STAT5020 Final Project

WU Zhihao

May 4, 2022

Contents

1 Introduction

Since obesity is a fundamental medical problem, which increases the risk of other diseases, such as heart diseases, diabetes, and high pressure, more and more people impose great importance on their weight.

There are many reasons why people have difficulties losing weight. In general, obesity results from inheritance, physical and mental condition, and diet. Since the gene data is complex and hard to collect, we choose the family history of being overweight as the factor representing inheritance. For physical condition, we consider the frequency of having physical activities and the total time of using technological devices. For diet, we consider three factors, the number of main meals each day, the consumption of water, and the frequency of having vegetables.

The **main question** we want to investigate is "How the family history influences the effects of diets and physical condition on obesity".

This project will consider structural equation models with **multisample** data, treating the family history of obesity as the group indicator. Besides, we will use the **Bayes Factor** to realize the model comparison.

2 Data

The data set is original for estimating obesity levels in individuals from the countries of Mexico, Peru, and California, based on their eating habits and physical condition, containing 17 attributes and 2111 records. Notice that only 23% of the data was collected directly from the users, and the other data was generated synthetically via some techniques, assigning values to the missing data.

Since one of the primary purposes is to present the realization of structural equation models with multisample data, we treat the whole data as our observed data. Due to the limitation of computing resources and brevity, we only contain attributes with continuous artificially assigned values for missing data and treat these attributes as continuous variables.

We construct one outcome latent variable, representing the level of obesity, and two explanatory latent variables, representing the conditions of eating diet and physical activities. The list of symbols and corresponding meanings follows:

- individual indicator: i, representing each individual.
- group indicator: g, representing the family history with overweight.
  - 1 for "No" and 2 for "Yes".
- latent variables: $\boldsymbol{\Omega} = (\boldsymbol{\omega}_i)$.
  - outcome latent variable: η_i, representing the factor of obesity.
  - explanatory latent variables: $\boldsymbol{\xi}_i$.
    - ξ_{1i}: the one related to eating diets.
    - ξ_{2i}: the one related to physical conditions.
- observed variables: $\boldsymbol{Y} = (\boldsymbol{y}_i)$.
  - y_{i1}: the mass body index, calculated by the weight (kg) dividing the square of height (m).
  - y_{i2}: (Weight) weight with the unit kilogram.
  - y_{i3}: (NCP) the number of main meals having daily, from 1 to 4.
  - y_{i4}: (CH2O) the consumption of water daily, from 1 to 3 (L).
  - y_{i5}: (FCVC) the frequency of having vegetables in your meals.
    - 1 for "Never", 2 for "Sometimes", and 3 for "Always".
  - y_{i6}: (FAF) the frequency of having physical activity.
    - from 0 to 3, and the larger value representing the more frequent exercise.
  - y_{i7}: (TUE) the total time of using technological devices.
    - from 0 to 2, and the higher score indicating the heavier use of electronic devices.

Remark:

1. For brevity, we ignore the group superscript index (g) in the description of symbols above.
2. The capital characters in the parentheses, e.g. NCP and CH2O, are the variable names in the .csv file.
3. Since the synthetically assigned values for missing data are continuous, then all these selected observed variables are treated as continuous.

3 Model

3.1 Model 0 (unisample model)

In **Model 0**, we don't take the group information into consideration.

The goal of this model is to get informative and proper priors for the latter Bayesian inference.

The measurement equation of **Model 0** is

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \\ y_{i5} \\ y_{i6} \\ y_{i7} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{73} \end{bmatrix} \begin{bmatrix} \eta_i \\ \xi_{i1} \\ \xi_{i2} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \\ \epsilon_{i5} \\ \epsilon_{i6} \\ \epsilon_{i7} \end{bmatrix}, \quad \epsilon_{ik} \sim \mathcal{N}\left(0, \psi_{\epsilon k}\right), \tag{1}$$

and the structural equation of **Model 0** is

$$\eta_i = \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \delta_i, \quad \begin{cases} \delta_i \sim \mathcal{N}\left(0, \psi_\delta\right), \\ \boldsymbol{\xi}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Phi}\right). \end{cases} \tag{2}$$

We fix some entries of the loading matrix in the measurement equation as 1 to deal with the identifiability problem.

To better interpret the structure of the model and the relationships between these latent variables and observed variables, we draw the path plot Figure 1.

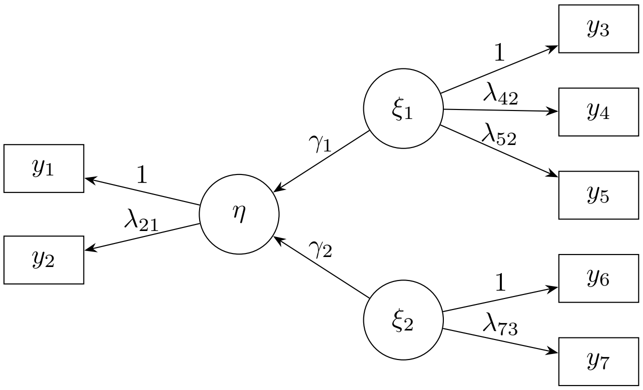


Figure 1: Path Diagram

In practice, we fit this model twice to get better convergence of the sample sequence.

