Predicting Obesity Status Based on Genetic and Lifestyle Factors

Muryam Hasan, Michael Jin, Serena Lu, Sydney Tomsick, Carrie Su

Table of Contents

Abstract

The goal of this study is to predict an individual's obesity status using a dataset of genetic, behavioral, and lifestyle predictors. We applied various statistical learning models, including logistic regression, decision trees, and random forests, to classify individuals as obese or not. The Random Forest model with 500 trees emerged as the best-performing model, achieving a testing accuracy of 99.943% and a low testing misclassification rate of 0.057%. Key predictors included frequent consumption of high-calorie food (FAVC), frequency of physical activity (FAF), daily water intake (CH2O), family history of being overweight, and the number of main meals (NCP). This study underscores the importance of integrating lifestyle, behavioral, and genetic factors in predictive modeling for obesity and offers insights for public health interventions.

Introduction

Obesity is a significant public health challenge affecting millions worldwide. It is associated with a wide range of chronic conditions, including diabetes, cardiovascular diseases, and certain types of cancer. Understanding the factors contributing to obesity and predicting an individual's obesity status is crucial for guiding early intervention efforts.

The objective of this study was to predict obesity status using a dataset containing a combination of genetic, behavioral, and lifestyle factors. The data provided included 42,686 observations and 29 predictors, capturing aspects such as dietary habits, physical activity, water consumption, and family medical history. This study applied statistical learning models to classify individuals as obese or not obese, using predictive accuracy as the key evaluation metric. Specifically, we sought to identify the most important predictors and build a robust model that minimizes misclassification rates while maintaining interpretability.

By leveraging a Random Forest model with 500 trees, this study highlights the critical predictors of obesity and provides practical insights for healthcare providers, policymakers, and individuals. The findings from this research contribute to the growing field of obesity prevention and management through data-driven approaches.

Data Analysis

We were provided with a training and testing dataset. The training set had 32,014 observations and the testing set had 10,672 observations. There were a total of 29 predictors, 11 being numerical and 18 being categorical. Below is a list of all of the predictors.

Numerical predictors:

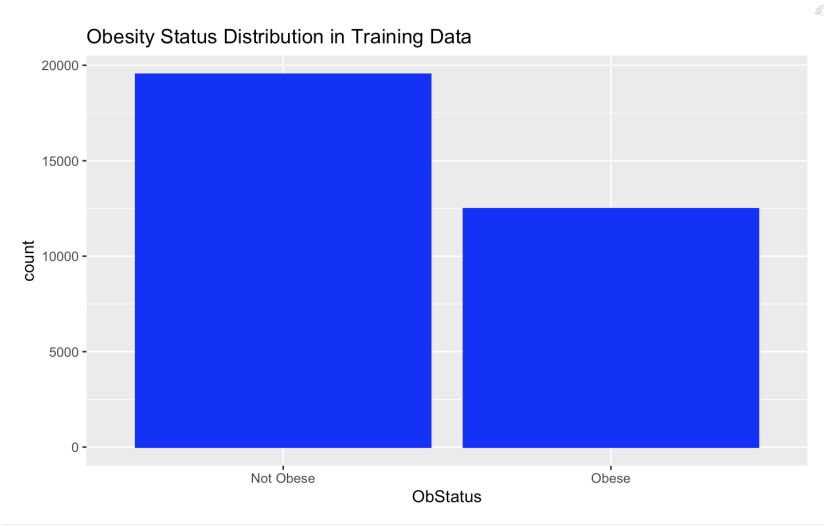
- Age
- Height
- FCVC: frequency of vegetable consumption
- NCP: number of main meals
- CH2O: daily water intake
- FAF: physical activity frequency
- TUE: time using technology devices
- RestingBP: resting blood pressure
- Cholesterol: cholesterol level
- MaxHR: maximum heart rate
- avg_glucose_level: the individual's average glucose level

Categorical predictors:

- Gender
- CALC: caloric intake

- FAVC: frequent consumption of high-calorie food
- SCC: consumption of sweet drinks
- SMOKE: smoking habits
- Family history of overweight: whether there's a family history of overweight
- CAEC: consumption of food between meals
- MTRANS: transportation method
- Race: the reported race of the individual
- FastingBS: fasting blood sugar
- Resting ECG: resting electrocardiogram
- ExerciseAngina: whether the individual exercises or not
- HeartDisease: whether the individual has a heart disease or not
- Hypertension: whether the individual has hypertension or not
- ever_married: whether the individual ever married or not
- work_type: the individual's type of work
- Residence_type: the individual's type of residence
- stroke: whether the individual has had a stroke or not

