

Stats 101C Final Project

Predicting Obesity Status

Chynna Arceno, Natasha Mauro, Gabriel Soria, Mari Yamamoto, Xinyi Yang

December 16, 2024

Abstract

The rising prevalence of obesity has highlighted the importance of leveraging data-driven approaches for early detection and prevention, as obesity in the US affects over 40% of adults and contributes to rising healthcare costs and long-term chronic diseases. This project explores the Obesity dataset, with the primary objective to predict an individual's obesity status.

Here, we will describe our methodology for this classification project, from exploratory data analysis and feature selection based on patterns in the data associated with obesity and chronic disease indicators, to building and evaluating models. Due to the dataset's scale, advanced machine learning models were developed to classify obesity status. We then discuss the model's performance and potential limitations and recommendations. The final model is a Random Forest model with 10 predictors, resulting in a perfect Kaggle score and a rank of 5th. This project demonstrates the potential of predictive modeling in addressing global health challenges related to obesity and chronic diseases, emphasizing the critical role of data science in public health initiatives.

I. Introduction

Obesity is one of the most increasingly prevalent health conditions in the United States, affecting more than 2 in 5 adults and representing a significant contributor to chronic diseases and rising healthcare costs. Every year, obesity-related health complications result in approximately \$173 billion (in 2019 dollars) in medical expenditures and reduce the quality of life for millions. Obesity is also often associated with more serious conditions such as Type 2 diabetes, heart disease, and hypertension. However, predicting and addressing obesity remains challenging due to its multifaceted nature and the influence of various demographic, socioeconomic, and behavioral factors, as well as genetic ones.

Existing methods for assessing obesity risk often involve clinical evaluation or self-reported measures, which may lack precision or scalability for large populations. Thus, identifying efficient data-driven models for predicting obesity status can play a pivotal role in public health strategies aimed at prevention and early intervention.

In this Kaggle project, the Obesity dataset provided us with 29 predictor variables describing a range of factors such as physical activity, eating habits, and chronic disease indicators. The dataset consists of both numerical and categorical predictors, with features capturing key aspects of individuals' lifestyles and health profiles. The training dataset contained 32,014 observations and the testing dataset contained 10,672 observations. Our objective was to develop a classification model capable of accurately predicting obesity status. Through exploratory data analysis, feature selection, and rigorous model evaluation, we identified key predictors and constructed a Random Forest model with 10 variables. This research highlights the potential of predictive analytics to address the growing global challenge of obesity and inform data-driven public health initiatives.

II. Data Analysis

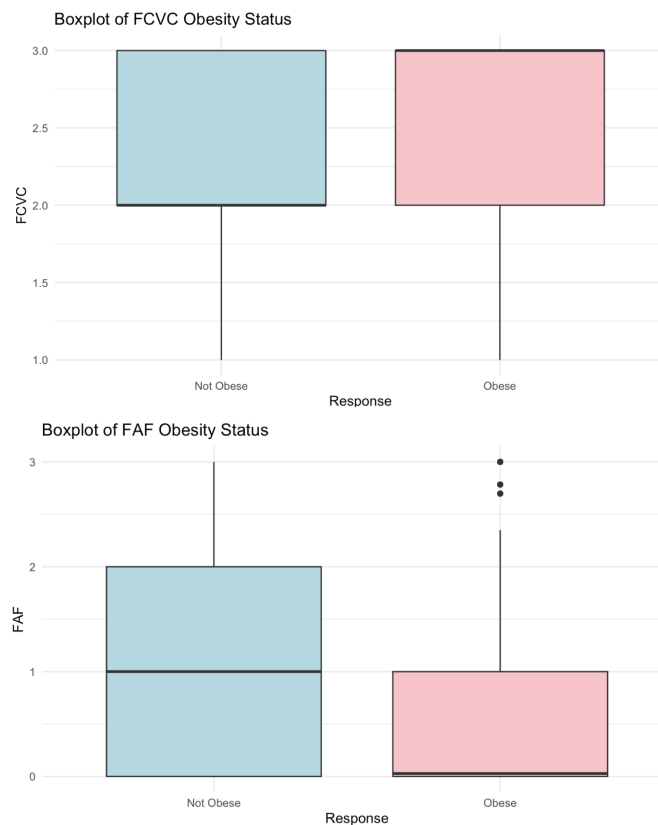
A. Exploratory Data Analysis (EDA)

