Estimation of obesity levels based on computational intelligence

Rodolfo Cañas Cervantes, Ubaldo Martinez Palacio *

Department of Computer Science and Electronics, Universidad de la Costa, CUC. Faculty Teacher Systems Engineering Program, Colombia

ABSTRACT

Obesity is a worldwide disease that affects people of all ages and gender; in consequence, researchers have made great efforts to identify factors that cause it early. In this study, an intelligent method is created, based on supervised and unsupervised techniques of data mining such as Simple K-Means, Decision Trees (DT), and Support Vector Machines (SVM) to detect obesity levels and help people and health professionals to have a healthier lifestyle against this global epidemic. In this research the primary source of collection was from students 18 and 25 years old at institutions in the countries of Colombia, Mexico, and Peru. The study takes a dataset relating to the main causes of obesity, based on the aim to reference high caloric intake, a decrease of energy expenditure due to the lack of physical activity, alimentary disorders, genetics, socioeconomic factors, and/or anxiety and depression. In the selected dataset, 178 students participated in the study, 81 male and 97 female. Using algorithms including Decision Tree, Support Vector Machine (SVM), and Simple K-Means, the results show a relevant tool to perform a comparative analysis among the mentioned algorithms.

1. INTRODUCTION

Excessive fat accumulation in various parts of the body presents a risk to the normal development of health and affects indistinctly people of all ages and gender [1]. According to the World Health Organization (WHO) [2] many persons aged over 18 years suffer from weight alteration problems, caused by several factors associated to food intake with high caloric content, sedentary nature, and modes of transportation. Authors [3] mention other determining factors for obesity: "being an only child, familiar conflicts such as the divorce of parents, depression or anxiety". According to Ref. [4], obesity is a disease with multiple factors; its main feature is the uncontrolled increase of weight due to an excessive proportion of fat, high caloric intake, and low energy consumption.

Based on prior studies, one can see that authors have invested important efforts in research [5–14]. They have performed analysis of the factors that impact the appearance of obesity, producing different computationally intelligent methods based on data mining techniques, discipline that involves the study of massive data sources, to search novel information and use it for decision-making processes.

In the USA, obesity prevalence based on self-reported data from BRFSS 2013 was 29%, in contrast to 34% using objectively-measured height and weight data from the National Health and Nutrition Examination Survey (NHANES). While NHANES is useful for monitoring national trends in obesity, its relatively small sample size (and lack of data collection in every state during each survey) is insufficient to produce a yearly state-specific estimate. As a result, no nationally-representative, objectively-measured BMI surveillance system exists that can provide unbiased estimates of state-specific obesity prevalence. This lack of accurate data limits states' ability to evaluate the health and economic effects of the obesity epidemic and to plan prevention policies and programs [15].

With the prevalence of overweight and obesity on the rise, the need to develop computational approaches to predict which individuals will benefit from dieting is also increasing. While weight loss shows success, we must balance it against any future weight regain. Similarly, a reduction in an individual's weight-related medication regimen may show the improvement of overall health, but we must maintain it to consider it as a long-term improvement. Besides the metric used to determine success, we must also consider how data is acquired. While in-lab tests may yield more reliable and detailed information, at-home monitoring involves less time and effort, especially for clinicians. Machine learning is a branch of computer science that explores algorithms meant to learn from the data. Widely applicable, various implementations of this approach have been used to examine a wide range of problems. From protein-protein interactions and tumour classification to detecting oil spills in satellite images, machine learning algorithms have a large impact on current science [35].

This study has the purpose of implementing methods and techniques

in data mining, based on supervised and unsupervised learning, to detect whether a person suffers from obesity. The structure of the research is: in section 2 the methodology used for the experimentation phase and the description of methods and techniques is presented, in section 3 are the results obtained from the techniques DT, SVM, and simple k-means, and finally in section 4 the conclusions of the study are provided.

2. RELATED WORKS

The authors in Ref. [5] present a logistic regression model to estimate the probability of mass body index in children with ages between 2 and 17 years, over small geographic areas. In Refs. [6] is presented a computational model using a fuzzy signature to understand and manage the intricacies of data in childhood obesity and one solution that could handle the associated risk with early obesity and motor development in childhood. In Refs. [7] is shown a framework with a hybrid approach based on Naïve Bayes for prediction, and genetic algorithms to optimize parameters applied to the prediction of pediatric obesity, with a low rate of negative samples compared to the positive ones.