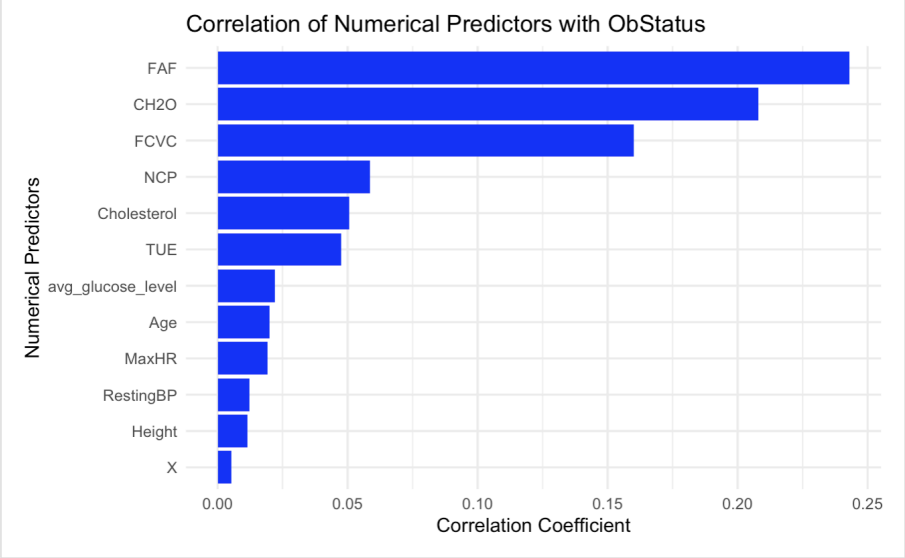
Our target variable is obesity status ("ObStatus"). Individuals were classified as either "Obese" or "Not Obese". Below is a bar graph showing the obesity status distribution in the training data. There are 12483 obese individuals and 19531 not obese individuals.



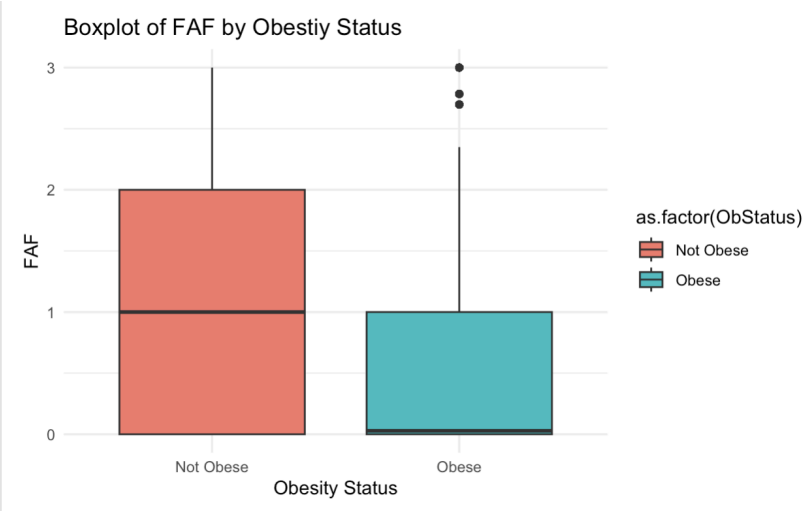
Before creating our models, we performed exploratory data analysis (EDA) to get a sense of which predictors have the largest correlation with obesity status. We used our imputed dataset to perform EDA so that there are no missing values. We discuss how we imputed our dataset in the Methods section.

Numerical Predictors

We computed the point-biserial correlation to find which numerical predictors have the highest correlation with obesity status. Below is a bar graph which ranks the predictors from most to least correlation. As we can see, the numerical predictors FAF, CH2O, FCVC, and NCP have the highest correlation with obesity status.

