Journal of Theoretical and Applied Information Technology
15th June 2021. Vol.99. No 11
© 2021 Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION AND FEATURES SELECTION METHOD FOR OBESITY LEVEL PREDICTION

[1]D.MOLINA, [2]A. DE-LA-HOZ, [3] F. MENDOZA

[1]Student, Universidad de la Costa, Department of Computer Science and Electronics, Colombia

[2]Assistant Professor, Universidad de la Costa, Department of Computer Science and Electronics, Colombia

[3]Associate Professor, Universidad de la Costa, Department of Computer Science and Electronics, Colombia

E-mail: [1]dmolina@cuc.edu.co, [2]adelahoz6@cuc.edu.co, [3]fmendoza@cuc.edu.co

ABSTRACT

Obesity has become one of the world's largest health issues, rich and poor countries, without exception, have each year larger populations with this condition. Obesity and overweight are defined as abnormal or excessive fat accumulation that may impair health according to the World Health Organization (WHO) and has nearly tripled since 1975. Data Mining and their techniques have become a strong scientific field to analyze huge data sources and to provide new information about patterns and behaviors from the population. This study uses data mining techniques to build a model for obesity prediction, using a dataset based on a survey for college students in several countries. After cleaning and transformation of the data, a set of classification methods was implemented (Logistic Model Tree - LMT, RandomForest - RF, Multi-Layer Perceptron - MLP and Support Vector Machines - SVM), and the feature selection methods InfoGain, GainRatio, Chi-Square and Relief, finally, crossed validation was performed for the training and testing processes. The data showed than LMT had the best performance in precision, obtaining 96.65%, compared to RandomForest (95.62%), MLP (94.41%) and SMO (83.89%), so this study shows that LMT it can be used with confidence to analyze obesity and similar data.

Keywords: *Data mining, Dataset, Obesity, Decision Trees, Support Vector Machines*

1. INTRODUCTION

Obesity has become one of the world's largest health issues, rich and poor countries, without exception, have each year larger populations with this condition. Obesity and overweight are defined as abnormal or excessive fat accumulation that may impair health according to the World Health Organization (WHO) [1] and has nearly tripled since 1975.

The Global Burden of Disease study accounted 4.7 million people died prematurely in 2017 as a result of obesity [2], and the World Health Organization (WHO) showed that 38.9% of adults worldwide were overweight by 2019 [3].

Being such a serious condition for health, many studies have been working with obesity influence factors, predicting risks and analyzing food intake behaviors, especially in the field of data mining, studies as [4], [5], [6], [7], [8], [9], [10], [11] and [12].

For this study, the dataset built by [13] was used, the authors searched in literature the main factors or habits that are associated with obesity, and their dataset has 18 variables that can establish if a subject suffers obesity. You can find other datasets to be used for data mining related to obesity in adults, adolescents and children such as [14] [15] and [16].

The goal of the current study is to build a model to predict obesity levels using a dataset based on a survey for college students in several countries. After cleaning and transformation of the data, the next step was to generate, select and extract the attributes, then, a set of classification methods were implemented (Logistic Model Tree - LMT, Decision Trees - RandomForest, Multi-Layer Perceptron - MLP and Support Vector Machines - SMO), combined with the methods for attributes selection InfoGain, GainRatio, Chi-Square and Relief, finally, crossed validation was performed for the training and testing processes. This paper is organized with the structure: section 2 shows the previous works related to the research and similar approaches, followed by section 3 Materials and Methods where you can find the information about the dataset used and the different methods used for the generation of the data mining model, then in section 4 you can find

all the information about the methodology applied for the experimentation, in section 5 you can see the results obtained by the study and finally in section 6 you can find the conclusions of the findings in the study. The current study shows a complete implementation of data mining methods over a obesity dataset, combined with several attributes selection methods, so far it does not include any modification of the algorithms used or optimization in the metrics applied to the dataset.

2. PREVIOUS WORKS

Researchers have predicting obesity using data mining for years, some of the most recent and relevant studies are shown next.

The goal of this study [17] is to simulate the risk factor by using statistical tools (SPSS), to predict the major risk factor of obesity by testing the class level attribute according to cross-sectional study with other attributes, using the collected data of 259 individuals from urban and rural areas. The study proposed a risk mining technique (PRMT) using different data mining classifiers, obtaining Naïve Bayes as the best classifier for the 10-fold cross-validation study.

The goal of this study [18] is to implement a collection of methods in data mining using supervised and unsupervised learning to detect individuals suffering obesity, the algorithms used were Decision Trees (DT), Support Vector Machines (SVM) and Simple K-Means through a dataset of 178 college students. The authors used the metrics Precision (98.5%), Recall (98.5%), TPR (98.5%), FPR (0.2%) and ROC area (99.5%), surpassing the results obtained in previous studies with precision levels of 75% and 85%.

The authors of this study [13] applied the SEMMA methodology to explore a dataset using several methods of data mining including: Decision Trees (J48), Bayesian Networks (Naïve Bayes) and Logistic Regression (Simple Logistic), with the metrics Precision, Recall, TPR and FPR, their results showed that J48 obtained the best results with a precision rate of 97.4%.

In this study [19], the authors proposed an automated diagnosis system for breast cancer, using data mining methods, specifically the decision trees algorithm is implemented showing the resulting rules visually, the dataset for the study was the Breast Cancer Coimbra Dataset, the data used for their approach includes leptin, resistin, body mass index and others, all associated with breast cancer. To test the accuracy of the model, it was used ROC Analysis, obtaining 90.52% by the Gini Algorithm, indicating that new automated tools can be safely used for an early diagnosis.

