Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 220 (2023) 978–983

Procedia Computer Science

www.elsevier.com/locate/procedia

The 1st International Workshop on Human-Centric Innovation and Computational Intelligence (IWHICI 2023)

Detection of variables for the diagnosis of overweight and obesity in young Chileans using machine learning techniques.

Mailyn Calderón-Díaz[abc*], Leonardo J. Serey-Castillo[d], Esperanza A. Vallejos-Cuevas[d], Alexis Espinoza[d], Rodrigo Salas[bc], Mayra A. Macías-Jiménez[e]

[a]*Facultad de Ingeniería, Universidad Andrés Bello, Antonio Varas 880 Providencia, Santiago 8320000, Chile.*
[b]*Doctorado en Ciencias e Ingeniería para Salud, Universidad de Valparaíso, Valparaíso 2340000, Chile.*
[c]*Millennium Institute for Intelligent Healthcare Engineering (iHealth), Valparaíso 2340000, Chile.*
[d]*Escuela de Kinesiología, Facultad de Salud, Universidad Santo Tomás, Av. Ejército Libertador 146, Santiago 8320000, Chile*
[e]*Departamento de Productividad e Innovación, Universidad de la Costa, Calle 58 55-66, Barranquilla 080001, Colombia*

Abstract

Overweight and obesity are considered epidemic problems. The number of factors involved in developing extra body fat makes harder the detection of this problem. Therefore, among the several variables and their levels presented in overweight and obese people, there is a need to improve the classification of people with these conditions. To this aim, in this paper, we conducted a variable analysis from biochemical and lipid profiles in young Chileans with normal weight, overweight, and obesity using machine learning techniques. XGBoost library was selected as the classifier. 21 variables (13 from biochemical and 8 from lipid profiles) were chosen as features. 100 iterations were conducted, and an 80% cross-validation was obtained. The variables with greater relevance in the classification task were total cholesterol, glycemia, LDH enzyme, bilirubin, and VLDL cholesterol. All of these, except bilirubin, are consistent with previous research in which these features have been used to assess risk factors for developing overweight or obesity. Then, further research must include a deep study regarding bilirubin's influence over these conditions.

Keywords: Machine learning, biochemical profiles, lipid profiles, normal-weight, overweight, obesity, classification.

* Corresponding author.
E-mail address: mailyn.calderon@unab.cl

10.1016/j.procs.2023.03.135

1. Introduction

Overweight and obesity represent a public health concern worldwide. According to the WHO (World Health Organization), overweight and obese people experience several risks to their health because of the extra body fat [1]. The prevalence of these conditions in children and youth has risen special attention from government and public organizations [2]. For instance, the available data show that a third of US children and adolescents are estimated to be overweight or obese [3]. Then, most developed countries consider overweight and obesity in youth as epidemic problems [4]. Therefore, to decrease the incidence of these phenomena, public policies have been focused on the promotion of acquiring healthier lifestyles including physical activities at an early age [5].

In Latin America, overweight and obesity rates have an increasing behavior in recent years [6; 7; 8]. In this zone, some factors influence childhood overweight and obesity, such as socio-economic conditions and even genetics [7; 8]. Particularly, Chile is in the 2nd place among the Organization for Economic Cooperation and Development (OECD) members in obesity prevalence [9]. The country has followed the same pattern as global performance, increasing its overweight and obesity rates in the last few years [10]. Despite the latent need to prioritize this situation, there are limited interventions for obesity prevention in Latin America. However, there is a consensus that these actions need to take into consideration multidimensional factors. Overweight and obesity used to be defined based on the body mass index (BMI), using boundaries of 25 kg/m^2 and 30 kg/m^2 as cut-off values, respectively. However, there are other indicators that, together with BMI, have improved the detection of overweight and obese people.

One of these is the biochemical and lipid profile analysis. The relationship between lipid profiles and overweight or obesity has been proven before [11; 12]. Techniques such as bivariate and multivariate analyses have been used to study this association [13]. But there is evidence arguing that traditional approaches are limited in identifying predictors efficiently [14].

Thus, in this article, we try to fill this gap by finding the critical variables from biochemical and lipid profiles to classify people with normal weight and altered BMI (overweight and obesity) in young people using machine learning (ML) techniques. Machine learning has been used extensively in the healthcare sector [15]. Nevertheless, its use for detecting obesity and overweight in youth is under-explored.

