Risk Assessment Analysis of Obesity with Machine Learning

Gulnaaz Shaikh (gs3058@rit.edu)

Section 1 – Introduction

Obesity is one of the most common health problems affecting millions of people throughout the country [1]. and, also contributing to chronic diseases in the long run. Often known as excess body fat, obesity is a complex condition that is also affected by genetics [2]. Many factors excluding food could also lead to obesity. Health care systems also face a significant amount of financial constraint as well when it comes to treating obesity. Majority of the people would not be able to afford the treatments and other procedures required to treat obesity, leading to other chronic diseases and eventually even death. Therefore, it is essential to analyze risk factors that can cause obesity for early assessment and time-effective treatment planning. This project aims to analyze obesity and the risk factors affecting obesity in the US with the help of machine learning models.

Section 2 – Problem

2.1Problem Definition

Since, obesity levels are rising across the country, better risk assessment analysis for predicting obesity are required. With the help of machine learning algorithms, the goal of this project is to identify various factors that need to be taken into consideration when analyzing the cause of obesity. The different risk factors that this project aims to analyze are **Demographics**, **Body Measures**, **Diabetes**, **Physical Activity**, **Sleep Disorders**, **Smoking**, and **Mental Health.** This project will primarily benefit health care professionals, individuals with obesity and other researchers.

2.2 Significance of the Problem

Obesity places a significant financial strain on healthcare systems and is associated with serious health problems. By facilitating early interventions, and an accurate risk assessment tool we can lower the health risks and related expenses associated with obesity. We can increase the prediction accuracy and enable more focused preventive actions by utilising machine learning techniques. With the help of this project, it aims to answer the following research questions:

1. What factors affect obesity? How can we avoid them?
2. What would be the relationship between different factors? Does one factor affect the other? Eg. How is BMI affected by other risk factors?
3. Which is the best model for prediction of risk factors?

Section 3 – Prior Work

3.1 Machine Learning Applications in Obesity Prediction

Over the years, machine learning has gained significant popularity in prediction. [3] site an in-depth analysis to predict the risk of obesity in Bangladesh using nine classification models. Their highest accuracy was achieved through logistic regression with 79.09%. Similarly, [4] utilized public datasets and performed statistical and ML techniques to identify the key factors. However, they faced challenges when it came to integration of heterogenous data sources. Additionally, big data analytics were explored for ML models to predict obesity trends by [2]. Furthermore, [5] delved into the psychological factors in obesity classification. They also emphasized advantages of machine learning over the traditional statistical methods.

3.2 Key Features influencing Obesity Risk

Biological, behavioral, and psychological aspects have been identified as influential key features in multiple studies. Behavioral interventions for the treatment of obesity were highlighted by [6], especially for children and young adults. Nevertheless, [5] summarizes the negative psychological factors associated with BMI in their research. They shed light on how weight related disorders are influenced. Additionally, [7] have performed investigation on the correlation between BMI and mental illness, which revealed a significant association between depression, anxiety, and obesity. Similarly, [2] also emphasized on how socioeconomic and demographic factors determine obesity; by incorporating these factors, they leveraged big data analytics for improvement in risk assessment. Based on these findings, we could interpret that the mental well-being of an individual is highly likely to affect their weight.

3.3 Challenges and Limitations in Existing Research

Despite, the progress made in machine learning for obesity risk management, there are various challenges that persist. One of the main issues with different studies was found the heterogenous nature of data. Multiple datasets happen to originate from different populations making the integration almost impossible. Furthermore, the limitations imposed due to sample size affect the models as highlighted by [3]. They also suggest that a larger dataset could improve their model accuracy in prediction and cover a broader spectrum of obesity categories. Moreover, [7] were challenged by the model's moderate accuracy, emphasizing the need for a larger dataset and more sophisticated algorithms. The need for more data-driven approaches and public health interventions for enhancing obesity prediction were stressed by [2]. Lastly, [4] proposed that the classification performance can be improved by enhancing the ML models with non-convex optimization and neural networks.

Section 4 – Proposed Methodology

4.1 Plan

The project follows a structured approach to develop and evaluate the risk assessment model for obesity, the plan would be as follows:

1. **Data Preprocessing:** This is one of the most important and vital steps especially in this project. The data that I will be using comes from https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023, this dataset contains large number of files in .XPT format for the various parameters like **Demographics**, **Body Measures**, **Diabetes**, **Physical Activity**, **Sleep Disorders**, **Smoking**, and **Mental Health** aspects. Combining this dataset based on the SEQN number would be the first step to obtain the correct data. Next, I would then handle, missing and null values if any to obtain the cleaned dataset. Furthermore, splitting the data into training and testing data would be followed.