The goal of this study [12] is to evaluate several feature selection methods based on different classifiers to detect obesity in children being 12 years old. The dataset used was provided by the SEGAK Assessment program including 153 primary schools, obtaining 4245 individuals. The classifiers included in this study were Bayesian Networks, Decision Trees, Neural Networks and Support Vector Machines (SVM), the results showed that Decision Trees (J48) and Support Vector Machines (SMO) obtained higher accuracy than the rest of the methods applied, with values over 82%.

In this study [20], authors use a data mining approach to identify salivary biomarkers and other factors associated with obesity in 700 children of Kuwait. The methods applied were Logistic Regression (Lasso), Multivariate Adaptive Regression Spline (MARS), Random Forest (RF) and Boosting classification trees (BT). To validate the performance, the study used a ROC Analysis. Their results showed that increased waist circumference is closely related to adipocyte, indicating a good measure to characterize obesity, being more sensitive than BMI.

In this study [21] their goal is to predict obesity in patients, comparing methods like Logistic Regression, Random Forest (RF), Decision Trees (DT) and Neural Networks, the dataset used was provided by the American Time Use Survey (ATUS) collected by the US Department of Labor. The results showed that Random Forest (RF) obtained the best accuracy with 72%, which it is considered not enough for these models, but these results can improve adding variables to the dataset.

The goal of this study [22] is to analyze different trends for each BMI category to predict obesity and its consequences, using Support Vector Machines (SVM) and ROC to calculate the accuracy of the categories. The dataset used was Standard Report of Non-Communicable Diseases collected by World Health Organization (WHO). Their results showed that BMI >= 25 had the highest accuracy (98.46%) and represents an alarming point that must be prevented to avoid chronic diseases due to the gradual increases in BMI.

The authors of this study [23] developed a web application based on data mining to diagnose obesity and provide information about the BMI ranges. The dataset used was provided by Size Korea including 6413 subjects. The methods used for the model were: Multiclass Random Forest, Multiclass Decision Jungle, Multiclass Neural Network and Multiclass Logistic Regression. ROC Curve was used to compare the different methods obtaining Random Forest (RF) as the best method based on its accuracy of 99%, and the web application to predict obesity is appropriate for the Korean characteristics, to analyze other countries and regions, it needs new data to improve their model.

The goal of this study [24] is to predict obesity prevalence using the perspective of the nutrition datasets, their experiment implemented the Decision Tree (J48) algorithm to predict the individual caloric intake, the dataset used included information of 35 households in Shah Alam Selangor with a total of 170 subjects, the results showed an accuracy of 89.41% using metrics such as TPR, FPR, F-Measure representing a valid model to predict obesity in the subjects.

The authors in [25] proposed a model based on data mining using K-Means algorithm to improve treatment for obesity, their dataset used samples of 200 subjects, and their goal was to improve the duration of it, their findings demonstrated the stability and accuracy of the algorithm and the model.

In [26] authors designed a portable phenotyping system with capabilities to store key components of rule-based systems to facilitate the reuse, adaptation and extension of NLP systems. Their solution was implemented in i2b2 Obesity Challenge, working with 1249 patient textual discharge summaries. Their study used four types of machine learning algorithms including LR, SVM, DT and RF and the results were evaluated with the metrics micro and macro-averaged precision, recall and f-measure. Their findings showed best results with DT algorithm with values of 95.38% for F-Micro with textual classification and 92.85% for intuitive classification.

The study in [27] implemented Fuzzy Cognitive Maps for the obesity problems, applying dimensionality reduction techniques. The authors used a prepared dataset, including the values proposed by the domain experts through a survey. The dataset included 23 concept values that define the obesity problem. The study applied the Non-Linear Hebbian Learning algorithm and their results confirmed it performed efficiently.

The study in [28] features Genome-Wide association studies to analyze Single Nucleotide Polymorphisms (SNPs) proposing a solution, SAERMA, to perform classification through a multi-layer perceptron neural network (MLPNN) for extreme obesity. Their results obtained 77% AUC with 204 SNPs compressed to 100 units. The cases dataset used was the eMerge Genome-Wide Association Studies of Obesity Project and the case control dataset used was the database from Genotypes and Phenotypes (dbGaP).

In [29] authors predicted smoking and obesity prevalence using a lasso-based variable selection and two-level random effects regression. The dataset used was provided by the Behavioral Risk Factor Surveillance System (BRFSS) from 1991 to 2010. Their findings showed 2% error in regions with BRFSS data (61.7% for smoking and 59% for obesity).

The study in [30] applied a data mining approach in data from 1140 children (from 9 to 13 years old), to derive dietary habits related to obesity. The dataset used was provided by the CYKIS study including 24 primary schools and 1140 children with a sample of 634 children because weight and height values were needed to calculate overweight/obesity. The algorithm used for classification was Decision Trees (C4.5), and their findings showed that the patterns revealed have potential to be used as a technique in the field of nutritional epidemiology.

In the study [31] authors created logistic regression and decision trees models for target overweight or obesity. The dataset used was collected by the US Lifestyle survey from the 2011 National Youth Risk Behavior Survey (YRBS). Their results showed that having physical activity, eating breakfast and avoiding tobacco are risk factors to develop obesity in high school students.

