EXPLORING THE RELATIONSHIP BETWEEN SUICIDE RATES AND MENTAL ILLNESS

Daniel Idogun,Sovanny Sreng, Hanh tran,Vasant Dave

Applied Arts and Technology
Algonquin College
Ottawa, Ontario
1385 Woodroffe Avenue, ON K2G 1V8

***Abstract*— In this paper, we identify the correlation between mental health and its cataclysmic outcome of suicide within United States of America. A Data-driven approach is used to identify some of the factors regarding mental health that affect suicide rates by integrating diverse datasets. Primarily three datasets are selected for analysis related to Suicide Rates, Indicators of anxiety or depression or both, and Mental health care received. Root cause analysis and Descriptive analysis techniques have been used to conduct a comprehensive examination of different information elements. Preliminary analysis signifies several trends and pattern changes for suicide rates during certain time periods. It also shows the association between certain demographic characteristics—such as age, gender, and socioeconomic status—and the prevalence of mental health disorders, including anxiety, depression, and substance abuse. Furthermore, our study highlights disparities between the number of People suffering from symptoms of anxiety or depression or both, and the actual number who utilize some of the methods to solve these disorders (prescription, counselling, or therapy). Leveraging machine learning techniques, we analyze demographic, clinical and socioeconomic factors to elucidate the underlying drivers of suicide rate. Using Algorithms such as Linear Regression, Logistic regression and Random Forest are used for different tasks related to future suicide rate prediction and mental health related classifications. Overall, this study contributes to a deeper understanding of the relationship between mental health and suicide rate within the USA, offering actionable insights for public health intervention and resolving a growing problem.**

Keywords- Suicide, Mental health, Suicide Prevention, Mental health treatment, Data analysis, Machine learning, AI, Python, Root cause Analysis.

I. INTRODUCTION

Mental health disorders are a significant challenge worldwide. There are more than 200 types of mental disorders.[1] Among the most severe consequences of untreated mental health disorders is suicide.[2] For the scope of this research, we have primarily emphasized our research on the two most impactful and widely seen mental disorders are anxiety and depression. Mental health is the state of wellbeing where an individual can recognize their own capabilities and deal with the normal stress of life.[3] It gives an individual strength to keep going and deal with any problems rationally and compassionately. Mental health illness not only affects the individual but also people near their proximity. When left untreated can lead to severe repercussions. A Canadian study showed that three percent of the violent crimes accruing to the sample data they collected were attributable to people with major mental disorders such as schizophrenia or depression. An additional seven percent were attributable to offenders with primary substance abuse disorder.[4] It has been observed that mental health is still a stigma in the modern information age. Firstly, it makes most people lose faith in themselves by society's labelling as weak or dangerous. Secondly it creates barriers and walls of confusion between the person suffering from the mental illness and the respondents. Inefficiency to rationalize their thoughts and delving into emotional depths finally result in isolation and alienation.[5] Self-stigma and social stigma both pose a deteriorating reverberation and traps people in vicious cycle where the outcomes are never in favor of goodness.

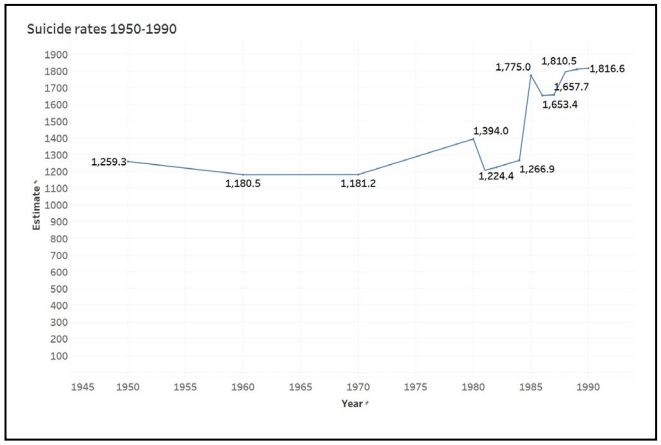
A. Research in United States of America

As for United states of America nearly 1 in 5 (19%) U.S adults experience some form of mental illness. [6] One in 24(4.1%) has a serious mental illness, one in 12(8.5%) has a diagnosable substance use disorder branching form mental illness. Furthermore, Lack of mental health treatment, facilities, awareness, social norms all worsen the effects of mental health illness. To this day the root cause of untreated mental health problems has been extensively debated and the result they give is negative implications.[7] It is seen that mental illness cases are rising at a faster pace than methods systematized to treat them. Met by high demand but faced by common barriers such as high cost and insufficient insurance coverage, Limited options and long wait times all multiply the negative effects of awareness and social norms. that mental health. Demographic factors such as ethnicity, gender or age play a key role in all observing the difference of likelihood where a subset of the population is more likely to have mental health disorders such as anxiety or depression, ultimately resulting in noticeable changes in the suicide rate among the related factors. [8]. The suicide rate among males in 2021 was approximately four times higher than the rate among females. Males make up 50% of the population but nearly 80% of suicides. There are many methods used for a tragic outcome as suicide, the most common of which is Firearms making approximate 56% (estimate count 26,328) of the method used for suicide. Followed up by 25.8% suffocation for the second most common and 11.6% Poisoning for the third most common. Additionally, During the last two decades suicide rate and statistics have seen many changes. Due to the stigma surrounding suicide, it is suspected that suicide is generally underreported. In April 2016, the CDC released data showing that the suicide rate in the United States had hit a 30-year high, and later in June 2018, released further data showing that the rate has continued to increase and has increased in every U.S. state except Nevada since 1999. [9] The total age-adjusted suicide rate in the United States in 2021 increased to 14.0 per 100,000. In 2021, the suicide rate among males was 4 times higher (22.8 per 100,000) than among females (5.7 per 100,000). [10]

II. ANALYSIS

A. Suicide rates analysis

In our study on suicide rates, indicators of anxiety or depression i.e. people having symptoms of anxiety or depression or both, and mental health care received through which methods and why, we employed a combination of root cause analysis and descriptive analysis methodologies to comprehensively investigate and understand the underlying factors contributing to this critical issue. Root cause analysis was utilized to delve deep into the fundamental reasons behind the occurrence of suicides, aiming to identify the primary drivers and triggers. This systematic approach allowed us to uncover the underlying societal, psychological,



and environmental factors that play a significant role in the incidence of suicides. Additionally, we conducted descriptive analysis to provide a detailed overview and summary of the

Fig. 1. Suicide rates between the years 1945-1990. Shows trendline change after 1980's.

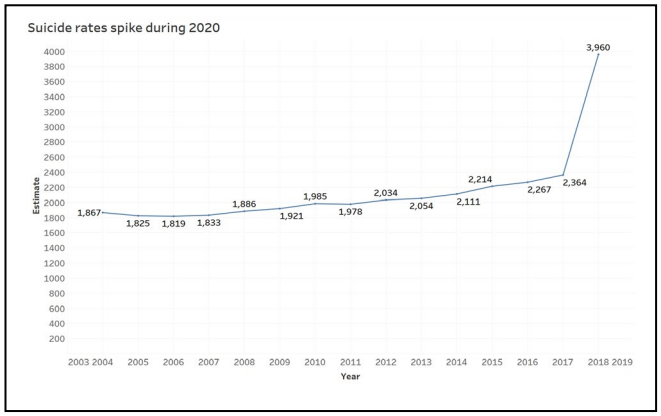


Fig. 2. Suicide rates between the years 2003-2019. A surge can be noticed after 2017.

prevailing trends, patterns, and demographics associated with suicide rates. By examining and organizing the available data, we were able to gain valuable insights into the characteristics and dynamics of suicide occurrences within the studied population. By integrating these two analytical approaches, our study not only helps explain the root causes of suicides but also offers a comprehensive understanding of the broader context in which these tragic events occur. This multifaceted analysis provides a solid foundation for the development of targeted interventions and strategies aimed at reducing suicide rates and promoting mental health and well-being in our society. In the case of suicide rates Together, these methodologies provide a comprehensive understanding of suicide rates, from patterns and trends to underlying causes. Through demographic analysis, such as age, gender, ethnicity, and socioeconomic status, descriptive analysis can pinpoint groups at higher risk of suicide. This information aids in tailoring prevention efforts to specific demographics, such as youth, veterans, or individuals in low-income communities. Root cause analysis goes beyond surface-level observations to unearth the systemic issues contributing to suicide rates. By examining factors such as access to care, socioeconomic disparities, and trauma, root cause analysis identifies the root causes of suicide.

