“BioDoopDetective: Predictive Analysis on Risk Factors Associated with Obesity/Overweight”

MSDSM Project Report

By
DIVYANSH GADWAL
(2104107017)
MSDSM Batch-1

INDIAN INSTITUTE OF TECHNOLOGY INDORE
INDIAN INSTITUTE OF MANAGEMENT INDORE

FEBRUARY 2024

"BioDoopDetective: Predictive Analysis on Risk Factors Associated with Obesity/Overweight"

A PROJECT

Submitted in partial fulfillment of
the requirements for Term VI
of
Master of Science in Data Science and Management

by
Divyansh Gadwal
(2104107017)
MSDSM Batch-1

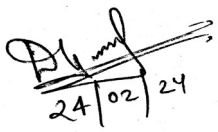
INDIAN INSTITUTE OF TECHNOLOGY INDORE
INDIAN INSTITUTE OF MANAGEMENT INDORE

FEBRUARY 2024

CANDIDATE'S DECLARATION

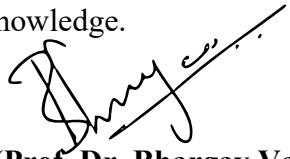
I hereby certify that the work which is being presented in the project entitled **BioDoopDetective: Predictive Analysis on Risk Factors Associated with Obesity/Overweight** in the partial fulfillment of the requirements for term six of **MASTER OF SCIENCE IN DATA SCIENCE AND MANAGEMENT, JOINTLTY OFFERED BY Indian Institute of TechnologyIndore** and **Indian Institute of Management Indore** is an authentic record of my own work carried out during the time period from November 2023 to January 2024 of MSDSM Project under the guidance of **Prof. Dr. Bhargav Vaidya**.

The matter presented in this project has not been submitted by me for the award of any other degree of this or any other institute.



(Divyansh Gadwal)

This is to certify that the above statement made by the student is correct to the best of my/our knowledge.



(Prof. Dr. Bhargav Vaidya)
Head, Max Planck Partner Group
Associate Professor
Department of Astronomy, Astrophysics, and Space Engineering (DAASE)
Indian Institute of Technology Indore

Divyansh Gadwal has successfully given his MSDSM Oral Examination held on **<Date>.**

Signature(s) of Guide
Date:

Signature of Faculty
Date:

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following individuals and institutions for their invaluable support and contributions to the completion of this research paper:

1. Prof. Dr. Bhargav Vaidya: Head-Max Planck Partner Group, an Associate Professor in the Department of Astronomy, Astrophysics, and Space Engineering (DAASE) at IIT Indore, has been a driving force throughout the project, thanks to his invaluable guidance.

2. Prof. Amit Vatsa and Prof. Parimal Kar: As coordinators of MSDSM programs, they have provided a platform for learning the intricacies of data science and its managerial dimensions. I am thankful for the remarkable opportunity to be part of this program under their guidance.

3. IIM Indore and IIT Indore: As eminent technical and managerial institutes in India, IIM Indore and IIT Indore have generously offered research resources and materials that have been invaluable during this research and throughout the preceding years. My sincere gratitude goes to these institutions.

4. Colleagues and Peers: I extend my utmost thanks to all my fellow MSDSM colleagues whose insights, feedback, and moral support have significantly enriched the research process.

5. Family and Friends: I thank my family members and friends for their steadfast support and encouragement.

This research owes its fruition to the collective efforts of these individuals and institutions. Their substantial contributions have played a pivotal role in shaping the outcomes of this study.

Abstract

Obesity is strongly associated with multiple risk factors. It is significantly contributing to an increased risk of chronic disease morbidity and mortality worldwide. There are various challenges to better understand the association between risk factors and the occurrence of obesity. The traditional regression approach limits analysis to a small number of predictors and imposes assumptions of independence and linearity. Machine Learning (ML) methods are an alternative that provides information with a unique approach to the application stage of data analysis on obesity. This study aims to assess the ability of ML methods, namely Logistic Regression, Random Forest, Support Vector Machines (SVM), Decision Trees, and XGBoost, to identify the risk factors associated with obesity using publicly available health data from National Health and Nutrition Examination Survey (NHANES), using a novel approach with sophisticated ML methods to find the relationship among different factors and to compare the performance of five different methods.

Meanwhile, the main objective of this study is to establish a set of risk factors for obesity in adults among the available study variables. After reviewing the data attributes, this study answers the following research questions:

1. Which variables are risk factors related to obesity?
2. What are the correlations between different risk factors and BMI?
3. Is mental health an important factor that correlates with obesity?
4. Which machine learning model can classify the dataset more accurately?

Age, Race, Country of Birth, Education, Family Income Ratio, Height, Diabetes, Moderate Work Activity, Moderate Recreational Activities, Depressed, Poor Appetite Over Eating, Smoking, Sleep Hours Weekdays, and Sleep Hours Weekend are significant in predicting obesity status in adults. Identifying these risk factors could inform health authorities in designing or modifying existing policies for better controlling chronic diseases, especially in relation to risk factors associated with obesity. Moreover, applying ML methods to publicly available health data, such as National Health and Nutrition Examination Survey (NHANES), is a promising strategy to fill the gap for a more robust understanding of the associations of multiple risk factors in predicting health outcomes.

TABLE OF CONTENTS

1. INTRODUCTION

Obesity constitutes a significant public health concern in the U.S. and globally. About 1 in 5 children and more than 1 in 3 adults struggle with obesity in the U.S. (CDC). Adults with obesity have a higher risk of developing heart disease, type 2 diabetes, and some types of cancer (CDC). According to the World Health Organization (WHO), 30% of global deaths will be caused by lifestyle diseases by 2030. From the research, a limited number of studies use machine learning to analyze obesity/overweight-related datasets in the U.S. Hence, I have chosen to research different risk factors related to obesity/overweight using data from the U.S.

source: CDC

Obesity is a major health problem strongly associated with many chronic illnesses with negative effects and long-term consequences, not only for the patients but also for their families. In Southeast Asia, problems related to nutrition or malnutrition are a double burden because the number of cases of malnutrition and malnourishment is still relatively high, and the number of cases of obesity has also increased significantly over time (1).

Innovations Risk factors for obesity have been studied extensively, and in general, they are divided into several categories: demographic and socio-economic factors (gender, age, education, income, marital status, and urban areas) (2, 4); lifestyle factors (consumption of fast food, stress, smoking, alcoholic drinks, and low level of physical activity) (4, 5); and genetic factors (obese parents) (2, 3). Among these risk factors, some can be changed or modified, while others cannot. Identifying modifiable risk factors for obesity at the individual and the population level is urgently required in order to implement an effective risk reduction strategy.

This project aimed to analyze the relationship between obesity and different risk factors such as BMI, race, gender, physical activities, mental health, education level, etc. This study will walk through the exploratory analysis of the dataset, the pre-modeling, and the development of different machine-learning models to classify the risk factors of obesity.

Based on previous research, ML approaches can increase the risk prediction of health outcomes compared to conventional approaches (7). Prediction of obesity using ML has been investigated by many researchers: Zhang et al. (8), Adnan et al. (9), Toschke et al. (10), Golino et al. (11), Dugan et al. (6), Zheng and Ruggiero (12), Chatterjee et al. (13), Singh and Tawfik (14), and Colmenarejo (15). The ML approach provides an alternative in providing information with a unique approach at the application stage of data analysis on obesity which is important in providing a better predictive solution to the likelihood of obesity (16).

The code for this project can be found in my **Github** repository.

2. Literature/ Industry research review

In a study conducted by the Technology and Health Departments of the University of Agder in Norway, researchers utilized advanced machine learning techniques to identify potential risk factors associated with obesity and overweight. Led by Chatterjee et al. in 2021, the study employed Support Vector Machines (SVM), Decision Trees, and Logistic Regression models to analyze complex datasets. These methods enabled the researchers to uncover significant factors contributing to obesity, shedding light on the multifaceted nature of this public health issue. The application of machine learning in this context not only facilitates the identification of risk factors but also provides insights crucial for developing targeted interventions and preventive strategies. By harnessing these advanced analytical tools, the study contributes to our understanding of obesity's complexities and offers valuable insights for guiding public health initiatives aimed at its prevention and management.

Study conducted by Daffodil International University in Dhaka, Bangladesh, represents a significant endeavor in leveraging machine learning (ML) algorithms to predict the risk of obesity. Ferdowsy and colleagues applied nine prominent ML algorithms to analyze data collected from diverse populations, encompassing individuals of various ages afflicted with both obesity and non-obesity. By employing a comprehensive approach, the research aimed to uncover patterns and factors contributing to obesity across a spectrum of demographics, providing valuable insights into the complex interplay of variables influencing weight status. Through the utilization of sophisticated ML techniques, including but not limited to decision trees, support vector machines, and logistic regression models, the study sought to not only identify risk factors associated with obesity but also to enhance predictive accuracy. By integrating data from a diverse range of individuals, the research contributes to a more nuanced understanding of obesity risk, potentially informing targeted interventions and public health initiatives tailored to specific demographic groups.

