Obesity Prediction Model for Individuals Residing in Colombia, Peru, and Mexico

Submitted by:

Laura Muñoz Mendoza

Santiago Arévalo

Submitted to:

Dr. Carlos Contreras

For Partial Fulfillment of the Requirement of

MATH 509

University of Alberta

Winter 2025

https://github.com/santi-proy/m509.git

Abstract

Objectives: The objective of this study is to develop a model capable of classifying an individual into a weight category given a set of basic sociodemographic and lifestyle attributes.

Methods: This secondary investigation is based on data collected from an online survey available to individuals between 14 and 61 years of age residing in Mexico, Colombia, and Peru. The survey generated a sample of 498 observations and 17 features. Two rebalancing procedures known as synthetic minority oversampling technique (SMOTE) and class weights (CW) were first used to preprocess the data. Then the data was split into training and testing sets, and a random forest analysis was performed to evaluate the classification of new observations using each rebalancing method. Lastly, the best performing model was chosen and tested on a new prediction.

Results: The random forest model achieved 92% accuracy with SMOTE rebalancing, highlighting weight, height, and age as key predictors, followed by physical activity and dietary habits.

Conclusion: Future studies should account for other factors including (but not limited to): body fat percentage, income, education, proximity to local markets, and the quality and type of ingredients or produce available to individuals in their respective regions. It is also important to know what kind of access, if any, individuals have to medical professionals such as nutritionists and dieticians, and whether special weight management programs exist in the regions of interest.

Keywords: Obesity, weight, random forest, SMOTE, class weights, classification

1. Introduction

Obesity is now considered to be a global endemic by the World Health Organization (WHO) resulting from sharp increases in obesity worldwide coupled with a lack of effective programming to mitigate this upward trend (Lingvay, Cohen, le Roux, & Sumithran, 2024, p. 1). Interestingly, between 1985 to 2017, this increase was most detected in rural areas (Lingvay et al., 2024, p. 1). It may be that the lack of dietary options and limited health care services in certain rural areas may create the perfect conditions for obesity to affect a greater proportion of individuals than those in urban centres.

Although this disease is broadly characterized as the storage of excess fat, the WHO also highlights that its complexity and chronic nature lead to adverse health effects (Lingvay et al., 2024, p. 1). The use of the terms *epidemic, disease*, and *chronic* are significant since they suggest that obesity is not an individual problem but one that requires a concerted effort from regional and local medical experts. In other words, while mainstream perceptions of obesity generally tend to shame the individual, the standpoint of the WHO suggests that obesity is a multifaceted disease that involves a more collaborative approach between patients and medical practitioners. This is supported by the fact that although bariatric surgery (the leading treatment) and some pharmacological agents have been developed, their use has been underutilized (Lingvay et al., 2024, p. 1). There is no definitive research to indicate why this is the case, but perhaps fear of stigmatization, cost, family commitments, or simply being unaware of these treatment options may be some of the reasons why they are not being accessed in greater numbers. Obesity deserves greater attention, and we lend our expertise in data analytics to study it further.

The central problem of this project is to determine if a reliable obesity prediction model, with an emphasis on accuracy, can be formulated from the data to aid medical professionals in potentially diagnosing the onset of obesity for the population of interest. The main questions of interest are framed from the perspective of medical practitioners to guide the development of the model and ensure that it can be informative to both patients and their treating physicians. These questions include (but are not limited to):

1. What are the key features in classifying an individual into a particular weight category?
2. Do rates of obesity differ between females and males?
3. Are there any unusual trends/anomalies in the data that should be explored further?

The dataset was obtained from Kaggle, and the observations were collected via an online survey encompassing residents in Mexico, Colombia, and Peru. A total of 498 observations and 17 features were collected. With this information in hand, the first step was to perform an exploratory data analysis of the data to obtain an overview of how the features were distributed and to identify any potential issues prior to formulating any model.

1.1 Exploratory Descriptive Analysis

1.1.1. Variable Description and Encoding

Table 1 describes the variables and their encoding, and the abbreviations *Cat* and *Num* refer to *categorical* and *numerical* respectively.

Table 1: Description of Features

Variable	Description/Question	Type	Encoding
gend	Gender	Cat	0: male, 1: female
age	Age	Num	Years
height	Height	Num	Metres
weight	Weight	Num	Kilograms
famhist	Has a family member suffered or suffers from being overweight?	Cat	0: No, 1: Yes
calH	Do you eat high caloric food frequently?	Cat	0: No, 1: Yes
veg	Do you usually eat vegetables in your meals?	Cat	0: Never, 1: Sometimes, 2: Always
meals	How many main meals do you have daily?	Cat	1, 2, 3, More than 3
snacks	Do you eat any food between meals?	Cat	0: No, 1: Sometimes 2: Frequently, 3: Always
smoke	Do you smoke?	Cat	0: No, 1: Yes
water	How much water do you drink daily?	Cat	1: Less than 1 litre, 2: 1-2 litres, 3: More than 2 litres
calM	Do you monitor the calories you eat daily?	Cat	0: No, 1: Yes
phys	How often do you have physical activity?	Cat	0: 0 days, 1: 1-2 days 2: 2-4 days, 3: 4-5 days
tech	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	Cat	0: 0-2 hours, 1: 3-5 hours, 2: More than 5 hours
alc	How often do you drink alcohol?	Cat	0: No, 1: Sometimes 2: Frequently, 3: Always
trans	Which transportation do you usually use?	Cat	1: Automobile, 2: Motorbike, 3: Bike, 4: Public Transportation, 5: Walking
obese	Obesity level (dependent variable)	Cat	1: Underweight, 2: Normal, 3: Overweight I, 4: Overweight II, 5: Obesity I, 6: Obesity II, 7: Obesity III

As Table 1 conveys, most of the features are categorical in nature and the response variable obese has seven levels of classifications. This immediately suggested that the formulation of a model should account for multi-class categorization where the majority of variables are categorical.

1.1.2. Distribution of Variables

The distribution of *obese* was then generated to gain a sense of how the responses compared according to each level. Figure 1 depicts these results.