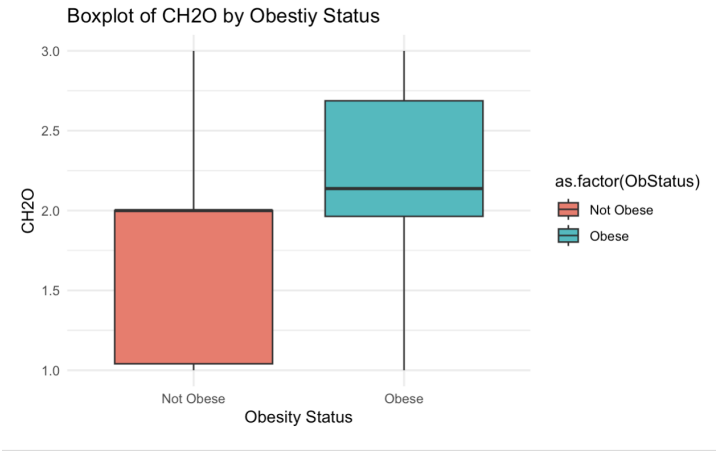


We then created boxplots of the numerical predictors with the highest correlation. First is a boxplot of FAF (physical activity frequency). We can see that individuals who are not obese tend to engage in physical activity more often than individuals who are obese. The median for the "Obese" box is extremely close to zero, telling us that half of the obese individuals do not engage in any physical activity.



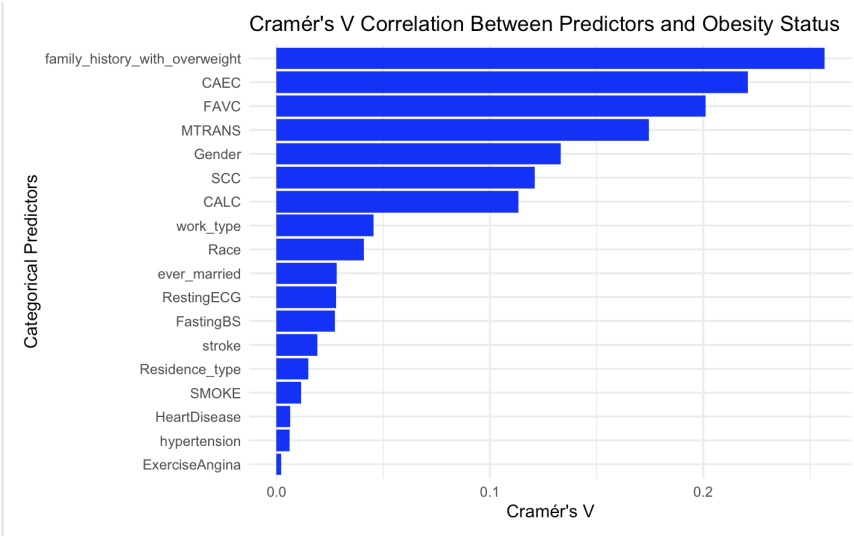
We also created a boxplot for CH2O (daily water intake). The middle 50% of non-obese individuals have a daily water intake of around 1-2, and the middle 50% of obese individuals have a daily water intake of 2-2.7. However, the median CH2Os of non-obese and obese individuals are very close with non-obese individuals having a median of 2 and obese individuals having a median of around 2.125. Also, both obese and non obese individuals have all levels of CH2O (from 1-3).

This tells us that obese individuals tend to have a slightly higher daily water intake. However, it is not uncommon for obese individuals to have low CH2O levels and non-obese individuals to have high CH2O levels. Daily water intake may play a factor in obesity status, but there are clearly other factors at play.

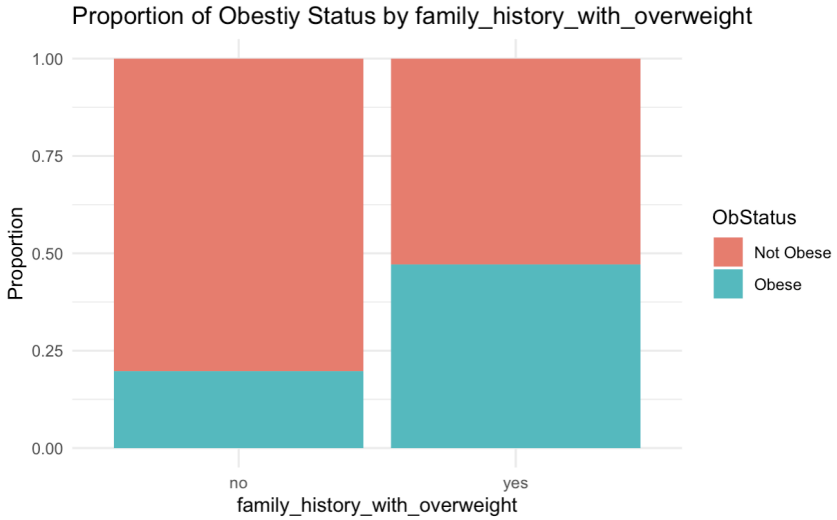


Categorical Predictors

To find the correlation between our categorical predictors and obesity status, we used Cramér's V Correlation. Below we have a bar graph that ranks the categorical predictors from highest to lowest Cramér's V Correlation. This shows us that the categorical predictors family_history_with_overweight, CAEC, and FAVC have the highest correlation with obesity status.