Another study was conducted by Delnevo et al. (2021) at the University of Bologna in Italy, in which machine learning techniques were employed to examine the predictive effects of emotional and affective variables on Body Mass Index (BMI) values. This research aimed to understand how emotional and affective factors contribute to variations in BMI, providing insights into potential predictors of obesity and overweight, which could inform interventions and preventive measures.

3. DATA

3.1 Data Source

The data used was aggregated from seven different datasets from the National Health and Nutrition Examination Survey dating March 2017 to 2020 Pre-pandemic. These datasets contain demographic, examination, laboratory, and questionnaire data.

In the final dataset, it consists of the following columns; Respondent Sequence Number, Gender, Race, Country of birth, Education Level, Ratio of family income to poverty, Body measures (Weight, Height, BMI), Diabetes status, Physical activity (Moderate work activity, recreational activity), Mental health (Depressed, Poor appetite or overeating), Smoking, and Sleep disorders (Sleep hours on weekdays and weekends). The size of the dataset is 12.4MB — XPT. files.

- https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2017-2020
- https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&Cycle=2017-2020
- https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2020

3.2 Data Attributes

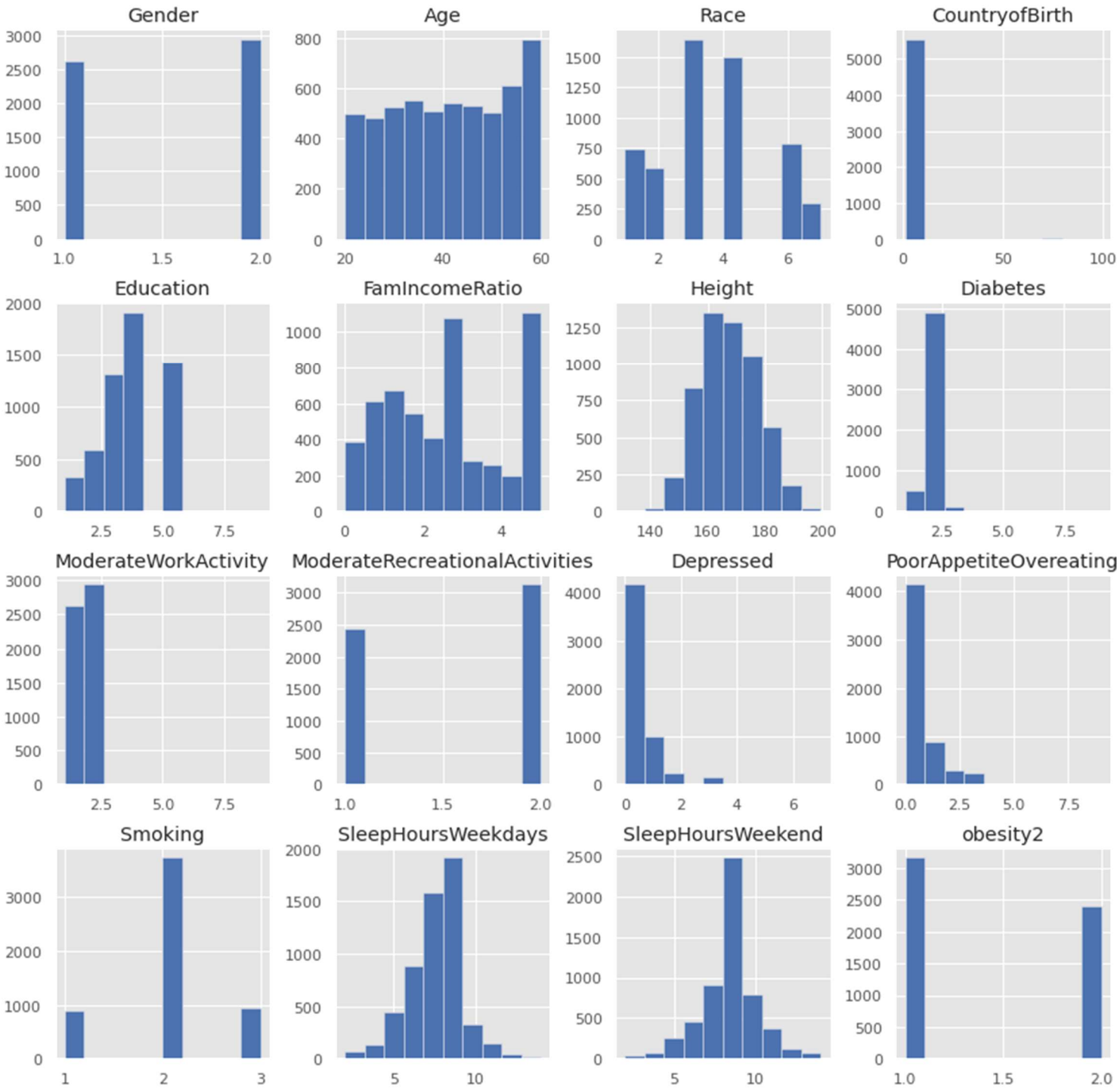
- **SEQN:** Respondent sequence number
- **Gender:** 1 = Male, 2 = Female
- **Age:** Age in years
- **Race:** 1= Mexican American, 2 = Other Hispanic, 3 = None-Hispanic White, 4 = None-Hispanic Black, 6 = None-Hispanic Asian, 7 = Other Race-Including Multi-Racial
- **Country of birth:** 1 = Born in 50 US states or Washington, DC, 2 = Other
- **Education level- Adults 20+:** 1 = Less than 9th grade, 2 = 9–11th grade (Includes 12th grade with no diploma), 3 = High School graduate/GED or equivalent, 4 = Some college or AA degree, 5 = College graduate or above,
- **Ratio of family income to poverty:** numerical values from 0 to 5.00
- Weight in kg
- Height in cm
- Body mass index — BMI
- **Doctor told you have diabetes:** 1 = Yes, 2 = No, 3 = Borderline
- **Moderate work activity:** 1 = Yes, 2 = No
- **Moderate recreational activities:** 1 = Yes, 2 = No
- **Feeling down, depressed, or hopeless:** 0 = Not at all, 1 = Several days, 2 = More than half the days, 3 = Nearly every day
- **Poor appetite or overeating:** 0 = Not at all, 1 = Several days, 2 = More than half the days, 3 = Nearly every day
- **Sleep hours — weekdays or workdays:** range of values
- **Sleep hours — weekends:** range of values
- **Do you now smoke cigarettes?** 1= Every day, 2 = Some days, 3 = Not at all

