Lifestyle Factors Related to Childhood Obesity

Taig Singh
MPCS 53120 - Applied Data Analysis
Final Report
June 3, 2021

Abstract

The goal of this project is to explore the different lifestyle factors that are related to obesity among high schoolers. Data was investigated to identify factors that may have positive or negative effects on adolescents' body mass index (BMI), and concurrent machine learning models were designed to create a predictive function for obesity. The data was extracted from the 2019 Youth Risk Behavior Surveillance System, which consists of behavioral data from over thirteen thousand high school students around the country along with their heights and weights (*YRBSS Overview* 2020). After the data was reformatted and cleaned, the initial step of the project involved exploratory data analysis. Conclusions from the initial exploration revealed that eating breakfast and increased physical activity positively influenced BMI, while smoking and prolonged periods of watching television influenced BMI in a negative manner. Subsequently, regression and classification models were built to predict BMI and obesity based on the factors studied. The most accurate model built during this stage was a k-nearest neighbors classification model with k=17. This model achieved a 5-fold cross-validation score of about 0.855.

Introduction

Childhood obesity is a serious disease that affects children and adolescents and is rapidly becoming more common in the United States. From a very young age, it pushes individuals down a road that leads to type 2 diabetes, high blood pressure, high cholesterol, and some cancers, not to mention serious mental health problems as well. All of these factors have a detrimental effect on health care costs and productivity (*About overweight and obesity* 2021). It is becoming increasingly vital to identify lifestyle factors that are related to childhood obesity and come up with strategies that adolescents can use to avoid this disorder.

The epidemic of childhood obesity can be tied to both biological and behavioral factors. Since genetics is a risk factor that can not be controlled, the obesity epidemic can be managed by employing useful data like behavioral patterns of individual children along with their respective BMI percentiles. One such data set was utilized in this project, taken from the CDC's Youth Risk Behavior Surveillance System (YRBSS). The purpose of the YRBSS is to observe the most prevalent factors that lead to death, disability, and social problems among children in the United States. These factors, measured by the system, include behaviors that lead to unintentional injuries and violence, sexual behaviors, alcohol and other drug use, tobacco use, dietary behaviors, and physical activity. Alongside these factors, the YRBSS measures the prevalence of obesity along with other health problems among individuals (*YRBSS Overview* 2020). The data set being used for this project is the national school-based survey from the YRBSS in 2019 which was given to high school students. This data was used to identify lifestyle factors that may cause childhood obesity and build a predictive model for this disease.

Related Work

There is a substantial body of research that focuses on factors that contribute to childhood obesity, however, a large portion of it merely considers genetic and weight data collected between an individual's birth and their second birthday. For this case study, only research that considers lifestyle factors later in a child's life to determine what causes childhood obesity will be reviewed.

One of the papers in this field is "Predictor factors for childhood obesity in a Spanish case-control study" by María Ochoa, María Moreno-Aliaga, Miguel Martínez-González, et al. The goal of this study was to determine the factors that put individuals at risk for childhood obesity. The factors examined were related to lifestyle along with dietary pattern, physical activity, family history of obesity, breastfeeding, sedentary behavior, and birth weight. These predictive factors were identified by conditional logistic regression on anthropometric data and data collected from personal interviews about lifestyle choices. The subjects included 185 obese individuals ranging from 6 to 18 years of age, and the control cases were matched by sex and age. Variables unrelated to diet, for example, family obesity and leisure time, were initially studied in a univariate model. Variables that initially approached statistical significance ($p < 0.25$) were included in a multivariate model. Subsequently, variables that were statistically significant ($p < 0.05$) were included in the final model. The results gathered from this study indicated that family history of obesity, watching television, consumption of sugar-sweetened beverages, and leisure-time physical activity are the most significant predictors of childhood obesity, while other factors such as sleep and birth weight are not significant (Ochoa et al., 2007).