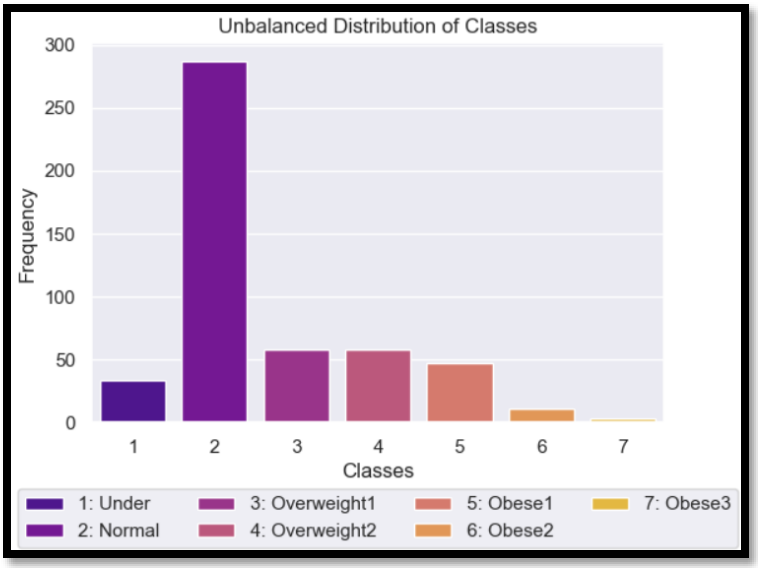


Figure 1: Unbalanced Distribution of Weight Classes

The classes of the response variable show a clear imbalance in the distribution, which creates challenges for building reliable predictive models. Most of the data is made up of *normal* weight individuals, while the obesity categories are underrepresented. This imbalance can lead to biased models, as standard statistical methods tend to favor the majority class, potentially missing important patterns in the less common obesity categories. To overcome this, we needed to use specialized techniques such as resampling methods (i.e. SMOTE) and class-specific weights. This will be further analyzed in the next sections.

The distribution of the features was also studied to see what anomalies may have present in the data. Table 2 provides a numerical summary of the three continuous variables, and Figure 2 depicts their distributions.

Table 2: Summary Statistics for Age, Height, and Weight

	Age (Years)	Height (m)	Weight (kg)
Mean	23.5	1.69	74.1
Median	21	1.68	67
Min	14	1.45	39
Max	61	1.98	173

A few notable observations stand out from these summaries. Although the age range was between 14 and 61, the average age of 23.5 indicated that most of the respondents were relatively young. This suggested two notions regarding age. First, the age of the participants was right-skewed, and second, the young mean age suggested that most of the participants may have been in good general health at the time of the survey. The numerical summaries of *weight* could be interpreted in a similar manner, but those of *height* hinted at a more balanced distribution.

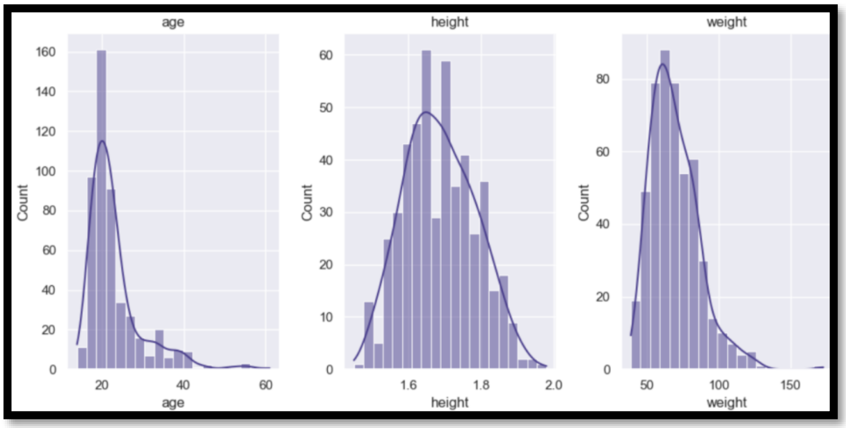


Figure 2: Histograms of the Numerical Features

The histograms show the distribution of the 3 numerical variables *age*, *height*, and *weight*., and confirmed our interpretation of their numerical summaries. *Age* is right-skewed, with most individuals concentrated around 18-25 years. *Height* follows an approximately normal distribution, centered around 1.7 meters. *Weight* is also right-skewed with most values clustered between 50 and 80 kg.

For clarity, only the distribution of *gend*, *smoke,* and *famhist* against *obese* are shown in Figure 3. Similar plots for the remaining figures can be found in *Appendix A.*

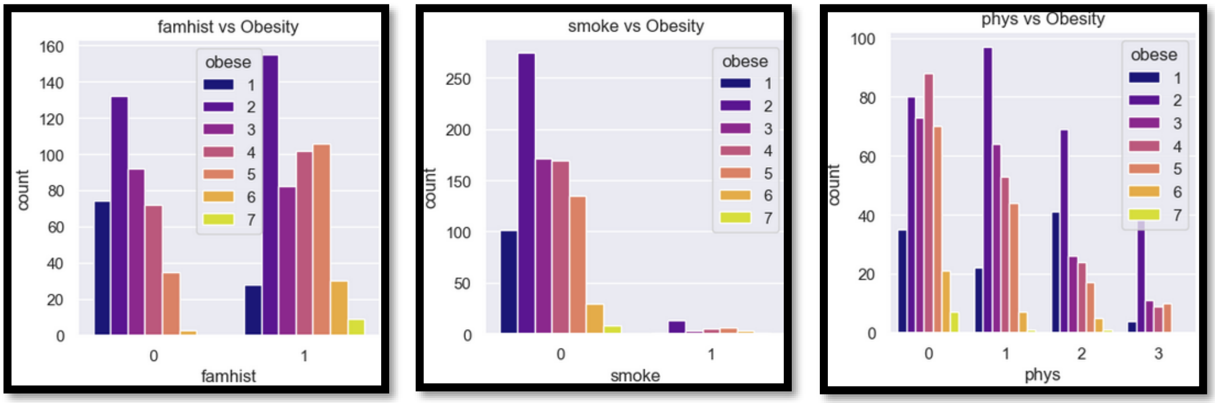


Figure 3: Distributions of famhist, smoke, and phys Across Obesity Levels

The graphs in Figure 3 along with those in *Appendix A* illustrate a relatively consistent pattern across most variables: the predominance of level 2 (Normal weight) respondents, confirming the unbalanced nature of the dataset. Some notable observations include:

- Family history appears to have a potential relationship with obesity, with those having family history showing higher counts in obesity categories.
- Smoking shows most respondents as non-smokers, making it difficult to draw conclusions about its relationship with obesity.
- Physical activity suggests a potential inverse relationship with obesity levels indicating that exercise may play a key role in reducing the incidence of obesity.

The unbalanced dataset with most *normal* weight respondents limits the statistical power for analyzing higher obesity levels (i.e. classes 5-7). Some categories have very small sample sizes for certain obesity levels, increasing the risk of drawing incorrect conclusions.

2. Formulation of Mathematical Algorithm

2.1. Assumptions and Restrictions

Since the data was collected from 3 cities in 3 different countries, regional and cultural differences may have affected some of the responses. To simplify the model-building process in such a way that captures the commonality among the survey respondents, the following assumptions were made:

- Survey respondents had access to the same quantity and type of food options from local stores, markets, and restaurants.
- Survey respondents had access to similar levels of medical care.

