# Heinz 95-845: Project Proposal

**Sarah Cho** jscho/jscho@andrew.cmu.edu
**Maggie Lu** yaol4/yaol4@andrew.cmu.edu
**Askari Shah** syedmehs/askari@cmu.edu
*Heinz College*
*Carnegie Mellon University*
*Pittsburgh, PA, United States*

## 1. Proposal Details

### 1.1 What is your proposed analysis? What are the likely outcomes?

There have been studies linking low socio-economic status(SES) with the prevalence of depression, see: Freeman et al. (2018) and Wirth et al. (2017). In the United States, we see the difference in food choices becoming more and more pronounced among those who have high disposable income and those who are living below the poverty line. Even though there have been numerous studies proving the link between nutrition and physiological health, the link between food and mental health has only been recently discussed (see Molendijk et al. (2018)).

For our project, we would like to examine how SES and nutrition affect one's mental health. We aim to explore the relationships and interactions between an individual's SES and his/her nutritional intake. Additionally, we would like to compare SES and nutritional factors on their effectiveness as predictors for depression. We predict that depression can be linked to both SES and nutrition, but we are not certain whether the combination of both can be a better predictor than either one.

### 1.2 Why is your proposed analysis important?

Our analysis would be able to provide a holistic view on how SES and nutrition affect one's mental health. Depression and mental health issues have been more and more pervasive in the U.S in recent years, with the Center for Disease Control and Prevention (CDC) reporting earlier this year that suicides are becoming more often in every age group (see Holly Hedegaard and Margaret Warner (2018)). We aim to discover new insights in the external factors that potentially correlate with mental health issues, and hopefully would add to the preventive measures against depression and other mental illness.

### 1.3 How will your analysis contribute to existing work? Provide references

Many studies explore the relationship between nutrition and depression in varying SES but have conducted limited predictive analysis. Our analysis will provide predictive capabilities to augment results from existing studies. For example, Leung and Kaplan evaluate the role of nutrition on perinatal depression and include SES as part of their social risk factor (see: Brenda M.Y. Leung N.D. (2009)). Additional studies linked depression and SES or nutrition and depression (see: Susan A. Everson (2005) and Jonathan E. Alpert M.D. (2009) ) but have not incorporated all three factors for predictive modeling.

### 1.4 Describe the data. If applicable, please also define Y outcome(s), U treatment, V covariates, and W population.

We collect demographic, nutritional and mental health data from NHANES (National Health and Nutrition Examination Survey) from year 1999 - 2016. The outcomes, Y, are a list of responses detailing depression/depression-like symptoms in data files that start with DLQ, which we will later simplify to a classifier. The treatment, U, will be nutrition data and SES data detailed in data files DBQ, DTQ, DSQ, RXQ (for dietary information) and DMQ, CBQ, FSQ, HIQ, HOQ, INQ (for SES information). The covariates V will include all other demographic information. NHANES is designed to be an approximate representation of the U.S population, therefore the population characteristics are in accordance with the U.S demographics, with occasional over-sampling of certain ethnic groups/races, which we will make sure to note in our final findings.

### 1.5 What evaluation measures are appropriate for the analysis? Which measures will you use?

Appropriate measures for analysis include confusion matrix (accuracy, precision, recall), ROC curve, Precision/Recall curves and statistical analyses techniques (Confidence intervals of errors, hypothesis tests of models). We intend to use confusion matrix, ROC curves and Precision/Recall curves. Depending on our data and class distributions along with our proposed use case, we will pick the most suitable/relevant measures from the listed appropriate measures.

### 1.6 What study design, pre-processing, and machine learning methods do you intend to use? Justify that the analysis is of appropriate size for a course project.

First, the data needs to be cleaned and re-organized so that only relevant factors are included. We will need to join tables and impute missing values accordingly. As of now, we do not intend to work with longitudinal data or impose time series on the data collected spanning from year 1999 to 2016. Instead, we would treat each entry as a separate individual in our final data table. After the data is cleaned, we want to separate our data into a training set and a test set and use cross validation to train our models. We intend to use the following machine learning methodologies: Decision Tree, Logistic Regression, Bagging, Boosting, K-Nearest Neighbor analysis, as well as Neural Networks. We want to compare and evaluate the effectiveness of each method and come up with an appropriate recommendation on which one to deploy for future analytical work on this topic.

### 1.7 What are possible limitations of the study?

One possible limitation comes from participants' reporting of their mental status. One example of a question to screen for depression is: Thinking about the last time you felt worried, nervous or anxious, how would you describe the level of these feelings?. The response is very subjective and this may cause biases and inadequate labeling of a participant as depressed.

Another limitation also arises from NHANES's survey data. The inputs to the analysis is dependent on the NHANES questionnaire. Therefore, we are limited to the information that is provided by NHANES where there may be other underlying SES factors. Additionally, the most recent set of survey results are from 2016. As there is a time delay of 2 years, the analysis may not be reflective of current ground truth.

### 1.8 Who will your analytic pipeline? In one or two sentences, describe an example of its use.

One potential user of the analytic pipeline is the Department of Health and Human Services (HHS). HHS can use the analytic pipeline to create relevant policies on providing additional nutrition supplements to clinics in lower SES brackets.

## References

Ph.D. Brenda M.Y. Leung N.D., M.Sc. Bonnie J. Kaplan. Perinatal depression: Prevalence, risks, and the nutrition linka review of the literature. 2009.

Aislinne Freeman, Stefanos Tyrovolas, Ai Koyanagi, Somnath Chatterji, Matilde Leonardi, Jose Luis Ayuso-Mateos, Beata Tobiasz-Adamczyk, Seppo Koskinen, Christine Rummel-Kluge, and Josep Maria Haro. The role of socio-economic status in depression: Results from the courage (aging survey in europe). *BMC Public Health*, 2018.

M.A. Holly Hedegaard, M.D.and Sally C. Curtin and Ph.D. Margaret Warner. Suicide rates in the united states continue to increase. 2018.

Ph.D. Maurizio Fava M.D. Jonathan E. Alpert M.D. Nutrition and depression: The role of folate. 2009.

Marc Molendijk, Patricio Molero, Felipe Ortuo Snchez-Pedreo, Willem Van der Does, and Miguel Angel Martnez-Gonzlez. Diet quality and depression risk: A systematic review and dose-response meta-analysis of prospective studies. *Journal of Affective Disorders*, 226:346–354, 2018.

John W. Lynch George A. Kaplan Susan A. Everson, Siobhan C. Maty. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. 2005.

Michael D. Wirth, Nitin Shivappa, James B. Burch, Thomas G. Hurley, and James R. Hbert. The dietary inflammatory index, shiftwork, and depression: Results from nhanes. *Health Psychology*, 2017.