Risk Assessment Analysis of Obesity with Machine Learning

Gulnaaz Shaikh (gs3058@rit.edu)

Table of Contents

Table of Figures

Tables

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

2.1 Problem 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**, **Examination dataset (blood pressure, etc), Laboratory dataset (cholesterol, blood profile etc), Dietary data** 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

0.8 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 – Methodology

This project incorporates a methodology that is structured in a multi-step process where I have used the data from NHANES 2021-2023 to prepare, explore, build predictive models, fine-tune it and then visualize the results for a better decision-making process.

4.1 Data Sources

The data is taken from NHANES 2021-2023 datasets, out of which I have utilized the following separate datasets:

a. Dietary Dataset
b. Laboratory Dataset
c. Examination Dataset
d. Demographics
e. Body Measures
f. Diabetes
g. Physical Activity
h. Sleep Disorders
i. Smoking

4.2 Merging

The 9 datasets taken from NHANES 2021-2023 were then merged into one dataset for further analysis. All the datasets were merged using the "SEQN" feature as it represented the respondent number taken down during data collection. The merged dataset contained 13137 rows and 18+ variables.

4.3 Data Cleaning

The ideal next step was cleaning the dataset after merging to perform the following:

a. Removed missing/null/duplicate values and imputed them using mean/mode
b. Categorize the non-numeric data into categories for better understanding
c. Focused on adults ages 20-60 years and eliminated children

4.4 Exploratory Data Visualization

Additionally, I explored the data and performed preliminary data visualizations to get a better understanding of the dataset and also gain insights into the otherwise non-visual datasets. Figure 1 depicts all the necessary data used initially, before adding new data from NHANES 2021-2023 i.e the Dietary Dataset, Laboratory Dataset and Examination Dataset. Figure 2 moves on to the newly added data and some EDA for the new datasets.

