8-2022

Effects of Macronutrients Intake and Physical Activity on Childhood Obesity of Hispanic Children

Prosanta Barai
The University of Texas Rio Grande Valley

EFFECTS OF MACRONUTRIENTS INTAKE AND PHYSICAL ACTIVITY ON CHILDHOOD OBESITY OF HISPANIC CHILDREN

A Thesis

by

PROSANTA BARAI

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Applied Statistics and Data Science

The University of Texas Rio Grande Valley

August 2022

EFFECTS OF MACRONUTRIENTS INTAKE AND PHYSICAL ACTIVITY ON CHILDHOOD OBESITY OF HISPANIC CHILDREN

A Thesis
by
PROSANTA BARAI

COMMITTEE MEMBERS

Dr. Xiaouhi Wang
Chair of Committee

Dr. Shantanu Chakraborty
Committee Member

Dr. Hansapani Rodrigo
Committee Member

Dr. Michael Machiorlatti
Committee Member

August 2022

ABSTRACT

Barai, Prosanta, Effects of Macronutrients Intake and Physical Activity on Childhood Obesity of Hispanic Children. Master of Science (MS), August, 2022, 88 pp., 22 tables, 30 figures, references, 100 titles.

Obesity has become more ubiquitous during the past few decades, and still, its prevalence is increasing. Although numerous studies are available in the literature contributing to diabetes and obesity prevention among Hispanic population, there is little to no study that has focused on early childhood, specifically the 4-6yo Hispanic population. Early prevention of obesity is the key for children health as well as their adulthood health. Our study aims to determine the causal path for this particular population's obesity using most recent three cycles (2014-2018) of NHANES' pre COVID pandemic data on physical activities, macronutrients intakes and socioeconomic factors. We also build a predictive machine-learning model to predict childhood obesity. Our analysis discovers a significant path from increased physical activity to decreased BMI. The Hispanic population has a significant positive causal path to BMI. Macronutrients, carbohydrate, fat, and sugar intakes show a positive path to BMI, while fiber intake shows a negative path to BMI. Interestingly, unlike other race/ ethnicity groups, Hispanic males and females have a significant positive path from carbohydrate intake to BMI. Among the machine-learning models we constructed, random forests model yields highest accuracy of eighty six percent, and the second highest accuracy of eight four percent from support vector machine algorithm.

Key Words: Obesity, Hispanic, Children, NHANES, Machine Learning, Path Analysis.

DEDICATION

I am grateful to my parents and brothers; without them, I would not have been the person I am today. Special thanks to my three elder brothers. I can't put it in words how much effort they put into me to guide in my academic life since childhood. Thank you, mom and dad, for raising me and always supporting my dream. I am proud to dedicate this work to my family.

ACKNOWLEDGMENTS

First, I thank Almighty Lord Krishna, who provided me with everything. Secondly, I want to thank Prof. Xiaouhi Wang, who has been a fantastic mentor to me during my time at UTRGV. She is simply outstanding, and her encouragement and support kept me going while I worked on this thesis. She is not only a brilliant individual with whom I am fortunate to have the opportunity to work but also a good human being. Still today, I can remember the first meeting we had after I arrived at UTRGV as an international student. I was going through a cultural transformation and jet-lag, but Dr. Wang guided me in every step to ensure my success while keeping my physical and mental health in mind. I can proudly say that working with Dr. Wang made my graduate journey smooth and made me an efficient and better person.

I'd also like to thank my honorable thesis committee members, Prof. Shantanu Chakraborty, Prof. Hansapani Rodrigo, and Prof. Michael Machiorlatti. Their guidance and support helped me write an impactful thesis. I'd also like to take this opportunity to thank my supervisor, Prof. George Yanev, who has been a fantastic mentor to me throughout my time at UTRGV. I am also grateful to prof. Tamer Oraby for his mentoring inside and outside the academic field. He has been one of the best instructors I had in my life. Thank you to Prof. Timothy Huber for your kindness and support. I was very satisfied whenever I contacted him with any necessity. I would also like to thank a very kind and polite person, Ms. Elda. She was always there for me with a beautiful smile on her face whenever I needed any assistance.

I am grateful to have Md. Salman Rahman as my friend and flatmate. He always encouraged me regarding academic difficulties. He was also there for me when I needed him as a friend. Without his support, this journey would have been much harder.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Overweight and obesity are defined as abnormal or excessive fat accumulation. A straight-forward indicator of that is body mass index (BMI), a simple weight-for-height index often used to describe overweight and obesity in adults. It is calculated as a person's weight in kilograms divided by the square of his height in meters (kg/m^2). According to WHO, adults with BMI greater than or equal to 25 are classified as overweight and obese if greater than or equal to 30. However, in children under 5, overweight is defined as weight-for-height greater than two standard deviations above WHO Child Growth Standards median and obese if it is greater than or equal to three standard deviations [1].

Obesity has become more ubiquitous during the past few decades and still, the prevalence of it is increasing. It is in every population in the world and all regions, including rural parts of low- and middle-income countries [2]. A 1997 WHO expert consultancy report warned of an escalating epidemic of obesity that would put the populations of most countries at risk of developing non-communicable diseases (NCDs). Since then the prevalence statistics on obesity have escalated rapidly in almost all countries, and these country-specific trends are now coalescing to create a true pandemic. In 2016, more than 1.9 billion (39%) adults aged 18 years and older were overweight globally. Over 650 million (13%) of those adults were obese, which caused the global obesity percentage to be tripled in 2016 than that of 1975 [3]. As far as the younger population concern, the prevalence of overweight and obesity among children and adolescents aged 5-19 has risen dramatically from just 4% in 1975 to just over 18% (340 million) in 2016 [4]. Besides, in 1975 less than 1% younger population (5-19 yo) were obese, while more than 124 million children and adolescents (6% of girls and 8% of boys) were obese in 2016 [5].

In the USA, regardless of age, the severity of obesity is no different from the global trend. Nationally, the prevalence of obesity has increased at an alarming rate over the past three decades. In 1988, among the population of age 20 and over, 23% were diagnosed obese (2.8% were severely obese). However, in 2017-2018 that prevalence roughly doubled and became 43% (9% was severely obese) [6]. Historically, obesity primarily affected adults, but childhood obesity has grown significantly in recent decades. From the mid-1970s to the mid-2015s, obesity roughly doubled among U.S. children ages 2 to 5 and approximately increased by more than four times (4% to 18%) among young people aged 6-11yo [7].

This massive prevalence of this chronic disease comes with massive costs. It not only affects our organ system but increases the risk of premature death [8]. Numerous studies found obesity has been linked to an increased risk of chronic disease, disability, death, decreased quality of life, and productivity [9, 10]. According to WHO overweight, and obesity are linked to more deaths worldwide than underweight [11]. Alone in the U.S., death caused by poor diet and physical activity increased by 33% in the past decade, and death attributed to other causes decreased [12]. Children are also greatly affected by it. Studies found that overweight children are more likely to develop heart disease and cause stroke before age 30 [13]. Health issues and obesity can cause high social, medical, and phycological costs [14, 15]. Passive costs caused by obesity (e.g., preventive, diagnostic) responsible for almost 7% of medical expenses in the U.S., and government health service programs cover half of it [16].

No way we can ignore the alarming surge of obesity in the U.S. younger population. Because as this generation of the population gets older, the U.S. healthcare system could meet an unexpected increase in the physical complications associated with diabetes, such as amputations, blindness, kidney failure, heart attacks [17, 18].

The rest of the chapters are organized as below. In Chapter II the previous research on obesity is presented. Also, some informative thoughtful public health expert opinion is shown for overall understanding of obesity. In chapter III we described the motivation and objectives that guided us to do the work. Chapter IV includes the description and structure of the data, as well as

data cleaning, preprocessing, and how we handled it. Chapter V describes the necessary statistical methods that we used in this study. Additionally, details of our study's findings are presented in the chapter VI. Chapter VII presents discussion and conclusion on findings. Chapter VIII is composed of limitations we had, and possible future works to address and improve those limitations of our study. Finally, the supporting supplementary materials were added in chapter A.

CHAPTER II

LITERATURE REVIEW

2.1 Public Health Expert's Opinion

Dietary patterns have a profound influence on human health, which has led the U.S. govern-ment to revise and reissue the Dietary Guidelines for Americans every five years [19, 20]. These guidelines emphasize a diverse and balanced diet primarily based on the consumption of whole grains (at least half of total grain intake), fresh fruits and vegetables (F&V), followed by the consumption of nutritious dairy, oils (mostly from oils in vegetable and foods) and lean animal protein foods. Moreover, the guidelines suggest adequate physical activity with moderate intake of energy, with the hope of helping to alleviate the severe obesity epidemic in the U.S. These guidelines specifically recommend limiting the consumption of sodium, refined sugar, saturated fat, trans-fatty acids, and avoiding overconsumption of total energy.

A significant growth spurt happens during puberty, usually between 8 to 13 years of age in girls and 10 to 15 years in boys. Puberty lasts about 2 to 5 years. By the time girls reach age 15 and boys reach age 16 or 17, the growth of puberty has ended for most, and they will have reached physical maturity. The following recommendations are critical to kids' overall health and wellness:

- **Enough rest:** Sleep patterns vary by age and the individual child. But most kids need an average of 10 to 12 hours of sleep per night. Sleep gives growing bodies the rest they need to grow well [21].
- **Good nutrition:** A balanced diet full of essential vitamins and minerals will help kids reach their full growth potential [22].
- **Regular exercise:** Because obesity is a problem for many kids, parents should ensure that

their kids exercise regularly. Bicycling, hiking, in-line skating, sports, or any enjoyable activity that will motivate kids to get moving will promote good health and fitness and help them maintain a healthy weight [23, 24].

2.2 Previous Works

Numerous relentless researcher has put their time and effort into studying obesity from different perspectives. Some focused on the causal relationship between diet and obesity, whereas some focused on physical activity. Some other studies tried to analyze the economic aspect associated with obesity [25]. For instance, Wang and Beydoun (2007) focused on predictive analysis of the prevalence of obesity by considering gender, age, socioeconomic, racial/ethnic, and geographic characteristics. The authors performed a systematic review and Meta-Regression analysis on multi-platform data (NHANES, BRFSS, and Youth Risk Behavior Surveillance System). The study projects an increasingly alarming rate of obesity for the next few decades . Additionally, another latest research focused on 2831 individuals from NHANES 2011-2014 data of mixed race (White, African American, and Mexican American) diagnosed with obesity [26]. The study focused on the causal relationships of nutrition intake and physical activity on the BMI of those individuals. Using multivariate analyses and path analyses, they conclude a significant causal path from increased physical activity to increased magnesium (Mg) intake to decreased BMI . Another study from 1999 focused more on the low-income Mexican American population. The obesity risk factors measured were body fat percentage, dietary fat intake, daily fruit and vegetable intake, and physical fitness [27].

There are quite a few studies that were more focused on how dietary intake (especially fiber) affects the metabolism of the elderly population. For example, in 2000, an in-depth analysis was performed focusing on the beneficial effects of high dietary fiber intake in patients with obesity. Thirteen patients were studied over more than a month and provided two different diets. Later they used repeated-measures analysis of variance and Wilcoxon signed-rank test to compare the two dietary periods. The study found that increased fiber intake was beneficial for patients with obesity (reduced BMI, reduced body fat percentage) [28]. A case-cohort study was conducted to determine

the association between dietary fiber intake and risk of obesity diabetes in the European population (n = 11,559) over 10.8 years. Country-specific HRs were estimated using Prentice-Weighted Cox-proportional hazards models and were pooled using a random-effects meta-analysis. Dietary fiber intake was associated with a lower risk of obesity [29].

As far as the younger population is concerned, some studies were conducted aimed at the younger population (2-18yo). In 2012, a study was conducted to explore the association of Fiber Intake with Childhood Obesity Risk (in 2-18yo, n = 2072) and Diabetes Risk of Adolescents (12-18yo, n = 2595) using NHANES 2003-2006 data. Another study was conducted on 2 - 18yo children. Both study discovered a beneficial association between dietary fiber intake and lower risk for overweight and obesity [30]. Another study was designed to identify dietary patterns and determine their relationships with obesity among Chinese children and adolescents. Data collected from 1282 children and adolescents aged 7-17yo. Three dietary patterns were identified: modern (high intakes of milk, fast foods and eggs), traditional north (high intakes of wheat, tubers and other cereals) and traditional south (high intakes of vegetables, rice and pork). As a result, the modern dietary pattern and the traditional north dietary pattern were associated with higher risk of obesity [31]. In 2007 another study was conducted on a cross sectional data of 2-4yo and found 'Overweight' children had less educated fathers and heavier parents than 'normal' weight children ($p < 0.05$).

Another study examined the relationship between dietary components, especially sugar intake, and insulin dynamics in overweight Latino youth (9-13yo, n = 63). Hierarchical regression analysis ascertained the potential independent relation between insulin dynamics and dietary components. Higher total sugar intake was significantly associated with lower acute insulin response (AIR) [32]. One study examined the association between overweight status and physical activity (PA) among gender and ethnic (Hispanic vs. non-Hispanic) sub-groups in elementary school-age children. There was a significant gender, ethnicity, and overweight interaction for total PA (combination of sedentary PA and moderate to vigorous PA) and MVPA (both $p < 0.01$) [33]. Another study focused on residential seatings effect on (rural or urban) obesity and related health

behaviors among children in the United States using NHANES data and found that significantly more rural children were obese than urban children [34]. In 2004 another study was conducted on 878 adolescents was to examine how diet, physical activity, and sedentary behaviors related to overweight status and found that of the seven dietary and physical activity variables examined in this cross-sectional study, insufficient vigorous physical activity was the only risk factor for higher body mass index for adolescent boys and girls [35]. Another study examined the relationship between television watching, energy intake, physical activity, and obesity status in 8-16yo US boys and girls. It found that the prevalence of obesity is lowest among children watching no more than one hours of television a day and highest among those watching four or more hours of television [36]. Another study found that being African American/ Hispanic, male, less physically active, and belonging to low-income families is more likely to be overweight [37]. Another study determined a negative association between fiber intake and obesity Using regression analysis. The total fiber in grams was inversely associated with BMI after adjusting for sex, age, education level, and income [38].

Another study was conducted on a sample of 5-6yo Hispanic (predominantly Mexican American) children in Chicago, Illinois, to see if overweight is more common in more highly acculturated immigrant families. Free access to food, drinking sweetened beverage was positively correlated with overweight but no significant relation with acculturation [39].

Besides, some other studies attempted to explore the effect of Sugar-Sweetened Beverages and dietary intake on obesity. A longitudinal 21 year follow-up study of Finnish children aged 3-18yo found the increase in consumption of sugar-sweetened soft drinks from childhood to adulthood was directly associated with BMI in adulthood in women (test statistics = 0.45, p = 0.0001) but not in men [40]. A study was conducted on one hundred sixteen 2 year old children and one hundred seven 5 year old children. Findings on juice consumption (sugar intake) have been more mixed; cross-sectional studies find a link [41], but some long-term studies do not. For example, a recent study looks at repeated cross-sections of 1,548 10 year old fifth graders in one school and finds a positive, but not significant, relationship between sweetened beverage consumption and BMI [42]. Less sugar sweetened beverage and more fiber intake were expectedly better for each case

[43, 44, 45, 46, 47, 48, 49].

CHAPTER III

SCOPES AND OBJECTIVES

Although numerous pieces of literature available that tried to find answers to some pressing issues like how obesity can be controlled and what risk factor is associated with it, using descriptive, unadjusted prevalence measure. Even some of the work tried to alarm us by forecasting the plausible scenario of a decade ahead. But none of them were aimed at the younger population (4-6yo), especially the Hispanic children, because of some limitations. In contrast, in the United States, the Hispanic population has a greater chance (11.7 percent) of being affected by obesity [50]. Our approach to analyzing obesity and the overlooked younger cohort narrowed us down to a unique perspective to put on obesity.

On top of that the global prevalence of obesity is projected to increase to 7079 individuals per 100,000 by 2030, reflecting a continued rise across all regions of the world [51]. So, this epidemic deserves more attention, and we should study this disease more meticulously than ever. Because if we can determine the cause and risk factors associated with it at an early age, it will strengthen us to fight the upcoming challenge of non-communicable chronic diseases like diabetes, cardiac complications, cancers, and musculoskeletal disorders. By keeping that in consideration, we set the purpose of this study as follows. Establish causal path models to analyze the effect of physical activity and macro nutrition intake on early childhood obesity among Hispanic children. Finally, we will implement cutting-edge machine learning algorithms (e.g., random forests, SVM, and decision tree) to classify and predict obesity among US children of age four to six.

CHAPTER IV

DATA AND VARIABLES

4.1 Brief Overview of NHANES Data

The National Health and Nutrition Examination Survey (NHANES) [52] program include a series of health examination surveys conducted in mobile examination units or clinics by the National Center for Health Statistics of the CDC. NHANES' main aim is to assess the health and nutritional status of a representative civilian US population using a multistage, stratified, clustered probability sampling design. The resulting data indicate Americans' nutritional and health status through dietary intake data, biochemical tests, physical measurements, and clinical assessments for evidence of nutritional deficiencies of individuals two years or older. The interviewer directly got answers to necessary questions from those aged 16 years or older. If the selected sample were less than sixteen years old, a SP (sampled person) or proxy interview was conducted and interviewer obtained responses from the parent or guardian. Here is a brief description of the components and their belongings data types.

1. **Demographic data:** The demographic file provides individual, family, and household-level information on several topics like age, education, family wealth status, ethnicity, gender of the respondent.

2. **Dietary data:** This component includes data regarding the individual's dietary intake. What respondents eat and what neuration they take from neuration supplements. However, the dietary data has two parts:

   (a) **Individual food file:** Contains detailed information about each food/ beverage item

(including the description, amount of, and nutrient content) reported by each participant is in the Individual Foods files.

(b) **Total Nutrient Intakes Files:** Contains each participant's daily total energy and nutrient intakes from foods and beverages, and whether the amount of food consumed was usual, much more than usual, or much less than expected, are included in the Total Nutrient Intakes files.

3. **Laboratory data:** Contains individual files on Urine Collection, Hepatitis A virus, HIV, Heavy Metals, Plasma Glucose, Total Cholesterol, Triglycerides, etc.

4. **Examination data:** Contains individual files on Audiometry, Blood Pressure, Body Measures, Muscular Strength, Oral Health, Vision Exam, etc.

5. **Questionnaire data:** Contains individual files on Physical Activity, Alcohol Use, Balance, Blood Pressure, Diabetes, Drug Use, Social Support, Vision, Weight History, etc.

4.2 NHANES Data Retrieving and Processing

Initially, we decided to use the latest data from NHANES in this study to get the most updated results. But Due to the coronavirus disease 2019 (COVID-19) pandemic, the NHANES program stopped conducting fieldwork in March 2020. As a result, data collection for the NHANES 2019-2020 cycle was not finished, and the data that was gathered was not representative of the entire country. Therefore, we decided to use the 2017-2018 cycle data. However, due to our defined cohort of 4-6yo, to produce a reasonable representative sample, we used data beyond the 2017-2018 cycle [53, 54]. Literature suggests, sample size can be augmented by combining previous cycle years data [34]. So, we decided to combine the NHANES cycles year "2014-2015," "2016-2017," and "2017-2018" together.

To construct a specific data set for this study, we had to collect data from more than one segment (see 4.1). For instance, age, gender, weight, and ethnicity are in the Demographics segment, while data about dietary intake (includes 65 nutritional intake variables) is in the dietary data section.

Table 4.1: List of variables, description and it's code name in NAHNES dataset that were used in the study.

Component	Variable Name	Description	NHANES variable code name
Demographics data	ID	Respondent sequence number	SEQN
	Gender	Gender	RIAGENDR
	Age	Age in years at screening	RIDAGEYR
	Race	Race/Hispanic origin	RIDRETH1
	SES	Income to poverty ratio	INDFMPIR
	Education	HH reference persons' education	DMDHREDU
Dietary Data(Total Nutrition Intake)	ID	Respondent sequence number	SEQN
	Fiber	Average dietary fiber intake (gm)	DR1TFIBE, DR2TFIBE
	Carb	Average carbohydrate intake (gm)	DR1TCARB, DR2TCARB
	Sugar	Average sugars intake (gm)	DR1TSUGR, DR2TSUGR
	Fat	Average fat intake(gm)	DR1TTFAT, DR2TTFAT
Examination Data	ID	Respondent sequence number	SEQN
	BMI	Body Mass Index (kg/m**2)	BMXBMI
Questionnaire data	ID	Respondent sequence number	SEQN
	Physical Activity (PA)	Days physically active at least 60 min.	PAQ706

BMI measurement data was collected from the examination data segment. Finally, data on physical activity was collected from the questionnaire segment. One variable that was identical in each segment in each year is the respondent sequence number (SEQN). We collect SEQN for each segment to merge the segment data files. Before merging, other necessary steps were performed as described in following sections. Table 4.1 shows the variables, key names, and descriptions we initially decided to include in our study upon reviewing numerous pieces of literature. This study will use combined data (cycle 2014-2015, 2015-2016, and 2017-2018)from NHANES.

4.2.1 Demographic Data

In the demographic data, age was recorded in years at the time of the interview. The race was coded as " 1 = Mexican American ", "2 = Other Hispanic", "3 = Non-Hispanic White", "4 = Non-Hispanic Black", and "5= Other Race - Including Multi-Racial". We combined "1 = Mexican American" and "2 = Other Hispanic" as Hispanic children. The Household's reference person's[1] education level was coded defiantly in cycles 2017-2018, 2013-2014, 2015-2016. Please refer to Table 4.2 to see the differences. We recoded the 2013-2014 and 2015-2016 coding per 2017-2018. So 2013-2014 and 2015-2016, so levels 1 and 2 merge into level 1 that is less than high school

[1]The household reference person is defined as the first household member 18 years of age or older listed on the household member roster, who owns or rents the residence where members of the household reside.

Table 4.2: Different levels of education in different cycle years.

Year	Levels	Value Description
2017-2018	1	Less than high school degree
	2	High school grad/GED or some college/AA degree
	3	College graduate or above
2013-14 and 2015-1016	1	Less Than 9th Grade
	2	9-11th Grade (Includes 12th grade with no diploma)
	3	High School Grad/GED or Equivalent
	4	Some College or AA degree
	5	College Graduate or above

degree (LHS). levels 3 and 4 merged into level 2 that is, high school grad/GED or some college/ AA degree (HSC). Finally, level 5 from the cycles 2013-2014 and 2015-2016 was recoded as 3. The variable socioeconomic status (SES) represents the measure of the ratio of family income to poverty threshold (HHS poverty guidelines specific to family size [55]). The value ranges from 0-4.99; the higher the value is better the socioeconomic status.

4.2.2 Dietary Data

In order to record dietary data (fiber, fat, carb, sugar), NHANES conducted two 24-hour dietary recall interviews where all foods and beverages consumed the previous 24 hours ending at midnight were solicited and recorded using the standardized Automated Multiple Pass Method. The first dietary recall was collected in person, and the second was collected by phone 3 to 10 days later, on a different day of the week. A set of food measurement guides were provided to participants for assistance in estimating portion sizes during both the in-person and phone recalls. For each participant, daily total energy and nutrient intakes from foods, beverages, and water (including tap and bottled water), are included in the Total Nutrient Intakes data. This data does not contain nutrients obtained from dietary supplement intakes, antacids, or medications. For example, "DR1TTFAT = Total fat (gm)" measures the amount of fat an individual has taken on the first day of the interview through the medium mentioned above. So we collected total dietary data from the first day (DR1TTFAT) and the second day (DR2TTFAT) and calculated the average of these two

measurements to calculate the "Fat" variable [34]. If either of those two measures (day one and day two) was missing, we took the available one. If both of them were missing, we recorded it as "NA." So the resulting "Fat" variable contains an average fat intake of two days. A similar approach was followed to calculate the variables "Fiber," "Carb," and "Sugar."

4.2.3 Examination Data

From this component, we extracted BMI information about sampled individuals. Body Mass Index (BMI) was calculated as weight in kilograms divided by height in meters squared and then rounded to one decimal place. The measurement was recorded for individuals of age two years or more. This variable was later used to classify an individual as obese or not obese according to CDCs guidelines [56]. According to CDC's new" BMI reference, provided in 2000, "Overweight" is defined as a BMI greater than or equal to the sex-age–specific reference 95th BMI percentile. However, the gender and age specific reference BMI percentile is available from age 24 months, 24.5 months, and then with one month increment in ages. So reference BMI percentile was not explicitly available for four years (48 months), five years (60 months), and six years (72 months). To find out the reference BMI, we performed linear interpolation using the nearest two BMI available. For example, BMI for age 60 was interpolated using the gender and age specific reference BMI for ages 59.5 and 60.5. The exact manner was followed for ages 48 months and 72 months.

