Prediction of Continuous Targets by Explainable Imbalanced Regression from Omics Data in Childhood Obesity

María Arteaga[1], Álvaro Torres-Martos[2,3], Augusto Anguita-Ruiz[4,2,3], Jesús Alcalá-Fdez[1,5], Rafael Alcalá[1,5,*] and María José Gacto[6,5]

[1]*Dept. of Computer Science and Artificial Intelligence, Univ. of Granada, Granada, Spain*
[2]*Dept. of Biochemistry and Molecular Biology II, "José Mataix Verdú Inst. of Nutrition and Food Technology (INYTA) and Center of Biomedical Research, Univ. of Granada, Inst. de investigación Biosanitaria ibs.GRANADA, Granada, Spain*
[3]*CIBER de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Inst. de Salud Carlos III, Madrid, Spain*
[4]*Barcelona Inst. for Global Health (ISGlobal), Barcelona, Spain*
[5]*Research Inst. in Data Science and Computational Intelligence (DaSCI), Univ. of Granada, Granada, Spain*
[6]*Dept. of Software Engineering, Univ. of Granada, Granada, Spain*

Abstract

Childhood obesity is a persistent challenge for society since it is highly related to insulin resistance and a wide range of other chronic diseases, which impair not only the health of the people, but also the health system itself due to extra costs of treating them. In this context, it is imperative to consider innovative approaches, like eXplainable Artificial Intelligence, to understand the factors underlying childhood obesity. Data imbalance has been thoroughly studied in classification, but scarcely studied in regression since classes do not exist. However, extreme values in continuous domains, which are usually minority are often the most clinically relevant. Facing the imbalance in regression while obtaining explainable models has never been studied. In this application work, we adopt a Machine Learning approach to obtain explainable regression models for imbalanced continuous targets in childhood obesity, as HOMA-IR and Waist Circumference. We consider 79 variables, including targeted metabolomics, targeted proteomics, exposomic data (e.g., the physical activity subdomain), hematological parameters and anthropometry of Spanish children from 3 to 18 years with significant imbalance ratios around 12% in HOMA-IR and Waist Circumference. Even though we mainly focus on extensions of linguistic fuzzy rule-based systems, particularly designed for explainability, we also consider highly accurate complementary approaches as Random Forest that could additionally provide contrast interesting information by post-hoc SHAP. The models so obtained are better considering the relevant minority information (17% improvements in F1). Moreover, they seem to properly explain biologically meaningful relations, as in the case of physical activity data or the one with the follicle stimulating hormone, among others.

Keywords

Childhood Obesity, Insulin Resistance, Accelerometry, eXplainable Regression, Imbalanced Regression

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain
*Corresponding author.
m.arteagajover@gmail.com (M. Arteaga); alvarotorresmartos@gmail.com (Á. Torres-Martos); augusto.anguita@isglobal.org (A. Anguita-Ruiz); jalcala@ugr.es (J. Alcalá-Fdez); alcala@decsai.ugr.es (R. Alcalá); mjgacto@ugr.es (M. J. Gacto)
0000-0001-6774-6219 (M. Arteaga); 0000-0003-3198-2556 (Á. Torres-Martos); 0000-0001-6888-1041 (A. Anguita-Ruiz); 0000-0002-6190-3575 (J. Alcalá-Fdez); 0000-0003-1140-6156 (R. Alcalá); 0000-0001-9895-9647 (M. J. Gacto)

1. Introduction

Childhood obesity entails a persistent challenge for society with consequences in the short and long term. It represents a risk factor for adult obesity, Insulin Resistance (IR) and the metabolic syndrome (metS), which pose an increased risk of cardiovascular diseases, type 2 diabetes, stroke, etc. Furthermore, it is also related to other chronic diseases, such as respiratory disorders, musculoskeletal disorders and liver disease. [1, 2, 3] As well as being harmful for health, childhood obesity is also detrimental for the health system itself due to the additional cost of treating the other disorder it causes.

Thus, the need for innovative and personalized approaches for the prevention and treatment of childhood obesity arises. This research involves the analysis of omics data (genomics, epigenetics, exposomics, metabolomics, etc.), and advanced tools, such as Machine Learning (ML), which enable the integration of these data and provide a more holistic vision of the issue. It is evident but crucial that the models generated in health-related issues must not only predict the variables of interest, but also provide detailed insights on the underlying factors. Following this idea, eXplainable Artificial Intelligence (XAI) [4] becomes an essential aspect.

In addition, real-world problems frequently involve imbalanced data due to limitations in data collection, the scarcity of cases, etc. For instance, in a study on childhood obesity, while most children may present a slightly elevated Body Mass Index (BMI), there may be a lack of children who are severely obese, the most clinically relevant. A considerable imbalance seriously impacts and biases the models, so it must be treated. This problem has been extensively studied in classification, but scarcely studied in regression since classes do not exist.

Some of the current proposals are pre-processing techniques, such as the over-sampling method SMOTE for regression [5], which is based on nearest neighbors, and its extension, SMOGN [6], which incorporates Gaussian noise. Other proposals are modified algorithms that address the imbalance problem within the learning process itself, such as the modifications of the linguistic algorithm FSMOGFS^e+TUN^e [7] (here referred to as LING for simplicity) and the purely approximate method METSK-HD^e [8], detailed in [9]. However, even in the case of the linguistic algorithm, interpretability aspects were not considered, so that it presents severe semantic problems (linguistic terms extreme overlapping, significant rule inconsistency, etc.).

In this application work, we explore a pioneering approach that, for the first time, concurrently addresses both explainability and imbalanced regression. Our aim is to develop interpretable regression models for two imbalanced continuous targets in childhood obesity: The HOMA-IR and Waist Circumference z-scores (zHOMA-IR and zWC, respectively). We mainly focus on the extension of a linguistic fuzzy rule-based system, originally called LING$_{CFLTS}$ [10] (Composed Fuzzy Linguistic Term Set based approach), which would be able to provide more interpretable information since it is particularly designed for explainability. For simplicity, we will call the original algorithm LING$_{eXplainable}$, and our extended proposal LING$_{eXplainable}$-ImbR. Alternatively, in this contribution, we also explore the use of Random Forest [11] combined with SMOGN pre-processing as complementary approach to contrast the results obtained in the proposed algorithm. While Random Forest generally outperforms the fuzzy rule-based models, it provides less directly interpretable information, relying solely on post-hoc SHapley Additive exPlanations (SHAP) analysis [12]. But contrasting the information obtained from both approaches could be still interesting.