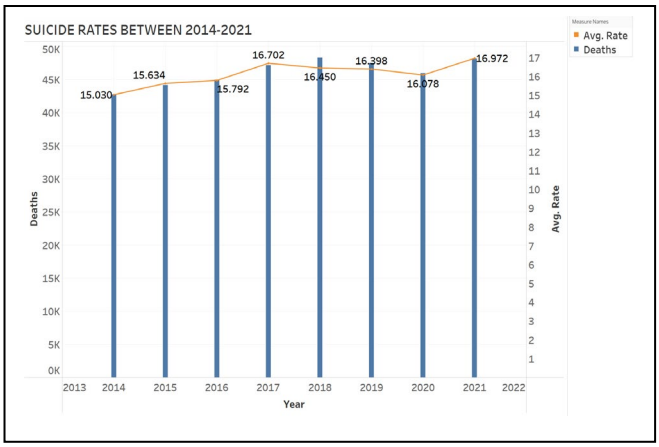


Fig. 3. Suicide rates between 2014-2021. Bars represent the deaths in estimate and orange trend line with labels represents the average rate.

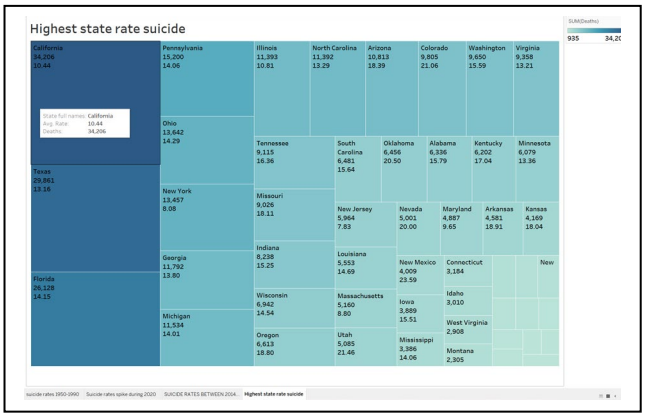


Fig. 4. Suicide rates and deaths by state. California, Texas, and Florida were the top 3 states for deaths.

Suicide has been a major issue for years, but suicides rates have not seen rapid changes throughout the years 1950 - 1980. [Fig. 1.] Rates increased 37% between 2000-2018 and decreased 5% between 2018-2020. However, rates nearly return to their peak in 2021. Certain trends and patterns have also been observed regarding suicide rates. Most notable of them being the sudden increase in suicide rates during the pandemic times. After which some variance can be observed. A huge spike in suicide rates can be seen just before the beginning of the pandemic era. [Fig. 2.] An evident parallel can be drawn between suicide rates and deaths, underscoring the profound impact of mortality on societal well-being.

Utilizing descriptive methods, we found that the same was the case. [Fig. 3.] Natural disasters and events in this period have also played a big role in shaping these rates. [11] Suicide rates increased in the four years after floods by 13.8 percent from 12.1 to 13.8 per 100,000. In the two years after hurricanes by 31.0 percent, from 12.0 to 15.7 per 100,000 and in the first year after earthquakes by 62.9 percent, from 19.2 to 31.3 per 100,000. The four-year increase of 19.7 percent after earthquakes was not statistically significant. Rates computed in a similar manner for the entire United States were stable. The increases in suicide rates were found for both sexes and for all age groups. The suicide rates did not change significantly after tornadoes or severe storms.[11] When examining suicide rates across states, California stood out with approximately 34,000 recorded deaths, marking it as having highest in the nation. [Fig. 4.] We can see most suicide rates have been seen in ages 75-84 years, followed by ages 85 years and over. [Fig. 5.]

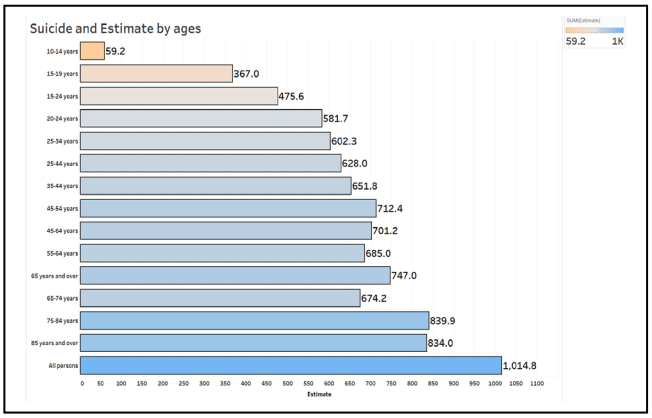


Fig. 5. Suicide rates and deaths by age. 75-84 age group observed to be the highest estimate count.

As males suicide counts were higher from 1950-2020, we wanted to compare it to the female count for the same frame to see how the trendline has changed. Male suicides saw a steep increase while for females it was a gradual increase. Not only that but male suicide rates changed notably over the years as compared to females. [Fig. 6.]

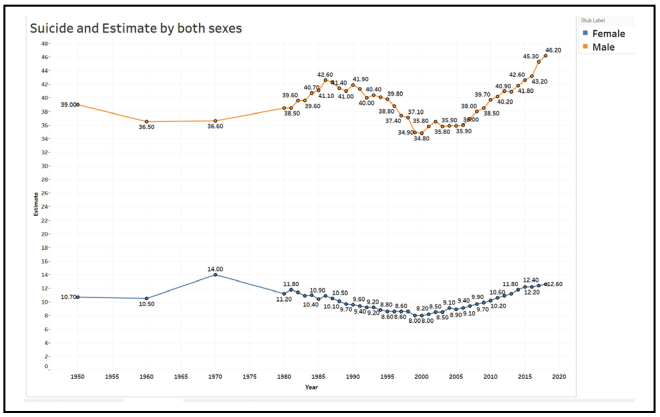


Fig. 6. Suicide rates and deaths by sex. Male group observed to be the highest estimate count.

To further examine the correlation between gender and ethnicity for suicides, another visualization is done. [Fig. 7.] In conclusion for the analysis for suicide rates, a definite line can be drawn to divide parallels that can be found with other datasets regarding correlated demographic factors which almost precisely matches the subsets and disparities explained through the analysis of mental care received.

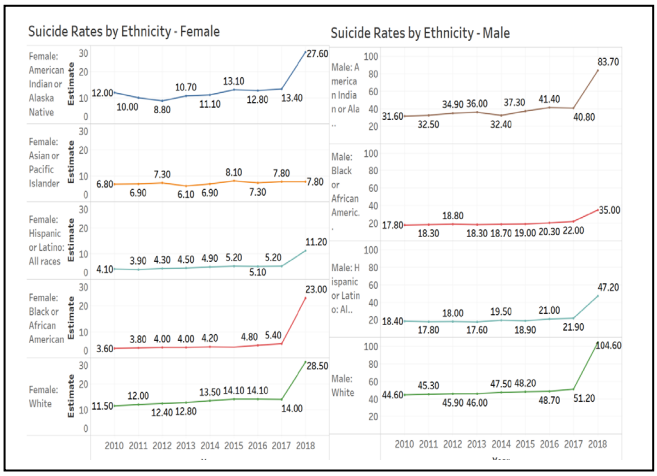


Fig. 7. Suicide rates by ethnicity and gender, most suicide rates for genders were observed in White races followed by American Indian or Alaska native in the second place.

B. *Anxiety and Depression analysis*

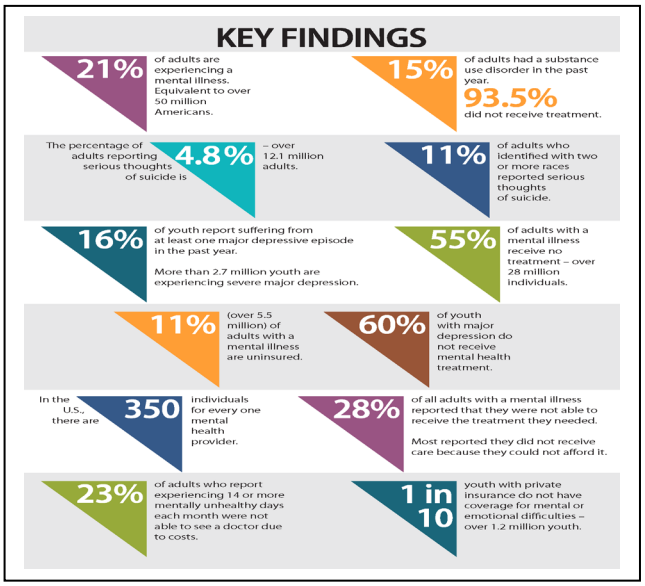


Fig. 8. Overview research of mental health disorders related with different aspects [12]

Overview research was done by a study on mental health disorders connecting to several different aspects of life highlighting the disparity mental health disorders and their lack of treatment. [Fig. 8.][12] Some demographics reported higher rates of anxiety and depression than others — including young adults, women, multi-racial Americans, and people without a bachelor's or higher degree. To delve deeper and comprehend this evident study, both analysis techniques again are utilized. Younger Americans reported the highest rates of anxiety and depression symptoms, with rates decreasing for each subsequent age group. [13] In the most recent round of the Household Pulse Survey from October 2023, over 48.4% of Americans ages 18–29 reported symptoms of anxiety or depression.