The goal of the study [32] is to provide dietary nutrition recommendations utilizing knowledge-based context data through a collaboration filtering method. The context information used in the dietary nutrition recommendation for obese included internal context information and external context information. Their findings showed 80% accuracy in menu recommendation for users.

The authors in [33] created a hybrid model based in data collected from sensors and a social network community, in general, their data comes from four sources: data entered by user, data calculated by the application, data retrieved from sensors and data collected from social network. The algorithm used for the prediction model was K-NN, specifically the FITCKNN MATLAB function.

Finally, in [34] the authors implemented data mining techniques to uncover critical interactions that may exist among drivers of obesity. The dataset used several sources: the Behavioral Risk Factor Surveillance System (BRFSS), the Council For Community and Economic Research (C2ER), the Quarterly Census of Employment and Wages (QCEW), the Bureau of Labor Statistics (BLS), and The Tax Burden on Tobacco data sources. The techniques used to analyze the data were LASSO and Regression Tree techniques. Their findings showed interactions between high income, regular physical exercise and those who experienced mental health challenges.

3. MATERIALS AND METHODS

3.1. Dataset

For this study, the dataset created by [13] was used, the authors searched in literature the main factors or habits that are associated with obesity, and their dataset has 18 variables that can establish if a subject suffers obesity. The information was collected as questions through a survey to college students in countries like Colombia, Mexico and Peru. The data include 178 individuals, 81 men and 97 women with ages between 18 and 25 years. In Table 1 you can see the attributes defined by the dataset.

Table 1: Attributes included in the obesity dataset [13].

Attributes	Values
Sex	H: Male
	M: Female
Age	Numeric Integer Values
Height	Numeric Integer Values (m)
Weight	Numeric Integer Values (kg)
Family with overweight/obesity	Yes
	No
Fast Food Consumption	Yes
	No
Frequency of Vegetables	S: Always
Consumption	A: Sometimes
	CN: Rarely
Number of Meals	1 - 2: UD
	3: TR
	> 3: MT
Intake between meals	S: Always
	CS: Usually
	A: Sometimes
	CN: Rarely
Smokes	Yes
	No
Daily liquid intake	MU: < 1 L
	UAD: 1 - 2 L
	MD: > 2 L
Daily Caloric Calculation	Yes
	No
Physical Activity	UOD: 1 - 2 days
	TAC: 3 - 4 days
	COS: 5 - 6 days
	NO: None
Number of hours for technology use	CAD: 0 - 2 hours
	TAC: 3 - 5 hours
	MC: > 5 hours
Alcohol consumption	NO: None
	CF: Rarely
	S: Weekly
	D: Daily
Transportation	TP: Public transportation
	MTA: Motorbike
	BTA: Bike
	CA: Walking
	AU: Automobile
BMI	WHO classification
Vulnerable	Based on WHO classification

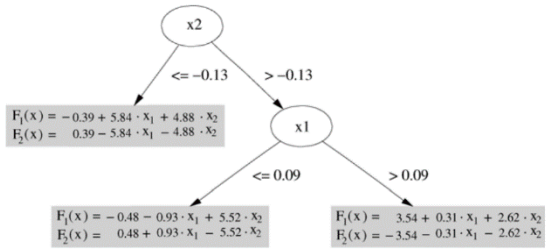
3.2. Classification Methods

3.2.1. Logistic Model Tree (LMT)

Logistic Model Tree (LMT) [31] is widely used in a classification model. LMT is a data mining algorithm that unites two classification algorithms namely Decision Tree (DT) and Logistic Regression (LR). In practice, LMT provides more accurate

results compared to similar algorithms, such as C4.5, CART and LE. You can see a view of an LMT in Figure 1

Figure 1: Logistic Model Tree created by the LMT algorithm for a polynomial dataset [32]



3.2.2. RandomForest - RF

A Decision Tree (DT) [33] is a classification procedure that partitions recursively a dataset in subdivisions, they are usually composed by a root node and a set of internal nodes created through data division and terminal nodes. Each node has a single parent and two or more descendants. A Decision Tree is used to support decision making processes and sets the probability for each possible choice, based on the context of the decision [34]. Decision Trees are commonly implemented through different algorithms such as C.4.5 and Random Forest. In Figure 2 you can see an example of the process for training and classification with a RandomForest Tree.

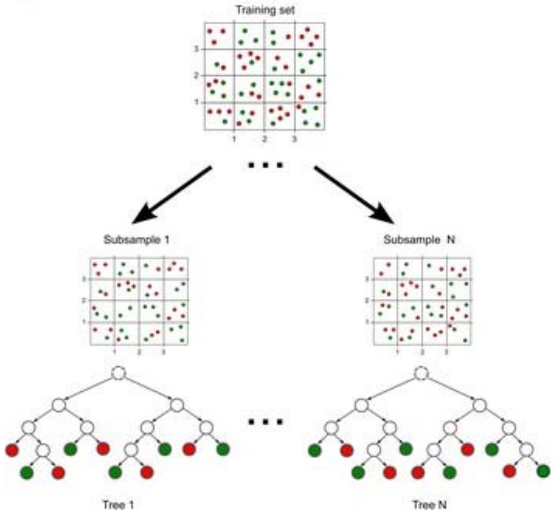


Figure 2: Example of training and classification processes using RandomForest [35]

3.2.3. Multi-Layer Perceptron (MLP)

Multi-layer perceptron (MLP) [36] is a supplement of feed forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer, as shown in Fig. 3. The input layer receives the input signal to be processed. The required tasks, such as prediction and classification, are performed by the output layer. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. Similar to a feed forward network in a MLP the data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the back propagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. The major use cases of MLP are pattern classification, recognition, prediction and approximation. In Figure 4 you can see the architecture of a multi-layer perceptron.

