Journal of Computer Science

Original Research Paper

Obesity Level Estimation Software based on Decision Trees

[1]Eduardo De-La-Hoz-Correa, [2]Fabio E. Mendoza-Palechor, [2]Alexis De-La-Hoz-Manotas, [2]Roberto C. Morales-Ortega and [2]Sánchez Hernández Beatriz Adriana

[1]Corporación Universitaria Americana, Colombia
[2]Universidad de la Costa, Colombia

Article history
Received: 20-06-2018
Revised: 20-11-2018
Accepted: 7-01-2019

Corresponding Author:
Roberto C. Morales-Ortega
Universidad de la Costa,
Colombia
Email: rmorales1@cuc.edu.co

Abstract: Obesity has become a global epidemic that has doubled since 1980, with serious consequences for health in children, teenagers and adults. Obesity is a problem has been growing steadily and that is why every day appear new studies involving children obesity, especially those looking for influence factors and how to predict emergence of the condition under these factors. In this study, authors applied the SEMMA data mining methodology, to select, explore and model the data set and then three methods were selected: Decision trees (J48), Bayesian networks (Naïve Bayes) and Logistic Regression (Simple Logistic), obtaining the best results with J48 based on the metrics: Precision, recall, TP Rate and FP Rate. Finally, a software was built to use and train the selected method, using the Weka library. The results confirmed the Decision Trees technique has the best precision rate (97.4%), improving results of previous studies with similar background.

Keywords: Obesity, Data Mining, Semma, Decision Trees, Naive Bayes, Logistic Regression, Weka, Java

Introduction

The World Health Organization (WHO) (OMS, 2016), describes obesity and overweight as excessive fat accumulation in certain body areas that can be harmful for health, the number of people that suffers from obesity has doubled since 1980 and also in 2014 more than 1900 million adults, 18 years old or older, are suffering from alteration of their weight. Some of the causes of overweight are the increase of intake of energy dense foods that are high in fat and decrease in physical activity due to the nature of a sedentary types of work, the new transportation modes and increasing urbanization.

According to (Gutiérrez, 2010), obesity is a public health problem worldwide and it can emerge in adults, teens and children.

Hernández (2011), the authors show that obesity can be considered a disease with multiple factors, having as symptom, the uncontrolled increase of weight, due excessive intake of fat and energy consumption.

Obesity can be caused by biological hazard factors such as hereditary background, so there can be several kinds of obesity as: Monogenic, leptin, polygenic and syndromic. Besides, there are other risk factors as social, psychological and eating habits as mentioned by (Zhingre and del Cisne, 2015). On the other side, authors as (Olmedo, 2011) propose other determining factors for obesity such as "being only child, family conflicts as divorce, depression and anxiety".

Based on the previous statements and the literature you can find in many studies working the obesity influence factors, they have implemented several data mining techniques as you can find in (Davila-Payan *et al*., 2015; Manna and Jewkes, 2014; Adnan and Husain, 2012; 2011; Adnan *et al*., 2010; Dugan *et al*., 2015; Zhang *et al*., 2009; Suguna, 2016; Abdullah *et al*., 2016). Data mining is a discipline that studies massive data sources, with the objective of obtaining new information from it, to support decision making.

Several authors have studies to analyze the disease and generate web tools to calculate the obesity level of a person, nevertheless such tools are limited to the calculation of the body mass index, omitting relevant factors such as family background and time dedicated to. Based on this, the authors considered an intelligent tool was needed to be able to detect obesity levels on people more efficiently.

This study had the objective of implementing several data mining techniques to determine if one person suffers from obesity. The methodology of the study was: Analysis of previous studies, creation of the dataset, analysis of data mining techniques, design and implementation of the estimation obesity tool, results and conclusions.

Previous Works

Obesity has become an area of interest for research and many studies can be found working with the factors that produce the disease. Next, you can find a brief review of works proposed by different authors that implement data mining techniques on datasets with attributes related with this health issue.

Davila-Payan *et al*. (2015), a logistic regression model was presented to estimate the probability of body mass index on children from 2 to 17 years old in small geographic areas. Their results confirmed that estimates in small geographic areas are essential to generate effective interventions and to help planning of possible solutions to the problem.

Manna and Jewkes (2014), a computational model was presented using fuzzy signature to understand and manage intricacies on the data of children obesity and a solution that could handle the risk associated with early obesity and children motor development. Their study used fuzz signatures based on fuzzy logic, a computational paradigm that provides a mathematical tool to handle uncertainty and imprecision, quite common in human reasoning.

Adnan and Husain (2012), a framework was presented with a hybrid approach, based on Naïve Bayes for prediction and genetic algorithms for parameter optimization, applied to the problem of predicting children obesity, with a low rate of negative samples compared to positive samples. As result, they obtained 19 parameters to be implemented in prediction with a precision of 75%.

Adnan and Husain (2011), they had an initial approach to the study of predicting children obesity, collecting information from primary sources: Parents, children and caretakers. The authors identified risk factors such as: Obesity and level of education of the parents, lifestyle and habits of the children and influence of environment. The proposed framework uses a hybrid technique of Naïve Bayes and decision trees called NBTree.

Adnan *et al*. (2010), the study used data mining to predict children obesity. The purpose of the proposed survey was to provide the necessary knowledge for the obesity problem, introduce data mining for prediction, describe the current efforts in that area and show the benefits and weaknesses of each technique used. The techniques involved were Neural Networks, Naïve Bayes and Decision Trees.

