Final Report: Obesity and Public Health

Submitted to:

Dr. Casey Bennett

School of Computing and Digital Media

DePaul University

Report Prepared By:

Jason Nowak,

Natalia Guzman

March 20, 2025

1. Introduction

The global rise in obesity poses a significant challenge to health systems, driven by its complex interplay of social, economic, genetic, and environmental factors. It is estimated that if current trends continue, 20% of the world population will be obese (Hruby & Hu, 2015). The health impacts of obesity are severe and can significantly impact quality of life. Obesity disrupts almost all physiological processes leading to a number of diseases. It is known to contribute to chronic diseases such as type 2 diabetes, hypertension, and some cancers. In addition to elevated risks of chronic disease, a recent meta-analysis showed obese patients in a trauma department had a 45% higher likelihood of mortality, extended stays in the ICU, and increased complication rates (Hruby & Hu, 2015).

In the most dire projections, the USA is estimated to have over 85% of adults being overweight or obese by 2030. Accordingly, health-care costs attributable to obesity and overweightness are projected to double every decade if current trends continue (Wang et al., 2008). In 2019, the costs of obesity-related care reached $173 billion, with individual costs reaching $1,861 per adult and $3,097 for those with severe obesity (Ward et al, 2021). These figures just relate to the projected cost of caring for individuals with obesity. When you factor in indirect costs such as decreased workforce productivity, the figures become more stark.

In order to address this alarming trend we must first understand what obesity is. The WHO defines obesity as “abnormal or excessive fat accumulation that presents a risk to health.” Another factor that is considered when diagnosing obesity is the body mass index (BMI). This is a measurement of a person's body weight in kilograms divided by the square of height in meters. A figure above 25 kg/m2 is considered overweight and above 30 kg/m2 is considered obese. It’s important to understand that BMI is not correlated with body fat so it is not a holistic measure of obesity. Obesity is multifactorial ultimately resulting from chronic positive energy balance. This occurs when your body consumes more energy than it expends. Socioeconomic factors also play a significant role in obesity, as individuals with lower income and education levels often have reduced access to healthy foods, safe spaces for physical activity, and healthcare resources. This disparity contributes to higher obesity rates in underserved communities, highlighting the need for targeted public health interventions.

Beyond the physical health implications, the stigma around obesity can have a negative impact on mental health, exacerbating some of the issues people with obesity face. Consequently, the social stigma can lead to poorer outcomes in the clinical setting due to a focus on weight vs. actual underlying medical conditions, poor relationship between patient and provider, and reduced health seeking behavior due to fear of judgement. Therefore, a holistic approach that addresses not only the physiological aspects of obesity but also the psychological and social dimensions, including the impact of stigma, is essential for effective intervention.

2. Literature Review

In a data brief titled, “Obesity and Severe Obesity Prevalence in Adults: United States, August 2021–August 2023,” the National Center for Health Statistics lays out recent data collected on the prevalence of adult obesity and severe obesity by age, sex, and education level. The prevalence of obesity among adults from 2021 to 2023 was 40.3% with men standing at 39.2% and women at 41.3% (Emmerich et al, 2024). So overall, there was no significant difference in obesity prevalence

when comparing men to women. When looking at ages broken down into groups of 20-39, 40-59, and 60 and older we can see that the prevalence was highest among the age group of 40-59 (46.4%). The data brief also discusses the prevalence of obesity among adults with different education levels. There was a significant difference in obesity prevalence among adults with a Bachelor's degree or more (31.6%) and adults with some college (45.0%) or a high school diploma or less (44.6%). One limitation of this data is that obesity is defined by BMI, which we discussed earlier does not provide the entire context of the disease. This is because it doesn't account for factors such as muscle mass, which contributes to overall weight.

A similar data brief titled "Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017–2018," looked at obesity prevalence among non-Hispanic whites, non-Hispanic blacks, and Hispanic adults. Non-Hispanic Black adults exhibited the highest obesity prevalence at 49.6%, which is 4.8 percentage points higher than Hispanics (44.8%) and 7.4 percentage points higher than non-Hispanic Whites (42.2%). Non-Hispanic Asians had the lowest obesity prevalence by a wide margin at 17.4%. When factoring in gender, non-Hispanic black women had the highest obesity prevalence at 56.9%. This is roughly a 15% difference than their male counterparts. When comparing males to females across the other races there was no significant difference. Finally, the authors discuss differences in severe obesity prevalence. Severe obesity, also known as morbid obesity, is classified as having a body mass index of 40 or higher.Women had a severe obesity prevalence of 11.5% while men were at 6.9%. Non-hispanic black adults had the highest prevalence at 13.8% and non-Hispanic asian adults had the lowest at 2.0%.

