**Final Paper**

**Beyond Infection:**

**Predicting Mental Health Impacts in Pandemics through Data Analytics**

**Mani Sai Deeraj Marru**

**Sai Chaitanya Avasarala**

**Sai Kamal Reddy Chintaguntla**

**University of North Texas**

**Toulouse Graduate School**

**ADTA 5940**

**Dr. Denise Philpot**

**Spring 2024**

Contents

[CHAPTER 1: INTRODUCTION 3](#_Toc165384431)

[EDA Questions 4](#_Toc165384432)

[Predictive Questions 4](#_Toc165384433)

[CHAPTER 2: LITERATURE REVIEW 5](#_Toc165384434)

[Introduction 5](#_Toc165384435)

[The Impact of Past Pandemics 6](#_Toc165384436)

[Predictive Modeling for Mental Health Outcomes 7](#_Toc165384437)

[Geographic Variations in Mental Health Impacts of Pandemics 8](#_Toc165384438)

[Longitudinal Studies of Mental Health During Pandemics 9](#_Toc165384439)

[Conclusion 12](#_Toc165384440)

[CHAPTER 3: METHODOLOGY 14](#_Toc165384441)

[Data Collection 14](#_Toc165384442)

[Variables 15](#_Toc165384443)

[Data Preparation 17](#_Toc165384444)

[Exploratory Data Analysis 27](#_Toc165384445)

[Data Visualization 28](#_Toc165384446)

[Analysis Techniques 32](#_Toc165384447)

[CHAPTER 4: RESULTS 51](#_Toc165384448)

[CHAPTER 5: INTERPRETATIONS 54](#_Toc165384449)

[CHAPTER 6: DISCUSSIONS/LIMITATIONS/SUMMARY 55](#_Toc165384450)

[Discussions 55](#_Toc165384451)

[Summary 56](#_Toc165384452)

[Limitations 57](#_Toc165384453)

[CHAPTER 7: REFERENCE 58](#_Toc165384454)

# CHAPTER 1: INTRODUCTION

We are currently working on research that analyses the anxiety and depression that individuals experienced during the COVID-19 epidemic. Everyone had an exceedingly difficult time during this period, and several individuals experienced extreme depression or anxiety. Our goal is to examine the information we gathered throughout this time to learn more about the reasons behind people’s feelings.

 For our project, We are concentrating on two main concepts . First, we’ll take a detailed look at the data we collected on the feelings of individuals during the COVID-19 epidemic. Our goal is to identify any patterns or explanations for why a greater number of individuals may have experienced anxiety or depression during that period. Finding patterns and determining what was happening are the main goals of this section.

Our second research component involves developing and predicting people’s likelihood of experiencing anxiety or depression in the event of a potential COVID-19 pandemic. To produce these predictions, this tool—which we will refer to as the model—will draw on the knowledge gathered during the first phase of our study. Our goal is to keep the general mood up even in difficult times, and by identifying these tendencies, the model might help to take action to assist individuals before they become depressed.

In other words, our project’s goal is to identify the reasons behind people’s pandemic-related emotions and then use that information to help avoid experiencing those same emotions in the future, should a circumstance such to this one arise.

Beyond simply pandemics, we also believe that our effort may be helpful in other difficult circumstances. By knowing how to avoid depression or anxiety, we can support individuals in maintaining their mental toughness in the face of adversity. In summary, the goal of our initiative is to identify strategies for maintaining people’s mental health during difficult times by using the lessons learned during the epidemic.

**Mission Statement**

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.

## EDA Questions

1. In which age range are anxiety and depression disorders most common?
2. How does the pandemic’s pattern of anxiety and depression disorders evolve over time?
3. Do any states or areas have greater than average prevalence of these disorders?

## Predictive Questions

1. Predicting the strongest factors among age, disability, gender, and race for developing anxiety and depressive disorders during different stages of the COVID-19 pandemic?
2. Predicting the likelihood of anxiety and depressive disorder based on the factors such as season and region.
3. Predicting the future trends of anxiety and depression disorder among people who are over the age of 50 years in the United States using time-series forecasting model.

# CHAPTER 2: LITERATURE REVIEW

## Introduction

Pandemics affect billions of people in turn around the world in their daily lives, places of work, or communities, thus creating a really 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. The ensuing literature review looks into the association of pandemics with mental health by the guidelines that have been determined from historical contextual analysis, approaches of predictive models, regional inequalities, longitudinal studies, and pattern of healthcare consumption, and finally, the drawing of ethical implications. In order to justify this declaration, we will first take a look at the general historical repercussions for pandemics, with consequences such as the Spanish Flu and, for the recent SARS outbreak, showing dramatically higher rates than the averages of anxiety, sadness, and post-traumatic stress disorder (PTSD) within the affected communities (Banna et al., 2023).

The psychological effect of previous crises on mental health was, accordingly, described by Wiedermann et al. and is necessarily known that makes preventive measures towards alleviating long-standing psychological consequences caused by pandemics. Now, let us speak about such predictive modeling methods by which mental health outcomes during crises can be forecast. Predictive modeling is one promising approach to guide targeted interventions that can inform efforts to support the pandemic response by identifying environmental, social, and demographic factors that might be associated with the prevalence of mental health disorders. Furthermore, further research within our proposal aims to identify regional differences that exist, considering the impact of pandemics on mental health indicators in relation to socioeconomic positioning, accessibility to health care, and population density variables. Such marked differences in the experience of those outcomes between regions suggest a role for individualized therapies towards the local need in order to strive for mental health equity at the community level. Far more is being exposed by the wider application of longitudinal studies about mental health problematics and their development through pandemics. In this perspective, longitudinal studies provide an unambiguous record of mental health conditions data during, before, and after epidemics, which are important in producing approaches to maintain mental care. In addition to this, being informed on such patterns about the mental health consumption of health care is supreme to having well-laid plans and proper resource management in pandemics. Last but not least, telehealth and increased accessibility to community-based services allow many of the persons in need to be supported on time (Levin,2023).

Last but not least, the development and application of prediction models of mental health outcomes in pandemics present an array of ethical precincts applicable to the whole process. Sustained revelation of ethical norms and trust for predictive modeling projects needs data privacy protectors, informed consent givers, and assure algorithmic fairness. This review will give a comprehensive view of the mental implications by synthesizing insights into pandemics and strategies for redress. It, therefore, is to mean that the above stakeholders in mental health support give a good source to the mental health of people and the community amidst crises, fostering resilience and recovery from adversities (Martín, 2023).

## The Impact of Past Pandemics

Severe long-term mental health sequelae after pandemics in history have been observed. A case study from the Spanish Flu and SARS has revealed extremely high escalations in mental health cases within the affected populations. Anxiety, sadness, PTSD all shot up right after these crises, which had left a psychological scar on the people and never worn off. Other causes for mental health deficit through pandemics are; change in rituals, fear of illness or death, social isolation, and fluctuating status linked to economic conditions. Moreover, mental health problems have long-lasting effects that always last longer after the crisis has passed. Studies have indicated that psychological distress and reduced functioning can then persist for several months to even years after the pandemic is over. As the authors state, evidence from research points towards what they believe to be one strong message: “Consciousness of the long-term impact of pandemics on mental health suggests a need for early efforts in developing the needed support of the affected individuals and the mitigation of long-term psychosocial consequences.” Different patterns of approaches are needed for pandemic-related mental health effects. Effective approaches will need to address inequalities in socio-economic status, poor social support networks, and resilience built by the various available activities, including minimal access to mental health services. When this is focused on and mental health is equally targeted, such as physical health may be in instances of pandemics, nations can better shield their interests among respective fellows and also build levels of robustness going forward (Wunderlich-Barillas, 2023).

