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.

The description of the numerical data is given in the table below.

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

The description of the categorical data is given in the table below.

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

5 Exploratory Data Analysis

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.

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.