4.2.4 Questionnaire Data

We extracted physical activity data from this segment. In NHANES interview, individuals of age 2-11 years were asked, "During the past seven days, on how many days were you/was SP physically active for a total of at least 60 minutes per day? Add up all the time you/he/she spent in any physical activity that increased your/his/her heart rate and made you/him/her breathe hard some of the time." Then the answer was recorded numerically, ranging from 0 days to 7 days. According to the guideline, a dichotomous variable was created as physically active if the record indicates vigorous activity $\geq$3 days; otherwise, not physically active ($<$ 3) days [57].

4.3 Examination of Three Cycles of NHANES

We collected the necessary data files from the NHANES website according to our demand and extracted the variables from each component listed in Table 4.1. Then we merged those four components variable using "ID" as a key variable. After that, we cleaned the data: removing missing values, taking care of special coded values, and any unusual observations. For example, the variable "education" had 7 and 9 coded for "refused" and "don't know" responses. After doing that, when we applied the age filter (children of age 4-6yo), we had 468, 425, and 317 observations in the 2013-2014, 2015-2016, and 2017-2018 cycle data, respectively. Combining those three cycles, in the end, gave us a cohort of 1210 children 4-6yo among them 348 observations was Hispanic. But before we did that, we validated that these three cycles could be merged and used as a single data set.

4.3.1 Comparison of Distribution

Figure 4.1 shows the grouped bar diagram of categorical variables for three cycles 2013-2014, 2015-2016, and 2017-2018. The proportion of male and female individuals was similar across three cycles, whereas non-Hispanic black children were the highest in 2017-2018. The proportion of physically active children was found to be much higher across all three cycles than not active.

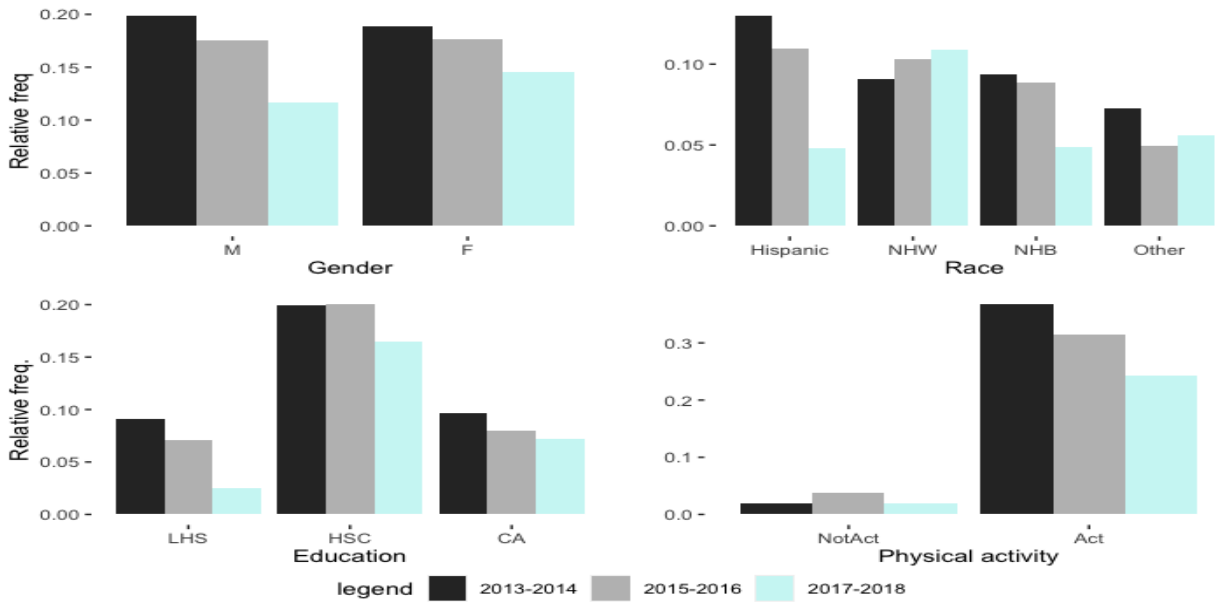

Figure 4.1: Bar-diagram of Gender, Race, Education, and Physical activity by cycles.

4.3.2 Test of Hypothesis

The Figure 4.2 shows the probability density functions of continuous variables in our data. Our visual inspection reveals that, there is not any considerable difference in the distribution of the variables across all three cycles. Despite of visual confirmation, we performed the K-S test to test the assumption that all three cycles of samples came from similar probability distribution under the null hypothesis. Test statistics and p-value was reported in Table 4.3. Analysis reveals that there is not enough evidence to assume their population differences. Additionally, we performed one way F test for multiple mean comparisons. At a 5% level of significance, the p-value of the F-test tells us we can assume the similarity in distributions. Finally, we performed Mood's median test to test the null hypothesis that the three population mean is equal. Combining Figure 4.2 and Table 4.4 we can confirm that there is no dissimilarity among the three cycles of data.

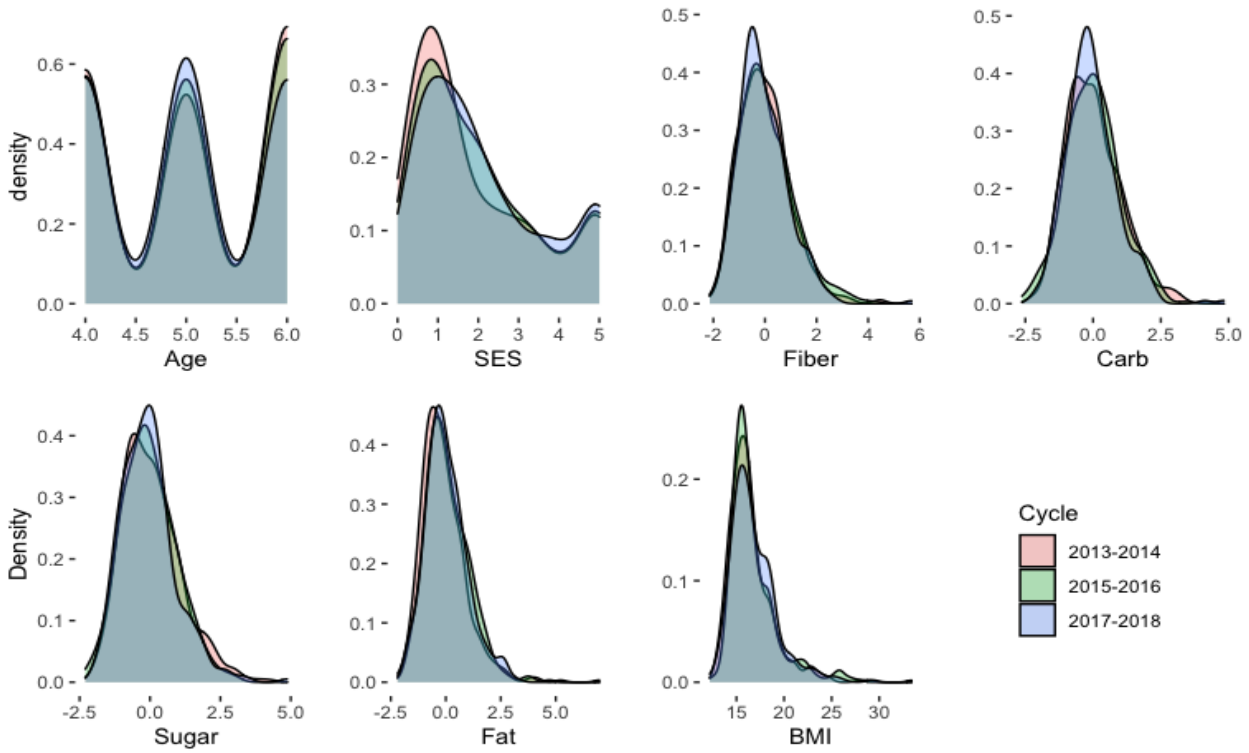

Figure 4.2: Probability density function of continuous variables for three cycles.

Table 4.3: Kolmogorov—Smirnov test, to pair wisely test whether three cohorts has different distributions. Value inside parenthesis are p-value to corresponding test statistic. Bolder values implies significant p-value.

Variables –>	**Age**	**SES**	**Fiber**	**Carb**	**Sugar**	**Fat**	**BMI**
Cycle (2013-2014, 2015-2016)	0.02 (1.000)	0.1 (**0.028**)	0.05 (0.631)	0.05 (0.747)	0.05 (0.757)	0.1 (**0.018**)	0.07 (0.186)
Cycles (2013-2014, 2017-2018)	0.06 (0.444)	0.15 (**0.000**)	0.06 (0.608)	0.07 (0.261)	0.08 (0.182)	0.11 (**0.023**)	0.04 (0.852)
Cycle (2015-2016, 2017-2018)	0.05 (0.804)	0.06 (0.555)	0.07 (0.274)	0.1 (0.071)	0.06 (0.462)	0.05 (0.743)	0.06 (0.445)

Table 4.4: Test of hypothesis for equality of population mean and median in three cycles. Bolder values implies significant p-value.

Variables	**F-test**[1]		**Mood's Median-test**[2]	
	Test statistic	**p-value**	**Test statistic**	**p-value**
Age	0.610	0.544	3.350	0.187
SES	2.350	0.096	16.790	**0.000**
Fiber	1.480	0.228	3.000	0.223
Carb	0.800	0.450	4.160	0.125
Sugar	1.460	0.233	0.290	0.863
Fat	4.060	**0.017**	8.510	**0.014**
BMI	2.170	0.115	1.230	0.542

[1] F-test for multiple mean comparison.
[2] Mood's median test: Test whether three cohorts has different Medians.

CHAPTER V

STATISTICAL METHODS

5.1 Path Analysis

Path analysis is a statistical technique used mainly to examine the comparative strength of direct and indirect relationships among variables. A series of parameters are estimated by solving one or more structural equations to test the fit of the correlation matrix in two or more causal models, which the researcher hypothesizes to fit the data. Geneticist Sewall Wright initially developed path analysis in the 1920s to examine the effects of hypothesized models in phylogenetic studies [58]. In Wright's analysis, the unknown parameters in the model were solved by developing a system of equations based on the correlations among variables impacting the outcome. The goal of the path analytic method was to discover "the extent to which any specific cause causes the variance of a given consequence" along each individual path in such a system [58].

A few decades later, along with social science and biology, other scientific fields adopted this efficient method to address causality in versatile problems, including health care and the latest health issue [59, 60]. Wu, Datta, and others used path analysis to examine the causal path among demographic and nutrition intake on type 2 diabetes [26]. Besides its applicability in numerous fields, we chose path analysis due to another advantage. Using path analysis, we could explicitly specify how the variables relate to one another and expect the development of detailed and logical theories about the processes influencing obesity. It will also be beneficial because it will allow us to break apart the causal linkages among various factors affecting obesity into direct effects and indirect components. Even though path analysis suits interval level data, its efficiency, structure, recent advancement[61] and wide availability of computational tools [62] motivated us to implement

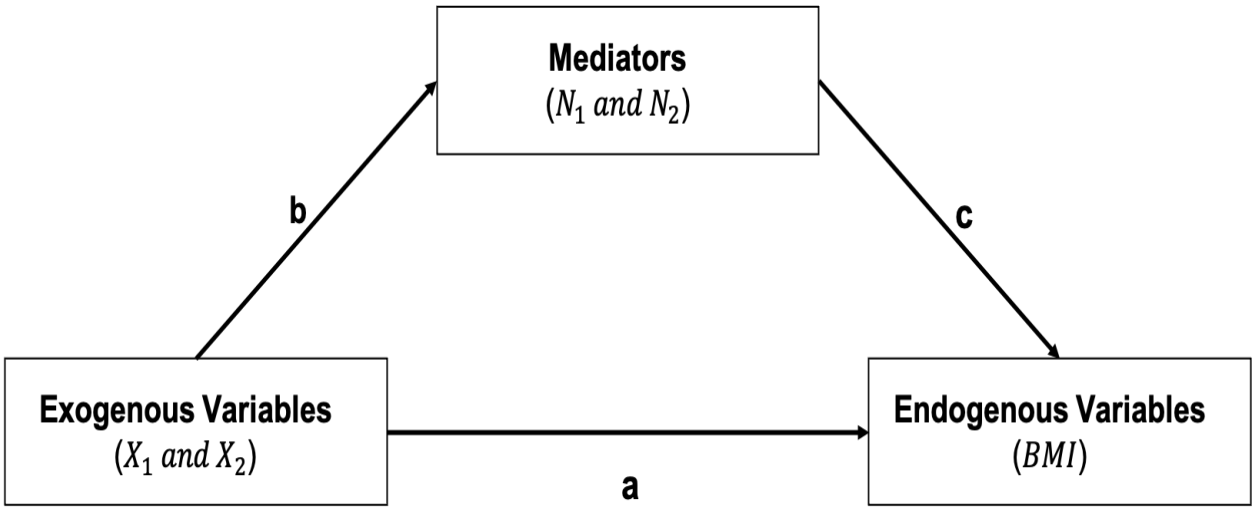

Figure 5.1: An example of Exogenous, Endogenous variables and Mediation effects.

this method in our study.

5.1.1 Exogenous, Endogenous Variables, and Mediation

Variables often play more than one role in path models and this is reflected in the analytic language used in path analysis. Exogenous variables are variables whose cause is external to the model and whose role is to explain other variables or outcomes in the model. In Figure 5.1, for example, the model does not explain the variability in mediator. However, these exogenous variables are hypothesized to account for differences dependent variables.

Endogenous variables are variables that are caused by one or more variables within the model. Endogenous variables have incoming arrows and can include outcome variables (only incoming arrows) and intervening causal variables. Endogenous variables, such as nutrient intake, have only incoming arrows. The hypothetical model in Figure 5.2 indicates that nutrient intake are influenced by other variables in the model (e.g., age, income to education), and in turn have an effect on BMI.

The overall association between two variables can be generated by several different causal relationships at the same time. The overall association between x and y could be due to the effects of a causal ancestor on a causal descendant, (and/ or) the effects of a common causal ancestor, (and/ or) due to an unresolved causal relationship. The effects of causal ancestors on causal descendants can

be breakdowns into direct and indirect effects. Indirect effects are the effects of a causal ancestor on its descendant that are completely transmitted through some other variable. This intervening variable is sometimes called a mediator of the causal effect. For instance, in Figure 5.1, the effect of X on *BMI* along the path $X \rightarrow N \rightarrow BMI$ is an indirect effect of X that is mediated by N. To quantify this effect, we have to multiply the path coefficients (**b**, **c**) along this path. This indirect effect measures how much *BMI* would change following a change in X if all causal parents of *BMI* except for N were held constant. In general, an indirect effect measures how much the effect variable would change following a change in the indirect cause when this effect is transmitted only along the path in question. However, the direct causal effects are the effect that goes directly from one variable to another without using any mediator. For example in Figure 5.1 the coefficient **a** express the direct effect of X on BMI ($X \rightarrow BMI$).

5.1.2 Hypothetical Causal Model

We will use Path Analysis[58, 63, 64] to determine the causal relationship between demographic variable, nitration intake, physical activity to the outcome variable obesity. According to studies[65, 66, 67], there is a strong correlation between body fat and BMI; it is the most accurate way to calculate body fat. So our hypothetical causal model is built under this assumption.

A hypothetical causal model often represented by path diagram which consisted rectangle (variables), oval (error variance among endogenous variables) and pointed arrow to portray a hypothetical causal path. So our hypothetical causal model is shown in Figure 5.2. From our knowledge, this method has not been used for this particular reason in this type of cohort. This method fit a multivariate non-experimental dataset to a complex causal model using causal structural equations [58, 63, 64]. This powerful technique analyzes and compares different complex hypothetical causal models afterward finds a model that consistently fits the data. Using Bonferroni correction, it estimates path coefficients thus, the causal importance of the variables in the pathways to the outcomes can be quantitatively estimated. We will initially hypothesize that physical activity (X_1) and demographic characteristics (X_2) affect nutrition intakes (N_1 and N_2), and they also affect BMI in other words, physical activity and dietary choices directly affect BMI. Here, N's, BMI

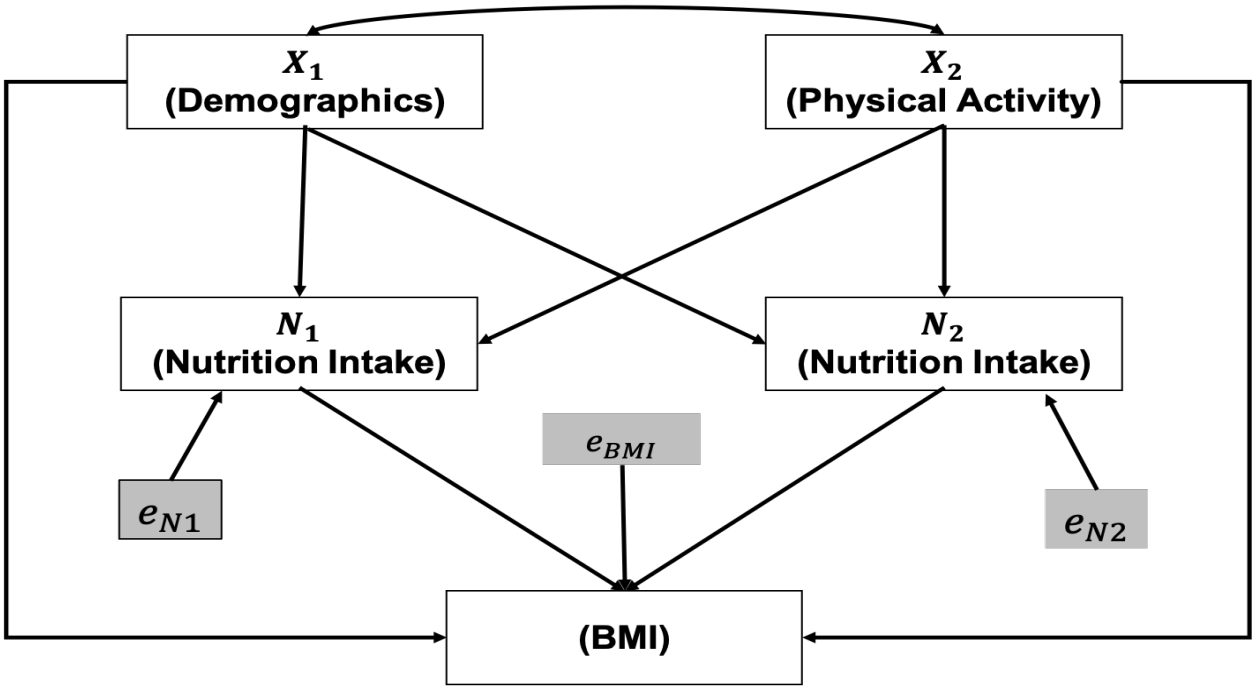

Figure 5.2: Hypothetical causal model: X_1 and X_2 indicate demographic characteristics (such as age, gender and ethnicity, physical activity). N_1 and N_2 are nutrition intake variables.

are endogenous variables (ie variances are partially explained by the other variables within the causal model) whereas, X's are exogenous variables (i.e. variances not explained by variables in the hypothetical causal model). As the variance of an endogenous variable is modeled by other variables, **e** here represents the residual error of that measurement.

5.1.3 Model Structure

For simplicity, first, we will define a simpler model with fewer variables and then move towards its generalization. For endogenous variables: $y_k, y_{j'}$, exogenous variables: $x_j, x_{j'}$, and errors: $\varepsilon_k, \varepsilon_{k'}$, the structural coefficients representing the direct (partial) effect of an exogenous variable x_j on an endogenous variable y_k is γ_{kj} (gamma) and of an endogenous variable $y_{k'}$ on another endogenous variable y_k is $\beta_{kk'}$ (beta). Covariances between two exogenous variables, x_j and $x_{j'}$ is $\sigma_{jj'}$ and two error variables, ε_k and $\varepsilon_{k'}$ is $\sigma_{kk'}$. Finally, the variance can be written as σ_j^2 or as σ_{jj} (i.e., the covariance of a variable with itself).

Consider we have four exogenous (x_1, x_2, x_3, x_4) variable and two endogenous variables y_1 and y_2.

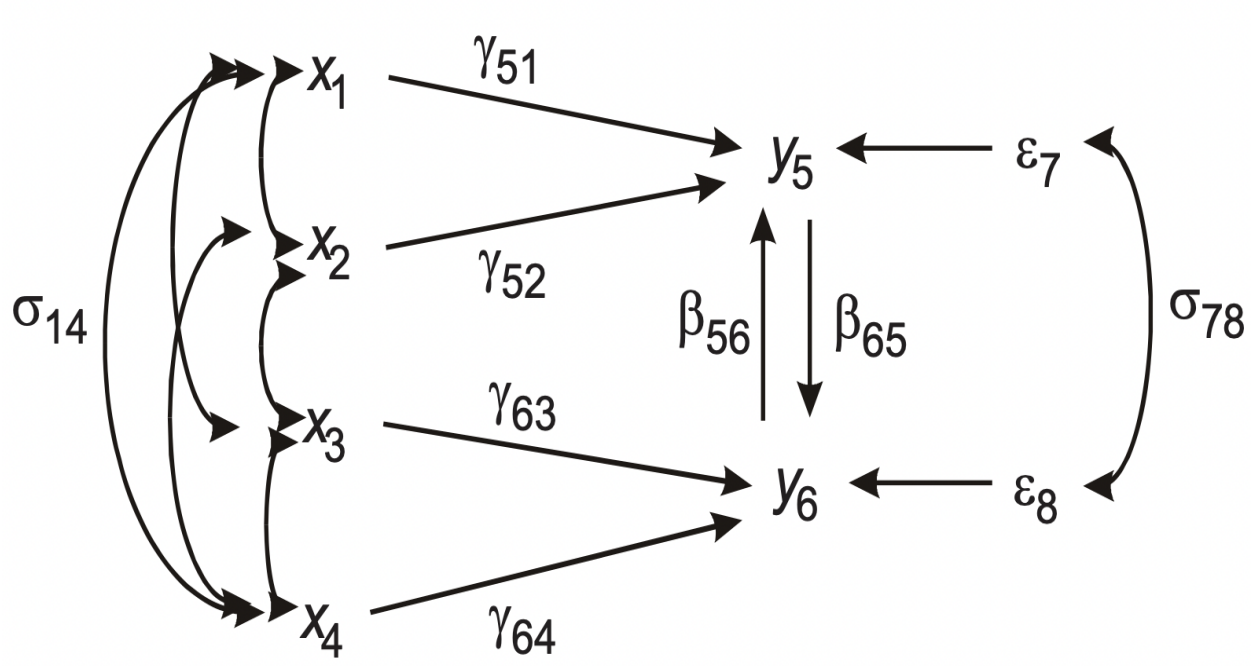

Figure 5.3: Causal model with four exogenous and two endogenous variables. Here γ is the path coefficients from exogenous variables to endogenous variables. β is the path coefficient for endogenous to endogenous variables.