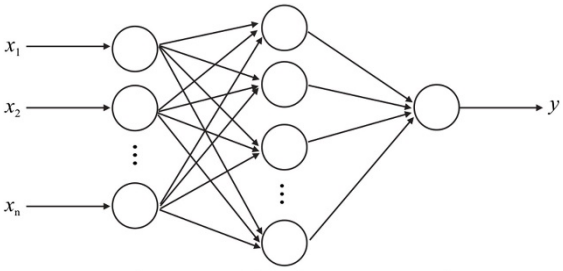


Figure 3: Architecture of a Multi-layer perceptron neural network. [37]

3.2.4. Support Vector Machines (SVM)

Support Vector Machines [38] is the most widely used Machine Learning technique-based pattern classification technique available nowadays. It is based on statistical learning theory and was developed by Vapnik in the year 1995. The primary aim of this technique is to project nonlinear separable samples onto another higher dimensional space by using different types of kernel functions. SVM has been applied extensively for researchers worldwide in different tasks such as handwritten digits recognition [39], object recognition [40], text classification [41] and human activities recognition [42]. In Figure 4, you can see an example of classification performed by Support Vector Machines.

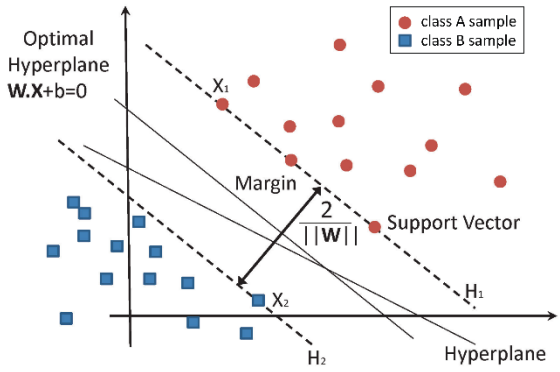


Figure 4: Classification of data by Support Vector Machines (SVM) [43]

3.3. Attribute Selection Algorithms

3.3.1. Infogain

Infogain is considered a symmetrical measure of the variation (reduction) of the entropy in the variable Y when increasing the information from variable X [44]. It has been used in studies that designed tools for indoor localization in wireless environments [45], chronic disease prediction (osteoporosis and arthritis) [46] and bioluminescent protein prediction (BLProt) [47].

3.3.2. GainRatio

GainRatio is a non-symmetric derivation of InfoGain designed to compensate its natural bias, it represents a normalized version of InfoGain [44]. It has been used in studies for classification algorithm creation based on weighted class distribution [48], tools for predicting configuration errors in a system (EFSPredictor) [49] and a predictive model for metabolism mediated by cytochrome P450s (CYP450s), 3A4, 2D6 and 2C9 [50].

3.3.3. Relief

Relief is a measure to evaluate the value of a feature when sampling an instance and compare it to the closest instance of similar and different classes. A weight is defined for each feature and those that surpass a threshold are selected as relevant features, such weight is calculated from probability of the closest neighbors in the same class and same value of the feature, which means that a greater difference between probabilities corresponds to greater the relevance of the feature. [44]

3.3.4. Chi-Square

Chi-Square is a measure to evaluate the value of a feature based on the chi-square statistic compared to its class. The initial hypothesis is always that features are not related, so the greater the value of chi-square, more evidence against the initial hypothesis. [44]

4. METHODOLOGY

The goal of the current study was to create a model to predict obesity levels using the dataset developed by [13]. The process started with the cleaning and transformation of the data, and generating, selecting and extracting the attributes, then the classification methods were implemented using crossed validation for the training and testing processes. Data mining methods literature is extense and diverse, to optimize the analysis for the current data, authors chose the methods Logistic Model Tree (LMT), Random Forest (RF), Multi-Layer Perceptron (MLP) and Support Vector Machines (SVM), due their widely use in previous studies and their different approaches to the data, to obtain a comparison more valuable and practical to other researchers. The addition of using a set of attribute selection algorithms was also considered to compare the specific combination of the classification and selection algorithms, in this area, only the authors considered InfoGain, GainRatio, Relief and Chi-Square, since they have been widely implemented in the literature.

4.1. Data Cleaning and Transformation

The data cleaning and transformation process involved identification of atypical values and detection of redundant and missing data, since the occurrence of these type of values, affects directly the learning process of the implemented algorithms.

4.2. Obesity Level Classification

The classification process implemented the feature selection methods InfoGain, GainRatio, Relief and Chi-Square combined with the algorithms LMT, RandomForest, MLP and SMO, to observe the behavior of the evaluation metrics when variables are excluded, for these several test scenarios were set, taking 5, 8, 12 and 16 features. Additionally, the cross-validation method with 10 folds was performed to divide the dataset with these iterations, one section for training and one section for the sets of the data created previously.

5. RESULTS

In Table 2, you can see the precision results for the algorithm LMT with the methods GainRatio, InfoGain, Relief and ChiSquare.

Journal of Theoretical and Applied Information Technology
15th June 2021. Vol.99. No 11
© 2021 Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: LMT Algorithm precision with classifiers.