The overarching restriction was that the analysis and results contained herein were restricted to the populations of the three participating cities and thus cannot be used to make inferences regarding any other populations. This study was also restricted by the available features. Although many of the variables were important, other sociodemographic variables including income and education level would have also been valuable.

2.2. Variables and Parameters

The features used in this investigation are summarized in Table 1 and were excluded here to avoid duplication. In the context of random forests, a set of hyperparameters were adjusted to control the overall functioning and structure of the random forest model. The hyperparameters are summarized in Table 3.

Table 3: Hyperparameters

Hyperparameter	Description
n_estimators	Number of decision trees in the random forest model.
max_depth	Maximum depth of each decision tree.
max_features	Maximum number of features considered at each split.

A few additional hyperparameters were available, but those described in Table 3 were the most fundamental in the model-building process since they controlled key aspects of the random forest classifier.

Often a mathematical model can be generated once the variables and parameters are identified as is the case with linear, logit, and probit models. The structure and processes of random forests, however, do not lend themselves to this type of compact representation. Thus, a different mathematical form was needed to convey how the assumptions, variables, and hyperparameters generate a classification result.

2.3. Justification of the Model

As mentioned earlier, this problem is unique in that the response variable consists of 7 classifications and the data is unbalanced. Standard approaches like multiple linear regression and logistic regression aren't suitable since the target variable is neither continuous nor binary. While it’s possible to group the minority categories together to make logistic regression work, doing so would lead to a significant loss of information. To overcome these limitations, this study employed a random forest methodology, which is better suited to this type of problem. While other methods could have been considered, and classification trees do have their own limitations, a careful review of the assumptions required for alternative models made classification trees the more flexible choice for this analysis.

Table 4: Classification Algorithm (b = 1 to B)

Step	Description
1. Sample	1. Draw a bootstrap sample of size *N* from the training data.
2. Iteration	2. Repeat the following steps recursively for each terminal node until a minimum node size is achieved: i) Randomly select *m* variables from the available *p* variables

	ii) Choose optimal variable among m iii) Split node into 2 nodes
3. Ensemble	3. Generate ensemble of decision trees generated: $\{T_b\}_1^B$
4. Prediction	4. Make a prediction at $x*$: i) Let $\hat{C}_b(x*)$ represent the predicted class of the b-th decision tree ii) $\hat{C}_{RF}^B(x*)$ = majority vote $\{\hat{C}_b(x*)\}_1^B$
Note: B = number of decision trees (See Table 3)	

Since a concise formulaic representation of random forests was not possible, we opted to represent the decision-making process of our model via a mathematical algorithm. While this lacks the visual appeal of a formula, it still provides a concise representation of how random forests perform classification. A general form of the algorithm is summarized in Table 4 (Hastie, Tibshirani, & Friedman, 2017, p. 588).

The fundamental principle of the algorithm is to derive several decision trees that will ultimately comprise the random forest classifier. For a new observation, $x*$, each decision tree generates a vote for a class. The class that receives the largest proportion of votes becomes the class for $x*$ and is denoted by $\hat{C}_{RF}^B(x*)$. The relationship between correlation and variance is also an important consideration. To reduce the correlation between the decision trees whilst keeping the increase in variance to a minimal, each decision node is limited to a set of $m \leq p$ features (Hastie et al., 2017, p. 589). Consequently, this prohibits decision nodes from relying on the same variables for each split. In other words, the decision node must rely on a subset of features thereby limiting the influence of a few significant features.

In short, Table 4 highlights that a random forest is an ensemble learning method that builds multiple decision trees to make predictions by randomly selecting subsets of the data and features for each tree. Each tree in the forest makes a prediction, and the final prediction is determined by taking a majority vote. This approach accounts for accuracy, reduces overfitting, and handles complex and non-linear relationships between features making it effective for predicting obesity levels based on the features available.

In addition, several options were available for mitigating the unbalanced data, but the appropriate choice of methodology needed to reflect the (i) nature of the data set and the questions of interest and; (ii) be grounded in appropriate qualitative and quantitative reasoning. One method suggested reducing observations from the majority class while another merely created duplicates of the existing minority classes. Both approaches were discarded since the former would have eliminated information while the latter would have introduced a greater degree of bias. The two prominent methods considered were class-weighting (CW) and synthetic minority oversampling technique (SMOTE).

- **Synthetic Minority Over-sampling Technique (SMOTE):** this technic generates synthetic samples for the minority class. Instead of simply duplicating minority class instances, SMOTE creates new data points by selecting existing minority instances, finding their nearest neighbors, and then generating new examples along the line segments connecting the original points and their neighbors. This helps balance the class distribution, allowing machine learning models to better learn about of the minority class. This method creates a new dataset.
- **Class Weights (CW):** this method adjusts the importance (weight) given to each class during model training. By assigning higher weights to the minority class, the model is penalized more for misclassifying these classes, effectively making them more influential in the learning process. This encourages the model to pay more attention to the underrepresented class, leading to better performance. This method keeps the original dataset as is.

Sections 1 and 2 provided us with three key insights that would be necessary in model-building phase. First, a better understanding of the features and their impact on the response variable was obtained via numerical and graphical representations. Next, a random forest model was proposed as a suitable framework for building a classification model. Lastly, two procedures were selected for attempting to rebalance the data. The subsequent step was to combine these results and use them to formulate a solution to our problem.

3. Solution of the Problem

As a starting point we conducted a series of data cleaning and pre-processing which involved among other things checking for missing values and/or extreme outliers, and encoding the features appropriately. The first step in the actual model-building process, however, was to address the issue of the unbalanced data via appropriate testing. Figure 4 provides an overview of the plan we implemented to ensure that our model was being developed according to appropriate statistical methodologies and sound diagnostic testing.