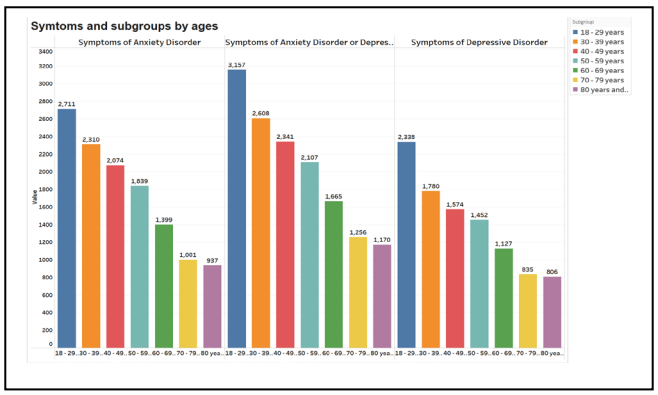


Fig. 9. Symptoms and subgroups by age. The 18-29 age group had the highest number of symptoms overall whether for anxiety or depression or both.

People 80 or older reported the lowest percentage at 18.8%. In October of 2023, nearly 5% more women than men reported symptoms of depression. Mental health illness for people having indicators such as anxiety or depression has also risen after 2020.[14] Pandemic has given a new rise in mental illness cases while the resources to offer required treatment to tackle these challenges stay unavailable or rise at a slower pace than mental illness cases. By age group a visualization was done to see what age groups are suffering the most from anxiety, depression, or both. [Fig. 9.] 18-29 age group reported the maximum number of symptoms for anxiety and depression.

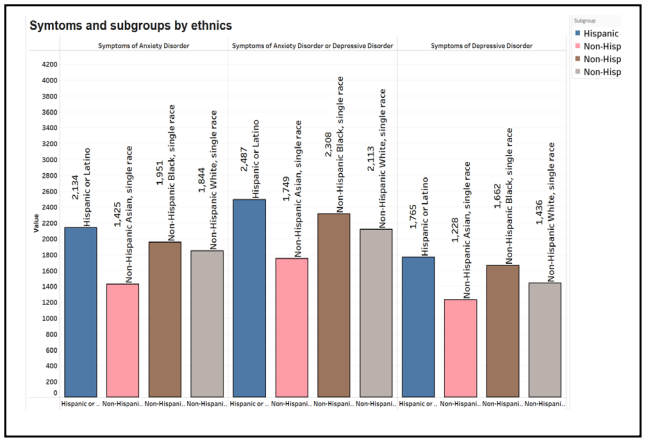


Fig. 10. Symptoms and subgroups by ethnicity. Hispanic or Latino race had the highest number of anxiety or depression symptoms.

The 30-39 and 40-49 age groups had the second most and third most symptoms count for anxiety or depression respectively. Analyzing ethnicity factor, Hispanic or Latino race had the highest count of symptoms followed by Non-Hispanic Black, Non-Hispanic White and Non-Hispanic Asian race in order. [Fig. 10.] Mental health illness due to stigma around it are many times not reported, thus the difference between suicide rates due to mental illness like anxiety and depression are far higher than the anxiety or depression cases. The people with anxiety or depression were also more likely to attempt suicide. [15] [16] 60% of suicides are linked with depression or anxiety. By examining ethnicity division for anxiety and depression we can clearly see how one race is at more risk of anxiety or depression than other races. Furthermore, one more demographic factor 'state' needed to be investigated. [Fig. 11.] The state of Louisiana had the Highest number of anxiety or depression symptoms with approximately 2332 cases, with Nevada (2286) and Mississippi (2236) in second and third positions respectively. Examining the gender comparison for symptoms of anxiety and depression. [Fig. 12.] Slightly a greater number of females were noted to have anxiety or depression symptoms, in some cases even both.

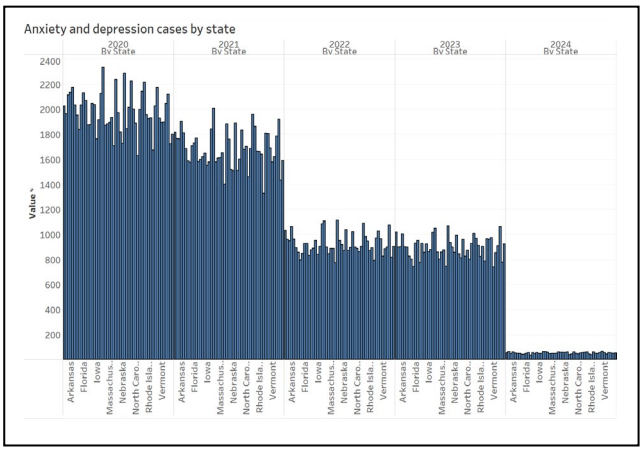


Fig. 11. Symptoms and subgroups by state. The state of Louisiana observed the most cases for symptoms of anxiety or depression followed by the state of Nevada and Mississipi in second and third place.

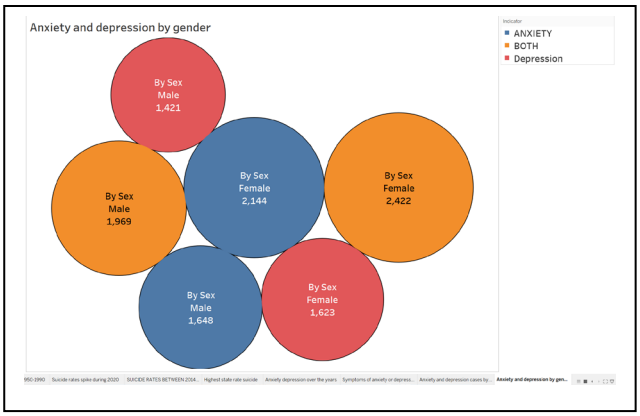


Fig. 12. Symptoms and subgroups by Gender. Female observations were slightly higher for anxiety or depression or both.

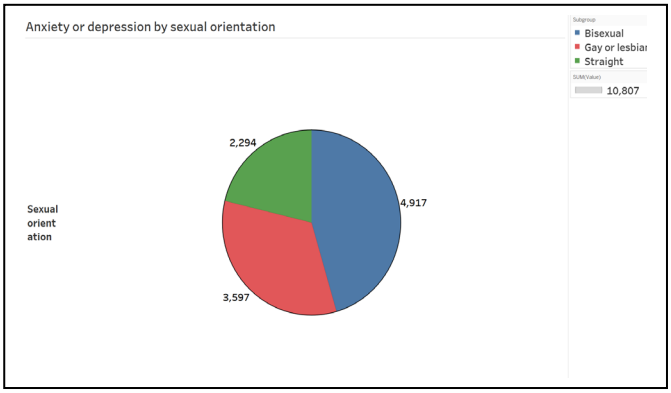


Fig. 13. Anxiety or depression by sexual orientation. Bisexual orientation group had the most number of symptoms 50% division.

While analyzing sexual orientation for symptoms of anxiety and depression revealed that bisexuality division as their sexual orientation reported more observations with an estimate 4917 (almost 50% of the value). [Fig. 13.] Summarizing mental health analysis. Our findings demonstrate that there are many underlying correlations between mental health illness and suicide rates based on demographic factors.

C. Mental health care analysis & conclusionary judgements

Mental health care and suicide rates all have one factor in common that was highlighted in bold during our research. It's the variance observed when examining mental health care received through various means among different segments. Discussing methods about how mental health care is received there are 2 primary methods observed in our dataset. It is either prescription by medicine or received counselling or therapy. Mental health care rise was seen between 2020-2021 for both methods took prescription or medicine or received therapy or counselling, but a similar value drop was noticed after 2021. [Fig. 14.]

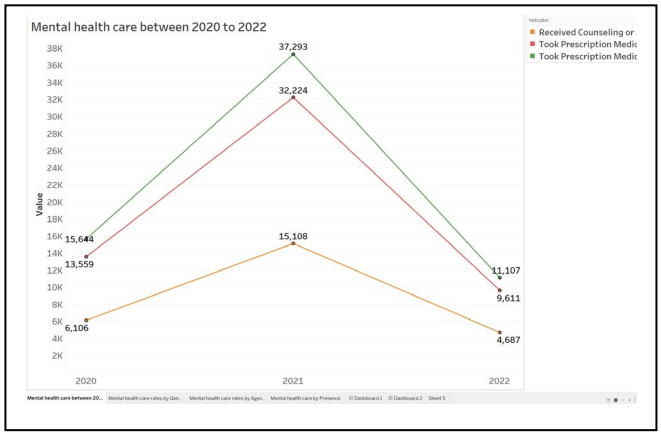


Fig. 14. Number of people that received mental health care between 2020 to 2022. Orange trendline represents people who received counselling as a way of care and red trendline represents people who received it through prescription, while green represents both.