Using the above notation, we can construct a diagram of this example model like Figure 5.3

The structural equation of the model shown in Figure 5.3 can easily be written as

$$\begin{aligned} y_{5i} &= \gamma_{50} + \gamma_{51}x_{1i} + \gamma_{52}x_{2i} + \beta_{56}y_{6i} + \varepsilon_{7i} \\ y_{6i} &= \gamma_{60} + \gamma_{63}x_{3i} + \gamma_{64}x_{4i} + \beta_{65}y_{5i} + \varepsilon_{8i} \end{aligned} \tag{5.1}$$

By excluding the intercept term and collecting variables and equation term on one side, we have the following model for an individual observation,

$$\begin{aligned} 1y_5 - \beta_{56}y_6 - \gamma_{51}x_1 - \gamma_{52}x_2 + 0x_3 + 0x_4 &= \varepsilon_7 \\ -\beta_{65}y_5 + 1y_6 + 0x_1 + 0x_2 - \gamma_{63}x_3 - \gamma_{64}x_4 &= \varepsilon_8 \end{aligned} \tag{5.2}$$

Note that the variables that don't have a path connection got 0 as a coefficient. Now the Equation 5.2 can be expressed as following matrix form,

$$\begin{bmatrix} 1 & -\beta_{56} \\ -\beta_{65} & 1 \end{bmatrix} \begin{bmatrix} y_5 \\ y_6 \end{bmatrix} + \begin{bmatrix} -\gamma_{51} & -\gamma_{52} & 0 & 0 \\ 0 & 0 & -\gamma_{63} & -\gamma_{64} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \varepsilon_7 \\ \varepsilon_8 \end{bmatrix} \tag{5.3}$$

In general the individual model can be expressed for **q** endogenous and **m** exogenous variables.

$$\underset{(q\times q)}{\mathbf{B}} \underset{(q\times 1)}{\mathbf{y}_i} + \underset{(q\times m)}{\Gamma} \underset{(m\times 1)}{\mathbf{x}_i} = \underset{(q\times 1)}{\varepsilon_i} \tag{5.4}$$

Finally, for n observations the model will be-

$$\underset{(n\times q)(q\times q)}{\mathbf{Y}} \underset{}{\mathbf{B}'} + \underset{(n\times m)(m\times q)}{X} \underset{}{\Gamma'} = \underset{(n\times q)}{E} \tag{5.5}$$

If **Y** and **E** follows the standard assumptions of linear models [68], the likelihood function corresponding to Equation 5.5 can be expressed as follows,

$$\begin{aligned} \log_e L(\mathbf{B},\Gamma,\Sigma_{\varepsilon\varepsilon}) =& n\log_e|\det(\mathbf{B})| - \frac{nq}{2}\log_e 2\pi - \frac{n}{2}\log_e\det(\Sigma_{\varepsilon\varepsilon}) \\ & - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{B}\mathbf{y}_i+\Gamma\mathbf{x}_i)'\Sigma_{\varepsilon\varepsilon}^{-1}(\mathbf{B}\mathbf{y}_i+\Gamma\mathbf{x}_i) \end{aligned} \tag{5.6}$$

Values of Γ and B that maximizes the function will be the estimates of the coefficients [69].

5.1.4 Estimation of Parameters in R-lavaan

We used R lavaan library to estimate the effects of exogenous and mediators on BMI. While accounting for slight violations of the normality assumption (6.2, we used the maximum likelihood function and NLMINB [70] optimization method to find the optimized value of parameters [71]. Then indirect effect was calculated by defining custom parameters as a function of the product of the path coefficients ($\rho_1 * \rho_2$) on the specified pathway. The total causal effects of demographic variables or physical activity on the outcome (BMI) are the sum of the direct and indirect effects

from all the paths that lead to the outcome [64]. Special Indirect Effect (SIE) in Table 6.3 denotes a variable's impact through a single mediator variable. And the Total indirect Effects (TIE) is the sum of all SIE. Finally, the Total Effects (TE) comprises the TIE and any direct effects a variable has on the outcome variable. The strength of the causality increases with the absolute value of the route coefficient. A positive path coefficient indicates that a causal variable increases the outcome value, and a negative path coefficient indicates a causal variable decreases the outcome value. An indirect path (i.e., the path that goes through mediators) was considered significant if the absolute value of the product of the single path coefficients is greater than 0.01 using the path product property [63, 64].

5.1.5 Assumptions

Like other modeling techniques, path analysis is based on a number of assumptions. When the analysis is done, it requires careful consideration if the model assumption is holding or not. A substantial violation in model assumption could lead to erroneous estimates of parameters resulting in biased high variance estimates [72]. Because the path analysis framework we present relies on the solution of multiple linear regression equations, the statistical assumptions underlying these procedures merit attention [73, 74].These assumptions are summarized as follows:

- All dependent variables should be roughly normally distributed, and all relationships between the variables are presumed to be causal, linear, and additive. It's also forbidden to have other curvilinear relationships or interactions.

- Residuals should not be correlated with the variables that predict the outcome variables toward which they point. This means that in Figure 5.3 ε_7 is not correlated with variables Y_5 and X_1 and X_2. That implies residual from the nutrition intake variables should not correlate with demographic variables like age, gender, race etc. This assumption implies that all relevant variables are included in the model, and any unmeasured variables are not correlated with the specified predictor variables.

- Causation flows in one direction; there are no feedback loops. Model is recursive.

- The variables are measured without error. Implies exogenous variables used in the model should be error free in terms of measurement error.
- Predictor variables may be continuous, ordinal categorical, or dichotomous.
- There is low multicollinearity among predictor variables in any of the structural equations.

5.1.6 Model Significance and Goodness-of-Fit

In statistical modeling, "goodness-of-fit" evaluates how well sample data fits a distribution from a population having a normal distribution. Simply put, it makes assumptions about whether a sample is representative of the data found in the actual population or is biased. The disparity between the actual values and those predicted by the model in the case of a normal distribution is established via goodness-of-fit. The chi-square is one of the techniques for figuring out goodness-of-fit. However, the ordered Mahalanobis distances vs. estimated quantiles (percentiles) for a sample of size n from a chi-squared distribution is an advanced variant of the standard chi-square test that can be very helpful. When using data from a multivariate normal distribution, this should resemble a straight line. The K-S test of distributional equality can be done to further check the normality assumption [75]. It basically takes the supremum of the difference of the empirical CDF of the first and the second sample, respectively, and compares it with test statistics to make a decision. A brief discussion on the K-S test is available in the supplementary material (see A).

We used the comparative fit index to evaluate the fit of our path analysis model. We chose this because of one advantage over other matrices. The comparative fit index (CFI) accounts for the sample size difficulties present in the chi-squared test of model fit and the normed fit index when analyzing the model fit by considering the difference between the null and the hypothesized model [76]. A null model is typically in which observed variables are constrained to covary with no other variables. Larger numbers indicate better fit, while CFI values range from 0 to 1. A CFI value of.90 or higher can be regarded as an indication of good model fit [77].

We also used the SRMR metric to evaluate the model fit. It is the average of the standardized residuals between the observed and hypothesized covariance matrices is represented by the term

"standardized root mean square residual" (SRMR) (Chen, 2007). This absolute fit index can be defined as follows

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left[\left(s_{ij}-\hat{\sigma}_{ij}\right)/\left(s_{ii}s_{jj}\right)\right]^2}{p(p+1)/2}}$$

where s_{ij} indicates a component of $\mathbf{S}$ sample covariance matrix and $\hat{\sigma}_{ij}$ shows a component of $\Sigma(\hat{\theta})$ hypothesized model whereas p is the number of observed variables.

Although SRMR can be used to determine an acceptable fit when it generates a value less than 0.10, an SRMR value less than .08 is widely used by researchers to consider the model fit as a good fit. Its relative independence from sample size is one of the reasons why researchers use the SRMR index in SEM modeling [78].

5.2 Supervised Learning

In the past few decades, statistical learning, in a general sense, improved significantly both in efficiency and accuracy. Most of it happened due to massive computational advancement and the relentless work of many researchers. Nowadays, it has become a sophisticated method to address many real-world situations like classification and predictive analysis. Supervised learning, popularly known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is distinguished by how it trains computers to accurately classify data or predict outcomes using labeled datasets. The model modifies its weights as input data is fed into it until the model has been properly fitted, which takes place as part of the cross-validation process. Usually, a training set is used in supervised learning to instruct models to produce the desired results. This training dataset has the right inputs and outputs, enabling the model to develop over time. The loss function verifies the algorithm's correctness, and iterations are made until the error is sufficiently reduced. When using data mining, supervised learning may be divided into two issues: regression and classification. The following sections 5.2.1 to 5.2.5 briefly describes the robust supervised learning method that we have used in this study for classification purposes.

5.2.1 Principle Component Regression (PCR)

Building the first M principal components, $Z_1, \ldots, Z_M$, and utilizing these components as predictors in a linear regression model which is fitted using least squares is known as the principal components regression (PCR) technique [79, 80, 81, 82]. The main point is that much of the variability in the data and the connection to the response can frequently be explained by a small number of principal components. In other words, we assume that the directions where X_1, ..., and X_p exhibit the greatest variance are the directions where Y is present. So, the fitted equation can be expressed as

$$\hat{\mathbf{y}}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

where $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle \,/\, \langle \mathbf{z}_m, \mathbf{z}_m \rangle$. Since the $\mathbf{z}_m$ are each linear combinations of the original $\mathbf{x}_j$, we can express the solution in terms of coefficients of the $\mathbf{x}_j$:

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

5.2.2 Support Vector Machine

Support vector machine (SVM) [83], an approach for classification developed in the computer science community in the 1990s that separates two (sometimes more) classes based on the principle of maximum margin classifier using a one/set of hyper plane in an high dimensional space, which is used for classification, regression, and outliers detection. In a p dimensional space, a hyper-plane is a flat affine subspace of dimension p-1. SVM separates data using a separating line and furthest away from the closet data point and this make SVM unique from other algorithm [81]. SVM are widely used as it can find a complex relationship between data without having lots of information about the data set. But SVM was an extension of the maximum margin classifier due to its limitation to the nonseparable class. Even though the support vector classifier can address the issue, the SVM can go beyond the linear margin and can work with more real-life linearly

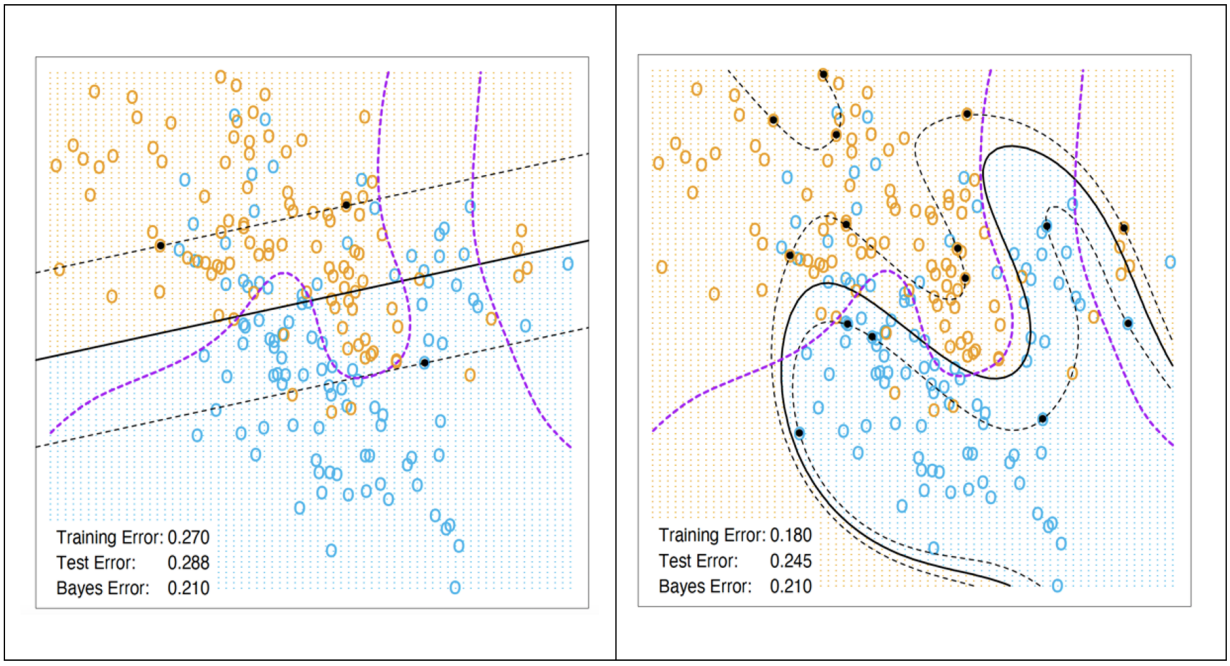

Figure 5.4: Example of linear and nonlinear SVM with overlapping classes. [1]

nonseparable classes. Figure 5.4 demonstrates how SVM separates a linearly nonseparable class. But if the number of features is bigger than the number of data point SVM trends to overfit.

The figure on the left showing the linear support vector boundary for with two overlapping classes and figure to the right showing the nonlinear SVM with a 4th degree polynomial kernel for some hypothetical data. The support vector machine classifies a test observation depending on which side of a hyperplane it lies. The hyperplane is chosen to correctly separate most of the training observations into the two classes. It is the solution to the optimization problem

$$\begin{aligned} &\underset{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n}{\text{maximize}} \quad M \\ &\text{subject to } \textstyle\sum_{j=1}^{p} \beta_j^2 = 1 \\ &y_i\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\right) \geq M\left(1-\varepsilon_i\right) \\ &\varepsilon_i \geq 0, \quad \textstyle\sum_{i=1}^{n} \varepsilon_i \leq C, \end{aligned} \tag{5.7}$$

Here **M** represents the margin Moreover, with the introduction of linear $K(x_i, x_{i'}) =$

[1] The figure was taken from the book "An Introduction to Statistical Learning [84]."

$\sum_{j=1}^{p} x_{ij}x_{i'j}$, and polynomial $K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij}x_{i'j}\right)^d$ kernel we could address the nonliniearity in the class boundary using SVM.

5.2.3 K-nearest Neighbors

K-nearest neighbors (KNN) is a supervised learning technique that determines the class of the desired data point based on the K closest point from it [85, 81]. The KNN algorithm determines a data point's class by using the majority vote principle. For example, if we set k as 7, the classes of 7 closest points are checked. Then KNN predicts the class according to the majority class out of the seven nearest classes. In general, say, K is a positive integer and x_0 is a test observation, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

$$\Pr(Y = j|_X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

5.2.4 Classification Tree

The classification tree is a popular supervised machine learning technique [86, 86]. As the name suggests, the branches of a classification tree stand for attributes, and the leaves stand for decisions. When in use, the decision tree procedure begins at the trunk and proceeds through the branches until it reaches a leaf. Hunt's, one of the earliest classification tree algorithms, helps us to understand this concept better (the latest algorithm includes CART, ID3, C4.5, etc.). Let D_t be the set of training records (samples) that reach a node t; then the tree grows using the following general procedure:

- If D_t contains records that belong the same class y_t, then t is a leaf node (pure) labeled as y_t. That is no further brancing is necessary ftom this node.
- If D_t is an empty set, then t is a leaf node labeled by the default class, y_d

- If D_t contains records that belong to more than one class (also known as an impure node), the algorithm will use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset until desired purity of the node is met.

Now, a few questions might come to our mind about how we can decide where to split the attributes or, in other words, what the best split is and when to stop splitting. To determine the best split, we can use misclassification error, Gini index, or cross-entropy function.

Since the algorithm aims to assign an observation in a given region to the most commonly occurring class of training observations in that region, the classification error rate is simply the fraction of the training observations in that region that do not belong to the most common class. So the classification error at a node t is :

$$\text{Error}(t) = 1 - \max_j P(j \mid t)$$

Which measures misclassification error made by a node, and its maximum is $(1 - 1/n_c)$ when records are equally distributed among all classes, implying the least interesting information. And minimum (0.0) when all records belong to one class, implying the most interesting information.

Gini Index for a given node t **:**

$$\text{GINI}(t) = 1 - \sum_j [p(j \mid t)]^2$$

Here $p(j \mid t)$ is the relative frequency of class j at node t). Maximum $(1 - 1/n_c)$ acquires when records are equally distributed among all classes, implying the least interesting information, and minimum (0.0) happens when all records belong to one class, implying most interesting information. A more advanced algorithm like CART and SLIQ do the splitting based on the Gini measures of the entire split, which is a weighted average of each child node. Say a parent node p is split into k partitions (children), then the quality of the split is computed as

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} \text{GINI}(i)$$

Where, n_i is the number of records at child i, and n is the number of records at node p.

Entropy at a given node t:

$$\text{Entropy } (t) = -\sum_i p(j \mid t) \log p(j \mid t)$$

Here $p(j \mid t)$ is the relative frequency of class j at node t). Maximum ($\log n_c$) acquires when records are equally distributed among all classes, implying most information, and minimum (0.0) happens when all records belong to one class, implying least information. Like Gini, the entropy measures of the entire split, which is a weighted average of each child node, can be calculated as

$$Entropy_{split} = -\sum_{i=1}^{k} \frac{n_i}{n} \text{Entropy}(i)$$

Where, n_i is the number of records at child i, and n is the number of records at node p.

However, an algorithm like ID3 and C4.5 uses the gain of a split to decide which split to go with. The gain of a split is often referred to as Information Gain and can be calculated as

$$\text{Gain } (\text{ split }) = E(\text{ Parent set }) - \sum E(\text{ all child sets })$$

$$GAIN_{\text{splut}} = \text{ Entropy } (p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} \text{Entropy}(i) \right)$$

The algorithm chose the splits that achieve the most reduction in entropy, i.e., split that maximizes the GAIN.

5.2.5 Random Forests

Random forests is widely used ensemble learning method where the classification output of the random forests is the class selected by most trees [87, 81]. Other ensemble learning methods

such as bagging methods get success for reducing the variance of an essential prediction function. Bagging especially works better for trees where variance is high and bias is low. Random forests is modification of the bagging technique that builds a large collection of de-correlated trees. The principle difference between bagging and random forests is the choice of predictor of subset size say m. For example, if a random forests model has $m = p$ then the model is simply to bagging. Random forests generally outperform than the decision trees but in comparison to accuracy it is lower than the gradient boosted trees. The random forests algorithm's steps are as follows:

- **Step 1:** From a data set with k predictors, n predictors are selected randomly and used in the Random forests algorithm.
- **Step 2:** Then, build a unique classification tree for each sample.
- **Step 3:** Each decision tree will produce an output.
- **Step 4:** For classification, the final result will be evaluated using a majority vote (or an average for regression problem).

5.2.6 Chi-square, p-value, and Level of Significance

Absolute Fit Indices assess how well a researcher's theoretical model fits the observed data they have collected. The only statistically based measure of fit is that the Likelihood Ratio Test, also known as the "chi-square" (CMIN) statistic and its associated "probability" or p-value. Since it is a test of statistical significance, the chi-square test stands out among other alternative measures of fit in SEM. A p-value can be computed using the model degrees of freedom and the chi-square value. This tests the null hypothesis that the observed data and anticipated model are equal, which shouldn't be statistically significant if the model fits the data well.

For example, consider testing the following hypothesis

$$\begin{aligned} H_0 &: \theta \in \Omega_0 \\ H_1 &: \theta \in \Omega_1 \end{aligned} \tag{5.8}$$

In Equation 5.8 Ω_0 represents the parameter space under null model whereas Ω_1 represents the parameter space under data model. If L_0 is the maximum likelihood function from equation 5.6 under null hypothesis and L_1 is the maximum likelihood function from Equation 5.6 under alternative hypothesis then the LR statistic can be defined by,

$$W = -2\log\left(\frac{L_0}{L_1}\right) \geq 0$$

If r_0 is the number of free parameters under the null hypothesis and let r is the number of free parameters under the alternative hypothesis, then for large sample size n, the statistic $W \sim \chi^2_{v=r-r_0,\alpha}$. For given $\alpha \in (0,1)$, $\chi^2_{v,\alpha}$ denotes the α level critical value of χ^2_v, the chi-square random variable or the chi-square distribution with v degrees of freedom. Here the α level is the probability of rejecting the null hypothesis when the null hypothesis is true. The standard is the α level, set of 0.05. The p-value is the smallest value for which the null hypothesis can be rejected. It indicates how extreme the data are. We compare the p-value with the α to determine whether the observed data are statistically significantly different from the null hypothesis. If the p-value is less than or equal to the α (p $<$.05), then we reject the null hypothesis, and we say the result is statistically significant. If the p-value is greater than α (p $>$.05), then we fail to reject the null hypothesis, and we say that the result is statistically insignificant ([88]). However, the χ^2 statistic is very sensitive to sample size, and we have to be very careful reporting this statistic, at the same time compare other Parsimony Adjusted Indices (CFI) and Predictive Fit Indices (SRMR).

5.3 Model Validation for Machine Learning Models

Model validation is a step that comes after model training and involves comparing the trained model to a test set of data. A portion of the training set's data set may or may not be included in the testing data. Any machine learning model's ultimate objective is to learn from examples in a way that allows it to generalize its knowledge to situations it has never encountered before. Finding the appropriate machine learning technique to build our model is crucial when we approach a problem with a dataset in hand.

5.3.1 Cross-Validation

In cross-validation, a subset of the sample would ideally be kept aside and used to evaluate the effectiveness of a prediction model, given that we have enough data. Based on the size of the subset, the validation method is named. For instance, a K=10 fold cross validation implies the training data was roughly divided into ten equal parts and passed through a model trained by the rest of the K-1 part of the data. For a data divided into K approximately equal sections, say $\kappa : \{1,\ldots,N\} \mapsto \{1,\ldots,K\}$ be an indexing function that indicates the partition to which observation i is allocated by the randomization. Denote by $\hat{f}^{-k}(x)$ the fitted function, computed with the k the part of the data removed. Then the cross-validation estimate of prediction error is

$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right)$$

Typical choices of K are 5 or 10. We used ten-fold cross-validation in this study whenever it was needed. Hyperparameter tuning is often done using cross-validation. For example, given a set of models $f(x,\alpha)$ indexed by a tuning parameter α, denote by $\hat{f}^{-k}(x,\alpha)$ the α the model fit with the k the part of the data removed. Then for this set of models the following function $\mathrm{CV}(\hat{f},\alpha)$ provides an estimate of the test error curve, and tuning parameter $\hat{\alpha}$ can be found by minimizing this function.

$$\mathrm{CV}(\hat{f},\alpha) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i,\alpha)\right)$$

5.3.2 Evaluation Metrics

An evaluation metric measures the performance of a model after training. You build a model, get proper feedback from the model is expected. Therefore, picking the appropriate metric when assessing an ML model is crucial. Selecting only one statistic may not always yield the greatest results; instead, it may be beneficial to combine several metrics. We used the following metrics to evaluate our implemented ML models.

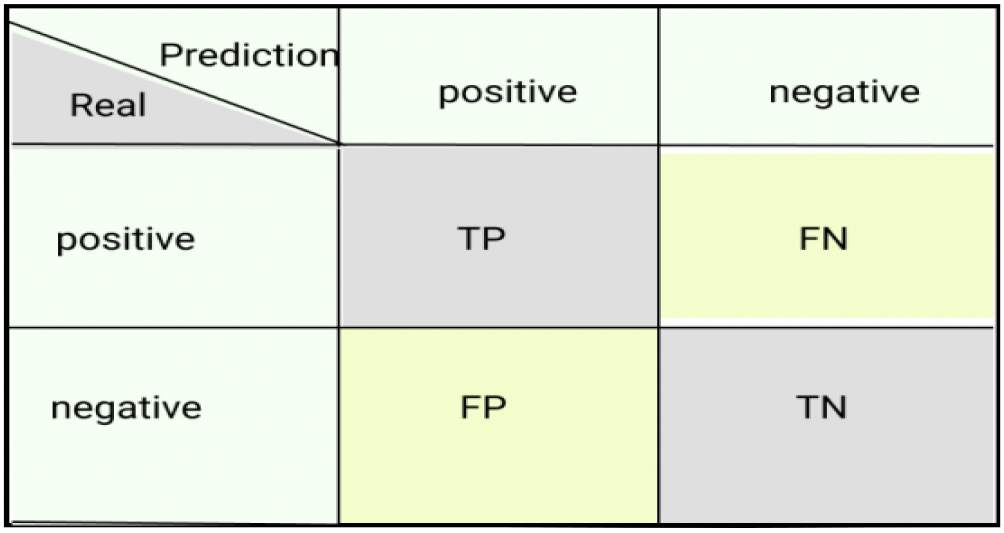

Figure 5.5: A general example of a confusion matrix.

5.3.2.1 Confusion Matrices. A table describing the distribution of classifier performance on the data is called a confusion matrix. It is a N $\times$ N matrix used to assess how well a classification model is working. It demonstrates the model's effectiveness and what areas require improvement. Consider the Figure 5.5.

Here:

- **TP:** true positive (the correctly predicted positive class outcome of the model),
- **TN:** true negative (the correctly predicted negative class outcome of the model),
- **FP:** false positive (the incorrectly predicted positive class outcome of the model),
- **FN:** false negative (the incorrectly predicted negative class outcome of the model).

5.3.2.2 Accuracy. Accuracy is a statistic that summarises how well a classification task was performed by dividing the total number of accurate predictions by the total number of predictions the model made. It is the proportion of all the data points that were successfully predicted.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FN+FP+TN)}$$

5.3.2.3 Balanced Accuracy. Both binary and multi-class classification uses balanced accuracy. Its application is when working with imbalanced data or when one of the target groups shows up much more frequently than the other. It is the arithmetic mean of sensitivity and specificity.

$$\text{Balanced Accuracy } = \frac{\text{sensitivity } + \text{ specificity}}{2}$$

Sensitivity is the proportion of actual positives that are accurately predicted out of all positive predictions produced by the model is measured by this, also known as the true positive rate or recall.

$$\text{Sensitivity } = \frac{TP}{TP+FN}$$

Specificity is the estimates of the ratio of correctly detected negatives to all of the model's predicted negative outcomes and is also referred to as the true negative rate.

$$\text{Specificity } = \frac{TN}{(TN+FP)}$$

5.3.3 Creating Training and Test data

Having representative class measurements i.e. uniform distribution over each target category is the core of ML performance. An ideal situation would be a 1:1 class ratio. However, it is not the case in the real world, and when it doesn't happen and any of the classes has small number of observation in it, it's called a class imbalance problem [89]. ML researchers had and still are hugely suffering from this class imbalance problem [90, 91, 92, 93]. However, to tackle this issue, several techniques have been developed. For example, in the deep learning field, data augmentation (axis rotation, contrast, sharpness modification in image data) is already a widely used established technique. But, due to the nature of our data, we will use force sampling or "SMOTE: synthetic minority over-sampling technique [94]" to overcome this issue. In this method, we will randomly choose 20-30 percent of the positive class sample into the test data set.