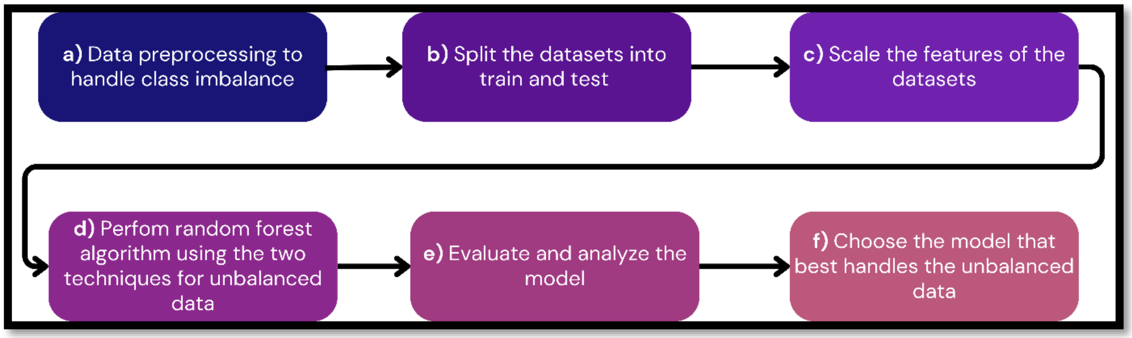

a) Data preprocessing to handle class imbalance
b) Split the datasets into train and test
c) Scale the features of the datasets
d) Perfom random forest algorithm using the two techniques for unbalanced data
e) Evaluate and analyze the model
f) Choose the model that best handles the unbalanced data

Figure 4: Model-Building Plan

The plan can be summarized as follows:

a) Data preprocessing to handle class imbalance (SMOTE and CW)

a) Split the data into training (80%) and testing (20%) sets.

b) Scale the features for consistency.

c) Run the random forest algorithm separately for SMOTE and CW.

d) Evaluate the resultant models using appropriate testing.

e) Select the model that performs best across the tests.

In the next section we carry out this process and provide an analysis of the results with the selected model.

4. Interpretation of Results

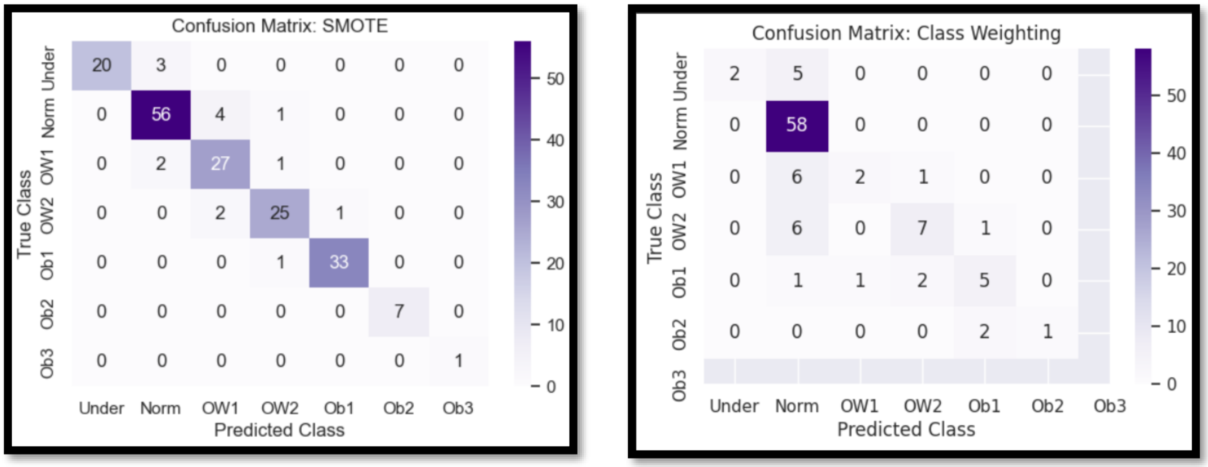
4.1. SMOTE v. CW

The results of the testing performed using both SMOTE and CW are found in Table 5 and Figure 5.

Table 5: Hyperparameter Results (SMOTE v. CW)

Hyperparameter	SMOTE	CW
n_estimators	300	1000
max_depth	12	15
max_features	$\sqrt{p}$	$\sqrt{p}$

Table 5 summarizes the results of a tuning procedure which determined the ideal hyperparameter values for the respective methods. It suggests that the hyperparameter configurations for SMOTE are superior to those of CW since less decision trees (300) and a lower decision tree depth (12) are required for optimal performance. Unlike the other two hyperparameters, however, the maximum value of $\sqrt{p}$ features were explicitly chosen for each decision node since this recommended value has a decorrelating effect leading to greater reliability and a reduction in the average variance of the decision trees (James, Witten, Hastie, & Tibshirani, 2023, p.344).



Confusion Matrix: SMOTE

True Class \ Predicted Class	Under	Norm	OW1	OW2	Ob1	Ob2	Ob3
Under	20	3	0	0	0	0	0
Norm	0	56	4	1	0	0	0
OW1	0	2	27	1	0	0	0
OW2	0	0	2	25	1	0	0
Ob1	0	0	0	1	33	0	0
Ob2	0	0	0	0	0	7	0
Ob3	0	0	0	0	0	0	1

Confusion Matrix: Class Weighting

True Class \ Predicted Class	Under	Norm	OW1	OW2	Ob1	Ob2	Ob3
Under	2	5	0	0	0	0	
Norm	0	58	0	0	0	0	
OW1	0	6	2	1	0	0	
OW2	0	6	0	7	1	0	
Ob1	0	1	1	2	5	0	
Ob2	0	0	0	0	2	1	
Ob3							

Accuracy = 0.92 Accuracy = 0.75

Figure 5: Confusion Matrix Results (SMOTE v. CW)

It should be noted that the number of observations tested in the respective confusion matrices are different since SMOTE generated new observations for testing while CW simply weighted the existing observations. When comparing the predicted versus the true classifications, SMOTE once again performed better with an accuracy of approximately 92% compared to an accuracy of 75% using CW. Thus, SMOTE was selected as the rebalancing procedure and its hyperparameter values were used to set up the random forest classification model. The resultant balanced data is shown in Figure 6.

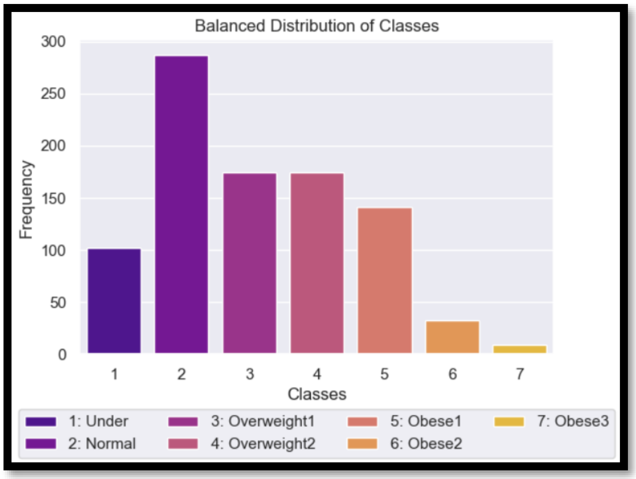


Figure 6: Balanced Distribution of Weight Classes

A threefold increase was applied to each minority class and although this scaling factor seems arbitrary, it generated a reasonable and interpretable distribution of the weight classes. For example, Figure 6 suggests that if an individual did not fit into the *Normal* weight class, the next likely classification could be *Underweight* or *Overweight1*. In the original study by Fabio Mendoza Palechor and Alexis de la Hoz Manotas, the researchers applied a balancing procedure that generated nearly identical counts for each class. Although their approach certainly balanced the data, it produced a distribution whereby each classification was equally probable. This was not realistic and thus we aimed to preserve the proportions of the original sample by applying a scaling factor that was more reasonable. In fact, the total cases of obesity (classes 5 to 7) comprise approximately 20% of the observations shown in Figure 6. This, in fact, is in line with estimates from the World Health Organization that indicate that obesity is prevalent within a range of 20 to 35% across the three countries (World Health Organization [WHO], 2025). In the final phase we tested the model and interpreted its output in the context of our original problem and the questions of interest.