2. **Exploratory Data Analysis:** The cleaned dataset would then be used for exploratory data analysis. Identifying correlations and implementing preliminary visualizations would be a part of this process. The data would then be normalized for model implementation using mean-max transformation.

3. **Model Development:** The data would be first split into training and testing sets. Additionally, machine learning algorithms would then be applied to the prepared data to analyze the risk assessment of obesity. I plan to implement Classification and Regression models like Logistic Regression, SVM, Random Forest, Decision Tree, and Gradient Boosting. A baseline classification model would also be set as a benchmark for comparison and evaluation purpose.

4. **Testing and Tuning:** The model would then be assessed with testing data. Based on the performance of the baseline model, tuning would be done to improve confusion matrices. This step is essential as tuning would improve the accuracy of models and provide better prediction.

5. **Model Assessment:** After implementation, I can identify the best model for this project and draw conclusions regarding the features. Model would then be evaluated based on the Precision, Recall and F-1 scores. Precision-Recall curves can be used to assess the performance of these algorithms.

6. **Dashboard:** If time permits, I would also like to implement a dashboard for an interactive experience where I can correlate different factors with each other and gain some visual insights.

7. **Conclusion:** I intend to summarize my findings and gain important insights from the predictions made.

4.2 Challenges or Barriers

One of the main challenges that I might face would be merging the said datasets. The datasets are voluminous and have large number of respondents. The challenge here would be to analyze and clean my dataset so that it would be beneficial to train the model and therefore, make predictions. This challenge can be overcome as every dataset has a sequenced number for each respondent. Merging and analyzing the data based on their sequence would be my approach to process this barrier. Furthermore, I think the accuracy of the said model should be high. While working on the development stage, it would be crucial to carefully implement the machine learning models to obtain a high accuracy.

4.3 Project Deliverables

The key project deliverables for this project are:

a. A cleaned dataset which is preprocessed for suitable machine learning modelling
b. Source code for all machine learning models implemented
c. Model evaluation and assessment metrics
d. Dashboard for risk factors

References

[1] C. L. Ogden, M. D. Carroll, B. K. Kit, and K. M. Flegal, "Prevalence of childhood and adult obesity in the united states, 2011-2012," JAMA, vol. 311, no. 8, pp. 806–14, 2014, doi: https://doi.org/10.1001/jama.2014.732.

[2] G. Vemulapalli, Sreedhar Yalamati, Naga Ramesh Palakurti, N. Alam, Srinivas Samayamantri, and Pawan Whig, "Predicting Obesity Trends Using Machine Learning from Big Data Analytics Approach," pp. 1–5, Jul. 2024, doi: https://doi.org/10.1109/apcit62007.2024.10673429.

[3] F. Ferdowsy, K. S. A. Rahi, Md. I. Jabiullah, and Md. T. Habib, "A machine learning approach for obesity risk prediction," Current Research in Behavioral Sciences, vol. 2, p. 100053, Nov. 2021, doi: https://doi.org/10.1016/j.crbeha.2021.100053.

[4] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview," Sensors, vol. 20, no. 9, p. 2734, May 2020, doi: https://doi.org/10.3390/s20092734.

[5] G. Delnevo, G. Mancini, M. Roccetti, P. Salomoni, E. Trombini, and F. Andrei, "The Prediction of Body Mass Index from Negative Affectivity through Machine Learning: A Confirmatory Study," Sensors, vol. 21, no. 7, p. 2361, Mar. 2021, doi: https://doi.org/10.3390/s21072361.

[6] D. E. Wilfley, J. F. Hayes, K. N. Balantekin, D. J. Van Buren, and L. H. Epstein, "Behavioral Interventions for Obesity in Children and adults: Evidence base, Novel approaches, and Translation into practice.," American Psychologist, vol. 73, no. 8, pp. 981–993, Nov. 2018, doi: https://doi.org/10.1037/amp0000293.

[7] Reya Pillai R, Suchitra Saravanan, and Gopal Krishna Shyam, "The BMI and Mental Illness Nexus: A Machine Learning Approach," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 526–531, Oct. 2020, doi: https://doi.org/10.1109/icstcee49637.2020.9277446.