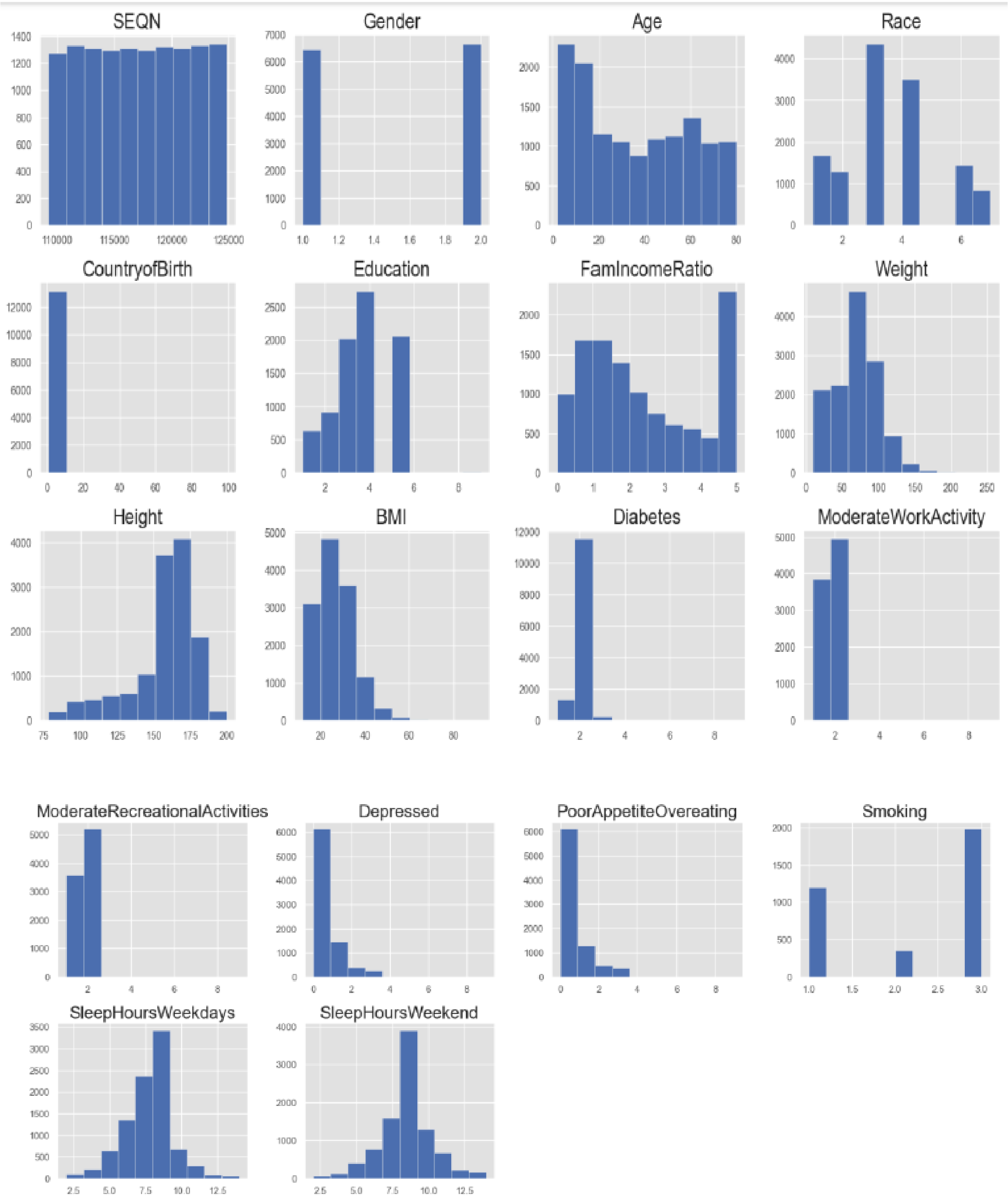


Figure 1: Preliminary Data Visualization

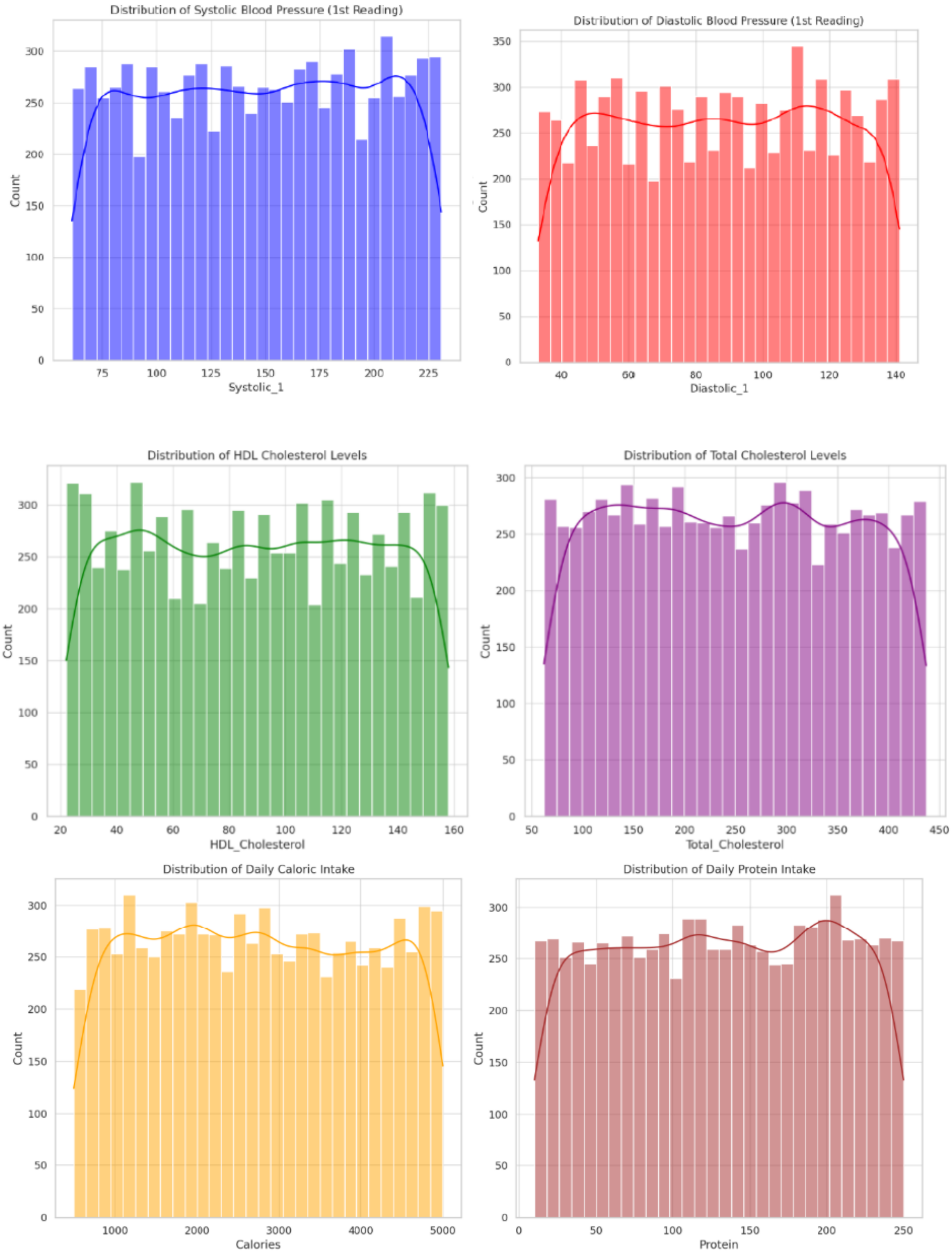


Figure 2: Preliminary Data Visualization on New Data

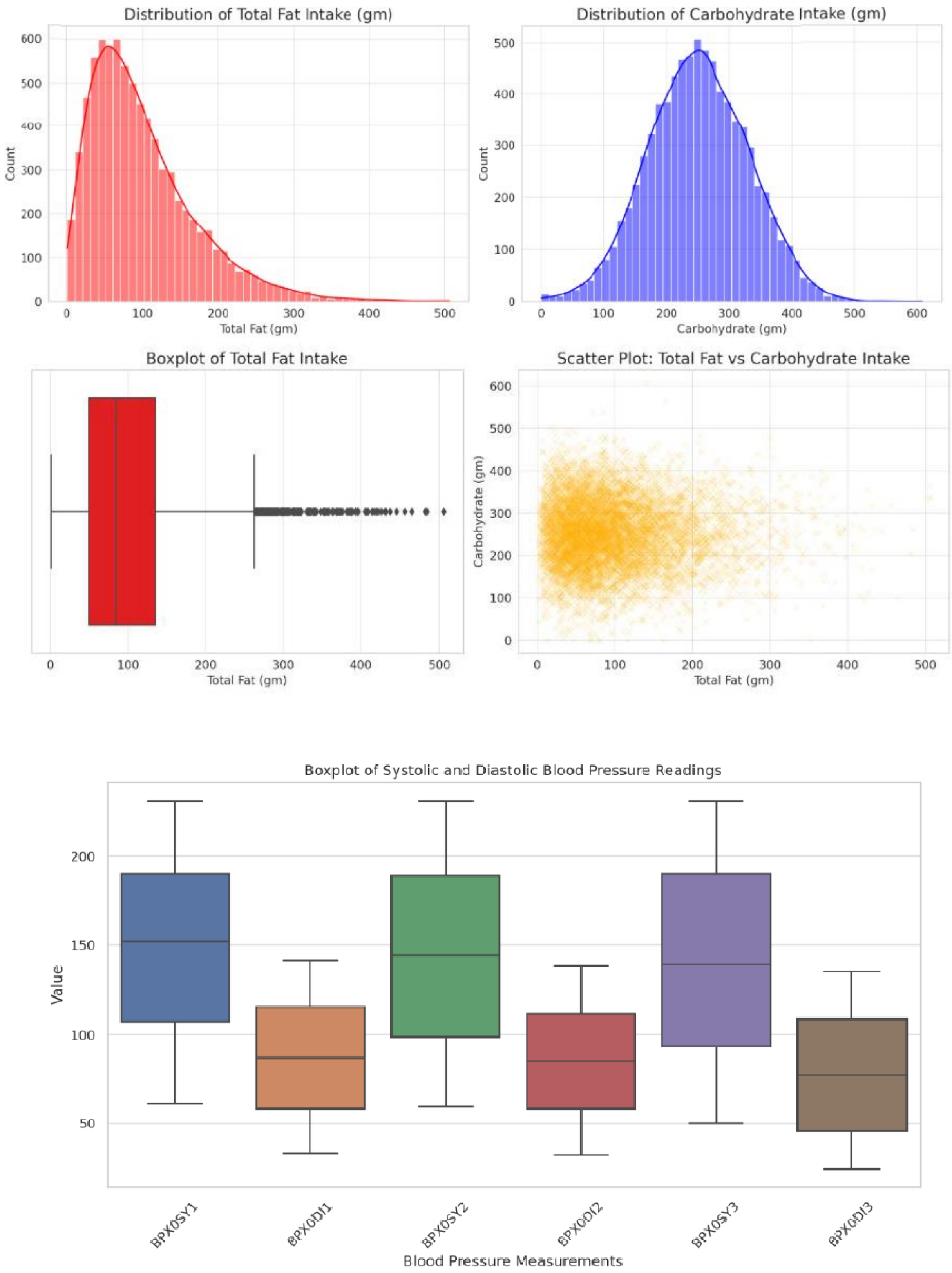


Figure 3: Preliminary Data Visualization on New Data

Preliminary Conclusions:

1. Compared to male respondents, female respondents have a somewhat higher BMI.
2. In this dataset, non-Hispanic Asians have a lower BMI than other racial groupings.
3. Those who have completed college or higher typically have lower BMIs.
4. Those who reported feeling down "more than half the days" had higher BMIs.
5. Participants with higher cholesterol levels tended to have higher BMI values, indicating a link between lipid profiles and obesity.
6. Individuals with lower levels of physical activity showed consistently higher rates of overweight and obesity.
7. Higher systolic blood pressure readings were moderately associated with higher BMI values.
8. Increased daily sugar intake, as reported in dietary records, was linked to elevated obesity risk across all age groups.

4.5 Feature Engineering and Selection

- **New Column created:**
  - Created categorical variable Obesity_Level based on BMI.
- **Feature Engineering:**
  - Performed categorical encoding for Gender, Race, Smoking status, etc.
  - Utilized MinMax normalization for numeric features.
- **Feature Selection:**
  - Post-modeling feature importance analysis revealed that Age, Dietary Habits (Sugar/Fat), Physical Activity Level, Sedentary Behavior, and Income Ratio were seen as top predictors.