The authors in Ref. [8] exhibit an initial approach to the study of predicting childhood obesity, collecting information from primary sources: parents, children, and caretakers. They identified risk factors such as obesity and education level of the parents, lifestyle and habits of the children, and environmental influence. In Ref. [9] is proposed the use of data mining for children obesity prediction. The purpose of the survey proposed is to provide the understanding needed for obesity as a disease.

Based on the literature found, is possible to see the effort of several authors to analyse the disease, even creating web tools like IMC calculation (2019), here one can calculate the obesity level of a person; nevertheless these tools are limited to the body mass index calculation, omitting other relevant factors such as if the individual has a family history with obesity, time dedicated to exercise routines, and other factors. Thus an intelligent tool is needed, capable of detecting efficiently obesity level.

In [10], the author aims to develop a model for the identification, analysis, and estimation of obesity, in which each user is considered a 'sensor' of the online social network that can provide valuable health information. Based on the detailed measurement of the correlation of obesity and the proposed characteristics, the analytical model of obesity of the NSO can estimate the rate of obesity in certain urban areas, and the experimental results demonstrate a high rate of estimation success.

The authors in Ref. [11] propose mid-level features, namely kernel-based features for prediction of health indices of populations from social media data. They extract the kernel-based features on the distributions of textual features overpopulation tweets and encode the relationships between individual textual features in a kernel function. They implemented features using three different kernel functions and applied them for two case studies of population health prediction: across-year prediction and across-county prediction.

In [12] the authors proposed a risk mining technique (PRMT) that foretells a model to analyse the risk factor of obesity class using different data mining classifiers, using WEKA to estimate the accuracy and error measurement. The outcome of this process via Naïve Bayes is the best classifier for the 10-fold cross-validation study.

The authors in Ref. [13] proposed a review to show Artificial Intelligent applications to obesity management and discussed their effectiveness. They performed the research in the following databases: Public Medline (PubMed), Web of Science, Biblioteca Regional de Medicina (BIREME), and Google Academic, by using the following keywords, "artificial intelligence" and "obesity". The results led to some Artificial Intelligence systems used in obesity handling, which were: the Decision Support System to bariatric surgery patients; the MOPET app to motivate physical activity; Parameter Decreasing Methods and Artificial Neural Network to correlate obesity to cardiovascular disease; Artificial Neural Network to predict resting energy expenditure; a Neuro-Fuzzy

Model to refine body mass index result; an Image Processing Algorithm; and a Support Vector Machine that monitors food intake.

In the article "Obesity Level Estimation Software based on Decision Trees" [14], the authors presented 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; they labelled the records 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. They generated 77% of the data synthetically using the Weka tool and the SMOTE filter, and they collected 23% of the data directly from users through a web portal. 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.

According to Ref. [15], accurate methods to measure food and energy intake are crucial for the battle against obesity. Providing users/patients with convenient and intelligent solutions that help them measure their food intake and collect dietary information are the most valuable insights toward long-term prevention and successful treatment programs. The authors proposed an assistive calorie measurement system to help patients and doctors succeed in their fight against diet-related health conditions. The system runs on smartphones, which allows the user to take a picture of the food and measure the amount of calorie intake automatically. To identify the food accurately in the system, they use deep convolutional neural networks to classify 10,000 high-resolution food images for system training. The results show that the accuracy of the method for food recognition of single food portions is 99%.

In [16], the authors propose a new dataset for the evaluation of food recognition algorithms that can be used in dietary surveillance applications. Each image represents a real dining tray with dishes and foods arranged in different ways. Each tray contains multiple instances of food types. The dataset contains 1027 dining trays for a total of 3616 instances of food belonging to 73 food types. The food on the tray images has been manually segmented using carefully drawn polygonal boundaries. They have made a comparative evaluation of the dataset by designing an automatic tray analysis pipeline that takes an image of the tray as input, finds the regions of interest, and predicts for each region the corresponding food type. They have experimented with three different classification strategies using also various visual descriptors. They achieved about 79% accuracy in food and tray recognition using features based on convolutional neural networks.