The remainder of the paper is organized as follows. A brief literature review is included in Section 2. Then, we describe the machine learning tool and data selected for conducting our research in Section 3. This is followed by the main results and implications in Section 4. Finally, the last section includes conclusions and further research opportunities.

2. Literature Review

The detection of overweight and obesity has been studied before in literature. Some methods, such as anthropometric assessment has been used to determine nutritional screening indirectly [16]. Moreover, machine learning models, including linear regression, Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Networks (ANN) have been used for this purpose. Among these models for predicting body weight, ANN and RF have shown better performance [17]. Also, machine learning-based clinical decision support systems have been developed to predict people at risk of some diseases where overweight and obesity have a critical influence [18].

Random Forest (RF), with other techniques such as Decision Trees (DT) and XG Boost, have been studied to predict heart diseases [19]. Using XG Boost has improved the models' performance and accuracy [20]. eXtreme Gradient Boosting (XG Boost) was first introduced as a scalable end-to-end tree-boosting system [21]. XG Boost has shown excellent performance in improving the diagnosis of primary lesions related to metastatic tumors [22], for the prediction of breast cancer [23; 24], and as a classifier for stunting among children in Zambia [25].

Despite the advantages of the XG Boost technique, their use for predicting body weight is negligible in the literature. Recently, Santisteban Quiroz [26] studied the performance of this technique in classifying the incidence of obesity based on dietary habits. Therefore, this paper presents an application of the XG Boost method to predict normal weight and altered BMI (overweight and obesity).

3. Method

3.1. Data.

A matrix with 21 variables was used for the lipid (L) and biochemical (B) profiles of 40 Chilean university students with ages 20 – 30 years. The sample selection criteria were their BMI. Also, people who follow any medical treatment for losing weight or athletes were excluded. Table 1 includes the chosen set of variables.

Table 1. Set of variables for lipid and biochemical profiles.

Feature	Variable	Profile type
F0	Basal insulin	Biochemical
F1	Total Cholesterol	Lipid
F2	Total proteins	Biochemical
F3	Albumin	Biochemical
F4	Uric acid	Biochemical
F5	Bilirubin	Biochemical
F6	Phosphorus	Biochemical
F7	Calcium	Biochemical
F8	Urea nitrogen	Biochemical
F9	Glycemia	Biochemical
F10	Alkaline phosphatase	Biochemical
F11	GOT transaminase	Biochemical
F12	LDH U/L (Lactate dehydrogenase)	Biochemical
F13	Alpha 1	Biochemical
F14	HDL (high-density lipoprotein) Cholesterol	Lipid
F15	LDL (low-density lipoprotein) Cholesterol	Lipid
F16	VLDL (very-low-density lipoprotein)	Lipid
F17	Triglycerides	Lipid
F18	Total cholesterol/HDL	Lipid
F19	LF/HF	Lipid
F20	SD1	Lipid

3.2. Classification procedure and validation:

XGBoost was selected as the classifier tool. XGBoost is more accurate than other machine-learning techniques in medical classification [27].
The rest of the procedure was structured according to the phases included in Fig. 1.

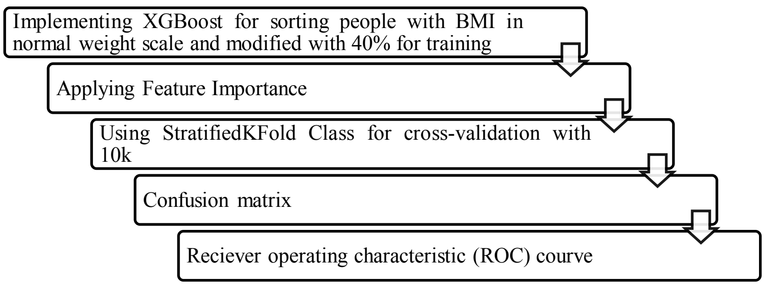

Fig. 1. Classifier implementation procedure.

4. Results

To conduct the analysis, Spyder 4.1.4 – Scientific Python and the toolkit of sci-kit-learn with XGBoost were used in an Intel® Core i5, 8GB RAM, Macbook Pro 13. After applying XGBoost a cross-validation of 80% was obtained.

In figure 2 (a) the best iteration is included with an 87.5% ROC curve and an area under the curve (AUC) of 0.818. Also, figure 2 (b) shows the confusion matrix for the best iteration where the classifier made two false negatives (type

II error), which means two observations where the actual classification was positive, and the predicted classification was negative.

