STATS 101C Final Project Paper - Predicting Obesity Status

Akshay Ashok, Dylan Winward, Junhyeong Park, Kyle Mukire, Nicole Lee

Abstract

The goal of this project is to predict obesity status, taking into account a host of variables that define an individual, their background, health, and lifestyle. This report gives a clear breakdown of how we cleaned and refined the large dataset, identified key features, constructed and explored different models, analyzed our results, and the conclusions we arrived at. The final model uses the RandomForests package with mtry = 5, which achieved a Kaggle score of 0.9912.

1. Introduction

Obesity is an epidemic affecting people of all ages, all over the world. The World Obesity Federation reported that in 2024, over 1 billion people were suffering from obesity worldwide (1), and this number is projected to increase significantly over the next few decades. Even more alarming is the fact that almost 160 million children and teenagers between the ages of 5-19 are suffering from obesity (1). Obesity can lead to various other health issues such as heart diseases, strokes, breathing problems, fatty liver diseases, and even some cancers (2). Because of its influence on such diseases and overall health, obesity is quickly rising to become one of the most prevalent, preventable causes of death, claiming 300,000 lives in the US annually (3). In comparison, tobacco use is associated with approximately 400,000 deaths per year in the US (3). These glaring figures highlight the urgency at which governments and healthcare officials need to act when addressing obesity. The first step is quicker identification and mitigating the risk at the earliest stage; to do this, we need to construct efficient models.

2. Exploratory Data Analysis

Prior to fitting models on the data, we wanted to explore some of the key attributes of individual variables and gain a greater understanding of our datasets. The Obesity dataset is split

75:25 between the testing and training set. In the training dataset, there are 32,014 observations across 30 variables, meaning we had 29 features to work with. In the testing dataset, there are 10,672 observations across the 29 feature variables. There are a mix of numerical and categorical predictors that could influence an individual's obesity status. The numerical variables give us an insight into the quantifiable factors like caloric intake and cholesterol that can influence obesity status, while the categorical variables show us the conditions/behavioural trends like race and stroke history that can influence obesity status.

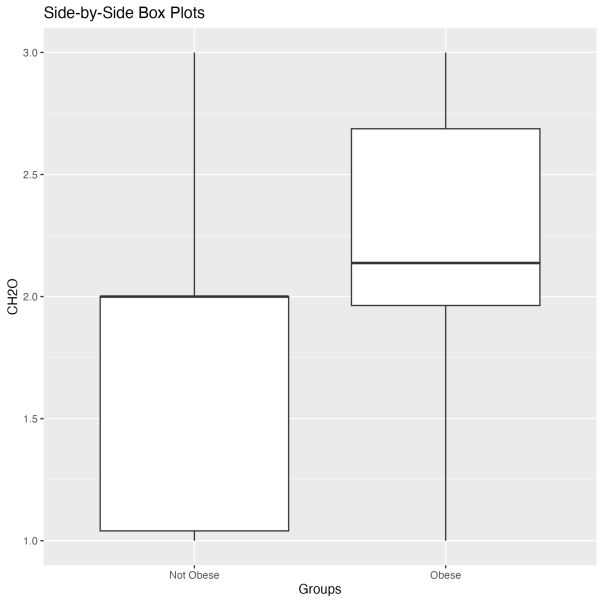
2.1 Categorical Analysis of Response Variable

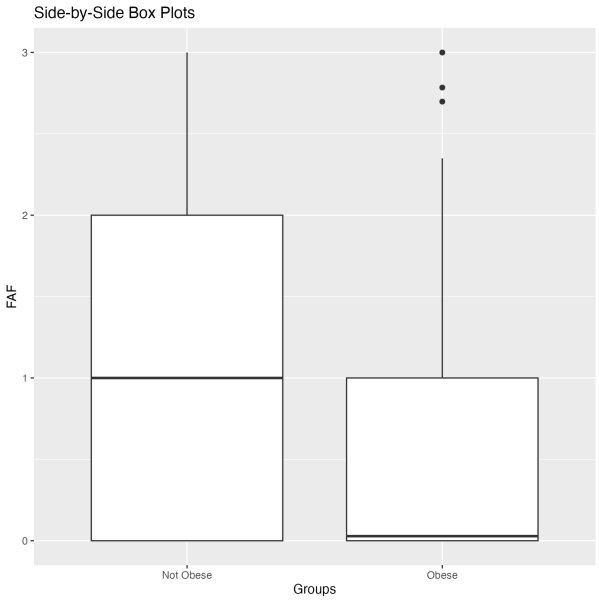
Around 39% of observations in our training dataset are categorized as obese. For the purposes of this project, we are assuming that data has been split into testing data and training data randomly so there are no patterns – this allows us to assume that 39% of observations in our test data should also be categorized as obese. As a result, our baseline error rate for any model testing will be 39%.

2.2 Numerical Analysis of Potential Features

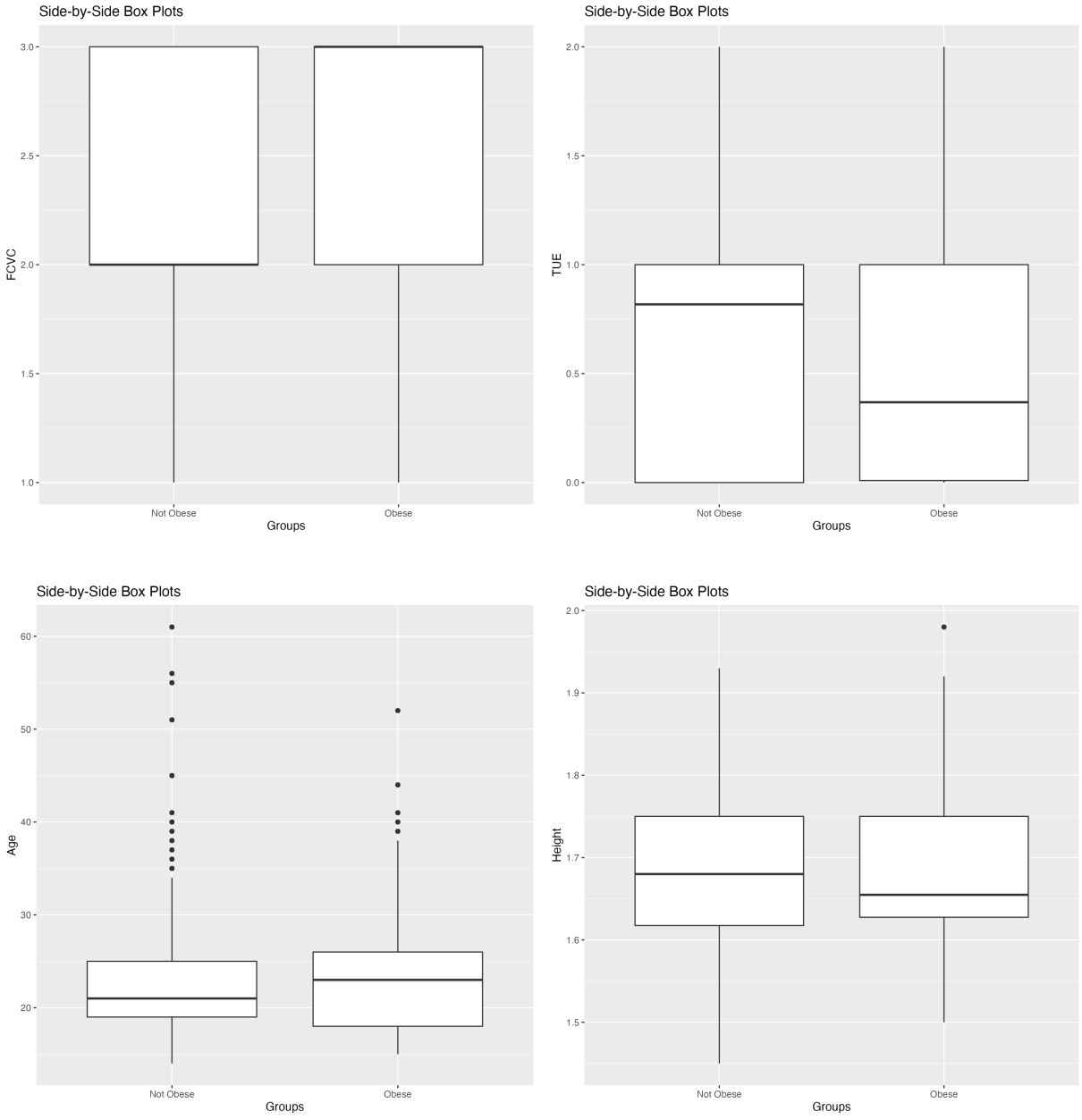
In the dataset, we had 11 numerical variables: the age of the individual to whom an observation corresponds, their height, the frequency of their vegetable consumption (FCVC), the number of main meals they have per day (NCP), their daily water intake (CH2O), the frequency of their physical activity (FAF), the amount of time they spend using technology devices (TUE), their resting blood pressure, their cholesterol, their maximum heart rate and their resting blood pressure.