CHAPTER VI

RESULTS

6.1 Descriptive Statistics

We used National Health and Nutrition Examination Survey (NHANES) dataset to investigate the obesity problem among 4-6yo U.S. children. Given the study structure, our data set includes the latest data we could find at the time of this study (2017-2018) and two years before that. Therefore, we had 1210 observations from the cycle years "2013-2014", "2015-2016," and "2017-2017." Table 6.1 gives us a quick overview of the data. It reveals that there is no missing value available in the data. The five-number summary also tells us the fiber intake is the lowest among all dietary variables, which is 12.89 gm. In the data there was around 23 percent (281) obese children and 51% (617) female and 49% male(593). Three of the most prevalent ethnic groups in the United States were represented in the study population: non-Hispanic whites made up 30% (367) of the population, non-Hispanic blacks accounted for 23% (279), Hispanics made up 29% (348), and other racial origins made up about 18% (216).

The correlation plot in Figure 6.1 shows that there are some degrees of multicollinearity in the data. For example, fiber intake is correlated with the carb, sugar, and fat intake having Pearson correlation coefficients of 0.63, 0.34, and 0.43, respectively. Overall the nutrient intake seems to be correlated with each other. The rest of the variables look fine with lower correlation coefficients. Also, we did the χ^2 association test between gender and education. That came insignificant with $\chi^2 = 2.314$, df = 3, and p-value = 0.314. This implies there is not enough evidence to say that gender and education is not associated with each other. Moreover, if we look at Figure 6.2 it looks like all the numerical variable are approximately normally distributed except SES. It is showing a

Table 6.1: Descriptive statistics for the variables included in the path analyses for the entire study cohort. five number summary and SD was reported for continuous variables and and the frequency and percentage (%) are reported for categorical variables.

Age	Gender	Race	Education	SES	Fiber	Carb	Sugar	Fat	BMI	PA	Obesity
Min: 4.0	Male: 593 (49%)	Hispanic: 348 (28.8%)	LHS: 227 (18.8%)	Min: 1.0	Min: 1.3	Min: 40.5	Min: 9.2	Min: 8.9	Min: 12.2	Not Act: 92 (7.6%)	Not Obese: 929 (76.8%)
1st Qu: 4.0	Female: 617 (51%)	NHW: 367 (30.3%)	HSC: 682 (56.4%)	1st Qu: 2.0	1st Qu: 9.1	1st Qu: 169.5	1st Qu: 71.6	1st Qu: 44.6	1st Qu: 15.1	Act: 1118 (92.4%)	Obese: 281 (23.2%)
Median: 5.0		NHB: 279 (23.1%)	CA: 301 (24.9%)	Median: 3.0	Median: 12.1	Median: 211.6	Median: 96.0	Median: 57.1	Median: 16.0		
Mean: 5.0		Other: 216 (17.9%)		Mean: 3.2	Mean: 12.9	Mean: 217.2	Mean: 100.1	Mean: 60.6	Mean: 16.6		
3rd Qu: 6.0				3rd Qu: 4.0	3rd Qu: 16.0	3rd Qu: 258.3	3rd Qu: 123.6	3rd Qu: 73.8	3rd Qu: 17.3		
Max: 6.0				Max: 5.0	Max: 43.7	Max: 542.7	Max: 292.6	Max: 225.1	Max: 33.4		
SD: 0.8				SD: 1.5	SD: 5.4	SD: 67.4	SD: 39.5	SD: 23.5	SD: 2.4		

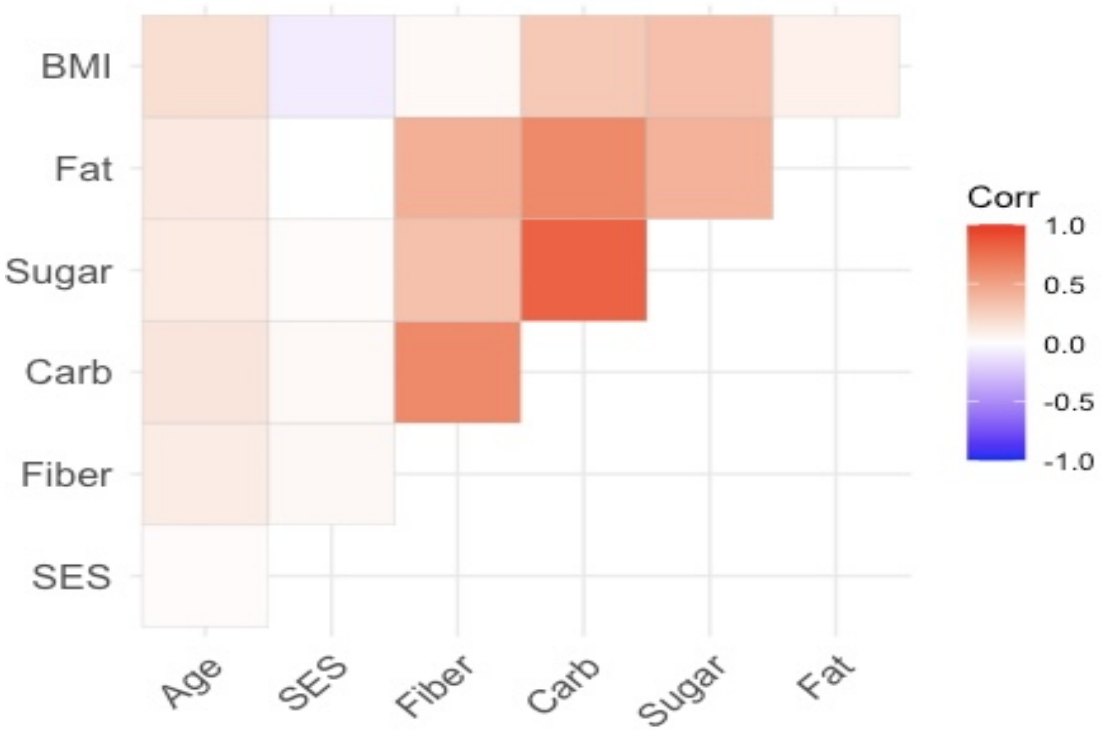

Figure 6.1: The correlation structure among numerical variables for the entire cohort.

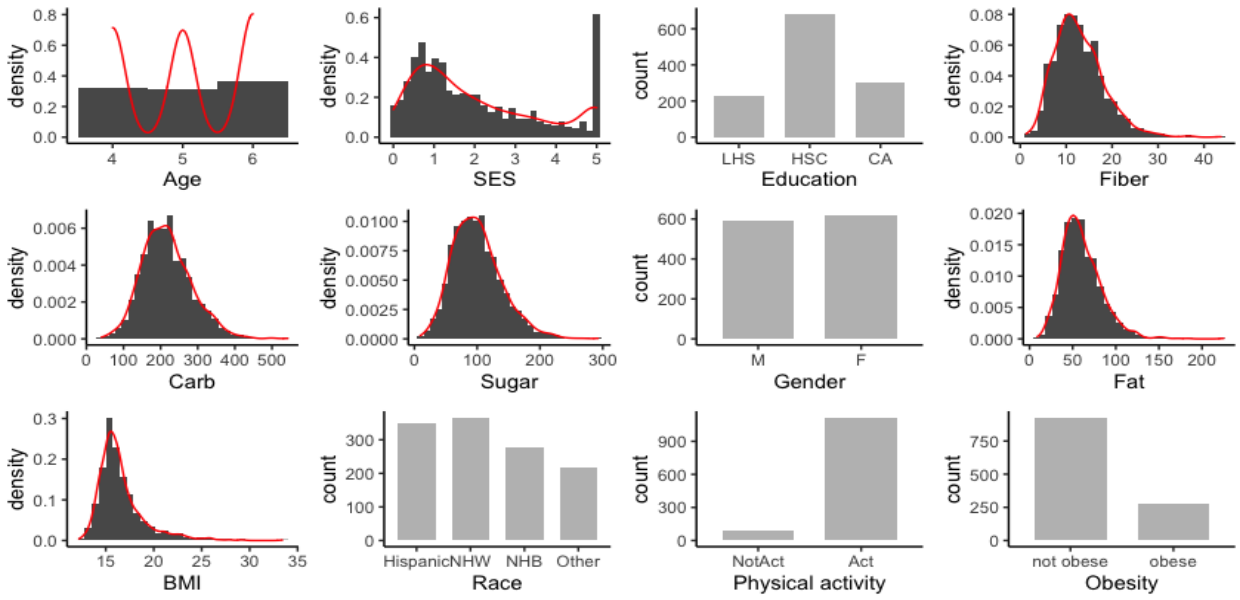

Figure 6.2: Density and bar plot of the variables that were used in the study.

slight right skewed pattern.

6.2 Path Analysis Results

6.2.1 Fit and Significance of Path Model

We will start this section by presenting the necessary assumption check and goodness-of-fit results. One of the critical assumptions in SEM path analysis is low multicollinearity among exogenous and endogenous variables. If we look at the correlation plot presented in Figure 6.1, we can see that there is some degree of multicollinearity among endogenous nutrition intake variables. Overall, the nutrient intake showed moderate multicollinearity with the maximum correlation between Carb and Sugar (0.82) with a variance inflation factor (VIF) of 1.258. Although there is no single cutoff point above which multicollinearity should be considered bad, there are widely accepted literature that suggests correlation coefficient < 0.8 and/ or VIF less than 5 or 10 depending on the situation [95, 96].

Regarding the normality assumption concern, a quick look at the distribution curve in Figure 6.2 confirms the normality of continuous variables in the data. However, a more thorough multivariate normality check was also performed. The Q-Q-plots of the Mahalanobis distance

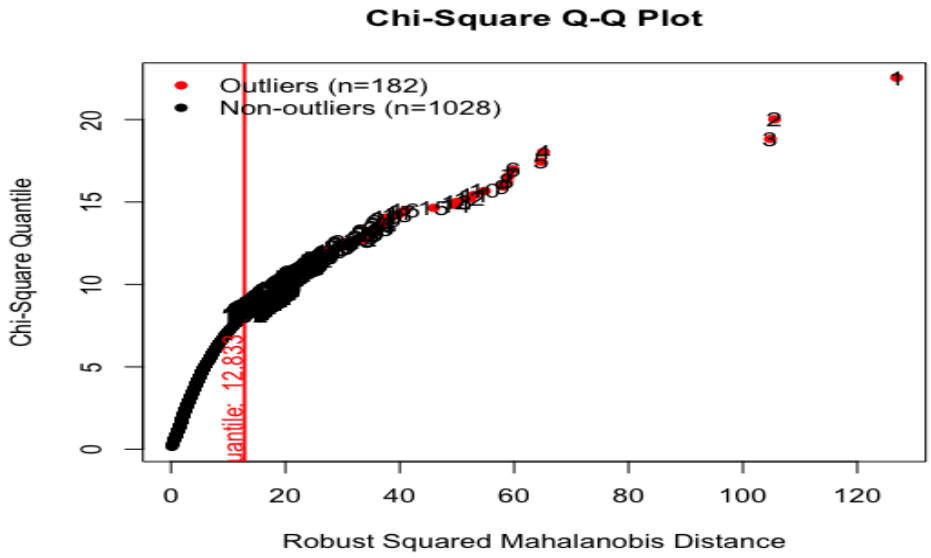

Figure 6.3: Q-Q plot of ordered Mahalanobis distances versus estimated chi-square quantiles (percentiles) to check the multivariate normality of the data.

is a widely accepted test for multivariate normality [97]. For multivariate data, we plotted the ordered Mahalanobis distances versus estimated quantiles (percentiles) for a sample of size n from a chi-squared distribution with degrees of freedom p in Figure 6.3. This should resemble a straight line for data from a multivariate normal distribution. Outliers will show up as points on the upper right side of the plot for which the Mahalanobis distance is notably greater than the chi-square quantile value.

After checking the assumption, we used T-Rule for the proper identification of the model. Here T is the number of free and unconstrained parameters. T has an upper bound defined by the number of exogenous and endogenous variables [T< 0.5 (p+q)*(p+q+1)] [74]. For our data, the upper limit of T was found to be 78. We also carefully looked for the under-identification of the SEM model. We used the Parsimony Adjusted Indices (CFI) to evaluate the fit of our path analysis model. Larger numbers indicate better fit, while CFI values range from 0 to 1. As we discussed in section 5.1.6 that a CFI value of.90 or higher could indicate a good model fit, our model's CFI was found to be 0.909. We reported standardized root mean square residual (SRMR) to investigate the model fit further. Our model's SRMR value was found to be 0.088, around the acceptable range of

Table 6.2: Table of the difference between the observed and implied covariance matrix of fitted model.

Variables	Fiber	Carb	Sugar	Fat	BMI	Age	Gender	Race	Education	SES	PA
Fiber	0.000										
Carb	0.307	0.310									
Sugar	0.339	0.200	0.000								
Fat	0.417	0.369	0.382	0.000							
BMI	0.114	0.076	0.060	0.010	0.001						
Age	0.000	0.000	0.000	0.000	0.000	0.000					
Gender	0.000	0.000	0.000	0.000	0.000	0.000	0.000				
Race	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000			
Education	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		
SES	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
PA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

less than 0.1.

Finally, we reported another essential measure which is the residuals covariance matrix. Table 6.2 shows the residuals of our fitted model. This is simply the difference between the observed and implied covariance matrix. An ideal situation would be all 0 elements. The table implies a good fit with fewer nonzero entries in nutrition intake cells.

The path diagram in Figure 6.4 shows the significant paths, and Table 6.3 shows the estimated significant path coefficients. Physical activity had a significant direct causal effect on BMI but did not have a significant another mediator. Through nutrition intake, physical activity had a total indirect effect of -0.004, but in total, it had a total effect of -1.168 ($\rho_1 = -0.004$ and $\rho_2 = -1.64$), decreasing BMI. If we held this indirect effect constant, then the resulting path coefficient -1.164 represents the direct effect of physical activity on reducing BMI. We can conclude that physical activity alone affects BMI more than other mediators. Age had increased BMI with a total causal effect of $\rho = 0.266$, which included a direct effect of $\rho = 0.232$ and a total indirect effect of $\rho = 0.035$ through nutrition intake. Similarly, a higher education level (CA) caused a decrease in BMI measurement with a direct effect of $\rho = -0.706$ but had a total indirect effect of $\rho = 0.08$ through nutrition intake, resulting in a total causal effect ($\rho = 0.627$) on BMI. Socioeconomic status has a direct impact of -0.036 on decreasing BMI but had a really low total indirect effect through nutrition intake ($\rho = 0.001$), resulting in a total decreasing effect of ($\rho = -0.035$) on BMI.

Table 6.3: Significant paths and summarized causal effects based on the path analysis for the entire population. A bold value indicates the total effect (both direct and indirect) on BMI.

Paths	**Intermediate Path Coefficients**		**Total Direct/ Indirect Effects**
	ρ_1	ρ_2	
Age			
(Age –> Fiber)			0.117
(Age –> Sugar)			0.132
(Age –> Fat)			0.144
SIE (Age –> Fiber –> BMI)	0.117	-0.089	-0.01
SIE (Age –> Carb –> BMI)	0.161	0.109	0.017
SIE (Age –> Fat –> BMI)	0.144	0.179	0.026
TIE (Age –> Nutrient Intake –> BMI)			0.035
TE (Age –> BMI)			**0.266**
Gender			
(Gender –> Fiber)			-0.154
(Gender –> Carb)			-0.219
(Gender –> Sugar)			-0.132
(Gender –> Fat)			-0.171
SIE (Gender –> Fiber –> BMI)	-0.154	-0.089	-0.031
SIE (Gender –> Carb –> BMI)	-0.219	0.109	-0.031
SIE (Gender –> Fat –> BMI)	-0.171	0.179	-0.031
TIE (Gender –> Nutrient Intake –> BMI)			-0.043
TE (Gender –> BMI)			**0.191**
Race			
Hispanic			
(Hispanic –> BMI)			0.329

Table 6.3 continued.

Paths	Intermediate Path Coefficients		Total Direct/ Indirect Effects
	ρ_1	ρ_2	
(Hispanic –> Fiber)			0.307
SIE (Hispanic –> Fiber –> BMI)	0.307	-0.089	-0.027
TE (Hispanic –> BMI)			**0.304**
Non Hispanic Black (NHB)			
(NHB –> Fiber)			0.156
(NHB –> Carb)			0.19
SIE (NHB –> Fiber –> BMI)	0.156	-0.089	-0.014
SIE (NHB –> Carb –> BMI)	0.19	0.109	0.021
SIE (NHB –> Fat –> BMI)	0.107	0.179	0.019
Other Race			
(Other Race –> BMI)			-0.383
(Other Race –> Sugar)			-0.316
(Other Race –> Fat)			-0.103
SIE (Other Race –> Fat –> BMI)	-0.103	0.179	-0.018
TE (Other Race –> BMI)			**-0.421**
Education			
High school and Some Collage (HSC)			
(HSC –> Fiber)			-0.384
(HSC –> Sugar)			0.347
(HSC –> Fat)			0.251
SIE (HSC –> Fiber –> BMI)	-0.384	-0.089	0.034
SIE (HSC –> Carb –> BMI)	0.119	0.109	0.013
SIE (HSC –> Fat –> BMI)	0.251	0.179	0.045

Table 6.3 continued.

Paths	Intermediate Path Coefficients		Total Direct/ Indirect Effects
	ρ_1	ρ_2	
TIE (HSC –> Nutrient Intake –> BMI)			0.098
TE(HSC –> BMI)			**-0.069**
Collage and Above (CA)			
(CA –> BMI)			-0.706
(CA –> Sugar)			0.348
(CA –> Fat)			0.248
SIE (CA –> Carb –> BMI)	0.206	0.013	0.022
SIE (CA –> Fat –> BMI)	0.248	0.179	0.044
Physical Activity (PA)			
(PA–> BMI)			-1.164
SIE (PA–> Fiber –> BMI)	0.112	-0.089	-0.01
TE (PA –> BMI)			**-1.168**

Among the four nutritional variables included in the path analysis (Fiber, Carb, Sugar, Fat), independently, Carb, Sugar, and Fat had an increasing effect on BMI with respective path coefficients of (ρ = 0.109, 0.018, and 0.179), but Fiber intake has a decreasing impact of -0.089 on BMI.

To investigate the difference in the causal structure between males and females, we performed separate path analyses (see Figure 6.5 and 6.6) on gender-specific samples. Our analysis reveals that physical activity directly affecting BMI for both males and females. In females, age played a more significant role than females affecting BMI. In males, age only affected BMI negatively via fiber, causing lower BMI with high fiber intake, whereas in males, age significantly affected BMI both directly and indirectly. In female sample, Non-Hispanic Black had negative (ρ_1 = 0.313, ρ_2 = 0.011) and positive (ρ_1 = 0.150, ρ_2 = 0.213) mediated through fiber and carb respectively

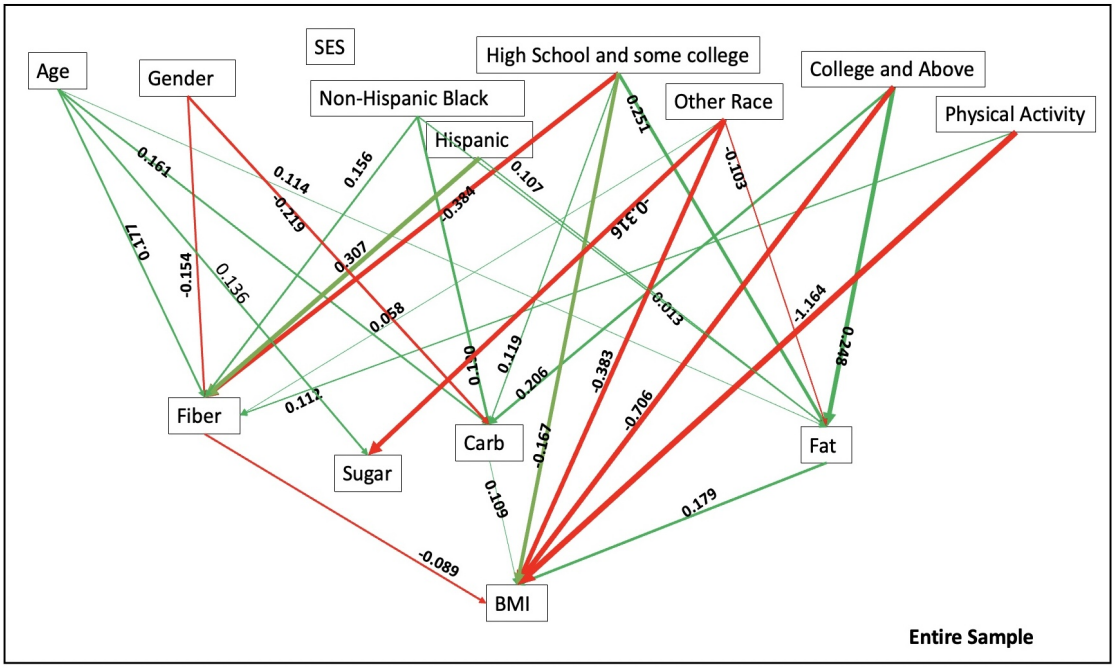

Figure 6.4: Path diagram for the entire sample. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

but our study did not find any significance in male sample (see Figure 6.5, 6.6). Education also played an important role in affecting BMI both in male and female children. In females, higher education of children's household reference members (usually parents) showed a negative causal effect on BMI through higher fiber intake. In contrast, the male had an opposite ($\rho = 0.268$) effect through the increased fat intake. Among Hispanic males and females, higher fiber intake caused decreased BMI, but only the female Hispanic cohort had direct causation to BMI (see 6.5, 6.6)).

We performed further gender-ethnicity-specific analysis To investigate these findings further. Based on the path analysis of different gender-ethnicity-specific subpopulations, a common significant nutritional variable was found to be carb intake (see Figure 6.7 and 6.8). No significant direct causation was found in Hispanic females, but in males, physical activity and higher education were found to be directly affecting BMI. In Non-Hispanic black children, age and sugar were found to be important for both males and females for increased BMI but through a different ancestor. As Figures 6.9 and 6.10 show, in Non-Hispanic Black females, education positively mediated through sugar,

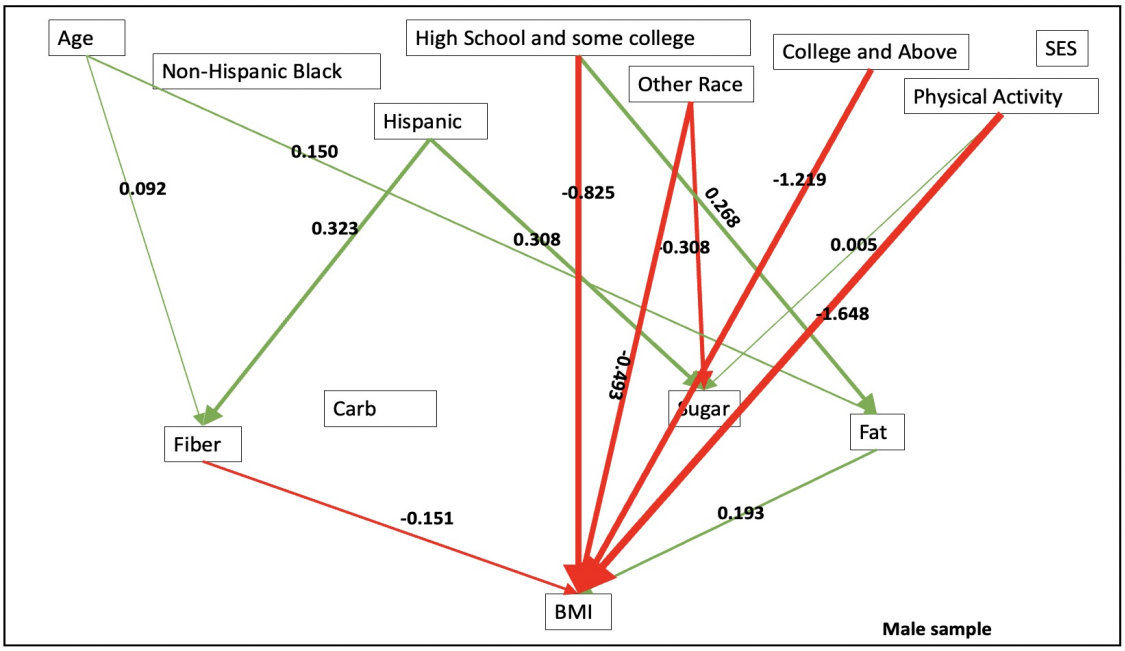

Figure 6.5: Path diagram for the males only. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

causing increased BMI. Still, in males, age was causing higher sugar intake resulting in higher BMI. Among other racial group (6.11, 6.12) socioeconomic status found to be directly affecting BMI (ρ = -0.291) in males in female age showed direct positive causation (ρ = 0.822). However, female socioeconomic status caused increased fat intake but did not significantly mediate BMI.