We consider 79 variables, including targeted metabolomics (e.g., cardiometabolic and hormonal biomarkers), targeted proteomics (e.g., inflammatory and hepatic biomarkers), some subdomains of the exposome (e.g., the physical activity subdomain, obtained by accelerometry, family history, and socioeconomic and demographic factors), as well as hematological parameters and anthropometry, of Spanish children from 3 to 18 years. Although all children are obese or overweight, zHOMA-IR and zWC present significant imbalance ratios around 12% due to a lack of representation of exceptionally high values, which are in fact the most clinically relevant cases. The said modification of the linguistic algorithm FSMOGFS^e+TUN^e proposed in [9] for actively addressing the imbalance will be also used for comparison in terms of Mean Square Errors (MSE) and imbalance consideration.

The models generated through the proposed approaches exhibit notable improvements in considering the relevant minority information, resulting in F1 [1] increases over 17%. These models effectively enlighten biologically significant relationships among the variables and the predicted targets, as in the case of physical activity data or the one with the follicle stimulating hormone (FSH), among others.

This work is organized as follows. In Section 2, we explore the concept of imbalanced regression and its implications in real-world problems. In Section 3, we focus on childhood obesity prediction, as we describe the case study (Section 3.1), the challenges of omics and data pre-processing (Section 3.2) and the importance of considering imbalance in this particular regression problem (Section 3.3). In Section 4, we present our linguistic rule-based system for explainable imbalanced regression. Section 5 describes the experimental design, while Section 6 presents the performance results obtained in both problems (Section 6.1) and explores the relevant variables and explains a model example (Section 6.2). Finally, Section 7 discusses the findings, limitations and future perspectives.

2. Imbalanced Regression in Real-World Problems

In classification, imbalance refers to an uneven representation of classes. As minority groups are easily identifiable, this issue has been thoroughly studied. However, in regression, imbalance refers to a skewed distribution of the data in the target variable, this is, an underrepresentation of certain specific subdomains, often those clinically relevant. Since classes do not exist, there is scarce emphasis on the need for a complete representation of the continuous variable domain. Identifying subdomains seems challenging, as they are usually not uniformly relevant and there is typically poor representation of the relevant subdomains, so it has been hardly studied.

Real-world problems frequently involve imbalanced data due to various factors, such as limitations in data collection or the scarcity of cases. For instance, in a study on childhood obesity, most children fmay present a slightly elevated Body Mass Index (BMI); in a study on type 2 diabetes, the majority of the patients may present moderately elevated levels of glucose; in a study in Chronic Obstructive Pulmonary Disease (COPD), most patients present a moderate reduction in pulmonary capacity. In all these examples, there is a lack of representation of the most clinically relevant cases (extreme values): children who are severely obese, patients with significantly elevated glucose levels, and patients with severe reductions in pulmonary capacity,

[1] Adaptation of the well-known F1 for classification.

respectively. A considerable imbalance seriously impacts the models, biasing towards the most frequent values and limiting generalization to underrepresented cases, so it must be treated.

In order to formally introduce the concept of imbalanced regression [6, 13], let us contemplate an unknown function $f(X_1, X_2, ..., X_p)$, where p denotes the number of predictor variables, aiming to approximate the output variable's defining function with maximal accuracy and proximity. Next, we possess a training dataset $D = \{\langle x_i, y_i \rangle\}_{i=1}^{n}$ with n instances from which we will derive our approximate function f, where x and y are the inputs and the continuous output values, respectively. To take into account a possible imbalance in Y, we need to define a relevance distribution function $\Phi(Y)$, so we can create a "minority" subset D_r of relevant data (i.e. those instances with a relevance value above t_r), as well as a "majority" subset D_n of non-relevant data (those with relevance values less than or equal to t_r). $\Phi(Y)$ is based on the concept of extreme Y values and established in the following sigmoid function by Torgo and Branco et al. as follows:

$$\Phi(Y) = \frac{1}{1 + e^{-s*(Y-c)}} \tag{1}$$

where c is the center of the sigmoid, that is, the value where $\Phi(Y) = 0.5$, and s is the shape of the sigmoid (see [13] for more details). Since both, low extreme values and high extreme values could exist, $\Phi(Y)$ is defined with two different sigmoid functions. Moreover, they also adapted some of the metrics used in imbalanced domains, as the well-known F-Measure, F1 when $\beta = 1$ (harmonic mean of *Recall* and *Precision*):

$$F = \frac{(\beta^2 + 1) * Precision * Recall}{\beta^2 * Precision + Recall} \tag{2}$$

This adaptation draws upon well-established classification concepts: *Recall* ($\frac{TP}{TP+FN}$ where TP represents *True Positives* and FN represents *False Negatives*) and *Precision* ($\frac{TP}{TP+FP}$ where FP represents *False Positives*). To achieve this, they establish a non-trivial connection between instances exhibiting an acceptable margin of error (effectively classified in classification problems) and their relevance, so that they can further quantify those relevant bad predictions (see [13] for further insights into this adaptation).

As stated previously, imbalanced regression has been hardly studied. The few existing techniques can be categorized as [9]:

- Pre-processing techniques: These re-sampling techniques balance data distribution via pre-processing so that algorithms focus on the most relevant instances. Their main advantage is that they can be applied to any existing algorithm, whereas their main disadvantages are that they are dependent on the data and sensitive to the data quality and parameters adjustment. These could be considered "passive" techniques since they do not affect the learning process.
- Algorithmic approaches: These are modified algorithms that address the imbalance problem within the learning process itself. These could be considered "active" techniques since they affect the learning process.

Within “passive” techniques, SMOTE for regression is outlined [5]. It is based on SMOTE (Synthetic Minority Over-sampling Technique) [14], an over-sampling method for classification that generates synthetic instances by interpolating between minority class samples based on their nearest neighbors and is usually combined with under-sampling. In SMOTE for regression, synthetic instances are generated by considering the regression line between neighboring minority class samples. SMOGN (Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise) [6] is an extension of SMOTE for regression that introduces Gaussian noise when interpolating between samples, so it preserves the diversity of the data. Regarding "active" approaches, the state-of-the-art is limited. We must highlight two algorithms based on fuzzy rules: a linguistic algorithm called FSMOGFS^{e}+TUN^{e} [7] (in this work referred to as LING for simplicity) and an approximate algorithm called METSK-HD^{e} [8], which were modified to actively address imbalance [9].

As said, even in the case of the linguistic algorithm, it was not particularly designed to provide fully transparent and therefore explainable regression models. In this work, we extend LING$_{CFLTS}$ [10] (Composed Fuzzy Linguistic Term Set based approach, in this work referred to as LING$_{eXplainable}$), which does not natively handle imbalanced regression, to actively address the imbalance, since this algorithm was particularly designed to obtain interpretable/transparent linguistic models.