Algorithm	ATTR	TPR (%)	FPR (%)	Precision (%)	Recall (%)
LMT + GainRatio + 10-fold Cross Validation	5	96,4	0,06	96,4	96,4
	8	96,9	0,05	96,9	96,9
	12	**97,3**	**0,04**	**97,3**	**97,3**
	16	97,1	0,05	97,1	97,1
LMT + InfoGain + 10-fold Cross Validation	5	97,4	0,04	97,4	97,4
	8	96,9	0,05	96,9	96,9
	12	**97,3**	**0,04**	**97,4**	**97,3**
	16	96,9	0,05	96,9	96,9
LMT + Relief + 10-fold Cross Validation	**5**	**97,1**	**0,05**	**97,1**	**97,1**
	8	96,8	0,05	96,8	96,8
	12	96,4	0,06	96,4	96,4
	16	97,0	0,05	97,0	97,0
LMT + ChiSquare + 10-fold Cross Validation	5	81,0	0,31	80,7	81,0
	8	97,3	0,04	97,3	97,3
	12	**97,4**	**0,04**	**97,5**	**97,4**
	16	97,4	0,04	97,4	97,4

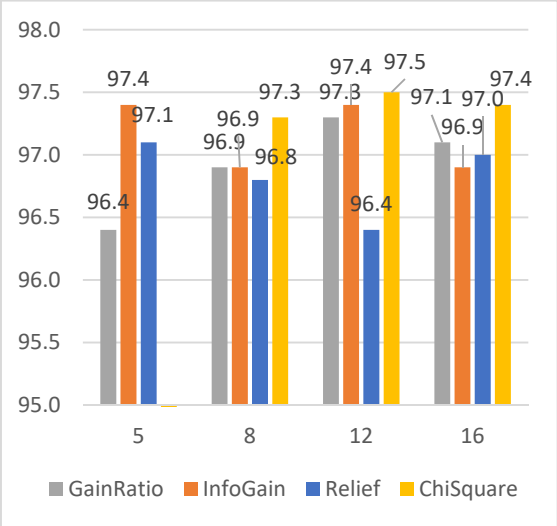


Figure 5: Precision for LMT Algorithm and classifiers

In Table 3, you can see the results for the algorithm DT with the methods GainRatio, InfoGain, Relief and ChiSquare.

Table 3: DT Algorithm precision with classifiers.

Algorithm	ATTR	TPR (%)	FPR (%)	Precision (%)	Recall (%)
	5	95,7	0,07	95,9	95,7
DT + GainRatio + 10-fold Cross Validation	8	95,5	0,07	95,7	95,5
	12	**95,8**	**0,07**	**96,0**	**95,8**
	16	95,5	0,07	95,7	95,5
DT + InfoGain + 10-fold Cross Validation	**5**	**95,9**	**0,07**	**96,1**	**95,9**
	8	95,5	0,07	95,7	95,5
	12	95,7	0,07	95,9	95,7
	16	95,5	0,07	95,8	95,5
DT + Relief + 10-fold Cross Validation	5	95,7	0,07	95,9	97,5
	8	**95,8**	**0,07**	**96,0**	**95,8**
	12	95,7	0,07	95,9	95,7
	16	95,6	0,07	95,8	95,6
DT + ChiSquare + 10-fold Cross Validation	5	90,5	0,16	90,5	90,5
	8	96,8	0,05	96,8	96,8
	12	95,9	0,07	96,0	95,9
	16	95,8	0,07	96,0	95,8

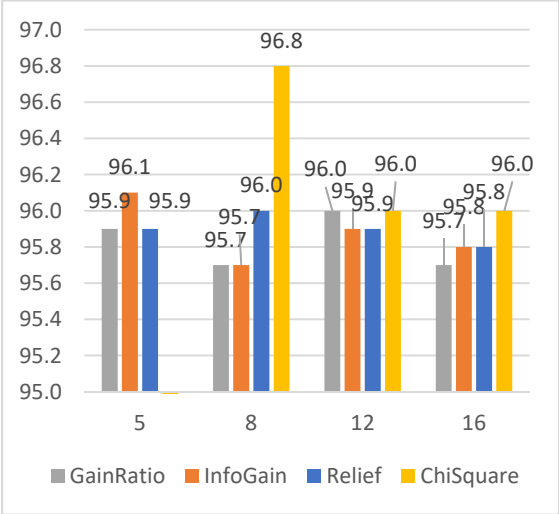


Figure 6: Precision for DT Algorithm and classifiers

In Table 4 you can see the results for the algorithm MLP with the methods GainRatio, InfoGain, Relief and ChiSquare.

Table 4: MLP Algorithm precision with classifiers.