For the first fitting, we use priors with large variance, which are (almost) noninformative. To be more specific,

$$\begin{cases} \mu_k \sim \mathcal{N}\left(\bar{y}_k, 0.25\right), \\ \lambda_{21}, \lambda_{42}, \lambda_{52}, \lambda_{73} \sim \mathcal{N}\left(0, 25\right), \\ \gamma_1, \gamma_2 \sim \mathcal{N}\left(0, 25\right), \end{cases} \tag{3}$$

and the rest of parameters follows the noninformative constant prior.

And for the second fitting, we consider conjugate priors with small dispersion based on the previous fitting:

$$\begin{cases} \mu_k \sim \mathcal{N}\left(\bar{y}_k, 0.25\right), \\ \psi_{\epsilon 2} \sim \mathrm{IG}\left(600, 50000\right), \\ \psi_{\epsilon k} \sim \mathrm{IG}\left(9, 4\right), k \neq 2, \\ \lambda_{21} \sim \mathcal{N}\left(3, 0.25 \times \psi_{\epsilon 2}\right), \\ \lambda_{42} \sim \mathcal{N}\left(1, 0.25 \times \psi_{\epsilon 4}\right), \\ \lambda_{52} \sim \mathcal{N}\left(0.9, 0.25 \times \psi_{\epsilon 5}\right), \\ \lambda_{73} \sim \mathcal{N}\left(0.05, 0.25 \times \psi_{\epsilon 7}\right), \\ \psi_{\delta} \sim \mathrm{IG}\left(300, 12000\right), \\ \gamma_1 \sim \mathcal{N}\left(24, 0.25 \times \psi_{\delta}\right), \\ \gamma_2 \sim \mathcal{N}\left(-4, 0.04 \times \psi_{\delta}\right), \\ \mathbf{\Phi} \sim \mathrm{IW}_2\left(\begin{bmatrix} 0.04 & 0.05 \\ 0.05 & 0.7 \end{bmatrix}, 5\right). \end{cases} \tag{4}$$

Figure 2 shows that **Model 0** sample traces of the chains are reasonable, then the results of Bayesian inference in Table 1 are reasonable.

	EST	SE		EST	SE
μ_1	29.703	0.1238	$\psi_{\epsilon 1}$	0.390	0.0906
μ_2	86.596	0.3755	$\psi_{\epsilon 1}$	83.136	2.1690
μ_3	2.686	0.0171	$\psi_{\epsilon 1}$	0.580	0.0184
μ_4	2.008	0.0134	$\psi_{\epsilon 1}$	0.334	0.0122
μ_5	2.419	0.0117	$\psi_{\epsilon 1}$	0.252	0.0094
μ_6	1.010	0.0184	$\psi_{\epsilon 1}$	0.289	0.0514
μ_7	0.658	0.0131	$\psi_{\epsilon 1}$	0.369	0.0113
λ_{21}	3.075	0.0257	ψ_{δ}	38.175	1.6965
λ_{42}	1.174	0.1294	ϕ_{11}	0.031	0.0044
λ_{52}	1.059	0.1259	ϕ_{21}	0.041	0.0083
λ_{73}	0.077	0.0283	ϕ_{12}	0.041	0.0083
γ_1	28.866	2.0459	ϕ_{22}	0.435	0.0544
γ_2	-5.360	0.6379			

Table 1: Bayesian estimates of parameters in **Model 0**.

3.2 Model 1 (mutli-sample model)

Model 1 is our main model, which takes the group information, family history of overweight, into consideration, and investigates the difference of parameter estimates in different groups.

The measurement equation of **Model 1** is

$$\begin{bmatrix} y_{i1}^{(g)} \\ y_{i2}^{(g)} \\ y_{i3}^{(g)} \\ y_{i4}^{(g)} \\ y_{i5}^{(g)} \\ y_{i6}^{(g)} \\ y_{i7}^{(g)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21}^{(g)} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{42}^{(g)} & 0 \\ 0 & \lambda_{52}^{(g)} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{73}^{(g)} \end{bmatrix} \begin{bmatrix} \eta_i^{(g)} \\ \xi_{i1}^{(g)} \\ \xi_{i2}^{(g)} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1}^{(g)} \\ \epsilon_{i2}^{(g)} \\ \epsilon_{i3}^{(g)} \\ \epsilon_{i4}^{(g)} \\ \epsilon_{i5}^{(g)} \\ \epsilon_{i6}^{(g)} \\ \epsilon_{i7}^{(g)} \end{bmatrix}, \quad \epsilon_{ik}^{(g)} \sim \mathcal{N}\left(0, \psi_{\epsilon k}^{(g)}\right), \tag{5}$$

and the structural equation is

$$\eta_i^{(g)} = \gamma_1^{(g)} \xi_{i1}^{(g)} + \gamma_2^{(g)} \xi_{i2}^{(g)} + \delta_i^{(g)}, \quad \begin{cases} \delta_i^{(g)} \sim \mathcal{N}\left(0, \psi_\delta^{(g)}\right), \\ \boldsymbol{\xi}_i^{(g)} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Phi}^{(g)}\right). \end{cases} \tag{6}$$