3. Childhood Obesity Prediction Problem

In this section, we will delve into the challenge of predicting continuous variables related to pediatric obesity in overweight and obese children. We will begin with a description of the case study and predictive targets, followed by a discussion on the challenges in omic data analysis and pre-processing. Finally, we will explore the relevance of considering imbalanced regression.

3.1. Description of the Case Study and Continuous Prediction Targets

Childhood obesity presents an enduring challenge for nowadays society given its strong relation with insulin resistance, metabolic syndrome, and numerous other chronic diseases. These not only affect the health of the population, but also the health system itself due to the additional costs of treating them. In this case study, we consider a database from national project of the Carlos III Health Institute [15]. It corresponds to the IBEROMICS project [16], which focuses on overweight or obese children between 3 and 18 years old in Santiago de Compostela and Zaragoza. Since we aimed to provide a comprehensive analysis that captures potential interactions and influences on the target variables, we consider 79 variables of interest, including targeted metabolomics (e.g., cardiometabolic and hormonal biomarkers), targeted proteomics (e.g., inflammatory and hepatic biomarkers), some subdomains of the exposome (e.g., the physical activity subdomain, obtained by accelerometry, family history, and socioeconomic and demographic factors), as well as hematological parameters and anthropometry. This approach let us explore the relationships between a wide range of variables. Fuzzy rule-based algorithms ensure that only the most relevant variables contribute to each rule, maintaining model interpretability and relevance, while Random Forest are known for their ability to handle high-dimensional data and their robustness to irrelevant features.

We established two variables as continuous prediction targets in childhood obesity: zHOMA-IR (n = 190) and zWC (n = 204). HOMA-IR (Homeostasis Model Assessment for Insulin Resistance) is an indicator of insulin resistance, which is strongly related to obesity since fat cells release biochemical substances that can interfere with the normal insulin response. Waist Circumference is an indicator to evaluate central obesity and visceral fat distribution. It is related to insulin resistance since visceral abdominal fat is metabolically active and produces several biochemicals, including adipokines and cytokines, that may contribute to it. A z-score is a statistical measure that represents the number of standard deviations a data point is from the mean of a reference population dataset: negative z-scores indicate below-average values and positive z-scores indicate above-average values, with the absolute value representing the deviation from the mean. This standardization ensures that comparisons are fair and meaningful, especially when analyzing variables where values can vary significantly based on age and height, as in this case. The HOMA-IR z-score was calculated using Stavnsbo as reference (see [17] for more details), whereas the Waist Circumference z-score was calculated using Ferrández as reference (see [18] for further insights).

This case study mirrors a real-world health-related scenario, necessitating interpretability (models must offer not just variables but also meaningful explanations) and a consideration of imbalanced data. This dual focus represents a novel approach.

3.2. Challenges in Omics ML and Data Pre-processing

The application of ML to Omics data implies unique challenges in pre-processing and analysis. Additionally, to ensure robust and reliable models, it was crucial to effectively handle missing values, noise, outliers and batch effects, among others.

In this case study, we conducted data cleaning to ensure the reliability of the dataset. Raw accelerometry data were transformed using the Actilife v.6.13.3 software into variables of interest, such as the number of steps or measures of sedentary lifestyle and exercise of different levels corresponding to weekdays, weekends, or in total. Also, several variables (e.g., HOMA-IR, BMI, triglycerides, blood pressure, etc.) were transformed to z-scores. Specifically, these variables are WC ([18]), HOMA-IR, HDL and TG ([17]), BMI ([19]), DBP and SBP ([20]). Additionally, to ensure robust and reliable models, it was crucial to effectively handle missing values, noise, outliers and batch effects, among others. Categorical variables were transformed by label encoding. Variables with more than 15% missing values were removed, followed by a non-parametric missing value imputation using the MissForest algorithm, based on Random Forest [21]. An exploratory analysis ensured that this imputation did not significantly affect data distribution.

3.3. Regression Imbalance Consideration

Real-world problems frequently involve imbalanced data due to various factors, such as limitations in data collection or the scarcity of cases. A considerable imbalance seriously impacts the models, biasing towards the most frequent values and limiting generalization to under-represented cases, so it must be treated. For example, if the majority of the children of a hypothetical dataset had an average body mass index (BMI), the model would be biased towards normal-weight children.

According to Section 2, the imbalance ratio is calculated as the percentage of instances with a phi(Y) value over 0.8, that is,

$$\frac{\text{number of instances with } \Phi(Y) > 0.8}{\text{total number of instances}} \cdot 100$$

In this case study, although all the children are overweight obese, there is still a lack of representation for those with exceptionally high values in zHOMA-IR and zWC, which present significant imbalance ratios ($\mathbf{12.105}\%$ and $\mathbf{12.745}\%$, respectively). This means that children with severe insulin resistance and children with a severe accumulation of abdominal fat, which are in fact the most clinically relevant cases, are underrepresented. Since it may appreciably affect and bias the models, the imbalance existent in these datasets must be considered for obtaining not only transparent/interpretable regression models, but also more reliable ones.

4. Linguistic Rule-based System for explainable imbalanced regression

In [10], the authors propose a novel method for evolutionary regression, which aims to predict numerical outcomes based on input variables while maintaining simplicity and transparency/interpretability, measured by two semantic interpretability indexes, Geometric Mean of 3 complementary Metrics (GM3M) and Rule Meaning Index (RMI). As stated before, in this work this method is referred to as LING$_{eXplainable}$. GM3M and RMI are two well-known interpretability measures from the specialized literature that account for the preservation of the initial linguistic terms definitions and/or the rule meaning/consistency, respectively.

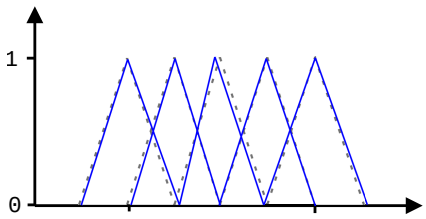

Figure 1: Illustration of linguistic partition with a GM3M value of 0.8 (blue) in comparison to the corresponding strong fuzzy partition (gray).