6.3 Predictive Results

This section will show how we implemented several ML methods and their results. Each model was trained using the same training data set we created at the beginning of the analysis. Later trained models were validated using testing data. Training and testing data was created following the method described in section 5.3.3. In our case, we used a 70:30 percent training-test split. And around 30 percent of the testing sample was chosen from the positive class (obese). By doing so, we had an 847: 363 training test split. Among the test data, 107 cases were from the obese class.

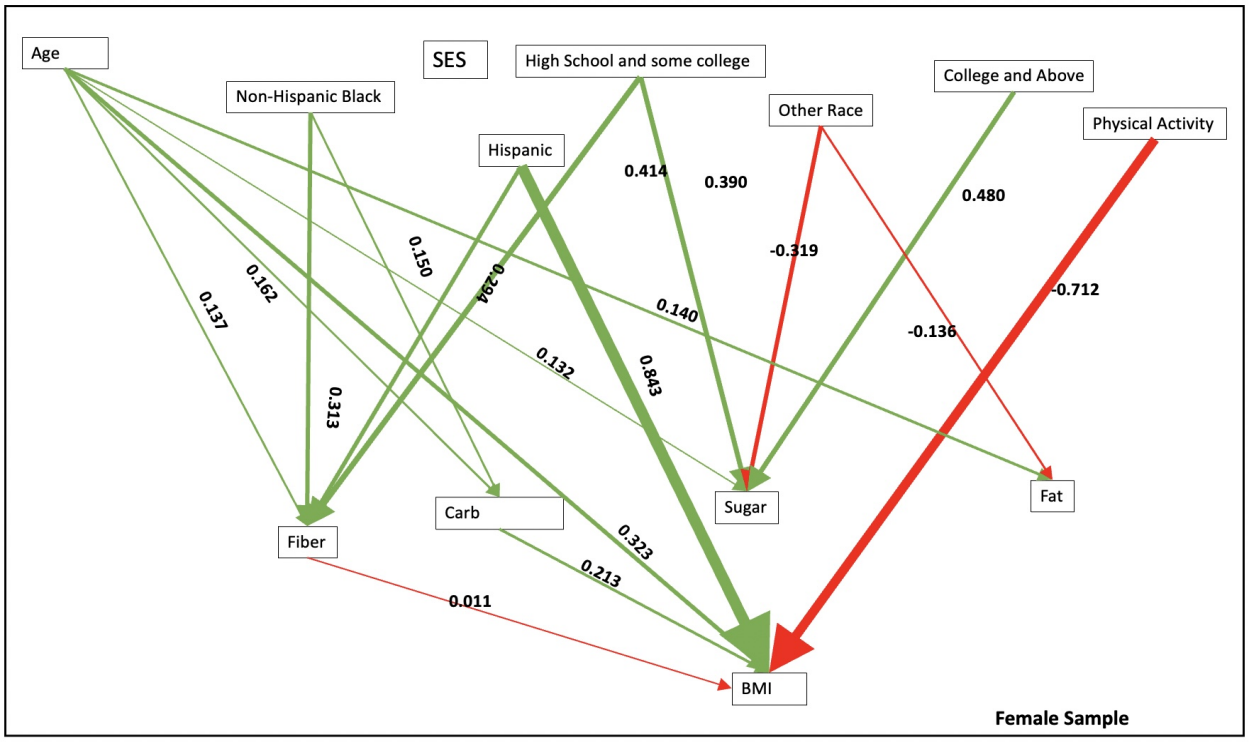

Figure 6.6: Path diagram for the females only. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

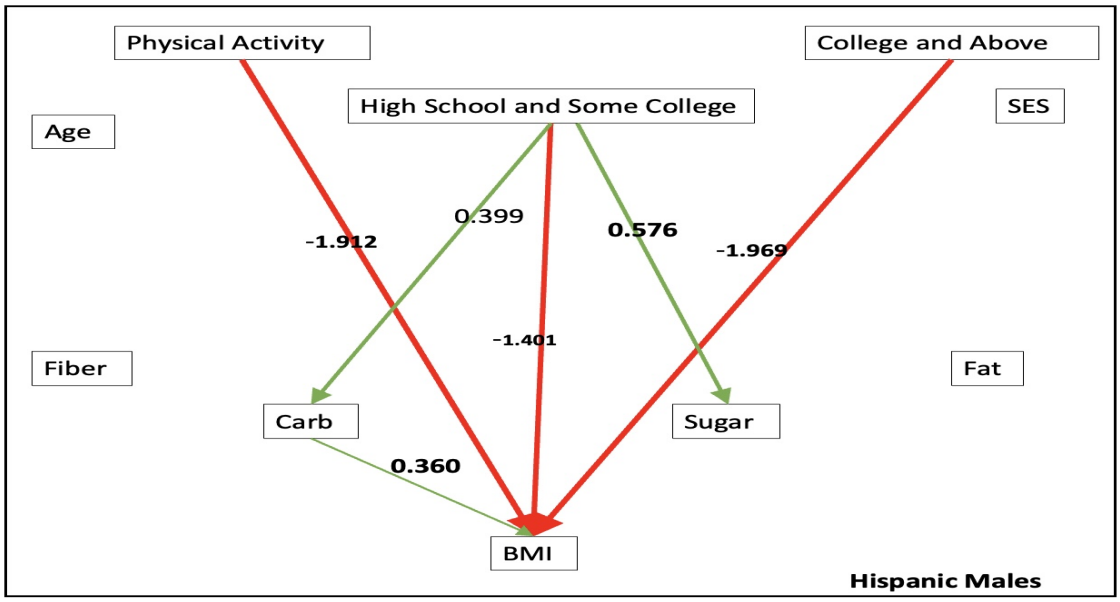

Figure 6.7: Path diagram for the Hispanic males. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

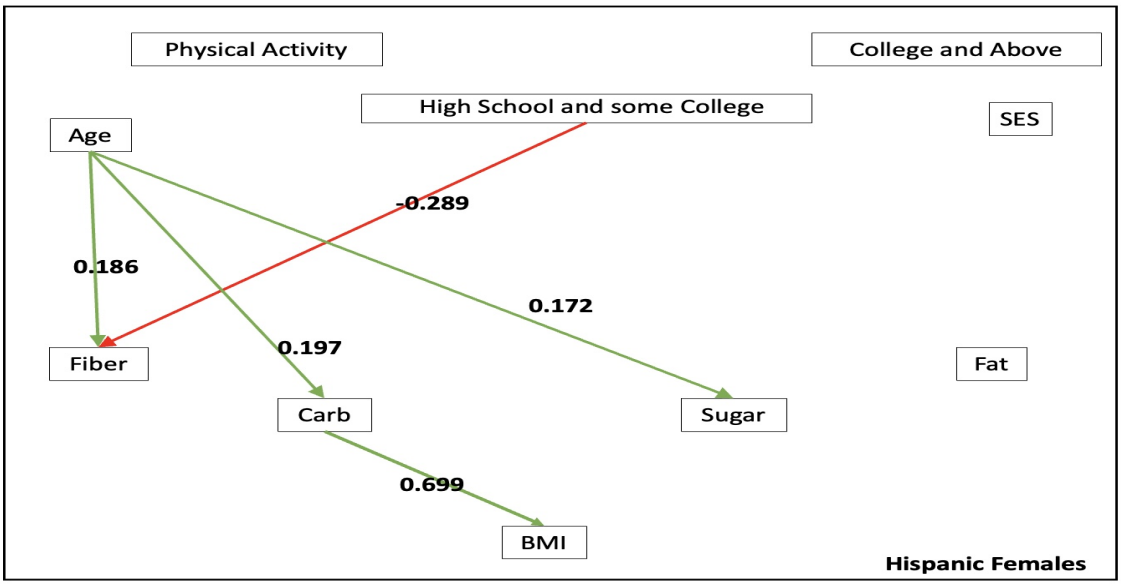

Figure 6.8: Path diagram for the Hispanic females. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

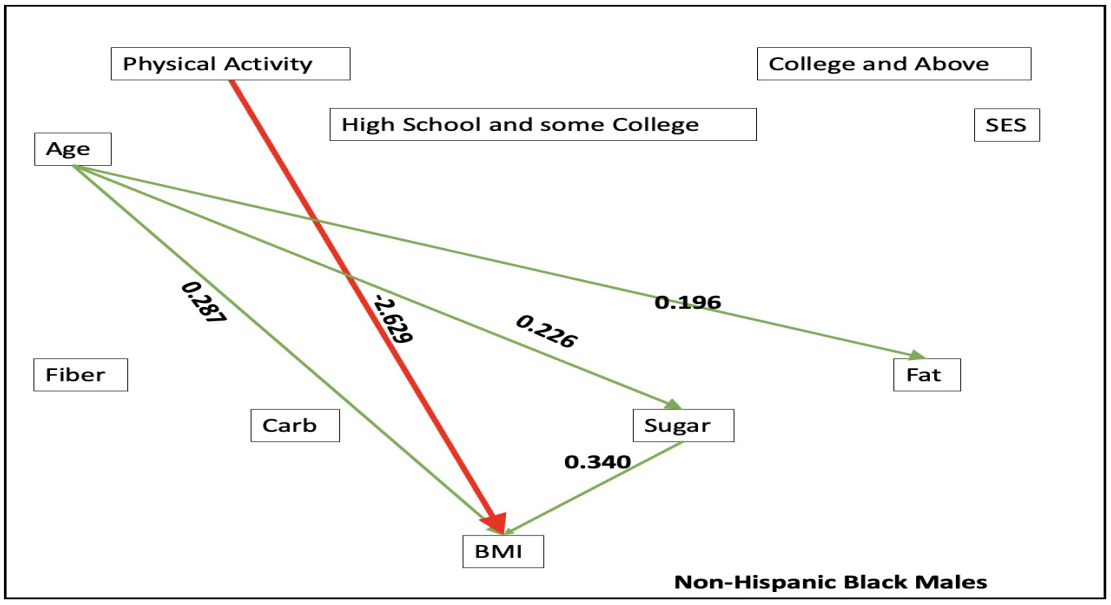

Figure 6.9: Path diagram for the Non-Hispanic Black males. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

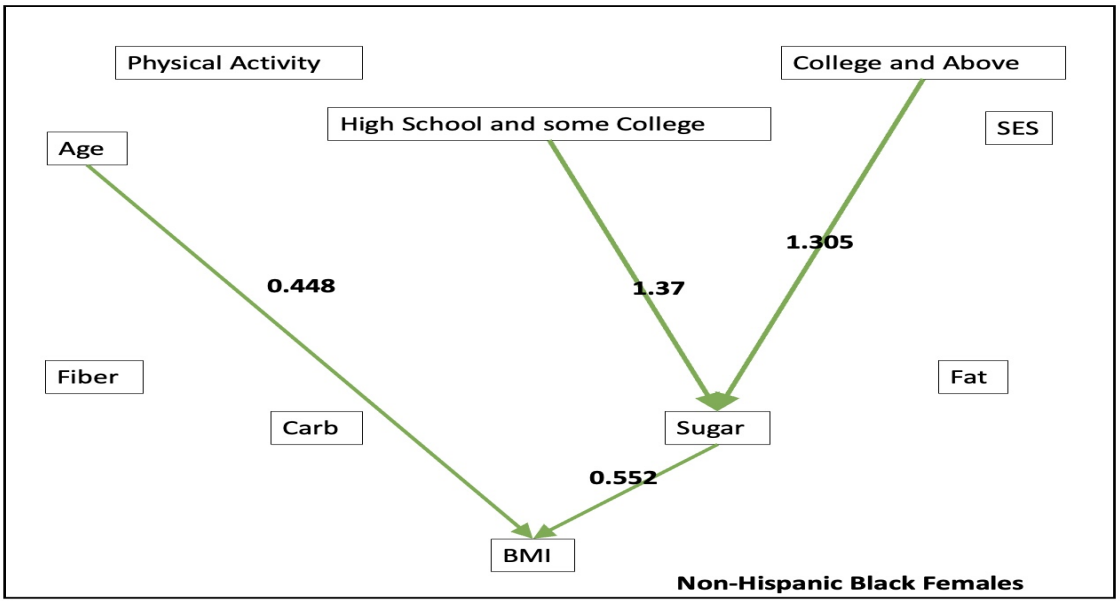

Figure 6.10: Path diagram for the non-Hispanic Black females. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

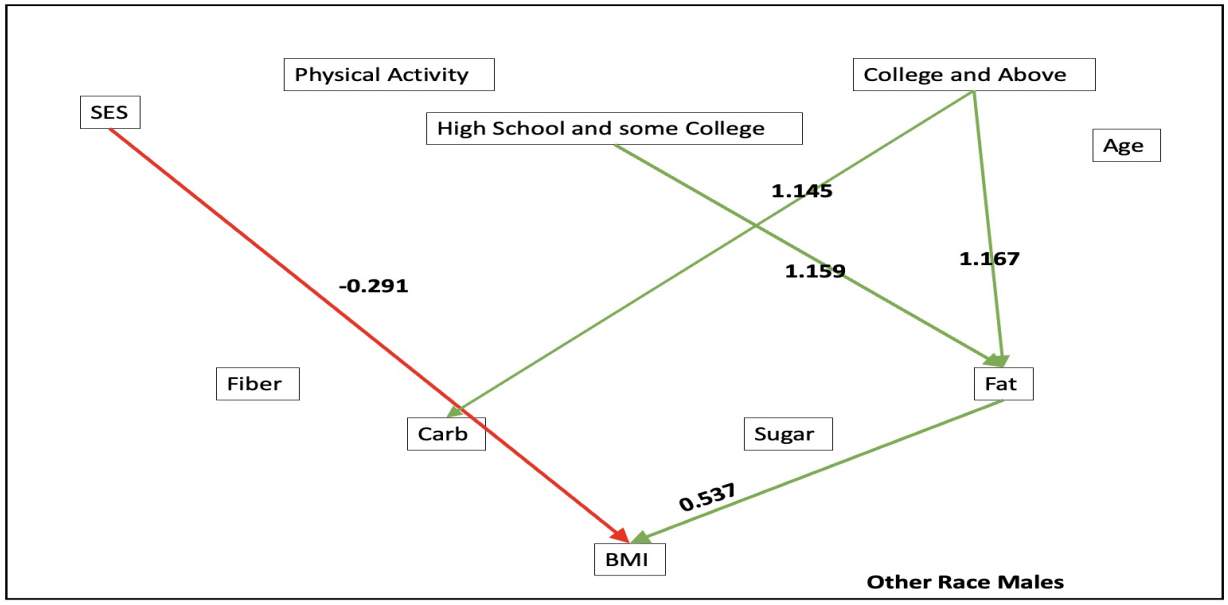

Figure 6.11: Path diagram for the Other Racial males. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

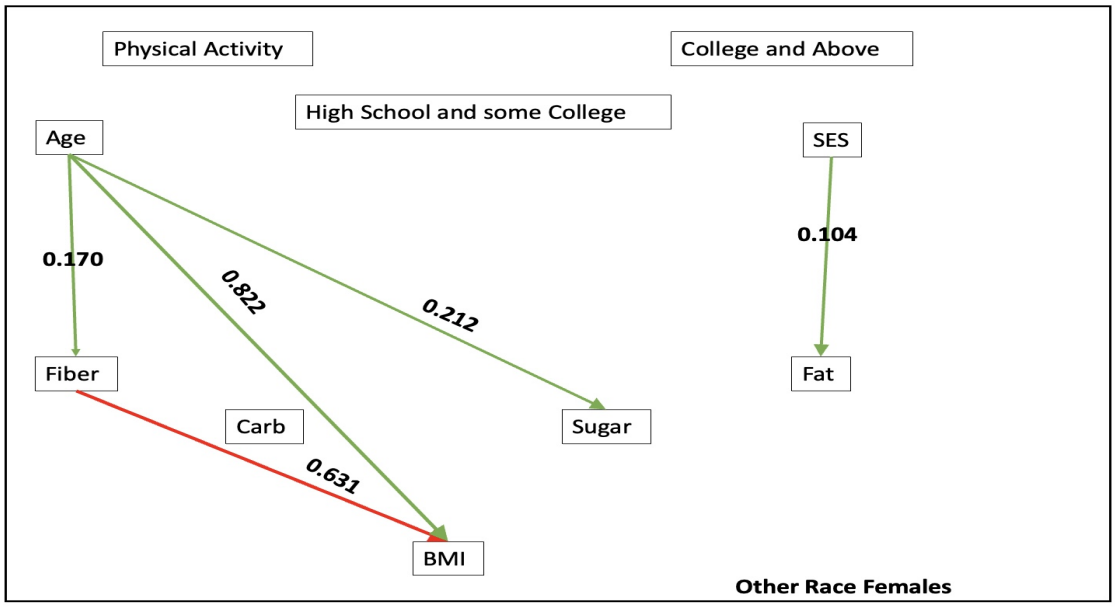

Figure 6.12: Path diagram for the Other Racial females. Paths of green indicate positive path coefficients, while paths of red line indicate negative path coefficients. The widths of the paths are related to the absolute values of path coefficients, and the wider of the line is the stronger causation. Only significant paths are shown in the diagram.

Table 6.4: The confusion matrix and accuracy of using principle component regression using test data.

Confusion Matrix			
Prediction/Reference label	**Not Obese**	**Obese**	**Total**
Not Obese	242	28	270
Obese	14	79	93
Total	256	107	363
Accuracy: 0.8843 95% CI: (0.8468, 0.9153); P-value: 0.000			
Sensitivity: 0.7383; Specificity: 0.9453			

6.3.1 Principle Component Regression Results

In this approach, we chose several linear combinations of our regressor (component). Because we can see from Figure 6.1 that there is low to moderate multicollinearity in our data, using a combination of colinear features will less number of components might perform better. As Table 6.5 below, a single component can explain 25 percent variability in obesity. Trends keeps improving until the 7th component. After that additional component does not capture any extra variability in the data given that our main purpose is to reduced the dimentionality as much as possible with lower cross validation error. Because, the standard approach here is to look for a low cross-validation error with a lower number of components than the number of variables in our dataset. If this is not the case or if the smallest cross-validation error results with a number of components close to the number of variables in the original data, then there is no dimensionality reduction. Additionaly, the test mean squared error(MSE) in crossvalidation starts with 0.4 when the first principal component is included followed by one point decrease in MSE when the second primary component is included. And after the 7th component no additional improvement is noticed.

Figure 6.13 shows a steep decline in validation error until the 7th component, and then the decrement is not much significant which is inclined with the previous findings. So we conclude that seven components are enough to explain more than 81% of the variation in the data, even the CV score is a little higher than with 7th or 8th components. Finally, note that ten components explain all the variability in obesity as expected.

Table 6.5: Cumulative percentage of training variance explained by several component in pcr.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
CV	0.403	0.401	0.379	0.363	0.317	0.311	0.309	0.272	0.272	0.272
adjCV	0.403	0.401	0.379	0.364	0.312	0.310	0.311	0.272	0.272	0.272
Var Explained	24.877	39.900	49.430	58.720	67.780	76.100	83.720	89.840	94.670	99.200
Obesity	1.365	12.300	18.980	40.400	41.870	42.070	55.030	55.530	55.570	55.570

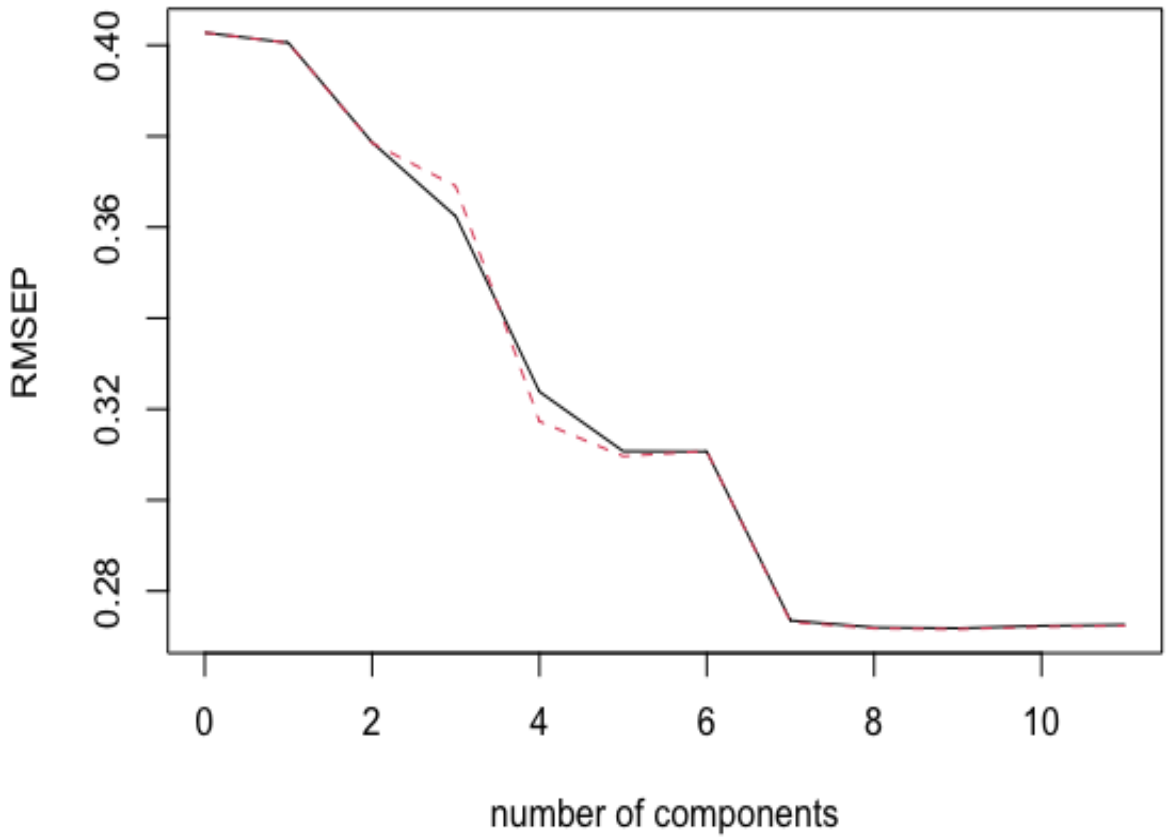

Figure 6.13: Mean square error vs number of component plot of PCR model.