Dugan *et al*. (2015), the authors generated a predictive study of children obesity with subjects older than 2 years old, using exclusively the data previous to their second birthday using a decision-making system called CHICA. The methods analyzed included: RandomTree, RandomForest, J48, ID3, Naïve Bayes and Bayes. Their results showed that ID3 had better behavior with 85% in precision and 89% in sensibility.

Zhang *et al*. (2009), the authors presented a comparison of logistic regression with six data mining techniques for children overweight and obesity prediction in 3-year-old subjects, using data at birth, at six weeks, at 8 months and two years old respectively. Authors noticed an improvement in the precision of prediction in the cases of 8 months and 2 years old in more than 10%. The techniques used were Decision Trees, Association Rules, Neural Networks, Naïve Bayes, Bayesian Networks and Support Vector Machines.

Suguna (2016), the authors provided a framework using the Child and Adolescent Health Measurement Initiative (CAHMI) dataset, that analyzed obesity in children between 10 and 17 years old. The proposed model uses Decision Trees with three different algorithms: Simple Cart, J47 and NB Tree.

Abdullah *et al*. (2016), the study showed a children obesity classification in grade school 6, from two different Malaysia districts. From the information collected, the authors created 4245 full datasets and they applied the classification techniques: Bayesian Networks, Decision Trees, Neural Networks and Support Vector Machines (SVM).

Husain *et al*. (2013), the authors presented MyHealthyKids, an intervention system for primary schools with the goal of handling and reducing children obesity problems. The system was composed of three modules: Obesity prediction, persuasion and recipe suggestion. The prediction module was based on Naïve Bayes to identify children that are prone to obesity. Tests showed that the system had a precision of 73.3% and great response from children.

Materials and Methods

This study used data related with young undergraduate students between 18 and 25 years old, including nationals from Colombia, Mexico and Perú. The size of the sample was 712 records, based on the surveys applied to 324 men and 388 women.

To initiate the process of collecting information, it was necessary to select the right number of students and they were surveyed with a series of questions to identify their obesity level, considering several factors such as age, weight, sex, physical activity frequency, fast food intake and others, that could help to describe the behavior of obese people. With the information gathered by the survey, it was possible to create a dataset and then the authors performed several types of analysis to discover patterns about the factors that influence the emergence of obesity in young students.

The methods and techniques used in the experimentation process of this study, refer to Decision Trees, Naïve Bayes and Logistic Regression.

To identify the obesity levels, we used the table provided by WHO (Table 1), to categorize correctly the data analyzed based on the BMI.

Dataset

The main causes for development of obesity are related to a high intake of calories, decrease of energy consumption (due to lack of physical activity), genetics, socio-economic factors and/or anxiety and depression, according to (Gómez and Ávila, 2008).

To create the dataset, first we searched for literary sources with the purpose of identify the main factors or habits that contribute to obesity. The dataset generated had 18 variables that make possible to determine if a person has the pathology, the information was collected, by the authors, through a survey and applied to undergraduate students of universities in Colombia, México and Perú.

In Table 2 you can see the factors considered to obtain the obesity levels with their corresponding values.

Table 1: BMI classification according to WHO and Mexican normativity (DO, 2010)

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

Table 2: Dataset description

Attributes	Values
Sex	H: Male M: Female
Age	Integer Numeric Values
Height	Integer Numeric Values (Mt)
Weight	Integer Numeric Values (Kg)
Family with overweight / Obesity	Yes No
Fast Food Intake	Yes No
Vegetables Consumption Frequency	S: Always A: Sometimes CN: Rarely
Number of main meals daily	1 to 2: UD 3: TR More than 3: MT
Food intake between meals	S: Always CS: Usually A: Sometimes CN: Rarely
Smoking	Yes No
Liquid intake daily	MU: Less than one liter UAD: Between 1 and 2 liters MD: More than 2 liters
Calories Consumption Calculation	Yes No
Physical Activity	UOD: 1 to 2 days TAC: 3 to 4 days COS: 5 to 6 days NO: No physical activity
Schedule dedicated to technology	CAD: 0 to 2 hours TAC: 3 to 5 hours MC: More than 5 hours
Alcohol consumption	NO: No consumo de alcohol CF: Rarely S: Weekly D: Daily
Type of Transportation used	TP: Public transportation MTA: Motorbike BTA: Bike CA: Walking AU: Automobile
IMC	WHO Classification
Vulnerable	Based on the WHO Classification

Finally, the dataset is a product generated based on the answers of the students who applied to the survey. Next, several data mining methods or techniques were applied to extract information that can be used to identify people with tendency to suffer obesity.

Decision Trees

Decision trees are considered classification algorithms with high performance, the most popular ones have been implemented in several tools and their names are ID3, C4.5, C5, BFTree and RandomForest. According to (Safavian and Landgrebe, 1991) decision trees can be used in many research areas as: To classify radar signals, text recognition, medical diagnoses, expert systems and others, with high levels of success.

Decision trees are classification methods that take the analyzed data and use a representation in a tree data structure, to provide better insight of the information from the data.

Martínez *et al*. (2009), a Decision tree was defined to perform inductive learning from observations and logical constructions, like the predictive systems based on rules, that allow to represent and categorize the data subject to analysis.

According to (Chang and Pavlidis, 1977), one of the main advantages of using Decision trees is that they can decompose a process that has many factors in a set of processes of less size and obtain solutions easier to interpret.