3.3 Data Exploratory Analysis

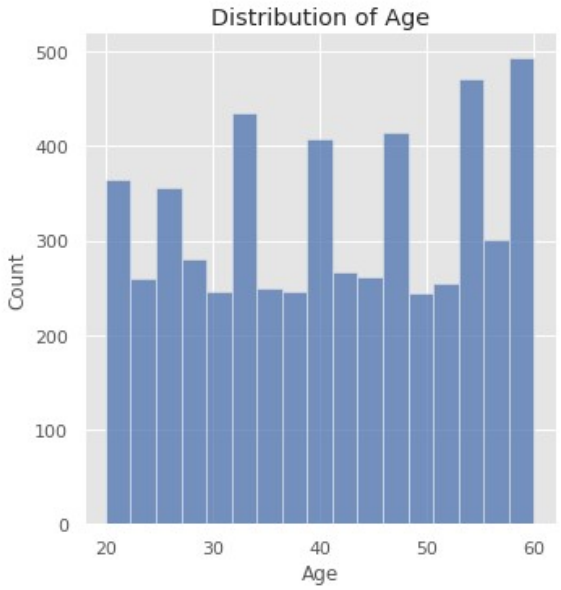
- After importing the datasets, I merged the seven datasets to a new dataset, renamed the columns, and checked the shape of the data frame which is 15,560 rows and 18 columns.
- Then I dropped the respondent data with missing BMI values, which changed the size of the dataframe to 13137 rows and 18 columns.
- The dataset contains a lot of missing values in the education, activity, depression, poor appetite/overeating, smoking, and sleep hours for weekdays and weekend columns.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13137 entries, 1 to 15559
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   SEQN                            13137 non-null  int64
 1   Gender                          13137 non-null  int64
 2   Age                             13137 non-null  int64
 3   Race                            13137 non-null  int64
 4   CountryofBirth                  13137 non-null  int64
 5   Education                       8381 non-null   float64
 6   FamIncomeRatio                  11443 non-null  float64
 7   Weight                          13137 non-null  float64
 8   Height                          13137 non-null  float64
 9   BMI                             13137 non-null  float64
 10  Diabetes                        13137 non-null  int64
 11  ModerateWorkActivity            8790 non-null   float64
 12  ModerateRecreationalActivities  8790 non-null   float64
 13  Depressed                       8203 non-null   float64
 14  PoorAppetiteOvereating          8202 non-null   float64
 15  Smoking                         3521 non-null   float64
 16  SleepHoursWeekdays              9188 non-null   float64
 17  SleepHoursWeekend               9183 non-null   float64
dtypes: float64(12), int64(6)
memory usage: 1.9 MB
```

- I decided to explore more about the data distribution before acting on the missing data.



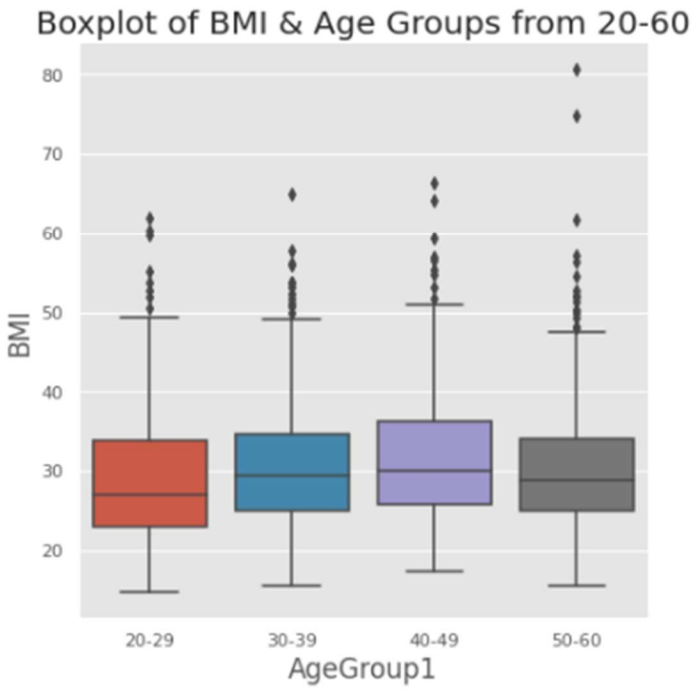
- From the histogram matrix, we can get an idea about how the values are distributed for each feature/factor in the dataset. Here, I specifically focused on the age distribution.

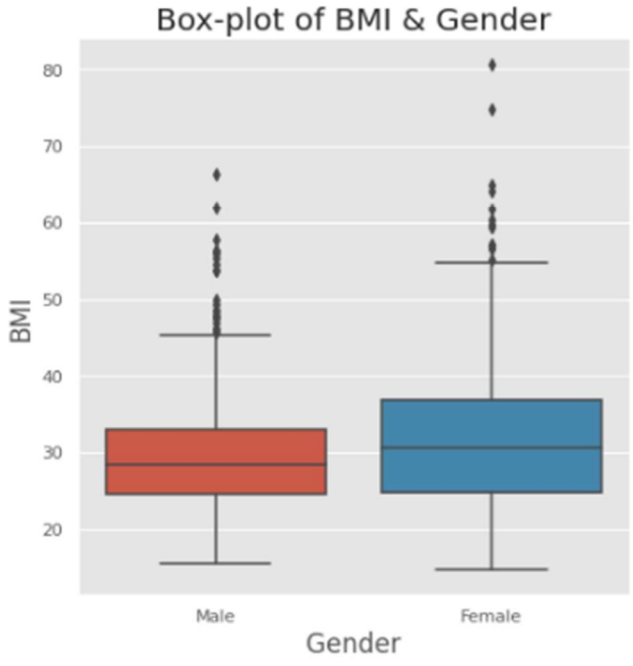


- Considering respondents below 20 or older than 60 years old might not have information such as education, family income ratio, diabetes, activities, eating disorder status, and smoking habit, I decided to extract the respondents' data between 20 to 60 years old to a new data frame.

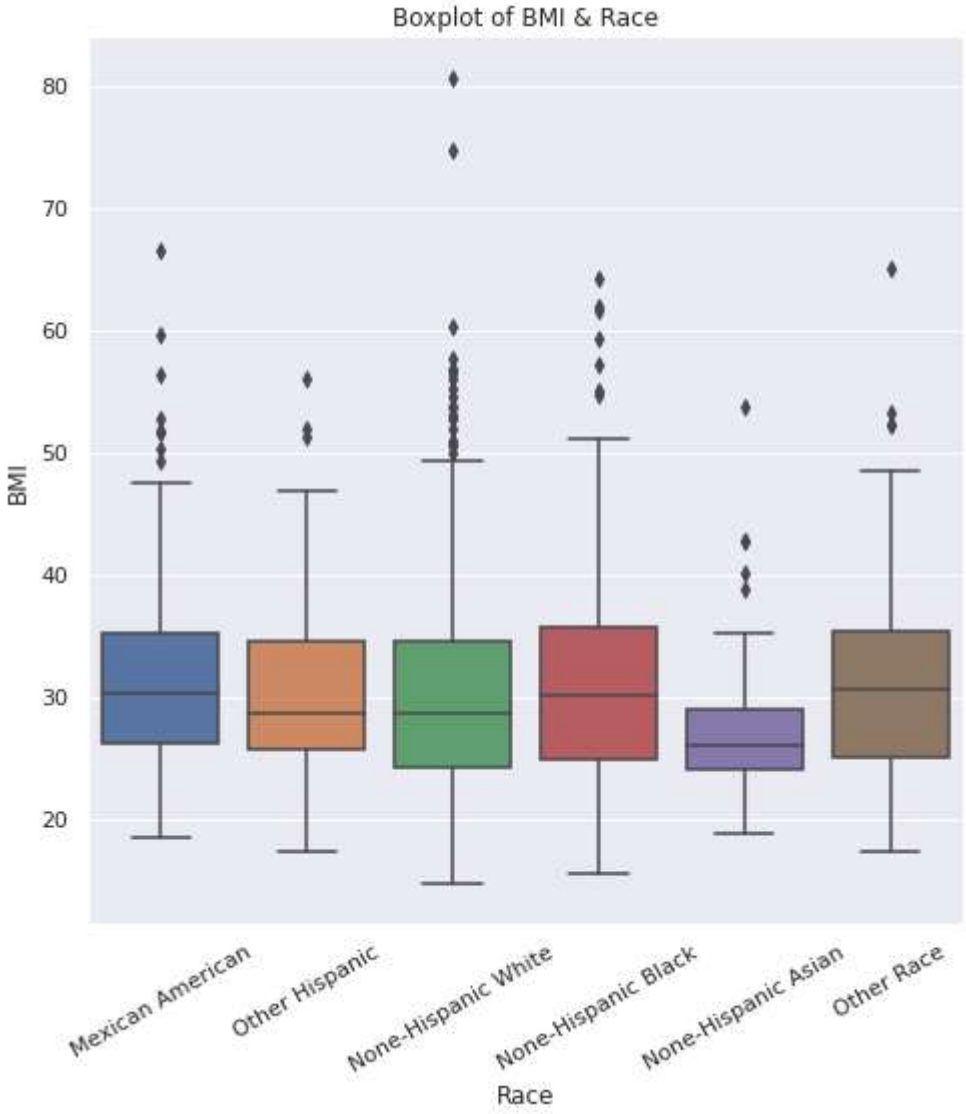
```
# Filter the age, extract respondents from 20-60 years old.
df_age_filter = df[(df['Age'] >= 20) & (df['Age'] <= 60)]
```

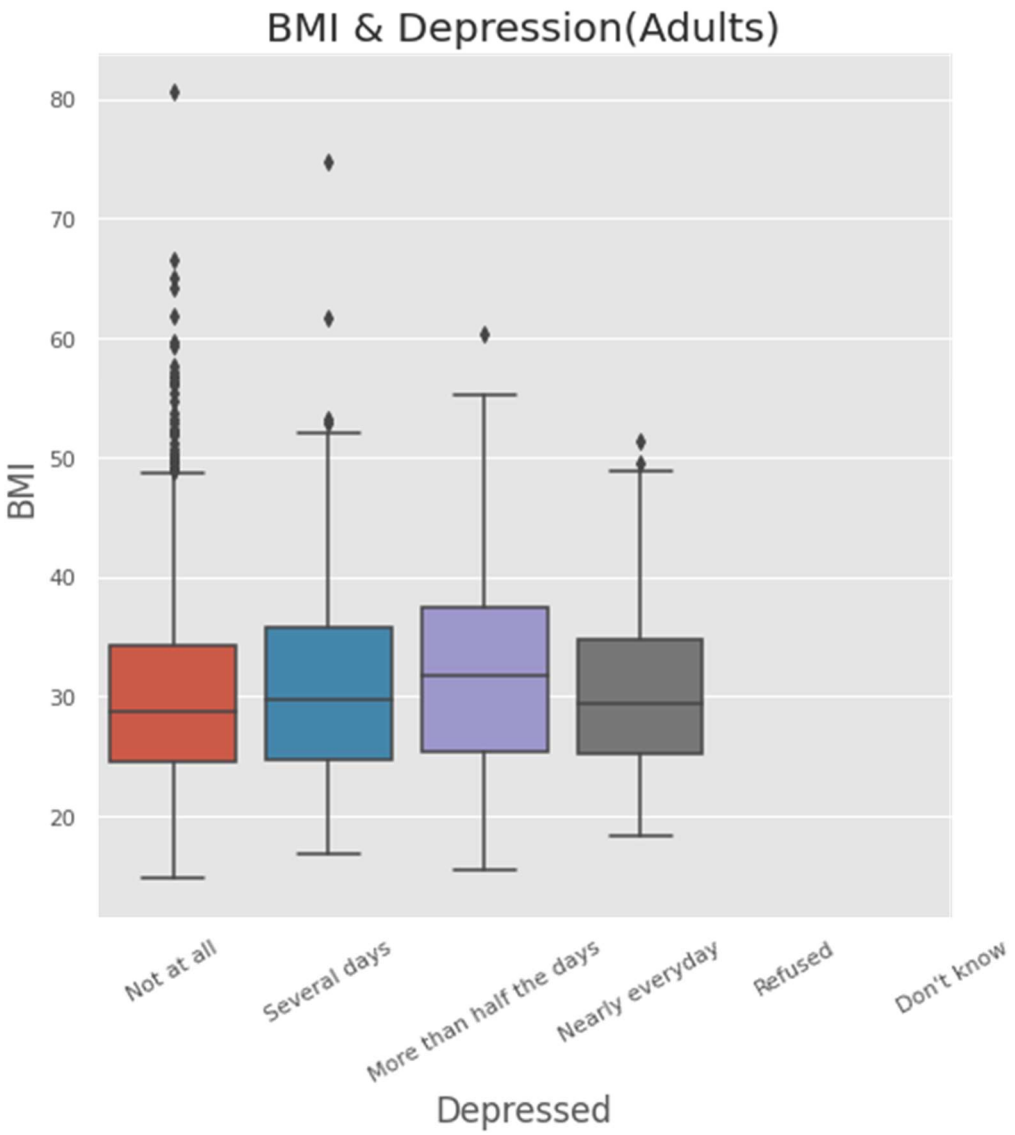
- After extracting respondents from 20 to 60 years old, I tried to visualize the relationship between different features/factors and BMI. Boxplot of BMI by age groups shows that respondents in the 40–49 age group have a higher BMI compared to other age groups. Females have a higher BMI (median) compared to males.





- Boxplot of BMI and race shows that Non-Hispanic Asians have a significantly low BMI (median) compared to other race groups. Respondents with depressed feelings more than half the day have a higher BMI (median).





- After visualizing the data, I created a new column to define respondents' weight level based on BMI guidelines from CDC (CDC, 2021).