As mentioned earlier, obesity is influenced by a complex combination of social and environmental factors. These are often referred to as the social determinants of health. This can include economic stability, education, healthcare access, neighborhood environment, social context, and race to name a few. Cultural differences also contribute to the disparity in obesity prevalence. In a study titled, "Social Determinants of Health, Health Disparities, and Adiposity," the authors looked at these socioeconomic factors and their relationship with obesity. Household socioeconomic status refers to income, occupation and education (Hattori & Sturm, 2023). Households with lower socioeconomic statuses tend to have a higher prevalence of obesity. Also, as we saw earlier, education has an inverse relationship with obesity prevalence as adults with more education tend to have lower obesity prevalence. The authors then discuss the neighborhood social environment's relationship with obesity. Perceived safety and social cohesion are particularly important, with studies showing that people who feel safer in their neighborhoods tend to have lower BMI over time. Social cohesion, trust among neighbors, and reduced physical disorder are consistently linked to lower obesity rates across racial and ethnic groups, especially among women. Additionally, neighborhoods with low access to healthy groceries have higher prevalence of obesity among children and adults. Residents in these neighborhoods tend to consume higher amounts of processed foods which are strongly linked to obesity.

Regarding biological and genetic risk for developing obesity, in recent years there was a discovery of the only genome-wide association study (GWAS) gene associated with obesity. The study of Loos et. al (2022), initially highlighted the importance of the fat mass and obesity-associated (FTO) gene in obesity and then revealed that individuals carrying specific variants in this gene tend to display higher BMI and body fat percentage compared to those

who do not. However, the authors also highlight that although the presence of FTO alleles can predispose individuals to obesity, it interacts with a wide range of environmental factors such as diet quality and frequency of physical activity. For example, populations with greater access to calorie-dense and nutrient-poor foods or limited opportunities for regular physical activities may be more vulnerable to the expression of this genetic risk into excess weight gain, while people with healthy eating habits and regular physical activity can mitigate some of the adverse effects of FTO variants. From a public health perspective, these findings emphasize that understanding genetic risk can help tailor interventions like a personalized dietary plan for those at high risk, but establishing how FTO variants interact with different environments is also crucial to developing precise preventive strategies.

Another study that supports this idea is *Epigenetics and Lifestyle* by Alegría-Torres et.al (2019). In this article, the authors explore how everyday exposures can influence gene expression through epigenetic modifications such as DNA methylation, histone modification, and non-coding RNA regulation that affects whether certain genes are on or off. The article shows that epigenetic changes might alter metabolic pathways, appetite regulation, or fat storage mechanisms. For example, the chronic consumption of high-calorie diets can produce lasting epigenetic marks that increase the possibility of energy storage in the adipose tissue. More importantly, these marks can sometimes be passed on to the next generations, meaning that environmental exposures can have intergenerational health consequences. These findings highlight once more the need for prevention strategies that account for inherited susceptibility and modifiable lifestyle factors to curb obesity risk.

Finally, translating all these genetic and epigenetic findings into effective programs for obesity prevention needs evidence-based guidelines. The US Preventive Services Task Force (USPSTF) provides a framework for evaluating and implementing interventions (JAMA, 2018). According to their recommendations, intensive behavioral counseling interventions (multiple sessions of dietary guidance, physical activity counseling, and self-management techniques) demonstrate moderate-high benefits. This kind of intervention leads to significant weight loss and risk reduction for conditions related to obesity like diabetes type 2 and cardiovascular disease. Furthermore, the USPSTF highlights that these interventions are more effective when they are frequent and sustained, using follow-up sessions to reinforce them. Additionally, regarding genetic and epigenetic factors, intensive counseling can be personalized to account for an individual's genetic predispositions that cause higher appetite or metabolic inefficiencies. At a population level, this also demonstrates the importance of insurance coverage for behavioral counseling, policies that support nutritional education, and environments that promote physical activity.