Naïve Bayes

A Bayesian network is considered, according to Edwards (1998; Edwards and Fasolo, 2001), a structure composed by four levels. In the higher level, you can find a set of variables represented by nodes and arrows that are related in terms of influence. In a lower level, you can find the levels or states, also known as space of states (Nadkarni and Shenoy, 2001; 2004) that can assume each of the variables of the model. In third place, the level is composed by a set of functions of conditional probability, one for each node, where you can find the probability of occurrence of each state of the variable, considering the possible values of the variables that determine their value. In the lower level you can find a set of algorithms that allow the network to recalculate the probabilities assigned to each level when there is new evidence about the model.

It is relevant to highlight that a Bayesian network is based on two elements, a qualitative dimension and a quantitative dimension (Martínez *et al*., 2003). The qualitative dimension is based on graph theory and probability theory (Ríos, 1995). According to (Spirtes *et al*., 2000) a Bayesian network is a type of graph called Acyclic Directed Graph (ADG).

There are three key elements that form the quantitative dimension of a Bayesian network: The probability concept, the Bayes Theorem and the probability conditional functions. The probability can be understood as something subjective, such as the level of belief of an event (Dixon and Pastor, 1970) and this concept of probability is called Bayesian and is derived from the principle of insufficient reasoning or uncertainty principle (Cowell *et al*., 1999).

The Bayes Theorem is deducted from the axiom that relates the probability of the event intersection and the conditional probability, which can help to work in an efficient way with the propagation of probabilities in graphic models in terms of conditional dependence or independence (Cowell *et al*., 1999). Next, you can see in Fig. 1, an example of the structure of a decision tree.

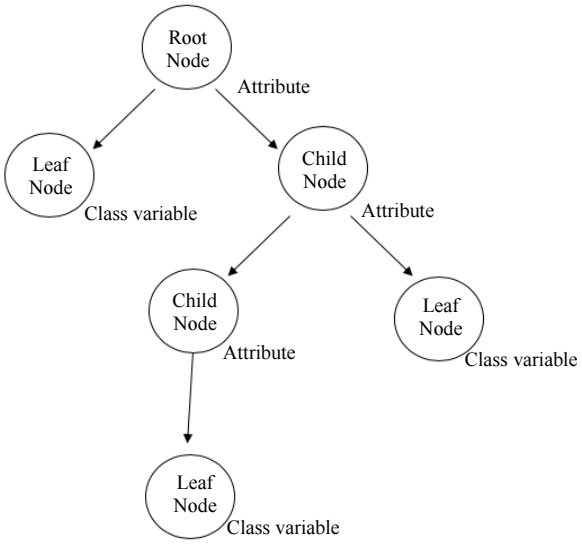

Fig. 1: Decision Tree Structure (Martínez *et al*., 2009)

A Bayesian network basically updates the probabilities inside an acyclic directed graph, considering the conditional independence principles when new evidence is added to the model.

A Bayesian network needs a set of conditional probability functions, one for each variable or node in the network, the ones that will be applied the Bayes rule. Specifically, each variable of the network is characterized by a conditional probability table that represents the values that can assume that variable considering the values of the set of variables that is dependent, following Cowell *et al*. (1999).

Logistic Regression

Logistic regression is a multivariant statistic technique that can estimate the existing relationship between a variable dependent non-metric, dichotomic and a set of variables independent of metric and non-metric. Systematically, the logistic regression has two objectives: The first, is to study the influence of the probability of occurrence of a specific event, the presence of several factors or not and the value or level of these; the second is to determine the better fit model that describe this relationship between the response variable and the set of variables to predict as mentioned by (Salcedo, 2002).

According to (Kurt *et al*., 2008), logistic regression is useful for situations where you need to predict the presence or absence of a feature or output, based on values of the set of variables to predict. It is similar to a regression linear model, but it is more appropriate for models where the dependent variable is dichotomic.

The main goal of logistic regression is to model the influence of the variables that need to be predicted related to the probability of occurrence of those variables.

SEMMA Methodology

The SEMMA Methodology was developed by the SAS Institute, including the processes of selection, exploration, modeling of big datasets to discover relevant information or patterns as mentioned by (SASI, 2017).

According to (Moine *et al*., 2011), the SEMMA methodology have five basic phases which are: Sample, Explore, Modify, Model and Assess. From (Olson and Delen, 2008) SEMMA facilitates the statistic exploration, the visualization techniques and the selection and transforming of the relevant variables in prediction, also can model the variables for prediction processes and later validate the precision of the model. In Fig. 2, you can see the relevant aspects and stages of the methodology.

In this study, each phase of the SEMMA methodology was implemented to obtain finer control of the activities developed starting with the data collecting stage to the results stage, so the authors could validate the capacity and quality of the proposed model.

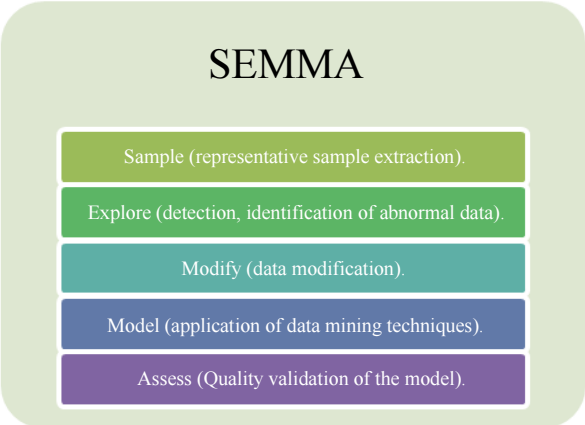

Fig. 2: SEMMA Methodology

Evaluation Methodology for Data Mining Methods and Techniques

This study proposed a software for prediction and detection of obesity levels in young people. With this goal in mind, we performed the stages based on the SEMMA methodology.