```
obese_condition = [(df_new['BMI'] < 18.5),
                  (df_new['BMI'] >= 18.5) & (df_new['BMI'] < 25.0),
                  (df_new['BMI'] >= 25.0) & (df_new['BMI'] < 30.0),
                  (df_new['BMI'] > 30.0)]

# 1 - Under Weight
# 2 - Healthy
# 3 - Overweight
# 4 - Obese
obese_value = [1, 2, 3, 4]

df_new['obesity'] = np.select(obese_condition, obese_value, default = 1)
```

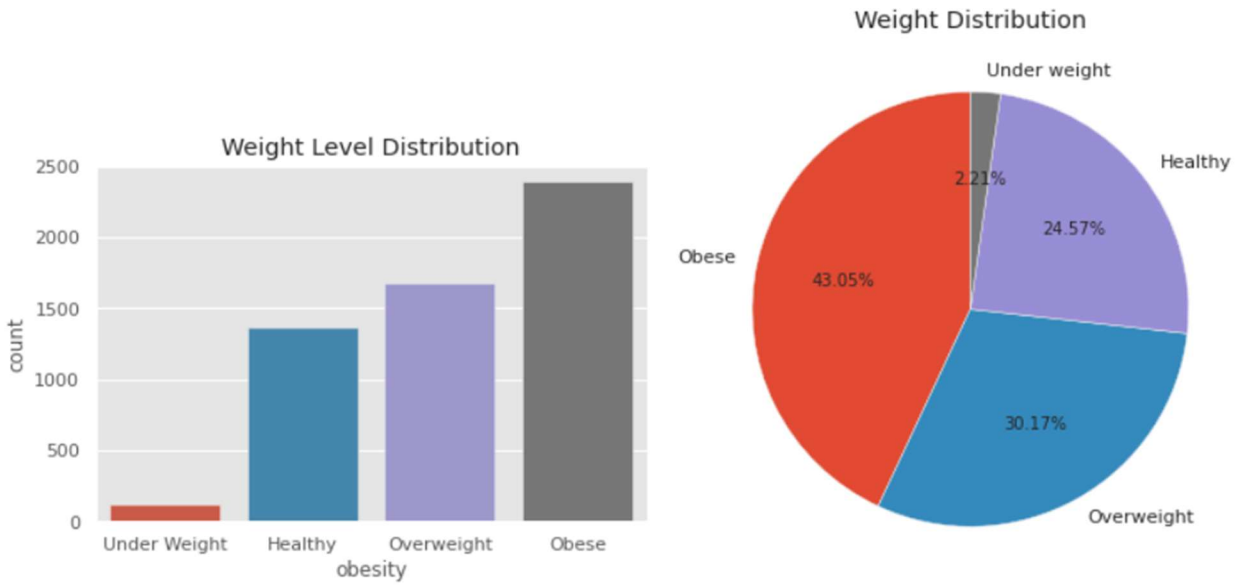
- After the data extraction, the dataset contains less missing values compared to the original dataset.

```
SEQN                                0
Gender                              0
Age                                 0
Race                                0
CountryofBirth                      0
Education                           0
FamIncomeRatio                    755
Weight                              0
Height                              0
BMI                                 0
Diabetes                            0
ModerateWorkActivity                0
ModerateRecreationalActivities      0
Depressed                         346
PoorAppetiteOvereating            346
Smoking                          3470
SleepHoursWeekdays                 41
SleepHoursWeekend                  48
dtype: int64
```

- Next, I filled the NaN values with mean and median values:

```
df_age_filter['FamIncomeRatio'].fillna(df_age_filter['FamIncomeRatio'].mean(), inplace=True)
df_age_filter['Depressed'].fillna(df_age_filter['Depressed'].median(), inplace=True)
df_age_filter['PoorAppetiteOvereating'].fillna(df_age_filter['PoorAppetiteOvereating'].median(), inplace=True)
df_age_filter['Smoking'].fillna(df_age_filter['Smoking'].median(), inplace=True)
df_age_filter['SleepHoursWeekdays'].fillna(df_age_filter['SleepHoursWeekdays'].mean(), inplace=True)
df_age_filter['SleepHoursWeekend'].fillna(df_age_filter['SleepHoursWeekend'].mean(), inplace=True)
```

- Checking the distribution of weight levels:



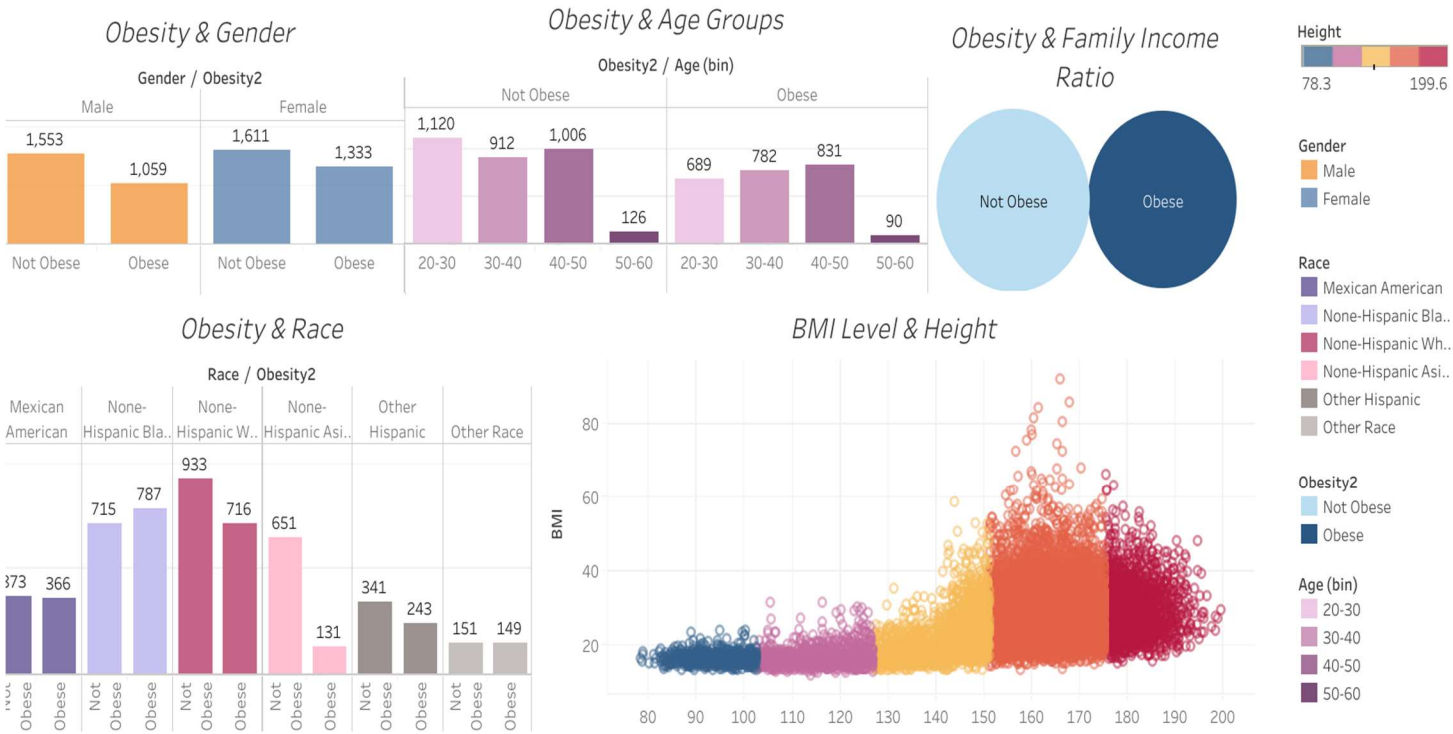
3.4 Class Imbalance

- From the plot we can see that data is not balanced, there are 43.05% respondents in the Obese level, which is about half of the dataset and indicates that there is a class imbalance problem in the dataset.
- To balance the dataset, I chose to combine the respondents from the underweight, healthy, and overweight groups together and keep the obese group separate. (1 = Not Obese, 2 = Obese)

- After rearranging the weight groups, the dataset looks more balanced.

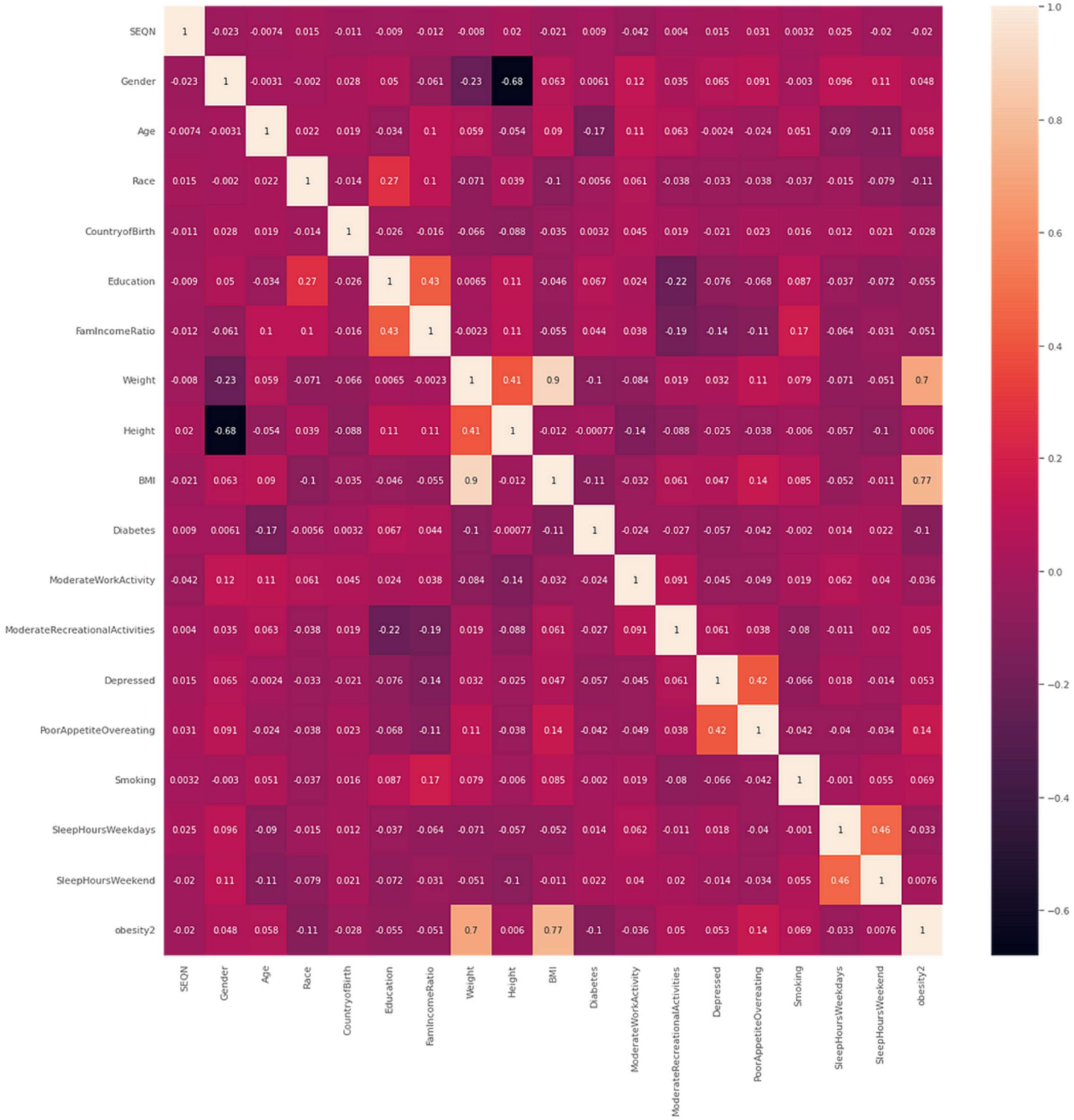
4. Tableau Dashboard

I created a ***tableau dashboard*** to visualize the obesity level versus gender, age groups, race, and average family income ratio to poverty, as well as BMI level versus height. This dashboard provides us a better visualization of the data values, and enables users to interact with different variables.



5. Correlation

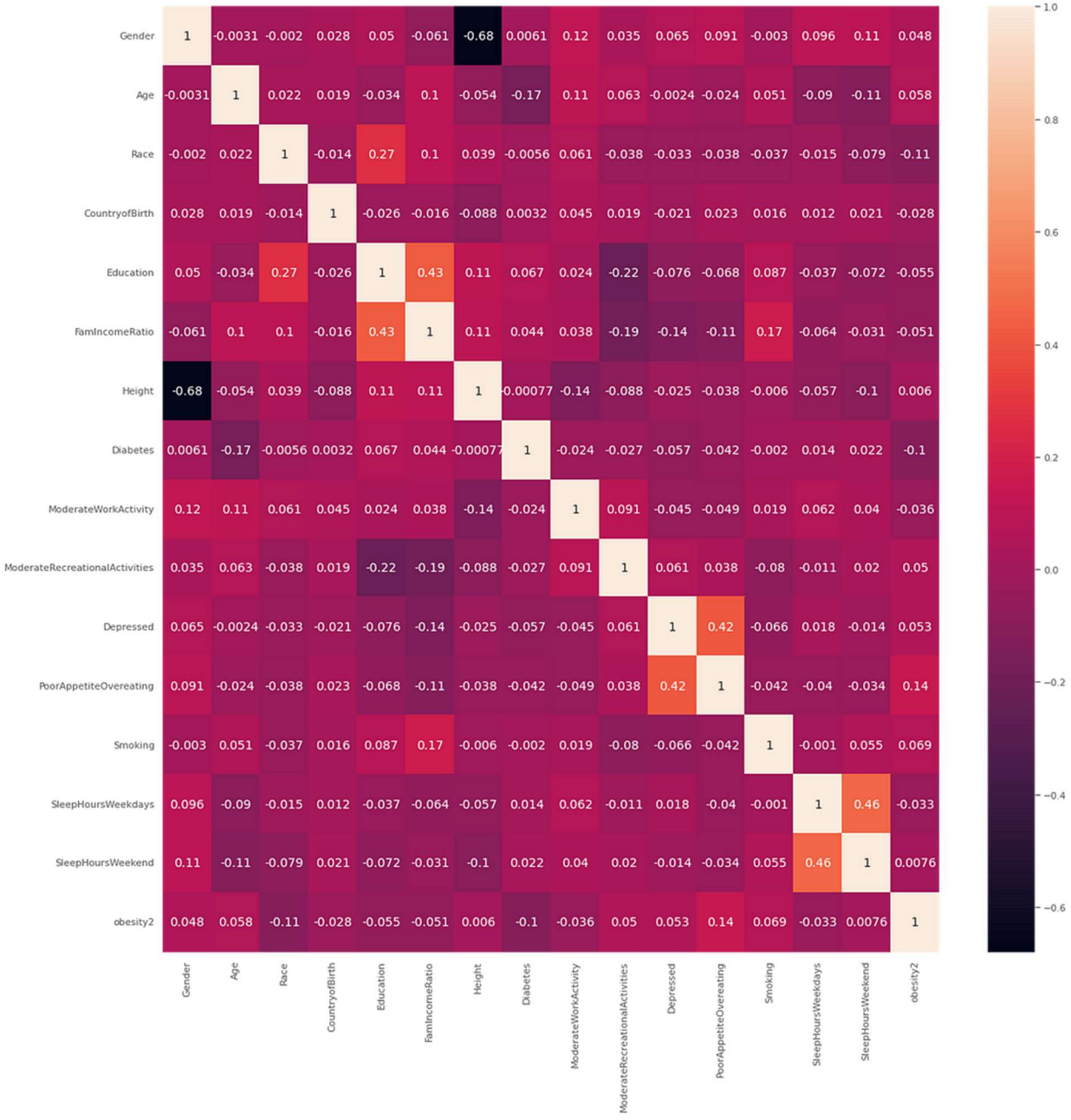
The final step of EDA is to check the correlation of different features/factors with the obesity level.



From the heatmap, we can see that obesity is highly correlated to Weight and BMI. By including these two features, the model accuracy could reach over 95% percent.

However, in order to figure out which other features, I decided to drop both of them to get a better idea of how other factors affect obesity level.