We consider informative priors for Bayesian inference:

$$\begin{cases} \mu_k^{(g)} \sim \mathcal{N}\left(\bar{y}_k, 0.25\right), \\ \psi_{\epsilon 2}^{(g)} \sim \mathrm{IG}\left(600, 50000\right), \\ \psi_{\epsilon k}^{(g)} \sim \mathrm{IG}\left(9, 4\right), k \neq 2, \\ \lambda_{21}^{(g)} \sim \mathcal{N}\left(3, 0.25\right), \\ \lambda_{42}^{(g)} \sim \mathcal{N}\left(1, 0.25\right), \\ \lambda_{52}^{(g)} \sim \mathcal{N}\left(0.9, 0.25\right), \\ \lambda_{73}^{(g)} \sim \mathcal{N}\left(0.05, 0.25\right), \\ \psi_\delta^{(g)} \sim \mathrm{IG}\left(300, 12000\right), \\ \gamma_1^{(g)} \sim \mathcal{N}\left(24, 1\right), \\ \gamma_2^{(g)} \sim \mathcal{N}\left(-4, 0.25\right), \\ \mathbf{\Phi}^{(g)} \sim \mathrm{IW}_2\left(\begin{bmatrix} 0.04 & 0.05 \\ 0.05 & 0.7 \end{bmatrix}, 5\right). \end{cases} \tag{7}$$

And there are some remarks about the choice of priors above:

1. Since currently, we have no information about the difference between groups, then we take the same priors for the same parameters in the different groups.
2. The information is from the Bayesian inference results of previous model.
3. For free parameters in the factor loading matrix and coefficients in the structural equation, e.g. $\lambda_{21}^{(g)}$ amd $\gamma_1^{(g)}$, we consider normal priors, instead of conjugate conditional normal priors given the variance of the corresponding error measurements, to avoid the potential problems induced by the structural equation model with multi-sample data under some constraints.

Table 2 presents the results of the Bayesian inference of **Model 1**.

	group 1	group 2		group 1	group 2
μ_1	27.721	30.433	$\psi_{\epsilon 1}$	0.371	0.424
μ_2	84.067	88.972	$\psi_{\epsilon 2}$	77.850	86.921
μ_3	2.628	2.700	$\psi_{\epsilon 3}$	0.850	0.508
μ_4	1.872	2.041	$\psi_{\epsilon 4}$	0.404	0.336
μ_5	2.388	2.410	$\psi_{\epsilon 5}$	0.354	0.202
μ_6	1.124	1.009	$\psi_{\epsilon 6}$	0.338	0.361
μ_7	0.628	0.667	$\psi_{\epsilon 7}$	0.407	0.360
λ_{21}	3.701	2.996	ψ_{δ}	41.391	36.304
λ_{42}	0.943	0.791	ϕ_{11}	0.055	0.028
λ_{52}	0.268	1.574	ϕ_{21}	0.131	0.016
λ_{73}	0.045	0.121	ϕ_{12}	0.131	0.016
γ_1	23.700	25.518	ϕ_{22}	0.538	0.330
γ_2	-5.475	-4.803			

Table 2: Bayesian estimates of parameters in **Model 1**.

3.3 Model Comparison

In this subsection, we consider an hypothesis testing problem:

$$H_0 : \boldsymbol{\Gamma}^{(1)} = \boldsymbol{\Gamma}^{(2)} = \boldsymbol{\Gamma} \quad \text{v.s.} \quad H_1 : \boldsymbol{\Gamma}^{(1)} \neq \boldsymbol{\Gamma}^{(2)} \tag{8}$$

The common strategy to solve this hypothesis testing problem is to convert this hypothesis testing problem to a model comparison problem. And we will realize model comparison via Bayes factor.

Rewrite **Model 1** in matrix form:

$$\begin{cases} \boldsymbol{y}_i^{(g)} &= \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} &= \boldsymbol{\Gamma}^{(g)} \boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}, \end{cases} \tag{9}$$

and consider **Model 2**:

$$\begin{cases} \boldsymbol{y}_i^{(g)} &= \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)} \\ \eta_i^{(g)} &= \boldsymbol{\Gamma} \boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}. \end{cases} \tag{10}$$

Since the distributions of error measurements and explanatory output latent variables are all the same in this subsection,

$$\begin{cases} \boldsymbol{\epsilon}_i^{(g)} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Psi_\epsilon}^{(g)}\right), \\ \delta_i^{(g)} \sim \mathcal{N}\left(0, \psi_\delta^{(g)}\right), \\ \boldsymbol{\xi}_i^{(g)} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Phi}^{(g)}\right) \end{cases} \tag{11}$$

then we will ignore these distributions when describing the model in this subsection.

Since it's not easy to link **Model 1** and **Model 2** directly, we introduce an auxiliary **Model a**,

$$\begin{cases} \boldsymbol{y}_i^{(g)} &= \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} &= \delta_i^{(g)}. \end{cases} \tag{12}$$

We first estimate the Bayes factor between **Model 1** (M_1) and **Model a** (M_a) via the path sampling method.

Construct a link model, **Model t1a** ($\mathrm{M}_{\mathrm{t1a}}$),

$$\begin{cases} \boldsymbol{y}_i^{(g)} &= \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} &= t \cdot \boldsymbol{\Gamma}^{(g)} \boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}, \end{cases} \tag{13}$$

and notice that $\mathrm{M}_{\mathrm{t1a}}$ reduces to M_1 when $t = 1$, while $\mathrm{M}_{\mathrm{t1a}}$ reduces to M_a when $t = 0$.