First, we proceeded to the dataset creation, from the information collected by the survey, as described in section 3.1.

After the dataset creation, we validated the data, looking for missing values, atypical data and the correlation level between variables, which it is lower than 0.5, so we can be sure that the stored data and the basis for the software implementation and the data mining methods, are correct.

Once the dataset was validated and prepared, the data mining techniques and methods were applied, using the Weka tool, that has a set of algorithms that can be applied to many situations. In this study, the methods used were Decision Trees (J48), Bayesian Networks (Naïve Bayes) and Logistic Regression (Simple Logistic).To validate the model and selecting the best technique, we used the precision metrics Recall, TP Rate and FP Rate. For the training process, we used crossed validation, as mentioned by (Palechor *et al*., 2015), to use part of the data for training and other part for testing, to guarantee optimal results and avoiding over training issues. The proposed model considers classes or categories, the values of underweight, normal, overweight, obesity level I, obesity level II and obesity level III, as you can see in Table 1.

Software Development

The proposed software was based on the dataset created and implements the best data mining technique of the study. Next, you can see the flow diagram of the development of the software in Fig. 3.

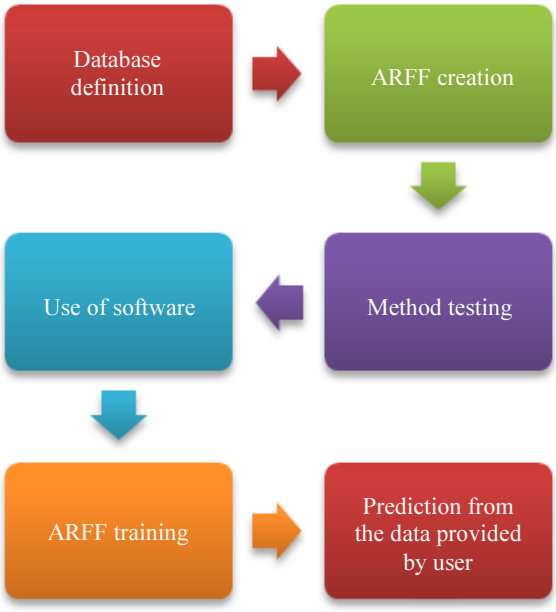

Fig. 3: Flow diagram for software development

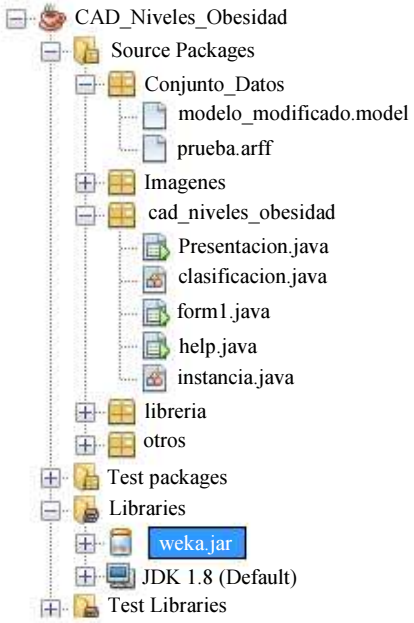

Fig. 4: Using the weka toolkit (weka.jar)

