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

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. A rebalancing procedure known as synthetic minority oversampling technique (SMOTE) was first applied to minimize the likelihood that a potential model would make predictions predominantly favouring the majority class. The data was then split into training and testing sets, and a random forest analysis was performed to evaluate the classification of new observations.

Results:

Conclusion: Future studies should consider accounting for other factors including (but not limited to): 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, 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 (6, p. 1). Interestingly, between 1985 to 2017, this increase was most detected in rural areas (6, 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 (6, 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 (6, 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. It is clear that obesity deserves greater attention and we lend our expertise in data analytics here.

The data file 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. Table 1 describes the variables and their encoding.

*Este espacio va a desaparecer cuando terminemos la parte de « Results » en el Abstract en la primer página y cuando incluya las fuentes (references)

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

Figure 1 depicts the distribution of the target variable obese across the different weight categories.

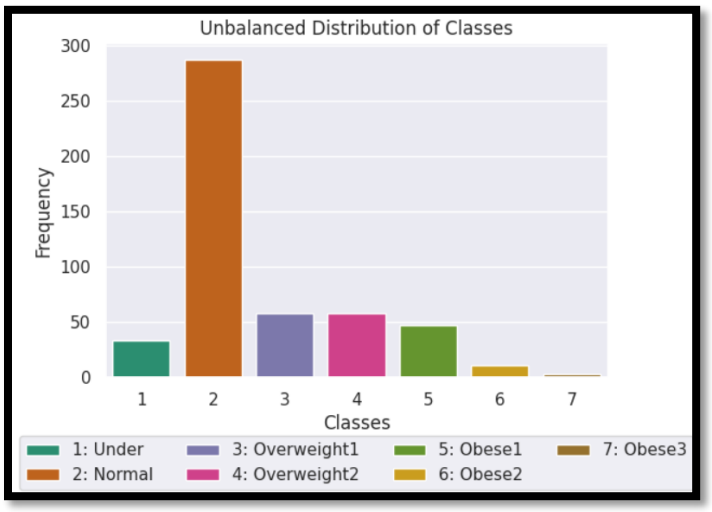

Figure 1: Unbalanced Distribution of Weight Classes

The unbalanced nature of the obese levels is clear; however, a statistical methodology to remedy this issue will be discussed in the *Formulation of Mathematical Model* section. Summary statistics for the continuous features are provided in Table 2 which give an overall sense of some of the main physical characteristics of the sample.

Table 2: Summary Statistics for Age, Height, and Weight			
	Age (Years)	Height (m)	Weight (kg)
Mean	23.5	1.69	74.1
Min	14	1.45	39
Max	61	1.98	173

Voy a agregar comentarios sobre Table 2

The central problem of this paper 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. As is evidenced by Figure 1, the problem is unique in that the response variable has 7 classifications and the data is unbalanced. Standard approaches such as multiple linear regression and logistic regression are unsuitable since the target variable is neither continuous nor binary. The minority categories could be grouped together thereby making logistic regression plausible, but this would result in a

significant loss of information. To address the shortcomings of the aforementioned models, this investigation utilized a classification tree methodology which was more apt for classification problem of this nature. Certainly other methodologies were available, and classification trees have their own inherent limitations; however, a careful review of the required assumptions for alternative models made classification trees more flexible for this study.

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? En todos los feature importance plots « gend » aparece bajo en la lista. Así que no sé si necesitamos esta pregunta o no. ¿Pensar en otra pregunta ?
3. *Eliminé la pregunta de interaction effects pq RF se encarga de eso ; ¿Pensar en otra pregunta ?

2. Formulation of Mathematical Algorithm

The model-building phase will be discussed in three components: assumptions and restrictions, variables and parameters, and a justification of the equations used.

2.1. Assumptions and Restrictions

Since the data was collected from 3 cities in 3 different countries, regional and cultural differences likely exist that 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 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 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

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. The algorithm is summarized in Table 4 (ESL, p. 588).

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)	

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 (ESL, 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. ¿Incluir una imagen de un decision tree aquí como ejemplo de su estructura básica ?