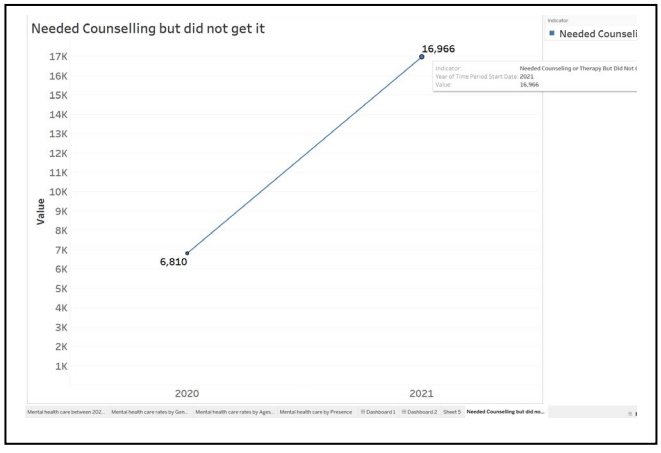


Fig. 15. The number of people that needed counselling or therapy but did not get it, which was 6810 in 2020, surged rapidly to unexpected approximate 17000 in just 1 year (2021).

Counselling or therapy is a prominent way to resolve mental health illnesses, especially in illnesses like anxiety or depression. When essential aid fails to reach those in need, problems escalate unabated, devoid of any solution to mitigate the incline. [Fig. 15.] More analysis shows that many people needed counselling but could not get it as not many public resources are available in place. There is a difference seen in rural and urban parts of America as well relative to the mental illness they have. The study mentions that 81 percent of rural counties do not have a psychiatric nurse practitioner. [17]. Adverse childhood experiences such as trauma or history of abuse, having feelings of isolation or loneliness, biological factors or chemical imbalances in the brain are all only some of the reasons for mental illness. [18] But when these reasons sprout in detrimental effects such as lack of awareness regarding mental health services, concern about confidentiality or the thought that an individual could address their problem by themselves, act as a barrier in providing the right treatment to the ones in need. [19] Cost affordability was the chief obstacle in providing counselling or therapy to the people. And again, as previously mentioned, stigmas and lack of awareness always play a major role in amplifying the setbacks. Moreover, we evaluated the mental health care received among different ages analogous to previous analysis. Among all age groups corresponding to our anxiety or depression analysis, the age group suffering the most from anxiety or depression is 18-29 years. Ultimately the same age group opting more for treatment. Therefore, the suicide rates for this age group (18-29) are relatively less compared to other age ranges. Although Prescription Or medicine is the optimal choice for many, 18-29 age subset inclines more towards counselling or therapy. [Fig. 16.] Expanding upon our earlier analysis, delving into gender correlations, we found that female category even though having slightly higher numbers of anxiety or depression were deciding to go for mental care through prescription or counselling. [Fig. 17.] This directly sustains the data why male suicide rates are relatively high in compared to females, as not many males opt for treatment due to various number of reasons (primarily seeking help may be viewed as a "weakness" leading man to be hesitant about seeking psychiatric help). [20] Ethnicity was the third factor that our analysis was focused on. There is a direct correlation as white or Caucasian ethnicity had the most suicide rates and the same subset of the population even selects and uses the methods of treatment. [Fig. 18.] Then one more question arises, because it was observed that Hispanic or Latino category for ethnicity had the greatest number of symptoms for anxiety, depression, or both. Then why is that portion of the population not electing treatment. Some studies have indicated that Latinos experience great difficulties in obtaining adequate access to mental health services and are underrepresented in mental health care settings. [21] Other studies have shown comparable levels of use of mental health services between Latinos and non-Latinos. Possible methodologic explanations for these divergent results are differences in the measures used to assess psychiatric disorders and service use response bias due to instrumentation differences in geographic locations, and differences in measures of access to mental health services. [21] To encapsulate our analysis, the issue of mental health is poised to escalate if not addressed with appropriate strategies and resources. Neglecting to tackle these conditions or address rising suicide rates could result in irreparable harm to society at large. In order to foresee

and classify with accuracy, we have utilized machine learning models to help achieve some of the predictions.

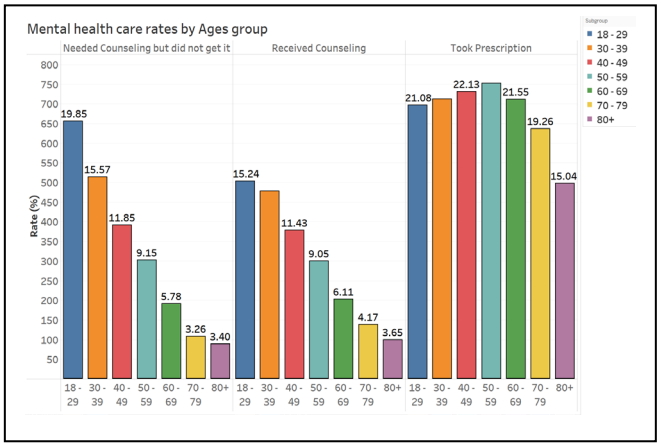


Fig. 16. Number of people that receive mental health care subsets by age. Higher numbers are seen using prescription as the method to go for and, average correlation with anxiety and depression analysis, the age group 18-29 had an average highest.

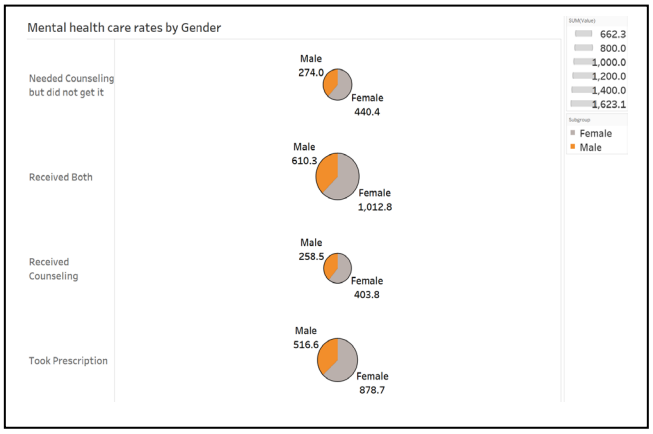


Fig. 17. Mental health care values divided by gender. Represents the Discrepancy between both the sexes when it comes to seeking treatment, with females utilizing treatment methods far more than males.

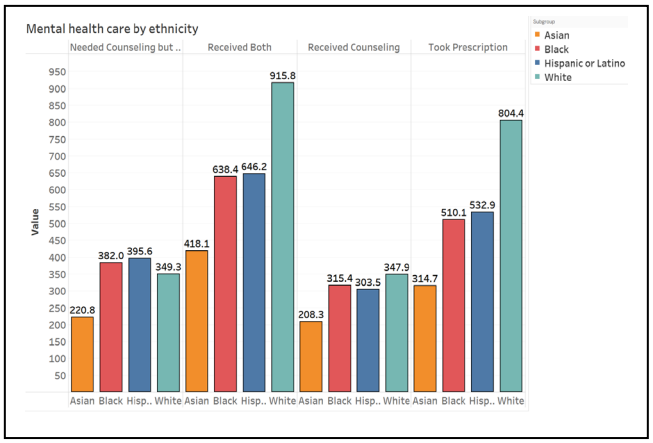


Fig. 18. Mental health care values divided by ethnicity. White or Caucasian ethnicity was the preeminent category for a singluar method of treatment or both combined.

III. DATA CLEANING

For an overview, we conducted a comprehensive data cleaning and preprocessing procedure across three distinct datasets focusing on suicide rates, indicators of anxiety or depression, and mental health care received. , which corresponded to the three sections for analysis that is Suicide rates, anxiety or depression, and Mental health care. [22-24] In all the three datasets , some of the common techniques such as, checking for outliers, validating the correct datatypes for columns (the first normalization form) , replacing missing values and dropping unecessary columns are used. In addition some feature engineering is done in some 2 of the datasets out the overall 3 as per the requirements, in order for the machine learning model to perform and predict better. Primarily used python libraries for cleaning and visualizing, such as Pandas,numpy ,seaborn and matplotlib, and Sklearn for machine learning algorithms imputations.

