Assignment 3

Vivian, 300525329 Saumya Sajwan, 300529034 Samanalie Perera, 300486075

2022-08-30

Background and Data

Datasets

Target Disease: Depression

NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Mental Health - Depression Screener datasets.

Predictors

Demographics NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Demographic Variables & Sample Weights datasets.

Sleep Disorders NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Sleep Disorders datasets.

Alcohol Use NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Alcohol Use datasets.

Smoking - Cigarette Use NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Smoking - Cigarette Use datasets.

Weight History NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Weight History datasets.

Sources

https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Questionnaire

https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics

Interest

We are interested in creating a machine learning model that finds and uses key features to identify individuals who are potentially suffering from depression.

Structure of the dataset

There are 29 variables in our dataset: 20 categorical variables and 9 numerical variables. Therefore, we will use all variables excluding id and target output as our predictors; each is a question the respondents need to answer to help us identify whether a person has depression.

Completeness of the dataset

- The dataset consists of data from 2005 to March of 2020. We were going to add in later data from 2004-1999, however the datasets for them were completely different. They had more questions and didn't match the 2005-2020 questions, so we decided to not add them in.
- For classifying whether or not an individual has depression, we calculate the score from each variable in the depression dataset. Therefore, we deleted all the NA values for more accurate output. This meant that our dataset went down from 43,928 to 26,627. 17, 301 values were taken out.
- For predictor datasets, we converted the "Refused" and "Don't know" categories to NA values for more accessible future data preprocessing.

Steps to integrate the datasets

We decided to use binary classification to predict our target disease, meaning 1 has depression and 0 is not. To get the target output column, here are the steps we did:

- Loaded the NHANES package on R. Only the datasets from 2005 to 2014 were part of the package. So for 2015 - March 2020, we had to read them using an R function (foreign::read.xport), which reads the data from a SAS XPORT file and returns it in a data frame.
- Adding more columns: filename, cycle, start year and end year, so the data from the newer datasets match the older ones.
- Added all the datasets together into one large dataset using rbind.
- To find out the total score each individual had, we added each of the columns together and added them to a new column called Score.
- We added another column for classifying whether or not an individual has depression. For that, we said that if the Score was 10 or higher, that person has depression. If not, they don't. We used this website for the score: https://bmcrheumatol.biomedcentral.com/articles/10.1186/s41927-021-00236-w#:~:text=Adults%20(%E2%89%A5%2018%20years)%20with,9%20(PHQ%2D9)

For generating the predictor variables, here is how we did it:

- First, using the same method we did for downloading the depression dataset. We decided to use the variables from Demographic Variables & Sample Weights, Sleep Disorders, Alcohol Use, Smoking - Cigarette Use and Weight History datasets.
- Each dataset has different variables every year. So for each dataset, we only use the variables that appear every year and then combine them as one dataset.
- Then we will look at the categorical variables. For example, the category "Refused" is 777 and "Don't know" is 999 in some datasets. We convert them to NA values, as the number doesn't have any meaning, which may affect our prediction accuracy in the future.
- Finally, we will combine all the datasets by their id numbers and add the target output column as our final dataset. It's to be used in later building the model.

Ethics, Privacy and Security

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.

Privacy concerns

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.

Exploratory Data Analysis

```
df = read.csv("project_data.csv")
#df.drop(c("id", "X")

dim(df)
```

```
## [1] 26627    30
```

There are 13471 rows and 12 features in our data. One of

Individual Contributions

Vivian, 300525329

- Background and Data: Datasets, Sources, Structure of the dataset, Completeness of the dataset, Steps to integrate the datasets.

Saumya Sajwan, 300529034

Samanalie Perera, 300486075