3. Dataset Description

The dataset we used for this project was obtained from Kaggle. It contains information about 2,111 individuals from an obesity study and includes 17 features that represent demographic characteristics, lifestyle habits, and obesity levels. Among the numerical features, you can get information about a person's age, height (in meters), weight (in kilograms), frequency of vegetable consumption (FCVC), number of main meals (NCP), daily water consumption in liters (CH2O), physical activity frequency in a week (FAF), and time spent using technology devices per day (TUE). On the other hand, the categorical variables include the individual's gender (male or female), family history of being overweight, frequent consumption of high-calorie food (FAVC), consumption of food between meals (CAEC), smoking status (SMOKE), calorie counting (SCC), alcohol

consumption frequency (CALC), means of transportation (MTRANS), and the obesity classification (NObeyesdad). The obesity classification ranges from insufficient weight to normal weight and various levels of overweight and obesity type I, II, or III.

4. Data Analysis and Results

4.1 Data Visualization Analysis

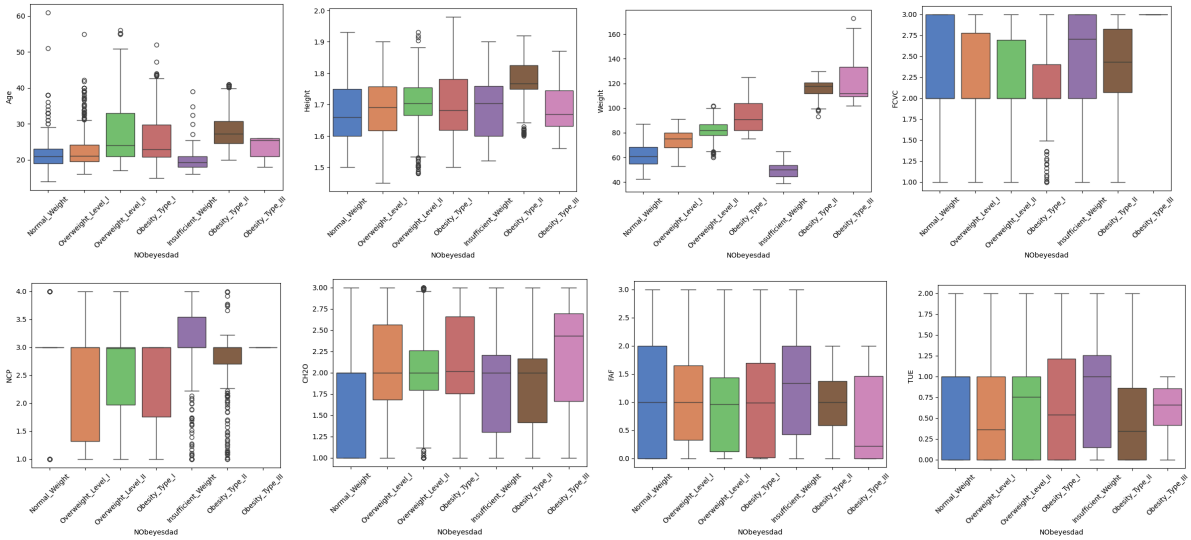

Figure 1. Numerical Variables vs. Obesity Level

We decided to analyze the distribution of our numerical variables using boxplots (Figure 1). When we look at the relationship between these variables and obesity levels we get some meaningful information. For age vs. obesity, we noticed a flat pattern meaning that individuals can be underweight or overweight regardless of the age, which makes us think that this is a problem that has to be tackled since the younger generations. We noticed a similar pattern for height, which means that this variable by itself has limited explanatory power. On the other hand, when we look at weight, as expected, we found a linear positive relationship with obesity since this is the most direct influencer of obesity classification. For frequency of vegetable consumption, we can notice that the groups normal weight and insufficient weight have a slightly higher consumption of vegetables, while obesity type I has the lowest median. However, the wide distribution across groups suggests that while a higher consumption of vegetables can support a healthy weight, it needs to be paired with other healthy lifestyle habits. When we analyze the number of main meals during the day, we see that normal weight and insufficient weight lean towards a median of 3 meals, while obesity type I, II, and III lean closer towards 2-3 meals. However, the overall results show that the number of mean meals is not by itself predictive of obesity. When we look at the consumption of water vs. obesity relationship, we notice that there are similar ranges across the groups, meaning that individuals of any weight level have different patterns of water intake. Regarding frequency of physical activity, we noticed a lower median in the obesity type II and III groups (obesity type III is even closer to 0 which means no exercise during the week), suggesting reduced overall activity among individuals in these categories. However, to our surprise, there are a lot of sedentary people regardless of the weight group which is concerning. Finally, the median of time spent using technology devices ranges from 0.5 to 1 hour in most groups, showing that this variable is not a strong predictor of obesity by itself.