According to Ref. [17], recent large-scale genome-wide association studies have identified tens of genetic loci robustly associated with Body Mass Index. They also found their associated gene expression profiles with BMI. However, accurate prediction of obesity risk utilising genetic data remains challenging. In a cohort of 75 individuals, the authors integrated 27 BMI-associated SNPs and obesity-associated gene expression profiles. They computed the genetic risk score by adding BMI-increasing alleles. They correlated the genetic risk score with BMI when they used an optimization algorithm that excluded some SNPs. They built linear regression and support vector machine models to predict obesity risk using gene expression profiles and the genetic risk score. They achieved an adjusted r^2 value of 0.556 and an accuracy of 76% for the linear regression and support vector machine models, respectively. The authors report a new mathematical method to predict obesity genetic risk. They constructed obesity prediction models based on genetic information for a small cohort. The computational framework serves as an example of using genetic information to predict obesity risk for specific cohorts.

According to Ref. [31], current methods of self-monitoring kilocalorie intake outside of laboratory/clinical settings suffer from systematic underreporting bias. Recent efforts to make kilocalorie information available have improved these methods to some extent, but it may be possible to derive an objective and more accurate measure of kilocalorie

intake from bite count. As an objective, this study sought to develop and examine the accuracy of an individualized bite-based measure of kilocalorie intake and to compare that measure to participant estimates of kilocalorie intake. They hypothesized that kilocalorie information would improve human estimates of kilocalorie intake over those with no information, but a bite-based estimate of kilocalorie intake would still outperform human estimates. Multiple regression statistics were used to determine whether height, weight, age, sex, and waist-to-hip ratio could predict an individual's mean kilocalories per bite for the training sample. The model was then validated with the test group, and the model-predicted kilocalorie intake was compared with the human-estimated kilocalorie intake.

The authors [32] focus on accurate EE estimation for tracking ambulatory activities (walking, standing, climbing upstairs, or downstairs) of individuals wearing mobile sensors. They use Convolution Neural Networks (CNNs) to detect automatically important features from data collected from triaxial accelerometer and heart rate sensors. Using CNNs, they find a significant improvement in EE estimation compared to other state-of-the-art models. They compare the results against state-of-the-art Activity-Specific Linear Regression and Artificial Neural Networks (ANN) based models. Using a universal CNN model, an overall low Root Mean Square Error (RMSE) of 1.12 is obtained, which is 30% and 35% lower than existing models. They calibrated the results against COSMED K4b2 indirect calorimeter readings.

The authors [33] performed a clinical experiment, using optimal design, in volunteers weighing 44–150 kg, to identify the effect of obesity on dapsone pharmacokinetic parameters based on the maximum-likelihood solution via the expectation-maximisation algorithm. They used artificial intelligence-based multivariate adaptive regression splines for covariate selection and identified weight and/or age as predictors of absorption, systemic clearance, and volume of distribution. These relationships occurred only between certain patient weight and age ranges, delimited by multiple hinges and regions of discontinuity, not identified by standard pharmacological approaches. Older and obese people have lower drug concentrations after standard dosing, but with complex patterns. Given that efficacy is concentration-dependent, they need to personalise optimal dapsone doses for obese patients.

The authors in Ref. [34] aimed at predicting different measures of obesity based on the plasma lipidome in a large population cohort using an advanced machine learning model. A total of 1061 participants from the FINRISK 2012 population cohort were randomly selected, and the levels of 183 species of plasma lipids were measured in a novel mass spectrometric approach. Multiple machine intelligence models were trained to predict the obesity estimates, i.e., body mass index (BMI), waist circumference (WC), waist-hip ratio (WHR), and body fat percentage (BFP), and validated in 250 randomly selected participants from the Malmö Cardiovascular Diet and Cancer Cohort (MDC-CC). A comparison of the different models revealed that the lipidome predicted BFP the best ($R2 = 0.73$), based on a Lasso model. In this model, the strongest positive and the strongest negative predictor were sphingomyelin molecules, which differed by only 1 double bond, implying the involvement of an unknown desaturase in obesity-related aberrations of lipid metabolism.