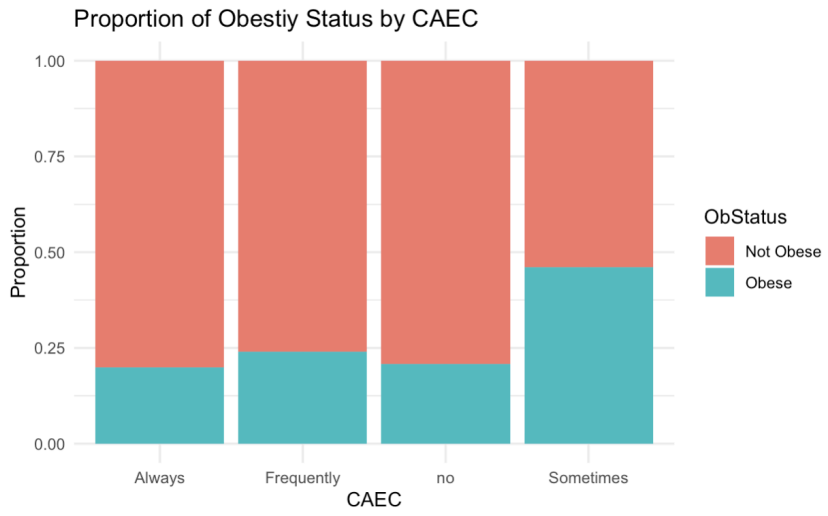


We then created proportional bar charts to look closer at the categorical predictors with the highest correlation with obesity. First, we looked at the predictor "family_history_with_overweight". We can see from the graph below that 50% of individuals who have a family history of being overweight are obese. Out of the individuals who do not have a family history of being overweight, only 25% are obese. This tells us that if an individual has a family history of being overweight, they are more likely to become obese.

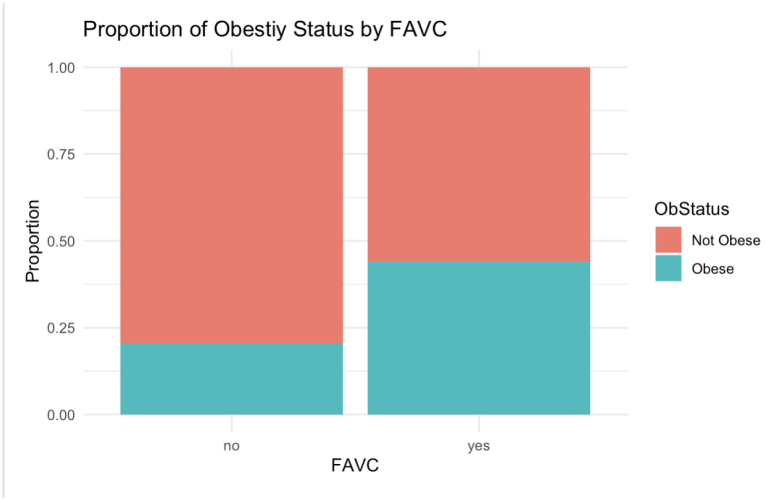


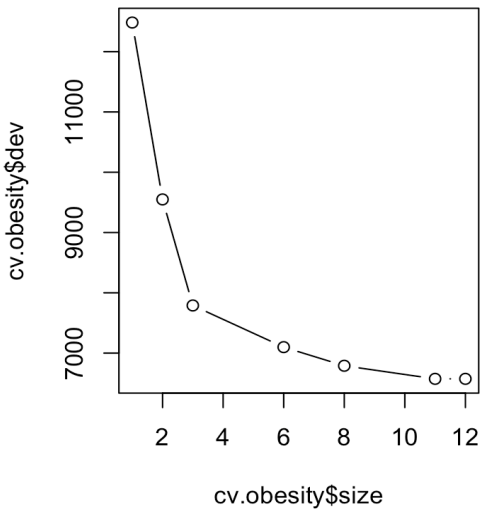
Looking at CAEC (consumption of food between meals), we see an interesting pattern. Out of the individuals who sometimes consume food between meals, a little under 50% of them are obese. Meanwhile, in each of the other categories ("always", "frequently", and "no"), a little under 25% of the individuals are obese. Therefore, individuals who only sometimes eat food between meals are more likely to become obese. This could be because it is very important to

have a consistent eating schedule. Eating at inconsistent times can disrupt one's circadian rhythm and lead to an increased risk of obesity, type 2 diabetes, and cardiovascular diseases (Nairn). For that reason, it may not matter whether an individual eats between meals, as long as they do it or don't do it consistently.