Another paper reviewed in the present field is "The International Study of Childhood Obesity, Lifestyle and the Environment (ISCOLE): design and methods" by Peter Katzmarzyk, Tiago Barreira, Stephanie Broyles, et al. The aim of this multi-national study was to identify the relationship between lifestyle factors and childhood obesity. This study also examined the influence of higher-order characteristics on these relationships, including behavioral settings along with the physical, social, and policy environments in the countries. The subjects included 6000 children of 10 years of age from 12 countries around the five major regions of the world. To generate the sample size necessary, a regression model was created to predict BMI from physical activity and self-reported caloric consumption among 10-year olds in the 2005/2006 U.S. National Health and Nutrition Examination Survey. Data was collected on an individual basis, and measurements were collected regarding anthropometry and bioelectrical impedance, demographics and family health history, diet and lifestyle information, accelerometry, assessment of biological maturity, neighborhood and home environment, and school environment. Data on individual attributes were summarized separately by sex as counts and percentages for categorical variables, and means and standard deviations for continuous variables. General linear and nonlinear models, including covariate-adjusted models, were utilized to investigate the relationship between obesity and these variables. This study found that the association of

moderate-to-vigorous physical activity and vigorous physical activity with obesity were significant in all 12 study sites, while the association between sedentary time and obesity was significant in only five of the 12 sites. The most significant behavioral risk factors of obesity found in this study were particularly low moderate-to-vigorous physical activity, short sleep duration, and high TV viewing (Katzmarzyk et al., 2013).

Materials and Methods

The national YRBSS data from 2019 was downloaded from the CDC's website. A data guide describing the format of the data and the questions corresponding to the question numbers was also downloaded here. This data was then reformatted and converted into two CSV files. The first, primary data set consisted of the raw responses to the survey converted from alphabetical multiple choice answers to integers, most often where increasing numbers represented increasing intensity of the answer to the survey question. The second data set was derived from the first and consisted of the dichotomous variables associated with each variable in the first data set. For each survey question, all responses that matched "responses of interest" were given the value of "1" and all other entries were given the value of "2." This data set also included extra, supplemental dichotomous variables that were calculated based on responses to multiple questions. Almost immediately, it was noticed that there was a great number of rows with missing values in this second data set. After all the rows with missing data were removed, the data set only consisted of just over a hundred responses. This is why the only column from this data set that was utilized was the "qnobese" variable, which was set to "1" if an individual's BMI landed at or above the 85th percentile for their age and sex, and was set to "2" otherwise. This row was concatenated with the first data set which was read in as a pandas data frame. Dummy variables were generated for the race and ethnicity variable, which was the only survey question that allowed multiple responses on the survey. Unnecessary columns were dropped, along with survey questions 4, 5, 6, 7, 67, and 68. Questions 4 and 5 asked the individual about their race and ethnicity, which was already represented by dummy variables. Questions 6 and 7 asked for height and weight which is what the "BMIPCT," or the BMI percentile variable used to calculate. Finally, questions 67 and 68 had to do with the individual's own perception of their weight, and this was chosen not to be included in the predictive model. Finally, all rows with missing data were dropped.

Once the data was reformatted and cleaned, it was split into predictive parameters and the outcome. The responses of the models consisted of the "BMIPCT" variable for regression models and the dichotomous "qnobese" variable for classification models. A correlation heatmap of the predictive variables was generated using the seaborn Python library to identify any close relationships within the data. Seaborn pair plots of individual question data and a linear regression model fit with the "BMIPCT" variable were then generated for every survey question in order to recognize any close significantly positive or negative relationships with an