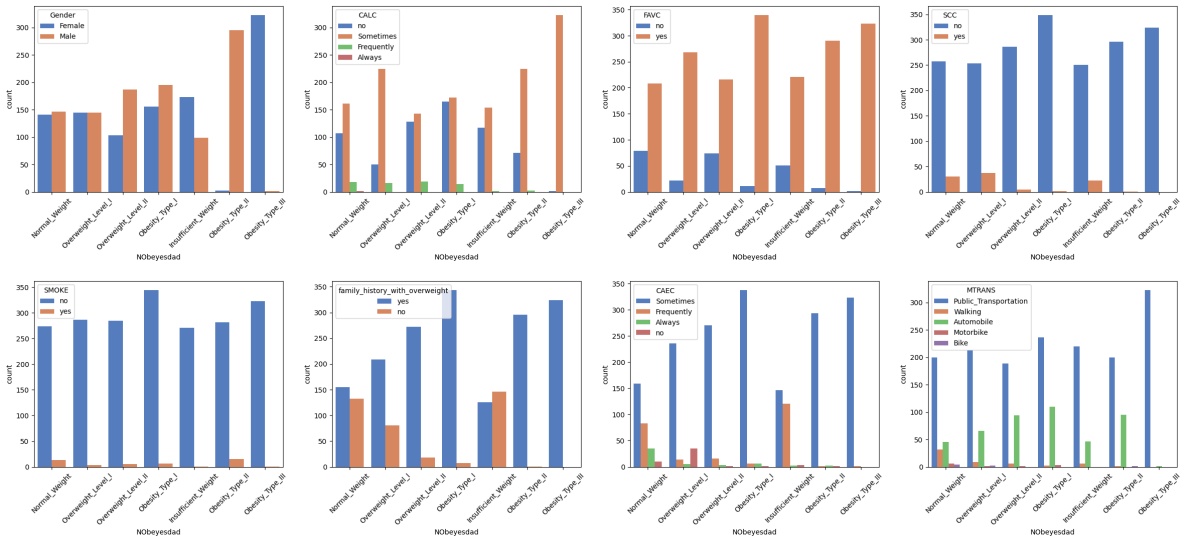

Figure 2. Categorical Variables vs. Obesity Level

We used count plots to analyze the relationship between the categorical variables and the obesity level (Figure 2). When we analyze the influence of gender in obesity, there is no clear pattern since we see for example a larger number of men in the group obesity type II compared to women, but the exact opposite in the group obesity type III. When we look at the alcohol consumption variable, we notice that regardless of the weight level, most people report drinking alcohol occasionally, while 'frequently' and 'always' are rare. For the variable consumption of high-calorie food, we noticed that it is frequent across all weight groups. However, the difference between 'yes' and 'no' is a lot higher in obesity type I, II, and III. We also noticed that calorie tracking is rare across all groups, since there majority of individuals in the dataset report 'no' to this variable. For the factor 'SMOKE', we noticed a similar pattern with most of the people reporting that they do not smoke, so there is no clear relationship between smoking and obesity level. When we see the family history of overweight, we can see that there is a larger amount of 'no' in the groups normal weight and insufficient weight. On the other hand, the proportion of 'yes' is significantly higher in the three groups of obesity, suggesting a stronger familial/genetic relationship to having obesity. When we look at the frequency of food consumption between meals, we see that occasional snacking is the most common pattern across all groups. The graph does not show that a positive relationship between snacking frequently and obesity. Finally, when we analyze the mean of transportation, we see that public transportation is the most common in all groups, followed by the use of automobiles. Although there is no clear pattern between the type of transportation and obesity, we did notice a considerable larger amount of individuals that walk across the normal weight group comparing to the other ones.