## Predictive Modeling for Mental Health Outcomes

According to White (2023), recent advances in predictive modeling hold promise for forecasting and addressing the various mental health implications of crises, including pandemics. The above statement implies that, through data analytical processes with the aid of machine learning algorithms, researchers can produce the material contributors to mental health crises—environment, other contributing demographics, and social attributions. Such predictive models aid authorities in the general pandemic response and support public health policymakers in more definite answers to be able to manage resources in relation to where most of the resources need to be put. The power of predictive modeling lies in hearing out who is at risk and tailoring interventions to that specific group. It entails detection of populations more prone to adversity from mental health from large data. This could be in the form of an enhanced prediction model, for probable use as an advanced warning system for pre-existing mental health cases, frontline healthcare workers, and social-economically weak sections who are most at risk to face pandemic experiences. On the other hand, the drawbacks, and ethical considerations in the event of predictive mental health outcomes completely fall under the knowledge flows. In this view, some of the critical aspects that need to be considered are the protection of an individual’s rights to data privacy, reduction in algorithmic bias, and creating more transparency in both the development and deployment of models. With respect to care given to mental health, some of the predictive models are to assist clinical judgment and qualitative impressions rather than to ever replace them. Integrating such predictive modeling into a comprehensive plan for pandemic response may help interested parties use data-driven insights to lessen the psychological impacts of crises and foster community resilience.

## Geographic Variations in Mental Health Impacts of Pandemics

The psychological impact of pandemics is not uniform and varied across regions; this variability could be related to a whole variety of characteristics, some of which were population density, access to health facilities, level of education, and SES position, among others. This research put forward, higher population densities would be expected to raise stress and anxiety due to greater viral exposure and greater difficulty with social distancing behaviors.

In the same vein, “mental health services might be less accessible in low access areas, so people in those kinds of communities would have both the most RI and the most RI inequality.”

On the highest rate of facts in such cases, the socioeconomic standing happens to be one of the highest determinants of crises with relation to mental health outcomes. Insecure housing, unreliable income, precarious employment, and lack of access to quality health care remain some of the common stressors in such instances. This goes on to be a clear demonstration of the needed importance that mental health interventions should be given, and further goes ahead to have such interventions be tailored to the needs encouraged by different geographical and demographic categorizations. The large gaps in mental health imply a response that can be targeted to individuals’ needs—both wide and deep—along with the systemic determinants of mental well-being. The strategies toward narrowing these gaps may involve several things: addressing the disparities within the socio-economic background, community-based intervening approaches that need to be taken in developing social support networks, and the expansion of mental health care in underserved areas. By applying a geographically informed approach to mental health intervention, policymakers and healthcare providers can better meet the varied needs of individuals affected by pandemics(Dewan,2022).

## Longitudinal Studies of Mental Health During Pandemics

Long-term studies can highlight patterns in the resistance, vulnerability, and recovery of affected communities as well as how mental health issues alter throughout pandemics. Researchers can uncover risk factors, psychological adjustment trajectories in response to crisis events, and protective variables by monitoring people’s psychological well-being before, during, and after an outbreak.

According to Jones,(2022) A systematic review and meta-analysis of longitudinal cohort studies comparing mental health before versus during the COVID-19 pandemic in 2020, while many people exhibit resilience and employ flexible coping mechanisms when faced with hardship, some may have prolonged psychological pain or experience a worsening of pre-existing mental health conditions. Identifying the elements of resilience can aid in the development of targeted interventions intended to fortify mental health in the event of a pandemic. In addition, longitudinal studies aid in the identification of long-term trends and needs in mental health, which is beneficial for allocating resources and organizing mental health care following the original crisis. These studies chronicle the changing nature of mental health consequences across time, contributing to a thorough understanding of the intricate interplay between individual, societal, and environmental elements during pandemics(Jones,2022).

**Healthcare Utilization for Mental Health Services in Pandemics**

Planning and resource allocation during pandemics require an understanding of patterns of healthcare usage for mental health services. Studies indicate that a number of issues, such as overburdened healthcare systems, the stigma associated with mental illness, and practical difficulties including limited transportation options and social distancing protocols, may make it difficult for people to receive mental health care during emergencies.   
Furthermore, unequal access to mental health services has the potential to exacerbate already-existing gaps, particularly for underprivileged or marginalized groups. Healthcare also denies other people to seek help, such as those with money problems, housing instability, or fighting with language, who show symptoms.

Such innovations in delivering mental health treatment may be either by means of community-based therapies, mobile outreach mechanisms, and telehealth platforms. Beyond that, the development of alternative care modalities, remote places or places poorly served, alternative care modalities, can significantly go a long way in providing solutions to innumerable hurdles in traditional “office-based” mental health service for reaching out to consumers in need.

In addition, the generalized health workforce is expanded to provide health care services, while primary care is integrated with mental health care services. The layered approach, that from one hand, it includes some of the more classic strategies and from the other innovatory approaches, is considered one of the many possible multi-pronged strategies in serving the bipolar or diverging needs from mental health clientele in a pandemic (Albarqouni, 2021).

**Ethical Considerations in Predictive Modeling for Mental Health during Pandemics**

Ethical considerations It goes directly to heart in is central to the development and application of prediction models for mental health outcomes in pandemics. The predictive model, although leading plausible possibilities in the way it provides a platform for improving resource allocation and initiating focused medical care, nevertheless raises concerns over privacy, informed consent, and the potential for algorithmic bias (Sodamade, 2023).

The entire operation of collecting and processing data to arrive at predictive modeling is sensitive and needs protection of the people’s privacy and confidentiality. Stipulations based on data security must be adhered to seriously so that data is not accessed or used by unauthorized persons hence re-identification. Informed consent remains another essential ethic, especially in the case of sensitive health data in predictive modeling. People engaged in it have to be appraised of all the objectives, risks, and potential benefits to be derived from their taking part in the collection of data activities. Additionally, their assent must be freely provided and not coerced(Dewan, 2024).

The findings of the predictive models need to be given a guarantee for fairness and to avert harm due to algorithmic bias. Utmost care should be taken in the assessment and correction of biased datasets, algorithms, and model outputs that are possibly leading to unintended consequences that can be harmful or scale up pre-existing inequalities, or even fuel prejudice. Moral predictive techniques must be accountable and transparent in modeling. It rests on the scientist to be honest and disclose such limitations and the related uncertainty that his or her model might bear, in order for bases to be set for transparent communication and collaboration with relevant parties, accountability and confidence-building in regard to predictive modeling initiatives. When the researchers were developing and testing predictive models of mental health during pandemics, related ethics always remained amiss but were especially important. In the long run, this will benefit the communities and people affected by crises (Zheng et.al,2023).