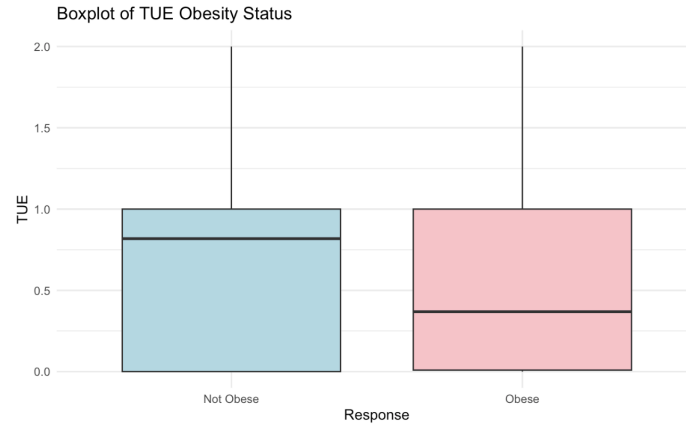
First, we investigate the data to see if there are certain patterns or relations between variables in the dataset. We will start with exploring the 11 numerical variables using boxplots and then break down the 18 categorical variables using frequency tables.

i. Numerical Variables

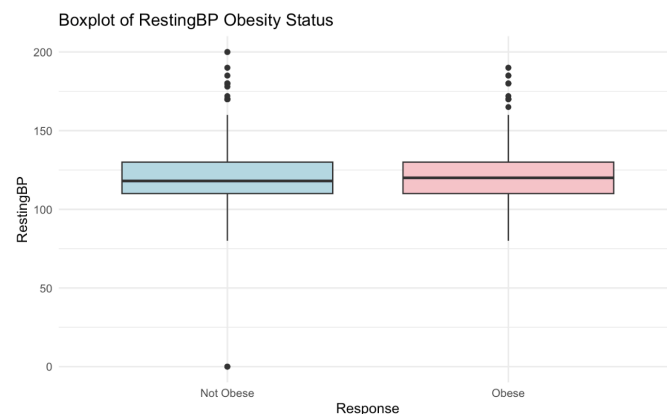
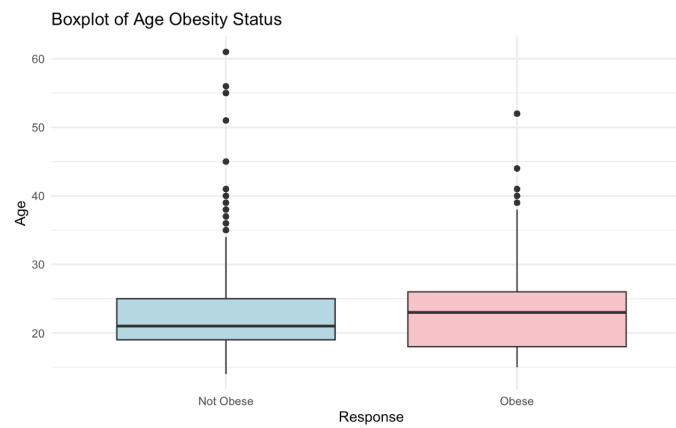
We found several patterns when comparing the boxplots.

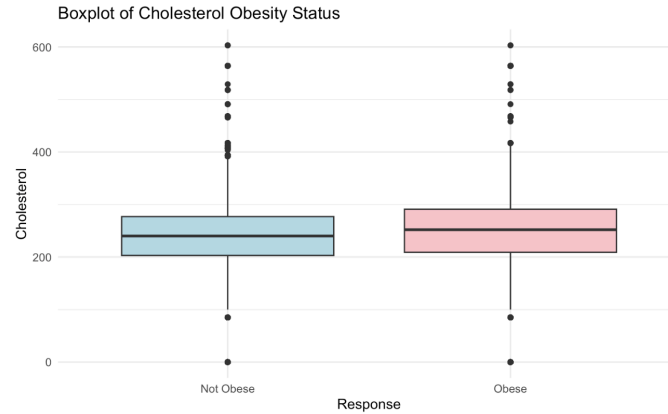
Pattern 1: FCVC, FAF, and TUE are variables with the median skewed to one end of the plot suggests asymmetric distribution. All of the variables here display difference in median, which could suggest an effect on obesity.