Lastly, we took a closer look at FAVC (frequent consumption of high-calorie food). Around 45% of individuals who frequently consume high-calorie food are obese, and around 20% of individuals who don't are obese. This suggests that people who frequently consume high-calorie food are more likely to become obese.





Methods

Imputation

The proportion of missing values per predictor ranges from 7.78% to 8.4% of values for a given predictor. This range does not warrant removing any variable entirely, so imputation for all predictors is possible. Random forest imputation, from the missforest package in R, was chosen as a nonparametric imputation method with high accuracy. Imputation through random forests occurs by predictor: the targeted variable is used as the response, while a model trained on observed values for that predictor is used to predict the missing values. Because of this, data was preprocessed for imputation by combining training and testing set predictors. Combining training and testing datasets and imputing missing values into this combined dataset helped increase imputation accuracy and consistency across training and testing datasets.

Logistic Regression

Logistic regression was chosen as an appropriate classification method because the predicted response variable of obesity status is binary. Obesity status was numerically coded: “Obese” was substituted with 1 and “Not obese” was substituted with 0. The classification threshold for predictions recorded as “Obese” is 0.5. The fitted logistic regression model had an R-squared value of 0.198 and a 24.9% training misclassification rate.

	Not Obese	Obese
Not Obese	16519	4957
Obese	3012	7526

Figure 1: Confusion Matrix for Logistic Regression Training

Decision Tree

A decision tree classification model splits the data into groups at each iteration until the data is impossible to split, resulting in terminal nodes of separate classifications. Cross-validation was used to balance model complexity and accuracy in determining the number of terminal nodes of the final model. As shown in Figure 2, the smallest model before steep increases in deviance has 6 terminal nodes. The final decision tree model has 6 terminal nodes. This model had a training misclassification rate of 22.093%. The final model is shown in Figure 4.

	Not Obese	Obese
Not Obese	17402	4944
Obese	2129	7539

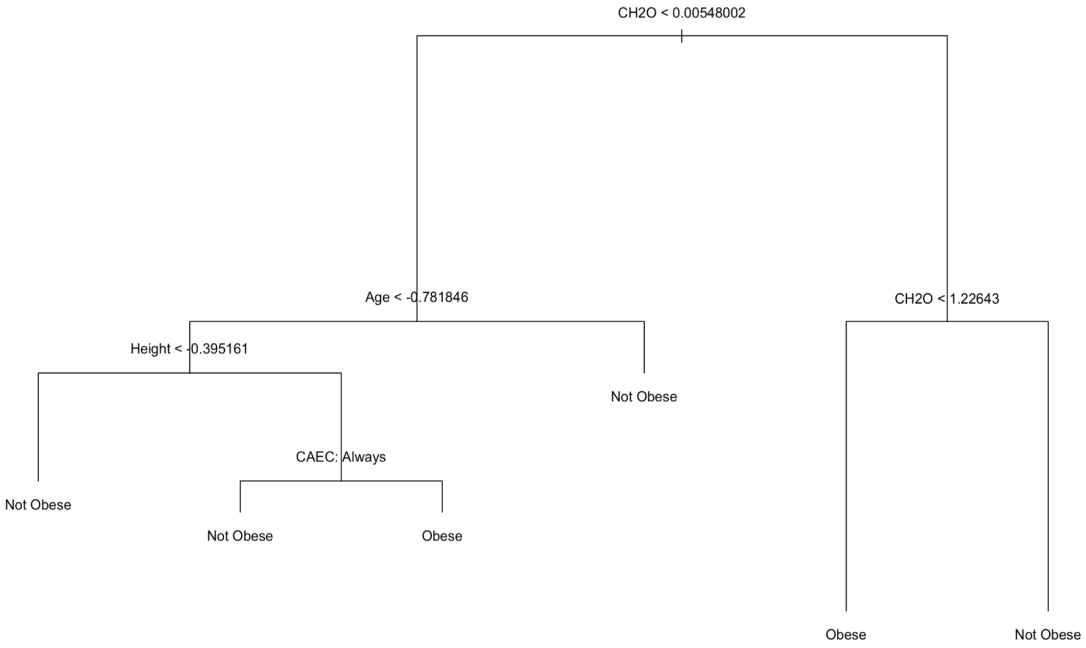
Figure 2: Plot of Decision Tree Size vs. Deviance in Cross-Validation Process

