Heinz 95-845: Project Proposal

Atsumi Kainosho **Haley Townsend**
Carnegie Mellon University, Heinz College

1. Project Background, Importance and Users

Body Mass Index (BMI) is a common and informative metric used by the Centers for Disease Control and Prevention (CDC). This indicator is often used to flag individuals who could have an unhealthy level of fat. Being overweight can lead to many health consequences and/or exacerbate existing conditions, such as diabetes. Therefore, reducing the number of overweight and obese persons in the country constitutes a top priority for the CDC. Specifically, the CDC states: *CDCs Obesity efforts focus on policy and environmental strategies to make healthy eating and active living accessible and affordable for everyone.*

The purpose of this project is to conduct a novel analysis using demographic data to classify BMI for the CDC. Specifically, this project is "novel" in its application of sophisticated machine learning algorithms to survey and observational policy data. Often, the most innovative machine learning research is conducted in the private sphere where companies are motivated to increase profits and funding is plentiful. Conversely, public sector agencies are often limited in technology, funding and the ability to perform complex data analyses. Therefore, we hope this project will provide important structure and insights to the CDC, the nation's health protection agency.

For example, the CDC could use this classifier to better target at-risk youth with additional health assistance (e.g. more food stamps). This project aims to not only provide insights and recommendations from the results, but also showcase the entire machine learning pipeline for this specific context. While we will be performing this analysis on data from the CDC, we intend for this pipeline to serve other health and nutrition agencies as well. Officials in the health care field can detect potentially unhealthy children based only on their socio-demographic backgrounds, which will help inform health care policy efforts.

2. Objectives and Analysis

The main objective of this project is to execute the machine learning pipeline from start to finish within the context of U.S. health and nutrition. In our analysis, we will build a classifier that will predict a child's BMI category (healthy or overweight) based only on their demographic information. More specifically, this project aims to:

- Work with SAS databases and .xpt files in R (which are common database/set types in the public sector)
- Clean, subset and manipulate the data for this use case
- Explore, summarize and visualize the data to get a sense of general trends
- Apply various machine learning techniques to classify a child's BMI
- Evaluate the performance of various machine learning algorithms
- Make policy recommendations to the CDC based on the results
- Suggest other possible use cases for the completed pipeline and insights

3. Existing Work

There is plentiful research related to BMI and socio-demographic factors for children. For example, J.Heerman et al. observed adverse family experiences and obesity in children in the United States (William J. Heerman (2016)). Ohri-Vchaspa et al. argues that a child's weight status is closely related to food and physical activity (Ohri-Vachaspati (2013)). However, previous studies, including the two just mentioned, have only used basic statistical methods, such as linear regression or descriptive statistics. When focusing on studies using the NHANES dataset, this trend remains. For instance, Skinner et al. published "Prevalence of Obesity and Severe Obesity in US Children, 1999 to 2016" that reports the prevalence estimates of overweight and obesity (class I, class II, and class III) by 2-year NHANES cycles and compares cycles by using adjusted Wald tests and linear trends by using ordinary least squares regression (Asheley Cockrell Skinner (2018)). In other words, these studies simply try to determine contributing factors of obesity, but they do not establish a model to predict whether a child has a higher BMI based on their socio-demographic factors. They did not use machine learning techniques in any form.

4. Data and Design

This project uses data from the 2012 National Health and Nutrition Examination Survey (NHANES) National Youth Fitness Survey (NNYFS). This one-time survey collected demographic, dietary, health and examination data on 1,576 children ages 3 to 15 years old across the United States. We define the following:

- Y: binary, whether the child is categorized as overweight or not. Right now our outcome variable has four classes, but we will condense these into two.
- U: number of days in the week the child is physically active at least 60 minutes. Please note: this is not a true treatment or intervention. Since this data is observational and comes from a self-reported survey, we can (at best) make associative statements between U and Y.
- V: major covariates include gender, age, race, country of birth, language, annual family income, household size, parent/guardian's marital status and a few others.
- W: 3 to 15 year-olds in the United States considered part of the non-institutionalized, resident population. This survey is considered "nationally representative."

Since some of the demographic columns have missing data, we will have to decide whether to impute the missing values, based on the nature of the missingness. We hope to build associative statements between BMI category and physical activity. Additionally, we want to build an interpretable model that can help show the relationship between the covariates and the outcome. To get there, we will apply a logistic regression, decision tree, random forest and potentially a neural network. This project is of appropriate size for this course, since we must work with a new database structure (from SAS), deal with missing data, apply various classifiers, evaluate performance and make interpretable statements. Having 1,571 rows of data allows us to accomplish these things without crashing our computers (since the dataset is smaller).

5. Evaluation Measures, Likely Outcomes and Limitations

To evaluate the performance of the machine learning algorithms, we will overlay ROC curves and compare AUCs. We expect the logistic regression or random forest will perform best on the test data. The decision tree may be too simple while the neural network is bound to be too complex and uninterpretable. If performance is similar across the logistic regression and random forest, we will likely select the logistic regression for our recommendation since it is more interpretable. In addition to evaluating the models, we will determine which variables are most important. We can find the variable importance of the features as MeanDecreaseGini in random forest. This knowledge can contribute to previous studies, since it provides new insights on the relationship between BMI and other factors that previous studies have tried to reveal.

The biggest possible limitation of this study is the small number of samples in the data set. The total number of samples without missing data is 1,571. With Ensemble methods, we will try to train an algorithm, but it can cause overfitting on this small amount of training data. We should check the appropriate balance between variance and bias using regularization. Also, we cannot argue about causalities of high BMI based on only our data set since it is not feasible to randomize samples. Since our objective to build a classifier for BMI, we do not need to consider causalities, but if we can do it, it would be more helpful for the government to tackle obesity problems.

References

Joseph A. Skelton Eliana M. Perrin Sarah C. Armstrong Asheley Cockrell Skinner, Sophie N. Ravanbakht. Prevalence of obesity and severe obesity in us children, 19992016. *Pediatrics*, 142(3):No page listed, 2018.

Delia D Tulloch D Yedidia MJ Ohri-Vachaspati, Lloyd K. A closer examination of the relationship between children's weight status and the food and physical activity environment. *Preventitive Medicine*, 57(3):162–167, 2013.

Shari L. Barkin Melissa McPheeters William J. Heerman, Shanthi Krishnaswami. Adverse family experiences during childhood and adolescent obesity. *Pediatric Obesity*, 24(3): 696–702, 2016.

No author listed. Healthy Weight: Body Mass Index (BMI). Centers for Disease Control and Prevention, U.S. Department of Health and Human Services. *See https://www.cdc.gov/healthyweight/assessing/bmi/index.html for full source.*

Additional information used:
https://www.cdc.gov/obesity/index.html
https://www.cdc.gov/nchs/nnyfs/data_faq.htm