Pattern 2: Age, RestingBP, Cholesterol, MaxHR, and avg_glucose_level are variables with large amounts of outliers, indicating non-normal distribution. Here are some of selected variables shown here to visualize:





Pattern 3: FCVC, CH2O, FAF, and TUE variables with greater variability, which may suggest heterogeneity.



ii. Categorical Variables

We have 19531 observations for not obese and 12483 observations for obese in the training set for our target variable obesity status.

Levels	Frequency
Not Obese	19531
Obese	12483

We explored the levels of each categorical variable in our dataset:

Variable	Levels
Gender	2
family__history__with__overweight	2
FAVC	2
CAEC	4
SMOKE	2
SCC	2
CALC	4
MTRANS	5
Race	4
FastingBS	2
RestingECG	3
ExerciseAngina	2
HeartDisease	2
hypertension	2
ever__married	3
work__type	6
Residence__type	3
stroke	2
ObStatus	2

The levels range from around 2 levels in most variables to 5 or 6 levels in MTRANS and work_type. We kept in mind these number of levels when evaluating the feature selection in the last part of this data analysis.

B. Clean Data

There are around 7% of missing values in both the training and testing data. Missing values can lead to biased predictions and unreliable results in the model that does not generalize well to unseen data. It is important to remove them effectively in order to train an accurate machine learning model. There are no missing values in the numerical variables, as most of these missing values reside in the categorical variables ever_married, work_type, and Residence_type. Across each of those variables, they have the same proportion of missing values. Here is the following missing values for both training and testing missing values:

Training	Testing
0.02077216	0.02276987

We will impute these missing values by replacing them with the most frequent categorical levels instead of removing them entirely. This is so we can preserve the original sample size of the data

and prevent any biases and loss of randomness. Here is the following values we replaced with each categorical variable:

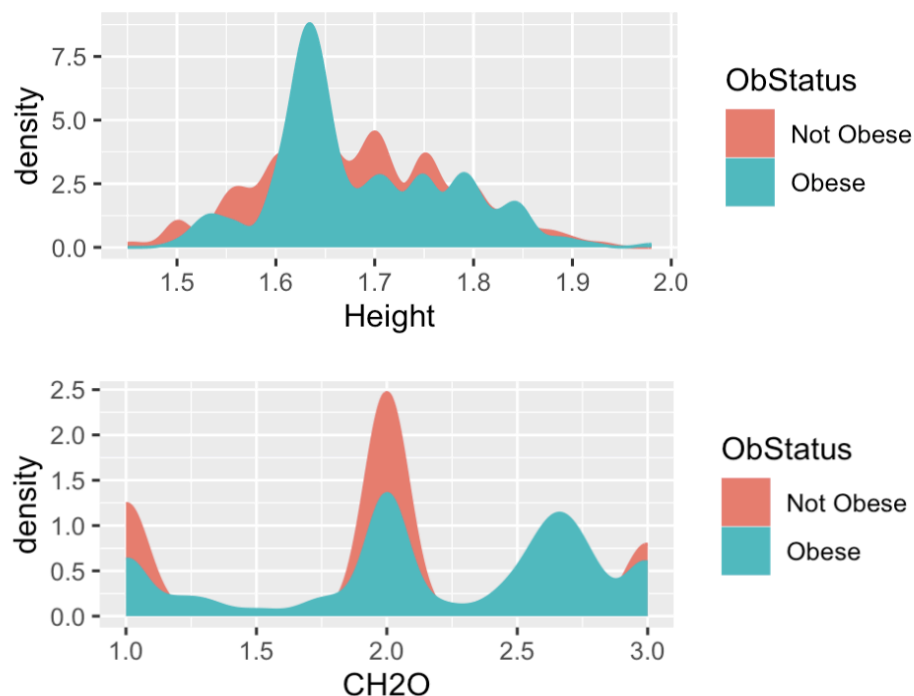
every_married	work_type	Residence_type
No	Private	Rural