```
df_age_filter.drop(['BMI'], axis=1, inplace=True)
df_age_filter.drop(['Weight'], axis=1, inplace=True)
df_age_filter.drop(['SEQN'], axis=1, inplace=True) #Drop the ID column
```



After removing BMI and Weight, we can see that obesity level is positively correlated to PoorAppetite/Overeating. And it is negatively correlated to Race, Diabetes. There are no features with a correlation of more than 0.2 with obesity level. There are a few features which have high correlation: Poor appetite/overeating and Depressed; Education and Family income ratio; Height and Gender.

6. Pre-modeling

6.1 Normalizing Data

Since the data is measured in different scales, then it should be normalized before splitting. We choose to normalize the data using MinMaxScaler in order to construct and run the models.

```
#normalizing the data
scaler = MinMaxScaler(feature_range = (0,1))
normalized_data = scaler.fit_transform(df_age_filter)
columns = ['Gender',  'Age',  'Race', 'CountryofBirth', 'Education',  'FamIncomeRatio', 'Height', 'Diabetes',
normalized_df = pd.DataFrame(normalized_data, columns=columns)
normalized_df['obesity2'] = normalized_df['obesity2'].astype(int)
normalized_df.head()
```

	Gender	Age	Race	CountryofBirth	Education	FamIncomeRatio	Height	Diabetes	ModerateWorkActivity
0	1.0	0.225	0.833333	0.010204	0.500	1.000000	0.424818	0.125	0.125
1	0.0	0.725	0.333333	0.000000	0.125	0.516536	0.747445	0.125	0.000
2	0.0	0.400	0.333333	0.000000	0.375	0.166000	0.775182	0.125	0.125
3	1.0	0.600	0.000000	0.010204	0.125	0.516536	0.315328	0.125	0.125
4	1.0	0.325	0.833333	0.010204	0.500	0.272000	0.398540	0.125	0.125

6.2 Splitting dataset to train and test set

After normalizing the dataset, I split the data into training and testing sets at an 80% to 20% ratio.

```
# splitting test and train data
X = normalized_df.iloc[:, :-1]
Y = normalized_df.iloc[:, -1]
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 2)
```

7. Modeling

I chose the following six models because they work well with binary classification problems, such as the obesity one this study is about.

7.1 Baseline Model

DummyClassifier is a classifier that makes predictions using simple rules. We use this to build a baseline model to compare with other models.

```
# Baseline classification accuracy
from sklearn.dummy import DummyClassifier

baseline_classifier = DummyClassifier(strategy = "most_frequent")

#fitting the dummy classifier to the training data
baseline_classifier.fit(X_train,Y_train)

# predicting the target values for the test data
Y_pred_base = baseline_classifier.predict(X_test)

# accuracy calculation
from sklearn import metrics

#predicting target values (Y_pred_base) with the actual target values (Y_test)
print("Accuracy:", metrics.accuracy_score(Y_test, Y_pred_base))
```