In general, by visually analyzing the different variables vs. the obesity level, we did not notice any clear relationship between them (except for the positive correlation between weight and obesity and a slight correlation between less physical activity and obesity type I and II). This reveals that there is not one single behavior that defines obesity risk and that obesity is a multifactorial problem that involves a combination of lifestyle, familial, and demographic factors.

4.2 Predictive Analysis - Logistic Regression Model

For our predictive analysis, we chose to create a logistic regression model due to its interpretability, efficiency, and speed, which makes it practical for our analysis. In the data preprocessing stage, we checked for missing data or any duplicated rows. We did not find any missing values, but we had 24 duplicated rows which we decided to eliminate. After this, we decided to perform label encoding in the categorical variables and separated the features and the target. Finally, we performed normalization by using 'StandardScaler' to ensure that the features are consistent and to make the training and model more stable.

To build our model, we split the data into 75% for the training set and 25% for the testing set. Then we used 5-fold cross-validation with the training set to evaluate our model and then checked its performance using accuracy. Finally, we made predictions on the test set and evaluated them with a confusion matrix (Figure 3) and the metrics of precision and recall.

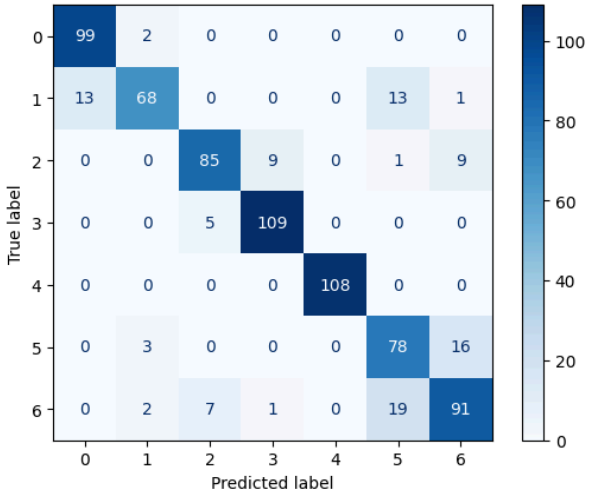

Figure 3. Confusion Matrix for Logistic Regression Model

In our results, we got an accuracy of 85% when we performed cross-validation which means that our model is consistent and has a strong predictive power. When we look at the confusion matrix, the diagonals represent the correct number of predictions for each class. As seen, every class has a high count of correct predictions, demonstrating that the model is classifying them accurately. Although there are some misclassifications in classes 5 and 6 (Obesity type II and III), this could mean there is some difficulty when differentiating between these two classes. Finally, our model got an average precision of 85.6% and an average recall of 85%, which makes it a good model for predicting obesity levels based on lifestyle, familial, and demographic factors.

5. Proposed Interventions

As we could notice in our analysis, obesity remains a problem that impacts individuals of all ages and demographics. Therefore, our interventions will adopt a community-based approach tailored to the general public, focusing on buckets two and three of the CDC's HI5 model.

Initially, to extend our intervention outside healthcare facilities, we plan to create an online website or mobile application that allows the public to self-assess their obesity risk

through a survey that will be included. This survey would include questions that mirror the features of our dataset like frequency of vegetable consumption, daily water intake, and physical activity. Once the individuals submit their responses, our model classifies their risk level and generates personalized recommendations. For example, a person who gets a result of 'high risk' for obesity would receive suggestions to go to a primary care physician, consume two liters of water per day, do at least 30 minutes of physical activity five times a week, and increase vegetable intake to around 3 servings daily. On the other hand, a person who gets 'moderate risk' would be recommended to log calorie intake, reduce sugar consumption, and suggestions for moderate workouts. Additionally, our platform would connect users to local resources based on their location, such as nearby farmers' markets, healthy restaurants, walking trails, gyms, and public parks. The platform would also send push notifications or emails to the users reminding them to meet daily goals. Finally, we would also incorporate a review section that allows the users to share common barriers that they face and ideas for overcoming them, which would be analyzed by health officials to tailor interventions more effectively.