GM3M measures (ranging from 0.0 to 1.0) the semantic interpretability at the level of linguistic partitions, where values close to 1.0 represent a really high proximity to the membership functions associated with the initial linguistic terms (in our case, equally distributed strong linguistic partitions). See an example of linguistic partition with GM3M equal to 0.8 in Figure 1. RMI measures (also raging from 0.0 to 1.0) the semantic interpretability at the rule base level, with values close to 1.0 representing the absence of inconsistent rules (i.e., the output of the rule is consistent with the whole model output in the corresponding rule core domain). GM3M or RMI associated to a whole model are computed as the worst case from all the linguistic terms or rules, respectively. Thus, having GM3M=0.7 and RMI=0.8 for one regression model means

that all membership functions have GM3M equal or higher than 0.7 and that all the rules have RMI values equal or higher than 0.8.

The algorithm combines two key elements: A new linguistic fuzzy grammar and an interpretable linear extension. This approach helps to effectively reduce the number of rules, maximize the semantic interpretability (both GM3M and RMI), minimize errors, and maintain the rule length (number of conditions) within reasonable bounds. It operates on a two-stage tree-based hybrid evolutionary multiobjective algorithm:

- First stage: This involves learning the initial linguistic partitions and associated rules. Here, an embedded multiobjective evolutionary learning of the linguistic partitions (number of linguistic terms and their definitions) is employed, aiming to minimize both number of rules and errors measured as the $MSE_{/2}$ (MSE divided by 2). It utilizes a multiobjective evolutionary approach to learn the linguistic partitions and applies a rapid method to derive rule sets for each definition of the linguistic partitions. It builds upon a hybridization with an innovative linguistic tree-based rule learning technique, extending the renowned M5-prime method to derive rule sets for each evolved linguistic partitioning.
- Second stage: This stage entails post-processing to further refine the learned solutions. It employs an advanced multiobjective evolutionary algorithm to fine-tune membership functions and for rule selection. This refinement process aids in minimizing number of rules, maximizing linguistic interpretability (GM3M and RMI), and reducing errors ($MSE_{/2}$) in the initial global structure obtained in the first stage, which was initially based on strong fuzzy partitions.

Unlike other methods, the adaptability of evolutionary algorithms allows them to adjust to specific problems in a versatile manner. Leveraging this characteristic, in this contribution, we introduce a solution called LING$_{eXplainable}$-ImbR, that guides the algorithm towards the said outcomes (number of rules, GM3M, RMI and $MSE_{/2}$) but by also considering the imbalance through a refined evaluation function enriched with additional information. The idea is to replace $MSE_{/2}$ in the second stage by an objective function maximizing the performance obtained in the set of relevant data D_r (guided by the relevance function), without unbalancing the performance that we obtain in the rest of the data D_n (guided by $MSE_{/2}$). In order to do this, we define a Weighted MSE ($MSE^W_{/2}$) attending to the following equations:

$$d_w = \Big(\big(\sum_{i=1}^{|D_r|} a + \Phi(y_i) \big) + |D_n| \Big). \tag{3}$$

d_w represents a weight factor that combines the contributions from the relevant and non-relevant data, since $\sum_{i=1}^{|D_r|} a + \Phi(y_i)\big)$ is the sum of a constant value (a) and the relevance function ($\Phi(y_i)$), and D_n is the number of instances in the non-relevant set. Based on another work [9], constant a has been set at 5, since it was the value that obtained the best results.

$$SE_{D_r} = \sum_{i=1}^{|D_r|} (F(x^i) - y^i)^2 * (a + \Phi(y_i)), \tag{4}$$

where SE_{D_r} is the sum of the squared errors for the relevant data, weighted by $a + \Phi(y_i)$.

$$SE_{D_n} = \sum_{i=1}^{|D_n|} (F(x^i) - y^i)^2, \tag{5}$$

and SE_{D_n} is the sum of the squared errors for the non-relevant data, without additional weighting.

$$MSE_{/2}^{W} = \frac{SE_{D_r} + SE_{D_n}}{d_w * 2} \tag{6}$$

where $MSE_{/2}^{W}$ is the weighted MSE, normalized by d_w.

5. Experimental design

In this work, we mainly focus on the proposed extension of the linguistic fuzzy rule-based system LING$_{eXplainable}$ [10], which could provide more interpretable information since it is particularly designed for explainability. To evaluate the performance of this algorithm, we apply various algorithms (Random Forest [11], LING [7], and LING$_{eXplainable}$ [10]) to the stratified partitions from the original dataset, so that imbalance is not considered, and again after applying SMOGN to the training sets, so that imbalance is considered via pre-processing the data. Moreover, we employ the modified version of LING [9] (LING-ImbR) and our proposal for LING$_{eXplainable}$, that actively address data imbalance (see Section 4), to the stratified partitions from the original dataset. Finally, the methods considered in the experiments are summarized in Table 1. Except for the number of trees in Random Forest, which was established to 500, the default parameters suggested by authors were used for all the algorithms.

Table 1
Methods considered in the experiments. MOEA: Multi-Objective Evolutionary Algorithm, TS: Tuning and Selection, ImbR: Imbalanced Regression, *: Proposed here.

Method	Ref.	Description
SMOGN+?	[6]	Preprocessing plus other method (?): *SMOTE for regression* with the introduction of Gaussian noise
RF	[11]	Random Forest Regressor
LING (FSmogfs^e+Tun^e)	[7]	MOEA for embedded DB Learning, RB wrapper generation and TS MOEA
LING-ImbR	[9]	MOEA for embedded linguistic partitions learning, rule base wrapper generation and TS MOEA & ImbR
LING$_{eXplainable}$	[10]	eXplainable-based MOEA for embedded learning of transparent linguistic partitions with wrapper linguistic tree-based rule base generation and TS MOEA
LING$_{eXplainable}$-ImbR	*	LING$_{eXplainable}$ & ImbR

Since error estimation depends on the training and test data, we used a repeated 10-fold stratified cross-validation to reduce the loss of diversity when partitioning the data and preserve the minority set representation. For each dataset (the dataset with zHOMA-IR as target and the one with zWC), we use a 10-fold cross-validation strategy, that is, the original data is divided into 10 subsets or folds. This is done two times. Thus, we created 20 partitions per dataset.

The stratification used in these regression problems consisted in the discretization of the continuous target variables into in c cuts, that is, a specified number of bins which result from dividing the number of examples of our dataset by the number of desired partitions. However, it must be applied separately to relevant and non-relevant data. In order to achieve this, based on another work [9], the threshold t_r for the relevance function was established at 0.8, so that higher values were considered relevant, and lower values, normal. Once relevant data were identified, stratification was applied to the relevant and non-relevant data separately. For each set (relevant and non-relevant), we divided the data into c cuts, ensuring that each bin had approximately the same number of examples (equal-frequency bins). After creating the bins, we randomly and proportionally distributed the relevant and non-relevant instances from each cut among the relevant and non-relevant data partitions, respectively, to ensure balanced representation. Finally, both relevant and non-relevant data partitions were joined.