## Conclusion

This is drawn as the inference from the literature analysis, showing a great mastery of the effects pandemics have on mental health and complex solutions to these problems. As such, each of the components has made invaluable contributions to the knowledge base on mental health in crises, be it the historical precedents of pandemics, new models to predict them, geographical variations, longitudinal studies, or health services used in crises and the pattern of that usage with related ethical implications. Pandemics have grossly and over the long run negatively affected mental health, bringing unfavorable levels of anxiety, depression, and even PTSD in most hit populations. However, initiative-taking measures must be factored into mental health planning during pandemics that can lower the long-term psychological aftermath and boost community resilience. Once the threshold ethical issues—data privacy, informed consent, algorithmic bias—are properly met, predictive modeling may well offer these innovators and others a new toolkit with which to forecast mental health outcomes and guide targeted therapies. This therefore only goes to recognize that regional disparities, in every effect, affect mental health, by doing so, it reinforces that the issue of specialized treatment approaches directed on disparities, in the region, can meet specific needs of some communities. Longitudinal studies give particularly good understanding of the way that mental health problems develop over time, which can feed construction to a long-term development in mental health care programs. In such times of pandemics, such understanding of the pattern of health care utilization for mental health services is required. This is for efficient planning and subsequently resource allocation. However, this very fact issues a warning that in some way, some form of creativity has been taking place already, like in the form of telehealth or community-based interventions to take down the barriers of access.

Lastly, for the case of ethical considerations, they will cut across the entire process of developing and deploying the prediction models of mental health outcomes for openness, accountability, and equity. “Responsible use of data can more properly serve the mental health needs of the people affected by earthquakes and the resilience and recovery of the community from adversities with such insights as integrated into the broader responses to disaster.

# CHAPTER 3: METHODOLOGY

**Software**

Data preparation such as data cleaning, data transformation, data preprocessing was performed using Python from Anaconda environment and Microsoft Excel. Data Analysis and Data Modeling were performed using R Studio and Python from Anaconda.

## Data Collection

The data that was collected for our research contains information related to anxiety and depression disorders. We have retrieved this data from HealthData.gov. Below is the website link through which you can find the data.

<https://healthdata.gov/dataset/Indicators-of-Anxiety-or-Depression-Based-on-Repor/xpsn-dxxd/about_data>

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.

Households were asked to participate by email and SMS in an online questionnaire format, which helped the survey accomplish its goal of delivering reliable and timely weekly data. Based on the Census Bureau's Master Address File Data, participants were chosen via a random method from dwelling units connected to at least one email address or cellphone number. One person was picked from each selected home to provide personal replies. Adjustments were performed for non-responses and to conform to the Census Bureau's demographic estimates regarding age, sex, race and ethnicity, and educational attainment in order to guarantee the representativeness of the findings. The information provided complies with the proportionate estimate presentation guidelines established by the National Centre for Health Statistics (NCHS).

The data consists of 15157 rows and 14 columns, and the data is collected in between 04/23/2020 and 03/04/2024.

A screenshot of a computer

Description automatically generated

Figure 1 Dataset Example

The dataset gives entire information about the symptoms of depressive and anxiety disorders by each group of each state with the value, low and high confidence interval, confidence intervals range and quartile range.

## 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.

|  |  |  |
| --- | --- | --- |
| **Name** | **Data Type** | **Description** |
| Indicator | text | The specific measure being observed. |
| Group | text | The category or classification the data belongs to. |
| State | text | The geographical location in USA. |
| Subgroup | text | A further division or classification within a group. |
| Phase | Number (Float) | The stage or period of the data collection or event. |
| Time Period | text | The duration over which the data was collected or applies. |
| Time Period Label | text | A descriptive name or label for the time period. |
| Time Period Start Date | calendar date | The starting date of the time period. |
| Time Period End Date | calendar date | The ending date of the time period. |
| Value | Number (Float) | The numerical measurement or value associated with the indicator. |
| Low CI | Number (Float) | The lower bound of the confidence interval for the value. |
| High CI | Number (Float) | The upper bound of the confidence interval for the value. |
| Confidence Interval | Number | The range within which the true value is expected to lie, with a given level of confidence. |
| Quartile Range | Number | A statistical range dividing the dataset into four equal parts, indicating variability or distribution. |

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 based on or grouped on state.

A screen shot of a computer

Description automatically generated

Figure 2 Total observations in each group

It is very crucial to know whether we have the target variables and predictor variables present in the dataset for our research. As a matter of fact, without proper variables our analysis and interpretation won’t be precise.

So, for all of the research questions that we are predicting or analyzing is performed on these variables, so the target variables and predictor variables need to be from these 14 variables or any other variables that we create or extract using these 14 variables as a part of feature engineering in the data preparation step. The predictor variables and target variable for the 3 predictive questions that are researched are as follows.

* For predicting the strongest factors among age, disability, gender, and race for developing anxiety and depressive disorders during different stages of the COVID-19 pandemic?

Predictor variables are Age, Disability, Gender, and Race.

Target Variables are depressive disorder and anxiety disorder

* For predicting the likelihood of anxiety and depressive disorder based on the factors such as season and region.

Predictor variables are Season, and Region.

Target Variables are depressive disorder and anxiety disorder

* For predicting the future trends of anxiety and depression disorder among people who are over the age of 50 years in the United States using time-series forecasting model.

For this time-series forecasting model, ‘Value’ and ‘Time Period’ are the variables.

As per the research and analysis that are to be performed, there are many variables that need to be extracted from the existing variables which will be evaluated or explained in the data preparation steps.

## Data Preparation

As every data that we get from the source won’t be refined and cleaned, preparing, and refining our raw or original dataset is especially important procedure and part of the entire life cycle. The data may have unnecessary information, variables with inappropriate format, observations with missing values. To improve the quality of the data and to structure the data in an appropriate format for the analysis preparing our data accordingly is recommendable. The more the data is prepared and refined, the more effective the statistical analysis will be. In addition to that, even the predictions, results, interpretations, and insights will be true, accurate and unbiased.

As a part of data preparation, there are many steps that can be or need to be performed such as:

* Removing unnecessary columns / Data Reduction
* Data Cleaning
* Feature Engineering/ Data Transformation.
* Outlier Detection

**Removing unnecessary columns / Data Reduction**

By performing this step, 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.

A screenshot of a computer

Description automatically generated

Figure 3 Example dataset after data reduction

**Data Cleaning**

By performing this step, we can get rid of observation 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*

A screenshot of a computer

Description automatically generated

Figure 4 Example dataset with inconsistent values in Phase variable

From the above Figure, it indicates that the Phase variable has 2101 observations with some inconsistent values of date making the numerical variable a string and difficult for the other manipulation operations. So, using Microsoft Excel replace option, we have replaced the date value with empty value.

From the below figure, we can see that inconsistency error in the Phase variable has been handled perfectly without removing or deleting any observations.

A screenshot of a computer

Description automatically generated

Figure 5 Example dataset after removing inconsistencies

*Handling Missing Values*

Among all these steps in data cleaning, identifying missing values and removing those missing values is an crucial step and we can handle these missing values in many ways, either by removing the total observation that consists a single missing value or filling the missing values with a default values or filling the missing value with average value of that columns if it’s an integer variables or filling the missing value with most repeated or above or below value if it’s a categorical variable depending on the requirement.

As per our research, 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.