Algorithm	ATTR	TPR (%)	FPR (%)	Precision (%)	Recall (%)
MLP + GainRatio + 10-fold Cross Validation	5	75,6	4,0	74,9	75,6
	8	79,5	3,3	79,5	79,5
	12	82,7	2,8	82,6	82,7
	16	**93,9**	**1,0**	**93,9**	**93,9**
MLP + InfoGain + 10-fold	5	70,6	4,9	70,2	70,6
	8	79,3	3,4	78,9	79,3

Cross Validation	**12**	**94,6**	**0,9**	**94,7**	**94,6**
	16	94,4	0,9	94,4	94,4
MLP + Relief + 10-fold Cross Validation	5	78,0	3,6	78,1	78,0
	8	82,0	3,0	81,7	82,0
	12	**93,7**	**1,0**	**93,7**	**93,7**
	16	93,6	1,0	93,6	93,6
MLP + ChiSquare + 10-fold Cross Validation	5	65,9	5,2	67,5	69,5
	8	**95,9**	**0,7**	**95,9**	**95,9**
	12	95,4	0,8	95,4	95,4
	16	94,4	0,9	94,4	94,4

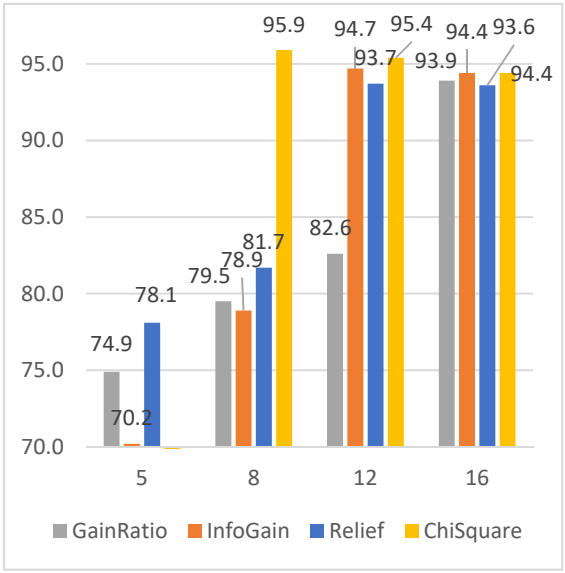


Figure 7: Precision for MLP Algorithm and classifiers

SMO + Relief + 10-fold Cross Validation	5	68,5	5,3	68,0	68,5
	8	71,4	4,8	70,5	71,4
	12	83,1	2,7	83,1	83,1
	16	**84,2**	**2,5**	**84,1**	**84,2**
SMO + ChiSquare + 10-fold Cross Validation	5	64,9	5,8	63,9	64,9
	8	81,8	2,9	83,2	81,8
	12	84,1	2,6	84,2	84,1
	16	**84,3**	**2,5**	**84,2**	**84,3**

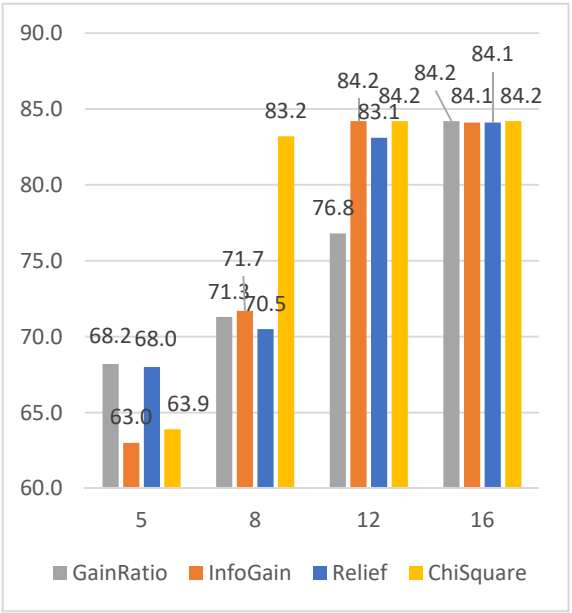


Figure 8: Precision for SMO Algorithm and classifiers

In Table 5 you can see the results for the algorithm RandomForest with the methods GainRatio, InfoGain, Relief and ChiSquare.

Table 5: SMO Algorithm precision with classifiers.

Algorithm	ATTR	TPR (%)	FPR (%)	Precision (%)	Recall (%)
SMO + GainRatio + 10-fold Cross Validation	5	69,6	5,1	68,2	69,6
	8	72,1	4,7	71,3	72,2
	12	77,3	3,8	76,8	77,3
	16	**84,2**	**2,5**	**84,2**	**84,2**
SMO + InfoGain + 10-fold Cross Validation	5	63,7	6,1	63,0	63,7
	8	73,0	4,4	71,7	73,0
	12	84,0	2,6	84,2	84,0
	16	**84,2**	**2,5**	**84,1**	**84,2**

In Table 6 you can see the final results for the set of algorithms implemented and the correspondent evaluation metrics and Figure 9 shows the precision for all algorithms with 16 attributes, being LMT with 97.50% in precision, the best performance for this study. Table 6 represents a comparison between the best results obtained combining the different attribute selection algorithms with the classification methods described in Sections 3.2 and 3.3, these methods were implemented under the same test scenario and using 10-fold cross validation.

Table 6: Final comparison results with 16 attributes.

Algorithms	ATR	TPR (%)	FPR (%)	PRECISION (%)	RECALL (%)
LMT + ChiSquare	12	97,4	0,04	97,5	97,4
Random Forest + ChiSquare	8	96,8	0,05	96,8	95,9
MLP + ChiSquare	8	95,9	0,7	95,9	95,9
SMO + ChiSquare	16	84,3	2,5	84,2	84,3

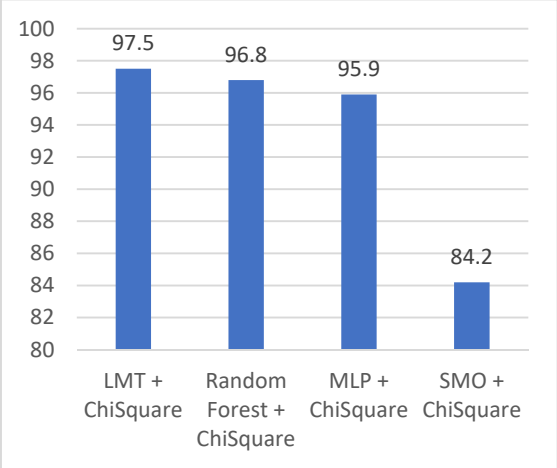


Figure 9: Precision for all algorithms with 16 attributes.