```
Accuracy: 0.5764388489208633
```

The baseline model has a 57.64% accuracy, which indicates the lowest possible prediction we can get. It is expected to get higher accuracy from other models that are selected.

7.2 Support Vector Machine (SVM) Model

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

Accuracy for SVM model = 0.6348920863309353

Cross validation score = 0.6090615075507162

	precision	recall	f1-score	support
0	0.63	0.87	0.73	641
1	0.64	0.32	0.43	471
accuracy			0.63	1112
macro avg	0.64	0.59	0.58	1112
weighted avg	0.64	0.63	0.60	1112

The accuracy score is 63.49%, which is higher than the baseline model. The cross-validation score has a 0.03 difference from the accuracy score, which indicates model overfitting.

7.3 Logistic Regression Model

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. It is mainly used when the target variable is categorical.

Accuracy of Logistic Regression Model = 0.6133093525179856
Cross validation score = 0.6061812171884113

	precision	recall	f1-score	support
0	0.63	0.81	0.71	641
1	0.57	0.35	0.43	471
accuracy			0.61	1112
macro avg	0.60	0.58	0.57	1112
weighted avg	0.60	0.61	0.59	1112

The accuracy score of Logistic Regression is 61.33%, it's lower than the accuracy of the SVM model.

7.4 Decision Tree Model

The goal of using a Decision Tree model is to create a training model that can be used to predict/classify the value of the target variable by learning simple decision rules inferred from training data.

Accuracy of Decision Tree Model = 0.5755395683453237
Cross Validation score = 0.5727069155486422

	precision	recall	f1-score	support
0	0.63	0.62	0.63	641
1	0.50	0.51	0.50	471
accuracy			0.58	1112
macro avg	0.57	0.57	0.57	1112
weighted avg	0.58	0.58	0.58	1112

The accuracy rate of the Decision Tree model is only 57.55%, with a 62% false positive rate, we can tell it is not a good model for our dataset.

7.5 Random Forest Model

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. Random Forest model does not suffer overfitting, and it can cancel biases from taking average predictions.

Accuracy of Random Forest Model = 0.6339928057553957

Cross Validation score = 0.6362376045109859

	precision	recall	f1-score	support
0	0.66	0.77	0.71	641
1	0.59	0.45	0.51	471
accuracy			0.63	1112
macro avg	0.62	0.61	0.61	1112
weighted avg	0.63	0.63	0.62	1112

The accuracy score of the Random Forest model is 63.40%, with a close cross-validation score. We can say that the Random Forest model has a better performance compared to SVM, Logistic Regression, and Decision Tree models.

7.6 XGBoost Model

XGBoost is a decision-tree-based ensemble ML algorithm. It uses a gradient boost framework, and delivers more accurate approximations by using the second order derivative of the loss function.

Accuracy of XGBoost Model = 0.6510791366906474

Cross Validation score = 0.6412722794737182

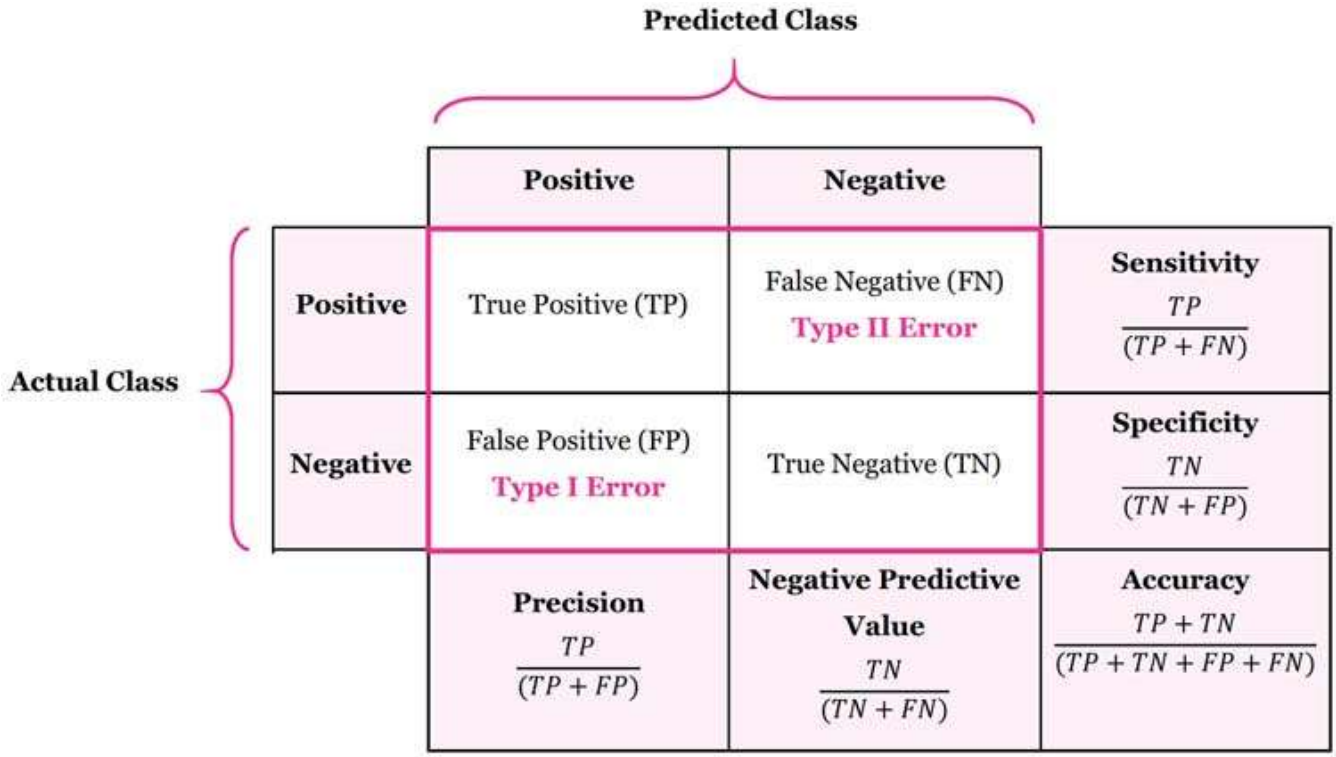
	precision	recall	f1-score	support
0	0.66	0.82	0.73	641
1	0.63	0.42	0.51	471
accuracy			0.65	1112
macro avg	0.65	0.62	0.62	1112
weighted avg	0.65	0.65	0.64	1112

The accuracy rate of the XGBoost model is 65.10%, and the cross-validation score is 64.12%, which indicates that there is almost no overfitting in this case.

The Baseline model shows 57.64% accuracy, the random forest model without tuning shows 63.40% accuracy. The XGBoost model has 65.10% accuracy which is the best model compared to all other models we selected.

8. Classifier Performance Evaluation

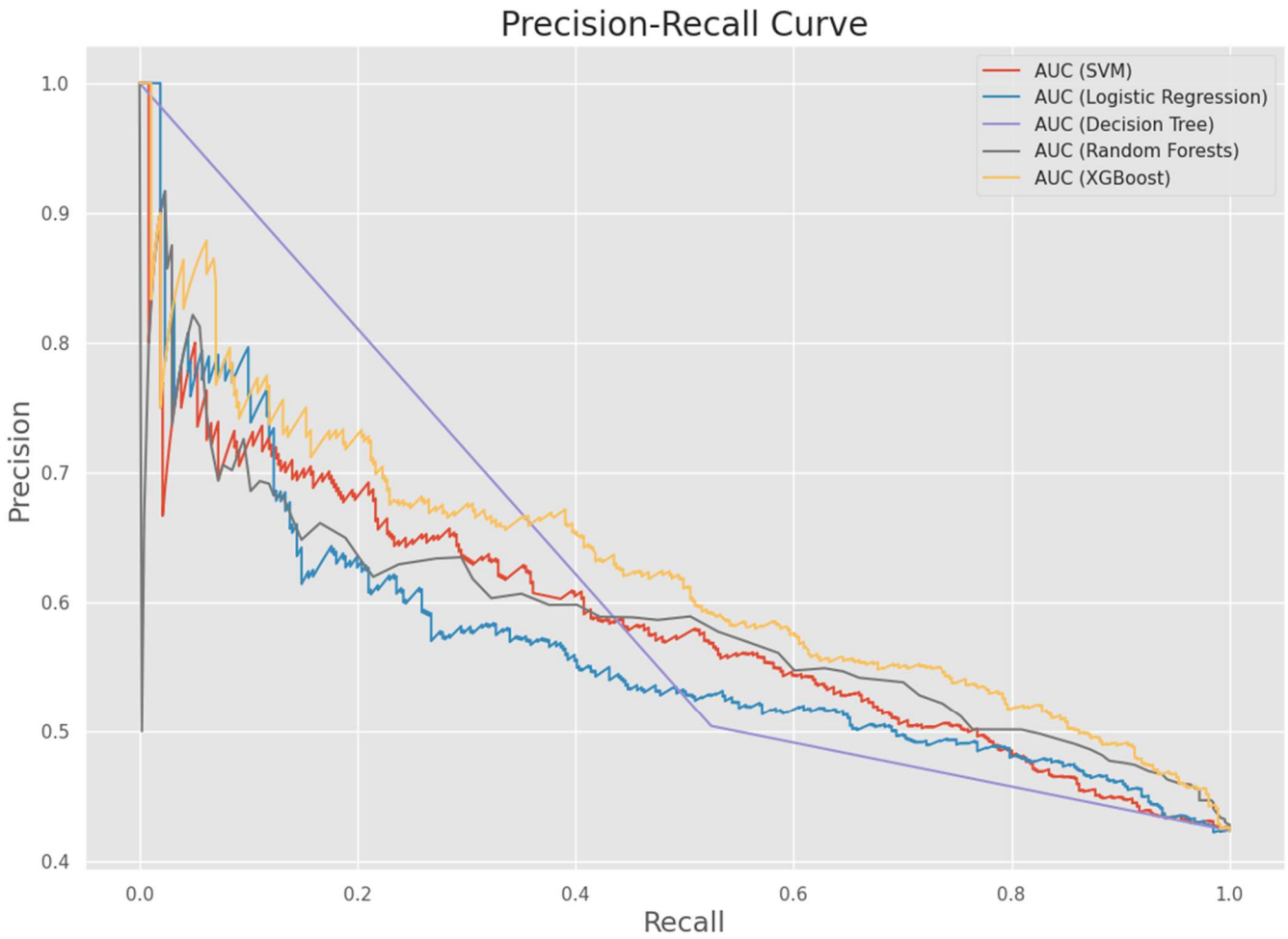
Now, let's use a classifier report to check the performance of each model.



Precision shows how much was correctly classified as positive out of all the positives. Recall of a classifier is the ratio between how much was correctly identified as positive to all the actual positives. Moreover, F1-score means the weighted average between precision and recall. Based on this research, F1-score is beneficial for imbalanced datasets. Since the dataset is balanced after combining the none-obese groups, we can use the accuracy score to choose the best model. The model that provides the best accuracy is the XGBoost model.

9. Precision-Recall Curve

Another method to evaluate the performance of classification algorithms is the Precision-Recall curve. Precision is a ratio of the number of true positives divided by the sum of the true positives and the false positives. It describes how good a model is at predicting the positive class. Precision is referred to as the positive predictive value.




```
AUC of Logistic Regression: 0.56
AUC of SVM: 0.58
AUC of Random Forest: 0.58
AUC of Decision Tree: 0.61
AUC of XGBoost: 0.62
```

From the Precision-recall curves, the XGBoost model has the highest AUC value, 0.62. Thus, the XGBoost model is the best model for this dataset.

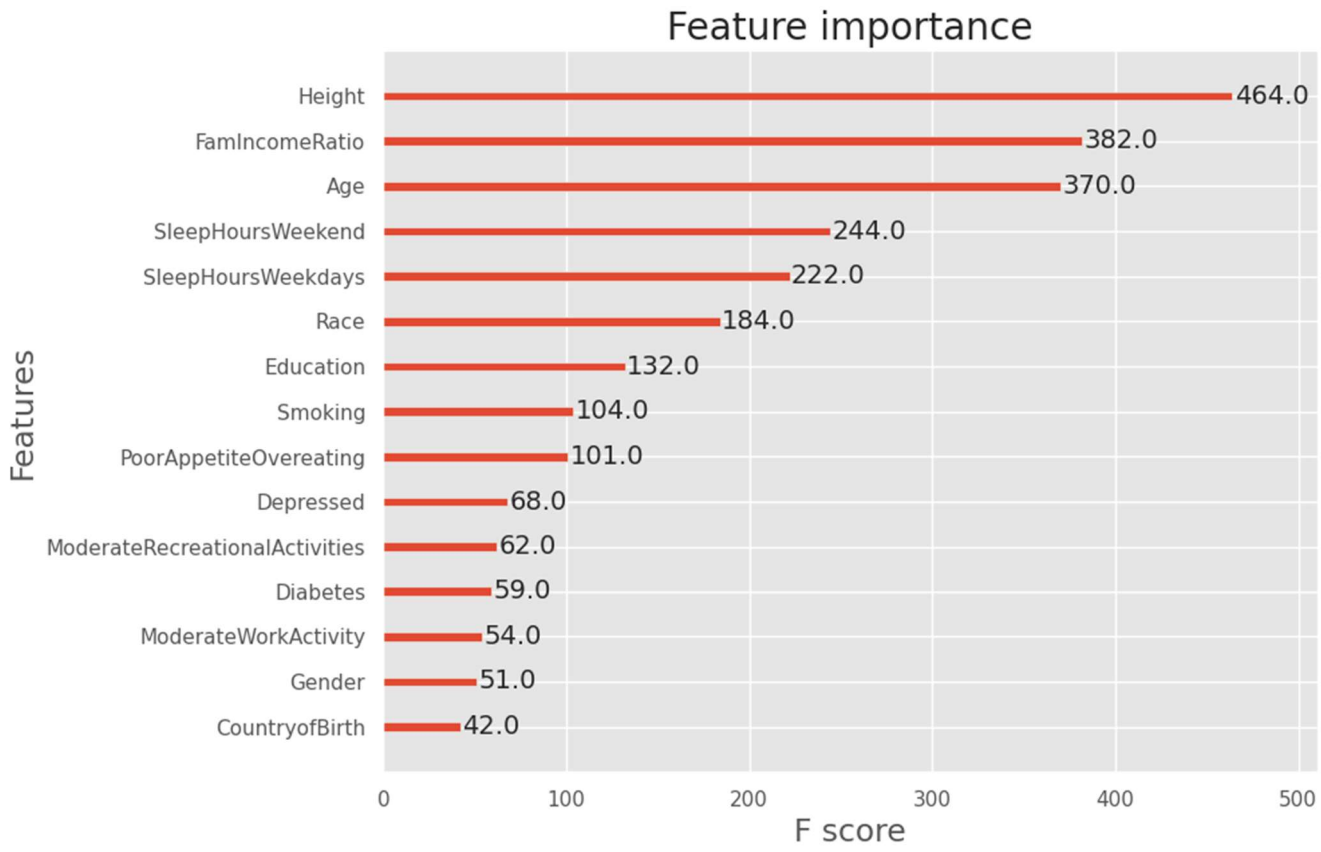
10. Parameter Tuning XGBoost Model

I used GridSearchCV to find the best estimator, and improved the accuracy to 65.55%, the cross-validation score is 63.60%.