From a box plot analysis, we saw a significant difference in some variables between the values associated with obesity and non-obesity. Specifically, the boxplots of daily water intake (CH2O) and frequency of physical activity (FAF) appear to imply that they would be strong predictors of obesity, as shown below.

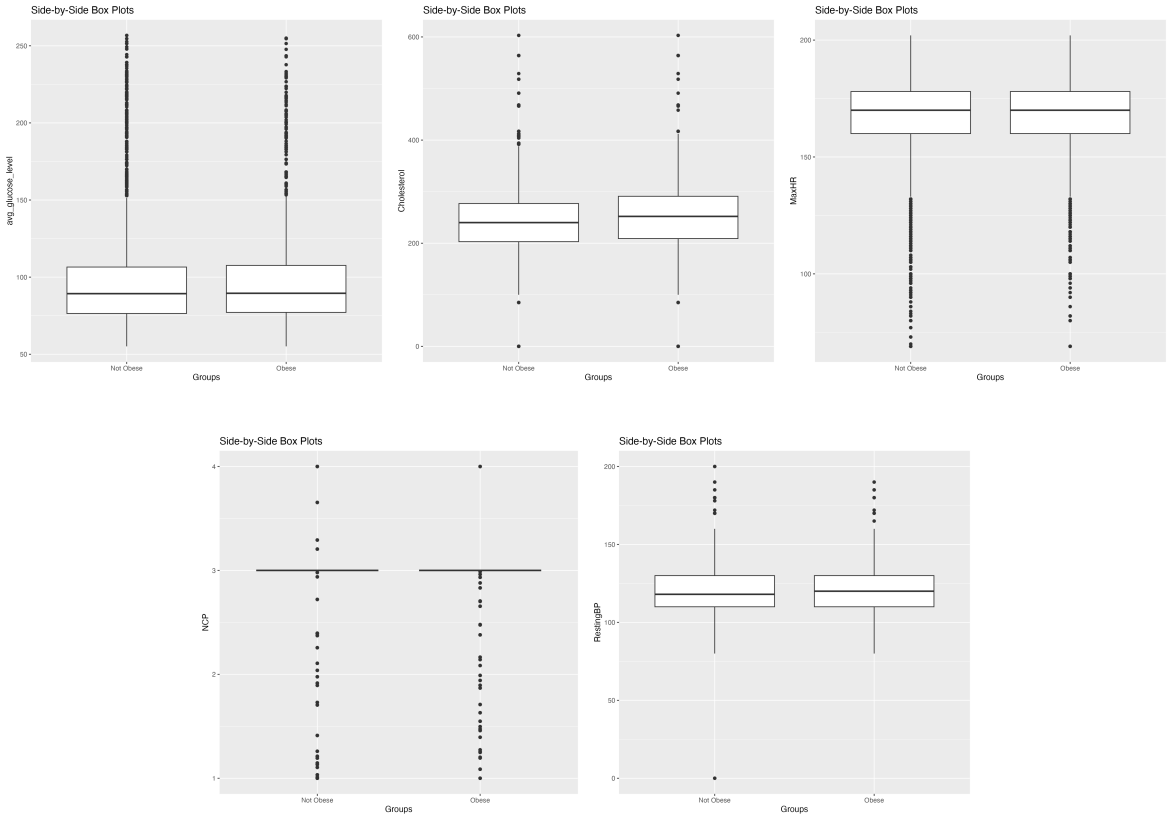




There is also a second group of numerical data in our study where there is a significant difference between the median values of the variable for our two response groups, but the overall distributions do not vary greatly. Frequency of vegetable consumption (FCVC), the amount of time the subject spends using technology devices (TUE), age, and height fall into this category, and are potential predictors worth exploring.



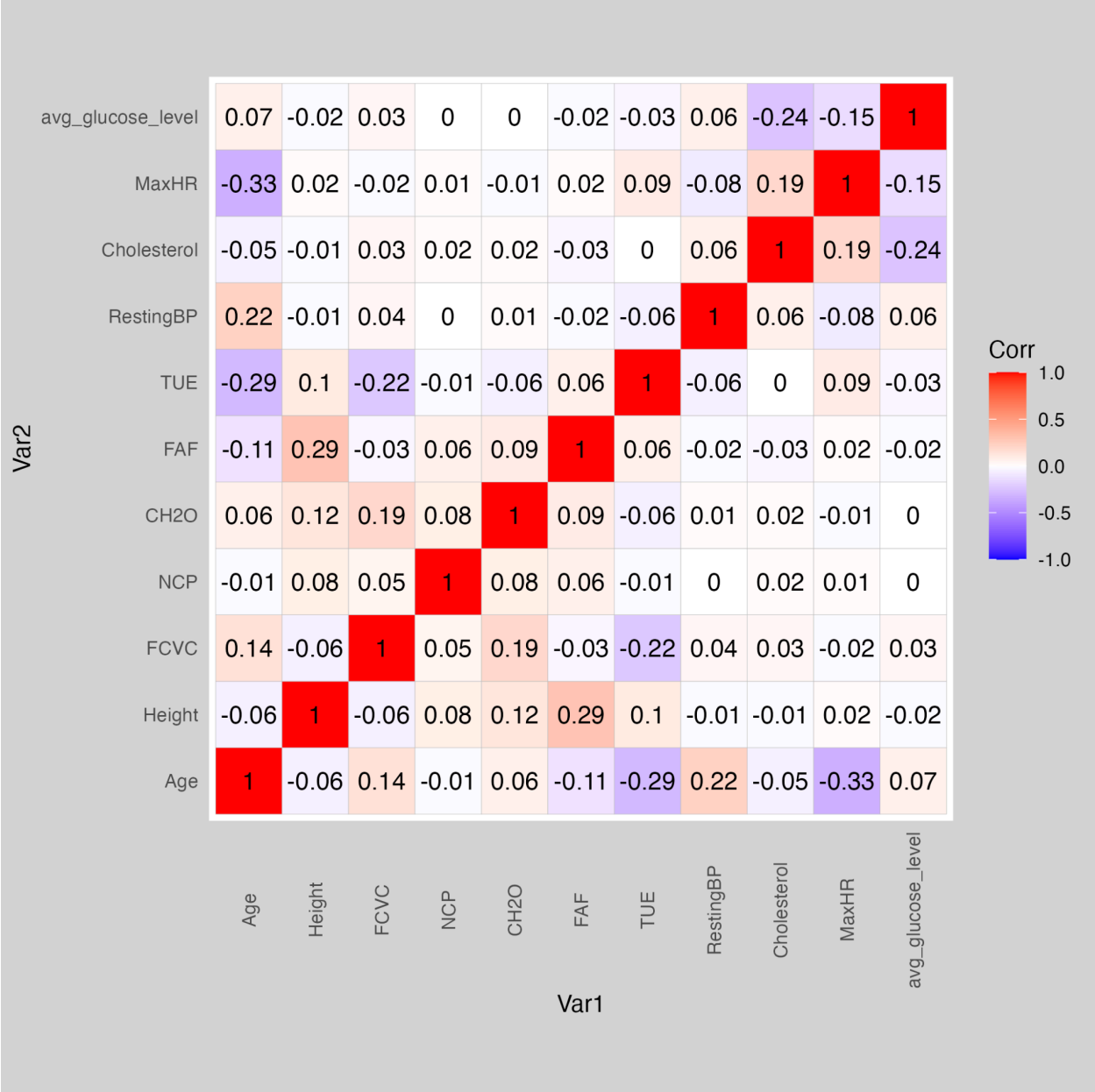
Finally, there is a group of variables where there appears to be no apparent difference between values associated with obesity and non-obesity. These variables include average glucose level of the test subject, their cholesterol, their maximum heart rate, the number of main meals they have per day and their resting blood pressure levels. These variables are unlikely to serve as good predictors for our final modeling.