4.6 Model Building and Training

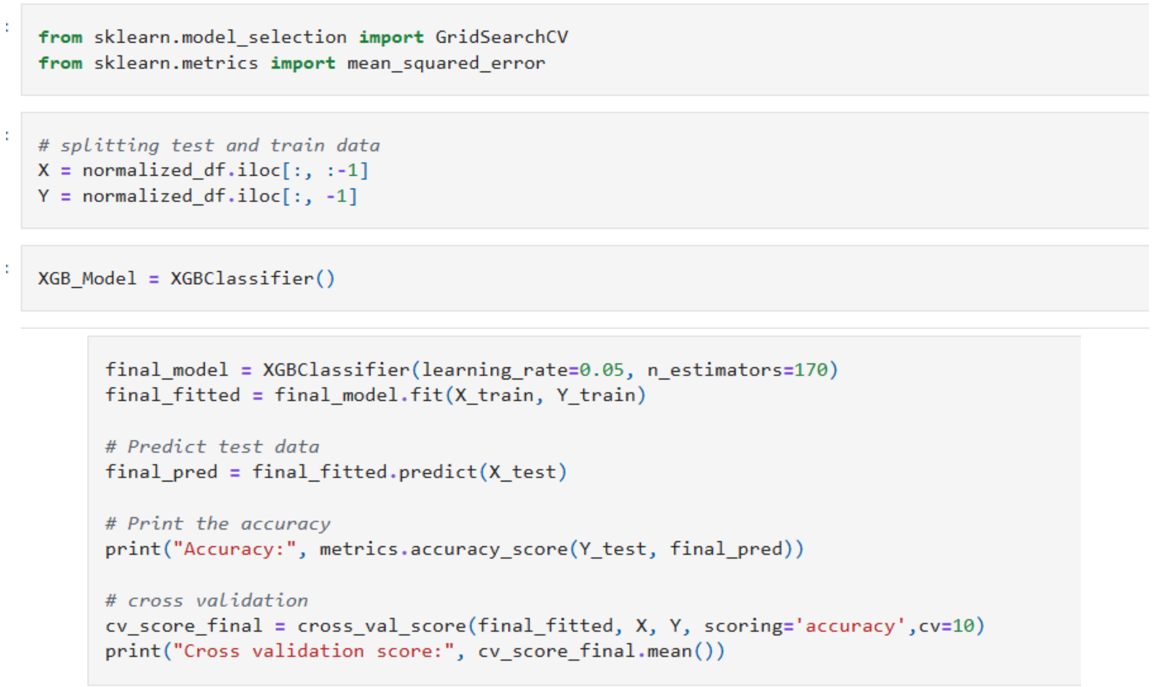
Machine learning models were trained for the purpose of predicting and assessing high risk metrics for obesity. The following are the 5 models implemented in this project:

- Logistic Regression: a simple method that gives prediction based on two categories- yes or no
- Support Vector Machine (SVM): finds the best boundary to separate different categories as clearly as possible
- Decision Tree: splits data into branches based on questions and helps in decision making step-by-step for final outcome
- Random Forest: builds on decision trees and combines the results later to improve accuracy
- XGBoost: uses multiple small trees sequentially while improving mistakes on previous ones for better accuracy

4.7 Hyperparameter Tuning

Hyperparameter tuning is implemented to achieve better and more accurate results on the models already implemented. Here, I have utilized GridSearchCV to optimize Random Forest and XGBoost parameters. For evaluation purpose, I have the following evaluation metrics to determine the best results:

- Accuracy, Precision, Recall, F1-score.
- Precision-Recall Curves for deeper performance insights



```
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_squared_error
```

```
# splitting test and train data
X = normalized_df.iloc[:, :-1]
Y = normalized_df.iloc[:, -1]
```

```
XGB_Model = XGBClassifier()
```

```
final_model = XGBClassifier(learning_rate=0.05, n_estimators=170)
final_fitted = final_model.fit(X_train, Y_train)

# Predict test data
final_pred = final_fitted.predict(X_test)

# Print the accuracy
print("Accuracy:", metrics.accuracy_score(Y_test, final_pred))

# cross validation
cv_score_final = cross_val_score(final_fitted, X, Y, scoring='accuracy',cv=10)
print("Cross validation score:", cv_score_final.mean())
```