```
final_model = XGBClassifier(learning_rate=0.05, n_estimators=50)
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())
```

```
Accuracy: 0.6555755395683454
Cross validation score: 0.6360545077451553
```

- **Feature Importance After parameter tuning**



11. CONCLUDING REMARKS

- Although the accuracy levels from the models were considerably low, they were higher than the baseline model.
- The XGBoost Model provided the best accuracy score compared to the other models so I decided to check the feature importance.
- The feature importance results from the XGBoost model indicated that the top five (5) risk factors besides BMI and Weight are:

    1. Height
    2. Family income ratio
    3. Age
    4. Sleep hours on weekdays
    5. Sleep hours on weekend.

  In contrast, factors such as Activity, Gender, and Country of Birth are the least relevant to obesity based on the dataset.
- In the literature review, I learned that depression can affect obesity levels. However, based on this dataset and the results, we cannot say that mental health is highly affecting obesity level.
- Identify potential limitations: Knowing a client's age and sleep habits can help Gym trainers tailor programs to their energy levels and injury risks. For example, shorter sessions or low-impact exercises might be suitable for individuals with less sleep or older age groups.

12. Limitations and Future Study

Limitations

- Limited access to robust open-source healthcare datasets due to US laws such as HIPAA (protects sensitive patient health information).
- This dataset does not include relevant features such as eating habits, family history of obesity, or family history of other diseases which might have varied the results.

Future Study

- The scalability and distributed processing capabilities of Hadoop can be employed to manage large volumes of datasets. Leveraging Hadoop's parallel processing power will enhance the efficiency of the machine learning model.
- Implement Neural Network Backpropagation to enable the model to self-learn and improve the accuracy while feeding in new data.
- Find a better dataset to do in depth research and build prediction models for other relevant diseases.
- Build a web interface/tool for disease prediction such as Diabetes.

13. REFERENCES

[1] ASEAN/UNICEF/WHO Regional Report. World Health Statistics 2016: Monitoring Health for the SDGs, Sustainable Development Goals. (2016). Available online at: https://www.who.int/about/licensing/copyright_form/en/index.html

[2] Roemling C, Qaim M. Obesity trends and determinants in Indonesia. Appetite. (2012) 58:1005–13. 10.1016/j.appet.2012.02.053 [PubMed] [CrossRef] [Google Scholar]

[3] Rachmi CN, Li M, Alison Baur L. Overweight and obesity in Indonesia: prevalence and risk factors-a literature review. Public Health. (2017) 147:20–9. 10.1016/j.puhe.2017.02.002 [PubMed] [CrossRef] [Google Scholar]

[4] Oddo VM, Maehara M, Rah JH. Overweight in Indonesia: an observational study of trends and risk factors among adults and children. BMJ Open. (2019) 9:e031198. 10.1136/bmjopen-2019-031198 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[5] Dewi NU, Tanziha I, Solechah SA, Bohari B. Obesity determinants and the policy implications for the prevention and management of obesity in Indonesia. Curr Res Nutr Food Sci J. (2020) 8:942–55. 10.12944/CRNFS.8.3.22 [CrossRef] [Google Scholar]

[6] Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. Appl Clin Inform. (2015) 6:506–20. 10.4338/ACI-2015-03-RA-0036 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[7] Selya AS, Anshutz D. Machine learning for the classification of obesity from dietary and physical activity patterns. In: Giabbanelli P, Mago V, Papageorgiou E, editors. Advanced Data Analytics in Health. Springer; (2018). p. 77–97. Available online at: http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-319-77911-9_5 [Google Scholar]

[8] Zhang S, Tjortjis C, Ze ng X, Qiao H, Buchan I, Keane J. Comparing data mining methods with logistic regression in childhood obesity prediction. Inform Syst Front. (2009) 11:449–60. 10.1007/s10796-009-9157-0 [CrossRef] [Google Scholar]

[9] Adnan MHBM, Husain W, Rashid NA. Parameter identification and selection for childhood obesity prediction using data mining. In: 2nd International Conference on Management and Artificial Intelligence. Singapore: (2012). p. 7. [Google Scholar]

[10] Toschke AM, Beyerlein A, Von Kries R. Children at high risk for overweight: a classification and regression trees analysis approach. Obes Res. (2005) 13:1270–4. 10.1038/oby.2005.151 [PubMed] [CrossRef] [Google Scholar]

[11] Golino HF, Amaral LSB, Duarte SFP, Gomes CMA, Soares J, Reis LA, et al.. Predicting increased blood pressure using machine learning. J Obes. (2014) 2014:637635. 10.1155/2014/637635 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[12] Zheng Z, Ruggiero K. Using machine learning to predict obesity in high school students. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas: (2017). p. 2132–2138. 10.1109/BIBM.2017.8217988 [CrossRef] [Google Scholar]

[13] Chatterjee A, Gerdes MW, Martinez SG. Identification of risk factors associated with obesity and overweight–a machine learning overview. Sensors. (2020) 20:2734. 10.3390/s20092734 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[14] Singh B, Tawfik H. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. Comput Sci ICCS 2020. (2020). 12140:523–35. 10.1007/978-3-030-50423-6_39 [CrossRef] [Google Scholar]

[15] Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. Nutrients. (2020) 12:2466. 10.3390/nu12082466 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[16] DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al.. A review of machine learning in obesity. Obes Rev. (2018) 19:668–85. 10.1111/obr.12667 [PMC free article] [PubMed] [CrossRef] [Google Scholar].

[17] Chatterjee A, Gerdes MW, Martinez SG. Identification of Risk Factors Associated with Obesity and Overweight — A Machine Learning Overview. Sensors. 2020; 20(9):2734. https://doi.org/10.3390/s20092734

[18] Wilfley, D. E., Hayes, J. F., Balantekin, K. N., Van Buren, D. J., & Epstein, L. H. (2018). Behavioral interventions for obesity in children and adults: Evidence base, novel approaches, and translation into practice. American Psychologist, 73(8), 981–993. https://doi-org.proxy-bc.researchport.umd.edu/10.1037/amp0000293

[19] CDC. (2021, March 1). Why It Matters. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/about-obesity/why-it-matters.html Centers for Disease Control and Prevention. (2021, August 27). About adult BMI. Centers for Disease Control and Prevention.

[20] Retrieved September 29, 2021, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html.

[21] Ferdowsy, F.,Rahi, K. S. A., Jabiullah, Md. I., Habib, Md. T. (2021, August 5). A Machine Learning approach for obesity risk prediction. Current Research in Behavioral Science, 2, 2021. https://doi.org/10.1016/j.crbeha.2021.100053

[22] Delnevo, G., Mancini, G., Roccetti, M., Salomoni, P., Trombini, E., & Andrei, F. (2021). The Prediction of Body Mass Index from Negative Affectivity through Machine Learning: A Confirmatory Study. Sensors, 21(7). https://doi.org/10.3390/s21072361

14. ACKNOWLEDGEMENT

This expression of appreciation signifies the deep gratitude I hold towards those who have assisted me in creating this report, enriching my experience along the way. I am thrilled to extend my heartfelt thanks to both the institutions involved and my esteemed project mentor, Dr. Bhargav Vaidya, for facilitating the formalization of my MSDSM project work, which has significantly expanded my knowledge. I wish to convey my profound appreciation to Dr. Bhargav Vaidya for his invaluable support, encouragement, supervision, and insightful suggestions throughout the duration of this project. Their unwavering support and continuous guidance were instrumental in the successful completion of my work. I am immensely thankful for the invaluable cooperation and unwavering encouragement from the Head of the Max Planck Partner Group, Dr. Bhargav Vaidya. His consistent advice greatly facilitated my progress and proficiency in the project. Lastly, I am grateful and indebted to all those who directly or indirectly contributed to the completion of this project report.