Another important aspect of numerical data analysis is checking for internal correlation between variables. While analytics like Variance Inflation Factor can be used later to check for correlated variables during the modelling stage, it can be useful to get an initial understanding of variable correlation.

The largest correlation between numerical variables in our data appears to be between the age of a test subject and their maximum heart rate, where there is a reasonable negative correlation. Similarly, there is a light negative correlation between a subject's time spent using technology devices and their age. A positive correlation between the frequency of a subject's physical activity and their height also appears to exist.

Correlation Plot



2.3 Categorical Analysis of Potential Features

In the dataset, we had 19 categorical variables, one of which was our response variable. As a result, there were 18 categorical variables that could be used as potential features.

Variable	Gender	Family History with Overweight	Frequent consumption of high-calorie	Consumption of food between meals	Whether someone smokes (SMOKE)

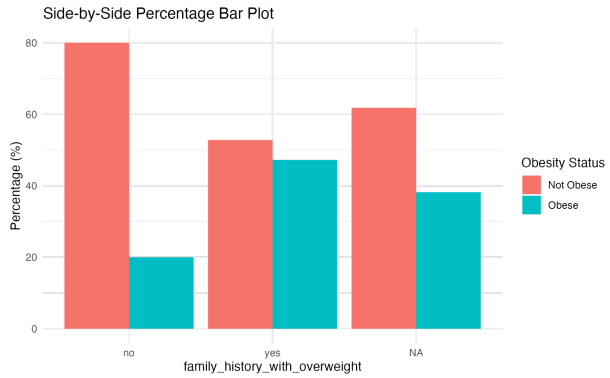
			food (FAVC)	(CAEC)	
Levels	2	2	2	4	2

Variable	Consumption of sweet drinks (SCC)	Caloric intake (CALC)	Method of transportation (MTRANS)	Race	Fasting Blood Sugar (FastingBS)
Levels	2	4	5	4	2

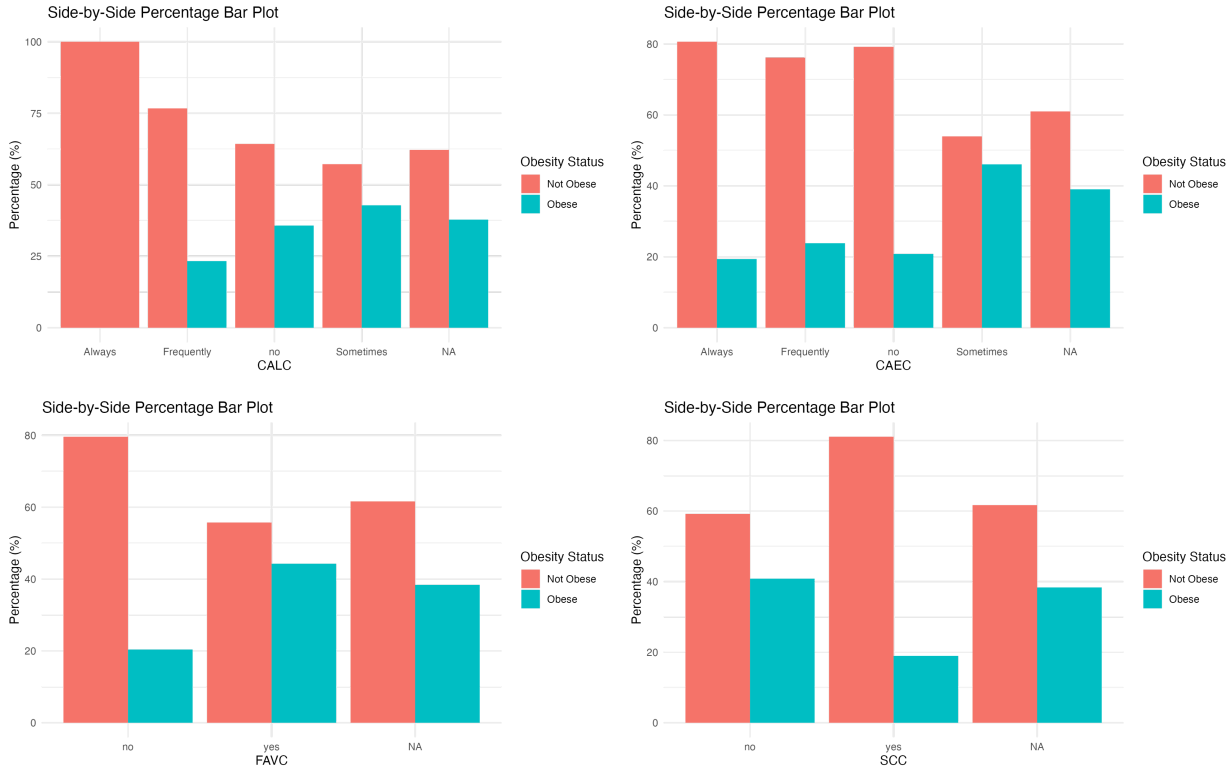
Variable	Resting ECG	Whether the individual exercises (Exercise Angia)	Heart Disease	Hypertension	Whether the individual has ever married (ever_married)
Levels	3	2	2	2	2

Variable	Work Type	Residence Type	Whether the individual has had a stroke (Stroke)
Levels	5	2	2

