Predicting Depression Using Machine Learning

Vivian Dong 300525329, Saumya Sajwan, Samanalie Perera

November 1, 2022

Contents

1 Executive Summary

2 Background

A study found that depression increased from 9% in 2017–2018 to over 14% in April 2020 among US adults during the COVID-19 pandemic. A study found that depression increased

from 9% in 2017–2018 to over 14% in April 2020 among US adults during the COVID-19 pandemic. Among different age groups, the mental health of young adults is most affected by the pandemic. This increased rate was not normal compared to other years, and there was not enough attention paid to data examining tools either. (Daly, Sutin, and Robinson 2021) Although there is an increasing number of adults with depression, the average treatment cost per individual decreased, which means more and more people have not been able to receive any treatment.(Greenberg et al. 2021) People who are depressed are also far more likely to be diagnosed with other diseases, such as heart disease, diabetes, and high blood pressure.

There are blood tests, brain scans and other medical examining methods for a depression diagnosis. In the end, the most affected way is to let the patients describe their symptoms. To achieve this, patients can answer a questionnaire like determining the frequency of depression symptoms over the past two weeks. This method may lead to subjective bias and imperfections in the diagnostic capabilities. (Buntinx et al. 2004) However, the cause of depression could be from many other things, such as sleep disorders, drug use, alcohol use and weight loss. For example, if a person has been working in a stressful environment with low income and not enough sleep, that person may have a higher chance than other people to have depression.

Healthcare data from the National Health and Nutrition Examination Survey database includes a wide range of concepts, like health records, genetic information and even demographic data. Furthermore, machine learning tools tend to perform better than humans at processing these big data sets and making use of it.(Beam and Kohane 2018)

Our primary goal in this project was to train a machine learning classification model to identify patients who suffer from depression using demographics and healthcare data from the NHANES database.

3 Data Description

The NHANES 2005 - March 2020 year range was selected as the data set for this project. We did not go before 2005 because the survey questions were different compared to later years. Demographics, sleep disorders, alcohol use, smoking (cigarette use) and weight history information are used as predictors since they are found to be primary factors for the cause of depression. We manually selected the variables with less than 50% missing rate from these data sets. For our target variable - depression, we used the Mental Health - Depression Screener data sets and PHQ-9 scoring system (Bhatt et al. 2016) to identify whether a person has depression. The final scores of 0–4, 5–9, 10–14, 15–19, and 20–27 are the ranges for none, mild, moderate, moderately severe and severe depression, respectively. In our project, we wanted to focus on building a binary classification model, so if the respondent has a total score that is greater than or equal to 10, then the individual is identified as having depression.

As a result, we have 29 variables in our data set, including id, depression result, and 27 predictors, including 15 numerical and 12 categorical variables. Table 1 and 2 show the descriptions of data.

Table 1: Description of the numerical data

Variable	Description
id	Unique identifier for each respondent
age	The age of the respondent
family_PIR	Poverty income ratio (PIR) - a ratio of family income to poverty threshold
sleep_hours	Total hours of sleep
drinks_per_occasion	Average drinks per day
SMD030	Age started smoking cigarettes regularly
SMD641	Number of days smoked cigarettes during the past 30 days
SMD650	Average number of cigarettes per day during past 30 days
SMD630	Age first smoked the whole cigarette
WHD010	Height of the respondent (inches)
WHD020	Weight of the respondent (pounds)
WHD050	Weight of the respondent (pounds) 1 year ago
WHD110	Weight of the respondent (pounds) 10 years ago
WHD120	Weight of the respondent (pounds) at the age of 25
WHD140	Respondent's heaviest weight (pounds)
WHQ150	Age of the respondent when heaviest weight

Table 2: Description of the categorical data

Variable	Description
result	Whether the respondent has depression (1=Yes, 0=No)
gender	Gender of respondent
race	Race of respondent
marital_status	The marital status of the respondent
education_level_adults	Highest level of education of the respondent
language	Language of the respondent
trouble_sleeping_history	Whether had trouble sleeping
SMQ020	Whether had smoked at least 100 cigarettes in life
SMQ040	Frequency of smoking cigarettes
SMQ670	Whether tried to quit smoking
WHQ030	How respondent consider their weight
WHQ040	Respondent likes to weigh more, less or the same
WHQ070	Whether the respondent tried to lose weight in the past year