3. Methodology

For this research, the authors referenced the study performed, which was used as a primary source of the data collected from a set of students of institutions of Colombia, México, and Peru [14]. The students had ages between 18 and 25 years.

3.1. Dataset

One of the major causes for the development of obesity refers to high caloric intake, a decrease of energy expenditure due to the lack of physical activity, alimentary disorders, genetics, socioeconomic factors, and/or anxiety and depression, based on [16].

According to Ref. [14], for creating their dataset, the authors searched in literature for the key factors or habits that are associated with obesity. The dataset has 18 variables that can determine if a person presents obesity. To collect the information, they presented the variables as questions through a survey, so as to apply to a group of college students in the countries of Colombia, Mexico, and Peru.

In the selected dataset, you can see that 178 students took part in the study, 81 men and 97 women with ages between 18 and 25 years.

3.2. Decision trees (DT)

A Decision Tree (DT) is defined as a classification procedure that partitions recursively a dataset into subdivisions; generally a decision tree is composed of a root node and a set of internal nodes created after data division and terminal nodes. Each node has a single parent and two or more descendants (see Friedl and Brodley, 1997). According to Magerma (1995), a decision tree is used to support decision-making processes and sets a probability for each of the choices based on the context of the decision. In Zhao and Koller (2001) is mentioned that decision tree represent a supervised approach for classification that has a simple structure formed by nodes or terminal nodes; nodes represent tests in one or more attributes and terminal nodes show the results of the decisions. The decision tree are implemented through different algorithms such as C4.5 and random forest.

3.3. Support Vector Machines (SVM)

SVM has a strong theoretical foundation and excellent empirical results. In Tong and Koller (2001), SVM has been applied by different agents in tasks such as digit recognition, also in Vapnik (1998), and in object recognition Papageorgiou et al. (1998), text classification [17] and human activities recognition [18]. These systems are based on the statistical learning system developed in Ref. [19]; they proposed a mathematical model for regression and classification problems In Ref. [20,21], and [22] it is considered a major advantage of SVM is the availability of powerful tools and algorithms to find solutions in an efficient and fast manner.

3.4. Simple K-Means

Simple K-Means is considered the most popular method for unsupervised learning tasks, and its foundation is the construction of a set of physical or abstract objects that has features or attributes in common [23]. Based on [24], from a good clustering, one can generate high-quality groups with high similarities for inter-classes and low similarities between classes. The implementation of algorithms of clustering for unsupervised analysis of the data has become a useful tool to explore and solve different problems in data mining. The cluster grouping methods [25,26] have been used to propose solutions to problems of many contexts and different disciplines.

3.5. Experimentation

This study proposes a method based on computational intelligence through supervised and unsupervised data mining techniques to predict or detect the obesity levels in young people. For this aim, the steps performed were as follows.

First, the dataset to be used was identified; the dataset selected was created by Ref. [14]. Once obtained, the process of preparation and transformation of the data was made, to train the data mining methods with clean information, avoiding missing data, atypical data, unbalanced classes and checking the correlation level between attributes.

With the dataset ready, the next step was to apply the data mining methods and techniques; for this the authors used the Weka tool, that

Table 1
Result comparison (DT, SMV and Crossed Validation).

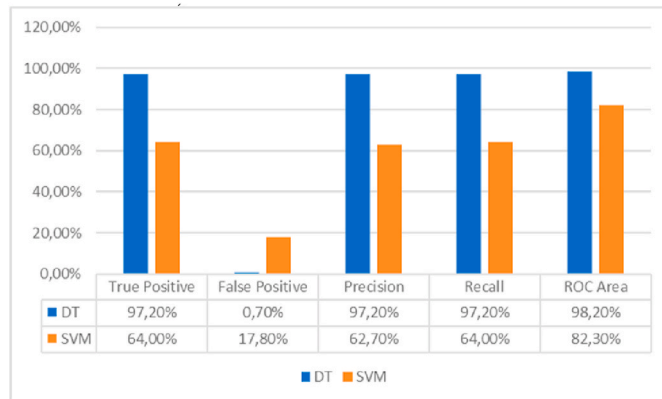| Method	True Positive	False Positive	Precision	Recall	ROC Area
DT	97.2%	0.7%	97.2%	97.2%	98.2%
SVM	64.0%	17.8%	62.7%	64.0%	82.3%