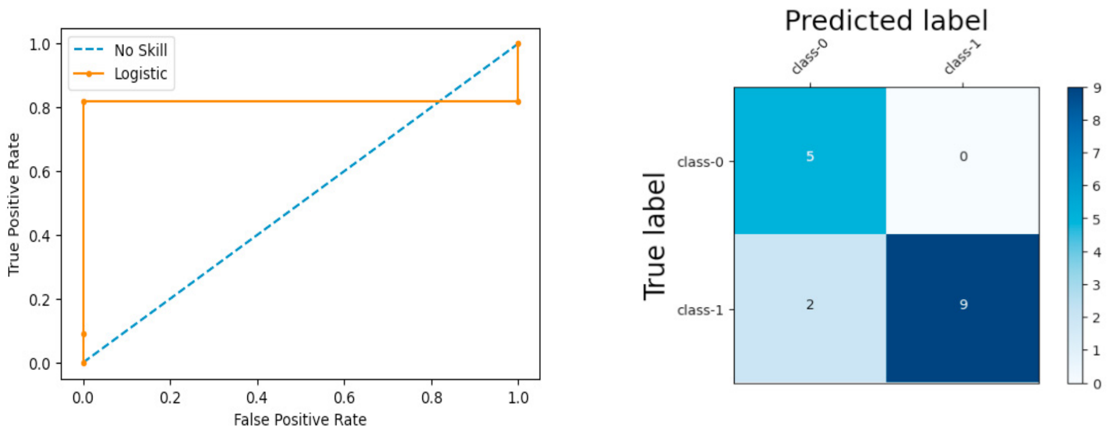

Fig. 2. (a) ROC for the best iteration. (b) Confusion matrix

Table 2. Set of sorted variables according to relevance after 100 iterations.

Feature	F1	F9	F12	F5	F16	F18	F0	F4	F6	F20	F10	F14	F3	F15	F19	F2	F7	F17	F11	F8
Relevance	61	47	37	33	32	29	26	25	23	22	17	16	15	15	14	12	10	10	8	4

Table 2 shows that the variables with greater impact on the classification process were: Total cholesterol (F1), glycemia (F9), LDH (F12), bilirubin (F5), and VLDL (F16). The priority level of each variable is congruent with previous research in which these variables have been used before to model overweight and obesity patterns.

First, the total cholesterol relevance found in our study agreed with the suggestions issued by the AHA (American Heart Association). The AHA recommends doing a total cholesterol assessment for cardiovascular disease prevention, where obesity plays a critical role [28]. However, using total cholesterol to identify people with obesity is still controversial since this variable considers as much HDL as LDL [29].

Moreover, our results suggest that glycemia (F9) has second place in importance. This is a common variable used in obesity studies due to its relationship with energy imbalance and comorbidities such as diabetes and insulin resistance [30]. For its part, LDH enzyme levels change in presence of obesity [31]. Therefore, it explains its 3rd place in importance in our analysis.

The last two variables: bilirubin and VLDL, are key indicators for overweight and/or obesity measurement. Besides, bilirubin has an inverse relationship with overweight and obesity [32]. It means that the higher the body weight, the lower the bilirubin level, and vice versa. Finally, VLDL has been set that contributes 55% to the total content of triglycerides in the blood circulation process [33]. Hence, its level influences weight measurements.

5. Conclusions

This paper has presented an application of the XGBoost tool as a machine-learning technique for youths' classification in two categories: normal weight and altered BMI (overweight and obesity). For this purpose, a set of 40 Chilean university students was selected, and 21 variables were chosen from lipid and biochemical profile tests. The classifier showed good performance with an 80% cross-validation and a confusion matrix suggesting just two false negatives.

The main variables affecting the presence of overweight were: Total cholesterol, glycemia, LDH (Lactate dehydrogenase), bilirubin, and VLDL (very low-density lipoprotein). Then, three variables from the biochemical profile and two from the lipid profile can be used to classify young people in the above-mentioned categories.

Four of these five variables are congruent with previous research. Therefore, further research opportunities include increasing the sample size and being focused on studying bilirubin levels for regulating body weight since the evidence related is limited.

Acknowledgments

This study was supported by ANID FONDECYT under grant number 1221938 and ANID - Millennium Science Initiative Program ICN2021-004.

