot of organizations incorporating these changes helps others and routine checks around these mental health campaigns helps organizations to perform in sync with the employees.

Limitations

* The scope of the dataset is set to be within the limits of the years ranging from 2020 – 2024 and the constant updates for this dataset might cause our analysis to be working only for the above-described variables and parameters, Though the extent of the project was vast the demographics and the location preferences will not be accurate with the constant updates and this dataset is focusing on disorders and mental health aspects which are impacting with pandemic. Though the range of the dataset is till 2024 primary focus of the project is to correlate the physical and mental aspects of the individual at the time of pandemic and to the normal years where it can predict in such a way that the results can be more insightful and useful.