Fig. 1. Results comparative graph (DT, SVM and Crossed validation).
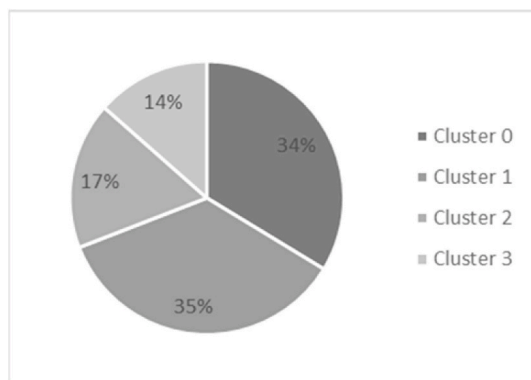

Fig. 2. The distribution of groups of records created with Simple K-Means.

contains a set of algorithms useful for many situations. In this study, the classification methods DT (J48), SVM (SMO) were used and Simple K-Means was selected as the clustering method for validation of the proposed model. To choose the best technique, the metrics used were precision, recall, True Positive Rate (TP Rate), False Positive Rate (FP Rate), and ROC area, since they are commonly used in previous studies such as [14,27–29], for the training process. Cross validation was used according to Ref. [30], creating a distribution of the data, one part for the training process and another for the testing process, guaranteeing optimum results and avoiding over-training issues.

Finally, after having the results of each classification algorithm, the best method was fused with the clustering algorithm, to generate an intelligent system based on the unsupervised technique of Simple K-Means.

4. Results

To develop the proposed tool, a comparative analysis was performed using the DT and SVM methods, testing them with the same training scenario (Table 1, Fig. 1).

Then, a segmentation process was performed based on the results of whether a person is obese or not, generating four groups described below in Fig. 2.

Table 2
The description of the groups generated with Simple K-Means.

Cluster	Description
Cluster 0	Sex: Male Average age: 21 years Average height: 1.67 m Weight: 66 kg Family with overweight problems, no fast food intake, usually consume vegetables, often have 3 meals, have snacks between meals, no smoking, drink more than 2 L of liquids daily, no daily caloric control, exercise 3 or 4 times weekly, use technology 0–2 h daily, no alcohol intake, use public transportation, BMI average 24.47, and they are considered people not prone to overweight problems
Cluster 1	Sex: Female Average age: 22 years Average height: 1.64 m Weight: 66 kg Family with overweight problems, no fast food intake, always consume vegetables, always have 3 meals, rarely have snacks between meals, no smoking, drink more than 2 L of liquids daily, daily caloric control, exercise 3 or 4 times weekly, use technology 0–2 h daily, no alcohol intake, use public transportation, BMI average 23.58, and they are considered people not prone to overweight problems
Cluster 2	Sex: Male Average age: 22 years Average height: 1.7 m Weight: 94 kg Family with overweight problems, fast food intake, rarely consume vegetables, always have 3 meals, rarely have snacks between meals, no smoking, drink 1–2 L of liquids daily, no daily caloric control, no exercise, use technology usually 3–4 h daily, rarely alcohol intake, use an automobile, BMI average 31,6, and they are considered people prone to overweight problems
Cluster 3	Sex: Female Average age: 22 years Average height: 1.6 m Weight: 72 kg Family with overweight problems, fast food intake, rarely consume vegetables, always have 3 meals, rarely have snacks between meals, no smoking, drink 1–2 L of liquids daily, no daily caloric control, no exercise, use technology usually 3–4 h daily, no alcohol intake, use public transportation, BMI average 27,9, and they are considered people prone to overweight problems

Table 3
Results obtained (DT, Simple K-Means and Cross Validation).

Method	True Positive	False Positive	Precision	Recall	ROC Area
DT + Simple k-Means	98.5%	0.2%	98.5%	98.5%	99.5%

In Table 2 you can find the descriptive characteristics of each of the clusters that are part of the study.

Finally, after the records were assigned to each cluster, the computational model was trained using the DT method based on the clusters generated by the method Simple K-Means, which produced the following results.

These results are evidence of improvement compared with the previous results obtained in Table 2, where the levels of recall reached by the DT method were exceeded by the recall levels of the method with Simple K-Means and DT combined in Table 3.