4.2. Random Forest Classification

The central problem was whether a reliable and accurate obesity prediction model could be formulated from the data set. The results up to this point were promising and suggested that a random forest classifier coupled with SMOTE rebalancing may be of benefit to medical professionals and patients in these regions. Thus, we tuned our model to the optimal hyperparameter values based on the SMOTE testing results and proceeded with answering the supporting questions.

First, we were interested in determining which features were most influential in predicting obesity outcomes. There are several ways to evaluate and compare feature performance, and Figure 7 illustrates the first approach called a Feature Importance Chart.

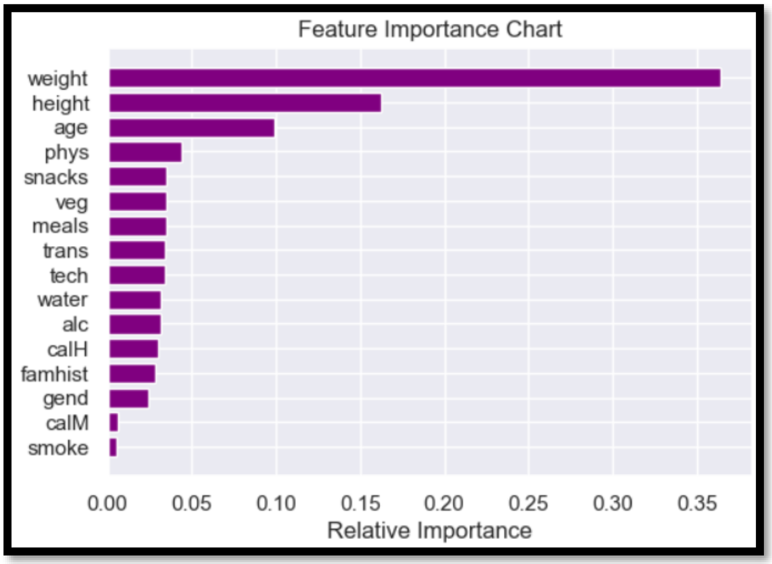


Figure 7: Feature Importance Chart

Figure 7 highlights what was expected: *weight*, *height*, and *age* are leading contributors to the prediction model. However, the chart reveals another important aspect of weight management. Following the first three features, the next features are *phys*, *snacks*, *veg*, and *meals* which when taken together highlight the role that both physical activity and eating habits play in classifying a patient into a weight class. Another common measure is SHAP values, and these can be seen in Figure 8.

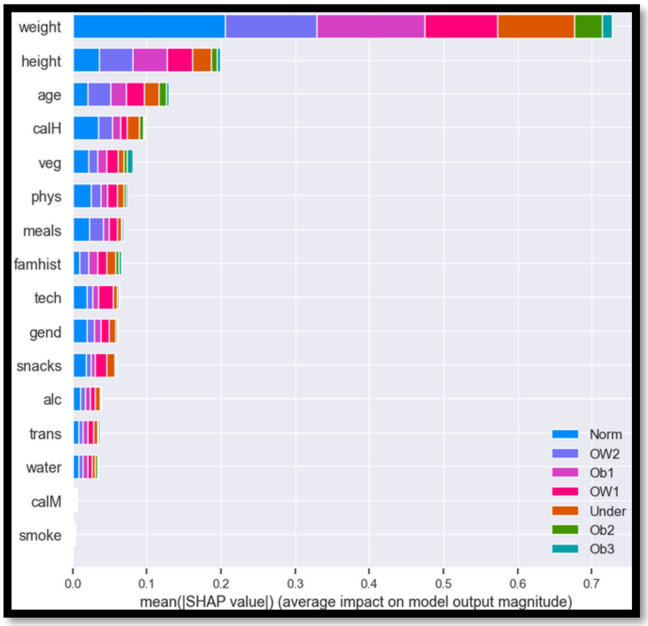


Figure 8: SHAP Values

Ideally SHAP values indicating whether a feature has a positive or negative effect would be obtained, but it was not possible to do so in this case. However, Figure 8 is still useful for two reasons. First, the colour encoding shows how much each feature contributes to each weight class. For example, the influence of weight on the *normal* group as is evidenced by the relatively large light blue bar in the first row. Similar interpretations hold for the other variables and weight categories. A second useful result from Figure 8 is that the order of importance is very similar to that of the feature importance chart found in Figure 7. One interesting takeaway is that *calM* – whether an individual monitors their calorie intake or not – has moved up in the list while *snacks* has fallen. At first glance this may seem like a contradiction to the results of Figure 7; however, calorie monitoring may also form part of a healthy dietary regime. In other words, it actually confirms the previous results and once again highlights the importance of healthy eating habits.

Figures 7 and 8 certainly provided us with a sense of which features were important, but they failed to give us an indication of how a classification could change as the value of a feature was adjusted. To gain this insight we generated decision boundaries. Figure 9 shows one such decision boundary for *weight* and *height*.

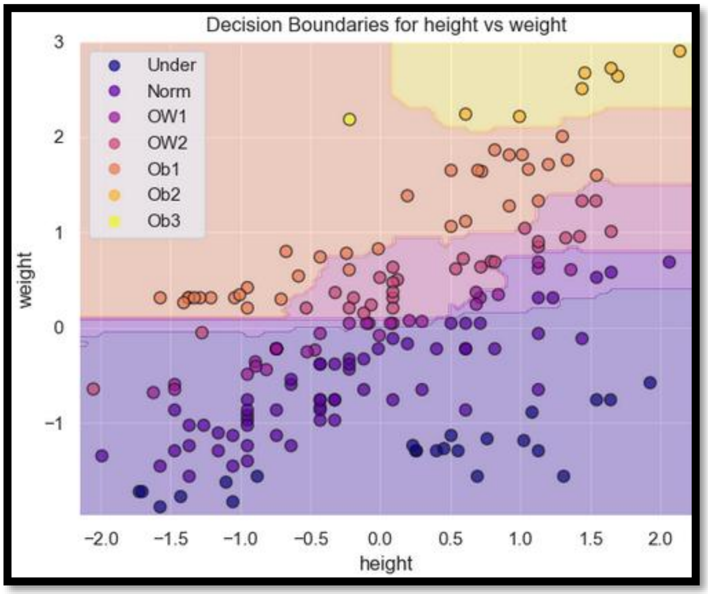


Figure 9: Decision Boundaries (height v. weight)

Figure 9 demonstrates how the classification changes as *height* and *weight* change. For example, a standardized height of 2.0 and standardized weight of 2.5 could potentially be classified as belonging to one of the three obesity categories. This, however, is just one decision boundary and others are possible. Although such a figure provides a numerical intuition of what is occurring, it does have drawbacks which we will return to.