C. Feature Selection

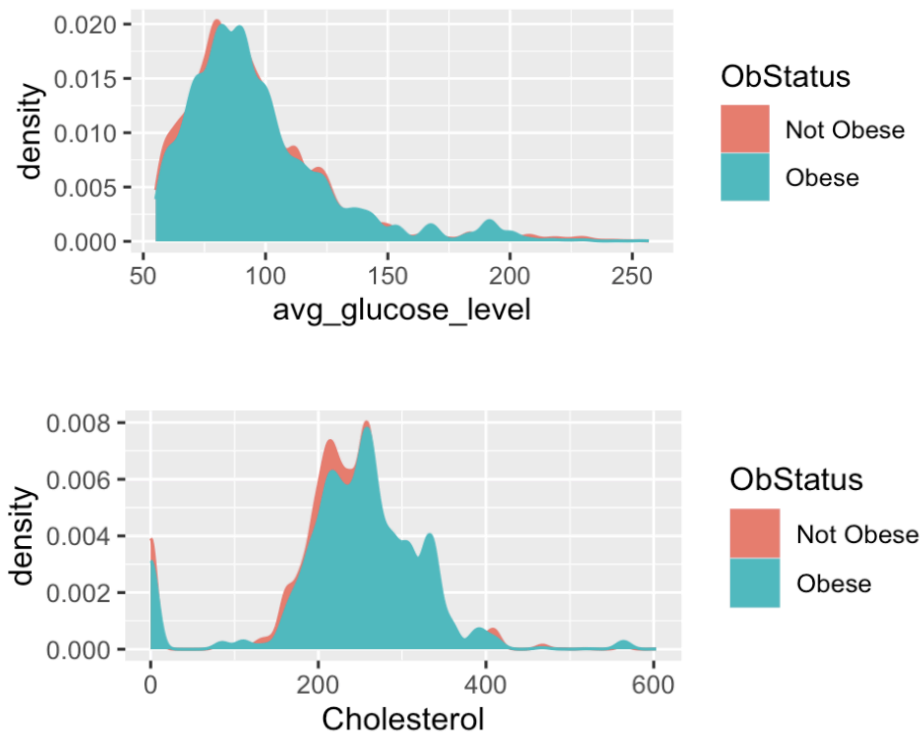
We will move on to selecting the variables that we would choose for our machine learning model after removing all the missing values. We will choose between numerical and categorical variables separately by keeping the significant predictors and removing the insignificant ones.

i. Numerical Variables

We chose density plots for selecting numerical variables because our target variable, obesity status, is a binary categorical variable. This helps us see the separability between the two categories among each numerical predictors, while also identifying any non-normal, overlapping distributions that signify potential insignificance. Here are examples of ones that best separate the categories:



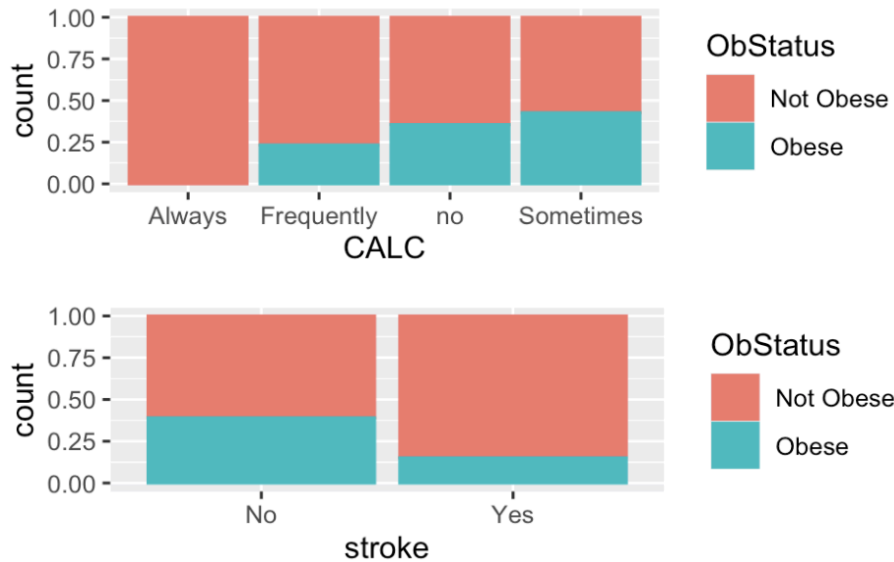
We can see the separation between the two curves more clearly compared to other density plots. This means that the 7 variables Height, CH2O, Age, FCVC, NCP, FAF, and TUE can be significant indicators of obesity status. We decided to keep these variables for our model. To see the difference and what variables we removed, here are examples of overlapping distributions:



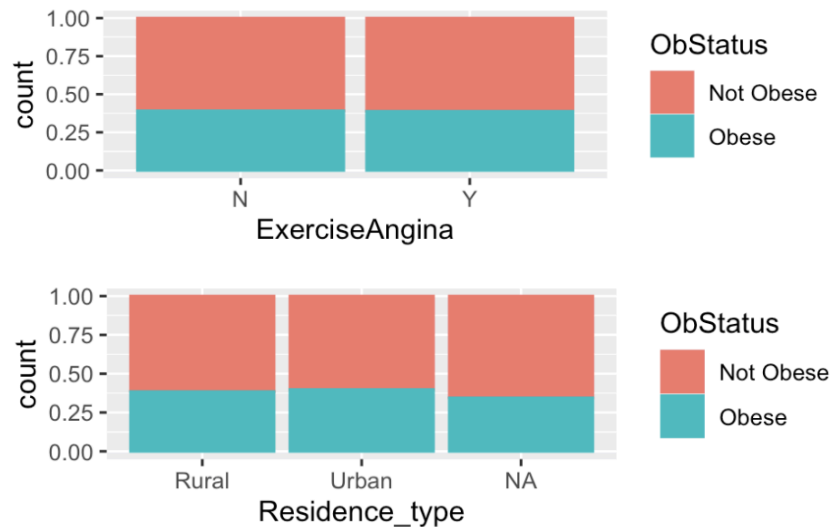
We removed 4 variables: avg_glucose_level, Cholesterol, MaxHR and RestingBP. Because they have a greater overlap in density curves, these variables have no difference in the probabilities of obese and non-obese status.

ii. Categorical Variables

Similar to the numerical variables, we chose to plot stacked bar charts for the categorical variables in order to evaluate if these variables are significant to obesity status. A stacked bar chart will help us compare the proportion within the binary category of obese and non-obese distributed across each level of the categorical variable. The height of each bar reveals if certain variables are more or less associated with the obesity status. Here are examples of high association:



There is a strong association between the categorical and obesity status since either obese or non-obese would dominate a specific level or levels relative to others levels. This suggests a relationship between the categorical variables (CALC, stroke, Gender, family_history_with_overweight, FAVC, CAEC, Smoke, SCC, MTRANS, Race, ever_married) and the target variable ObStatus. We will select these 11 categorical variables. Here are examples of less association:



ExerciseAngina, Residence_type, and other variables (FastingBS, hypertension, RestingECG, HeartDisease, work_type) have similar proportions of obesity status across each level, indicating weak or no association between the two variables because the probabilities of obese and non-obese does not change between differing levels. We decided to remove these 7 categorical variables.

We finalize our variables for our model with 18 from 29 predictors.