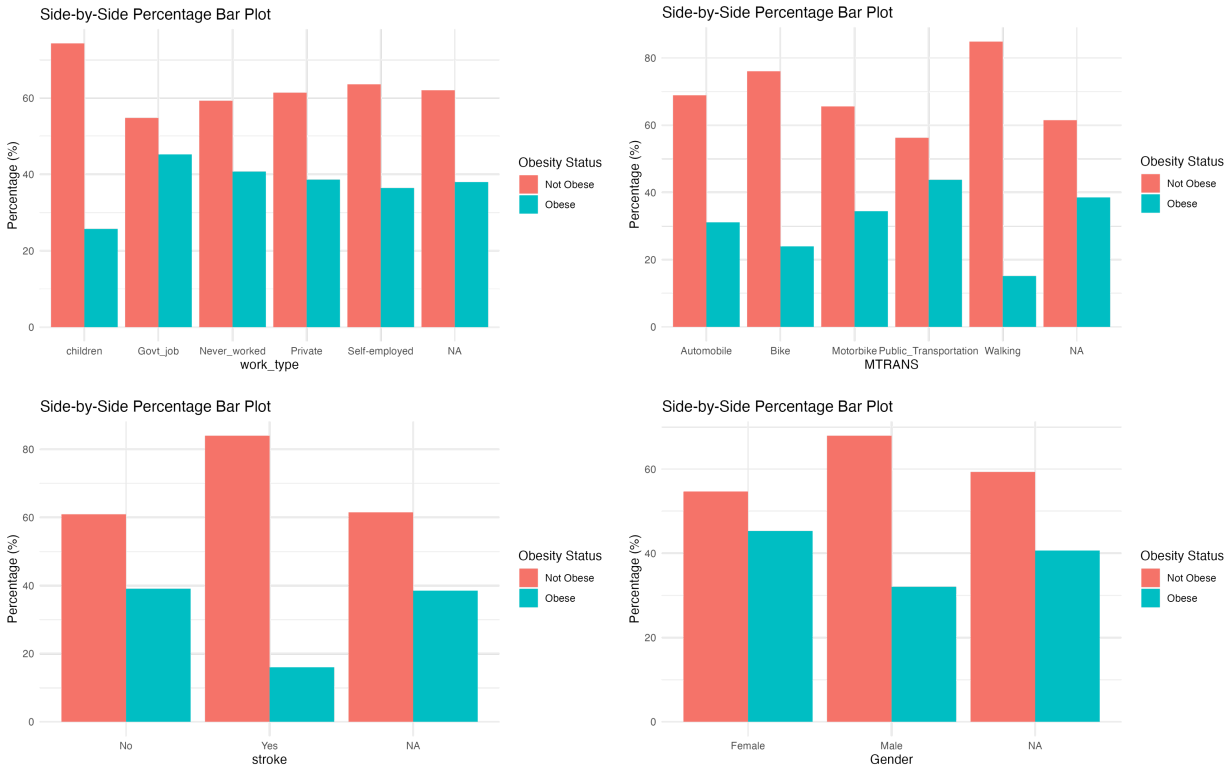
We can then investigate the frequency with which each of these variables appears to relate to obesity. One of the areas where there is a stark difference when it comes to likelihood of being obese is family history, with a subject being twice as likely to be obese if they have someone in their family who was.



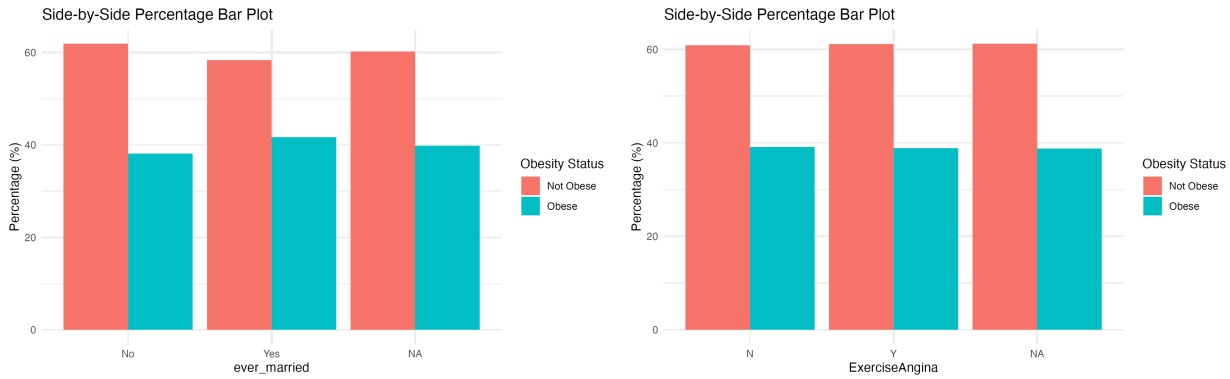
There are also a number of variables that appear to have a significant relationship with obesity status. For example, the caloric intake (CALC), consumption of food between meals (CAEC), frequent consumption of high calorie meals (FAVC) and consumption of sweet drinks (SCC) all appear to have a significant impact on someone's likelihood of being obese. Intuitively, many of these make sense as they are related to direct consumption habits.

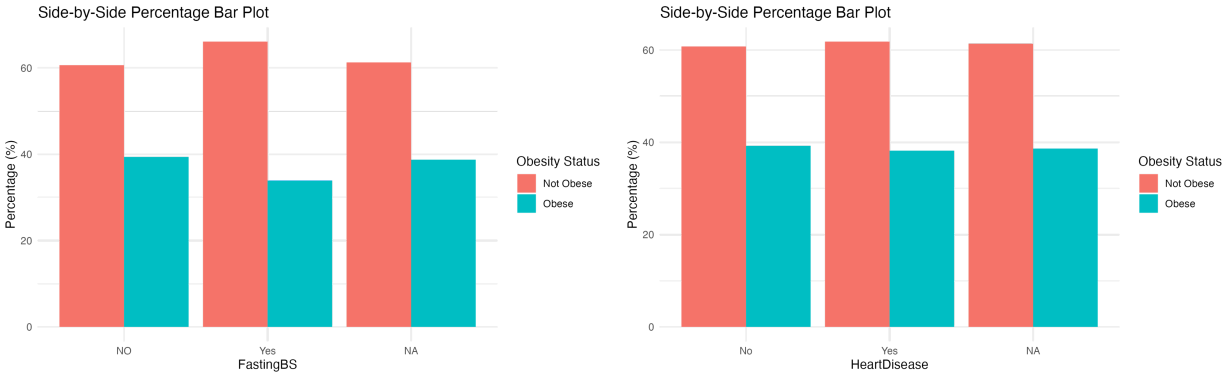


For some variables, there appeared to be a smaller but still significant difference between the frequency of variables when it came to obesity or non-obesity. From our side by side proportion barplots, people's work type, transportation method, gender and whether they have had a stroke appear to be potential predictors of obesity status.

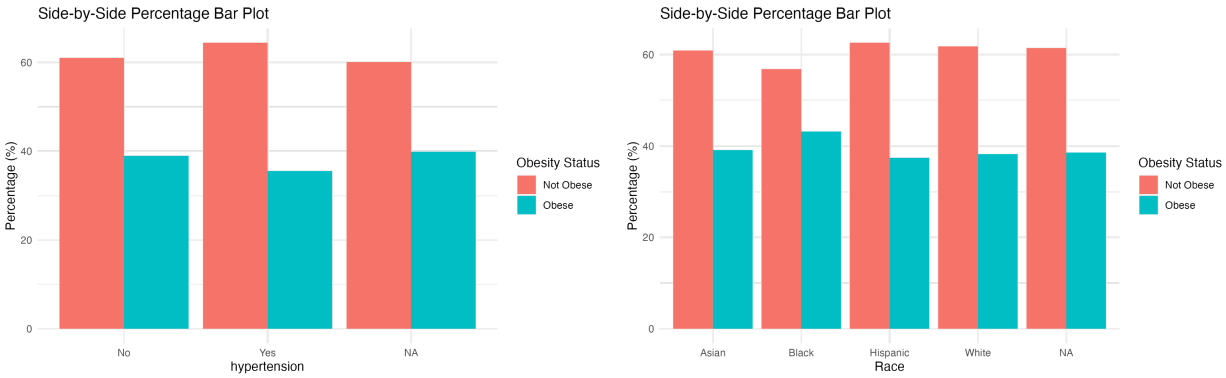


Finally, there are a number of variables – including marital status, whether someone has had heart disease, fasting blood sugar and whether the individual exercises or not – that do not appear to significantly relate to obesity status.