A screenshot of a computer program

Description automatically generated

Figure 6 Variables with missing values

The above figure shows that there are 703 missing values in each variable of Value, Low CI, and High CI. We want to handle these 703 missing value observations by removing these 703 observations from our total observations. Below figure is the dataset after removing missing values. A screenshot of a computer

Description automatically generated

Figure 7Example dataset after removing missing values

Below figure indicates that 703 observations from 15156 have been removed from the dataset as part of handling missing values and there are only 14453 observations present in the dataset which is a good number of observations for the analysis and modeling.



Figure 8 Shape of the dataset before and after handling missing values

*Data Transformation*

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.

Feature Engineering

In this step, we will create new features, columns or variables based on existing variables. As per our requirements and research questions, firstly, we want to create 2 new features from the Indicator column namely Depressive disorder and Anxiety disorder that has binary variables such as 0 and 1 as the values indicating whether they have depressive disorder or anxiety disorder.

Below figure is the example of the dataset after creating the two new features Depressive disorder and Anxiety Disorder.

A screenshot of a computer

Description automatically generated

Figure 9 Example dataset after creating depressive and anxiety disorder variables

Also, looking into the dataset, we do not have separate variables or columns for demographic factors such as Age, Gender, Education, Disability, State, and Race. We have created all these features into new columns/variables from the Group Variable as shown below.

A screenshot of a computer

Description automatically generated

Figure 10 Example Dataset after creating demographic variables

To research on the time series analysis model, we need two columns namely Year and Month. Using the existing variable ‘Time Period Start Date,’ we have created these two columns.

A screenshot of a computer

Description automatically generated

Figure 11 Example Dataset after creating Year and Month variables

For the research question, predicting the likelihood of anxiety and depressive disorder based on the factors such as season and region we need the season and region columns. So, we have created these two columns from the month column. If the month value is 3,4,5 we have assigned Spring, if the month value is 6,7,8 we have assigned Summer, if the month value is 9,10,11 we have assigned Autumn and the rest all to winter.

For the creation of region column, we have used State column, divided the states into the following regions: Northeast, Midwest, South, West and Other. Below figure is the image with the new feature variables such as Year, Month, Season, and Region.

A screenshot of a computer

Description automatically generated

Figure 12 Example Dataset after creating feature variables such as Year, Month, Season, and Region.

Overall, we have created the new variables or features that we want for our analysis on our research questions in this feature engineering process.

*Outlier Detection*

Detecting outlier is also a crucial part of the data analysis project. Outliers indicate data points that are present extremely 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. It completely depends on the requirements of the project that we are researching on whether to remove outliers or not if detected. Outliers can be detected through statistical methods or through box plots and scatter plots. Outliers are usually present or seen in numerical variables, date variables rather than categorical variables. According to the dataset, Phase, Time Period, Time Period Start Date, Time Period End Date, Value, Low CI, High CI, Depression disorder, Anxiety Disorder, and Month are the numerical and date variables.

The below 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.

A graph with different colored squares and lines

Description automatically generated with medium confidence

Figure 13 Box plot to check outliers for numerical variables in the dataset

As discussed earlier, every outlier that is detected need not to be bad and sometimes very important data point can also be detected as an outlier. So, it is especially important to be careful while dealing with outliers.

## Exploratory Data Analysis

Exploratory data analysis (EDA) is used to understand and examine the dataset characteristics. After examining the dataset characteristics, with the general understanding of the dataset we can start performing data analysis or data modeling. Statistical analysis, correlation analysis, visualizing the data using charts, and many more comes under Exploratory Data Analysis.

*Correlation Analysis*

Correlation Analysis is especially important to know the correlation or linear relationship between the variables. As we have 24 variables after performing data transformation, it is particularly 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 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.

A screenshot of a computer

Description automatically generated

Figure 14 Correlation plot between numerical variables

Above image indicates the correlation between the numerical variables. Th 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.

It’s better to perform correlation between the categorical variables, as we are using these variables in our analysis so that we can understand the relationship between those variables too. Below is the image that we have got after performing Chi-Square test between the variables ‘By Age’ , ‘By Disability Status,’ ‘By Race/Hispanic ethnicity,’ ‘By Sex,’ ‘By State’ and ‘Region.’ 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.

A screenshot of a computer

Description automatically generated

Figure 15 Correlation between categorical variables using Chi-Square test

## Data Visualization

With the help of data visualization, we can visualize the whole data in the form of charts or graphs such as histogram, box plot, bar plot, scatter plot, etc. Through these charts, we can discover many patterns and gain enough information about the relationship between the data and even distribution of the data. Below are the research questions that I want to gain insights about the data using data visualization.

*Questions*

1. In which age range are anxiety and depression disorders most common?

With the help of data visualization, we want to analyze or visualize and discover the pattern on which age range, the anxiety disorders and depression disorders are most common in the United States.

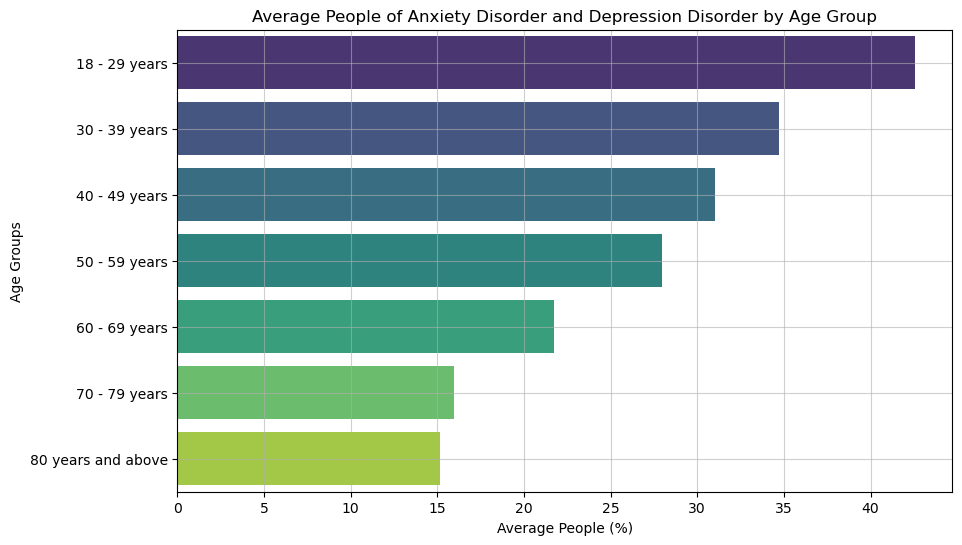


Figure 16 Bar chart to visualize the most common age ranges that have anxiety and depressive disorders

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. From this visualization, we can also understand or discover that most of the young people are more prone to have anxiety disorders and depressive disorders as an assumption of them being more stressed because of studies and career pressure.

1. How does the pandemic’s pattern of anxiety and depression disorders evolve over time?

With the help of data visualization, we want to visualize and discover the pandemic’s pattern of the anxiety disorders and depression disorders over the period of time from 2020 to 2024.

A graph of different colored lines

Description automatically generated

Figure 17 Line chart to visualize the patterns of anxiety and depression orders over the time

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. The chart also indicates that there are more people who have both the symptoms of anxiety and depression disorder combined with a substantial dip in 2022 and a higher rise in 2022-07. From this visualization, we can also understand or discover that most of the people are more prone to have both anxiety disorders and depressive disorders combined.