Consider the grids from 0 to 1 with step 0.05, and on each fixed grid $t_{(s)}$, generate observations

$$\left(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}\right) \sim p\left(\boldsymbol{\Omega}, \boldsymbol{\theta} \,|\, \boldsymbol{Y}, t_{(s)}\right), \quad j = 1, \cdots, J = 1000, \tag{14}$$

after 500 warm-up iterations using MCMC methods.

Then calculate $U(\boldsymbol{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)})$ by

$$\begin{aligned} U(\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t) &= \frac{d}{dt} \log p\left(\boldsymbol{\Omega}, \boldsymbol{\theta} \,|\, \boldsymbol{Y}, t\right)\Big|_{t=t_{(s)}} \\ &= \sum_{g=1}^{2} \sum_{i=1}^{n_g} \frac{1}{\psi_\delta^{(g)}} (\eta_i^{(g)} - t_{(s)} \cdot \gamma_1^{(g)} \xi_{i1}^{(g)} - t_{(s)} \cdot \gamma_2^{(g)} \xi_{i2}^{(g)})(\gamma_1^{(g)} \xi_{i1}^{(g)} + \gamma_2^{(g)} \xi_{i2}^{(g)}) \end{aligned} \tag{15}$$

then

$$\bar{U}_{(s)} = \frac{1}{J} \sum_{j=1}^{J} U(\boldsymbol{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}). \tag{16}$$

and then the Bayes factor between M_1 and $\mathrm{M_a}$ is estimated by

$$\widehat{\log \mathrm{BF}}_{1a} = \frac{1}{2} \sum_{s=0}^{S} (t_{(s+1)} - t_s)(\bar{U}_{s+1} - \bar{U}_s). \tag{17}$$

Similarly, we construct link model, **Model t2a** ($\mathrm{M_{t2a}}$),

$$\begin{cases} \boldsymbol{y}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} = t \cdot \boldsymbol{\Gamma} \boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}, \end{cases} \tag{18}$$

and based on the same procedure, we can get the estimate $\widehat{\log \mathrm{BF}}_{2a}$.

The Bayes factor between M_1 and M_2 is estimated by

$$\widehat{\log \mathrm{BF}}_{12} = \widehat{\log \mathrm{BF}}_{1a} - \widehat{\log \mathrm{BF}}_{2a}. \tag{19}$$

In realization, we replicate the calculation for $\bar{U}_{(11)}$ ($t = 0.5$) 5 times to verify that the number of iterations is enough to get a stable results. Numeric results follow:

$$\begin{pmatrix} 65.82619 & 66.57391 & 67.88888 & 66.22054 & 62.90564 \end{pmatrix}.$$

The estimated $\widehat{\log \mathrm{BF}}_{12}$ is equal to 0.9906.

4 Conclusion

Based on **Model 1**, we can conclude that

1. Not only the mass body index, but also the weight are significantly related to the outcome latent variable interpreted as the level of obesity.

2. All the coefficients in the measurement equation and structural equation, except $\lambda_{52}^{(1)}$ and $\lambda_{73}^{(1)}$, are significant.

3. Notice that $\gamma_1^{(2)}(= 25.52) > \gamma_1^{(1)}(= 23.70) > 0$, then the effect of eating diets on obesity in the second group is stronger than the one in the first group.

   Hence, having the same eating diet, people with family obesity history are more likely to being overweight than those people without family obesity history.

4. Notice that $\gamma_2^{(1)}(= -5.48) < \gamma_2^{(2)}(= -4.80) < 0$, then the effect of physical activity on obesity in the second group is weaker than the effect in the first group.

   Therefore, it's harder for those people with family obesity history losing weight via taking more physical activities then those without family obesity history.

Based on the estimate of Bayes factor between **Model 1** and **Model 2**, since $2\widehat{\log \mathrm{BF}}_{12} = 1.98$, then M_1 is slightly better than M_2, the data provide weak evident to support the alternative hypothesis $H_1 : \mathbf{\Gamma}^{(1)} \neq \mathbf{\Gamma}^{(2)}$. Hence, we can conclude that the coefficients in the structural equation in different group are different.