A. Cleaning and Preprocessing

1) Dataset 1 (Death rates by suicide)

This dataset had 12 columns and partial null in columns “estimate”. [Fig. 19.] Almost all the values in one column were missing(“FLAG”) and its information was not required, so it was dropped. In addition to that, we checked for outlier’s box plot for numerical values and count plot for categorical values (Step stays consistent for All datasets). [Fig. 20.] After which missing values must be handled. For numerical columns we had two choices to impute and replace the missing values with mean or median. Median is less sensitive to outliers but used when there are a smaller number of observations, and the skewness is not generalized, or data is heavily skewed. [25]

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6390 entries, 0 to 6389
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   INDICATOR       6390 non-null   object
 1   UNIT            6390 non-null   object
 2   UNIT_NUM        6390 non-null   int64
 3   STUB_NAME       6390 non-null   object
 4   STUB_NAME_NUM   6390 non-null   int64
 5   STUB_LABEL      6390 non-null   object
 6   STUB_LABEL_NUM  6390 non-null   float64
 7   YEAR            6390 non-null   int64
 8   YEAR_NUM        6390 non-null   int64
 9   AGE             6390 non-null   object
 10  AGE_NUM         6390 non-null   float64
 11  ESTIMATE        5484 non-null   float64
 12  FLAG            906 non-null    object
dtypes: float64(3), int64(4), object(6)
memory usage: 649.1+ KB
```

Fig. 19. Dataset 1 (suicide rates) Information about the columns before cleaning.

Fortunately for our data we didn’t have outliers, or the one outlier seen in visualization (‘year’) is justified. [Fig. 20.] In all three datasets we had a normal distribution of data. After careful consideration we replaced the missing values with

mean (column average) for estimate column over median (This step also stays consistent with all three datasets because we have replaced missing values with mean in all three datasets). Overall, our cleaning for dataset 1 was completed and only one step remained, that was feature engineering for this dataset. [Fig. 21.]

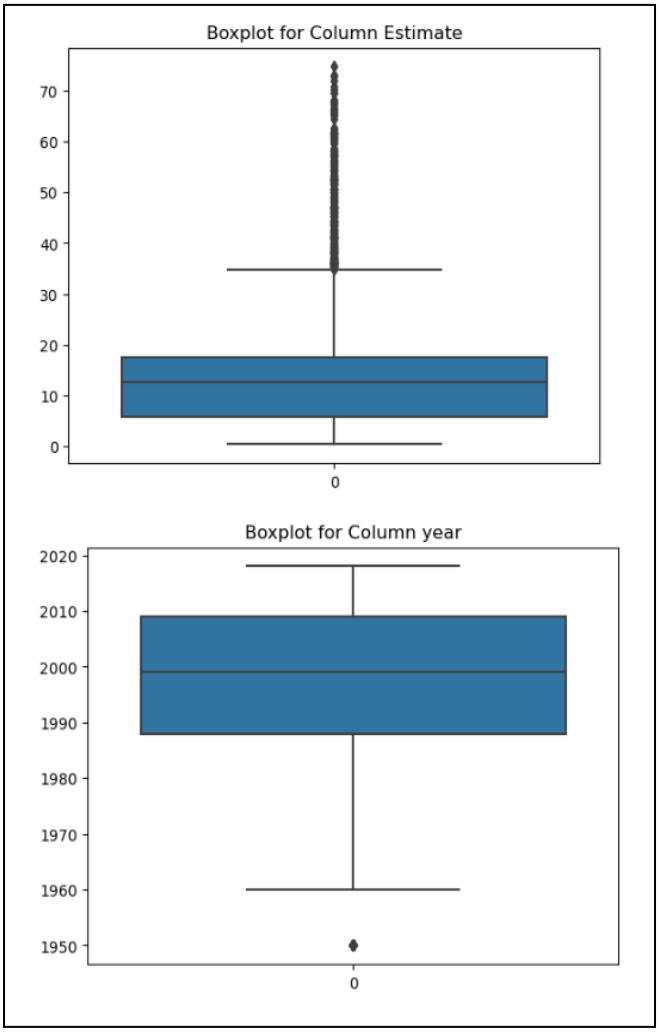


Fig. 20. Outliers check for columns estimate and year used for data modelling, which are used for data modelling after.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6390 entries, 0 to 6389
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   INDICATOR       6390 non-null   object
 1   UNIT            6390 non-null   object
 2   UNIT_NUM        6390 non-null   int64
 3   STUB_NAME       6390 non-null   object
 4   STUB_NAME_NUM   6390 non-null   int64
 5   STUB_LABEL      6390 non-null   object
 6   STUB_LABEL_NUM  6390 non-null   float64
 7   YEAR            6390 non-null   int64
 8   YEAR_NUM        6390 non-null   int64
 9   AGE             6390 non-null   object
 10  AGE_NUM         6390 non-null   float64
 11  ESTIMATE        6390 non-null   float64
dtypes: float64(3), int64(4), object(5)
memory usage: 599.2+ KB
```

Fig. 21. Dataset 1 (suicide rates) after cleaning

2) Dataset 2 (Indicators of Anxiety or depression)

Indicators of anxiety or depression contained many missing values in different columns and there were many columns which had less information about the dataset or were not required for predictive modelling. [Fig. 22.] Two Unnecessary columns with missing values were dropped: "Confidence Interval" and "Quartile range". Outliers were collectively checked as previously mentioned using Box plots for numerical values [Fig. 23.] and count plots for categorical values. [Fig. 24]

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14922 entries, 0 to 14921
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Indicator              14922 non-null  object
 1   Group                  14922 non-null  object
 2   State                  14922 non-null  object
 3   Subgroup               14922 non-null  object
 4   Phase                  14922 non-null  object
 5   Time Period            14922 non-null  int64
 6   Time Period Label      14922 non-null  object
 7   Time Period Start Date 14922 non-null  object
 8   Time Period End Date   14922 non-null  object
 9   Value                  14220 non-null  float64
 10  Low CI                 14220 non-null  float64
 11  High CI                14220 non-null  float64
 12  Confidence Interval    14220 non-null  object
 13  Quartile Range         9792 non-null   object
dtypes: float64(3), int64(1), object(10)
memory usage: 1.6+ MB
```

Fig. 22. Dataset 2 Information about the columns before cleaning.

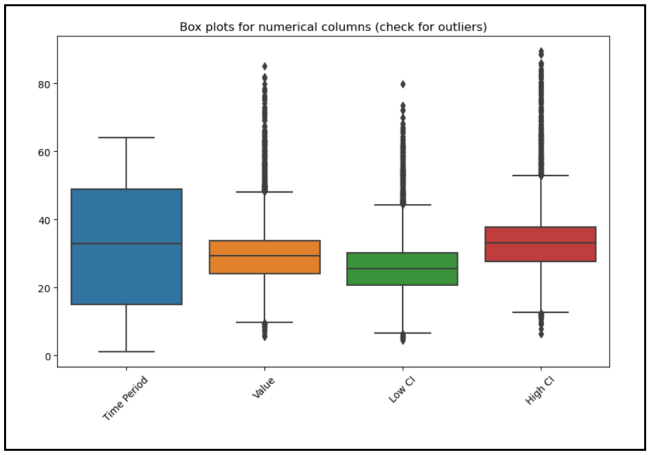


Fig. 23. Dataset 2 Information about outliers' numerical columns, Time period, value, low CI & high CI

Missing values again were replaced by the mean values. Therefore, three columns "Value", "Low CI", "High CI", mean for the columns was calculated and used to replace the null values. No feature engineering was done related to this dataset as it was not required. Occasionally adding features worsens the datasets and working for the model after inputting the dataset in the algorithms. There is a thin layer of difference in knowing what to keep, what to remove and what to add when handling a particular dataset. That is where understanding the data is helpful, and we completely grasped what needed to be done so that we can achieve good results. [26] Overall, the cleaning and preprocessing for this

dataset was complete and the only part left was to implement it in the data modelling. [Fig. 25]

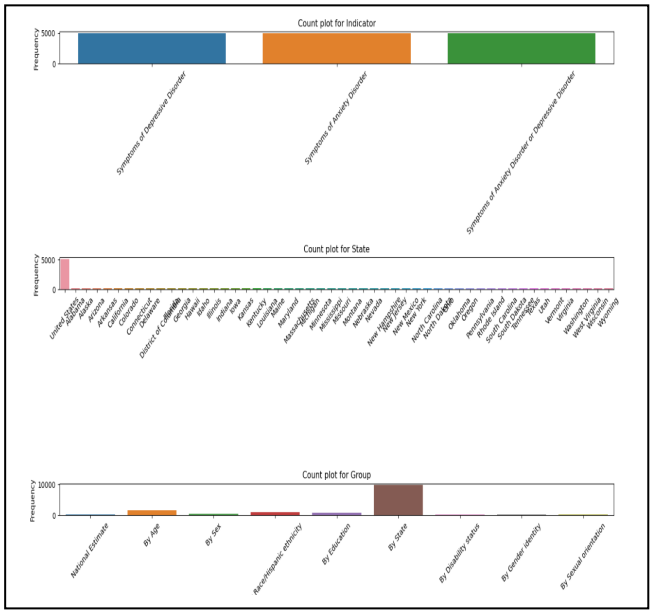


Fig. 24. Dataset 2 Categorical columns, count plots.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14922 entries, 0 to 14921
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Indicator               14922 non-null  object
 1   Group                   14922 non-null  object
 2   State                   14922 non-null  object
 3   Subgroup                14922 non-null  object
 4   Phase                   14922 non-null  object
 5   Time Period             14922 non-null  int64
 6   Time Period Label       14922 non-null  object
 7   Time Period Start Date  14922 non-null  object
 8   Time Period End Date    14922 non-null  object
 9   Value                   14922 non-null  float64
 10  Low CI                  14922 non-null  float64
 11  High CI                 14922 non-null  float64
dtypes: float64(3), int64(1), object(8)
memory usage: 1.4+ MB
```