Also, we need to consider metrics to properly evaluate the algorithms performance. In this study, we mainly consider two metrics: the said adaptation of F1 for regression (see Section 2) and the Mean Square Error (MSE). F1 is a metric that allows us to evaluate how well are the models addressing the imbalance problem. MSE is a classical metric to evaluate accuracy. It measures the average squared difference between the predicted and the actual values. Considering both will help us assess the models performance not only on the relevant data but also on the overall set.

6. Main Results and Insights: zHOMA-IR and zWC

In this section, we will delve into the performance results obtained in both problems (zHOMA-IR and zWC). Also, we will explore the relevant variables and explain a model example.

6.1. Performance results obtained in both problems

Firstly, we will compare the results for all the algorithms, which are shown in Table 2. Each row corresponds to an algorithm, with those actively addressing imbalanced marked in grey. The table is divided in two sections, one per problem (zHOMA-IR and zWC). Within each section, for each algorithm, the first column shows the number of variables (when applicable) and the second column, the number of rules (when applicable). The third and forth columns show the average MSE and F1 values in the test sets, respectively.

The results show that when SMOGN is applied, the F1 for all the algorithms values are better compared to those obtained with the original data. This means that addressing imbalance through the previously mentioned passive imbalance pre-processing technique improves the performance in the relevant set. However, the results show that applying SMOGN implies a cost in MSE, that is, an increased error in predicting our continuous outcomes. The results obtained in F1 by the algorithms that actively address imbalance within the learning process

Table 2
Results obtained by the studied algorithms. MSE and F1 values are the average errors calculated on the corresponding test sets (generalization). Those actively considering imbalanced regression are marked in grey.

	zHOMA-IR				zWC			
Methods	**#Vars.**	**#Rules**	**MSE$_{tst}$**	**F1$_{tst}$**	**#Vars.**	**#Rules**	**MSE$_{tst}$**	**F1$_{tst}$**
RF	-	-	**1.018**	0.00	-	-	**0.944**	0.55
SMOGN+RF	-	-	1.241	0.45	-	-	1.210	0.71
LING	6.7	23.5	1.662	0.09	3.9	8.7	1.200	0.48
SMOGN+LING	5.8	27.7	2.905	0.15	4.8	20.5	1.931	0.49
LING-ImbR	6.7	21.7	2.070	0.12	3.9	8.0	1.402	0.52
LING$_{eXplainable}$	3.4	6.7	1.804	0.36	**2.2**	4.5	1.329	0.62
SMOGN+LING$_{eXplainable}$	**2.8**	7.0	3.647	**0.50**	2.4	5.7	2.217	0.73
LING$_{eXplainable}$-ImbR	3.4	**6.4**	2.305	0.40	**2.2**	**4.3**	1.329	**0.75**

(those marked in gray) are slightly worse or better (depending on the problem) than those obtained when using SMOGN, whereas MSE values are notoriously better, similar to the results obtained without considering the imbalance, specially for zWC. LING$_{eXplainable}$-ImbR presents the best values for F1 in both problems. While Random Forest outperforms the fuzzy rule-based models in MSE in both problems, it provides less detailed information, relying solely on post-hoc SHapley Additive exPlanations (SHAP) analysis. For both linguistic algorithms, the number of variables and rules are similar in the original algorithm, the SMOGN approach and the extended version. Nevertheless, the number of variables and rules is lower in LING$_{eXplainable}$ compared to LING, which contributes to the interpretability of the model, since LING$_{eXplainable}$ is particularly designed for explainability. Furthermore, the errors are in 10 to 15% of the output domains, so the models explain from 85 to 90% of the domain.

Table 3
GM3M and RMI: Interpretability metrics obtained by the linguistic fuzzy algorithms

	zHOMA-IR		zWC	
Methods	**GM3M**	**RMI**	**GM3M**	**RMI**
LING	0.2	0.3	0.2	0.5
SMOGN+LING	0.2	0.3	0.2	0.4
LING-ImbR	0.2	0.4	0.2	0.5
LING$_{eXplainable}$	**0.7**	**1.0**	**0.8**	**1.0**
SMOGN+LING$_{eXplainable}$	**0.7**	0.9	0.7	1.0
LING$_{eXplainable}$-ImbR	**0.7**	**1.0**	0.7	**1.0**

Secondly, we will compare the linguistic algorithms. Table 3 shows the interpretability metrics obtained by the linguistic fuzzy algorithms. Each row corresponds to an algorithm. The columns are also divided in two sections: the results for the zHOMA-IR and the zWC dataset. In each section, the first columns represents the GM3M value and the second column,

the RMI value. As we stated previously, GM3M measures the semantic interpretability at the level of linguistic partitions, whereas RMI measures the semantic interpretability at the rule base level. Thus, having GM3M=0.7 and RMI=1.0 for one algorithm in the table means that, in average the 20 runs, all membership functions have GM3M equal or higher than 0.7 and that all the rules have RMI values equal to 1.0 (all the rules in the 20 runs are fully consistent rules). For both zHOMA-IR and zWC, LING, LING using SMOGN and the modified version of LING have consistently low GM3M and RMI values, indicating the existence of significant interpretability problems. In contrast, LING$_{eXplainable}$, LING$_{eXplainable}$ using SMOGN and LING$_{eXplainable}$-ImbR show quite high GM3M and RMI values, indicating a very high level of transparency. GM3M values for this algorthim, which are close to 1.0, represent a high proximity to the membership functions associated with the initial linguistic terms. RMI values, which are even 1.0 in the case of LING$_{eXplainable}$ and LING$_{eXplainable}$-ImbR, represent the absence of inconsistent rules.

Table 4
Variables that appear in the models shown by importance: More to less times used #T and best to worst SHAP ranking/position respectively. Those related to physical activity are marked in blue.

zHOMA-IR			zWC		
	LING$_{eXplainable}$-ImbR	SMOGN+RF		LING$_{eXplainable}$-ImbR	SMOGN+RF
#T	Vars.	Vars.	#T	Vars.	Vars.
15	zBMI	zTriglycerides	20	zBMI	zBMI
13	Adiponectin-leptin ratio	MCH	18	Age	LightWEd
11	TSH	SedentaryWd	11	FSH	LH
9	Urea, Calcium	zBMI	7	%Fat_mass	cpmWd
8	Age, Origin, StepsWd, zHDL, zTrygliceride	LDL	3	Creatinine, Erythrocytes, Light, Sex	Iron
		zWC	2	%ACT, ALT, GGT, Haemoglobin,	FSH
7	Sedentary, zWC	MCV		LightWEd, Moderate, Origin,	Age
6	Tanner_stage	Calcium		Tanner_stage, Testosterone, zSBP,	Cortisol
5	Cortisol, SedentaryWEd	zHDL		zTriglycerides	zDBP
4	CpmWd, Haemoglobin, ModerateWd,	ALP			Monocytes
	#days, SedentaryWd	Total_Bilirrubin			Erythrocytes
3	LDLc, LightWEd, mvpaWd, MCV	Creatinine			Calcium

6.2. Relevant Variables and Explained Model Example

Table 4 shows the variables that appeared in the models shown by importance. For LING$_{eXplainable}$, variables are shown from more to less times used (#T), whereas for Random Forest, variables are shown from more to less importance for the model output based on SHAP rankings.