A Figure

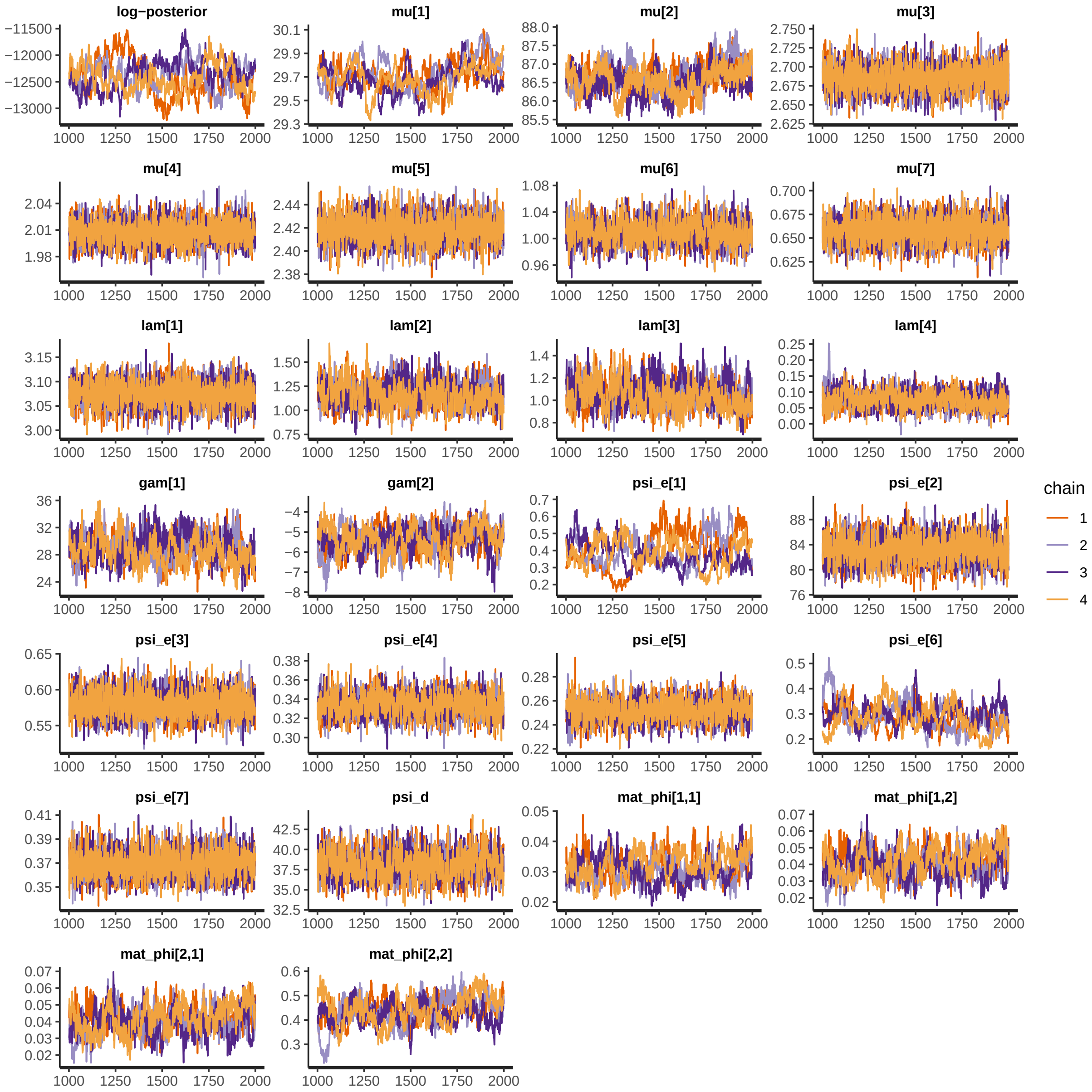


Figure 2: Trace Plot of Model 0