It is tempting to assume that the random forest classifier only uses *weight* and *height* to make predictions since their feature importance and SHAP values are quite large. If so, our model would be biased and unhelpful. Thus, a better understanding of how a classification is made is warranted. As a starting point, let us consider part of a decision tree (See Figure 10).

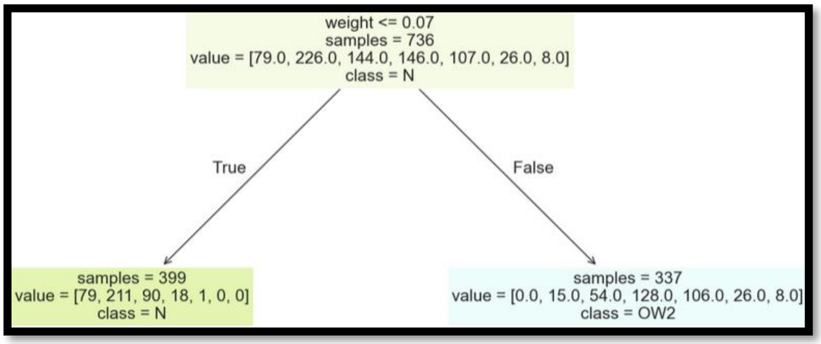

weight <= 0.07
samples = 736
value = [79.0, 226.0, 144.0, 146.0, 107.0, 26.0, 8.0]
class = N
True
False
samples = 399
value = [79, 211, 90, 18, 1, 0, 0]
class = N
samples = 337
value = [0.0, 15.0, 54.0, 128.0, 106.0, 26.0, 8.0]
class = OW2

Figure 10: Partial Decision Tree

In this case, the decision tree randomly started off by making a decision based on *weight* $\leq$ *0.07* scaled units. The key to ensuring that *weight* is not solely used through the decision tree goes back to our original condition that the maximum number of features at each decision/internal node is $m = \sqrt{p}$. This means that a randomly drawn subset of *m* features from the *p* available will be provided to each decision/internal node as the decision tree is built. In short, this prevents the decision tree from always using *weight* to split a node. This has the effect of ensuring that the trees are decorrelated and that all the features are randomly considered at different stages in the random forest (Hastie et al., 2017, p. 589).

The random forest classification model generated in this study could be an informative diagnostic tool to help medical practitioners treat patients who are or are on the verge of developing obesity. However, our model is inherently flawed in other areas. In the final section we critique our model and highlight both its usefulness as well as its shortcomings. We will also use the questions posed at the beginning of the study to aid in the development of this discussion.

5. Critique of the Model

One of our first questions was to explore what features were important in the classification process, and it was certainly no surprise to see *weight* and *height* at the top of the list. Moreover, the decision boundaries confirmed that a classification of obesity was more likely when an individual was taller and heavier. However, let us consider a new observation and use our model to make a prediction (See Table 6).

Table 6: New Observation and Classification

Characteristics	Classification
age=34, height=1.55m, weight=75kg, male, *no family history of obesity, does not consume a high-calorie diet,* *sometimes eats vegetables, eats 2 meals/day,* *does not snack between meals, nonsmoker,* *consumes 1-2 L of water/day, monitors calories,* *engages in physical activity 4-5 days/week,* *uses tech 3-5 hours/day, nondrinker, uses an automobile daily*	Obese 1

Our model predicted that an individual with these traits could be classified as someone who is in the first stage of obesity. However, this individual possesses many healthy traits including a low-calorie diet, possibly some fasting (2 meals/day), and a physically active lifestyle. The problem is that this individual has a weight that is close to the mean, but he is significantly shorter than the mean height. Thus, the model perceives this individual to be too heavy given his height. This highlights two problems with our model. One, many of the healthy habits seem to be largely ignored, and two, the weight of the individual could actually come from muscle instead of visceral fat. Our model, however, does not have an input for body fat percentage, so it cannot view the concept of weight from alternative perspectives. This is a significant shortcoming since it could mistakenly classify an otherwise healthy, active, and muscular individual as being obese even when there could be evidence to the contrary.

Our last two questions focused on whether differences in obesity rates existed amongst women and men, and if any noteworthy trends/anomalies could be detected. The first question is particularly important since women's health is generally underrepresented in mainstream health discourses. That is, other health considerations regarding women (i.e. hormonal changes) can be often overlooked when treating female patients. The distribution of the weight classes amongst men (0) and women (1) in our model is shown in Figure 11.

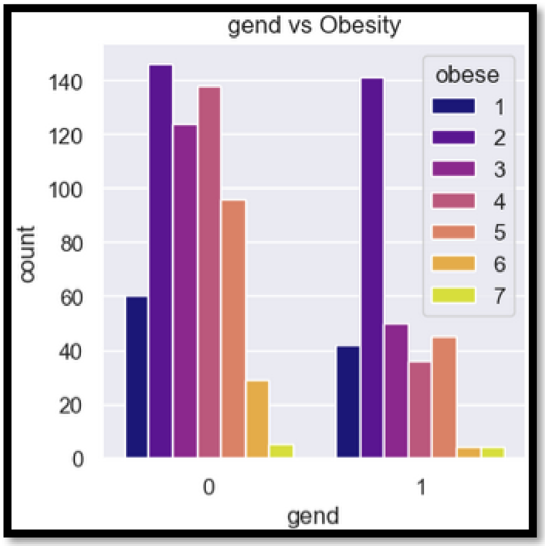


Figure 11: Distribution of gend v. obese

Figure 11 certainly suggests that incidence of obesity may be higher in males than in females. Although we cannot infer that this is the case, our model does seem to indicate that a difference between men and women may be present signalling that it may be necessary to design unique obesity treatment plans for the two sexes. However, the use of the terms *gender* and *sex* should not be viewed as interchangeable, and thus adds another layer of difficulty when it comes to drawing inferences from our model. For example, *gender* is often described as fluid meaning that an individual could be biologically male, but their identity – including their mannerisms, habits, and relationships – could mimic those of the opposite sex. Moreover, some individuals may even take hormone supplements as a means of further identifying with another sex. In this context, such information is invaluable because it would mean that a male could choose to following the dietary habits and physical activity regime of women as a means of identifying with that sex. Unfortunately, *gend* does not capture this information nor does any other feature.