5. Conclusions

Data mining is a discipline responsible for performing exploratory analysis of data to identify patterns or behaviours on information Through this discipline, researchers have created solutions for many problems in society, such as disease identification based on historical data. The analysis of obesity level is a current need for society, since it is present worldwide and can affect people regardless of age or gender.

Based on the literature analysis presented in this study, many authors have invested efforts and resources to identify this pathology, which makes this theme one of constant evolution.

In this research, a method based on computational intelligence is created, using supervised and unsupervised techniques including K-Means, DT, and SVM. The methods were compared through evaluation metrics of precision, recall, true positive rate, false-positive rate, and ROC area. The training and classification processes were performed using the data mining tool Weka after a preparation and transformation data process was made to identify missing data, atypical data, and correlation analysis. Finally, the results obtained by the DT and Simple K-Means methods in precision (98.5%), recall (98.5%), true positive rate (98.5%), false-positive rate (0.2%) and, ROC area (99.5%), surpass the results obtained in previous studies as [7,10,14] that had values of 75% and 85% in precision level.

The results presented can be useful to analyse the relevance of methods based on computational intelligence to study different diseases or pathologies, early and proper detection of them, and minimise the impact of those diseases on society.

Declaration of competing 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.

Acknowledgements

REFERENCES

[1] Gutiérrez HM. Diez problemas de la Población de Jalisco: una perspectiva sociodemográfica. In: Edición Primera, editor. Guadalajara, México. Dirección de Publicaciones del Gobierno de Jalisco; 2010.

[2] OMS. Organización mundial de la Salud. 2016. http://www.who.int/mediacent re/factsheets/fs311/es/.

[3] Olmedo, M. V. "La obesidad: un problema de salud pública. 2016 Revista de divulgación científica y tecnológica de la Universidad Veracruzana. Reference to a journal publication with an article number.

[4] Hernández GM. Prevalencia de sobrepeso y obesidad, y factores de riesgo, en niños de 7-12 años, en una escuela pública de Cartagena. Universidad Nacional de Colombia (Colombia), septiembre - octubre; 2011.

[5] Davila-Payan C, DeGuzman M, Johnson K, Serban N, Swann J. Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. Prev Chronic Dis 2015;12:140229. https://doi. org/10.5888/pcd12.140229.

[6] Manna, S., & Jewkes, A. M. "Understanding early childhood obesity risks: an empirical study using fuzzy signatures", In Fuzzy systems (FUZZ-IEEE). 2014 IEEE international conference on (pp. 1333-1339). IEEE.

[7] Adnan MHBM, Husain W. 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. p. 281–5.

[8] Adnan MHM, Husain W. A framework for childhood obesity classifications and predictions using NBtree". In: Information technology in asia (CITA 11), 2011 7th international conference on. IEEE; 2011. p. 1–6.

[9] Adnan, M. H. B. M., Husain, W., & Damanhoori, F. "A survey on utilization of data mining for childhood obesity prediction", In Information and telecommunication technologies (APSITT). 2010 8th asia-pacific symposium on (pp. 1-6). IEEE.

[10] Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. Appl Clin Inf 2015;6(3):506–20.

[11] Zhang ML, Zhou ZH. Multi-instance clustering with applications to multi-instance prediction. Appl Intell 2009;31(1):47–68.

[12] Suguna M. Childhood obesity epidemic analysis using classification algorithms. Int. J. Mod. Comput. Sci 2016;4(1):22–6.

[13] Abdullah FS, Manan NSA, Ahmad A, Wafa SW, Shahril MR, Zulaily N, Ahmed A. Data mining techniques for classification of childhood obesity among year 6 school children. In: International conference on soft computing and data mining. Cham: Springer; 2016. p. 465–74.

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

[15] Ward ZJ, Long MW, Resch SC, Gortmaker SL, Cradock AL, Giles C, et al. Redrawing the US obesity landscape: bias-corrected estimates of state-specific adult obesity prevalence. PloS One 2016;11(3):e0150735. https://doi.org/10.1371/journal. pone.0150735.

[16] Gómez M, Ávila L. La obesidad: un factor de riesgo cardiometabólico. In: Medicina de Familia, vol. 8; 2008. p. 91–7. N°. 2.