References

[1] WHO. (2021). *Obesity and overweight*. WHO. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
[2] Fan, H., & Zhang, X. (2022). "Influence of parental weight change on the incidence of overweight and obesity in offspring". *BMC Pediatrics*, *22*(1). https://doi.org/10.1186/s12887-022-03399-8
[3] Nelson, T. D., Haugen, K. A., Resetar Volz, J. L., Zhe, E. J., Axelrod, M. I., Spear Filigno, S., Stevens, A. L., & Lundahl, A. (2015). "Overweight and obesity among youth entering residential care: Prevalence and correlates". *Residential Treatment for Children and Youth*, *32*(2), 99–112. https://doi.org/10.1080/0886571X.2015.1043786
[4] Lafontaine, T. (2008). "Physical Activity: The Epidemic of Obesity and Overweight Among Youth: Trends, Consequences, and Interventions". *American Journal of Lifestyle Medicine*, *2*(1), 30–36. https://doi.org/10.1177/1559827607309688
[5] Skogen, I. B., & Høydal, K. L. (2021). "Adolescents who are overweight or obese-the relevance of a social network to engaging in physical activity: a qualitative study". *BMC Public Health*, *21*(701). https://doi.org/10.1186/s12889-021-10727-7
[6] Banna, J. (2019). "Obesity Prevention in Children in Latin America Through Interventions Using Technology". *American Journal of Lifestyle Medicine*, *13*(2), 138–141. https://doi.org/10.1177/1559827618823320
[7] Cuevas, A., Alvarez, V., & Olivos, C. (2009). "The emerging obesity problem in Latin America". *Expert Review of Cardiovascular Therapy*, *7*(3), 281–288. https://doi.org/10.1586/14779072.7.3.281
[8] Corvalán, C., Garmendia, M. L., Jones-Smith, J., Lutter, C. K., Miranda, J. J., Pedraza, L. S., Popkin, B. M., Ramirez-Zea, M., Salvo, D., & Stein, A. D. (2017). "Nutrition status of children in Latin America". *Obesity Reviews*, *18*, 7–18. https://doi.org/10.1111/OBR.12571
[9] OECD. (2021). *Overweight or obese population (indicator)*. https://www.oecd-ilibrary.org/social-issues-migration-health/overweight-or-obese-population/indicator/english_86583552-en
[10] Vio, F., & Kain, J. (2019). "Descripción de la progresión de la obesidad y enfermedades relacionadas en Chile". *Revista Médica de Chile*, *147*(9), 1114–1121. http://dx.doi.org/10.4067/s0034-98872019000901114
[11] Alghamdi, A. S., Yahya, M. A., Alshammari, G. M., & Osman, M. A. (2017). "Prevalence of overweight and obesity among police officers in Riyadh City and risk factors for cardiovascular disease". *Lipids in Health and Disease*, *16*(79). https://doi.org/10.1186/s12944-017-0467-9
[12] Yin, R., Wang, X., Li, K., Yu, K., & Yang, L. (2021). "Lipidomic profiling reveals distinct differences in plasma lipid composition in overweight or obese adolescent students". *BMC Endocrine Disorders*, *21*(201). https://doi.org/10.1186/s12902-021-00859-7
[13] Pengpid, S., & Peltzer, K. (2015). "Overweight and obesity and associated factors among school-aged adolescents in six pacific island countries in Oceania". *International Journal of Environmental Research and Public Health*, *12*(11), 14505–14518. https://doi.org/10.3390/ijerph121114505
[14] Dunstan, J., Aguirre, M., Bastías, M., Nau, C., Glass, T. A., & Tobar, F. (2020). "Predicting nationwide obesity from food sales using machine learning". *Health Informatics Journal*, *26*(1), 652–663. https://doi.org/10.1177/1460458219845959
[15] Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). "Involvement of Machine Learning Tools in Healthcare Decision Making". *Journal of Healthcare Engineering*, *2021*, 1–20. https://doi.org/10.1155/2021/6679512
[16] Fernández-Juan, A., Ramírez-Gil, C., & van der Werf, L. (2016). "La valoración antropométrica en el contexto de la escuela como medida para detectar y prevenir efectos a largo plazo de la obesidad y del sobrepeso en niños en edad escolar". *Revista Colombiana de Cardiologia*, *23*(5), 435–442. https://doi.org/10.1016/j.rccar.2016.06.007
[17] Babajide, O., Tawfik, H., Palczewska, A., Gorbenko, A., Astrup, A., Martínez, A., Oppert, J. M., & Sørensen, T. I. A. (2020). "A machine learning approach to short-term body weight prediction in a dietary intervention program". *Computational Science – ICCS 2020*, *12140 LNCS*, 441–455. https://doi.org/10.1007/978-3-030-50423-6_33
[18] Du, Y., Rafferty, A. R., McAuliffe, F. M., Wei, L., & Mooney, C. (2022). "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus". *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-05112-2
[19] Mahadevaswamy, U. B., Keerthana, R., Pooja, B. B., Sangatya, V., & Supritha, S. (2022). "A Hybrid Model approach for Heart Disease Prediction". *2022 IEEE 2nd Mysore Sub Section International Conference*, 1–6. https://doi.org/10.1109/MysuruCon55714.2022.9972516
[20] Omkar, M., & Nimala, K. (2022). "Machine Learning based Diabetes Prediction using with AWS cloud". *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 1–7. https://doi.org/10.1109/icses55317.2022.9914160
[21] Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system". *KDD '16: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785
[22] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
[23] Likitha, B., Nakka, J., Verma, J., & Naik, N. S. (2021). "Prediction of Breast Cancer Analysis Using Machine Learning Algorithms and XGBoost Technique". In K. R. Venugopal, P. D. Shenoy, R. Buyya, L. M. Patnaik, & S. S. Iyengar (Eds.), *Data Science and Computational Intelligence* (pp. 298–313). Springer International Publishing.
[24] Huo, L., Tan, Y., Wang, S., Geng, C., Li, Y., Ma, X., Wang, B., He, Y., Yao, C., & Ouyang, T. (2021). "Machine learning models to improve