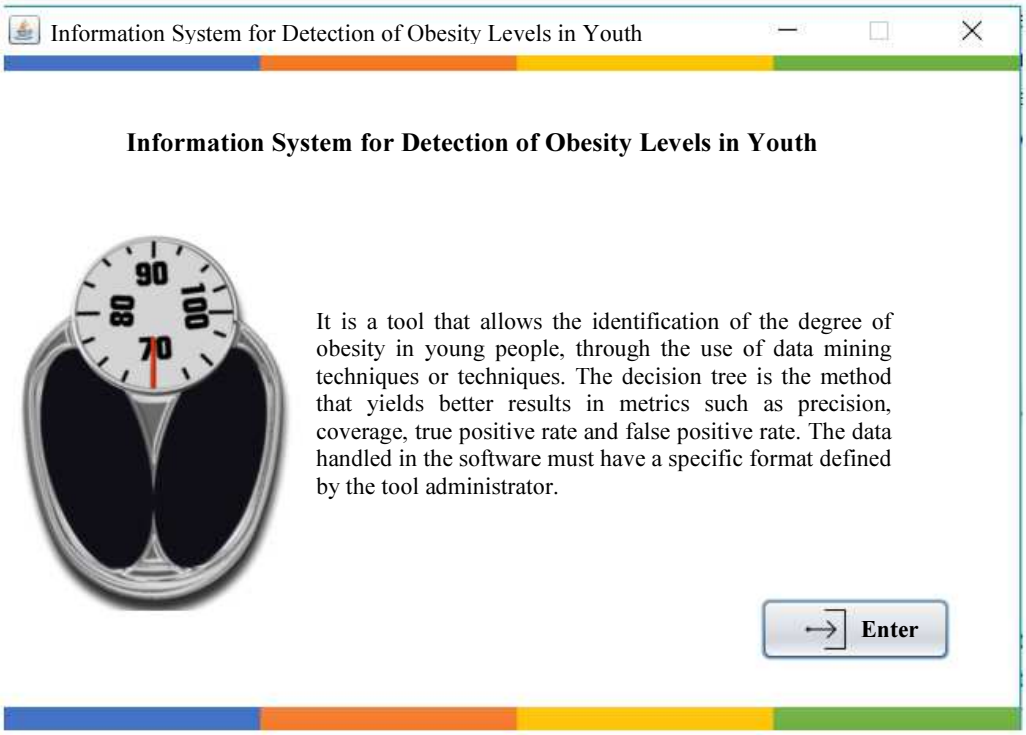

Fig. 5: Presentation Form of the proposed software

For the development of the software we used the NetBeans IDE, based on Java. To be able to use the data mining methods, we added the Weka Toolkit (weka.jar), in Fig. 4 you can see the library import in the tool used for it.

Once the library was imported, we proceeded to coding the classes, methods, procedures and forms as you can see in Fig. 3.

First, we designed the presentation form, to help describe the tool. You can see it in Fig. 5.

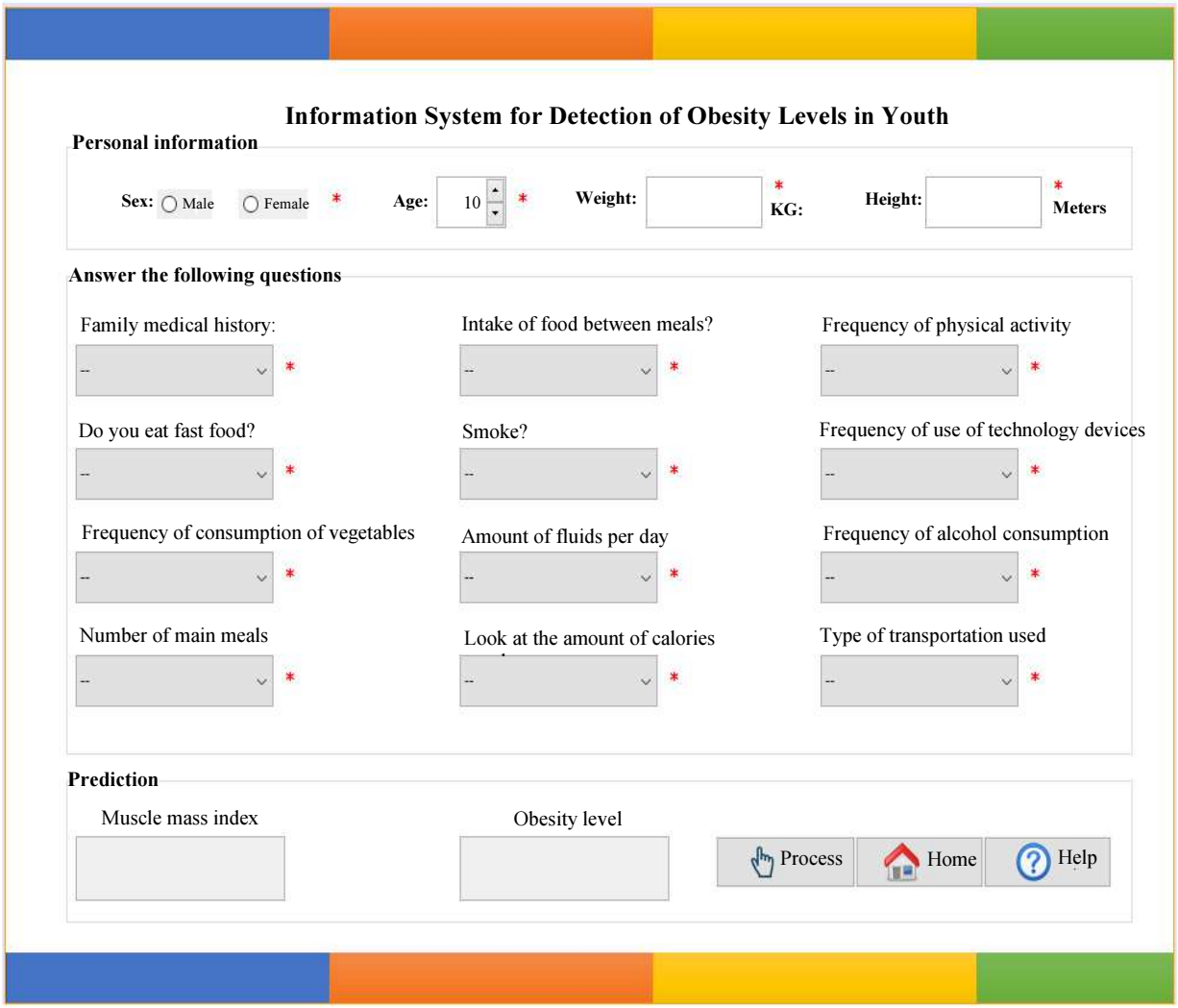

Fig. 6: Form for estimating obesity levels

```
String ruta_dataset="src\\COnjunto_Datos\\prueba.arff";
String ruta_modelo="src\\COnjunto_Datos\\modelo_modificado.model";
```

Fig. 7: Path to the dataset

After the presentation form, the user will find the input form, to receive the variables that we considered as factors for obesity levels. You can see it in Fig. 6.

To avoid missing data, all fields in the input form were validated and are mandatory to make a correct prediction. The tool also must have access to the dataset or training model, which it is loaded automatically by the software.