1. Do any states or areas have greater than average prevalence of these disorders?

With the help of data visualization, we want to visualize and discover the number of states in the United States that have greater than or equal to average prevalence of the anxiety and depression disorders.

A graph of a graph with different colors

Description automatically generated with medium confidence

Figure 18 Bar chart to visualize the states that have equal or greater than the average anxiety and depression disorders.

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%. This chart also indicates that there almost half of the states in the US that have greater than or equal to national average which suggests that the people living in those states need more help from the peers in the form of therapy and from the government in terms of awareness campaigns.

## Analysis Techniques

As the data preparation is perfectly done as per the requirements of our 3 research statements, understanding the characteristics of the dataset through exploratory data analysis and from the patterns that we had uncovered through the visualization, we are able to completely understand about the nature of the data we are using.

With this information, we can do the other important step modeling through which we can make the predictions from the results that we get by using Machine Learning models such as Linear regression, Logistic regression, Random Forest, Decision tree, Gradient Boosting Model, LSTM neural networks, ANN and many more. It is very crucial and important to choose an ML model for our modeling based on the type of research we are performing as each model works completely differently as they have different approach and different pros and cons. So, according to the research questions, these are the models that are being used as part of modeling.

1. Predicting the strongest factors among age, disability, gender, and race for developing anxiety and depressive disorders during different stages of the COVID-19 pandemic?

Analysis Techniques: Random Forest, Decision Tree.

1. Predicting the likelihood of anxiety and depressive disorder based on the factors such as season and region.

Analysis Techniques: Random Forest, Decision Tree

1. Predicting the future trends of anxiety and depression disorder among people who are over the age of 50 years in the United States using time-series forecasting model.

Analysis Techniques: LSTM technique.

**Predicting the strongest factors among age, disability, gender, and race for developing anxiety and depressive disorders during different stages of the COVID-19 pandemic?**

By performing modeling, we had predicted the most important or strongest factors among age, disability, gender, and race that have an impact of developing the symptoms such as anxiety disorder and depressive disorder during the pandemic.

For this research question, ML models such as logistic regression, random forest, and decision tree can be used. We have performed Random Forest and Decision tree model among all the models.

*Random Forest*

Random Forest Model is an analysis technique that is used to perform both classification and regression tasks. We have chosen Random Forest as our first model as we are performing classification task, as the target variables that we are having for this research questions are Anxiety Disorder and Depressive Disorder that consists of Boolean values such as 0 and 1. We split the dataset with 30% testing data and 70% training data.

**A screenshot of a computer

Description automatically generated**

Figure 19 Evaluation Metrics of Random Forest Model

The Below image indicates that 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 extremely low, and it indicates that the complexity of the model in predicting both the target variables as each label is low. As we are performing multi-classification using random forest model, the accuracy can be resulted low compared to the individual accuracies for each variable and it is not a bad indicator.

Other evaluation metrics such as Precision, Recall and F1 score have also been calculated for each class. For the depressive order class, we got a precision value of 0.68 which suggests that the model is correctly predicting the positive class. The recall value of 1.00 suggests that the recall value is exceedingly high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.81.

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 exceedingly high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.80.

AUC score and ROC curve has also been plotted to understand and evaluate the random forest model performance. ROC curve is used to plot the graph between the true positive rates and false positive rates. If the ROC curve passes through 0 and 1, we can identify that the random forest model is perfect. In addition to ROC curve, if the AUC score is closer to 1, then the random forest model is performing perfectly and if it is closer to 0, them the random forest model is performing poorly.

Below ROC curve and AUC score of 0.50 is for depressive disorder, it identifies that the model is performing moderate in predicting depressive order and can perform more well.

A graph of a person's body

Description automatically generated

Figure 20 ROC curve of Depressive Disorder

Below ROC curve and AUC score of 0.50 is for anxiety disorder, it identifies that the model is performing moderate in predicting anxiety order and can perform more well.

A graph of a positive and negative rate

Description automatically generated with medium confidence

Figure 21 ROC curve of Anxiety Disorder

The assumption to be having moderate AUC score for both the variables may be because of the dataset having the same distributions for both variables. Below feature importances bar plot indicates that the ‘By Sex’ / Gender variable is main influence factor influencing the Depressive and anxiety disorder over the time in the United States with an importance score of 0.35. From this, we can assume that the gender plays a crucial role in getting affected with these disorders during pandemic and now.

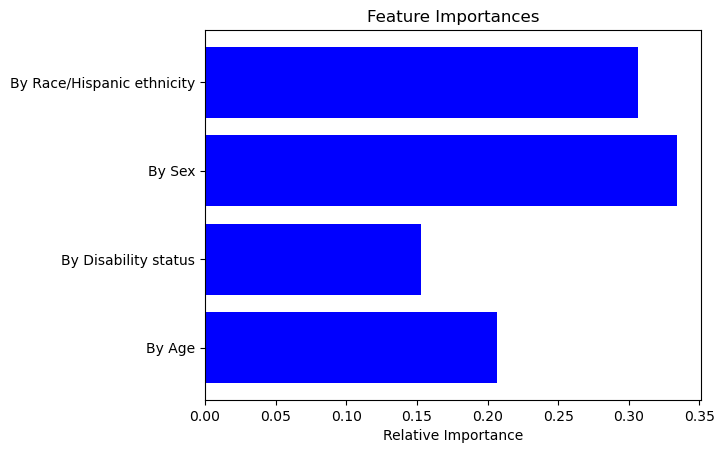


Figure 22 Importance Score of variables

*Decision Tree*

Decision Tree is an analysis technique that is used to perform both classification and regression tasks. As decision tree provides a tree structure plot with some nodes and branches, it is amazingly easy to understand the insights and patterns that are resulted. We have chosen decision tree as our second model as we are performing classification task, as the target variables that we are having for this research questions are Anxiety Disorder and Depressive Disorder that consists of Boolean values such as 0 and 1. We split the dataset with 30% testing data and 70% training data as well.

The Below image indicates that the accuracy of the decision tree for depressive disorder is 68% which is a good indicator that our model is performing well. Other evaluation metrics such as Precision, Recall and F1 score have also been calculated for each disorder. For the depressive order class, we got a precision value of 0.68 which suggests that the model is correctly predicting the positive class. The recall value of 1.00 suggests that the recall value is remarkably high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.81. We have also calculated AUC score and ROC curve, for depressive disorder, AUC score of 0.50 identifies that the model is performing moderate in predicting depressive order and can perform more well.

A graph of a positive and negative disorder

Description automatically generated with medium confidence

Figure 23 ROC curve for depressive disorder based on Decision Tree

The Below image indicates that 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 extremely high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.80. We have also calculated AUC score and ROC curve, for anxiety disorder, AUC score of 0.50 for anxiety disorder identifies that the model is performing moderately in predicting anxiety order and can perform more well.

A graph of a positive and negative disorder

Description automatically generated with medium confidence

Figure 24 ROC curve for Anxiety disorder based on Decision Tree

The assumption to be having moderate AUC score for both the variables may be because of the dataset having the same distributions for both variables. Below feature importances bar plot also indicates that the ‘By Sex’ / Gender variable is main influence factor influencing the Depressive and anxiety disorder over the time in the United States with an importance score of 0.37. From this, we can assume that the gender plays a crucial role in getting affected with these disorders during pandemic and now.