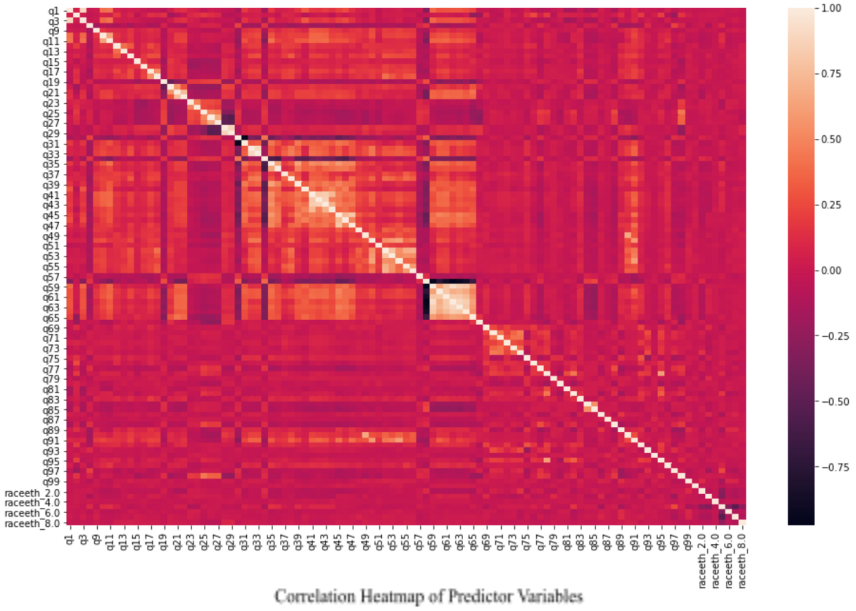
individual's BMI. Observed significant relationships were then displayed and analyzed separately.

After the brief section on exploratory data analysis, regression and classification models were built to predict BMI percentile and obesity. Forward stepwise selection using linear regression was performed on the predictors to reduce the number of predictors to the model. The R^2 and adjusted R^2 were used as scoring metrics through this algorithm. A simple linear regression model was then fit on this subset of predictors and the "BMIPCT" column. This model was evaluated with 5-fold cross-validation using R^2 as the scoring metric. Lasso and Ridge regression models were also trained and evaluated using the data. Cross-validation scores for these models were very low. Following this, the response variable was switched to the dichotomous "qnobese" variable and classification models were trained. Forward stepwise selection using logistic regression was then performed on all of the predictors, and a logistic regression model was trained. This model was evaluated using 5-fold cross-validation using classification accuracy as the scoring metric. A neural network was then fit on the data and evaluated similarly. The number of layers and layer size was optimized by trial and error which is not represented in the code. Finally, a k-nearest neighbors classification model was fit on the data. The number of neighbors was tested from one through fifty, and the model that returned the maximum cross-validation score was used.

Results

From the research and analysis done in the present project, lifestyle factors related to childhood obesity were identified and a successful obesity classification model was built.

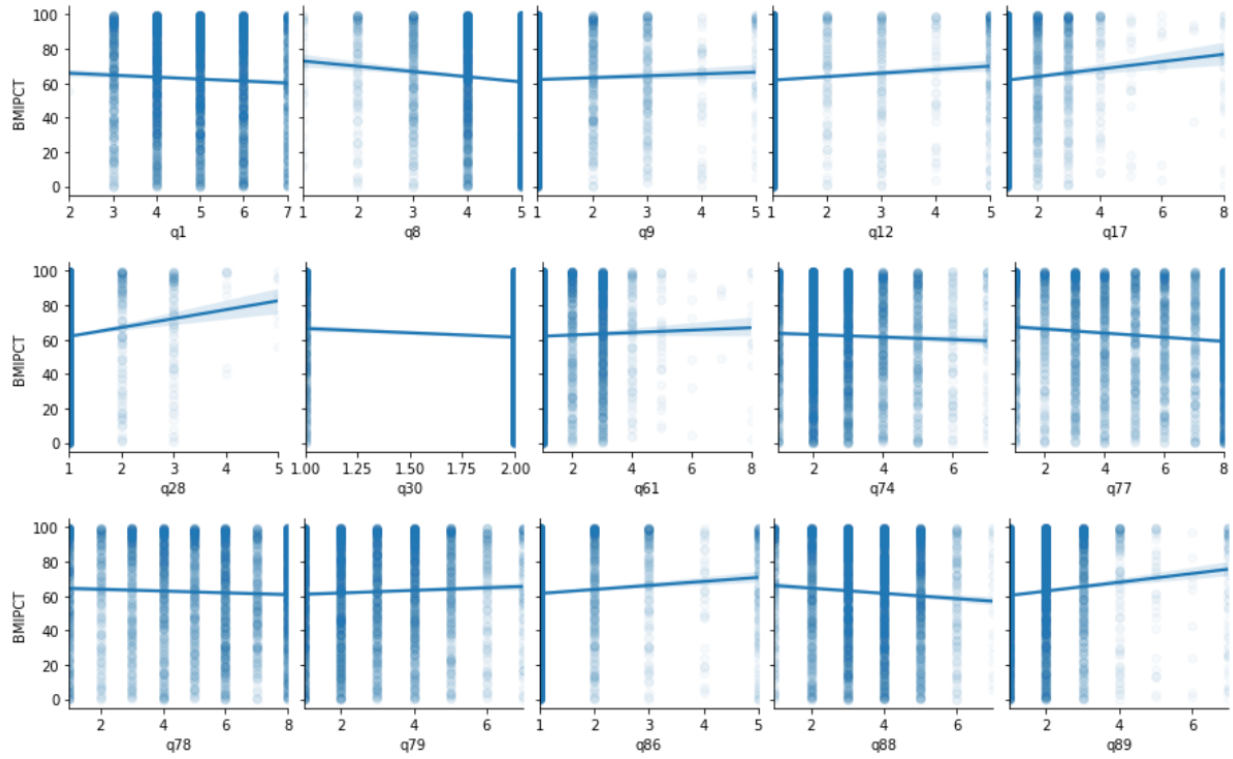
To the right is the correlation heatmap that was generated for the predictive variables. The main feature to pay attention to here is the highly correlated rectangles with corners on the central diagonal. These rectangles are so bright and noticeable because the survey is structured such that questions dealing with similar topics are grouped together. For example, questions 58 through 66 all deal with sexual behaviors and sexuality. The high correlation between these variables is represented by the bright



