Caroline Sklaver
Capstone Project Proposal
06/04/2020

In this project I aim to build a machine learning model to predict depression in the US adult population. According to the CDC, about 1 in every 6 adults in the US will experience depression during their lifetime, which is about 16 million adults each year. While there are many factors that play into depression, the exact causes are unknown (1). Identifying the importance of specific indicators and building a machine learning model to predict depression may help diagnose or inform treatment plans for Major Depressive Disorder (MDD).

In this project I will be using the National Health and Nutrition Examination Survey (NHANES) data (2). This survey includes both interview and physical examinations. Interviews consist of demographic, socioeconomic, dietary, and health-related questions, while the examination includes medical, dental, and physiological measurements, as well as laboratory tests. Participants are a nationally representative sample of about 5,000 persons each year. Data is available (at varying degrees of completeness and uniformity) from 1999 to 2018.

After reviewing the existing literature, I will narrow down the variables I would like to include in my analysis. From there, I will explore the data with basic statistics and visuals. Next, dimensionality reduction may be necessary, for which I may use correlation plots, feature selection techniques (classification trees), and/or feature extraction (Principal Component Analysis). Then, I plan to build and evaluate multiple models. This will include Logistic Regression, SVM, Random Forest, KNN, Bagging, Boosting, MLP, and any other models I may find in my literature review that have been successful in similar projects. I will be using Python3 (both PyCharm and Jupyter notebook) for this project.

To judge the performance of these models, I will look at the classification sensitivities, specificities, precision, and F-1 Score. I will also visualize outcome accuracies using Receiver Operating Characteristic (ROC) curves. With this type of data it will be important to keep in mind the real-world health outcomes and meaning of each of these measures.

Though this exact analysis has not been done using NHANES data, many other publications are available for reference. In the literature, many associations are made between certain variables and depression using NHANES data (3,4,5). Additionally, researchers have used this data to predict both heart disease (6) and diabetes (7,8). And lastly, machine learning for medicine and public health is a growing field with many resources I may reference throughout this project.

Rough schedule for completing this project:

Project Plan:	Deadline:
- Have all variables chosen - All data imported & merged - Handled missing values	Week of June 15th

- EDA, visuals, statistics	Week of June 22nd
- Prepare data for modeling - Preliminary models	Week of July 2nd
- Preliminary Presentation	Due July 9th
- All models & hyperparameter tuning completed - Evaluated models	Week of August 3rd
- Mock Presentation - Much of journal submission completed	Due August 13th
- Final Project Presentation - Journal Submission completed	Due August 20th

References

1. https://www.cdc.gov/tobacco/campaign/tips/diseases/depression-anxiety.html
2. https://www.cdc.gov/nchs/nhanes/index.htm
3. https://www.sciencedirect.com/science/article/pii/S0091743511002775?casa_token=cRGji2oDdO4AAAAA:CI75fmmDOngBn-M6m_3nt91Rps0R20Km50_hLTxGkoNf9T9xhkUimjiA8Ez0YIXwy2v3Oi05DQ
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4309548/
5. https://journals.sagepub.com/doi/abs/10.1177/003335490512000112
6. https://github.com/ben-bay/NHANES-disease-predictor/blob/master/predicting-cardiac-disease.pdf
7. https://github.com/semerj/NHANES-diabetes/blob/master/report/report.pdf
8. http://worldcomp-proceedings.com/proc/p2013/DMI8087.pdf