The next step was to determine if a random forest could be generated given the outline of the model provided in this section. Although their flexibility does make them suitable for classification problems, it is still imperative to run through appropriate tuning, training, and testing protocols to ensure that the classification results are reliable. Although no formal guidelines for the reliability and validity of classification trees, we erred on the side of caution by analyzing the results of the most current diagnostic tests for random forests.

3. Solution of the Problem

The first problem to consider was the highly unbalanced distribution of the weight classes. 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 of these 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). The basic framework of CW is to assign larger weight factors to minority classes and smaller weight factors to majority classes. SMOTE, on the other hand, generates synthetic samples from the data, and the desired number of instances in each class can be configured.

As a starting point we conducted a series of data cleaning and pre-processing which involved: checking for missing values and/or extreme outliers, encoding the features

appropriately, separating the data into training (80%) and testing (20%) sets, and finally standardizing the features for consistency. The train-test split and standardization steps generated two new processed sets of data: one with CW and the other relying on SMOTE. Next, the random forest classifier needed to be created so that testing could be performed on the two rebalancing methods.

A tuning procedure was first carried out to determine the ideal hyperparameter values for the data. The results suggested that a random forest composed of 300 trees with a maximum depth of 12 for each decision tree was one possible configuration for maximizing performance. Additionally, a maximum value of $\sqrt{p}$ features was chosen at each split which produced a decorrelating effect leading to greater reliability and a reduction in the average variance of the decision trees (ISLR, p.344). Thus, the random forest was set up according to these values and several diagnostics were run to test model efficacy between the two methods. The results are provided in Table 4.

Table 4: Model Evaluation Scores

Test	SMOTE	CW
Accuracy	0.918	0.770

The accuracy score is a basic measurement that compares how predictions made from a model compare to the actual values. In this case, the classification model based on SMOTE had an accuracy of approximately 92% which was significantly larger than the 77% accuracy of the CW approach. While the accuracy score is a useful statistic, a confusion matrix provides a better representation of how well each model performed by comparing the predicted and true values across all levels of obesity (See Figure 3).

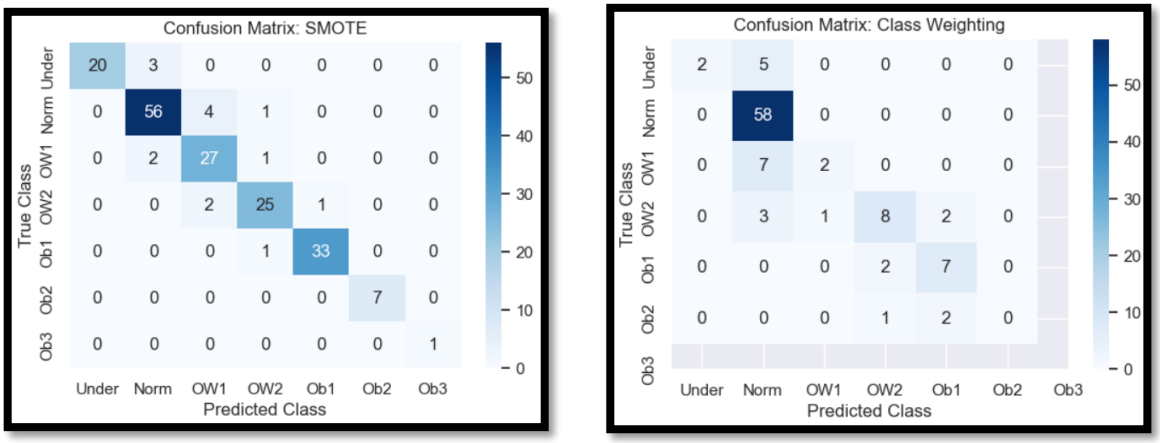

Figure 2: Confusion Matrices for SMOTE and CW

Amiga, ¿pq hay menos observaciones en el CW matriz ? Es por la forma en cómo elige la class ?