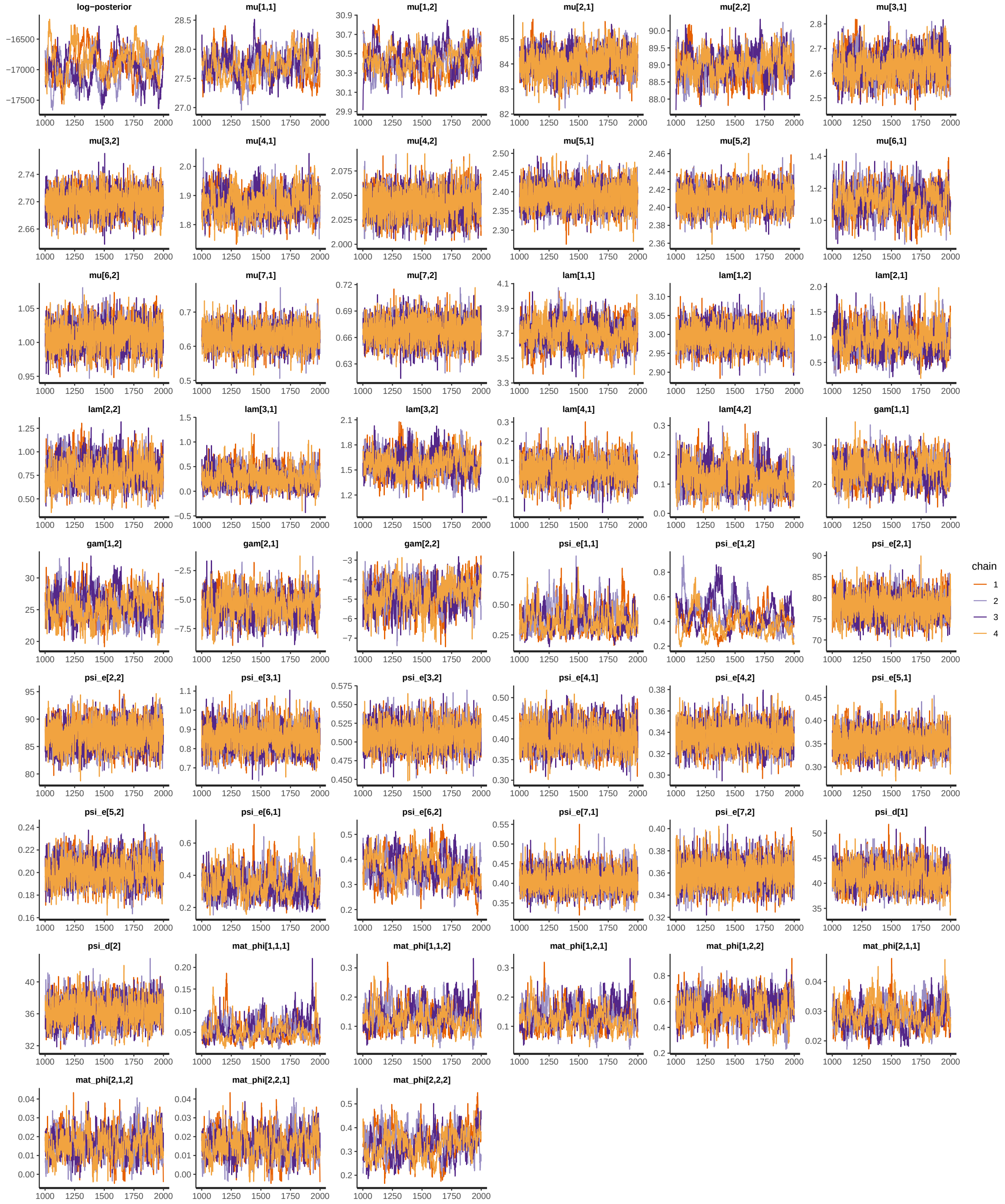
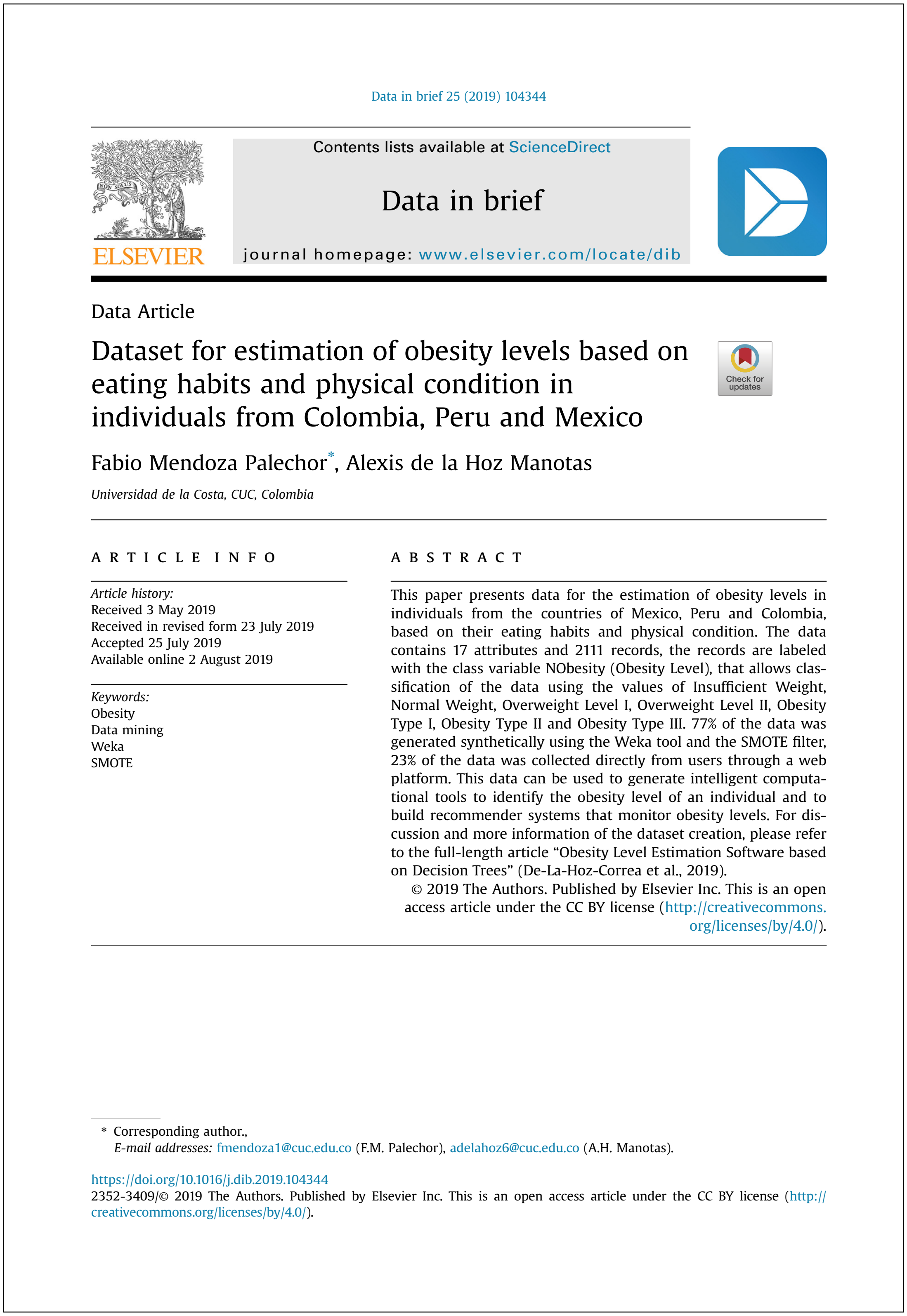