In the zHOMA-IR dataset, it is noteworthy that all but one of the rule bases include a variable associated with accelerometry (see those in blue in Table 4). However, in the zWC dataset, accelerometry variables are present in only half of the generated rule bases, suggesting that these variables may not always be significant. This suggests complex factors beyond physical

activity influence in waist circumference during childhood obesity. Further analysis and context exploration are necessary for deeper insights into this fact.

Moreover, the rules for zWC consistently incorporate the variable zBMI in each line. However, in the zHOMA-IR dataset, the variable zBMI appears 15 times, while the variable Adiponectin-leptin ratio appears 13 times, indicating the presence of specific rule bases containing both zBMI and Adiponectin-leptin ratio. The Adiponectin-leptin ratio has been associated with insulin resistance in overweight and obese children [22]. The thyroid-stimulating hormone (TSH) appears 11 times, as abnormal levels of TSH seem positively correlated to insulin resistance [23] independently of the body status [24].

SHAP rankings let us contrast to the most variables used in Random Forest models. For the zHOMA-IR problem, the z-score of triglycerides is the most important variable in the Random Forest models, followed by the mean cell hemoglobin (MCH), an accelerometry measure, zBMI, which is the most repeated variable for the linguistic models, etc. For the zWC problem, the most important variables for the output are zBMI, some accelerometry measures, the follicle stimulating hormone (FSH), which is the third most repeated variable for the linguistic models, etc.

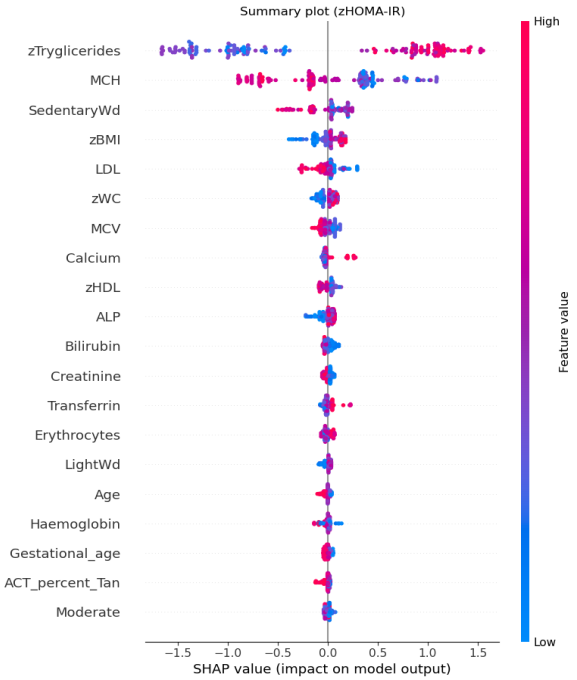

Figure 2: Top 20 variable's contribution in the SMOGN+RF model in the zHOMA-IR problem (SHAP)

Figure 2 represents a summary graph for a model generated via Random Forest in the zHOMA-IR dataset. It shows the contribution of the 20 most important variables, ordered from more important (top) to less important (bottom). Each point represents an instance, where its horizontal position indicates the impact that the feature has in the model output (positive or negative), and its color indicates the feature value (blue for low values, red for high

values). Triglycerides, MCH, and SedentaryWd are the most influential features for predicting zHOMA-IR. Thus, both clinical measures (e.g., lipids, blood counts) and lifestyle factors (e.g., sedentary physical activity) are important, demonstrating the multifactorial nature of insulin resistance. Nevertheless, other characteristics with lower SHAP values contribute cumulatively, emphasizing the need for a broader view in the prediction and management of insulin resistance.

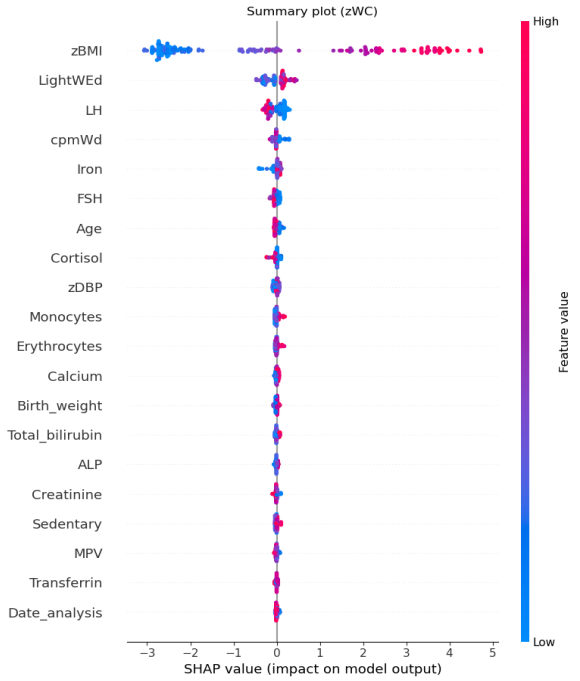

Figure 3: Top 20 variable's contribution in the SMOGN+RF model in the zWC problem (SHAP)

Figure 3 represents a summary graph for a model generated via Random Forest in the zWC dataset. This graph provides an overview of how features affect the model predictions, with zBMI being notoriously the variable with the highest impact on the model output. However, this does not imply that the other variables are irrelevant, but that their individual impacts are smaller compared to zBMI. This means that these features contribute to the model through their cumulative effects, even if their individual SHAP values are smaller. Furthermore, features can interact with each other in complex ways, so a small individual SHAP value might still be crucial when considered in combination with other features.