In Fig. 7, you can see the sentence of code to access the model or dataset to train the tool.

Next, you can find the ARFF file that was used to generate the model and train the tool, as depicted in Fig. 8.

After the training and classification process, you can find the predictions produced by the tool as shown in Fig. 9.

Next, you can see in Fig. 10 the result shown by the tool, after data input to the forms in Fig. 9.

```
@relation obesidadEntrenamiento

@attribute Sexo {H,M}
@attribute Edad numeric
@attribute Estatura numeric
@attribute Peso numeric
@attribute Antecedentes {Si,No}
@attribute C_Rapida {Si,No}
@attribute Verduras {A,S,CN}
@attribute C_Principal {TR,UD,MT}
@attribute E_Comidas {A,S,CS,N}
@attribute Fuma {Si,No}
@attribute Liquidos {UAD,MD,MU}
@attribute Observar {Si,No}
@attribute Ejercicio {NO,UOD,TOC,COS}
@attribute Horario {CAD,TAC,MC}
@attribute Alcohol {NO,CF,SM,D}
@attribute Transporte {TP,CA,AU,MTA,BTA}
@attribute IMC numeric
@attribute Vulnerable {PESOBAJO,NORMAL,SOBREPESO,OBESUNO,OBESDOS,OBES

@data
M,21,1.62,64,Si,No,A,TR,A,No,UAD,No,NO,TAC,NO,TP,24.4,NORMAL
M,21,1.52,56,Si,No,S,TR,A,Si,MD,Si,COS,CAD,CF,TP,24.2,NORMAL
H,23,1.8,77,Si,No,A,TR,A,No,UAD,No,TOC,TAC,SM,TP,23.8,NORMAL
H,20,1.8,87,No,No,S,TR,A,No,UAD,No,TOC,CAD,SM,CA,26.9,SOBREPESO
```

Fig. 8: ARFF File

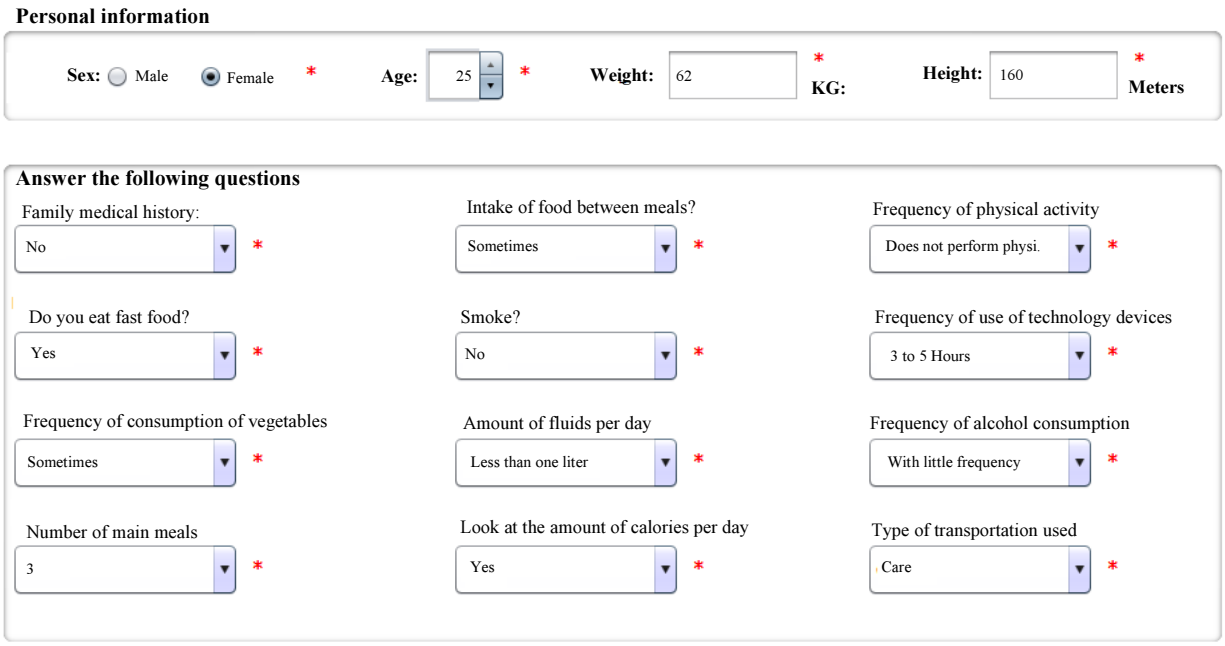

Fig. 9: Form with example data

Fig. 10: Form with results given by the tool

Results and Discussion

Based on the data shown in Table 3 and Fig. 11, the technique with best results was Decision Trees, so we chose the J48 algorithm to be the one selected to implement in the proposed software.