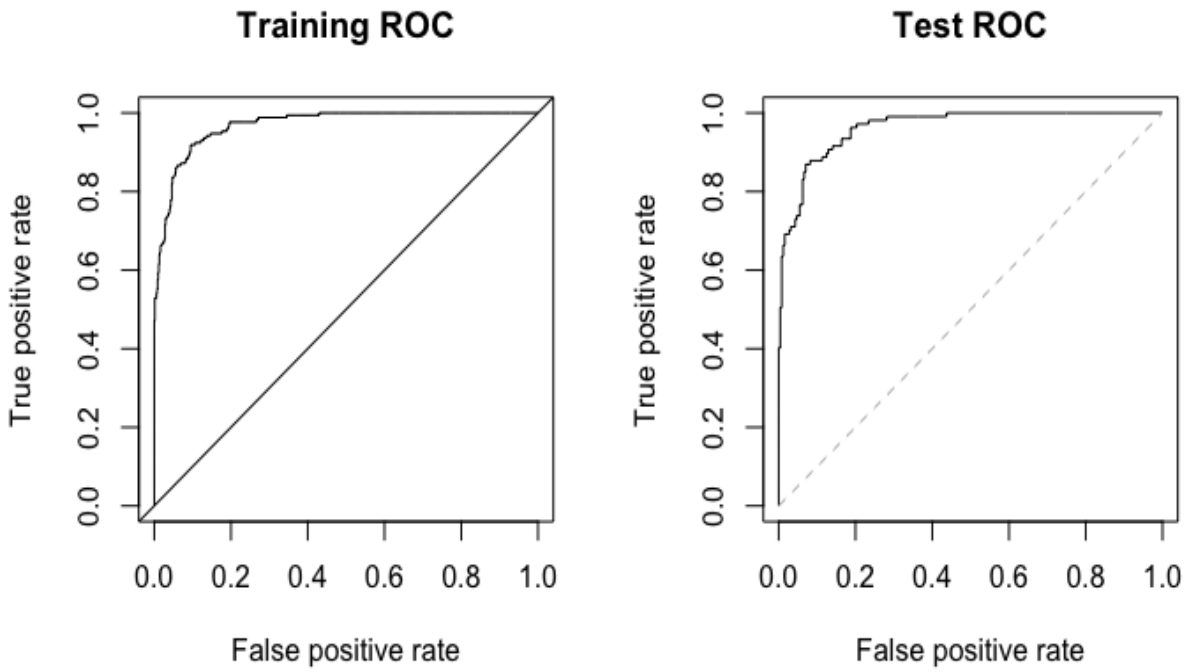

Figure 6.14: ROC curve of PCR model when applied on test and training dataset.

Therefore, using the first seven components, we proceed further and train and test our model. The Table 6.4 summarize the testing results. It reveals that the trained model is accurately predicting 321 observations; that is, the accuracy is around 88 percent. But As we discussed in section 5.3.2 it is better to use the balanced accuracy, and the balance accuracy is 84 percent.

6.3.2 Support Vector Machine(SVM) Results

To implement a support vector machine to our data, we used the e1071 library in R, which contains implementations for several statistical learning methods. In particular, the svm() function was used to fit a support vector classifier with the argument kernel="linear" was used. A cost argument in the library allows us to specify the cost of a violation to the margin. But before we did our final modeling, several pilot models were run on the training data for multiple cost ranges with linear kernel. 10-fold cross-validation was considered to find out the best-performing cost parameter. The cost and associated error rates are shown in Table 6.6. It shows that the minimum cost 0.211 is associated with the cost value 1. Figure 6.15 shows how error rate is changing with

Table 6.6: Hyper parameter estimation of SVM using 10 fold cross validation.

Cost	0.001	0.010	0.100	1.000	5.000	10.000	100.000
Error	0.187	0.178	0.118	0.073	0.077	0.082	0.125

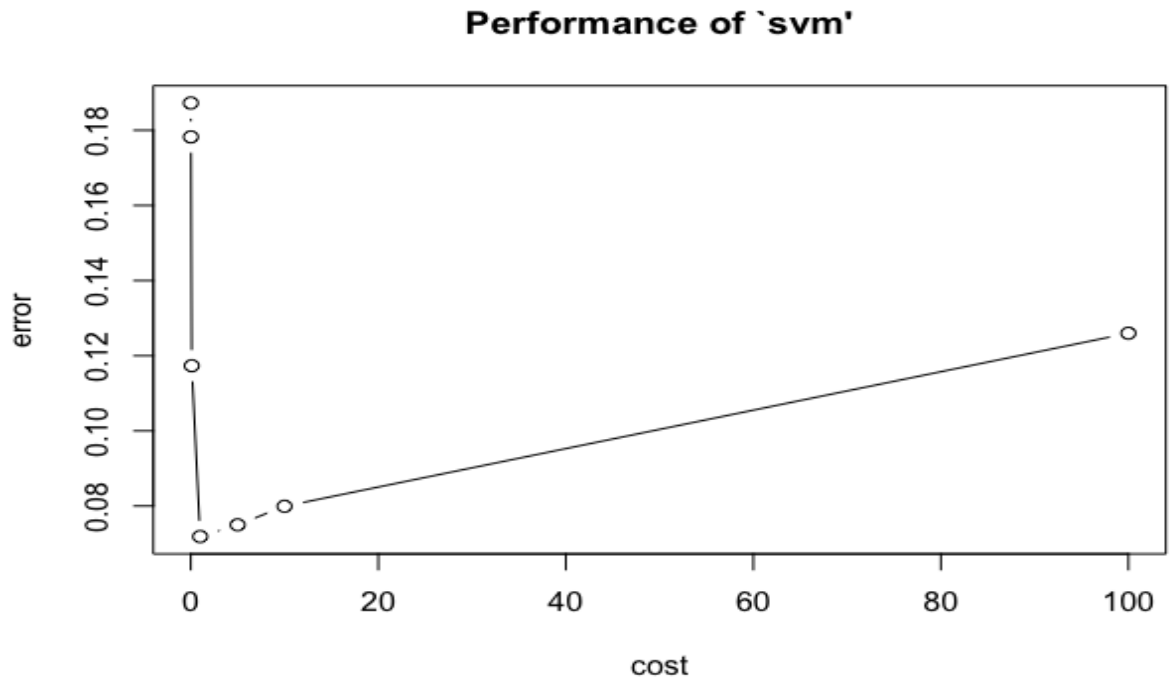

Figure 6.15: Cost vs error rate of SVM's hyper parameter estimation.

increment of costs.

For the optimum value of the cost parameter, we fit the support vector classifier to our data. Our model used 138 support vectors to draw the hyperplane between the two classes while using a "linear" kernel. This model performed really well on training data with a misclassification error rate of 0.068, but the performance decreased by approximately 4% when the model was applied to unseen test data (0.096). Detailed performance of the model is given in the confusion matrix Table 6.7. As our data is a class imbalance, we are interested in knowing whether our model's accuracy is higher than the percentage of the data that belongs to the class with the majority. And the significant p-value (0.0007) reveals that the 90% accuracy we have is better than no information rate. Again the balanced accuracy is showing an expectedly smaller accuracy of 85%. However, we tested

Table 6.7: The confusion matrix and accuracy of support vector machine.

Confusion Matrix			
Prediction/Reference label	**Not Obese**	**Obese**	**Total**
Not Obese	249	28	277
Obese	7	79	86
Total	256	107	363
Accuracy: 0.9036; 95% CI : (0.8685, 0.9319); P-Value : 0.0007			
Sensitivity: 0.7383; Specificity: 0.9727			

several other kernel options such as "polynomial," "radial bias," and "sigmoid." Even though some kernels performed better than the linear kernel during training, misclassification errors increased when encountered unseen test data.

In Figure 6.16 the ROC curve shows the true positive rate (TP), i.e., the number of correctly predicted positive class outcomes of the model against the false positive rate (FP), i.e., he incorrectly predicted the positive class outcome of the model. Using multiple cutoff probability of class allocation creates thousands of TP and FP. In our favorable scenario, we expect a well-performed model to show greater TP against the corresponding FP. According to the ROC curve, Our trained model is behaving well on the training data, but the model slightly deviates from the training when applied to testing data.

6.3.3 K-nearest Neighbors (KNN) Results

While doing KNN, we had to consider the fact that the scale of the variables matters because the KNN classifier determines the observations that are closest to a given test observation to predict the class of that observation. Any large-scale factors will have a significantly more significant impact on the distance between the observations and, thus, on the KNN classifier than small-scale variables. So, we standardize the variables that were larger in magnitude such as Carb, Sugar, and Fat (see the summary Table 6.1).

As we described in section 5.2.3 how KNN determines the class of the desired data point based on the K closest point from it. As a result, it might be mistaken to keep K larger than usual.

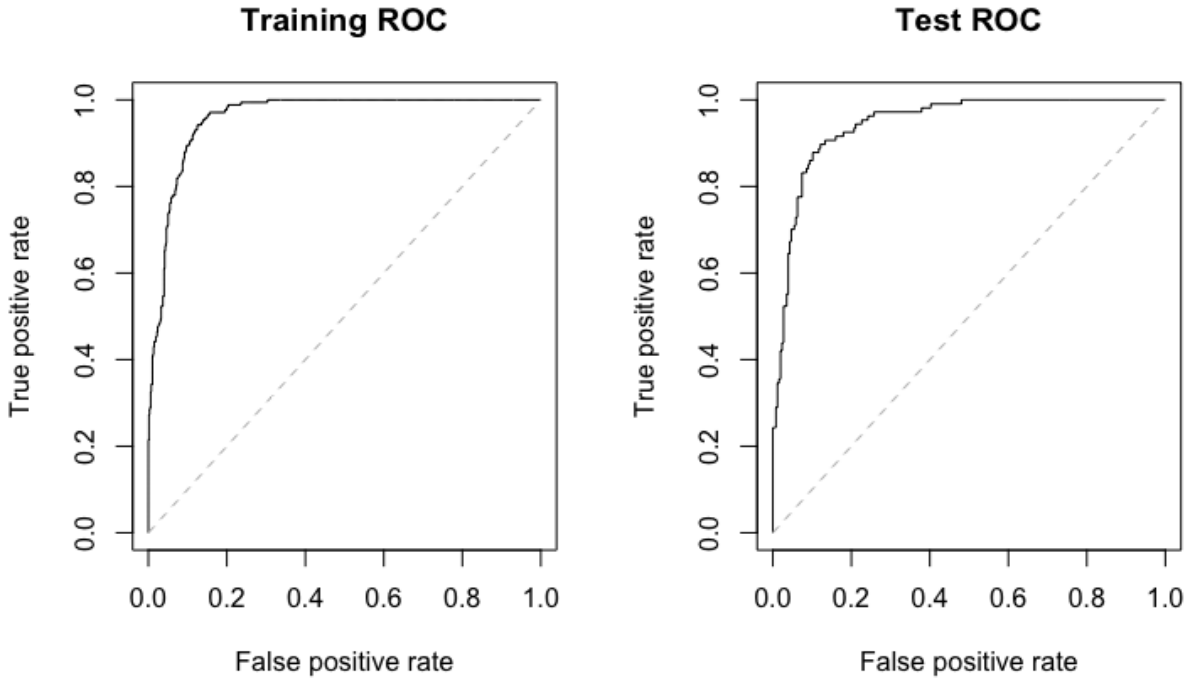

Figure 6.16: ROC curve of SVM model when applied on training and test data.

Then we could make a more informed decision about an observation's actual class. But as we discovered, this is always not the case. As the K value gets larger, there is an increasing risk of using features value that might be appropriate for other classes rather than the actual class. Figure 6.17 reveals how classification error changes with the value of K. We performed the parameter tuning to find the optimum K that minimizes the error, at the same time, addresses the issue described above. After doing a 10-fold cross-validation 1000 times, we found the optimum K to be 84, which gave us the minimum cross-validation error. However, we picked K value as 30 because the error rate after K = 30 does not improve significantly enough.

Therefore, using K = 30, we move forward and train our model with the KNN algorithm. Despite doing the force sampling, the KNN algorithm fails to predict the positive class (obese). One possible explanation could be a large amount of negative class data and KNN's nearest voting algorithm. We think due to the excessive amount of negative class data in training, the nearest vote is resulting in not being obese anyway. This explains the low specificity of our model. Additionally,

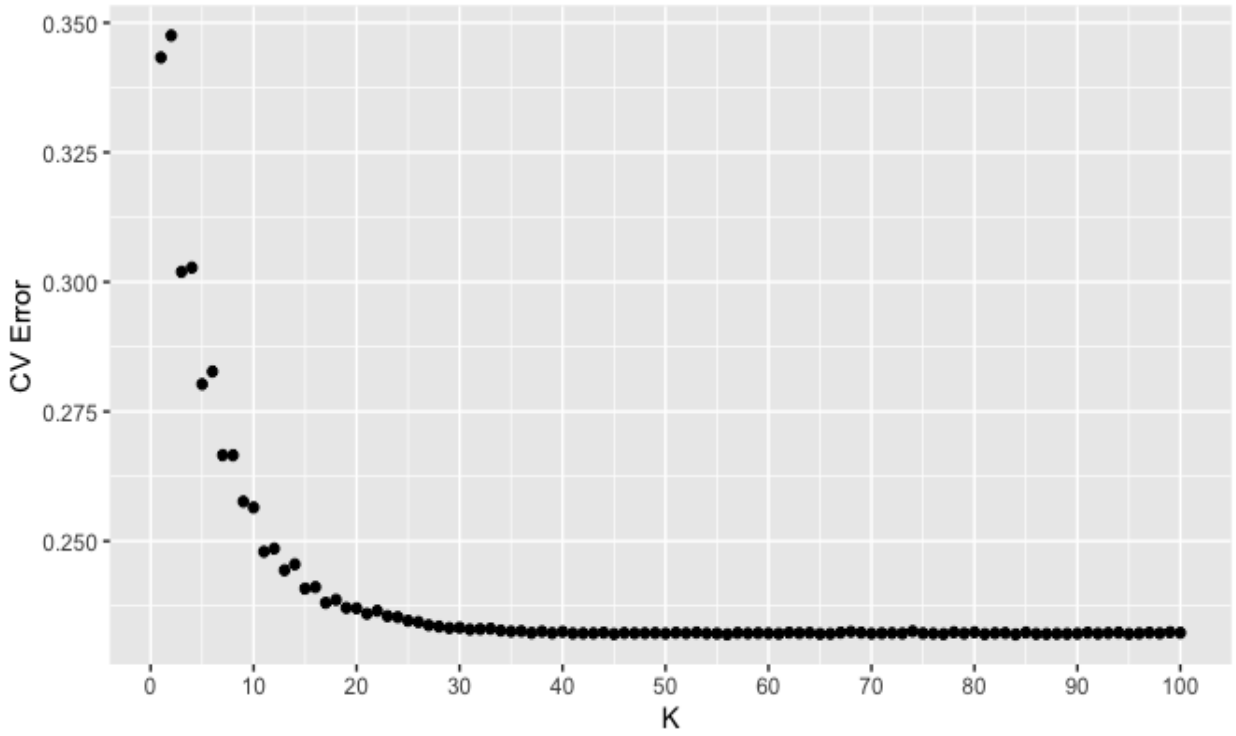

Figure 6.17: Corss-validation error vs K plot of K-nearest neighbor algorithm.

for the first time, we are seeing the advantage of using balanced accuracy because the overall accuracy is 80% in most ML practices which might be acceptable. However, the balanced accuracy is 65%, meaning decision-making with this model will be misleading.

6.3.4 Classification Tree Results

We used R's "tree" library to construct a classification tree for our data. But, before the analysis was performed, we had to consider one significant issue with the classification tree method.

Table 6.8: The confusion matrix and accuracy of K-nearest neighbor.

Confusion Matrix			
Prediction/Reference label	**Not Obese**	**Obese**	**Total**
Not Obese	256	74	242
Obese	0	33	33
Total	256	107	363
Accuracy : 0.7961; 95% CI : (0.751, 0.8364); P-value: 0.000			
Sensitivity : 0.3084; Specificity : 1.0000			

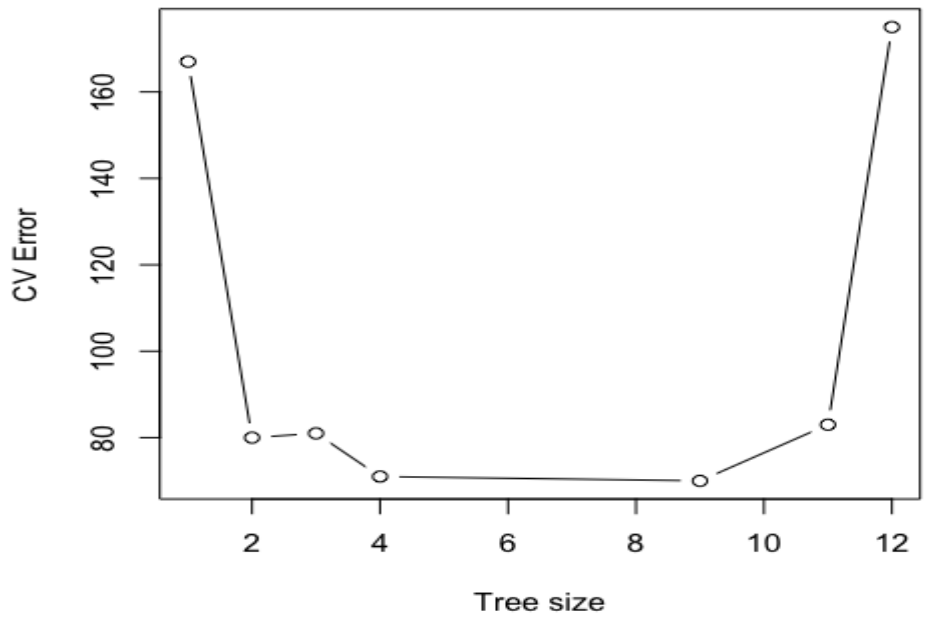

Figure 6.18: Graph of number of effective end nodes vs corss-validation error of classification tree.

The issue is when to stop growing the tree. In theory, it might seem that growing the tree until the end should be the primary choice because it will allow the tree to classify every observation. Such a tree will learn the pattern in the training data so well that there will be no misclassification. In this case, we will face the classical overfitting problem with machine learning.

On the other hand, if we don't let the tree grow sufficiently enough, it will cause an oversimplified model that is of no use to us. Therefore, we decided to prune the tree in order to determine the optimal level of tree complexity and cost complexity. So, rather than deviance, we performed a classification-error-guided cross-validation and pruning process to select a sequence of trees for consideration. The figure 6.18 shows the number of terminal nodes of each tree considered as well as the corresponding error rate. The tree with 9 terminal nodes results in the lowest cross-validation error rate, with 70 cross-validation errors.

Then we refit our tree with nine terminal nodes, and the algorithm used the top six significant features "PA," "Fat," "SES," "Fiber," "Carb," and "Race" to classify the obesity status. One of

Table 6.9: The confusion matrix and accuracy of classification tree.

Confusion Matrix			
Prediction/Reference label	**Not Obese**	**Obese**	**Total**
Not Obese	253	33	286
Obese	3	74	77
Total	256	109	363
Accuracy: 0.9008; 95% CI: (0.8653, 0.9296); P-value: 0.000			
Sensitivity: 0.6916; Specificity: 0.9883			

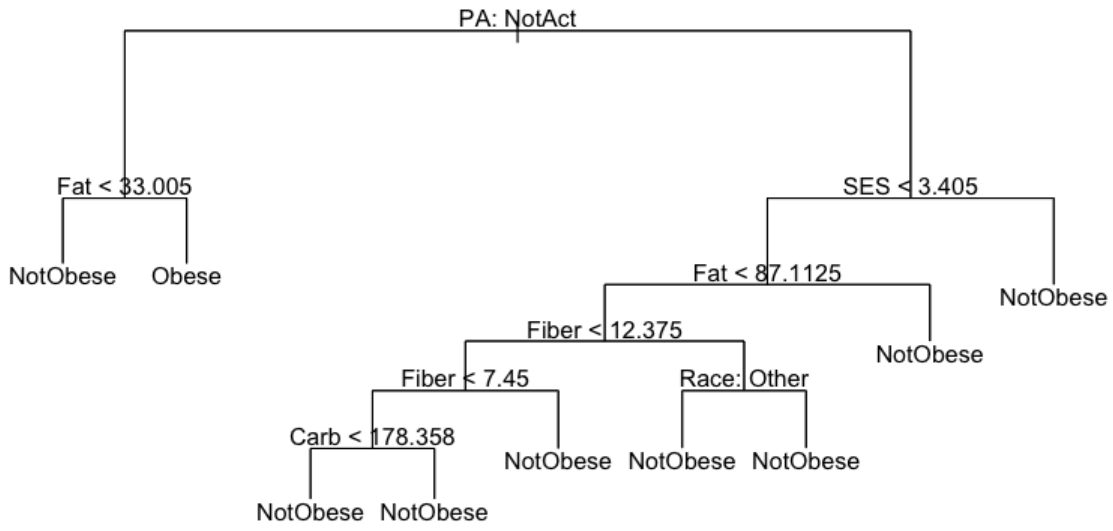

Figure 6.19: Final rendered classification tree with nine terminal nodes. Left branch from any given node implies satisfied node condition.

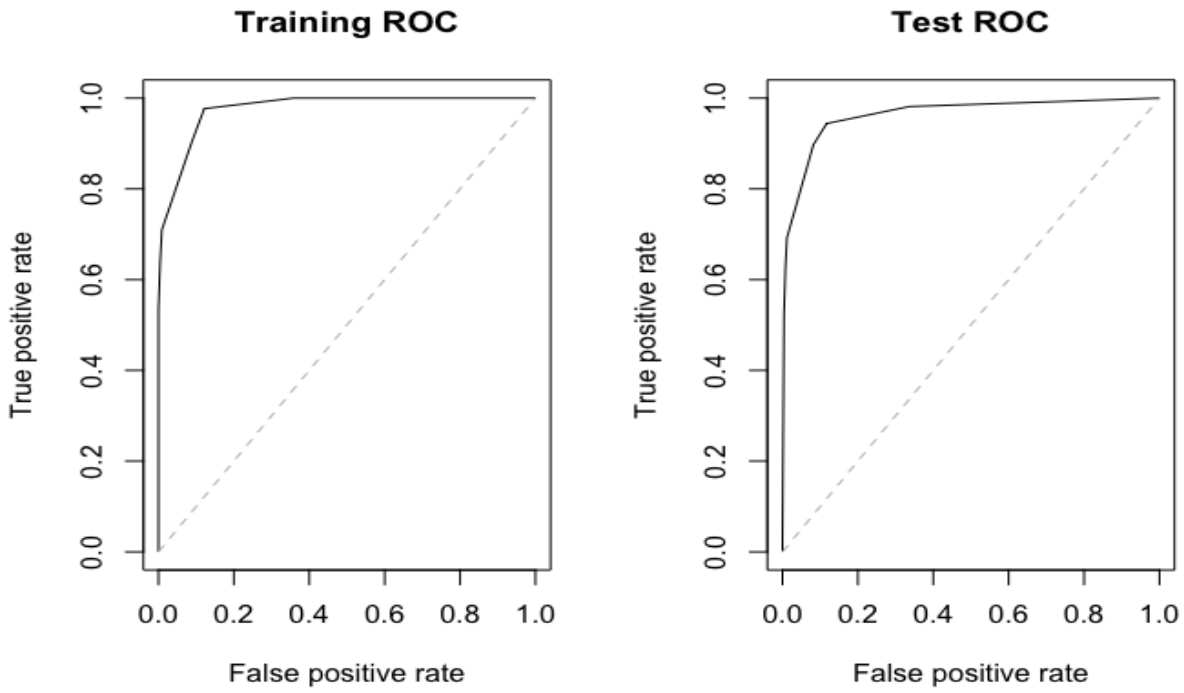

Figure 6.20: Training and testing ROC curve of decision tree algorithm.

the most attractive properties of trees is that they can be graphically presented to display the tree structure, specific split, node labels, and tree growing. Figure 6.19 portrays the graphical layout of the grown tree. According to figure 6.19 the most critical indicator of Obesity appears to be physical activity since the first branch differentiates physically active from not physically active status. Another interesting observation is the algorithm left branch ended quickly into pure node (that is, homogeneous class label) at the Fiber split < 33 gm of average sugar intake.

Regarding accuracy prediction, the fitted model can classify 90 percent of the observations during testing, and the model misclassifies 33 true obese cases as not obese resulting 69 percent sensitivity rate, i.e., 31 percent of the true obese cases were misclassified by the model. During testing, the model's sensitivity was found to be 69 percent, i.e., 69 percent of the cases was truly classified obese by the model that were actually obese. Let us have a look at the ROC curve in Figure 6.20. It represents the plot of the false positive rate against the true positive rate. The ROC curve reveals that the model is performing well during training and with slightly poor performance

Table 6.10: Detailed feature importance table of random forests algorithm.

Variables	Not obese	Obese	Mean Decrease (Accuracy)	Mean Decrease (Gini)
Age	-0.56	0.56	-0.14	11.80
Gender	-0.56	-0.08	-0.42	7.35
Race	-0.40	1.14	0.22	18.50
SES	1.93	2.81	3.21	56.93
Education	5.03	2.78	6.15	19.52
Fiber	1.79	1.85	2.48	55.29
Carb	6.19	-4.51	4.77	51.85
Sugar	5.65	-3.26	4.22	55.13
Fat	2.27	2.10	3.58	60.92
PA	9.10	10.65	13.22	15.85

during testing with a AUC value 0.839.

6.3.5 Random Forests Results

As described in section 5.2.5, random forests provide an improvement over bagged trees by way of a minor tweak that helps to decorrelates the trees. In other words, building a random forest at each split in the tree prevents the algorithm from using high correlated predictors all the time.