Figure 4 represents an example of a linguistic model obtained for zWC. The arrangement of variables in the figure corresponds to the sequence of splits generated in the decision tree during the rule-learning process. Each split is a division of the data, from more general to more specific in each split. Colors are used to aid in distinguishing the various cases depicted by the rules, with each split and variable sharing the same color. Gray text elements are not part of the rules, but provide extra information, as well as percentages of covered instances and the GM3M and RMI values. These values provide insights into the semantic quality of each partition and rule, respectively. Notably, all rules exhibit an RMI value of 1.0, indicating that any

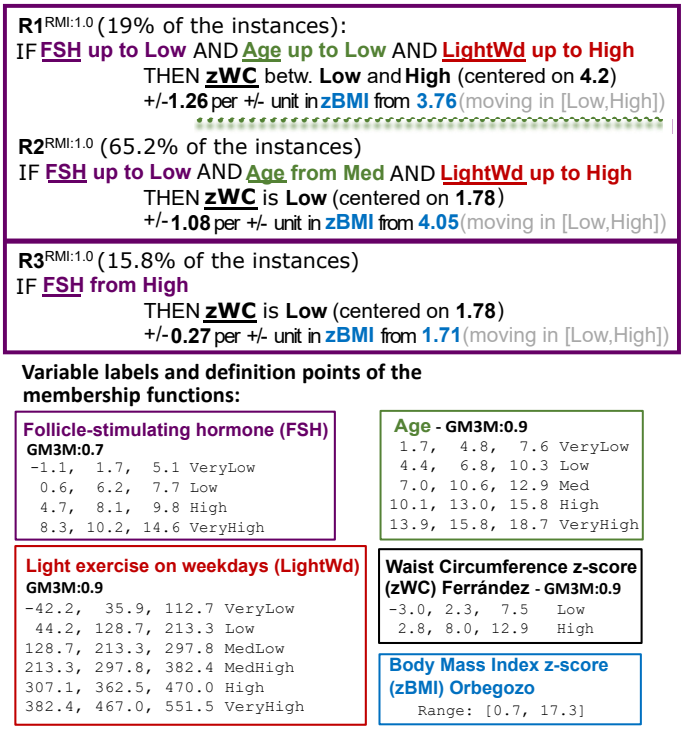

Figure 4: Linguistic model obtained by LING$_{eXplainable}$-ImbR in the zWC problem. MSE$_{tst}$ = 0.572, F1$_{tra}$ = 0.85 and F1$_{tst}$ = 0.76.

single rule assert actually matches the real output of the model. Moreover, all variables display GM3M values exceeding 0.7, and even reaching 0.9, suggesting equidistant strong linguistic partitions. The MSE in test (0.572) is almost a half of the one obtained by RF in average without imbalance consideration (0.944) while F1 in test is over 0.75. Consequently, it seems that the method effectively characterizes the overall behavior of the dataset and actually improves the consideration of the relevant data, which could help to appropriately discern the principal relationships among predictor variables.

The model was able to accurately retrieve the zWC of children with obesity using just 4 variables (which included the Follicle stimulating hormone (FSH) blood concentrations, the child age (decimal age), the amount of light physical activity, and the zBMI). Interestingly, despite the presence of clear zWC indicators among predictors (such is the case of zBMI), the model chose a non-typical metabolic risk marker to be placed in the first level of division (at the rule tree-based generation splitting) for each rule; the FSH hormone. FSH showed an intriguing and non-obvious relationship with the zWC in this population. As can be seen in rule R3, high and very high ("from High") values of FSH were directly linked to low zWC values, with a zBMI-mediated effect modification of +/- 0.27 per unit +/-. Individuals with higher concentrations of this hormone will therefore tend to present less central obesity, with a small contribution of increases in their zBMI.

In rules R1 and R2, on the other hand, the opposite zWC-FSH relationship was evidenced. While in R1, "up to Low" values of FSH were linked to medium ("between Low and High") zWC

(as expected from R1), in R2, these same values (up to low values of FSH) were linked to "Low" zWC, with child's age jumping on the scene as an effect modifier in this case (i.e., younger children with up to low FSH tend to present "between Low and High" zWC, while medium aged and older children with up to low FSH present low zWC). Interestingly, the cutoffs in the age variable proposed by the method for generating the fuzzy intervals that differentiate these two rules expand exactly in the estimated ages for normal puberty onset, suggesting a plausible involvement of the sexual maturation procedure in the FSH-zWC relationship.

Scientific literature has extensively pointed FSH as a significant player in human metabolic disorders. Epidemiological studies have established a strong correlation between FSH levels and metabolic diseases, while experimental research has delved into the underlying mechanisms both in vivo and in vitro [25]. From in vitro and in vivo studies, now we know that FSH is a risk marker for metabolic dysfunction given its direct role in promoting adipogenesis and insulin resistance. Nevertheless, epidemiological studies have shown an inverse relationship; individuals with obesity tend to present lower FSH values (which goes in line with our findings from R3 and R1). To account for this apparent contradiction, a hypothesis has been proposed: FSH stimulates obesity, leading to elevated estrogen levels, which subsequently diminish FSH levels through negative feedback mechanisms. Moreover, since FSH is a sexual hormone, its production gets increased after puberty initiation, and its effects could be influenced by even more complex feedback mechanisms, highlighting the relevance of this period for the study of its effects in obesity and cardiometabolic health (as evidenced by our model in R2) [26].

Our model therefore adds some evidence to the existing findings relating high FSH with lower obesity (R1 and R3), reporting in this case an important and novel contribution of sexual development in this relationship (R2). Further epidemiological studies and in vitro studies would be needed to deepen into presented hypotheses.

7. Conclusions

In this application work, we have adopted a Machine Learning approach to obtain explainable regression models for two imbalanced continuous targets in childhood obesity, the z-score of HOMA-IR and the Waist Circumference, with significant imbalance ratios around 12%. The models obtained by the proposed approaches are better considering the relevant minority information, with F1 improvements over 17%. Furthermore, the linguistic fuzzy models generated by LING$_{eXplainable}$ and the proposal, LING$_{eXplainable}$-ImbR exhibit superior interpretability (elevated GM3M and RMI and low number of variables and rules), since they have been specifically designed for explainability. LING$_{eXplainable}$-ImbR presents the best values for F1, that is, the metric that allows us to evaluate the models in the imbalanced set, while relatively preserving the MSE value. Moreover, they seem to properly explain biologically meaningful relations among the involved factors and the predicted targets. The identification of accelerometry variables, specially in zHOMA-IR, and novel relationships, such as FSH levels in predicting zWC, could inform clinical practice by highlighting new biomarkers and pathways for intervention in childhood obesity.

8. Acknowledgements

This research was supported by the Instituto de Salud Carlos III co-funded by ERDF - A way of making Europe - and the European Union (grant numbers PI20/00711, PI20/00563, PI23/00129, PI23/00165 and PI23/00028).