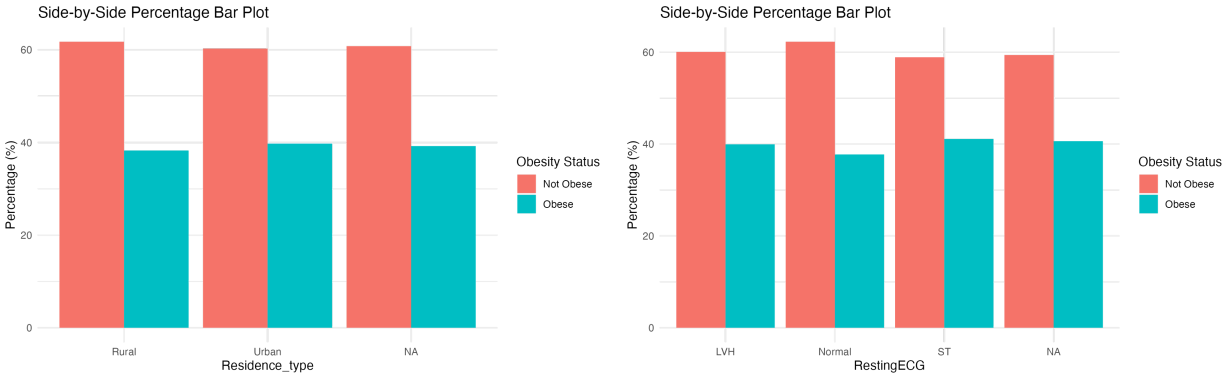



Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
NO
Yes
NA
FastingBS
Obesity Status
Not Obese
Obese

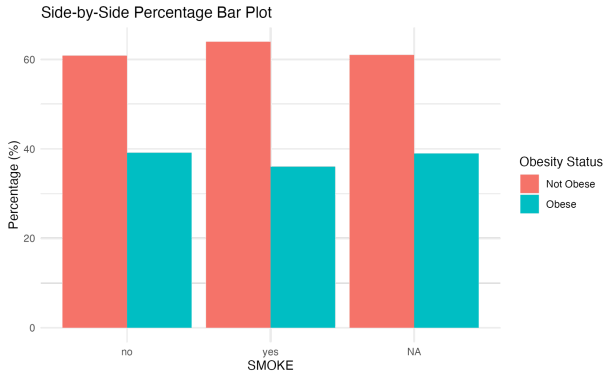
Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
No
Yes
NA
HeartDisease
Obesity Status
Not Obese
Obese


Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
No
Yes
NA
hypertension
Obesity Status
Not Obese
Obese

Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
Asian
Black
Hispanic
White
NA
Race
Obesity Status
Not Obese
Obese


Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
Rural
Urban
NA
Residence_type
Obesity Status
Not Obese
Obese

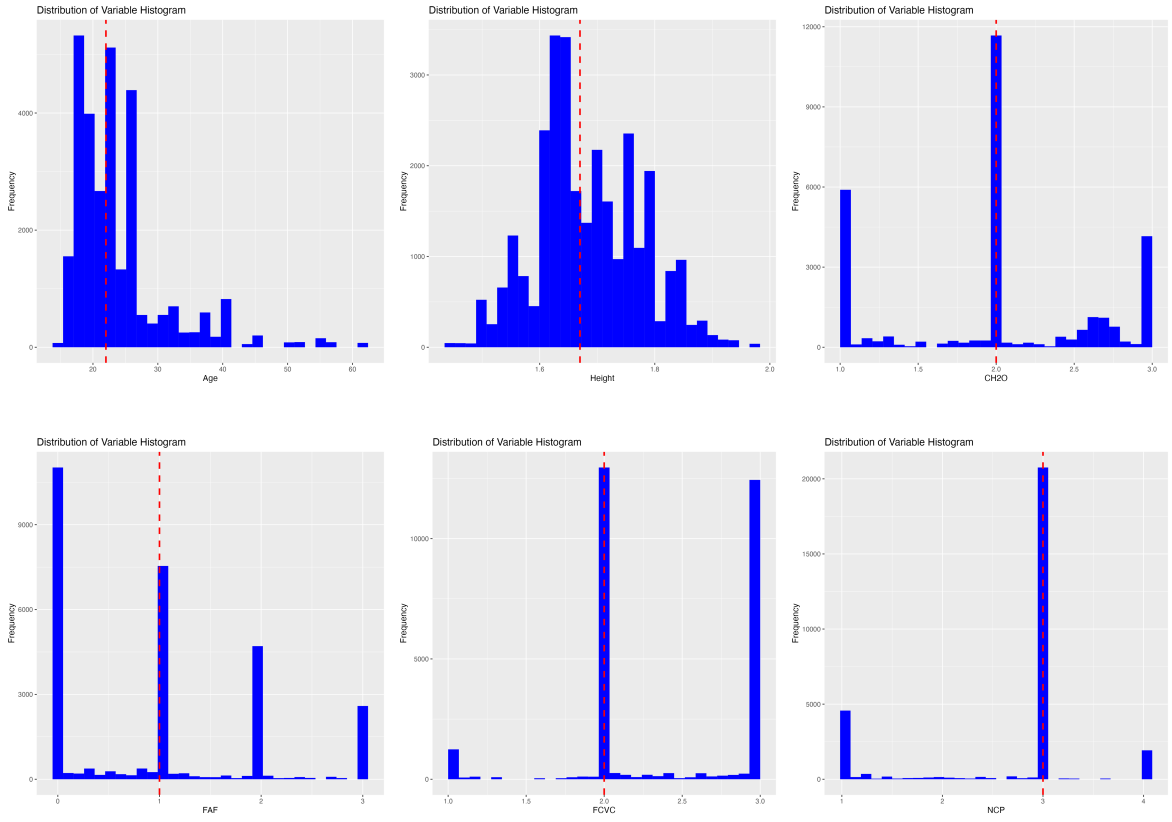
Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
LVH
Normal
ST
NA
RestingECG
Obesity Status
Not Obese
Obese