Fig. 25. Dataset 2 after cleaning

3) Dataset 3 (Mental health care received)

In this dataset we have worked the most as some of these were essential for the algorithm used in predictive modelling. Starting with basic information [Fig. 26] once more the dataset had some missing values and some unnecessary columns with null values as well. In total three columns were dropped this time around as two were unnecessary and not useful, meanwhile one column was missing 98% values in rows ("Suppression flag"). [Fig. 26.] Missing values were handled in the same way as previous datasets, that is replacement with mean(average) of each individual column. Some of the row names were renamed as well for clear readability. For outliers' same technique as dataset 1 and 2 was used. In this dataset one more thing we observed was the datatype format for time period start date and time period end date did become a hindrance when implementing it for modelling. So, we changed the "time period start date" and "time period end date" datatypes to datetime. One feature engineering extract was done in this dataset as it was used for algorithm as an independent variable.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10404 entries, 0 to 10403
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Indicator               10404 non-null  object
 1   Group                   10404 non-null  object
 2   State                   10404 non-null  object
 3   Subgroup                10404 non-null  object
 4   Phase                   10404 non-null  object
 5   Time Period             10404 non-null  int64
 6   Time Period Label       10404 non-null  object
 7   Time Period Start Date  10404 non-null  object
 8   Time Period End Date    10404 non-null  object
 9   Value                   9914 non-null   float64
 10  LowCI                   9914 non-null   float64
 11  HighCI                  9914 non-null   float64
 12  Confidence Interval     9914 non-null   object
 13  Quartile Range          6732 non-null   object
 14  Suppression Flag        22 non-null     float64
dtypes: float64(4), int64(1), object(10)
memory usage: 1.2+ MB
```

Fig. 26. Dataset 3 before cleaning and information.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10404 entries, 0 to 10403
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Indicator               10404 non-null  object
 1   Group                   10404 non-null  object
 2   State                   10404 non-null  object
 3   Subgroup                10404 non-null  object
 4   Phase                   10404 non-null  object
 5   Time Period             10404 non-null  int64
 6   Time Period Label       10404 non-null  object
 7   Time Period Start Date  10404 non-null  datetime64[ns]
 8   Time Period End Date    10404 non-null  datetime64[ns]
 9   Value                   10404 non-null  float64
 10  LowCI                   10404 non-null  float64
 11  HighCI                  10404 non-null  float64
 12  Year                    10404 non-null  int32
dtypes: datetime64[ns](2), float64(3), int32(1), int64(1), object(6)
memory usage: 1016.1+ KB
```

Fig. 26. Dataset 3 after cleaning and information.

B. Feature engineering

1) Dataset 1 (suicide rates) (column extraction)

We have conducted feature engineering to create a new column termed "Gender" derived from the existing "Stub Label" column. This column contained various types of information beyond just gender-related details like ethnicity. To tackle this, we carefully analyzed the data to identify elements related to gender, such as "Male," "Female," or "Other". By categorizing the data in this way, I've made it simpler for researchers and readers to interpret and utilize the gender information within the dataset. [Fig. 27]

2) Dataset 3 (mental health care) column extraction

In the mental health care dataset, a combination between the time period started and the end dates is used to get a time range. And from that time range median, the initial year is used for all the values in the Year column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6390 entries, 0 to 6389
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   INDICATOR       6390 non-null   object
 1   UNIT            6390 non-null   object
 2   UNIT_NUM        6390 non-null   int64
 3   STUB_NAME       6390 non-null   object
 4   STUB_NAME_NUM   6390 non-null   int64
 5   STUB_LABEL      6390 non-null   object
 6   STUB_LABEL_NUM  6390 non-null   float64
 7   YEAR            6390 non-null   int64
 8   YEAR_NUM        6390 non-null   int64
 9   AGE             6390 non-null   object
 10  AGE_NUM         6390 non-null   float64
 11  ESTIMATE        6390 non-null   float64
 12  Gender          6390 non-null   object
dtypes: float64(3), int64(4), object(6)
memory usage: 649.1+ KB
```

Fig. 27. Dataset 1 after cleaning and feature engineering.

IV. DATA MODELLING AND MACHINE LEARNING

In our study, we have made three models for different problems and goals. All the objectives for each model were to predict based on historical data, with relevant variables and inputs. Regarding Suicide Rates our objective was to predict future suicide rates for the next three years (2025,2026,2027) based on historical trends and relevant predictors. After trial and error with different algorithms such as random forest regression our model was not performing as it should, and the prediction was near unexplainable, unjustifiable, and inaccurate. Linear regression was chosen as the best algorithm to predict suicide rates. [Fig. 28.] The annual suicide rate (number of suicides per population) for each year had to be the optimal choice for target variable. On the other hand, for the indicators of anxiety or depression dataset and model, objective was to predict the likelihood of an individual experiencing symptoms of anxiety or depression based on demographic, socioeconomic, and other relevant factors. Logistic Regression was utilized for this classification process. Target Variable was the indicator showing whether an individual exhibits symptoms of anxiety or depression (e.g., diagnosed with anxiety or depression disorder). Concerning the third model, the goal was to predict the type of mental care an individual is likely to receive based on their characteristics, symptoms, and other relevant predictors. The target variable was the type of mental care received by individuals (e.g., prescription medication, counseling). For this random Forest was selected initially no matter how much manual configuration for hyperparameters was done the result always ended in overfitting which is a difference between the accuracy observed in training set in comparison to test set. The training performed better with 80% accuracy and the test set resulted in only 55%. This is where we integrated grid search CV which checks and automatically finds the best configurations from a given set. Grid search CV is a technique for finding the optimal parameter values from a given set of parameters in a grid. This was very helpful in hyperparameter tuning and then the result was a generalized output. In predicting future suicide rates using linear regression, the aim is to anticipate the annual number of suicides per population for the years 2025, 2026, and 2027 by leveraging historical trends and pertinent predictors. Linear regression proves advantageous due to its ability to model linear relationships between the target variable (suicide rate) and various factors such as demographic factors and timeframe. For linear regression our accuracy for the model is approximately 60%. This approach enables the identification of significant predictors and trends, aiding in proactive intervention strategies and policy formulation to address mental health challenges. Meanwhile, the prediction of symptoms of anxiety or depression using logistic regression involves assessing the likelihood of individuals experiencing these symptoms based on demographic, socioeconomic, and other relevant factors. Logistic regression is apt for this task as it can estimate the probability of a binary outcome—whether an individual exhibits symptom of anxiety or depression—by considering multiple predictor variables. Accuracy for logistic regression was between 65-70% considering both sets. [Fig. 29.] This facilitates the identification of high-risk individuals who may benefit from targeted interventions or support services. Additionally, predicting the type of mental care individuals are likely to receive using random forest with integrated grid search cross-validation (CV) entails leveraging a more complex algorithm capable of handling non-linear relationships and interactions among predictors. [Fig. 30] By considering characteristics, symptoms, and other relevant predictors, random forest can provide insights into the most suitable types of mental care for individuals, such as prescription medication or counseling, thereby facilitating personalized treatment plans and improving mental health outcomes.

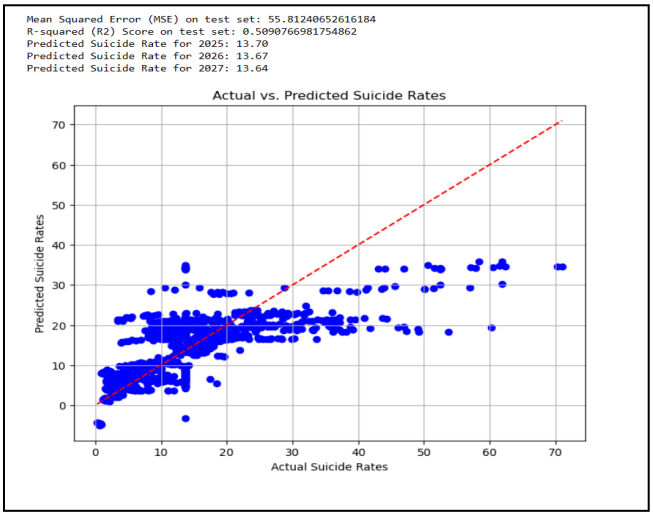
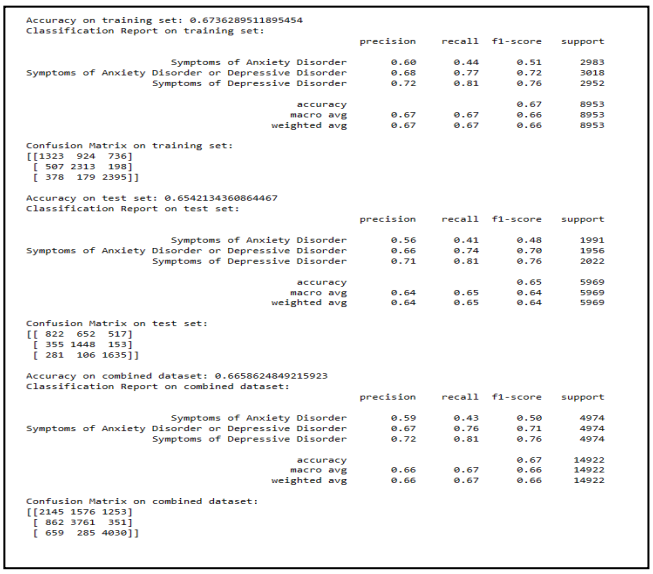


Fig. 28. Model 1 Linear Regression, choosing three independent variables (Year, Age and Gender) and Target variable (Estimate). Near accurate predictions when compared to websites doing the same.

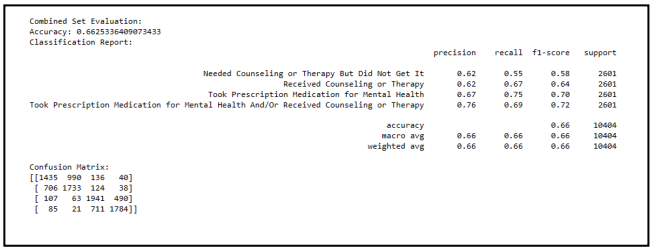
Many evaluation metrics were used related to models such as r2 and mean squared error (MSE) for regression and accuracy, precision, recall and support for classification and some formulars were kept as a reference to see how machine learning outperforms those formulas, so our r2 score is better while also being understandable. [Fig. 31] One of the key hurdles encountered during our predictive modeling endeavors was the constraint imposed by limited data availability, which ultimately hindered our ability to achieve accuracy rates exceeding 70% across all models. This limitation significantly impacted the robustness and

reliability of our predictions, as the models lacked sufficient data points to effectively capture the complexities and nuances of the underlying patterns. Despite employing various techniques and methodologies to optimize model performance, including feature engineering and hyperparameter tuning, the inherent scarcity of data acted as a bottleneck, preventing us from attaining higher levels of accuracy. Moving forward, addressing this challenge necessitates exploring avenues to augment the dataset through data collection efforts, leveraging alternative data sources, or implementing advanced techniques such as transfer learning to extract meaningful insights and improve predictive performance. Overall, these predictive modeling approaches contribute to more informed decision-making in mental health policy, intervention strategies, and personalized care delivery.