A blue and white bar graph

Description automatically generated

Figure 25 Importance score of variables

We had also plotted decision tree to visualize the tree model as below. It indicates that the first split is made on ‘By Sex’ as it has a greater number of samples on the branch and later the splits are made on ‘By Race/Hispanic Ethnicity,’ ‘By Age,’ ‘By Disability Status’

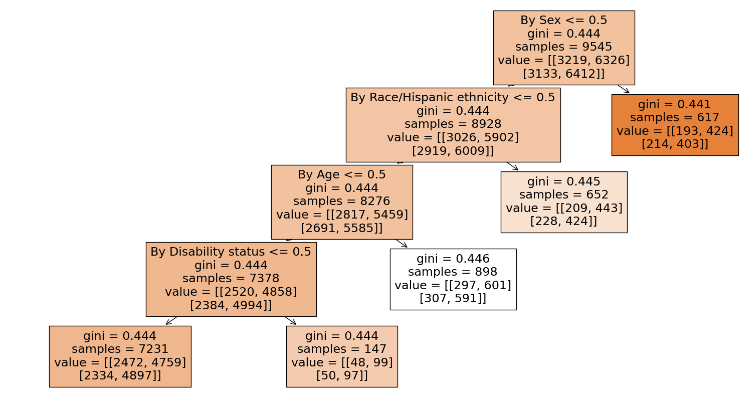


Figure 26 Decision Tree

So, overall, we can conclude from both the models that from all of the demographic factors, ‘By Sex’ is the strongest factor and predictor in developing anxiety and depressing disorders. This is maybe because of the difference in the immunity of the both the genders in majority cases.

**Predicting the likelihood of anxiety and depressive disorder based on the factors such as season and region.**

By performing modeling, we had predicted the likelihood of anxiety disorder and depressive disorder based on the factors season and region to evaluate whether is there any relationship between the weather and location for a person to be having disorder symptoms.

For this research question, we have performed Random Forest and Decision tree model among all the models.

*Random Forest*

Random Forest Model is used to perform both classification and regression tasks, we have chosen Random Forest as our first model, as the target variables that we are having for this research questions are Anxiety Disorder and Depressive Disorder that consists of Boolean values such as 0 and 1. We split the dataset with 30% testing data and 70% training data.

**A screenshot of a computer

Description automatically generated**

Figure 27 Evaluation Metrics of Random Forest Model

The Below image indicates that 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 exceptionally low, and it indicates that the complexity of the model in predicting both the target variables as each label is low. As we are performing multi-classification using random forest model, the accuracy can be resulted low compared to the individual accuracies for each variable and it is not a bad indicator.

Other evaluation metrics such as Precision, Recall and F1 score have also been calculated for each class. For the depressive order class, we got a precision value of 0.68 which suggests that the model is correctly predicting the positive class. The recall value of 1.00 suggests that the recall value is exceedingly high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.81.

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 extremely high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.80.

AUC score and ROC curve has also been plotted to understand and evaluate the random forest model performance. ROC curve is used to plot the graph between the true positive rates and false positive rates. If the ROC curve passes through 0 and 1, we can identify that the random forest model is perfect. In addition to ROC curve, if the AUC score is closer to 1, then the random forest model is performing perfectly and if it is closer to 0, them the random forest model is performing poorly.

Below ROC curve and AUC score of 0.50 is for depressive disorder, it identifies that the model is performing moderate in predicting depressive order and can perform more well.

A graph of a person's body

Description automatically generated

Figure 28 ROC curve of Depressive Disorder

Below ROC curve and AUC score of 0.50 is for anxiety disorder, it identifies that the model is performing moderate in predicting anxiety order and can perform more well.

A graph of a positive and negative rate

Description automatically generated with medium confidence

Figure 29 ROC curve of Anxiety Disorder

The assumption to be having moderate AUC score for both the variables may be because of the dataset having the same distributions for both variables. Below feature importances bar plot indicates that the ‘Region’ variable is main influence factor influencing the Depressive and anxiety disorder over the time in the United States with an importance score of 0.5. From this, we can assume that the region plays a crucial role in predicting the likelihood of anxiety and depression disorders among the people.

A blue bar graph with white text

Description automatically generated

Figure 30 Importance Score of variables

*Decision Tree*

As decision Tree is an analysis technique that is used to perform both classification and regression tasks providing a tree like structure plot with some nodes and branches, it is extremely easy to understand the insights and patterns that are resulted. We have chosen decision tree as our second model. We split the dataset with 30% testing data and 70% training data as well.

The Below image indicates that the accuracy of the decision tree for depressive disorder is 68% which is a good indicator that our model is performing well. Other evaluation metrics such as Precision, Recall and F1 score have also been calculated for each disorder. For the depressive order class, we got a precision value of 0.68 which suggests that the model is correctly predicting the positive class. The recall value of 1.00 suggests that the recall value is extremely high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.81. We have also calculated AUC score and ROC curve, for depressive disorder, AUC score of 0.50 identifies that the model is performing moderate in predicting depressive order and can perform more well.

A graph of a positive and negative disorder

Description automatically generated with medium confidence

Figure 31 ROC curve for depressive disorder based on Decision Tree

The Below image indicates that 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 exceedingly high and indicates that the random forest is identifying all the actual instances of the depressive order successfully. As the precision and recall value is high, the F1- score has also resulted in a high value if 0.80. We have also calculated AUC score and ROC curve, for anxiety disorder, AUC score of 0.50 for anxiety disorder identifies that the model is performing moderately in predicting anxiety order and can perform more well.

A graph of a positive and negative disorder

Description automatically generated with medium confidence

Figure 32 ROC curve for Anxiety disorder based on Decision Tree

The assumption to be having moderate AUC score for both the variables may be because of the dataset having the same distributions for both variables.

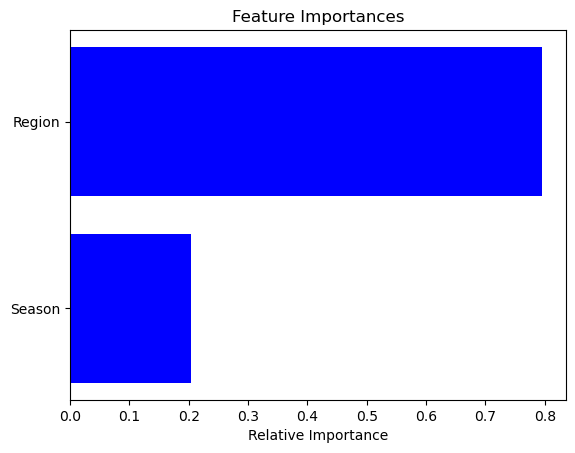


Figure 33 Importance score of variables

Above feature importances bar plot also indicates that the Region variable is main influence factor influencing the Depressive and anxiety disorder over the time in the United States with an importance score of 0.8. From this, we can assume that the Region plays a crucial role in getting affected with these disorders during pandemic and now.