We had 43,928 entries when we first combined all the data sets, but we need the respondent to answer every single question in the Mental Health - Depression Screener Survey to calculate the score. Therefore, we had to remove the respondent who did not complete the survey, which gave us 26,473 data at the end. (17,455 respondents were taken out)

The structure of the missing data in our data set varied from variable to variable, so we had to find the data description for each variable from the NHANES website and convert them to NA. For example, 7, 777 and 7777 could all be refused to answer the survey question; both "-1" and "." mean no answers.

Another problem we had was that the variables in the data set from each year may be different or have different names. We ended up choosing the variables that appeared every year and changing them to have the same names.

4 Ethics, Privacy and Security

4.1 Ethical Considerations

There are a number of ethical concerns that need to be considered when developing and deploying a machine learning algorithm that predicts if an individual is depressed. First, it is important to consider the potential impact of false positives and false negatives. If the algorithm incorrectly predicts that an individual is depressed, this could lead to them being unnecessarily treated or stigmatized. On the other hand, if the algorithm fails to predict that an individual is depressed, this could result in them not receiving the help they need.

4.2 Privacy Concerns

4.2.1 Access and use

Whether we got permission and notified every respondent before using their information is essential, especially under legislation in different states. Also, our model requires access to many respondents' data, and we need to prevent the data from being used in different ways over time.

4.2.2 Re-identification

Another concern with healthcare data is whether we can protect patients' information. A lot of research shows that people can use different techniques to re-identify individuals in the data. However, on the other hand, too much de-identification may diminish the clinical utility of the data, but too little de-identification may lead to a breach of privacy. (Emam et al. 2009)

4.3 Security concerns

If the algorithm is made public via a data leak, then anyone could use it to find out which individuals are more likely to be depressed. This information could be used to target those individuals with ads or content that exploits their vulnerabilities. For example, an advertiser could show ads for antidepressant medications to someone who is predicted to be depressed.

4.3.1 Steps to secure data

The steps we can take to maintain integrity and confidentiality is making sure that only authorized users have access to the data. To do that, we can:

- Make the GitHub repository private. That way only authorized people have access (group members)
- Password can protect our data- with a password that is only distributed to the users who are authorized access
- To protect the integrity of the data, we can make sure only to have information that is needed for this analysis, as well as not re-identifying the data (which is mentioned in privacy concerns)

5 Exploratory Data Analysis

- Revise the style and content of the EDA section from the Phase 1 group report in response to the feedback you received about that report
- Select a subset of the results from the Phase 1 group report. You do not need to include every plot from the EDA assignment. You may want to choose particular plots that turned out to be most relevant for the detailed analysis.
- A few summary tables and/or plots, each describing one, two or three variables in the data that you thought were interesting, or were relevant for the detailed analysis
- Tables can be summary statistics of individual variables, contingency tables for pairs of categorical variables, or results tables for basic analyses
- Explain the definitions of the variables in each table/plot
- Comment on the main features of each plot
- Include suitable labels and keys for each plot
- Make sure all plots would be readable if printed in black & white, and adjust the point sizes and/or line thicknesses to improve readability
- Lay out all tables so that they are clearly readable and clearly labelled, and do not use excessive significant figures