III. Methods and Models

A. Why We Chose Random Forest

Although we tested out other models on our data, we selected the Random Forest model because it offers several advantages that fit our dataset well. Random Forest is an ensemble learning method that builds multiple decision trees using bootstrap samples of the data. It then aggregates the results to improve predictive accuracy and mitigate overfitting. Some key reasons we chose Random Forest are:

- **Handles Mixed Data Types:** Since our dataset contains both numerical and categorical predictors, we needed a model that worked with both types. Random Forest can work with both variable types easily.
- **Feature Importance:** Random Forest illustrates which predictors are the most important for making accurate predictions. Since the dataset contains a variety of predictors, we wanted to ensure only the most influential were included in the model to prevent overfitting.
- **Reduces Overfitting:** By combining multiple decision trees, Random Forest reduces the chances of overfitting compared to using a single tree.
- **Scales Well:** Random Forest works efficiently even with large datasets, which is important considering the size of our data.

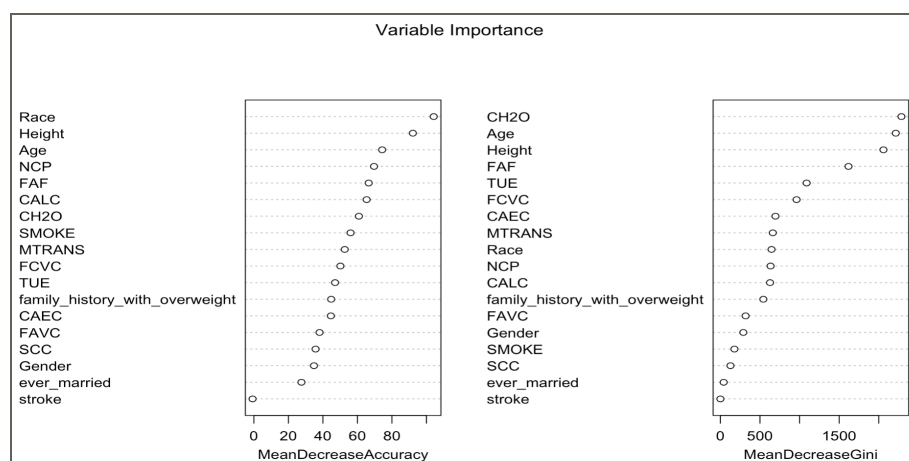
B. Model Implementation

We tested two versions of the Random Forest model to compare their performance: one with all predictors and another with a reduced set of predictors.

Full Model

Initially, we implemented a Random Forest model that included all relevant predictors after data preprocessing. This model resulted in a total of 18 predictors with the number of trees set to 500 to ensure robust performance.

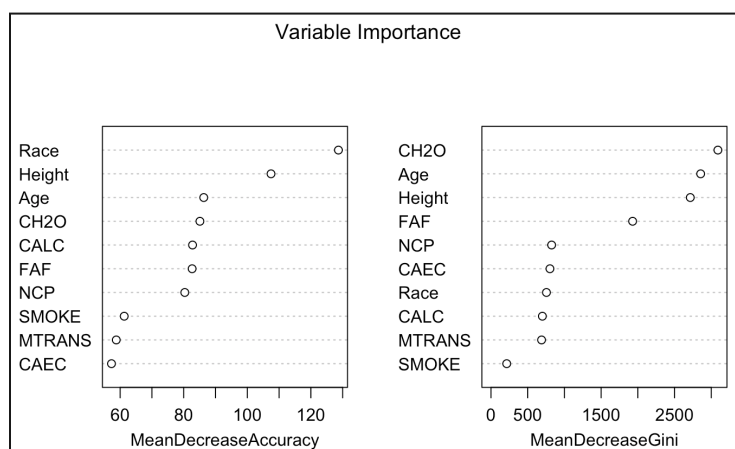
The full model performed well, achieving an Out-of-Bag (OOB) error rate of 0.31% and a misclassification rate of 0.003 on the training data. When evaluated on the test set, it achieved a Kaggle score of 0.99306. However, upon examining the variable importance plot, we found that not all predictors were equally useful. Some features contributed minimally to the model's performance, suggesting that the model might benefit from excluding these less significant predictors to avoid overfitting and improve generalization.



Reduced Model

Based on the insights from the full model's variable importance analysis, we decided to try a model with a reduced set of predictors. We explored different subsets of the most important predictors according to the Mean Decrease Accuracy plot and ultimately decided on a model using 10 of some of the top predictors. These features were selected for their high importance scores, allowing for simplicity of the model that does not compromise its predictive power. For this model, we again used 500 trees.