References

[1] A. K. Leung, A. H. Wong, K. L. Hon, Childhood obesity: an updated review, Current Pediatric Reviews 20 (2024) 2–26. doi:10.2174/1573396318666220801093225.

[2] D. Drozdz, J. Alvarez-Pitti, M. Wójcik, C. Borghi, R. Gabbianelli, A. Mazur, V. Herceg-Čavrak, B. G. Lopez-Valcarcel, M. Brzeziński, E. Lurbe, et al., Obesity and cardiometabolic risk factors: from childhood to adulthood, Nutrients 13 (2021) 4176. doi:10.3390/nu13114176.

[3] A. Horesh, A. M. Tsur, A. Bardugo, G. Twig, Adolescent and childhood obesity and excess morbidity and mortality in young adulthood—a systematic review, Current obesity reports 10 (2021) 301–310. doi:10.1007/s13679-021-00439-9.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82 – 115. doi:10.1016/j.inffus.2019.12.012.

[5] L. Torgo, R. P. Ribeiro, B. Pfahringer, P. Branco, Smote for regression, in: Portuguese conf. on artificial intelligence, 2013, pp. 378–389. doi:10.1007/978-3-642-40669-0_33.

[6] P. Branco, L. Torgo, R. P. Ribeiro, Smogn: a pre-processing approach for imbalanced regression, in: First international workshop on learning with imbalanced domains: Theory and applications, 2017, pp. 36–50.

[7] R. Alcalá, M. J. Gacto, F. Herrera, A fast and scalable multiobjective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems, IEEE Transactions on Fuzzy Systems 19 (2011) 666–681. doi:10.1109/TFUZZ.2011.2131657.

[8] M. J. Gacto, M. Galende, R. Alcalá, F. Herrera, Metsk-hde: A multiobjective evolutionary algorithm to learn accurate tsk-fuzzy systems in high-dimensional and large-scale regression problems, Information Sciences 276 (2014) 63–79. doi:10.1016/j.ins.2014.02.047.

[9] M. Arteaga, M. J. Gacto, M. Galende, J. Alcalá-Fdez, R. Alcalá, Enhancing soft computing techniques to actively address imbalanced regression problems, Expert Systems with Applications 234 (2023) 121011. doi:10.1016/j.eswa.2023.121011.

[10] C. Biedma-Rdguez, M. J. Gacto, A. Anguita-Ruiz, J. Alcalá-Fdez, R. Alcalá, Transparent but accurate evolutionary regression combining new linguistic fuzzy grammar and a novel interpretable linear extension, Int. J. Fuzzy Syst. 24 (2022) 3082–3103. doi:10.1007/s40815-022-01324-w.

[11] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[12] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery., Nat Biomed Eng 2 (2018) 749 – 760. doi:10.1038/s41551-018-0304-0.

[13] L. Torgo, R. Ribeiro, Precision and recall for regression, in: Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12, Springer, 2009, pp. 332–346. doi:10.1007/978-3-642-04747-3_26.
[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority oversampling technique, J. Artif. Intell. Research 16 (2002) 321–357. doi:10.1613/jair.953.
[15] Carlos iii health institute, 2024. URL: https://www.isciii.es/Paginas/Inicio.aspx.
[16] ibs Granada, pi20 / 00563 - omics and artificial intelligence as tools to understand molecular mechanisms of insulin resistance in obese children during puberty, https://www.ibsgranada.es/en/proyectos/pi20-00563-omics-and-artificial-intelligence-as-tools-to-understand-molecular..., 2018. [Acceso: 30 de mayo de 2024].
[17] M. Stavnsbo, G. K. Resaland, S. A. Anderssen, J. Steene-Johannessen, S. L. Domazet, T. Skrede, L. B. Sardinha, S. Kriemler, U. Ekelund, L. B. Andersen, et al., Reference values for cardiometabolic risk scores in children and adolescents: Suggesting a common standard, Atherosclerosis 278 (2018) 299–306. doi:10.1016/j.atherosclerosis.2018.10.003.
[18] A. Ferrández, L. Baguer, L. C. Labarta Jl, E. Mayayo, B. Puba, C. Rueda, M. Ruiz-Echarri, Longitudinal study of normal spanish children from birth to adulthood (anthropometric, pubertal, radiological and intellectual data, Pediatr Endocr Rev 2 (2005) 423–559.
[19] B. Sobradillo, A. Aguirre, U. Aresti, A. Bilbao, C. Fernández-Ramos, A. Lizárraga, H. Lorenzo, L. Madariaga, I. Rica, I. Ruiz, E. Sánchez, C. Santamaría, J. Serrano, A. Zabala, B. Zurimendi, M. Hernández, Curvas y tablas de crecimiento (estudios longitudinal y transversal), 2004.
[20] N. H. B. P. E. P. W. G. on High Blood Pressure in Children, Adolescents, The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents, Pediatrics 114 (2004) 555–576.
[21] D. J. Stekhoven, P. Bühlmann, Missforest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012) 112–118. doi:10.1093/bioinformatics/btr597.
[22] C. Frithioff-Bøjsøe, M. A. Lund, U. Lausten-Thomsen, P. L. Hedley, O. Pedersen, M. Christiansen, J. L. Baker, T. Hansen, J.-C. Holm, Leptin, adiponectin, and their ratio as markers of insulin resistance and cardiometabolic risk in childhood obesity, Pediatric diabetes 21 (2020) 194–202. doi:10.1111/pedi.12964.
[23] P. Zhu, X. Liu, X. Mao, Thyroid-stimulating hormone levels are positively associated with insulin resistance, Medical science monitor: international medical journal of experimental and clinical research 24 (2018) 342. doi:10.12659/MSM.905774.
[24] M. I. Santos, C. Limbert, F. C. Marques, F. Rosário, L. Lopes, Childhood obesity, thyroid function, and insulin resistance–is there a link? a longitudinal study, Journal of Pediatric Endocrinology and Metabolism 28 (2015) 557–562. doi:10.1515/jpem-2014-0319.
[25] D. Sun, M. Bai, Y. Jiang, M. Hu, S. Wu, W. Zheng, Z. Zhang, Roles of follicle stimulating hormone and its receptor in human metabolic diseases and cancer, American Journal of Translational Research 12 (2020) 3116.
[26] B. K. Aydin, R. Stenlid, I. Ciba, S. Y. Cerenius, M. Dahlbom, P. Bergsten, R. Nergårdh, A. Forslund, High levels of fsh before puberty are associated with increased risk of metabolic syndrome during pubertal transition, Pediatric Obesity 17 (2022) e12906. doi:10.1111/pedi.12964.