Figure 2 assessed model performance by comparing how the predictions faired against the true classifications. The different shades of blue are indicative of the quantity of observations in each frame. That is, the darker the shade, the more observations occurred in that frame. For example, in the SMOTE confusion matrix, the intersection of the predicted *OB1* versus the true *OB1* contains 33 accurate predictions shaded in a medium blue colour with 1 inaccurate prediction at the intersection of *OB1-OB2* in that same column. Most of the counts fell along the diagonal elements suggesting that the model using SMOTE performed well. However, a few entries fell outside of the diagonal which represented misclassifications. Amiga, ¿pq hay menos observaciones en el CW matriz ? Es por la forma en cómo elige la class ? No comenté sobre el CW matriz pq no estaba seguro de esto.

Thus, SMOTE was selected as the rebalancing procedure for the data set and Figure 3 depicts the rebalanced distribution of weight classes after its application.

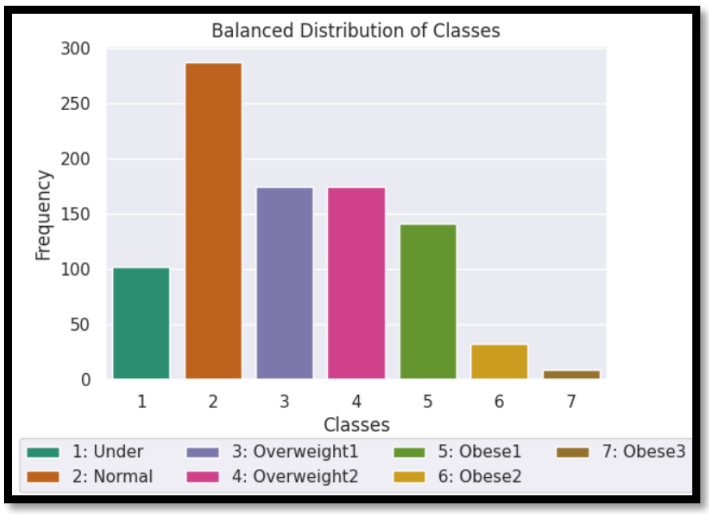

Figure 3: Balanced Distribution of Weight Classes

A threefold increase was applied to each minority class and although this scaling factor seemed arbitrary, it generated a reasonable and interpretable distribution of the weight classes. For example, Figure 3 suggested 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. Our analysis, however, intended to capture a more real-world scenario where cases of severe/extreme obesity (*Obese3*) still remain rare in comparison to the other weight classifications.

Through data preprocessing, parameter tuning, and testing we were able to formulate an obesity prediction model that could be used as an additional diagnostic tool to identity individuals who may be at risk of developing obesity. For the sake of comparison, however, two additional analyses were carried out: (i) multinomial logistic regression and; (ii) reduced random forest (See Appendices X & Y). The rationale for this was that fitting and assessing multiple models was part of our due diligence as researchers, and it assisted us in validating our model

and consequently the results obtained from it. – Esta última frase - Voy a dejar esto para mientras. Tal vez hacemos un multinomial pero no es importante en este momento creo yo

4. Interpretation of Results

Interpretation of results. Answer the questions posed in the introduction, and summarize major findings. Use non-technical terms.

The first question that we needed to answer was whether the model could provide some sort of assessment of how each feature contributes to the classification process. Several measures exist for ranking the importance of each feature; however, for the purposes of clarity, Figure 5 summarizes two/three of these methods.

ANDRES

1. Do I still need to perform feature selection? No. I can use the RED section in my notes to justify what happens when we remove 2 (doesn't make much difference) – so justifies not removing vars.

2. Do I need to use/write a pruning function or do random forests already know to apply that in the background processes? Pruning for individual tree; model starts with high V and low Bias; Averaging effect reduces V

3. Is there a way of showing in Python (or R) how random forests arrives at its classification (i.e. a visual or summary table) for a new prediction? Should be able to obtain proportion demonstrating why class chosen in new observation; Consider fitting 1 DT, interpret and compare to RF)

4. Do random forests account for interaction effects or do I have to use/write a function for this? Inherently captured!

5. Methods like random forests do not have a mathematical model; correct? Maybe Gaussian-esque (confirm); structure itself is the model

Suggestion: Run Multinomial since very sim to RF and compare