The reduced model showed a slight increase in the OOB error rate to 0.75% and a misclassification rate of 0.017. Despite this, the reduced model achieved a perfect Kaggle score of 1.00, indicating superior generalization to the unseen test data. The reduced model was also simpler, faster to train, and easier to interpret due to the smaller number of predictors.



Model Comparison

The full model included 18 predictors and achieved an OOB error rate of 0.31% with a misclassification rate of 0.003. The reduced model, on the other hand, included only 10 predictors and had a higher OOB error rate of 0.75% with a misclassification rate of 0.017.

However, the reduced model outperformed the model on the Kaggle test set, achieving a score of 1.00 compared to the full model's score of 0.99306.

	Full	Reduced
Number of Predictors	18	10
OOB Error	0.31%	0.75%
Misclassification Rate	0.003	0.017
Kaggle Score	0.99306	1.00

While the full model fit the training data better, the reduced model demonstrated better generalization. By focusing on fewer, more influential predictors, the reduced model avoided overfitting and provided a more efficient and interpretable solution, which is why we chose this model as our final model. This model effectively balanced performance, efficiency, and interpretability, making it the optimal choice.

IV. Discussion and Limitations

Given that our raw dataset had a total of 29 predictors, we were able to simplify it through variable selection to 10 predictors. Our final random forest model includes the 10 predictors: NCP, Smoke, CAEC, FAF, CALC, MTRANS, Race, CH2O, Age, and Height. Based on our variable importance plot, Height (meters), Age, and CH2O (daily water intake) are the top 3 most influential predictors. This suggests that when predicting a patient's obesity status, height, age, and daily water intake are the most significant factors.

We settled on using the random forest model because it is able to handle large datasets with multiple predictors of different types such as numerical and categorical data as well as being able to handle missing values. However, we opted to replace the missing values with the mean and the most frequently occurring level. Another benefit is that the model is less sensitive to outliers and noise within the data by creating multiple trees thus, outliers are less likely to impact the model's performance.

When comparing our two random forest models, one applied bagging while the other used limited predictors. For our first model we applied bagging with the mtry parameter being set to the default value of 0.01 to reduce overfitting. As a result 18 predictors were used. For the reduced model, we opted to only include the important 10 predictors with ntree = 500. By simplifying the number of predictors, we were able to improve the model accuracy and interpretability.

While the random forest model performed exceptionally well, there were some areas for improvement in our approach. A notable limitation was our group's focus on reducing the number of predictors to simplify the model. While this helped improve interpretability and precision for certain predictions, it diverted attention from fine-tuning the number of trees in the random forest. A high number of trees, while improving stability, can lead to overfitting and increased computational complexity. By not fully exploring the impact of tree reduction, we may have limited the model's generalizability and efficiency. Future efforts should prioritize balancing predictor selection with hyperparameter tuning, ensuring the model is both robust and computationally efficient.

V. Conclusion and Recommendations

The primary goal of this report was to develop a model capable of accurately predicting an individual's obesity status. After testing multiple models, we found that the random forest model performed the best, achieving a misclassification rate of 0.017 and a perfect Kaggle score of 1.00. While exploring different approaches, we discovered that reducing the number of predictors does not always improve accuracy; however, it can enhance the model's precision when predicting actual obesity status. In our analysis, the most influential predictors for obesity were identified as height, age, and daily water intake. This insight emphasizes the role of fundamental lifestyle and demographic factors in obesity risk.

While our focus was primarily on reducing the number of predictors to simplify the model, future work could benefit from additional optimization of parameters such as the number of trees in the random forest. Balancing model complexity with predictive performance could enhance its robustness and real-world applicability.

Real-world applications of this model include raising public awareness about health risks associated with obesity and guiding interventions to promote healthier lifestyles. These findings underscore the potential of data-driven approaches to address significant public health challenges and support informed decision-making for both individuals and policymakers.

References

https://bruinlearn.ucla.edu/courses/195305/files/18631568?module_item_id=7021333