the differentiation between benign and malignant breast lesions on ultrasound: A multicenter external validation study". *Cancer Management and Research*, *13*, 3367–3379. https://doi.org/10.2147/CMAR.S297794

[25] Chilyabanyama, O. N., Chilengi, R., Simuyandi, M., Chisenga, C. C., Chirwa, M., Hamusonde, K., Saroj, R. K., Iqbal, N. T., Ngaruye, I., & Bosomprah, S. (2022). "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia". *Children*, *9*(7). https://doi.org/10.3390/children9071082

[26] Santisteban Quiroz, J. P. (2022). "Estimation of obesity levels based on dietary habits and condition physical using computational intelligenc e". *Informatics in Medicine Unlocked*, *29*. https://doi.org/10.1016/j.imu.2022.100901

[27] Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring". *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

[28] Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., Ferranti, S. de, Després, J. P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Isasi, C. R., Jiménez, M. C., Judd, S. E., Kissela, B. M., Lichtman, J. H., Lisabeth, L. D., Liu, S., … Turner, M. B. (2016). "Heart Disease and Stroke Statistics—2016 Update". *Circulation*, *133*(4), e38–e48. https://doi.org/10.1161/CIR.0000000000000350

[29] Hwang, S., Kim, H.-E., M.D., J. J., & Kim, H.-J. (2016). "A novel approach for tuberculosis screening based on deep convolutional neural networks". *Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis*, *9785*, 750–757. https://doi.org/10.1117/12.2216198

[30] Hjorth, M. F., Zohar, Y., Hill, J. O., & Astrup, A. (2018). "Personalized Dietary Management of Overweight and Obesity Based on Measures of Insulin and Glucose". *Annual Review of Nutrition*. https://doi.org/10.1146/annurev-nutr-082117

[31] Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). "Identification of risk factors associated with obesity and overweight—a machine learning overview". *Sensors (Switzerland)*, *20*(9). https://doi.org/10.3390/s20092734

[32] Kwon, Y. J., Lee, H. S., & Lee, J. W. (2018). "Direct bilirubin is associated with low-density lipoprotein subfractions and particle size in overweight and centrally obese women". *Nutrition, Metabolism and Cardiovascular Diseases*, *28*(10), 1021–1028. https://doi.org/10.1016/j.numecd.2018.05.013

[33] Nascimento, H., Alves, A. I., Coimbra, S., Catarino, C., Gomes, D., Bronze-Da-Rocha, E., Costa, E., Rocha-Pereira, P., Aires, L., Mota, J., Ferreira Mansilha, H., Rêgo, C., Santos-Silva, A., & Belo, L. (2015). "Bilirubin is independently associated with oxidized LDL levels in young obese patients". *Diabetology and Metabolic Syndrome*, *7*(1), 1–5. https://doi.org/10.1186/1758-5996-7-4/TABLES/2