We had also plotted decision tree to visualize the tree model as below. It indicates that the first split is made on season <= 2.5, as it has a greater number of samples on the branch and later the splits are made on ‘By Race/Hispanic Ethnicity,’ ‘By Age,’ ‘By Disability Status’

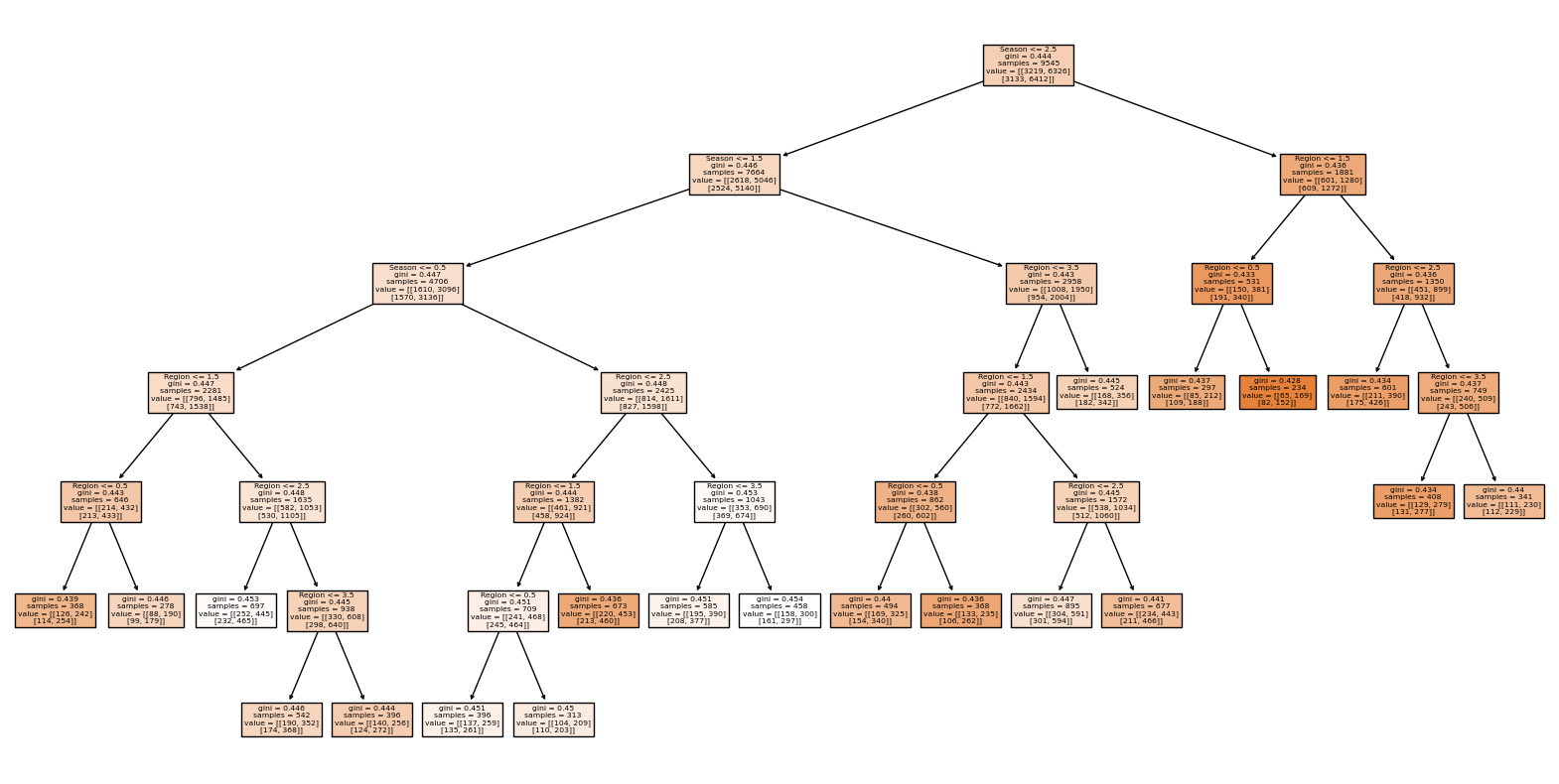


Figure 34 Decision Tree

So, overall, we can conclude from both the models that from Region and Season, likelihood of anxiety and depressive disorders is based on Region rather than the season. This can also be because of the difference in eating habits, lifestyle, and weather conditions of the regions. So, changes in habits, weather, and lifestyle can also cause us anxiety and depression disorders.

**Predicting the future trends of anxiety and depression disorder among people who are over the age of 50 years in the United States using time-series forecasting model.**

By applying LSTM neural network model, we had predicted the future trends of anxiety and depressive disorder among people who are over the age of 50. For this research question, we have firstly filtered our dataset by ‘Subgroup’ containing 50-59 years, 60-69 years, 70-79 years, and 80 years and above as seen in the below figure.

A screenshot of a computer

Description automatically generated

Figure 35 Example dataset after filtering Subgroup to age >= 50

Later, the variable ‘Value’ has been normalized using MinMaxScaler, the input sequence has been set to 12 which indicates that it takes 12 data points to predict the next data point, number of features to 1. Before building the LSTM model, the data was split into 30% testing data and 70% validation data. The LSTM model has been built using 2 LSTM layers with 100 and 50 neurons and 1 dense layer.

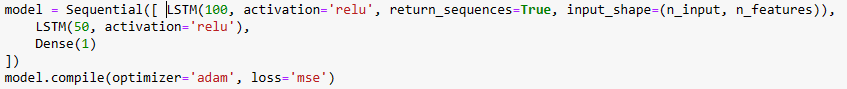


Figure 36 Python Code to build LSTM model.

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 similar to 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.

A graph showing a graph of blue and orange lines

Description automatically generated

Figure 37 Forecast vs Actuals Values Comparison

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. 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. As the dataset does not have these 15 days actual values, it is not fair to let people know whether the LSTM model had future 15 days actual values correct.

A graph with red lines

Description automatically generated

Figure 27 Value for Future 15 days value of the indicators

Overall, we can say that the LSTM model was successfully built and forecasting the future patterns of anxiety and depression disorders for the people who are greater than or equal to 50 years.

# 

# CHAPTER 4: RESULTS

The informational insights with Exploratory Data Analysis are quite easy to understand with the visuals which are appealing to everybody and the information from each of the visualizations using line chart, and bar chart. Line charts are of great use especially when we want trends, seasonality’s or sudden spikes or patterns in the data which are great usage factors examples include: Stocks, Weather, Comparison. There are specific use cases which need respective visualization to depict the information exactly and accurately. Using the appropriate visualization can reveal much more statistics which will lead to actionable insights.

Bar charts are used specifically for comparison or for comparing any data among different groups or similar groups, also it indicates the exact data points from the given data which results in excellent representation. Bar chart is widely used as it is quite simple and doesn’t need any complex variables for implementation.

Based on the first EDA question we can easily depict that the age groups of 18-29 are more prone to the anxiety and depression disorders then the age groups of 30-39 followed by 40-49. Which is indicating that it is showing a linear growth through age groups and the younger the age group the higher are the chances for anxiety and depression disorders. The usage of bar chart is significantly the best representation for this question which has different age groups to be visualized based on the same category which is about disorders.

Based on the second EDA question, finding the patterns and trends for the disorders through the years ranging from 2020-2024, Which is clearly stating that covid has definitely made an impact in the disorders and the specific symptoms for depressive disorders are spiking in 2022 but lately the trend is stagnating which indicates there are less and less percentage of symptoms affecting the people of US, But when anxiety and depressive disorders are combined it is resulting in higher chances of adverse effects summing up to higher spike in 2022. So, the trends are aligning to a point where anxiety is the threatening factor which is leading all these other factors based on the trends.