Lastly, we aimed to determine if the data set exhibited trends/anomalies that warranted further investigation. Throughout the exploratory data analysis and model-building phases there was nothing that immediately stood out in this regard. However, a follow-up test on the significance of the features via a permutation importance method provided another perspective on the validity of our model (See Figure 12).

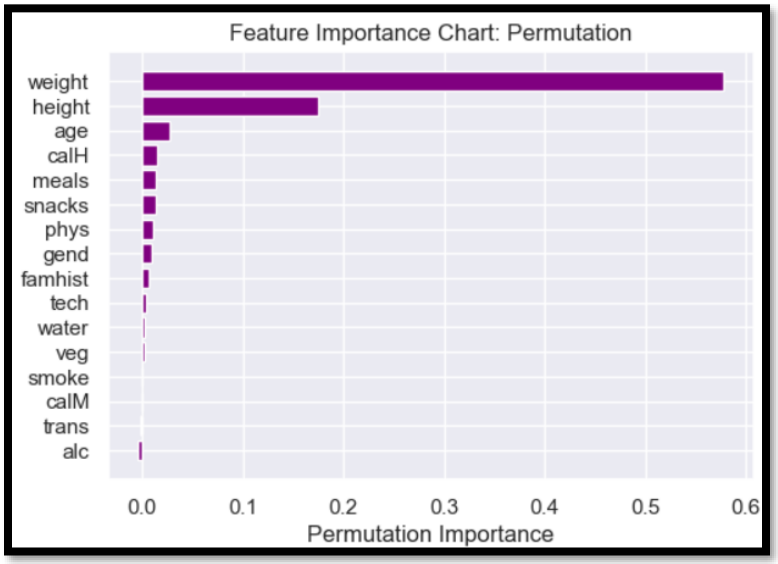


Figure 12: Permutation Importance Chart

The permutation importance graph in Figure 12 looks similar to the feature importance graph in Figure 7 in that *weight* and *height* continue to be the dominant features. A noticeable distinction between the two tests, however, is that the permutation method indicates that apart from *weight* and *height*, the influence of the remaining features is almost completely insignificant as their permutation scores are near 0. At least based on this test, it is unusual that many other important features (i.e. water consumption, vegetable consumption, transportation mode) have no effect on the classification outcome. The results of Figure 12 certainly demand further study.

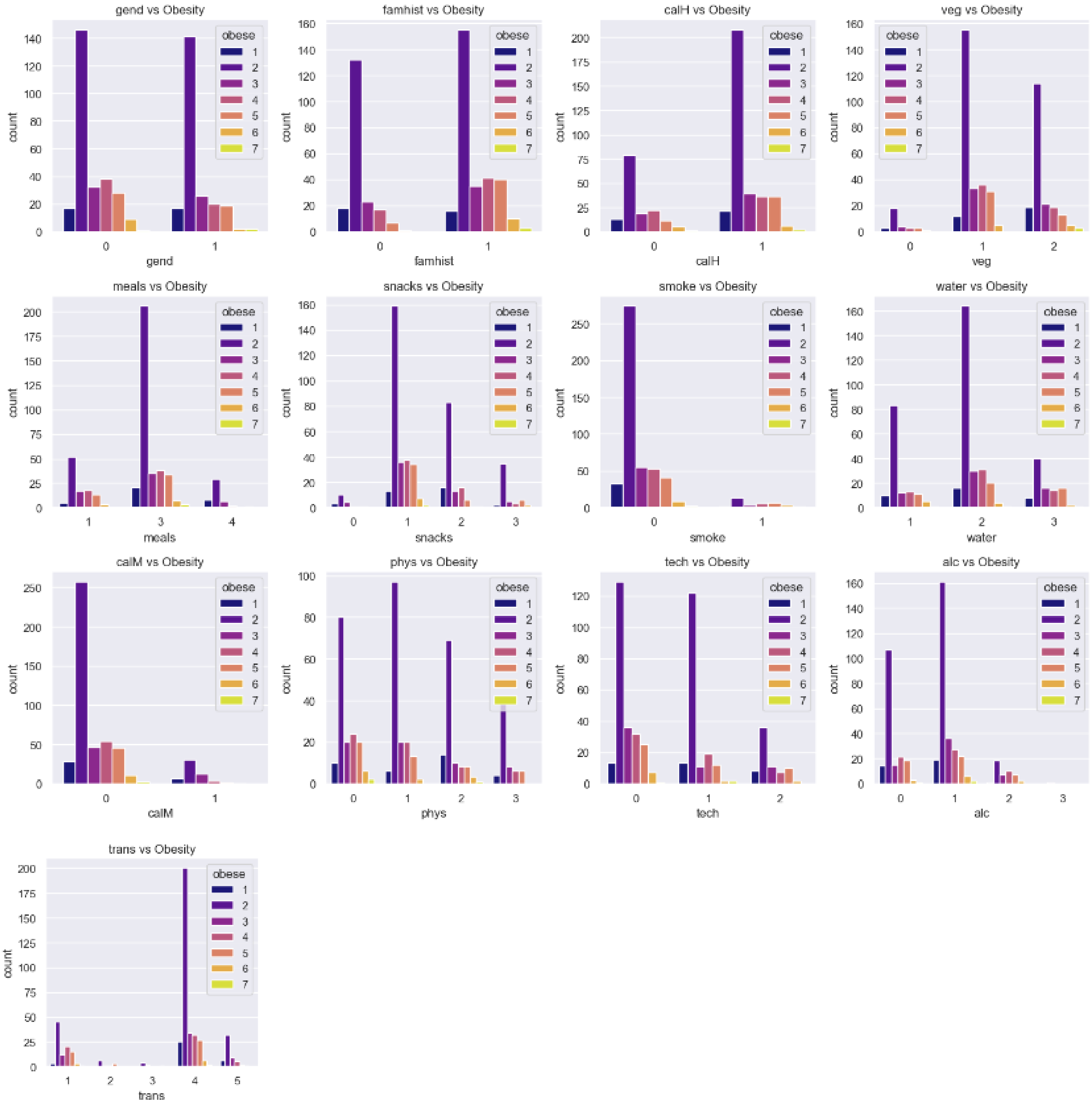
The critique in this section brings us to the most important limitation of this study: a poor sampling methodology. It is our view that the initial investigators erred in *what* data they collected and *how* it was collected. Although 17 valuable features were available, other important ones are missing. Family income and education, for example, could be important determinants of healthy lifestyles. For example, a wealthy individual might use their income to purchase high-quality foods, join sports clubs, and pay for gym memberships and/or personal trainers thereby reducing their likelihood of developing obesity. If individuals knew their body fat composition, that would have also been useful in distinguishing between weight attributed to fat and weight attributed to muscle. Another error in their methodology was in *how* they collected the data. They collected the data via an online survey that was available for approximately 30 days. There are several problems with this, but let us focus on two of them. First, weight estimates from individuals can be highly unreliable due to rounding and guessing. For example, an individual might recall weighing about 60 kilograms at one point (i.e. from a previous doctor's visit); however, their current weight could be very different now even if they do not think they have gained or lost weight. In other words, a person could believe they still weight about 60 kilograms, but they may in fact be significantly heavier or lighter. A second problem is that this was an online survey meaning that households without internet access were automatically excluded from participating. In other words, instead of applying a random sampling methodology, the original researchers inadvertently targeted a specific subset of the population. All of these shortcomings imply that no matter how we rebalance the data or what statistical methodology we employ, our results will always be biased and our model could never be used to make meaningful inferences about the populations they were intending to study. This emphasizes the importance of designing a study properly right from the beginning. If this is not performed correctly, then it does not matter what type of model is selected since the results will never be viable. Thus, if we had more time, we would propose the following:

- The application of a random sampling methodology that captures data on a wider segment of the population.
- The inclusion of other features (i.e. income, body fat percentage, etc.).
- Greater collaboration between health care providers, statisticians, and patients to ensure that the data can be collected anonymously and accurately.
- Testing of other models/methodologies (i.e. XGBoost, multinomial logistic regression).

A preliminary analysis using XGBoost and multinomial logistic regression can be found in *Appendix B.*

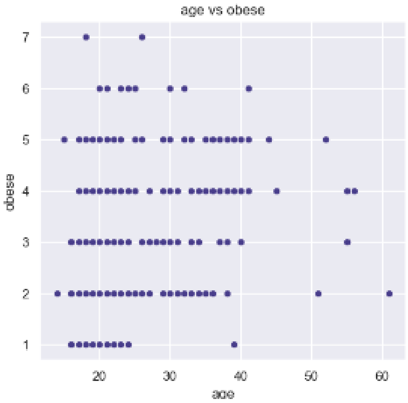
Appendix A: Supporting Plots/Figures (Exploratory Data Analysis I)

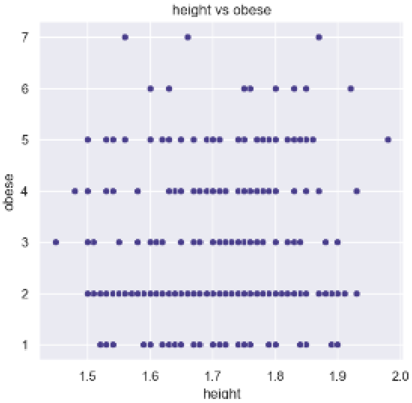
Categorical Features

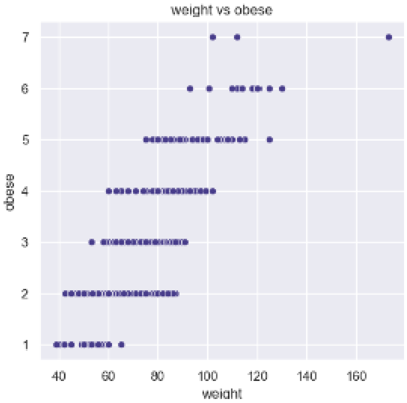


Appendix A (Continued)

Numerical features

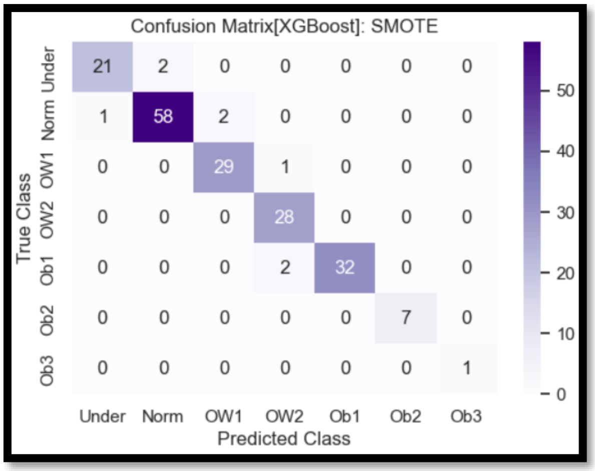






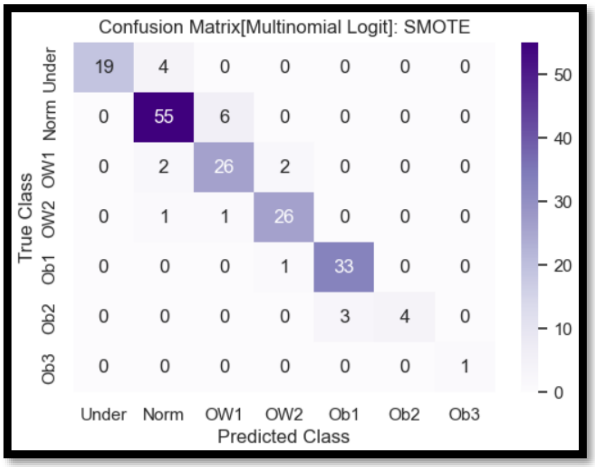
Appendix B: XGB & Multinomial Logistic Regression

i) XGBoost



Applying XGBoost produced an accuracy score of approximately 96%. This is an improvement over our original 92%.

i) MLR



Applying XGBoost produced an accuracy score of approximately 89%. This is slightly lower than our original 92%.

Although these model summaries are helpful, further analysis would be required to properly compare the models to determine which is the most appropriate, accurate, and interpretable. Consideration should also be given to overfitting, bias, and necessary assumptions.

References

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York, USA: Springer Science+Business Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R (2nd ed.).* New York, USA: Springer Science+Business Media.

Lingvay, I., Cohen, R., le Roux, C.W., & Sumithran, P. (2022). Obesity in Adults. *The Lancet,* 1-16. doi: https://doi.org/10.1016/S0140-6736(24)01210-8

Mendoza Palechor, F., & de la Hoz Manotas, A. (2019). *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru, and Mexico*. Colombia. https://doi.org/10.1016/j.dib.2019.104344

World Health Organization. (2025). Prevalence of obesity among adults [Web page].
Retrieved from: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-obesity-among-adults-bmi--30-(age-standardized-estimate)-(-)