Figure 3: Trace Plot of Model 1

B Data Set Official Document



Data in brief 25 (2019) 104344

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico

Fabio Mendoza Palechor[*], Alexis de la Hoz Manotas

Universidad de la Costa, CUC, Colombia

ARTICLE INFO

Article history:
Received 3 May 2019
Received in revised form 23 July 2019
Accepted 25 July 2019
Available online 2 August 2019

Keywords:
Obesity
Data mining
Weka
SMOTE

ABSTRACT

This paper presents data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform. This data can be used to generate intelligent computational tools to identify the obesity level of an individual and to build recommender systems that monitor obesity levels. For discussion and more information of the dataset creation, please refer to the full-length article "Obesity Level Estimation Software based on Decision Trees" (De-La-Hoz-Correa et al., 2019).

[*] Corresponding author.,
E-mail addresses: fmendoza1@cuc.edu.co (F.M. Palechor), adelahoz6@cuc.edu.co (A.H. Manotas).

https://doi.org/10.1016/j.dib.2019.104344

Specifications table

Subject area	*Biology*
More specific subject area	*Obesity, cardiovascular risk*
Type of data	*Text, table, figure*
How data was acquired	*Survey (see Table 1)*
Data format	*Raw and Analyzed*
Experimental factors	Data was retrieved from online survey and preprocessed including missing and atypical data deletion, and data normalization
Experimental features	*Labeling process was performed based on WHO and Mexican Normativity. Balancing class was performed using the SMOTE filter using the tool Weka. Features were chosen based on literacy analysis.*
Data source location	*Barranquilla – Colombia, Lima – Peru, City of Mexico - Mexico*
Data accessibility	*Data is within this article*
Related research article	*E. De-La-Hoz-Correa, F. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. Morales-Ortega, B. Sánchez Hernández. Obesity Level Estimation Software based on Decision Trees, Journal of Computer Science, 67, 2019 [6]*

Value of the data

- This data presents information from different locations such as Mexico, Peru and Colombia, can be used to build estimation of the obesity levels based on the nutritional behavior of several regions.
- The data can be used for estimation of the obesity level of individuals using seven categories, allowing a detailed analysis of the affectation level of an individual.
- The structure and amount of data can be used for different tasks in data mining such as: classification, prediction, segmentation and association.
- The data can be used to build software tools for estimation of obesity levels.
- The data can validate the impact of several factors that propitiate the apparition of obesity problems.

1. Data

This paper contains data for the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition as mentioned by [1], data was collected using a web platform with a survey (see Table 1) where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records, after a balancing process described in Figs. 1 and 2. The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other variables obtained were: Gender, Age, Height and Weight. Finally, all data was labeled and the class variable NObesity was created with the values of: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III, based on Equation (1) and information from WHO and Mexican Normativity. The data contains numerical data and continous data, so it can be used for analysis based on algorithms of classification, prediction, segmentation and association. Data is available in CSV format and ARFF format to be used with the Weka tool.

2. Experimental design, materials, and methods

The initial recollection of information was made through a web page using a survey where users had evaluated their eating habits and some aspects that helped to identify their physical condition. The survey was accesible online for 30 days. In Table 1, the questions of the survey are presented.

After all data was collected, then data was preprocessed, so it could be used for different techniques of data mining. The number of records was 485 records, and the data was labeled using equation (1).

Table 1
Questions of the survey used for initial recollection of information.

Questions	Possible Answers
¿What is your gender?	• Female • Male
¿what is your age?	Numeric value
¿what is your height?	Numeric value in meters
¿what is your weight?	Numeric value in kilograms
¿Has a family member suffered or suffers from overweight?	• Yes • No
¿Do you eat high caloric food frequently?	• Yes • No
¿Do you usually eat vegetables in your meals?	• Never • Sometimes • Always
¿How many main meals do you have daily?	• Between 1 y 2 • Three • More than three
¿Do you eat any food between meals?	• No • Sometimes • Frequently • Always
¿Do you smoke?	• Yes • No
¿How much water do you drink daily?	• Less than a liter • Between 1 and 2 L • More than 2 L
¿Do you monitor the calories you eat daily?	• Yes • No
¿How often do you have physical activity?	• I do not have • 1 or 2 days • 2 or 4 days • 4 or 5 days
¿How much time do you use technological devices such as cell phone, videogames, television, computer and others?	• 0–2 hours • 3–5 hours • More than 5 hours
¿how often do you drink alcohol?	• I do not drink • Sometimes • Frequently • Always
¿Which transportation do you usually use?	• Automobile • Motorbike • Bike • Public Transportation • Walking

$$Mass\ body\ index = \frac{Weight}{height*height} \tag{1}$$

After all calculation was made to obtain the mass body index for each individual, the results were compared with the data provided by WHO and the Mexican Normativity [7].