```
Accuracy on training set: 0.6736289511895454
Classification Report on training set:
                                                precision    recall  f1-score   support

                     Symptoms of Anxiety Disorder       0.60      0.44      0.51      2983
Symptoms of Anxiety Disorder or Depressive Disorder       0.68      0.77      0.72      3018
                  Symptoms of Depressive Disorder       0.72      0.81      0.76      2952

                                         accuracy                           0.67      8953
                                        macro avg       0.67      0.67      0.66      8953
                                     weighted avg       0.67      0.67      0.66      8953

Confusion Matrix on training set:
[[1323  924  736]
 [ 507 2313  198]
 [ 378  179 2395]]

Accuracy on test set: 0.6542134360864467
Classification Report on test set:
                                                precision    recall  f1-score   support

                     Symptoms of Anxiety Disorder       0.56      0.41      0.48      1991
Symptoms of Anxiety Disorder or Depressive Disorder       0.66      0.74      0.70      1956
                  Symptoms of Depressive Disorder       0.71      0.81      0.76      2022

                                         accuracy                           0.65      5969
                                        macro avg       0.64      0.65      0.64      5969
                                     weighted avg       0.64      0.65      0.64      5969

Confusion Matrix on test set:
[[ 822  652  517]
 [ 355 1448  153]
 [ 281  106 1635]]

Accuracy on combined dataset: 0.6658624849215923
Classification Report on combined dataset:
                                                precision    recall  f1-score   support

                     Symptoms of Anxiety Disorder       0.59      0.43      0.50      4974
Symptoms of Anxiety Disorder or Depressive Disorder       0.67      0.76      0.71      4974
                  Symptoms of Depressive Disorder       0.72      0.81      0.76      4974

                                         accuracy                           0.67     14922
                                        macro avg       0.66      0.67      0.66     14922
                                     weighted avg       0.66      0.67      0.66     14922

Confusion Matrix on combined dataset:
[[2145 1576 1253]
 [ 862 3761  351]
 [ 659  285 4030]]
```

Fig. 29. Model 2 Logistic Regression, choosing four independent variables (State, Group, Time Period Start Date, Value) and Target variable (Indicator).



```
Combined Set Evaluation:
Accuracy: 0.6625336409073433
Classification Report:
                                                                precision    recall  f1-score   support

                                Needed Counseling or Therapy But Did Not Get It       0.62      0.55      0.58      2601
                                                Received Counseling or Therapy       0.62      0.67      0.64      2601
                                            Took Prescription Medication for Mental Health       0.67      0.75      0.70      2601
Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy       0.76      0.69      0.72      2601

                                                                     accuracy                           0.66     10404
                                                                    macro avg       0.66      0.66      0.66     10404
                                                                 weighted avg       0.66      0.66      0.66     10404