6. DISCUSSION

The detection of problems related to obesity, is a research field with increasing interest by data analysis scientists, using historic information to create intelligent tools to detect if an individual suffers from a disease. A critical factor to build data mining models, is the amount of data required to make estimations properly, the attribute selection algorithms have a relevant role in this process, since they allow to choose the right number of variables to used in a specific model.

The study obtained a comparative analysis in the accuracy of the classification methods in combination with the attribute selection techniques. The algorithm with best results was ChiSquare, with higher precision levels with the LMT, RF, MLP and SMO algorithms, as you can see in Table 6. The best results for all algorithms implemented, based on the validation metrics TPR, FPR, Precision and Recall, was the LMT Algorithm, using only 12 features of 17 possible and available in the original dataset.

The results of the study demonstrated that LMT was the best algorithm with 97.5% in precision, this result is quite close to other studies such has [17] with accuracy of 99.2% using Naïve Bayes, [18] with precision of 98.5% using Decision Trees combined with Simple K-Means, [13] with precision of 97.4% using Decision Trees, [22] with accuracy of 98.46% using Support Vector Machines (SVM) and [23] using RandomForest and obtaining precision of 99%.

These results show clearly that the current implementation in this study, have good performance and to the same level of precision than previous studies in the literature, and confirms the resulting model, can be used to predict correctly obesity levels, nevertheless, these results improved the findings in other studies such as [19] implementing Decision Trees (GINI) with 90.5% accuracy, [12] using Decision Trees (J48) with 82.72% accuracy, [21] using RandomForest with 91% accuracy and [24] using Decision Trees (J48) with 89.41% in accuracy.

7. CONCLUSIONS

The goal of the current study was to create a model to predict obesity levels using the dataset based on a survey for college students in several countries. After cleaning and transformation of the data, the next step was to select and extract the attributes, then, a set of classification methods were implemented (LMT, RandomForest, MLP and SMO), combined with the methods for attributes selection InfoGain, GainRatio, Chi-Square and Relief, finally crossed validation was performed for the training and testing processes. The results showed that using data mining classification methods combined with attribute selection methods can improve the precision of the algorithms, but the number of attributes chosen, it does not always reflect an improvement of the performance, it needs a further study of the effects of the number of attributes in the selection to determine their effect and the causes why the performance varies from set to set. The data showed than LMT had the best performance in precision, obtaining 97.5%, compared to RandomForest (96.8%), MLP (95.9%) and SMO (84.2%).

REFERENCES:

[1] WHO, «Obesity and overweight,» 2020. [En línea]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] O. W. i. Data, «Obesity - Our World In Data,» 2020. [En línea]. Available: https://ourworldindata.org/obesity#:~:text=13%25%20of%20adults%20in%20the,of%20energy%20intake%20and%20expenditure..

[3] Statista, «Overweight prevalence by age,» 2020. [En línea]. Available: https://www.statista.com/statistics/1065605/prevalence-overweight-people-worldwide-by-age/.

[4] C. Davila-Payan, M. DeGuzman, K. Jhonson, N. Serban y J. Swann, «Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data,» *Preventing Chronic Disease,* vol. 12, 2015.

[5] S. Manna y A. Jewkes, «Understanding early childhood obesity risks: An empirical study using fuzzy signatures,» *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* 2014.

[6] T. Dugan, S. Mukhopadhyay, A. Carroll y S. Downs, «Machine learning techniques for prediction of early childhood obesity,» *Applied Clinical Informatics,* 2015.

[7] M. H. Muhammad Adnan, W. Husain y N. Abdul Rashi, «A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction,» *2012 International Conference on Computer & Information Science (ICCIS),* 2012.

[8] M. H. Muhammad Adnan, W. Husain y N. Abdul Rashi, «A framework for childhood obesity classifications and predictions using NBTree.,» *Information Technology in Asia (CITA 11),* pp. 1-6, 2011.

[9] M. H. Muhammad Adnan, W. Husain y F. Damanhoori, «A survey on utilization of data mining for childhood obesity prediction,» *Information and Telecommunication Technologies (APSITT),* pp. 1-6, 2010.

[10] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan y J. Keane, «Comparing data mining methods with logistic regression in childhood obesity prediction,» *Information Systems Frontiers,* pp. 449-460, 2009.

[11] M. Suguna, «Childhood obesity epidemic analysis using classification algorithms,» *Int. J. Mod. Comput. Sci,* Vols. %1 de %222-26, 2016.

[12] F. Abdullah, N. Manan, A. Ahmad, S. Wafa, M. Shahril, N. Zulaily y A. Ahmed, «Data Mining Techniques for Classification of Childhood Obesity Among Year 6 School Children,» *International Conference on Soft Computing and Data Mining,* pp. 465-474, 2016.

[13] E. De-La-Hoz-Correa, F. Mendoza-Palechor, A. De-la-Hoz-Manotas, R. Morales-Ortega y B. Sanchez Hernandez, «Obesity Level Estimation Software based on Decision Trees,» *Journal of Computer Science,* vol. 15, 2019.