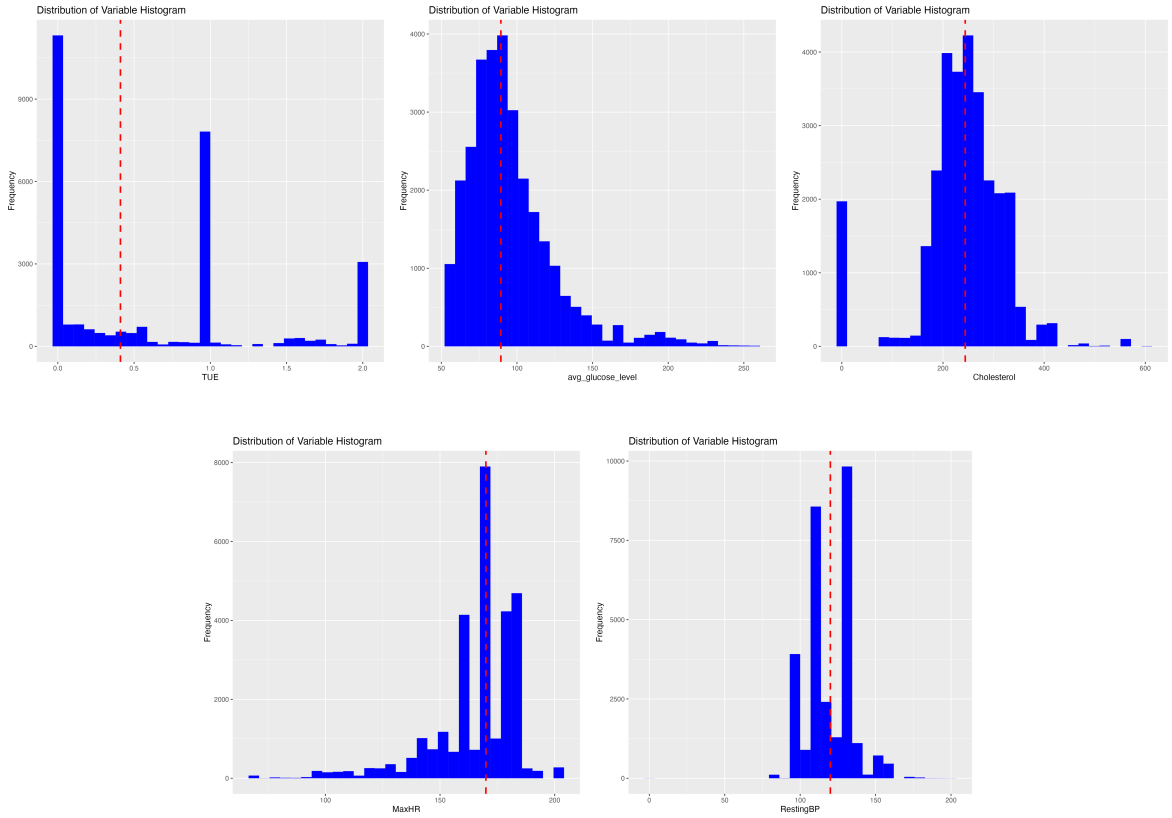

Side-by-Side Percentage Bar Plot
Percentage (%)
0
20
40
60
no
yes
NA
SMOKE
Obesity Status
Not Obese
Obese

2.4 Variable Data Cleaning

Within our training dataset, there were 74,272 missing values which required imputation in advance of modelling, since many of our candidate models cannot function on incomplete data. Helpfully, none of those missing values were in the response variable, so we did not have to remove any observations from our data. There were also 24,759 missing values in our test data, which also required imputation.

Across many of the 11 numerical variables, histograms showed us that the variables across observations did not appear to follow central distributions. Some variables, including age, average glucose level and maximum heart rate followed skewed distributions, meaning that median imputation would likely be inappropriate. Moreover, we do not know how outlier data might impact resulting predictions so would prefer not to take the risk of median imputation.





Many of our categorical variables also have more than two categories, with a lack of ordinal structure to their categorization, making modal imputation for categorical variables high-risk. Intuitively, many of the categorical variables also appear to be the those where there would be some amount of inter-variable correlation – for example, we think it is likely that someone with unhealthy eating habits may also have unhealthy drinking habits. As a result, we decided to use KNN imputation for our data, using information we had to predict other feature variables for a given observation.

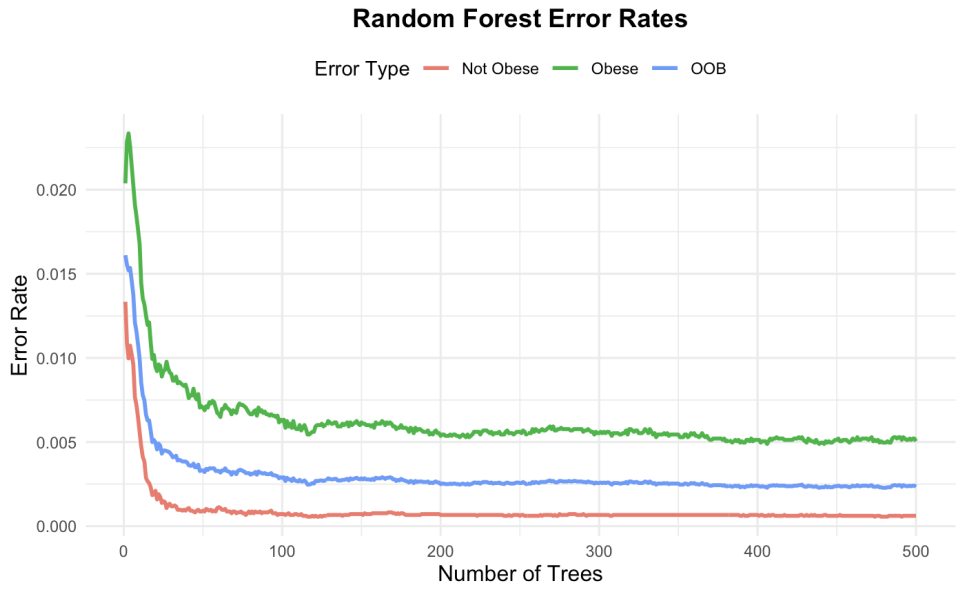
3. Methods and Models

Given our new, cleaned data sets using KNN imputation, we used the training data to develop several models using different algorithms and created predictions using the testing data. The effectiveness of these models were considered in terms of both the training and testing accuracy rates, and the complexity of the models were noted during the selection of the best

model. Algorithms considered in this study include Random Forest, Generalized Boosting, Extreme Gradient Boosting, KNN, Logistic Regression, and LDA.

3.1 Random Forest Analysis

We implemented a Random Forest model to classify obesity status in our dataset. Its nonlinear and robust architecture made it well-suited to our dataset's characteristics, such as high dimensionality, missing data, and non-normal distributions. This algorithm effectively minimized overfitting while providing valuable insights into variable importance.



We trained the model using an optimal mtry value of 5 and incrementally increased the number of trees to observe performance trends. The Out-of-Bag (OOB) error rate served as a key measure of model stability. As shown in the OOB error plot (Figure above), the overall error decreased as the number of trees increased, stabilizing around 200 trees. Beyond this point, the error plateaued, indicating diminishing returns on accuracy. The final OOB error rate was impressively low, falling under 1%, which reflects strong overall model performance. However,

this low error may indicate a bias toward the majority class, revealing a potential class imbalance in our dataset.

Further examination of class-specific OOB errors provided additional insights. The "Not Obese" class, which represents the majority, exhibited a notably lower error rate compared to the "Obese" class. Although the error for the "Obese" class decreased as more trees were added, it plateaued at a higher level, suggesting the model struggled to accurately classify the minority class. Addressing this imbalance through methods such as oversampling, undersampling, or class weighting could enhance predictive accuracy for the minority class and improve overall model performance.

3.1.1 Variable Importance

Variable importance for the Random Forest model was determined using two primary metrics: 1) Mean Decrease in Accuracy, and 2) Mean Decrease in Gini Impurity. The first evaluates the impact of permuting each variable on the model's accuracy. Higher mean decreases indicate greater dependence of the model on that variable for predictions. Predictors such as Age, Height, CH2O, and FAF consistently showed significant contributions, aligning with their biological relevance to obesity. The second metric reflects a variable's contribution to reducing impurity in decision tree splits. Higher values correspond to greater discriminatory power. Variables that reduce Gini impurity the most are critical for partitioning the data effectively. Both metrics affirmed the importance of certain predictors, particularly those indicative of physiological or behavioral patterns, such as physical activity and caloric intake. These findings suggest that these predictors not only influence obesity classification but may warrant further investigation in obesity-related research.