Figure 3: Confusion Matrix for Single Decision Tree Training (Best = 6)

Figure 4: Decision Tree (Best = 6)

Random Forest

Random Forest classification improves on the previous model of a single decision tree, creating multiple decision trees based on different subsets of predictors and random samples from the data. This method was chosen because of its high accuracy, compatibility with diverse variable



types, and multiple computational parameters that enable hyperparameter tuning of the model. Notable parameters for this model are the number of variables subsetted to grow each tree (`mtry`) and the number of trees (`ntree`). The former was held constant at 5 predictors, the square root of the total number of predictors which is the default for classification models. The latter, the number of trees, was varied: fitted models had either 500, 1000, or 2000 trees. The models with 1000 and 2000 trees had an estimated out-of-bag (OOB) error rate of 0.07%, while the model with 500 trees had an OOB error rate of 0.06%. All Random Forest models had a 0% training misclassification rate, demonstrating high accuracy and potential evidence of overfitting.

	Not Obese	Obese
Not Obese	19531	0
Obese	0	12483

Figure 5: Confusion Matrix for Random Forests Model Training

Results

The random forests model with random forest imputation had a testing prediction accuracy of 99.943% with 500 trees, 99.934% with 1000 trees, and 99.925% with 2000 trees. The logistic regression model with random forest imputation had a 75.037% testing prediction accuracy. The testing error for a single decision tree with random forest imputation is 77.164%.

Discussion

The final model chosen in this study is the random forest model with 500 trees. This model was chosen for its high accuracy, with a 0.057% misclassification rate and 6th place on the Kaggle leaderboard. The ultimate goal of this study is to predict obesity status, so accuracy was the main factor in choosing a final model.

Random forest models have a few key limitations. Firstly, random forest classification is often seen as a "black box," meaning it is difficult to see its mechanisms or attribute prediction results to any specific causes. Secondly, this model is highly complex, accounting for the output of 500 decision trees. This is further compounded by the inclusion of all predictors in our final model. High complexity increases variance and the chance of overfitting, which is already a known limitation of the random forest model. Further exploration may use variable importance in the random forest model to select only the variables that contribute most to the final model.

In considering the outputted variables of importance, it was noted that FAVC (frequency of high-calorie foods consumed), NCP (number of main meals), family history with being overweight, FAF (frequency of physical activity), and CH2O (daily water intake) were the predictors with highest importance. Looking specifically at FAVC, it can be observed that FAVC has the highest impact on predicting obesity status for our constructed model. This aligns with background knowledge of high-calorie foods, as they are usually high in sugar and fat, which contributes to excess calorie intake beyond daily needs and increases the likelihood of obesity.

Conclusion

In predicting obesity status, our chosen random forest model ultimately highlights environmental, behavioral, and genetic risk factors for obesity. This can inform healthcare providers of which patients would most benefit from potential preventative care interventions,

guide public health professionals in developing programs or policies to curb rising obesity rates, and empower people with knowledge about their health. In the future, this model could be expanded to predict who will develop obesity in the future, rather than who is already obese. The significance of our model lies primarily in the public health and medical fields, especially for individuals looking to better understand and take care of their health.

References

Almohalwas, Akram Mousa. “Using Kaggle plus some CDC health related Data Set for Predictive Analysis about obesity status”. 30 Nov. 2024.

Kumar, Satyam. "Predict Missing Values in the Dataset." *Towards Data Science*, 26 Jul. 2020, https://towardsdatascience.com/predict-missing-values-in-the-dataset-897912a54b7b

IBM Cloud Learn Hub. "Exploratory Data Analysis." 25 Aug. 2020, https://www.ibm.com/cloud/learn/exploratory-data-analysis

Nairn, Rayven. “Timing is Everything: Why Eating on a Regular Schedule Supports Overall Well-Being”. *John Hopkins University: Student Well-Being Blog*. 9 Dec. 2022, https://wellbeing.jhu.edu/blog/2022/12/09/timing-is-everything-why-eating-on-a-regular-schedule-supports-overall-well-being/