Additionally, we propose policy and environmental strategies that focus on the community. First, urban planning interventions can increase green spaces, ensure sidewalks and bike lanes are well maintained, and spread recreational activities like yoga or Zumba in the park across various neighborhoods, taking advantage of the public parks around the cities. We think that supporting these amenities will encourage people to do activities like walking, jogging, and cycling, which reduce the sedentary behavior that we saw in our analysis. Through our data, we also noticed that obesity is present in younger generations as well. Therefore, we also propose that schools offer healthier meal options and interventions like early education on nutrition and a healthy lifestyle. Finally, we would have to launch large-scale awareness campaigns that reinforce key healthy lifestyle habits. In this section, it would also be important to include a dashboard (Figure 4) that helps public health authorities to monitor obesity trends and lifestyle factors, as well as measure the impact of community interventions over time.

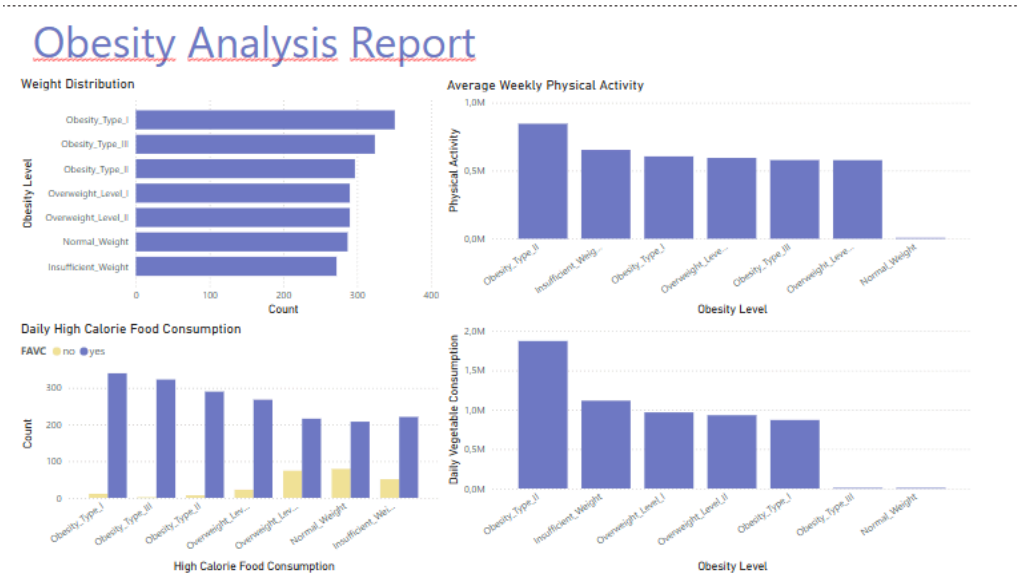

Figure 4. Dashboard Report Draft

Finally, to support these interventions, we propose a data pipeline (Figure 5) that collects user survey responses into a secure server. Then, an ETL (Extract, Transform, Load)

process would extract survey data from the platform, transform it by encoding categorical features and scaling numeric values to match the logistic regression model's training format, and load standardized data into a central database for long-term storage. Once the data is loaded, the model deployment layer runs each response through the logistic regression algorithm, generating an instant result. The system stores results to allow ongoing retraining or fine-tuning of the model as more users participate. Over time, this feedback loop ensures the predictions remain accurate for diverse populations and behaviors. Additionally, the results would constantly update the dashboard used by public health officials for them to get insights, track progress, and improve policies.