- Underweight Less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III Higher than 40

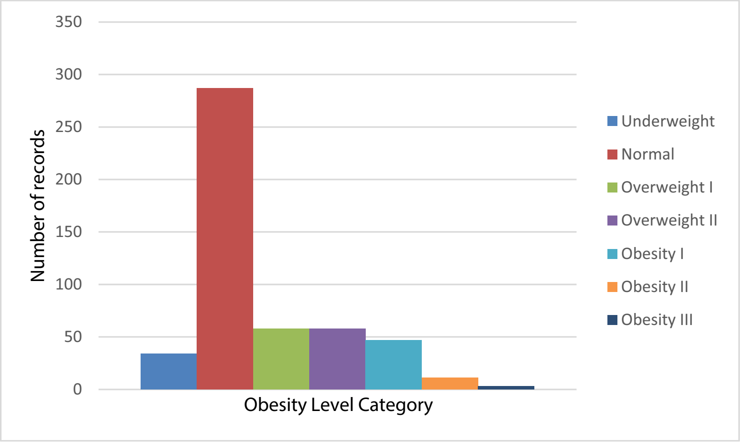


Fig. 1. Unbalanced distribution of data regarding the obesity levels category.

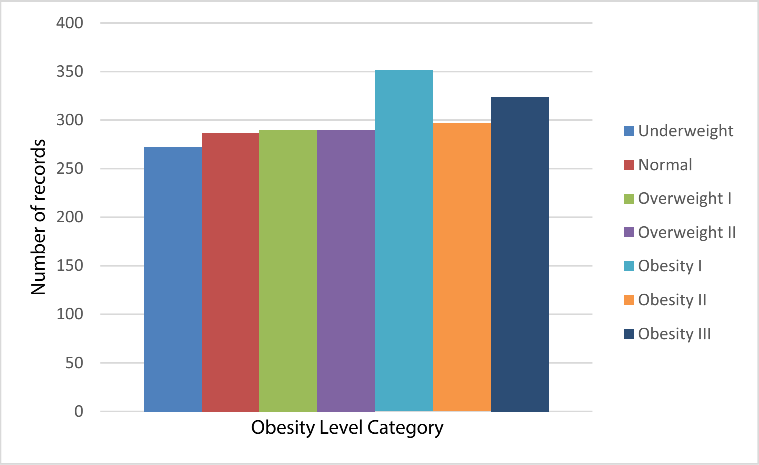


Fig. 2. Balanced Distribution of data regarding the obesity levels category.

After the labeling process was finished, the categories of obesity levels were unbalanced (as shown in Fig. 1), and this presented a learning problem for the data mining methods, since it would learn to identify correctly the category with most records compared with the categories with less data. In [8], you can see a dataset is unbalanced if the classification categories are not represented equally.

After the balancing class problem was identified, synthetic data was generated, up to 77% of the data, using the tool Weka and the filter SMOTE proposed by [8]. The filter required to indicate the class for generation of synthetic data, the number of nearest neighbors used, the percentage that you need to increase the selected class and the random seed used for random sampling. Other aspects analyzed were the identification of atypical and missing data. Finally, after the filter was applied to each category, the final result were 2111 records. Next, in Fig. 2 you can see the final distribution of the data after the balancing process was completed.

It is important to notice that data must be preprocessed (delete missing data, atypical data, data normalization, etc.) before using SMOTE, since the neighbor selected to generate the synthetic data could contain noise or disturbances, and the data produced would have low quality. Nevertheless, using the filter SMOTE has a positive impact when data is unbalanced, since the balancing process decrease the probability of skewed learning on favor of a majority class.

Features included in the data were chose based in literacy analysis such as [1–6,8], and there is a noticeable relationship between weight and height given by Equation (1).

This data contributes to build tools using computational intelligence for detecting obesity levels based on eating habits and physical conditions, seeing that sometimes available data lack of the necessary number of records, they are not publicly accessible, and their structure makes difficult the application of several methods on the data.

Acknowledgments

Universidad de la Costa (CUC) has provided technical support for this project. Authors would like to thank Beatriz Sanchez Hernandez for her technical support.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.V. Olmedo, La obesidad: un problema de salud pública. Revista de divulgació científica y tecnológica de la Universidad Veracruzana, 2011. Recuperado de: https://www.uv.mx/cienciahombre/revistae/vol24num3/articulos/obesidad/.

[2] C. Davila-Payan, M. DeGuzman, K. Johnson, N. Serban, J. Swann, Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data, Prev. Chronic Dis. 12 (2015).

[3] S. Manna, A.M. Jewkes, Understanding early childhood obesity risks: an empirical study using fuzzy signatures, in: Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on, IEEE, 2014, July, pp. 1333–1339.

[4] M.H.B.M. Adnan, W. Husain, A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction, in: Computer & Information Science (ICCIS), 2012 International Conference on vol. 1, IEEE, 2012, June, pp. 281–285.

[5] T.M. Dugan, S. Mukhopadhyay, A. Carroll, S. Downs, Machine learning techniques for prediction of early childhood obesity, Appl. Clin. Inf. 6 (3) (2015) 506–520.

[6] Eduardo De-La-Hoz-Correa, Fabio E. Mendoza-Palechor, Alexis De-La-Hoz-Manotas, Roberto C. Morales-Ortega, Beatriz Adriana Sánchez Hernández, Obesity level estimation software based on decision Trees, J. Comput. Sci. 15 (Issue 1) (2019) 67–77, https://doi.org/10.3844/jcssp.2019.67.77.

[7] DO, NORMA Oficial Mexicana NOM-008-SSA3-2010, Para el tratamiento integral del sobrepeso y la obesidad, Diario Oficial, 2010.

[8] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.