Based on the third EDA question, Disorder comparison over the states which involves the usage of bar chart, which is very insightful as discussed earlier and here the visualization is using heat and cool which is showing the highest prevalence with cool (blue) color and lowest prevalence with heat (red) color. The visualization indicates that Louisiana has the highest disorder prevalence with Mississippi, Nevada, Oklahoma, and West Virginia. The lowest disorder prevalence states include Colorado, Ohio, Utah, Indiana, and Alaska.

Based on the first Predictive Question, Using the Random Forest and Decision Trees has showed us that the accuracy for both the models is resulting around 68% suggesting that the various factors like age, gender, disability, and race plays a crucial role for the disorders and it can also suggest that these may arise through hereditary as well.

Based on the second Predictive Question, Since the data with which we are analyzing is mostly revolving around distribution of categorical and numerical variables considering Random Forest and Decision Trees is a wise approach as these can be used for both regression and classification and the accuracy for this question is also summing up to 68% with random forest performing better than decision tree where decision tree is indicating more patterns with insights and F-1 scores better than the Random Forest.

Based on the third Predictive Question, The prediction is about future trends in the disorders who are of age group above 50, For the future trends prediction we are going to use the sequential data to find the future trends. So for this Recurrent Neural Networks are a best use case for the sequential data which can reveal more insights with the activation functions and the usage of LSTM(Long Short Term Memory) which consists of gates through which the feedback from the arised training data is given so that it can store and process it accordingly which can mitigate future discrepancies with similar issue, The trend patterns are identified accurately with the help of RNN with LSTM and likewise, The usage of the specific visualization and algorithms helps in revealing better actionable insights which can work well with the real world scenarios.

# 

# CHAPTER 5: 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 high in the young age groups of 18 – 29 which suggests 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 also suggest that the states which have huge party cultures and lavish lifestyle are 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.

This indicates major analysis on young adults and stress is also an added factor for the young adults with lot of things running on their mind as they are living independently the options are unlimited to choose. The all results are suggesting that prevalence of the disorders is being affected by the demographics, locality, already having a disorder and all these factors are exactly correlating with the given information which has revealed best informational insights, and the accuracies are at the best for the presented data.

# CHAPTER 6: DISCUSSIONS/LIMITATIONS/SUMMARY

## Discussions

All the factors with results and interpretations are summing up to the one and only concern of mental health impacts with anxiety which should be diagnosed as early as possible to have the right medication and be controlled at the right stages which will not lead to other dis orders and most of the anxiety is seeming to be around young adults of 18-29 age group suggests that there needs to be proper awareness on mental health as well as medical awareness on health check-up every year which can be of great use for everyone and act on it faster.

The data is also suggesting that there are multiple factors which might be an Add-ons which can contribute to the anxiety disorder which includes demographics like race, gender, region, state, age. These are the factors which might influence one’s conditioning through which there might also arise economic conditions which suggests that there are higher chances the disorders can be influenced through these factors but proper choices relating to health and lifestyle can be a drastic impact which might put us in better standards where the disorders might not affect us.

The stress factors impacting the youth is also a peak concern which can lead to mental health problems leading to disorders, Usually these stress factors arise at the middle age or about to start middle age with family around 35-49, But with current technological advances and also the substance usages with stress might lead to the prevalence of these disorders and these stress can cause due to academics, social life, work life, economic conditions and we don’t know what conditions are affecting one’s life. So, we have to make sure that planning, scheduling, and sticking to a routine helps everybody in check. Discipline and Motivation from that age helps individuals take better actions on mental health and cope with the situations better ultimately controlling the situations better than earlier.

## 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 ridiculously 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.

The locations and demographical impacts are to be taken into consideration and necessary steps have to be taken at the right time which can have profound impact on the information which is revolving around these factors, There are hereditary patterns, much more aspects to other factors which might be relating to these disorders and the location preferences might be the fast paced living and also the work-life balance which might lay a foundation for these. These factors can be retrospective and controlled to an extent**,** but it all sums up to the individuals and their behaviors on managing these factors can pave their direction. Similarly, the other factors include the trends in the years, As in the current 2024 we can see the less impact on these disorders which might suggest that with the current technology everyone are considering making necessary changes to make best out of what they have.

## Limitations

The scope of the dataset is set to be within the limits of the years ranging from 2020 – 2024 and also 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 particular dataset is focusing on disorders and also 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. These regression and classification algorithms used are the ones which are better suited for this numerical and categorical data than other algorithms and the accuracy scores are around 70%, The vast amount of the data has been cut shorted in data pre-processing where there are huge amounts of null and missing values. The models which we have used have better correlation and no outliers which resulted in the best ROC curves. The visualizations used perfectly depict the information which has to be represented with the given information.

# CHAPTER 7: REFERENCE

Banna, M. H. A., Ghosh, T., Nahian, M. J. A., Kaiser, M. S., Mahmud, M., Taher, K. A., ... & Andersson, K. (2023). A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data. *IEEE Access*, *11*, 77009-77022.

Wiedermann, C. J., Barbieri, V., Plagg, B., Marino, P., Piccoliori, G., & Engl, A. (2023, May). Fortifying the foundations: a comprehensive approach to enhancing mental health support in educational policies amidst crises. In *Healthcare* (Vol. 11, No. 10, p. 1423). MDPI.

Levin, M. (2023). Testing the applicability of idionomic statistics in longitudinal studies: The example of ‘doing what matters’. *Journal of Contextual Behavioral Science*, 100728.

Wunderlich-Barillas, T. (2023). Beyond COVID-19: the impact of recent pandemics on medical students and their education: a scoping review. *Medical education online*, *28*(1), 2139657.

White, M. (2023). Impact of the fear of catching COVID-19 on mental health in undergraduate students: A Predictive Model for anxiety, depression, and insomnia. *Current Psychology*, *42*(16), 13231-13238.

Dewan, A. (2022). Geography of children’s worry during the COVID-19 pandemic: insights into variations, influences, and implications. *Children's Geographies*, *22*(1), 116-133.

Jones, A. (2022). A systematic review and meta-analysis of longitudinal cohort studies comparing mental health before versus during the COVID-19 pandemic in 2020. *Journal of affective disorders*, *296*, 567-576.

Zheng, W., Chen, Q., Yao, L., Zhuang, J., Huang, J., Hu, Y., ... & Wang, Y. (2023). Prediction models for sleep quality among college students during the COVID-19 outbreak: cross-sectional study based on the internet new media. *Journal of Medical Internet Research*, *25*, e45721

Sodamade, O. (2023). Data analytics in public health, A USA perspective: A review. *World Journal of Advanced Research and Reviews*, *20*(3), 211-224.

Coley, R. L., Carey, N., Baum, C. F., & Hawkins, S. S. (2023). COVID-19 vaccinations and mental health among US adults: Individual and spillover effects. *Social science & medicine*, *329*, 116027.

Martin, J. (2023). Mental health and life satisfaction among those advised to shield during the COVID-19 pandemic in the UK: a secondary analysis of the understanding society longitudinal study. *Frontiers in public health*, *11*, 1235903.