Our final result using the Random Forest model yielded a Kaggle score of 0.9912.

3.2 Boosted Model Analysis

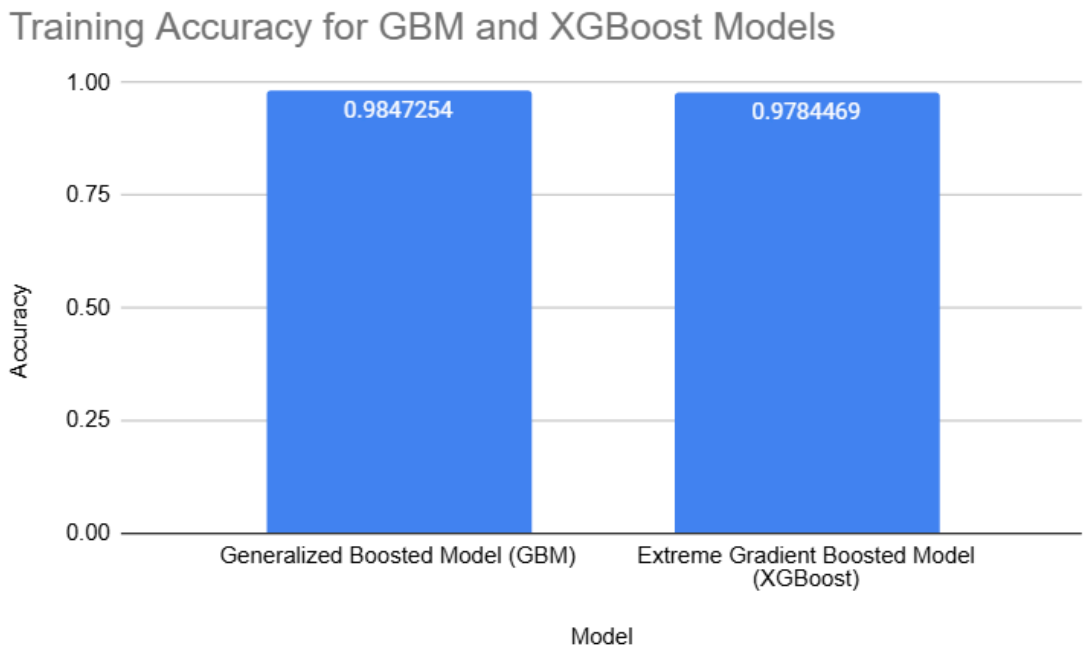
Similar to Random Forests, boosting is another ensemble method that is particularly effective in capturing complex, non-linear relationships between the predictors in a dataset. Boosted models are created by sequentially training weak decision trees and combining them into a single model. In doing so, boosted models effectively reduce bias while improving accuracy.

3.2.1 Generalized Boosted Model (GBM)

The implementation of a Generalized Boosted Model is an ensemble learning method that sequentially combines weak decision trees where each tree tries to correct the errors made by the previous ones. In order to create a GBM model, parameters were optimally selected using 5-fold cross validation: number of trees (n.trees): 500, interaction depth (interaction.depth): 4, learning rate (shrinkage): 0.1, and minimum observations in terminal node (n.minobsinnode): 5. The model was then trained using 10-fold cross validation and applied on the training dataset to achieve a prediction accuracy of 98.47%. While the high accuracy rates on the training dataset indicate a strong model performance, this could also reflect overfitting on our training data.

3.2.2 Extreme Gradient Boosted Model (XGBoost)

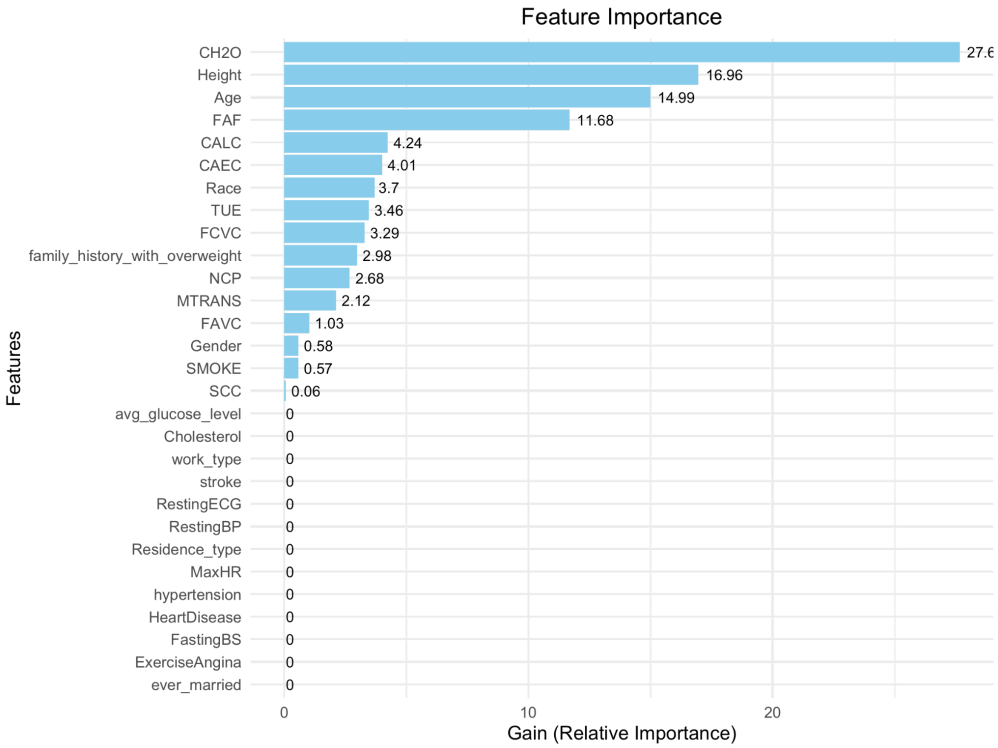
Extreme Gradient Boosted Model is an optimized implementation of gradient boosting with built in features like regularization and tree pruning which removed the need for extensive hyperparameter tuning. Like the GBM model, the XGBoost model was trained using 10-fold cross validation and applied to the training dataset achieving a prediction accuracy of 97.84%.