For instance, say there is one very strong predictor in the data set, along with several other moderately strong predictors. Then in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Therefore, all the bagged trees will look quite similar to each other, resulting in highly correlated trees with high within-tree variance. So using R's "randomForest" library, we estimated prediction error and out-of-bag error (OOB) for a set of variable selections. We repeated each instance 1000 times, and figure 6.21 shows the MSE and OOB for the number of variables we want to choose in each split. Here we are looking for the combination of low MSE and OOB to determine the number of variables we should randomly sample as candidates at each split. Our estimates suggest we should use six randomly selected variables at each split to achieve the best results.

After we train our model with six variables at each tree split, we calculated the importance of the variables used in this study. Four measures of variable importance are reported in the table 6.10. The first two columns represent the class-specific performance of each feature. That is, what

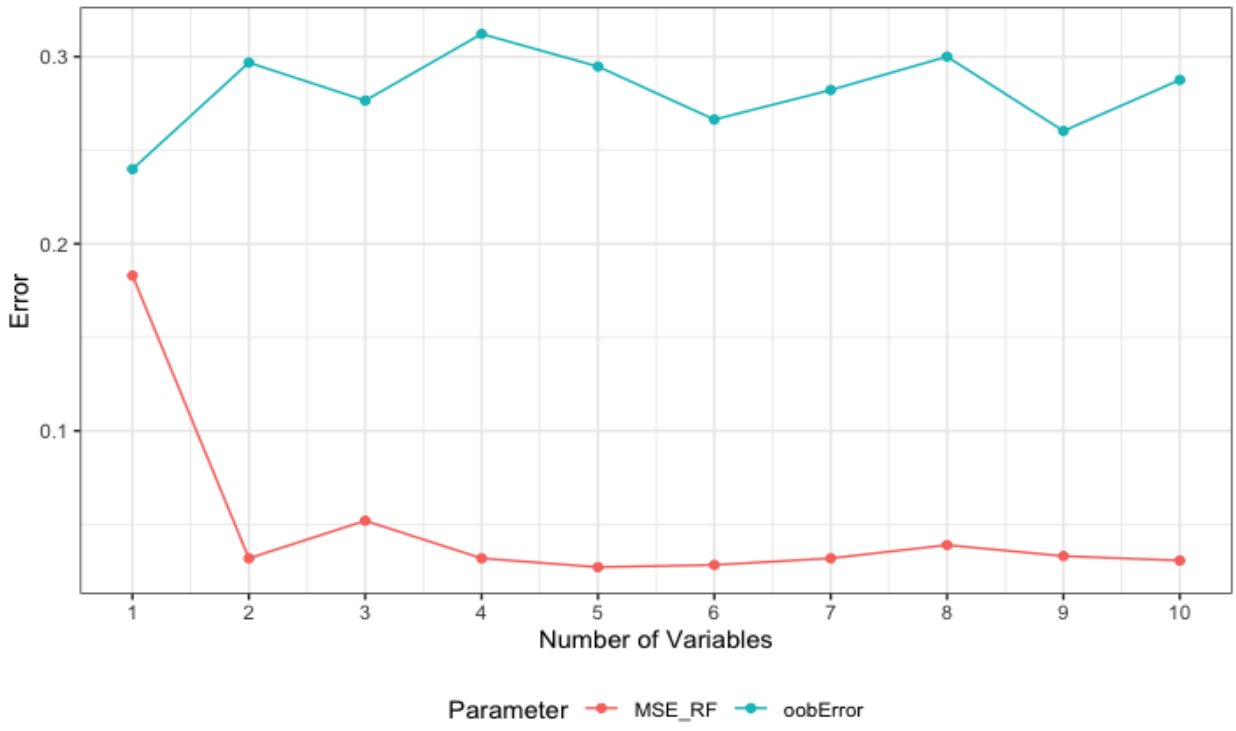

Figure 6.21: MSE and OOB error with respect to number of variables selected as sample candidates at each split.

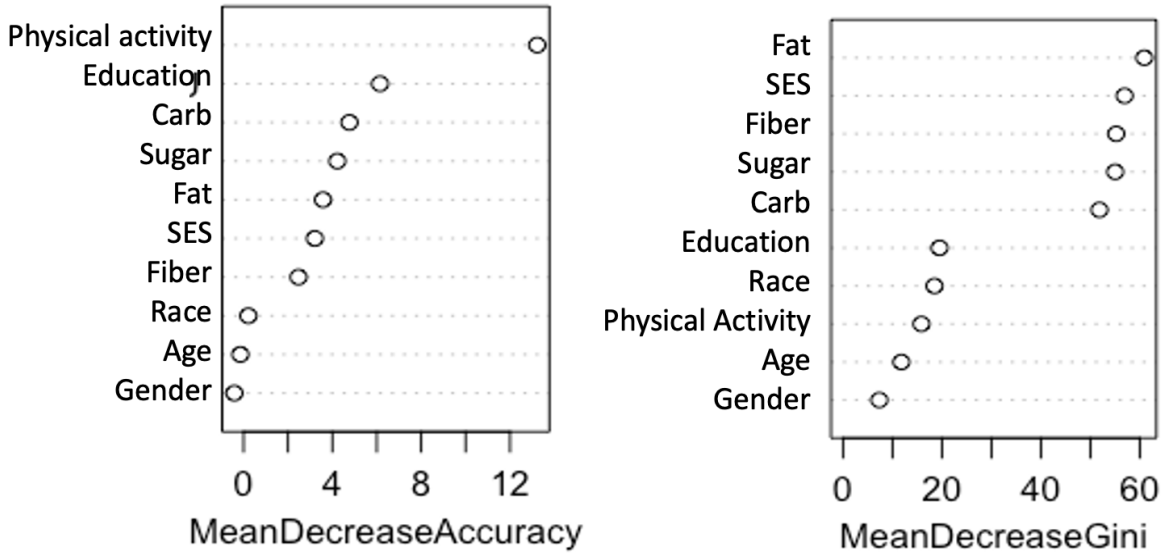

Figure 6.22: Graph of mean decrease in accuracy and Gini for each variables in random forests classification.

Table 6.11: The confusion matrix and accuracy of random forests classification.

Confusion Matrix			
Prediction/Reference label	**Not Obese**	**Obese**	**Total**
Not Obese	249	27	276
Obese	7	80	87
Total	256	107	363
Accuracy: 0.9063; 95% CI: (0.8716, 0.9343); P-value:0.000			
Sensitivity: 0.7477; Specificity: 0.9727			

effect the added variable has on the class prediction accuracy. However, as our data resembles a class imbalanced random forest, in the third column mean decrease in accuracy of a certain variable was reported. This gives us the overall effect a variable has on predicting class. Finally, the fourth column represents the average decrease in Gini impurity measures. As we already discussed in section 5.2.5 that a smaller Gini impurity means a better split, so the third and fourth column essentially represents how effective a variable was in splitting the nodes. The results indicate that across all of the trees considered in the random forest, the variable SES, Education, and Race are by far the three most important variables from demographic data. Moreover, among nutrition intake variables, all four of them have a measurable effect on predicting a child's obesity status. Figure 6.22 summarizes the mean decrease in accuracy and Gini impurity for each feature in the data. In general, variables with a low mean decrease in accuracy should be on the far left of the first figure and to the far right of the second figure of figure 6.22. A quick look at the figure 6.22 corroborates this hypothesis.

The Table 6.11 reflects the random forests classifications performance when unseen test observations feed into the model. The diagonal elements of the confusion matrix table 6.11 indicate correct predictions, while the off-diagonals represent incorrect predictions. Hence our model correctly predicted 249 observations as not obese and 80 observations as obese, for a total of 320 correct predictions. However, when a 10 fold cross validation was performed on the the 100 times the average testing error found to be 0.164 (16.4 percent). Additionally, the balanced accuracy was found to be 86 percent with a sensitivity of 75 percent. And during testing the area under the ROC

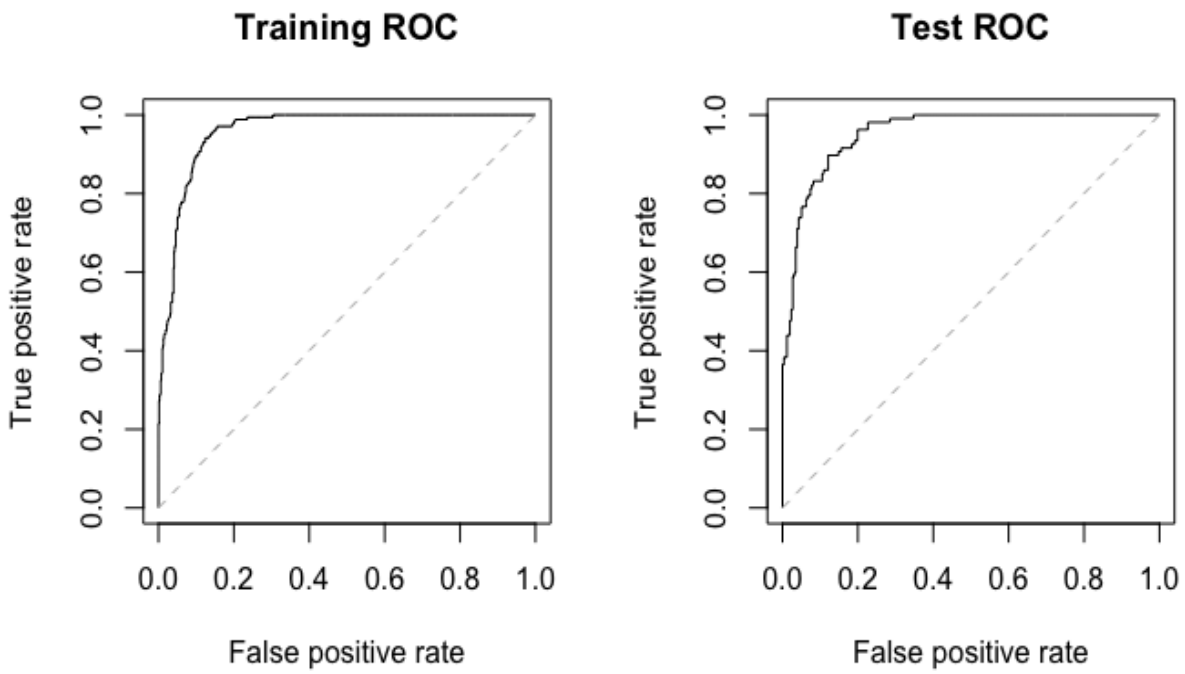

Figure 6.23: ROC curve of random forests model when applied on training and test data.

Table 6.12: The summary of all the machine learning models we used and their corresponding accuracy, balanced accuracy, sensitivity, speceficity, AUC, and 95% CI.

Models	**Accuracy**	**Balanced Accuracy**	**Sensitivity**	**Speceficity**	**AUC**	**95% CI of accuracy**
PCR	0.880	0.840	0.730	0.940	0.831	(84, 91)
SVM	0.900	0.850	0.740	0.970	0.855	(86, 93)
KNN	0.790	0.650	0.310	1.000	0.654	(75, 83)
Classification Tree	0.900	0.840	0.700	0.980	0.839	(86, 92)
Random Forests	**0.910**	**0.860**	**0.750**	**0.970**	**0.859**	(87, 93)

curve was found to be 0.859 (Figure 6.23).

Finally, Table 6.12 shows us the summary of all the models we used and their corresponding balanced accuracy, Sensitivity AUC, etc. It shows that the random forest was the best performing model among all, having 86% balanced accuracy. And the sensitivity of the ransom forests algorithm was found to be 0.75; that is, the model could accurately identify 75% of the obese cases as obese. Across all the models, we experienced high model specificity meaning all models were good at identifying negative class, in this case, non-obese children. One reason we could think of the high volume of negative class data in the training and testing sample. The support vector machine and

the principal component algorithm also performed well during both training and testing.

CHAPTER VII

CONCLUSION AND DISCUSSION

Our study comprises two parts. First, we studied the effect of demographics and micronutrient intake on 4-6yo U.S. children using Path Analysis. Then combining knowledge from our study and literature knowledge, we made an informed decision about what causes obesity in the U.S. younger cohort. Second, using these factors use implemented sophisticated machine learning algorithm to build an ML model to predict obesity.

Our study discovered a significant causal relationship between gender, race, and ethnicity in obesity. Hispanic ethnicity independently showed a significant positive causal path to BMI. A household's reference person's low education level showed positive causation to BMI. In most instances, age also showed significant positive causal relation to BMI. Age consistently significantly increased BMI in all analyses. But through higher fiber intake, age showed a decreasing causal relationship with BMI during the study with the entire and gender-race specific samples. Especially in females, age directly caused increased BMI with $p = 0.323$. In non-Hispanic black males and females, we also found the direct causal effect of age on BMI. There is a strong chance that it's due to biological growth, but this hypothesis again can't explain why age has a decreasing effect on BMI mediated through fiber. Our study showed a substantial direct impact of PA on decreasing BMI across both gender subgroups. In some gender-race-specific subgroups effect of PA was not significant such as in non-Hispanic black children and Hispanic females. Our study found lower BMI was associated with increased physical activity.

Among nutrition intake, the significant path was passed through all nutrition intake variables but in different instances. In females, fiber and carb intake were found to be substantial, whereas fiber and fat intake was significant for males. Carb intake was found to be significant in both

Hispanic males and females. Sugar played a significant role in increasing BMI in non-Hispanic black males and females. Among other racial origins, our study found a significant causal pathway between fiber intake and BMI. We developed a handful of hypothesized causal models and tested them but only reported the best-performing one.

In summary, micronutrient intake, education, race, and physical activity showed significant causation to BMI. Hispanic ethnicity was found to be positively associated with BMI. In every sample and homogenous subsample where the effect of physical activity was present, it was always found to be negative. Age showed increasing causation to BMI.

We implemented several machine learning models on our data set. But here, we would like to mention the data structure from an ML perspective carefully. Our data was class imbalance data, meaning obese class frequency was lower than usual compared to the non-obese class. We used SMOTE technique to address this issue. Among the applied ML algorithms, the tree-based algorithm seems to fit our data well. We trained two tree-based methods classification of trees and random forests. Both models performed well during testing, but the random forest was the best performing one, with around 86 percent of balanced accuracy during testing. Besides these two algorithms, we also applied SVM, PCR, and KNN. SVM and PCR performed competitively with tree based method with respective balanced accuracy of 85%, 84% respectively, but KNN performed the worst among all (balanced accuracy 65%).

CHAPTER VIII

LIMITATIONS AND FUTURE WORK

The biggest challenge we faced doing this study was preparing the data to produce an acceptable sample size. There were lots of missing values in the data set across different IDs and variables. Depending on how many missing values we can estimate in future studies, using missing value estimation methods like bootstrap, interpolation, and EM algorithm results could be improved. The second biggest challenge we overcome is understanding and taking appropriate action regarding the class imbalance problem. In the future, if more data becomes available, our study framework can be applied and evaluated. Even though we applied established methodologies to overcome the class imbalance problem, there will always remain scope to improve the data due to that our predictive model was performing marginally better on predicting positive class (obese). Again the scope of improvement is here based on data availability. We only used demographic characteristics and nutrition intake to study obesity. A better result could be expected if genetic factors were taken into consideration. According to the definition of psychical activity and the nature of the question NHANES asked individuals, we conclude that the PA data for the younger population is unreliable. So it was challenging to perform statistical analysis and produce acceptable results. The physical activity data's reliability was made based on the NHANES variable "perfect recall." It reflects how reliable the recall information is. We wish to build our possible future research from this work by combining path analysis with machine learning to explore more complex path networks that fit the real-world context. The motivation behind this idea was taken from Yu and Yan (2022) [98] approach to Combining machine learning and main path analysis to analyze citation network. [77, 76]

REFERENCES

[1] W. H. Organization *et al.*, *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*. World Health Organization, 2006.

[2] W. H. Organization *et al.*, "Diagnosis and management of type 2 diabetes (hearts-d)," *World Health Organization: Geneva, Switzerland*, 2020.

[3] W. H. Organization, *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization, 2016.

[4] W. H. Organization *et al.*, "Unicef/who/the world bank group joint child malnutrition estimates: levels and trends in child malnutrition: key findings of the 2020 edition," *Available from: https://apps.who.int/iris/bitstream/handle/10665/331621/9789240003576-eng.pdf*, 2020.

[5] W. H. Organization *et al.*, "Levels and trends in child malnutrition: Unicef," *Available from: https://apps.who.int/iris/bitstream/handle/10665/331621/9789240003576-eng.pdf*, 2021.

[6] C. D. Fryar, M. D. Carroll, and J. Afful, "Prevalence of overweight, obesity, and severe obesity among adults aged 20 and over: United states, 1960–1962 through 2017–2018," *NCHS Health E-Stats*, 2020.

[7] C. L. Ogden, M. D. Carroll, H. G. Lawman, C. D. Fryar, D. Kruszon-Moran, B. K. Kit, and K. M. Flegal, "Trends in obesity prevalence among children and adolescents in the united states, 1988-1994 through 2013-2014," *Jama*, vol. 315, no. 21, pp. 2292–2299, 2016.

[8] G. Roglic *et al.*, "Who global report on diabetes: A summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, p. 3, 2016.

[9] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The disease burden associated with overweight and obesity," *Jama*, vol. 282, no. 16, pp. 1523–1529, 1999.

[10] D. B. Allison, K. R. Fontaine, J. E. Manson, J. Stevens, and T. B. VanItallie, "Annual deaths attributable to obesity in the united states," *Jama*, vol. 282, no. 16, pp. 1530–1538, 1999.

[11] "Obesity and overweight — who.int." `https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight`. [Accessed 28-Jul-2022].

[12] A. H. Mokdad, J. S. Marks, D. F. Stroup, and J. L. Gerberding, "Actual causes of death in the united states, 2000," *Jama*, vol. 291, no. 10, pp. 1238–1245, 2004.

[13] H. Cristol, "Trends in global obesity," *The Futurist*, vol. 36, no. 3, p. 10, 2002.

[14] D. B. Allison, R. Zannolli, and K. Narayan, "The direct health care costs of obesity in the united states.," *American Journal of Public Health*, vol. 89, no. 8, pp. 1194–1199, 1999.

[15] J. Rippe, L. Aronne, V. Gilligan, S. Kumanyika, S. Miller, G. Owens, *et al.*, "Public policy statement on obesity and health from the interdisciplinary council on lifestyle and obesity management," *Nutr Clin Care*, vol. 1, pp. 34–7, 1998.

[16] D. Rowland and B. Lyons, "Medicare, medicaid, and the elderly poor," *Health care financing review*, vol. 18, no. 2, p. 61, 1996.

[17] N. S. Wellman and B. Friedberg, "Causes and consequences of adult obesity: health, social and economic impacts in the united states," *Asia Pacific journal of clinical nutrition*, vol. 11, pp. S705–S709, 2002.

[18] O. of the Surgeon General (US); Office of Disease Prevention, H. P. U. C. for Disease Control, and P. (US);, "National institutes of health (us). surgeon general's call to action to prevent and decrease overweight and obesity. rockville (md): Office of the surgeon general (us);," *Available from: https://www.ncbi.nlm.nih.gov/books/NBK44206/*, 2001.

[19] U. D. of Agriculture, U. D. of Health, and H. Services., *US Department of Agriculture and US Department of Health and Human Services . Dietary Guidelines for Americans, 2020–2025. 9th ed.* US Department of Agriculture and US Department of Health and Human Services., 2020.

[20] U. D. of Health, H. Services, *et al.*, "Dietary guidelines for americans 2005," *http://www. health. gov/dietaryguidelines/dga2005/document/default. htm*, 2005.

[21] "Kids and Sleep (for Parents) - Nemours KidsHealth — kidshealth.org." `https://kidshealth.org/en/parents/sleep.html`. [Accessed 24-Jul-2022].

[22] "MyPlate Food Guide (for Parents) - Nemours KidsHealth — kidshealth.org." `https://kidshealth.org/en/parents/myplate.html`. [Accessed 24-Jul-2022].

[23] "Overweight and Obesity (for Parents) - Nemours KidsHealth — kidshealth.org." `https://kidshealth.org/en/parents/overweight-obesity.html`. [Accessed 24-Jul-2022].

[24] "Kids and Exercise (for Parents) - Nemours KidsHealth — kidshealth.org." `https://kidshealth.org/en/parents/exercise.html`. [Accessed 24-Jul-2022].

[25] Y. Wang and M. A. Beydoun, "The obesity epidemic in the united states—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis," *Epidemiologic reviews*, vol. 29, no. 1, pp. 6–28, 2007.

[26] Y. Wu, S. Datta, B. B. Little, and M. Kong, "Magnesium dietary intake and physical activity in type 2 diabetes by gender in white, african-american and mexican american: Nhanes 2011-2014," *Endocrinology, Diabetes & Metabolism*, vol. 4, no. 1, p. e00203, 2021.

[27] R. P. Treviño, R. M. Marshall, D. E. Hale, R. Rodriguez, G. Baker, and J. Gomez, "Diabetes risk factors in low-income mexican-american children.," *Diabetes care*, vol. 22, no. 2, pp. 202–207, 1999.

[28] M. Chandalia, A. Garg, D. Lutjohann, K. Von Bergmann, S. M. Grundy, and L. J. Brinkley, "Beneficial effects of high dietary fiber intake in patients with type 2 diabetes mellitus," *New England Journal of Medicine*, vol. 342, no. 19, pp. 1392–1398, 2000.

[29] I. C. anneleen. kuijsten@ wur. nl, "Dietary fibre and incidence of type 2 diabetes in eight european countries: the epic-interact study and a meta-analysis of prospective studies," *Diabetologia*, vol. 58, pp. 1394–1408, 2015.

[30] M. Brauchla, W. Juan, J. Story, and S. Kranz, "Sources of dietary fiber and the association of fiber intake with childhood obesity risk (in 2–18 year olds) and diabetes risk of adolescents 12–18 year olds: Nhanes 2003–2006," *Journal of nutrition and metabolism*, vol. 2012, 2012.

[31] J. Zhang, H. Wang, Y. Wang, H. Xue, Z. Wang, W. Du, C. Su, J. Zhang, H. Jiang, F. Zhai, *et al.*, "Dietary patterns and their associations with childhood obesity in china," *British journal of nutrition*, vol. 113, no. 12, pp. 1978–1984, 2015.

[32] J. N. Davis, E. E. Ventura, M. J. Weigensberg, G. D. Ball, M. L. Cruz, G. Q. Shaibi, and M. I. Goran, "The relation of sugar intake to β cell function in overweight latino children," *The American journal of clinical nutrition*, vol. 82, no. 5, pp. 1004–1010, 2005.

[33] C. Byrd-Williams, L. A. Kelly, J. N. Davis, D. Spruijt-Metz, and M. I. Goran, "Influence of gender, bmi and hispanic ethnicity on physical activity in children," *International Journal of Pediatric Obesity*, vol. 2, no. 3, pp. 159–166, 2007.

[34] A. M. Davis, K. J. Bennett, C. Befort, and N. Nollen, "Obesity and related health behaviors among urban and rural children in the united states: data from the national health and nutrition examination survey 2003–2004 and 2005–2006," *Journal of pediatric psychology*, vol. 36, no. 6, pp. 669–676, 2011.

[35] K. Patrick, G. J. Norman, K. J. Calfas, J. F. Sallis, M. F. Zabinski, J. Rupp, and J. Cella, "Diet, physical activity, and sedentary behaviors as risk factors for overweight in adolescence," *Archives of pediatrics & adolescent medicine*, vol. 158, no. 4, pp. 385–390, 2004.

[36] C. J. Crespo, E. Smit, R. P. Troiano, S. J. Bartlett, C. A. Macera, and R. E. Andersen, "Television watching, energy intake, and obesity in us children: results from the third national health and nutrition examination survey, 1988-1994," *Archives of pediatrics & adolescent medicine*, vol. 155, no. 3, pp. 360–365, 2001.

[37] M. N. Lutfiyya, R. Garcia, C. M. Dankwa, T. Young, and M. S. Lipsky, "Overweight and obese prevalence rates in african american and hispanic children: an analysis of data from the 2003–2004 national survey of children's health," *The Journal of the American Board of Family Medicine*, vol. 21, no. 3, pp. 191–199, 2008.

[38] M. A. H. Alfieri, J. Pomerleau, D. M. Grace, and L. Anderson, "Fiber intake of normal weight, moderately obese and severely obese subjects," *Obesity research*, vol. 3, no. 6, pp. 541–547, 1995.

[39] A. J. Ariza, E. H. Chen, H. J. Binns, and K. K. Christoffel, "Risk factors for overweight in five-to six-year-old hispanic-american children: a pilot study," *Journal of Urban Health*, vol. 81, no. 1, pp. 150–161, 2004.