Correlation Heatmap of Predictor Variables

rectangle along the diagonal line slightly to the right of the center of the heatmap. One issue that this could introduce is multicollinearity, as some predictive variables are highly correlated with each other. However, this was potentially resolved by selecting variables one at a time based on R^2 through the forward stepwise selection algorithm.

The seaborn pair plots of individual question data and a linear regression model fit with the BMI percentile variable led to conclusions regarding lifestyle factors that were correlated with obesity. The most significant plots are displayed below.



Pair Plots of Individual Question Data and a Linear Regression Model Fit with BMI Percentile

Survey questions:

- Q1 - Age ranging from *12 years old or younger* to *18 years old or older*. Negatively correlated with BMI.
- Q8 - Commonality of wearing a seatbelt in a car ranging from *Never* to *Always*. Negatively correlated with BMI.
- Q9 - Number of times riding in a car driven by someone who had been drinking alcohol in the last 30 days ranging from *0 times* to *6 or more times*. Positively correlated with BMI.
- Q12 - Number of days carrying a weapon in the last 30 days ranging from *0 days* to *6 or more days*. Positively correlated with BMI.
- Q17 - Number of times in a physical fight in the last 12 months ranging from *0 times* to *12 or more times*. Positively correlated with BMI.

- Q17 - Number of times attempted suicide in the last 12 months ranging from *0 times* to *6 or more times*. Positively correlated with BMI.
- Q30 - Tried cigarette smoking ranging from *Yes* to *No*. Negatively correlated with BMI.
- Q61 - Number of people had sexual intercourse with in the last 3 months ranging from *I have never had sexual intercourse* to *6 or more people*. Positively correlated with BMI.
- Q74 - Number of times eaten vegetables other than green salad, potatoes, or carrots in the last 7 days ranging from *I did not eat other vegetables during the past 7 days* to *4 or more times per day*. Negatively correlated with BMI.
- Q77 - Number of days eaten breakfast in the last 7 days ranging from *0 days* to *7 days*. Negatively correlated with BMI.
- Q78 - Number of days physically active for 60 minutes or more ranging from *0 days* to *7 days*. Negatively correlated with BMI.
- Q79 - Number of hours watching TV on a school day ranging from *I do not watch TV on an average school day* to *5 or more hours per day.* Positively correlated with BMI.
- Q86 - Length of time since the last appointment for dental work ranging from *During the past 12 months* to *Not sure*. Positively correlated with BMI.
- Q88 - Number of hours of sleep on a school day ranging from *4 or less hours* to *10 or more hours.*
- Q89 - School grades during the past 12 months ranging from *Mostly A's* to *Not sure*. Positively correlated with BMI.