Confusion Matrix:
[[1435  990  136   40]
 [ 706 1733  124   38]
 [ 107   63 1941  490]
 [  85   21  711 1784]]
```

Fig. 30. Model 3 Random Forest, choosing three independent variables (Year, Group, Value) and Target variable (Indicator).

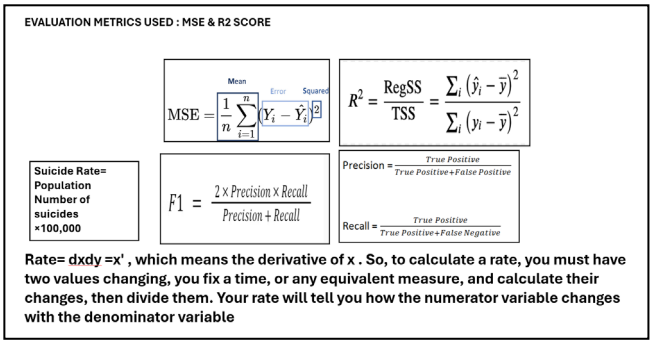


Fig. 31. Model 3 Random Forest, choosing three independent variables (Year, Group, Value) and Target variable (Indicator).

V. CONCLUSION

Mental health is an integral part of life and making sure that all the subsets of the populace are in serenity of mind and well-being is our responsibility. Furthermore, when we scrutinized the fourth dataset, we sought to understand how suicide rates are changing by that geographic variability, observing how in some places mental health is more neglected and finally resulting in a catastrophic end such as suicide [27]. As society and culture have a huge impact on mental health, some positive while some negative, it the society's responsibility is to negate the causes and help in whatever way possible. [28] Our collective aim in society should be the overall welfare and happiness of its members. And it all starts with us as individuals in every aspect of day-to-day life. Studies have shown that mental health is a public health issue not only affecting the relationships close to oneself but also the people who the person meets for the first time. [29] Programs for children, programs for older adults, violence prevention initiatives and policy advocacy are all some of the ways to spread awareness while also upholding values and principles of the society, simultaneously making the future a better place for everyone. Many of the adverse consequences we see in our lives are connected to the fragility of the human mind, which if disregarded for too long, takes an abominable form. [30] In summary our analysis highlights the complex nature of mental health, showing how different factors like age, ethnicity, gender, state, and community support all affect the wellbeing of a subset. We have identified significant trends and disparities in mental health outcomes, emphasizing the importance of targeted interventions and comprehensive support systems. Our findings emphasize the need for holistic approaches that address not only individual symptoms but also underlying social determinants and systemic barriers. Moving forward, fostering greater awareness, promoting destigmatization, and prioritizing accessible mental health services are essential steps towards fostering a society where mental well-being is valued and supported for all individuals. Ultimately, our goal should be to create a society where everyone's mental well-being is valued and nurtured without any judgement or lack of resources.

REFERENCES

[1] Cleveland Clinic, "Mental Health Disorders: Types, Diagnosis & Treatment Options," Cleveland Clinic, Jan. 24, 2022. https://my.clevelandclinic.org/health/diseases/22295-mental-health-disorders (accessed Apr. 07, 2024).

[2] L. Brådvik, "Suicide Risk and Mental Disorders," *International Journal of Environmental Research and Public Health*, vol. 15, no. 9, p. 2028, 2019, doi: https://doi.org/10.3390/ijerph15092028.

[3] I. of C. Energetics, "The Importance of Mental and Emotional Well-Being," *Institute of Core Energetics*, Sep. 10, 2023. https://www.coreenergetics.org/the-importance-of-mental-and-emotional-well- being/?gad_source=1&gclid=EAIaIQobChMIvd2bu6GkhQMVW1VHAR0ksQnAEAAYASAAEgJxD_D_BwE (accessed Apr. 08, 2024).

[4] H. STUART, "Violence and Mental illness: an Overview," *World Psychiatry*, vol. 2, no. 2, pp. 121–124, Jun. 2003, Accessed: Apr. 05, 2025.[Online].Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525086/#:~:text=Three%20percent%20of%20the%20violent

[5] P. W. Corrigan and A. C. Watson, "Understanding the impact of stigma on people with mental illness," *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, vol. 1, no. 1, pp. 16–20, Feb. 2002, Accessed: Apr. 06, 2024.[Online].Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489832/

[6] "What Is Mental Illness?," *www.psychiatry.org*. https://www.psychiatry.org/patients-families/what-is-mental-illness#:~:text=In%20a%20given%20year%3A (accessed Apr. 04, 2024).

[7] "Study Reveals Lack of Access as Root Cause for Mental Health Crisis in America," *National Council for Mental Wellbeing*. https://www.thenationalcouncil.org/news/lack-of-access-root-cause-mental-health-crisis-in-america/?gad_source=1&gclid EAIaIQobChMI0Luy-p-khQMVMEpHAR1-0ANBEAAYASAAEgLQCvD_BwE (accessed Apr. 09, 2024).

[8] Centers for Disease Control and Prevention, "Suicide Data and Statistics," *www.cdc.gov*, Aug. 10, 2023. https://www.cdc.gov/suicide/suicide-data-statistics.html (accessed Apr. 08, 2024)

[9] "Suicide in the United States," *Wikipedia*, May 08, 2023. https://en.wikipedia.org/wiki/Suicide_in_the_United_States#:~:text=From%202000%20to%202020%2C%20more (accessed Apr. 07, 2024).

[10] "Suicide - National Institute of Mental Health (NIMH)," *www.nimh.nih.gov*, Feb. 2024. https://www.nimh.nih.gov/health/statistics/suicide#:~:text=100%2C000%20in%202020.- (accessed Apr. 08, 2024).

[11] E. G. Krug *et al.*, "Suicide after Natural Disasters," *New England Journal of Medicine*, vol. 338, no. 6, pp. 373–378, Feb. 1998, doi: https://doi.org/10.1056/nejm199802053380607.

[12]"Mental Health Disorder Statistics,"*www.townsendla.com*. *https://www.townsendla.com/blog/mental-health-disorder-tatistics* (accessed Apr. 07, 2024).

[13] "Who experiences anxiety and depression in the US?," *USAFacts*. https://usafacts.org/articles/who-experiences-anxiety-and-depression-in-the-us/#:~:text=The%20gender%20gap%20in%20rates (accessed Apr. 09, 2024).

[14] A. Vahratian, "Symptoms of Anxiety or Depressive Disorder and Use of Mental Health Care Among Adults During the COVID-19 Pandemic — United States, August 2020–February 2021," *MMWR. Morbidity and Mortality Weekly Report*, vol. 70, no. 13, Apr. 2021, doi: https://doi.org/10.15585/mmwr.mm7013e2.

[15] J. Zhang, X. Liu, and L. Fang, "Combined Effects of Depression and Anxiety on suicide: a case-control Psychological Autopsy Study in Rural China," *Psychiatry Research*, vol. 271, pp. 370–373,Jan. 2019, doi: https://doi.org/10.1016/j.psychres.2018.11.010.

[16] L. Holmes, "Rates and Statistics for Suicide in the U.S.," *Verywell Mind*, Jun. 24, 2021. https://www.verywellmind.com/suicide-rates-overstated-in-people-with-depression-2330503 (accessed Apr. 09, 2024).

[17] Rural Minds, "Serving Rural America," *Rural Minds*, 2024. https://www.ruralminds.org/serving-rural-america?gad_source=1&gclid=EAIaIQobChMI3camgKukhQMV4HJHAR0_FQeAEAAYAiAAEgKhf_D_BwE (accessed Apr. 12, 2024).

[18] CDC, "About Mental Health," *www.cdc.gov*, Apr. 28, 2023. https://www.cdc.gov/mentalhealth/learn/index.htm#:~:text=More%20than%201%20in%205 (accessed Apr. 11, 2024).

[19] J. Conroy, L. Lin, and A. Ghaness, "Why People aren't Getting the Care They Need," *Apa.org*, 2022. https://www.apa.org/monitor/2020/07/datapoint-care#:~:text=People%20cited%20many%20reasons%20for (accessed Apr. 11, 2024).

[20] S. Siddharth K., "Why Men Don't Go to Therapy," *www.orlandohealth.com*. https://www.orlandohealth.com/content-hub/why-men-dont-go-to-therapy#:~:text=Seeking%20help%20may%20be%20viewed (accessed Apr. 09, 2024).

[21] M. Alegría *et al.*, "Mental Health Care for Latinos: Inequalities in Use of Specialty Mental Health Services Among Latinos, African Americans, and Non-Latino Whites," *Psychiatric Services*, vol. 53, no. 12, pp. 1547–1555, Dec. 2002, doi: https://doi.org/10.1176/appi.ps.53.12.1547.

[22] U.S. Department of Health & Human Services, "Death rates for suicide, by sex, race, Hispanic origin, and age: United States," *Data.gov*, Apr. 27, 2022. https://catalog.data.gov/dataset/death-rates-for-suicide- by-sex-race-hispanic-origin-and-age-united- states-020c1 (accessed Apr. 11, 2024).

[23] U.S. Department of Health & Human Services, "Mental Health Care in the Last 4 Weeks," *Data.gov*, May 18, 2022. https://catalog.data.gov/dataset/mental-health-care-in-the-last- 4-weeks (accessed Apr. 11, 2024).

[24] U.S. Department of Health & Human Services, "Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days," *Data.gov*, Jan. 06, 2021. https://catalog.data.gov/dataset/indicators-of-anxiety-or-depression-based-on-reported-frequency-of-symptoms-during- last-7- (accessed Apr. 11, 2024).

[25] T. Firdose, "Filling missing values with Mean and Median," *Medium*, May 29, 2023. https://tahera- firdose.medium.com/filling-missing-values-with-mean-and-median-76635d55c1bc (accessed Apr. 11, 2024).

[26] K. Srivastava, "What to Keep and What to Remove," *The Startup*, Aug. 22, 2020. https://medium.com/swlh/what-to-keep-and-what-to-remove-74ba1b3cb04 (accessed Apr. 12, 024).

[27] Centers for Disease Control and Prevention, "Suicide Rates by State," Centers for Disease Control and Prevention, USA, 2021. https://www.cdc.gov/suicide/suicide-rates-by-state.htmL

[28] O. of the S. General (US), C. for M. H. Services (US), and N. I. of M. Health (US), "Chapter 2 Culture Counts: the Influence of Culture and Society on Mental Health," *www.ncbi.nlm.nih.gov*, Aug. 01, 2001. https://www.ncbi.nlm.nih.gov/books/NBK44249/#:~:text=For%20example%2C%20supportive%20families%20and (accessed Apr. 11, 2024).

[29] Tulane University, "Understanding Mental Health as a Public Health Issue," *publichealth.tulane.edu*, Jan. 13, 2021. https://publichealth.tulane.edu/blog/mental-health-public-health/ (accessed Apr. 11, 2024).

[30] "Unattended Mental Health's Impact on Society," Tacoma-Pierce County Health Department , USA, Jan. 2016. https://tpchd.org/wp-content/uploads/2023/12/Unattended-Mental-Health-Impact-on-Society.pdf