[40] K. Nissinen, V. Mikkilä, S. Männistö, M. Lahti-Koski, L. Räsänen, J. Viikari, and O. T. Raitakari, "Sweets and sugar-sweetened soft drink intake in childhood in relation to adult bmi and overweight. the cardiovascular risk in young finns study," *Public Health Nutrition*, vol. 12, no. 11, pp. 2018–2026, 2009.

[41] B. A. Dennison, H. L. Rockwell, and S. L. Baker, "Excess fruit juice consumption by preschool-aged children is associated with short stature and obesity," *Pediatrics*, vol. 99, no. 1, pp. 15–22, 1997.

[42] R. Rajeshwari, S.-J. Yang, T. A. Nicklas, and G. S. Berenson, "Secular trends in children's sweetened-beverage consumption (1973 to 1994): the bogalusa heart study," *Journal of the American Dietetic Association*, vol. 105, no. 2, pp. 208–214, 2005.

[43] V. S. Malik, B. M. Popkin, G. A. Bray, J.-P. Després, and F. B. Hu, "Sugar-sweetened beverages, obesity, type 2 diabetes mellitus, and cardiovascular disease risk," *Circulation*, vol. 121, no. 11, pp. 1356–1364, 2010.

[44] E. E. Ventura, J. N. Davis, K. E. Alexander, G. Q. Shaibi, W. Lee, C. E. Byrd-Williams, C. M. Toledo-Corral, C. J. Lane, L. A. Kelly, M. J. Weigensberg, *et al.*, "Dietary intake and the metabolic syndrome in overweight latino children," *Journal of the American dietetic association*, vol. 108, no. 8, pp. 1355–1359, 2008.

[45] E. Ventura, J. Davis, C. Byrd-Williams, K. Alexander, A. McClain, C. J. Lane, D. Spruijt-Metz, M. Weigensberg, and M. Goran, "Reduction in risk factors for type 2 diabetes mellitus in response to a low-sugar, high-fiber dietary intervention in overweight latino adolescents," *Archives of pediatrics & adolescent medicine*, vol. 163, no. 4, pp. 320–327, 2009.

[46] F. B. Hu and V. S. Malik, "Sugar-sweetened beverages and risk of obesity and type 2 diabetes: epidemiologic evidence," *Physiology & behavior*, vol. 100, no. 1, pp. 47–54, 2010.

[47] N. Fidler Mis, C. Braegger, J. Bronsky, C. Campoy, M. Domellöf, N. D. Embleton, I. Hojsak, J. Hulst, F. Indrio, A. Lapillonne, *et al.*, "Sugar in infants, children and adolescents: a position paper of the european society for paediatric gastroenterology, hepatology and nutrition committee on nutrition," *Journal of pediatric gastroenterology and nutrition*, vol. 65, no. 6, pp. 681–696, 2017.

[48] J. A. Welsh, M. E. Cogswell, S. Rogers, H. Rockett, Z. Mei, and L. M. Grummer-Strawn, "Overweight among low-income preschool children associated with the consumption of sweet drinks: Missouri, 1999–2002," *Pediatrics*, vol. 115, no. 2, pp. e223–e229, 2005.

[49] D. S. Ludwig, K. E. Peterson, and S. L. Gortmaker, "Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis," *The lancet*, vol. 357, no. 9255, pp. 505–508, 2001.

[50] R. Pérez-Escamilla and P. Putnik, "The role of acculturation in nutrition, lifestyle, and incidence of type 2 diabetes among latinos," *The Journal of nutrition*, vol. 137, no. 4, pp. 860–870, 2007.

[51] M. A. B. Khan, M. J. Hashim, J. K. King, R. D. Govender, H. Mustafa, and J. Al Kaabi, "Epidemiology of type 2 diabetes–global burden of disease and forecasted trends," *Journal of epidemiology and global health*, vol. 10, no. 1, p. 107, 2020.

[52] C. for Disease Control and P. (CDC), "National health and nutrition examination survey.."

[53] K. Siddiqui, "Heuristics for sample size determination in multivariate statistical techniques," *World Applied Sciences Journal*, vol. 27, no. 2, pp. 285–287, 2013.

[54] E. J. Wolf, K. M. Harrington, S. L. Clark, and M. W. Miller, "Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety," *Educational and psychological measurement*, vol. 73, no. 6, pp. 913–934, 2013.

[55] G. M. Fisher, "Poverty guidlines for 1992," *Soc. Sec. Bull.*, vol. 55, p. 43, 1992.

[56] W. H. Dietz, "The response of the us centers for disease control and prevention to the obesity epidemic," *Annu Rev Public Health*, vol. 36, no. 1, pp. 575–596, 2015.

[57] T. Armstrong and F. Bull, "Development of the world health organization global physical activity questionnaire (gpaq)," *Journal of Public Health*, vol. 14, no. 2, pp. 66–70, 2006.

[58] S. Wright, "The method of path coefficients," *The annals of mathematical statistics*, vol. 5, no. 3, pp. 161–215, 1934.

[59] K.-H. Huarng, T. H.-K. Yu, and C. fang Lee, "Adoption model of healthcare wearable devices," *Technological Forecasting and Social Change*, vol. 174, p. 121286, 2022.

[60] F. Huang, W. Sun, L. Zhang, H. Lu, and W.-T. Chen, "Depressive symptoms mediate covid-associated stigma and quality of life: Stigma instrument validation and path analysis," *Journal of affective disorders*, vol. 297, pp. 269–275, 2022.

[61] J. Jaccard, C. K. Wan, and J. Jaccard, *LISREL approaches to interaction effects in multiple regression*. No. 114, sage, 1996.

[62] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.

[63] C. C. Li *et al.*, *Path Analysis-a primer.* The Boxwood Press., 1975.

[64] B. Shipley, *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press, 2016.

[65] J. S. Garrow and J. Webster, "Quetelet's index (w/h2) as a measure of fatness.," *International journal of obesity*, vol. 9, no. 2, pp. 147–153, 1985.

[66] M. M. Finucane, G. A. Stevens, M. J. Cowan, G. Danaei, J. K. Lin, C. J. Paciorek, G. M. Singh, H. R. Gutierrez, Y. Lu, A. N. Bahalim, *et al.*, "National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9· 1 million participants," *The lancet*, vol. 377, no. 9765, pp. 557–567, 2011.

[67] D. Gallagher, M. Visser, D. Sepulveda, R. N. Pierson, T. Harris, and S. B. Heymsfield, "How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups?," *American journal of epidemiology*, vol. 143, no. 3, pp. 228–239, 1996.

[68] S. Weisberg, *Applied linear regression*, vol. 528. John Wiley & Sons, 2005.

[69] G. Casella and R. L. Berger, *Statistical inference*. Cengage Learning, 2021.

[70] D. M. Gay, "Usage summary for selected optimization routines," *Computing science technical report*, vol. 153, pp. 1–21, 1990.

[71] K. Gana and G. Broc, *Structural equation modeling with lavaan*. John Wiley & Sons, 2019.

[72] W. D. Berry, *Understanding regression assumptions*, vol. 92. Sage, 1993.

[73] G. W. Bohrnstedt and T. M. Carter, "Robustness in regression analysis," *Sociological methodology*, vol. 3, pp. 118–146, 1971.

[74] K. A. Bollen, *Structural equations with latent variables*, vol. 210. John Wiley & Sons, 1989.

[75] H. C. Thode, *Testing for normality*. CRC press, 2002.

[76] X. Fan, B. Thompson, and L. Wang, "Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes," *Structural equation modeling: a multidisciplinary journal*, vol. 6, no. 1, pp. 56–83, 1999.

[77] L.-t. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural equation modeling: a multidisciplinary journal*, vol. 6, no. 1, pp. 1–55, 1999.

[78] F. F. Chen, "Sensitivity of goodness of fit indexes to lack of measurement invariance," *Structural equation modeling: a multidisciplinary journal*, vol. 14, no. 3, pp. 464–504, 2007.

[79] I. T. Jolliffe, "A note on the use of principal components in regression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 31, no. 3, pp. 300–303, 1982.

[80] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[81] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[82] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[83] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[84] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[85] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[86] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[87] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[88] T. Dahiru, "P-value, a true test of statistical significance? a cautionary note," *Annals of Ibadan postgraduate medicine*, vol. 6, no. 1, pp. 21–26, 2008.

[89] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*, vol. 26. Springer, 2013.

[90] T. Fawcett and F. J. Provost, "Combining data mining and machine learning for effective user profiling.," in *KDD*, vol. 96, pp. 8–13, 1996.

[91] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001.

[92] M. Kubat, S. Matwin, *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97, p. 179, Nashville, USA, 1997.

[93] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on Artificial Intelligence*, vol. 56, pp. 111–117, Citeseer, 2000.

[94] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[95] W. D. Berry, S. Feldman, and D. Stanley Feldman, *Multiple regression in practice*. No. 50, Sage, 1985.

[96] J. Neter, M. H. Kutner, C. J. Nachtsheim, W. Wasserman, *et al.*, "Applied linear statistical models," 1996.

[97] B. R. Sinco and P. L. Chapman, "Adventures in path analysis and preparatory analysis," *Dimuat turun daripada http://www. mwsug. org/proceedings/2013/AA/MWSUG-2013-AA05. pdf*, 2013.

[98] D. Yu and Z. Yan, "Combining machine learning and main path analysis to identify research front: from the perspective of science-technology linkage," *Scientometrics*, pp. 1–24, 2022.

[99] A. M. Mood, "On the asymptotic efficiency of certain nonparametric two-sample tests," *The Annals of Mathematical Statistics*, pp. 514–522, 1954.

[100] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

APPENDIX A

APPENDIX A

SUPPLEMENTARY MATERIALS

Moods Median Test Mood's median test is a nonparametric test to compare the medians of two independent samples [99]. Hypotheses are defined as

1. **Null hypothesis:** The medians of the populations from which the groups were sampled are equal.

2. **Alternative hypothesis(two-sided):** The medians of the populations from which the groups were sampled are not all equal.

Kolmogorov-Smirnov Test The Kolmogorov-Smirnov test (K-S test) is a nonparametric test of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test) [100]. Hypotheses are defined as

1. **Null hypothesis:** The probability distribution of the populations from which the groups were sampled are similar.

2. **Alternative hypothesis (two-sided):** The probability distribution of the populations from which the groups were sampled are NOT similar.

For this test the empirical distribution function F_n for n independent and identically distributed (i.i.d.) ordered observations X_i and can be defined as

$$F_n(x) = \frac{\text{number of (elements in the sample } \leq x)}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{[-\infty,x]}(X_i)$$

where $1_{[-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

In that case, the Kolmogorov-Smirnov statistic is

$$D_{n,m} = \sup_x \left| F_{1,n}(x) - F_{2,m}(x) \right|,$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively, and sup is the supremum function. For large samples, the null hypothesis is rejected at level α if

$$D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1+\frac{m}{n}}{2m}}$$

Table A.1: All individual paths and summarized causal effects based on the path analysis for the entire population.

Paths	**Estimates**	**Std Error**	**P-value**	**Std.lv**	**Std.all**
Fiber					
(Age –>Fiber)	0.117	0.033	0.000	0.117	0.096
(Gender –>Fiber)	-0.154	0.055	0.005	-0.154	-0.077
(Hispanic –>Fiber)	0.307	0.069	0.000	0.307	0.139
(Non Hispanic Black –>Fiber)	0.156	0.078	0.047	0.156	0.066
(Other Race –>Fiber)	0.058	0.084	0.049	0.058	0.022
(SES –>Fiber)	0.008	0.024	0.736	0.008	0.012
(High school and Some Collage –>Fiber)	-0.384	0.130	0.003	-0.384	-0.182
(Collage and Above –>Fiber)	-0.076	0.148	0.606	-0.076	-0.033
(Physical Activity–>Fiber)	0.112	0.104	0.278	0.112	0.030
Carb					
(Age –>Carb)	0.161	0.030	0.000	0.161	0.160
(Gender –>Carb)	-0.219	0.059	0.000	-0.219	-0.132

Table A.1 continued.

Paths	Estimates	Std Error	p-value	Std.lv	Std.all
(Hispanic –>Carb)	0.017	0.076	0.241	0.017	0.009
(Non Hispanic Black –>Carb)	0.190	0.093	0.041	0.190	0.096
(Other Race –>Carb)	-0.083	0.081	0.311	-0.083	-0.038
(SES –>Carb)	0.014	0.024	0.555	0.014	0.026
(High school and Some Collage –>Carb)	0.119	0.123	0.335	0.119	0.068
(Collage and Above –>Carb)	0.206	0.150	0.171	0.206	0.107
(Physical Activity–>Carb)	-0.007	0.092	0.936	-0.007	-0.002
Sugar					
(Age –>Sugar)	0.132	0.029	0.000	0.132	0.109
(Gender –>Sugar)	-0.132	0.056	0.019	-0.132	-0.066
(Hispanic –>Sugar)	-0.080	0.079	0.315	-0.080	-0.036
(Non Hispanic Black –>Sugar)	0.015	0.090	0.871	0.015	0.006
(Other Race –>Sugar)	-0.316	0.082	0.000	-0.316	-0.121
(SES –>Sugar)	0.011	0.024	0.636	0.011	0.017
(High school and Some Collage –>Sugar)	0.347	0.112	0.002	0.347	0.164
(Collage and Above –>Sugar)	0.348	0.140	0.013	0.348	0.150
(Physical Activity–>Sugar)	-0.107	0.119	0.368	-0.107	-0.028
Fat					
(Age –>Fat)	0.144	0.036	0.000	0.144	0.119
(Gender –>Fat)	-0.171	0.054	0.002	-0.171	-0.086
(Hispanic –>Fat)	0.013	0.077	0.866	0.013	0.006
(Non Hispanic Black –>Fat)	0.107	0.082	0.190	0.107	0.045
(Other Race –>Fat)	-0.103	0.079	0.031	-0.103	-0.039
(SES –>Fat)	0.002	0.024	0.920	0.002	0.004

Table A.1 continued.

Paths	Estimates	Std Error	p-value	Std.lv	Std.all
(High school and Some Collage –>Fat)	0.251	0.115	0.029	0.251	0.119
(Collage and Above –>Fat)	0.248	0.134	0.064	0.248	0.107
(Physical Activity–>Fat)	0.048	0.090	0.594	0.048	0.013
BMI					
(Fiber –>BMI)	-0.089	0.096	0.035	-0.089	-0.036
(Carb –>BMI)	0.109	0.190	0.568	0.109	0.037
(Sugar –>BMI)	0.018	0.139	0.898	0.018	0.007
(Fat –>BMI)	0.179	0.096	0.044	0.179	0.073
(Age –>BMI)	0.231	0.079	0.004	0.231	0.078
(Gender –>BMI)	0.234	0.139	0.093	0.234	0.048
(Physical Activity–>BMI)	-1.164	0.240	0.000	-1.164	-0.126
(Hispanic –>BMI)	0.329	0.192	0.047	0.328	0.061
(Non Hispanic Black –>BMI)	-0.083	0.194	0.667	-0.083	-0.014
(Other Race –>BMI)	-0.383	0.213	0.050	-0.383	-0.060
(SES –>BMI)	-0.036	0.063	0.564	-0.036	-0.023
(High school and Some Collage –>BMI)	-0.167	0.308	0.588	-0.167	-0.032
(Collage and Above –>BMI)	-0.706	0.373	0.048	-0.706	-0.125

Table A.2: All individual Special effects, Total Indirect Effects, and Total effects based on the path analysis for the entire population.

Mediations	Estimate	Std Error	p-value	Std.lv	Std.all
Age					
SIE (Age –>Fiber –>BMI)	-0.010	0.011	0.336	-0.010	-0.003
SIE (Age –>Carb –>BMI)	0.017	0.030	0.555	0.017	0.006

Table A.2 continued.

Mediations	Estimate	Std Error	p-value	Std.lv	Std.all
SIE (Age –>Sugar –>BMI)	0.002	0.018	0.898	0.002	0.001
SIE (Age –>Fat –>BMI)	0.026	0.015	0.048	0.026	0.009
TIE (Age –>Nutrient Intake –>BMI)	0.035	0.017	0.039	0.035	0.012
TE (Age –>BMI)	0.266	0.076	0.000	0.266	0.090
Gender					
SIE (Gender –>Fiber –>BMI)	0.014	0.016	0.394	0.014	0.003
SIE (Gender –>Carb –>BMI)	-0.024	0.043	0.585	-0.024	-0.005
SIE (Gender –>Sugar –>BMI)	-0.002	0.019	0.902	-0.002	0.000
SIE (Gender –>Fat –>BMI)	-0.031	0.018	0.041	-0.031	-0.006
TIE (Gender –>Nutrient Intake –>BMI)	-0.043	0.024	0.038	-0.043	-0.009
TE (Gender –>BMI)	0.191	0.136	0.039	0.191	0.039
Race					
Hispanic					
SIE (Hispanic –>Fiber –>BMI)	-0.027	0.031	0.041	-0.027	-0.005
SIE (Hispanic –>Carb –>BMI)	0.002	0.018	0.219	0.002	0.000
SIE (Hispanic –>Sugar –>BMI)	-0.001	0.014	0.321	-0.001	0.000
SIE (Hispanic –>Fat –>BMI)	0.002	0.018	0.091	0.002	0.000
TIE (Hispanic –>Nutrient Intake –>BMI)	-0.025	0.031	0.433	-0.025	-0.005
TE (Hispanic –>BMI)	0.304	0.191	0.032	0.304	0.056
Non Hispanic Black					
SIE (Non Hispanic Black –> Fiber –>BMI)	-0.014	0.019	0.459	-0.014	-0.002

Table A.2 continued.

Mediations	Estimate	Std Error	p-value	Std.lv	Std.all
SIE (Non Hispanic Black –> Carb –>BMI)	0.021	0.042	0.622	0.021	0.004
SIE (Non Hispanic Black –> Sugar –>BMI)	0.000	0.015	0.987	0.000	0.000
SIE (Non Hispanic Black –> Fat –>BMI)	0.019	0.019	0.319	0.019	0.003
TIE (Non Hispanic Black –> Nutrient Intake –>BMI)	0.026	0.031	0.404	0.026	0.005
TE (Non Hispanic Black –>BMI)	-0.057	0.200	0.775	-0.057	-0.010
Other Race					
SIE (Other Race –>Fiber –>BMI)	-0.005	0.014	0.710	-0.005	-0.001
SIE (Other Race –>Carb –>BMI)	-0.009	0.027	0.740	-0.009	-0.001
SIE (Other Race –>Sugar –>BMI)	-0.006	0.044	0.899	-0.006	-0.001
SIE (Other Race –>Fat –>BMI)	-0.018	0.019	0.032	-0.018	-0.003
TIE (Other Race –>Nutrient Intake –>BMI)	-0.038	0.036	0.291	-0.038	-0.006
TE (Other Race –>BMI)	-0.421	0.210	0.043	-0.421	-0.066
Socioeconomic Status					
SIE (SES –> Fiber –>BMI)	-0.001	0.004	0.846	-0.001	0.000
SIE (SES –> Carb –>BMI)	0.002	0.007	0.819	0.002	0.001
SIE (SES –> Sugar –>BMI)	0.000	0.004	0.964	0.000	0.000

Table A.2 continued.

Mediations	Estimate	Std Error	p-value	Std.lv	Std.all
SIE (SES –> Fat –>BMI)	0.000	0.005	0.926	0.000	0.000
TIE (SES –> Nutrient Intake –>BMI)	0.001	0.006	0.819	0.001	0.001
TE(SES –>BMI)	-0.035	0.062	0.575	-0.035	-0.022
Education					
High school and Some College					
SIE (High school and Some Collage –> Fiber –>BMI)	0.034	0.042	0.420	0.034	0.007
SIE (High school and Some Collage –> Carb –>BMI)	0.013	0.039	0.738	0.013	0.002
SIE (High school and Some Collage –> Sugar –>BMI)	0.006	0.048	0.897	0.006	0.001
SIE (High school and Some Collage –> Fat –>BMI)	0.045	0.038	0.234	0.045	0.009
TIE (High school and Some Collage –> Nutrient Intake –>BMI)	0.098	0.060	0.049	0.098	0.019
TE(High School and Some Collage –>BMI)	-0.069	0.312	0.826	-0.069	-0.013
College and Above					
SIE (Collage and Above –> Fiber –>BMI)	0.007	0.022	0.760	0.007	0.001
SIE (Collage and Above –> Carb –>BMI)	0.022	0.051	0.663	0.022	0.004

Table A.2 continued.

Mediations	Estimate	Std Error	p-value	Std.lv	Std.all
SIE (Collage and Above –> Sugar –>BMI)	0.006	0.052	0.905	0.006	0.001
SIE (Collage and Above –> Fat –>BMI)	0.044	0.038	0.047	0.044	0.008
TIE (Collage and Above –> Nutrient Intake –>BMI)	0.080	0.052	0.129	0.080	0.014
TE(Collage and Above –>BMI)	-0.627	0.373	0.033	-0.627	-0.111
Physical activity					
SIE (Physical activity–> Fiber –>BMI)	-0.010	0.019	0.592	-0.010	-0.001
SIE (Physical activity–> Carb –>BMI)	-0.001	0.020	0.967	-0.001	-0.000
SIE (Physical activity–> Sugar –>BMI)	-0.002	0.020	0.923	-0.002	-0.000
SIE (Physical activity–> Fat –>BMI)	0.009	0.019	0.654	0.009	0.001
TIE (Physical activity–> Nutrient Intake –>BMI)	-0.004	0.032	0.897	-0.004	-0.000
TE (Physical activity –>BMI)	-1.168	0.249	0.000	-1.168	-0.126

Table A.3: The goodness of fit measures (adjusted R^2, CFI, SRMR of gender specific path model.

Adjusted R^2	Estimate
Fiber	70
Carb	46
Suger	55
Fat	33
BMI	78
CFI: 0.924; SRMR: 0.032	

Table A.4: The residuals correlation difference of our gender-specific fitted model. This is simply the difference between the observed and implied covariance matrix. An ideal situation would be all 0 elements.

Variables	Fiber	Carb	Suger	Fat	BMI	AGE	Race	SES	Education	PA
Fiber	0.000									
Carb	0.612	0.000								
Suger	0.326	0.765	0.000							
Fat	0.443	0.580	0.356	0.000						
BMI	0.069	0.027	0.081	0.042	0.000					
AGE	0.000	0.000	0.000	0.000	0.000	0.000				
Race	0.000	0.000	0.000	0.000	0.000	0.000	0.000			
SES	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		
Education	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
PA	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000

Table A.5: The goodness of fit measures (adjusted R^2, CFI, SRMR of female race-specific path model.

Adjusted R^2	Estimate
Fiber	81
Carb	47
Suger	65
Fat	53
BMI	71
CFI: 0.871; SRMR: 0.062	

Table A.6: The residuals correlation difference of our gender race-specific fitted model. This is simply the difference between the observed and implied covariance matrix. An ideal situation would be all 0 elements.

Variables	Fiber	Carb	Suger	Fat	BMI	AGE	Race	SES	Education	PA
Fiber	0.000									
Carb	0.612	0.000								
Suger	0.326	0.765	0.000							
Fat	0.443	0.580	0.356	0.000						
BMI	0.069	0.027	0.081	0.042	0.000					
AGE	0.000	0.000	0.000	0.000	0.000	0.000				
Race	0.000	0.000	0.000	0.000	0.000	0.000	0.000			
SES	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		
Education	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
PA	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000

BIOGRAPHICAL SKETCH

Prosanta Barai was born in a small village of Barishal, Bangladesh, in 1995. His early childhood flourished with dynamic memory of nature and rural Bangladesh. To attend college, he moved to the capital city Dhaka and joined one of the most esteemed universities in Bangladesh, the University of Dhaka, and studied Applied Statistics. During college, he was awarded several private and public awards for his outstanding academic record. After completing college in Feb 2018, he worked as a research assistant in a statistical consultancy firm in Dhaka for a short period. But the quest for higher studies never ended for him, and in the spring of 2021, he joined the UTRGV to complete his master's degree in Applied Statistics and Data Science. UTRGV has awarded him with the prestigious presidential graduate research assistantship (PGRA) award for his previous accomplishment and support of his future glory. He completed his master's in applied statistics and data science in August 2022 at UTRGV. His research has always focused on how statistics and artificial intelligence can be applied to solve critical public health and health care problems. His future research will continue to be focused on developing AI-based solutions to real-world problems that otherwise will be difficult to address. He is joining the University of Arizona as a MIS Ph.D. student, where he aims to research NLP and healthcare. He can be reached at baraip007@gmail.com.