Table 3: Results of the implemented techniques

Method	Precision	Recall	TP. Rate	FP. Rate
J48	97,4%	97,8%	97,8%	0,2%
Naive Bayes	90,1%	91,1%	91,1%	6,0%
Simple Logistic	90,4%	91,6%	91,6%	4,1%

Table 4: Confusion matrix

a	b	c	d	e	f	Classification
40	4	0	0	0	0	a = Underweight
0	360	0	0	0	0	b = Normal
0	0	180	4	0	0	c = Overweight
0	0	0	92	8	0	d = Obesity level I
0	0	0	0	20	0	e = Obesity level II
0	0	0	0	4	0	f = Obesity level III

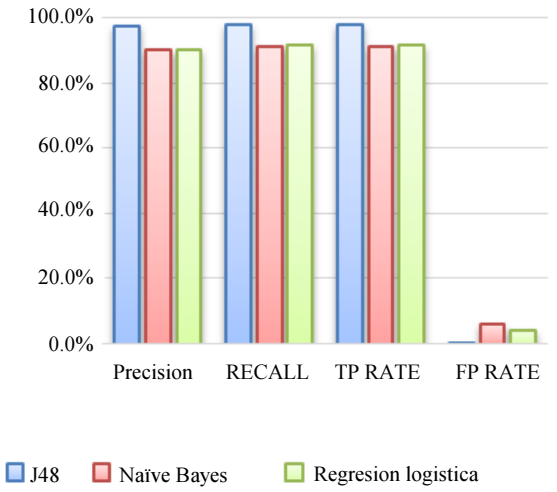

Fig. 11: Results of the implemented techniques

Next, you can see in Table 4, the confusion matrix, where you can find the number of records organized by category.

Conclusion

Obesity is a disease with worldwide exposure, no matter social or cultural level of the people, it is a disease that has doubled since 1980, in 2014 more than 1900 million of adults suffered from it. To help fight this disease, several tools and solutions have been developed to be able to detect or predict the appearance of the disease.

Data mining is an essential tool that allow us to discover information, in our study we used different techniques to achieve best precision rates to detect obesity. According to this, the Decision Trees method obtained 97.4% precision levels to classify users that carry the disease, also the technique shows a TP Rate of 97.8%, which guarantees a high percentage of success to classify data, finally have a FP Rate of 0.2%, a correct value for it. The technique also obtained better results than the values from techniques such as Bayesian Networks and Logistic Regression. The proposed method also surpasses the results obtained in (Adnan and Husain, 2012) that had 75% in precision, (Dugan *et al*., 2015) that obtained 85% and (Husain *et al*., 2013) that showed 73.3%.

The software created in this study allows to classify patients with obesity and it is a clear integration between Weka, Java and NetBeans, integration that can generate many tools to analyze diseases that affect a group of the population, which represents a positive advance in this research area.

Acknowledgement

The authors are grateful to the collaboration from Ana Isabel Oviedo Carrascal, her efforts and counseling were keys to achieve the expected results and the divulgation of these, through this paper.

The authors would like to thank the support of the Universidad de la Costa, this study would have not been possible without it.

Author's Contributions

Dr. Eduardo De la Hoz Correa: Lead research, coordinate developer, doing experiments, adapt analysis and writing the manuscript.

Roberto Morales Ortega and Fabio Mendoza Palechor: Advise research, adapting analysis for the data mining methods part and writing manuscript and proof reading.

Alexis De la Hoz Manotas and Beatriz Sánchez Hernandez: English proof reading and software adapting analysis and result verifications.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that the coauthor has read and approved the manuscript and there are no ethical issues involved.

References

Abdullah, F.S., N.S.A. Manan, A. Ahmad, S.W. Wafa and M.R. Shahril *et al*., 2016. Data mining techniques for classification of childhood obesity among year 6 school children. Proceedings of the International Conference on Soft Computing and Data Mining, Aug. 18-20, Springer, Cham, pp: 465-474. DOI: 10.1007/978-3-319-51281-5_47

Adnan, M.H.B.M. and W. Husain, 2012. A hybrid approach using Naïve Bayes and genetic algorithm for childhood obesity prediction. Proceedings of the International Conference on Computer and Information Science, Jun. 12-14, IEEE Xplore Press, Kuala Lumpeu, Malaysia, pp: 281-285. DOI: 10.1109/ICCISci.2012.6297254

Adnan, M.H.B.M., W. Husain and F. Damanhoori, 2010. A survey on utilization of data mining for childhood obesity prediction. Proceedings of the 8th Asia-Pacific Symposium on Information and Telecommunication Technologies, Jun. 15-18, IEEE Xplore Press, Kuching, Malaysia, pp: 1-6.

Adnan, M.H.M. and W. Husain, 2011. A framework for childhood obesity classifications and predictions using NBtree. Proceedings of the 7th International Conference on Information Technology in Asia, Jul. 12-13, IEEE Xplore Press, Kuching, Sarawak, Malaysia, pp: 1-6. DOI: 10.1109/CITA.2011.5999502

Calabria-Sarmiento, J. C., Ariza-Colpas, P., Pineres-Melo, M., Ayala-Mantilla, C., Urina-Triana, M., Morales-Ortega, R. and I. Echeverri-Ocampo, 2018. Software Applications to Health Sector: A Systematic Review of Literature.

Chang, R.L. and T. Pavlidis, 1977. Fuzzy decision tree algorithms. IEEE Trans. Syst. Man Cybernet., 7: 28-35. DOI: 10.1109/TSMC.1977.4309586

Cowell, R.G., A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter, 1999. Probabilistic networks and expert systems.

Davila-Payan, C., M. DeGuzman, K. Johnson, N. Serban and J. Swann, 2015. Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. Prevent. Chronic Dis., 12: E32-E32. DOI: 10.5888/pcd12.140229

Dixon, J.R. and V.M. Pastor, 1970. Introducción a la Probabilidad: Texto Programado. 1st Edn., Limusa-Wiley, ISBN-10: 9681807200, pp: 418.