5.1 Read Data

```
##      id result age gender race marital_status family_PIR education_level_adults
## 1 31139      0  18      2    2              5       4.91                     NA
## 2 31143      0  19      1    3              5       5.00                     NA
```

```
## 3 31150      0  79      1    3              3       1.22                      3
## 4 31151      0  59      2    4              1       3.03                      3
## 5 31152      0  27      2    1              1       1.76                      3
## 6 31153      0  44      2    5              5       1.03                      3
##   language sleep_hours trouble_sleeping_history drinks_per_occasion SMQ020
## 1        1           6                         2                  NA     NA
## 2        1           7                         2                  NA     NA
## 3        1           8                         2                   3      1
## 4        1           2                         1                  NA      1
## 5        1           8                         2                  NA      2
## 6        1           7                         1                  NA      1
##   SMD030 SMQ040 SMD641 SMD650 SMD630 SMQ670 WHD010 WHD020 WHQ030 WHQ040 WHD050
## 1     NA     NA     NA     NA     NA     NA     62    160       3      3    190
## 2     NA     NA      2      2     17      2     73    160       3      3    150
## 3     14      3     NA     NA     NA     NA     69    190       3      3    190
## 4     25      3     NA     NA     NA     NA     67    180       1      2    250
## 5     NA     NA     NA     NA     NA     NA     59    214       1      2    214
## 6     15      3     NA     NA     NA     NA     66    230       1      2    280
##   WHQ070 WHD110 WHD120 WHD140 WHQ150
## 1     NA     NA     NA    190     17
## 2      2     NA     NA    160     19
## 3      2    190    170    227     38
## 4      2    365    145    365     57
## 5      1     NA    195    214     26
## 6     NA    280    145    280     43
```

id	result	age	gender	race	marital_status	family_PIR	SMQ020
31139	0	18	2	2	5	4.91	NA
31143	0	19	1	3	5	5.00	NA
31150	0	79	1	3	3	1.22	1
31151	0	59	2	4	1	3.03	1
31152	0	27	2	1	1	1.76	2

id	education_level_adults	language	sleep_hours	trouble_sleeping_history
31139	NA	1	6	2
31143	NA	1	7	2
31150	3	1	8	2
31151	3	1	2	1
31152	3	1	8	2

id	drinks_per_occasion	SMD030	SMQ040	SMD641	SMD650	SMD630	SMQ670
31139	NA	NA	NA	NA	NA	NA	NA
31143	NA	NA	NA	NA	NA	NA	NA
31150	NA	NA	NA	NA	NA	NA	NA
31151	NA	NA	NA	NA	NA	NA	NA
31152	NA	NA	NA	NA	NA	NA	NA

id	WHD010	WHD020	WHQ030	WHQ040	WHD050	WHD020	WHQ030	WHQ040

id	WHD050	WHQ070	WHD110	WHD120	WHD140	WHQ150

6 Detailed Analysis Results

7 Conclusions and Recommendations

Reference List

Beam, Andrew L., and Isaac S. Kohane. 2018. "Big Data and Machine Learning in Health Care." *JAMA* 319 (13): 1317. https://doi.org/10.1001/jama.2017.18391.

Bhatt, Kunal N., Andreas P. Kalogeropoulos, Sandra B. Dunbar, Javed Butler, and Vasiliki V. Georgiopoulou. 2016. "Depression in Heart Failure: Can PHQ-9 Help?" *International Journal of Cardiology* 221 (October): 246–50. https://doi.org/10.1016/j.ijcard.2016.07.057.

Buntinx, Frank, Jan De Lepeleire, Jan Heyrman, Benjamin Fischler, Dirk Vander Mijnsbrugge, and Marjan Van den Akker. 2004. "Diagnosing Depression: What's in a Name?" *European Journal of General Practice* 10 (4): 162–65. https://doi.org/10.3109/13814780409044305.

Daly, Michael, Angelina R. Sutin, and Eric Robinson. 2021. "Depression Reported by US Adults in 2017–2018 and March and April 2020." *Journal of Affective Disorders* 278 (January): 131–35. https://doi.org/10.1016/j.jad.2020.09.065.

Emam, Khaled El, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mark Lysyk. 2009. "Evaluating the Risk of Re-Identification of Patients from Hospital Prescription Records." *The Canadian Journal of Hospital Pharmacy* 62 (4). https://doi.org/10.4212/cjhp.v62i4.812.

Greenberg, Paul E., Andree-Anne Fournier, Tammy Sisitsky, Mark Simes, Richard Berman, Sarah H. Koenigsberg, and Ronald C. Kessler. 2021. "The Economic Burden of Adults with Major Depressive Disorder in the United States (2010 and 2018)." *PharmacoEconomics* 39 (6): 653–65. https://doi.org/10.1007/s40273-021-01019-4.