[17] Joachims T. Text categorization with support vector machines. Proceedings of the European C~njerence on machine learning. Springer-Verlrtg; 1998.

[18] Kim Y, Ling H. Human activity classification based on micro-Doppler signatures using a support vector machine". IEEE Trans Geosci Rem Sens 2009;47(5): 1328–37.

[19] Da Silva F, Niedermeyer E. "Electroencephalography: basic principles. 1993. Clinical applications, and related fields", William & Wikins, Baltimore.

[20] Parsons TD, Rizzo AA. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: a meta-analysis. J Behav Ther Exp Psychiatr 2008;39 (3):250–61.

[21] De la Hoz E, de la Hoz E, Ortiz A, Ortega J, Martínez-Álvarez A. Feature selection by multi-objective optimisation: application to network anomaly detection by hierarchical self-organising maps. Knowl-Based Syst 2014;71:322–38.

[22] Bekele E, Wade J, Bian D, Fan J, Swanson A, Warren Z, Sarkar N. Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (ASD). In: 2016 IEEE virtual reality. VR; 2016. p. 121–30. https://doi.org/10.1109/VR.2016.75046_95.

[23] Han J, Kamber M. Data mining: concepts and techniques. second ed. San Francisco: Morgan Kaufmann Publishers; 2001.

[24] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference and prediction. New York: Springer; 2001.

[25] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans Neural Network 2005;16(3):645–78.

[26] Zhang ML, Zhou ZH. Multi-instance clustering with applications to multi-instance prediction. Appl Intell 2009;31(1):47–68.

[27] Palechor FM, De la Hoz Manotas A, Colpas PA, Ojeda JS, Ortega RM, Melo MP. Cardiovascular disease analysis using supervised and unsupervised data mining techniques. J SW 2017;12(2):81–90.

[28] Mendoza-Palechor F, Menezes ML, Sant'Anna A, Ortiz-Barrios M, Samara A, Galway L. Affective recognition from EEG signals: an integrated data-mining approach. Journal of Ambient Intelligence and Humanized Computing 2018:1–20.

[29] Fabio Mendoza-Palechor, Alexis De la Hoz-Manotas, Roberto Morales-Ortega, Ubaldo Martinez-Palacio, Jorge Diaz-Martinez, Harold Combita-Nino. Designing A method for alcohol consumption prediction based on clustering and support vector machines. Res J Appl Sci Eng Technol 2017;14(4):146–54.

[30] Palechor FM, Manotas ADLH, Franco EDLH, Colpas PA. Feature selection, learning metrics and dimension reduction in training and classification processes in intrusion detection systems. J Theor Appl Inf Technol 2015;82(2):291.

[31] Salley James N, Hoover Adam W, Wilson Michael L, Muth Eric R. Comparison between human and bite-based methods of estimating caloric intake. J Acad Nutr Diet 2016;116(Issue 10):1568–77. https://doi.org/10.1016/j.jand.2016.03.007. ISSN 2212-2672, http://www.sciencedirect.com/science/article/pii/S2212267 216300090.

[32] Zhu J, Pande A, Mohapatra P, Han JJ. Using deep learning for energy expenditure estimation with wearable sensors. In: 2015 17th international conference on E-health networking. Boston, MA: Application & Services (HealthCom); 2015. p. 501–6. https://doi.org/10.1109/HealthCom.2015.7454554.

[33] Hall R, Pasipanodya J, Swancutt M, Meek C, Leff R, Gumbo T. Supervised machine-learning reveals that old and obese people achieve low dapsone concentrations. CPT Pharmacometrics Syst Pharmacol 2017;6:552–9. https://doi.org/10.1002/ psp4.12208.

[34] Gerl MJ, et al. Machine learning of human plasma lipidomes for obesity estimation in a large population cohort. PLoS Biol Oct. 2019;17(10):e3000443. https://doi. org/10.1371/journal.pbio.3000443 [Online]. Available:.

[35] Craig A. Biwer. Computing obesity: signal processing and machine learning applied to predictive modeling of clinical weight-loss. 2017. https://deepblue.lib.umich.ed u/bitstream/handle/2027.42/140907/cbiwer_1.pdf?sequence=1&isAllowed=y.