The predictive regression models predicting BMI percentile were very inaccurate. The 5-fold cross-validation scores using R^2 as a scoring metric for these models ranged from 0.026 to 0.049. However, the classification models predicting obesity were far more accurate. The 5-fold cross-validation scores using classification accuracy as a scoring metric for these models hovered around 0.85 with the most accurate model being the k-nearest neighbors model with k=17. This model achieved a cross-validation score of 0.855, making it a far more accurate predictor of childhood obesity.

Discussion

Many of the lifestyle factors found to be related to childhood obesity in this project were also discovered by other research in this field. The aforementioned ISCOLE study also found low physical activity, high amounts of TV viewing, and short sleep duration to be highly correlated with increasing BMI among children (Katzmarzyk et al., 2015). “Predictor factors for childhood obesity in a Spanish case-control study” also found watching TV, an unhealthy diet, and low physical activity to be predictors of childhood obesity. However, this study found that sleep duration is not significant (Ochoa et al., 2007). Other studies have found that low sleep duration is highly correlated with obesity among adults and children, so the research in this project is still corroborated.

Although the regression models generated in this study were inaccurate and unusable, the final KNN classifier is substantially accurate and helpful. It can be implemented to help predict obesity among children and build strategies to avoid this disease and its corollaries.

Future work in this field may include collecting even larger data sets with less missing data. Given the number of predictors in the data set, the low amount of data, especially after removing observations with missing data, affected the performance of the predictive models and the identification of related lifestyle factors. Secondly, building a more successful regression model to successfully predict the continuous BMI variable is important. The regression models trained in this project were unsuccessful in accurately predicting an individual's BMI percentile. Finally, making this information more widely available and the classification models easily accessible is imperative to the impact of this research.

References

Centers for Disease Control and Prevention. (2020, August 20). YRBSS Overview. Centers for Disease Control and Prevention. https://www.cdc.gov/healthyyouth/data/yrbs/overview.htm.

Centers for Disease Control and Prevention. (2021, April 7). About overweight and obesity. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/about-obesity/index.html.

Katzmarzyk, P. T., Barreira, T. V., Broyles, S. T., Champagne, C. M., Chaput, J.-P., Fogelholm, M., Hu, G., Johnson, W. D., Kuriyan, R., Kurpad, A., Lambert, E. V., Maher, C., Maia, J., Matsudo, V., Olds, T., Onywera, V., Sarmiento, O. L., Standage, M., Tremblay, M. S., … Church, T. S. (2013). The International Study of childhood obesity, lifestyle and the environment (ISCOLE): Design and methods. BMC Public Health, 13(1). https://doi.org/10.1186/1471-2458-13-900

Katzmarzyk, P. T., Barreira, T. V., Broyles, S. T., Champagne, C. M., Chaput, J.-P., Fogelholm, M., Hu, G., Johnson, W. D., Kuriyan, R., Kurpad, A., Lambert, E. V., Maher, C., Maia, J., Matsudo, V., Olds, T., Onywera, V., Sarmiento, O. L., Standage, M., Tremblay, M. S., … Church, T. S. (2015). Relationship between lifestyle behaviors and obesity in children ages 9-11: Results from a 12-country study. Obesity, 23(8), 1696–1702. https://doi.org/10.1002/oby.21152

Mayo Foundation for Medical Education and Research. (2020, December 5). Childhood obesity. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/childhood-obesity/symptoms-causes/syc-20354827.

Ochoa, M. C., Moreno-Aliaga, M. J., Martínez-González, M. A., Martínez, J. A., & Marti, A. (2007). Predictor factors for childhood obesity in a Spanish case-control study. Nutrition, 23(5), 379–384. https://doi.org/10.1016/j.nut.2007.02.004