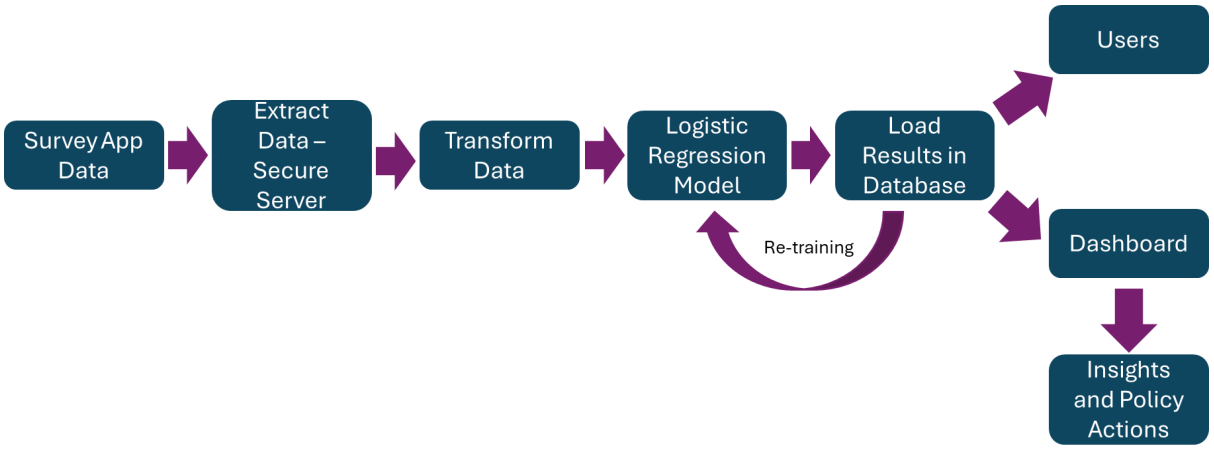

Figure 5. Data Pipeline

6. Conclusions

This analysis highlights the multifactorial nature of obesity and the continued burden it has had on our society. In addition to its impact on physical health, our review of the literature highlights the broader social and economic consequences associated with the increasing prevalence of obesity in the US. Our data analysis underscores that while certain lifestyle choices (e.g., physical activity) correlate with obesity status, no single factor alone fully predicts an individual's weight classification. Instead, obesity arises from a confluence of factors including familial and genetic predispositions, socioeconomic constraints, and everyday behaviors. With this information, we were able to develop a logistic regression model with 85% accuracy through cross-validation, demonstrating that data-driven methods can effectively classify obesity levels based on demographic and lifestyle features.Through our research and data analysis, we sought to develop a comprehensive, community-wide intervention designed to achieve measurable positive health outcomes within a five-year timeframe. Our goal was also to achieve long term cost savings for the population. With those goals in mind, we proposed an intervention that includes the development of an online platform and mobile app for self-assessment and personalized recommendations. By deploying this, we can empower individuals to understand and manage their obesity risk. In addition, we proposed policy and environmental changes to provide communities with more access to safe healthy living spaces. Our holistic approach addresses the complex interplay of individual behaviors, social influences, and environmental conditions. With constant monitoring and revisions of this intervention, we should achieve healthy outcomes. As more individuals and communities engage, the collective insights gained will help pave the way for sustainable, long-term success in obesity prevention and management.

7. References

Alegría-Torres, José A., et al. “Epigenetics and Lifestyle.” *Epigenomics*, 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC3752894/

Emmerich SD, Fryar CD, Stierman B, Ogden CL. Obesity and severe obesity prevalence in adults: United States, August 2021–August 2023. NCHS Data Brief, no 508. Hyattsville, MD: National Center for Health Statistics. 2024. DOI: https://dx.doi.org/10.15620/cdc/159281.

Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity and severe obesity among adults: United States, 2017–2018. NCHS Data Brief, no 360. Hyattsville, MD: National Center for Health Statistics. 2020.

Key J, Burnett D, Babu JR, Geetha T. The Effects of Food Environment on Obesity in Children: A Systematic Review. *Children (Basel)*. 2023;10(1):98. Published 2023 Jan 3. doi:10.3390/children10010098

Loos, Ruth J. F., and Giles S. H. Yeo. “The Bigger Picture of FTO—the First GWAS‐Identified Obesity Gene.” *Nature Reviews Endocrinology*, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC4188449/

Pavela G, Lewis DW, Locher J, Allison DB. Socioeconomic Status, Risk of Obesity, and the Importance of Albert J. Stunkard. *Curr Obes Rep*. 2016;5(1):132-139. doi:10.1007/s13679-015-0185-4

U.S. Preventive Services Task Force. “Behavioral Weight Loss Interventions to Prevent Obesity-Related Morbidity and Mortality in Adults: Updated Recommendation Statement.” *JAMA*, 2018. https://jamanetwork.com/journals/jama/fullarticle/2702878

Wang Y, Beydoun MA, Liang L, Caballero B, Kumanyika SK: Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. Obesity (Silver Spring). 2008, 16:2323-30. 10.1038/oby.2008.351