Figure 4 : Hyperparameter Tuning

4.8 Dashboard

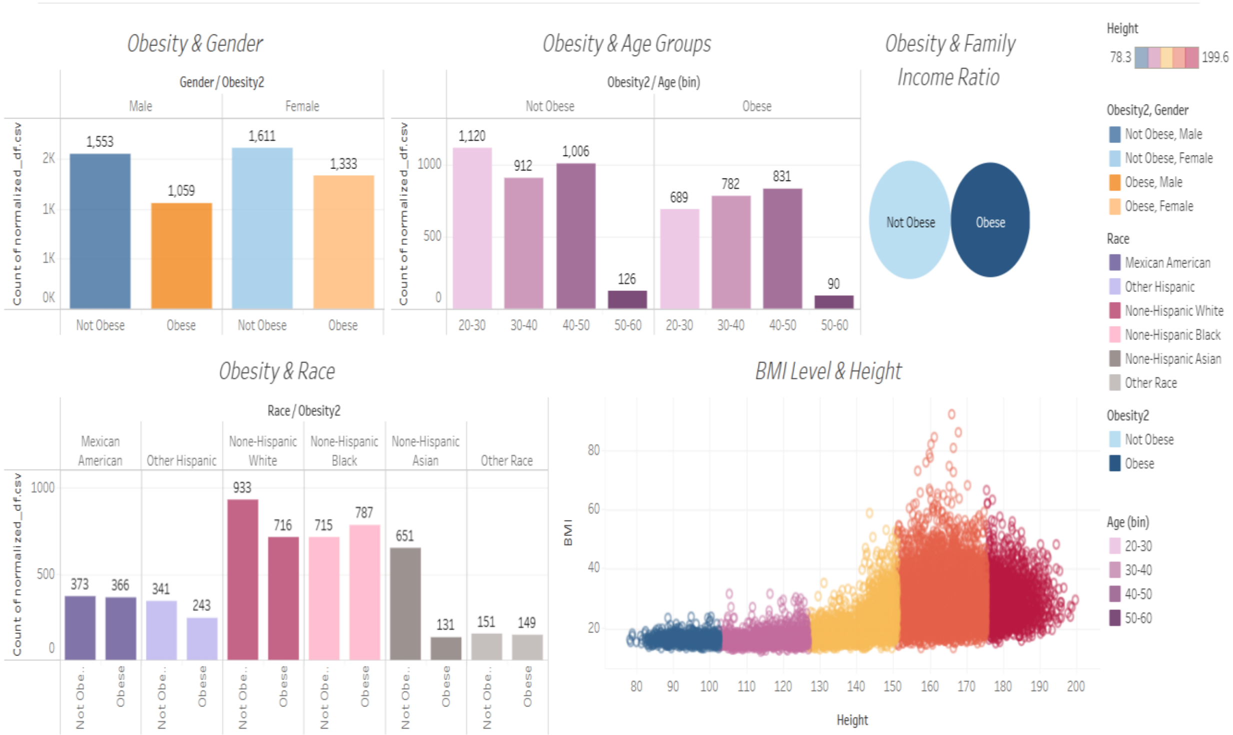


Figure 5: Tableau Dashboard

Results from visualizations and model prediction helped in creating an interactive dashboard for identifying the causes and insightful information that leads to obesity in adults. The dashboard was implemented using Tableau and it visualizes the relationships between various risk factors and obesity levels. Interactivity is added on filtering age group, activity levels, gender, and dietary patterns.

Section 5 – Experiments and Results

Finally, after model implementation and fine-tuning here are the results from all the experiments done during the course of this project. For a better understanding I have used Table 1 for summarized results throughout the project. The evaluation metrics are F1 score, cross Validation score and Accuracy.

Model	Accuracy	Cross-validation Score
Baseline Model	57.64%	
SVM	68.23%	0.6545
Logistic Regression	65.51%	0.6328
Decision Tree	61.25%	0.5956
Random Forest	75.35%	0.7375
Random Forest using GridSearchCV	76.55%	0.7474
XGBoost	78.65%	0.7725
XGBoost with tuning	79.81%	0.7870

Table 1: Summary Results

Furthermore, I implemented a graph as shown in Figure 6 for precision-recall curves. This helps understand the best model for this project.