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

Dugan, T.M., S. Mukhopadhyay, A. Carroll and S. Downs, 2015. Machine learning techniques for prediction of early childhood obesity. Applied Clin. Inform., 6: 506-520. DOI: 10.4338/ACI-2015-03-RA-0036

Edwards, W. and B. Fasolo, 2001. Decision technology. Annual Rev. Psychol., 52: 581-606. DOI: 10.1146/annurev.psych.52.1.581

Edwards, W., 1998. Hailfinder: Tools for and experiences with Bayesian normative modeling. Am. Psychol., 53: 416-416. DOI: 10.1037/0003-066X.53.4.416

Gómez, M. and L. Ávila, 2008. La obesidad: un factor de riesgo cardiometabólico. Medicina de Familia.

Gutiérrez, H.M., 2010. Diez problemas de la población de jalisco: Una perspectiva sociodemográfica (Primera Edición ed.). Dirección de Publicaciones del Gobierno de Jalisco, Guadalajara, México.

Hernández, G.M., 2011. Prevalencia de sobrepeso y obesidad, y factores de riesgo, en niños de 7-12 años, en una escuela pública de Cartagena septiembre - octubre de 2010. Universidad Nacional de Colombia, Bogota – Colombia.

Husain, W., M.H.M. Adnan, L.K. Ping, J. Poh and L.K. Meng, 2013. My healthy kids: Intelligent obesity intervention system for primary school children. Proceedings of the 3rd International Conference on Digital Information Processing and Communications, (IPC' 13), The Society of Digital Information and Wireless Communication, pp: 627-633.

Kurt, I., M. Ture and A.T. Kurum, 2008. Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease. Expert Syst. Applic., 34: 366-374. DOI: 10.1016/j.eswa.2006.09.004

Manna, S. and A.M. Jewkes, 2014. Understanding early childhood obesity risks: An empirical study using fuzzy signatures. Proceedings of the IEEE International Conference on Fuzzy Systems, Jul. 6-11, IEEE Xplore Press, Beijing, China, pp: 1333-1339. DOI: 10.1109/FUZZ-IEEE.2014.6891838

Martínez, F., M.C. Díaz, M.T. Martín, V.M. Rivas and L.A. Ureña, 2003. Aplicación de redes neuronales y redes bayesianas en la detección de multipalabras para tareas IR. Artculo presentado en las II Jornadas de Tratamiento y Recuperacin de la Informacin, Madrid.

Martínez, R.E.B., N.C. Ramírez, H.G.A. Mesa, I.R. Suárez and M.D.C.G. Trejo *et al*., 2009. Árboles de decisión como herramienta en el diagnóstico médico. Revista Médica de la Univ. Veracruzana, 9: 19-24.

Moine, J.M., A.S. Haedo and S.E. Gordillo, 2011. Estudio comparativo de metodologías para minería de datos. 13th Workshop de Investigadores en Ciencias de la Computación.

Nadkarni, S. and P.P. Shenoy, 2001. A Bayesian network approach to making inferences in causal maps. Eur. J. Operat. Res., 128: 479-498. DOI: 10.1016/S0377-2217(99)00368-9

Nadkarni, S. and P.P. Shenoy, 2004. A causal mapping approach to constructing Bayesian networks. Dec. Support Syst., 38: 259-281. DOI: 10.1016/S0167-9236(03)00095-2

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

Olson, D.L. and D. Delen, 2008. Advanced Data Mining Techniques. 1st Edn., Springer Science and Business Media, Berlin, ISBN-10: 354076917X, pp: 180.

OMS, 2016. Organización Mundial de la Salud. Obesidad y sobrepeso.

Palechor, F.M., A.D.L.H. Manotas, E.D.L.H. Franco and P.A. Colpas, 2015. Feature selection, learning metrics and dimension reduction in training and classification processes in intrusion detection systems. J. Theoretical Applied Inform. Technol., 82: 291-298.

Palechor, F.M., A. De la Hoz Manotas, P.A. Colpas, J.S. Ojeda and R.M. Ortega *et al*., 2017. Cardiovascular disease analysis using supervised and unsupervised data mining techniques. JSW, 12: 81-90.

Ríos, S., 1995. Modelización. Alianza Universidad, Madrid.

Safavian, S.R. and D. Landgrebe, 1991. A survey of decision tree classifier methodology. IEEE Trans. Syst. Man Cybernet., 21: 660-674. DOI: 10.1109/21.97458

Salcedo, C.M., 2002. Estimación de la Ocurrencia de incidencias en declaraciones de pólizas de importación. Universidad Nacional Mayor de San Marcos, Lima.

SASI, 2017. Data mining and the case for sampling. SAS Institute.

Spirtes, P., C.N. Glymour and R. Scheines, 2000.

Causation, Prediction and Search. 1st Edn., MIT Press, Cambridge, Mass, ISBN-10: 0262194406, pp: 543.

Suguna, M., 2016. Childhood obesity epidemic analysis using classification algorithms. Int. J. Mod. Comput. Sci., 4: 22-26.

Zhang, S., C. Tjortjis, X. Zeng, H. Qiao and I. Buchan *et al*., 2009. Comparing data mining methods with logistic regression in childhood obesity prediction. Inform. Syst. Frontiers, 11: 449-460. DOI: 10.1007/s10796-009-9157-0

Zhingre, O. and P. del Cisne, 2015. Factores de riesgo que influyen en los estudiantes 10-13 añosde la Institución Educativa Mater Dei para desarrollar sobrepeso y obesidad en la vida adulta (Bachelor's Thesis).