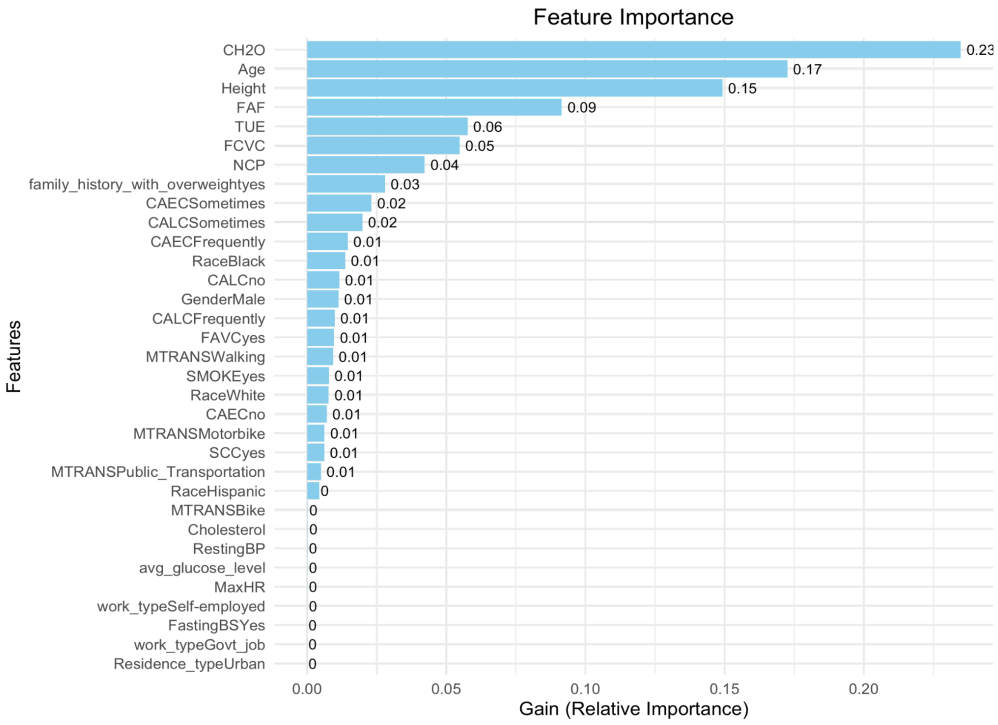
3.2.3 Relative Importance of Features

The plots show the relative importance/influence of features for the boosting models. The importance is calculated based on the number of times the variable is used to split the trees, the improvement in purity with each split, and the total gain in the model's predictive capabilities attributed to the feature. Results show that predictors such as Age, Height, FAF, and CH2O were key contributors for obesity classification in both GBM and XGBoost models. When compared to the other models we developed, it can be noted that these variables were found to be significant by the other models as well.

Relative Importance of Features for GBM



Relative Importance of Features for XGBoost



3.3 Other Models

We also tried implementing Logistic Regression, Linear Discriminant Analysis, and KNN on our data. We found that the first two methods were rather inefficient in terms of their training accuracy rates when compared to other models (0.747 and 0.746 for Logistic Regression using all and just the best 20 features respectively, and 0.743 for LDA). Despite the benefit that comes with the relative lack of complexity in using these models, because their accuracy rates were significantly lower, we did not consider these models in the final model selection. KNN had a high training accuracy rate (0.953) given the optimal k value of 3, but performed poorly with the testing data, indicating a case of overfitting. As such, we also excluded KNN from our final evaluations**.**

4. Discussion

Among all of our models, Logistic Regression and Linear Discriminant Analysis (LDA) performed the worst. Logistic Regression models assume linear relationships between the predictors and target variable, while LDA uses linear decision boundaries and assumes normality in the covariance matrices. While these two models are simple and easier to interpret, they were not well-suited for our dataset, which contains complex relationships and predictors with non-normal distributions. Reflected by their low accuracy rates, we can see that these models had limited effectiveness.

Our KNN model with k=3 might perform well on the training data because it closely fits the individual points, effectively "memorizing" the training set. However, this can lead to overfitting, making the model highly sensitive to noise, outliers, and small variations in the data. On the testing data, this lack of generalization often results in poor performance compared to models like Random Forest or boosted models, which are designed to handle noise and

generalize better. Because KNN, especially with a small k, struggles with capturing complex relationships, it is understandable why the training and testing accuracy rates had such a disparity.

The model that performed the best was our Random Forest model. Random Forests are effective for large, non-normal, non-linear datasets with outliers or missing values. As such, these models are resistant to overtraining and are well suited for complex datasets. However, these models are computationally expensive and often difficult to interpret.

Lastly, the most complex model out of the three, boosted models, are effective against imbalanced, non-normal, non-linear datasets. Unlike Random Forests which fit separate and independent decision trees to multiple copies of the data to create a model, boosted models have a slower learning method. Each decision tree is fit sequentially: a new tree is grown from the previously modified dataset of the original. This makes boosted models like GBM and XGBoost models the most computationally expensive and more prone to overfitting than Random Forests.

When choosing our final model, we wanted to strike the perfect balance between simplicity, accuracy, and variable selection. Considering all of this, we chose our final model to be the Random Forest model using mtry = 5.

5. Limitations

While the Random Forest model provided exceptional accuracy rates and robustness, several limitations in our methodology and dataset must be acknowledged.

5.1 Class Imbalance

A key limitation of the Random Forest model is its sensitivity to class imbalances. In our dataset, the "Not Obese" class was the majority, leading to skewed results where the model

achieved lower accuracy for the "Obese" class. Additional balancing techniques like oversampling the minority class, undersampling the majority class, or incorporating class weights during training can help in reducing these skewed results.

5.2 Interpretability Challenges:

Despite its high predictive power, the Random Forest model operates as a "black box," where decision-making processes are not easily interpretable. While variable importance plots provide some insight into key predictors, they do not fully explain the causal relationships between variables. This ambiguity limits the model's applications where interpretability is needed.

5.3 Bias from Imputed Data:

Missing values in our dataset were imputed using KNN imputation, introducing potential biases. This approach estimates missing values based on observed data patterns, which might not accurately reflect real-world variability. Although this method improved model performance, it might have skewed the distribution of key predictors, thus impacting the generalizability of our results.

5.4 Computational Complexity:

The Random Forest model, particularly with an increased number of trees and predictors, required significant computational resources. With this complexity, the issue of scalability for larger datasets or real-time applications arises.

5.5 OOB Error and Model Generalization:

While the OOB error rate provided a reliable estimate of model accuracy, its low value

(<1%) raises concerns about overfitting the training data. Potential future iterations could incorporate additional validation techniques or external test sets to better assess generalizability.

Acknowledging these limitations is vital for interpreting the results of our study and for guiding future improvements in the predictive modeling of obesity status.

6. Conclusion

In this study, we aimed to classify obesity status by leveraging various statistical methods. Through extensive exploratory data analysis and preprocessing, we addressed issues such as missing values and variable distributions to prepare our dataset for modeling. While simpler models like logistic regression and linear discriminant analysis provided valuable benchmarks for comparisons, they fell short of capturing the complexity of our data. The Random Forest model is the most effective approach, offering the best accuracy and robustness, specifically for high-dimensional, non-linear data with NA values. In optimizing hyperparameters such as mtry and thoroughly analyzing the OOB error rate, we found the optimal count to be 200 trees, balancing performance and computational efficiency.

Furthermore, variable importance measures highlighted key predictors such as age, height, daily water intake, and frequency of physical activity, underscoring their relevance in obesity classification. Although several of the models indicated the significance of said variables, more research is needed to determine the relevance of these variables in comparison with typical indicators of obesity such as cholesterol levels, glucose levels, or blood pressure.

It is important to acknowledge the limitations such as class imbalance sensitivity, interpretability challenges, and computational demands. These issues raise the need for further

improvements to mitigate class imbalance and explore modified approaches that improve the balance of accuracy and interpretability. Ultimately, our findings contribute to supporting the use of predictive models for public health and medical research. By identifying critical predictors and refining modeling techniques, future studies can enhance the accuracy and applicability of predictive models for obesity classification.

7. Acknowledgement

We would like to thank Professor Akram Mousa Almohalwas for his guidance and instruction throughout the duration of this quarter.

References

1. World Obesity Federation. (n.d.). Prevalence of obesity. World Obesity. Retrieved December 15, 2024, from https://www.worldobesity.org/about/about-obesity/prevalence-of-obesity
2. National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). *Health risks of being overweight*. U.S. Department of Health and Human Services. Retrieved December 16, 2024, from https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/health-risks#:~:text=Having%20overweight%20or%20obesity%20increases,the%20cells%20in%20your%20body.
3. Allison, D. B., Fontaine, K. R., Manson, J. E., Stevens, J., & VanItallie, T. B. (1999). Annual deaths attributable to obesity in the United States. *Journal of the American Medical Association, 282*(16), 1530-1538. Retrieved December 16, 2024, from https://pmc.ncbi.nlm.nih.gov/articles/PMC99423/