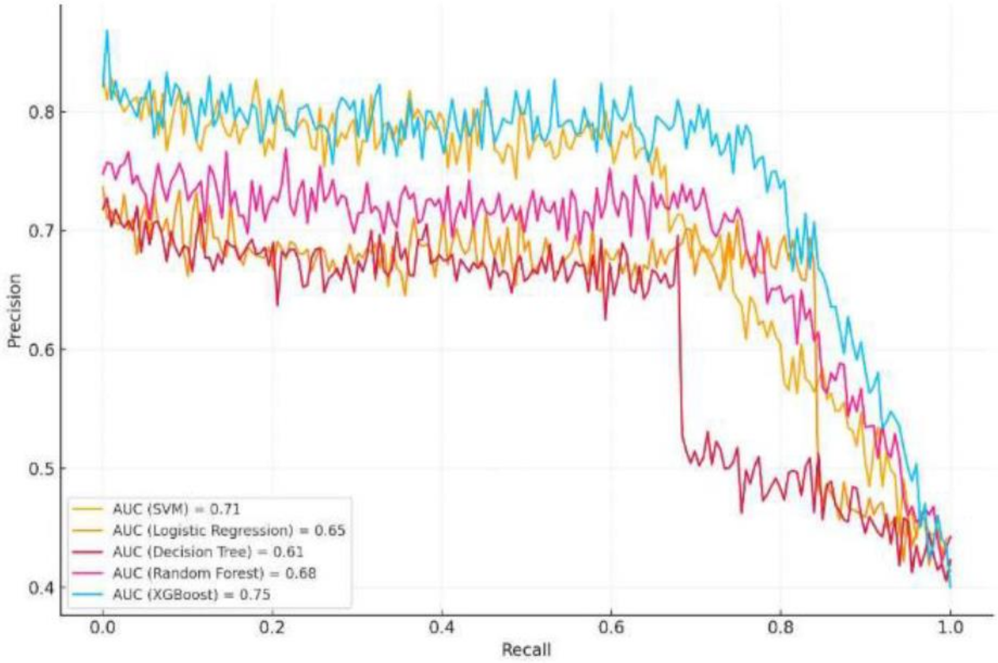


Figure 6: Precision-Recall Curve

Final Model - XGBoost:

- Highest overall accuracy (79.81%) post-tuning.
- Top Features influencing obesity:
    - Age
    - Total Fat/Sugar Intake
    - Physical Activity
    - Sedentary Behavior
    - Income Ratio

Model Summarization:

- Logistic Regression performs lower, which is expected since it’s a simpler model.
- Decision Trees perform worse due to overfitting, which is common.
- Random Forest & XGBoost perform the best, as ensemble methods tend to perform better as they combine multiple decisions which means less overfitting and better generalization.

Section 6 – Conclusions and Future Work

XG Boost Model provided the best accuracy score compared to other models and fine tuning improved the accuracy from 78.65% to 79.81%. The top features using this model were identified as Age, Dietary Habits (sugar/Fat), Physical Activity level, Sedentary Behavior, and Income Ratio.

Machine learning, especially ensemble techniques like XGBoost, offers a strong foundation for evaluating obesity risk. Important modifiable risk variables include physical activity, socioeconomic status, and dietary practices. The project's over 80% accuracy rate shows that machine learning algorithms can help with early obesity prevention initiatives.

Moreover, future work can be done by incorporating time-series data for accurate trend analysis, more investigation can be done on neural network designs for increasing accuracy as well as real-time prediction dashboards can be implemented. This project serves as a stepping stone in a long line of work that can benefit the future generations and prevent the risk of obesity in individuals.

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.

[8] Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey: 2021–2023 Data Documentation, Codebook, and Frequencies. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023