[14] A. Aora, «Obesity among adults by country, 1975-2016,» 2020. [En línea]. Available: https://www.kaggle.com/amanarora/obesity-among-adults-by-country-19752016.

[15] Eurostat, «Obesity rate by Body Mass Index,» 2020. [En línea]. Available: https://data.europa.eu/euodp/en/data/dataset/A2eMGcMJTMLVVWbsvAlr8w.

[16] Center for Disease Control and Prevention, «Nutrition, Physical Activity, and Obesity - Behavioral risk factor surveillance system,» 2020. [En línea]. Available: https://healthdata.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system.

[17] R. Hossain, S. Hasan, M. A. Hossin, S. R. Haider Noori y H. Jahan, «PRMT: Predicting risk factor of obesity among middle-aged people using data mining techniques,» *Procedia Computer Sciences,* vol. 132, pp. 1068-1076, 2018.

[18] R. Cañas Cervantes y U. Martinez Palacio, «Estimation of obesity levels based on computational intelligence,» *Informatics in Medicine Unlocked,* 2020.

[19] S. Akben, «Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System,» *IRBM,* vol. 40, 2019.

[20] S. Ping y M. Goodson, «A Data Mining Approach Identified Salivary Biomarkers That Discriminate between Two Obesity Measures,» *Journal of Obesity,* 2019.

[21] A. Joshi, T. Choudhury, A. S. Sabitha y K. Srujan Raju, «Data Mining in Healthcare and Predicting Obesity,» *Proceedings of the Third International Conference on Computational Intelligence and Informatics,* vol. 1090, 2020.

[22] M. Siddiqui, R. Morales-Menendez y A. Sultan, «Application of Receiver Operating Characteristics (ROC) on the Prediction of Obesity,» *Brazilian Archives of Biology and Technology,* vol. 63, 2020.

[23] C. Kim y S. Youm, «Development of a Web Application Based on Human Body Obesity Index and Self-Obesity Diagnosis Model Using the Data Mining Methodology,» *Sustainability,* vol. 12, 2020.

[24] N. A. Daud, N. L. Mohd Noor, S. Aljunid, N. Noordin y N. I. Fahmi Teng, «Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity,» *2018 IEEE Conference on Big Data and Analytics (ICBDA),* 2018.

[25] A. Bu y L. Wang, «Research on the Rule of Acupuncture and Moxibustion for Treatment of Obesity Based on Data Mining,» *2016 International Conference on Smart City and Systems Engineering (ICSCSE),* 2016.

[26] H. Sharma, C. Mao, Y. Zhang, H. Vatani, L. Yao, Y. Zhong, L. Rasmussen, G. Jiang, J. Pathak y Y. Luo, «Portable Phenotyping System: A Portable Machine-Learning Approach to i2b2 Obesity Challenge,» *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W),* 2018.

[27] N. Nadar Selvin y A. Srinivasaragahavan, «Dimensionality reduction of inputs for a Fuzzy Cognitive Map for obesity problem,» *2016 International Conference on Inventive Computation Technologies (ICICT),* 2016.

[28] C. Curbelo, P. Fergus, C. Chalmers, N. Hassain Malim, B. Abdulaimma, D. Reilly y F. Falciani, «SAERMA: Stacked Autoencoder Rule Mining Algorithm for the Interpretation of Epistatic Interactions in GWAS for Extreme Obesity,» *IEEE Access,* 2020.

[29] A. Ortega Hinojosa, M. Davies, S. Jarjour, R. Burnett, J. Mann, E. Hughes, J. Balmes, M. Turner y M. Jerrett, «Developing small-area predictions for smoking and obesity prevalence in the United States for use in Environmental Public Health Tracking,» *Environmental Research,* 2014.

[30] C. Lazarou, M. Karaolis, A.-L. Matalas y D. Panagiotakos, «Dietary patterns analysis using data mining method. An application to data from the CYKIDS study,» *Computer Methods and Programas in Biomedicine,* 2012.

[31] A. Pochini, Y. Wu y G. Hu, «Data Mining for Lifestyle Risk Factors Associated with Overweight and Obesity among Adolescents,» *Data Mining for Lifestyle Risk Factors Associated with Overweight and Obesity among Adolescents,* 2014.

[32] H. Jung y K. Chung, «Knowledge-based dietary nutrition recommendation for obese management,» *Information Technology and Management,* 2016.

[33] S. Harous, M. A. Serhani, M. El Menshawy y A. Benharref, «Hybrid obesity monitoring model using sensors and community engagement,» *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC),* 2017.

[34] R. Salehnejad, R. Allmendiger, Y.-W. Chen, M. Ali, A. Shahgholian, P. Yiapanis y M. Mansur, «Leveraging data mining techniques to understand drivers of obesity,» *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),* 2017.

[35] M. Firman Maulana y M. Defriani, «Logistic Model Tree and Decision Tree J48 Algorithms for predicting the length of study period,» *Journal Penelitian Ilmu Komputer, System Embedded & Logic,* vol. 8, pp. 39-48, 2020.

[36] N. Landwehr, M. Hall y E. Frank, «Logistic Model Trees,» *Machine Learning,* vol. 59, 2005.

[37] M. Friedl y C. Brodley, «Decision tree classification of landcover from remotely sensed data,» *Remote sensing of environment,* vol. 61, nº 3, pp. 399-409, 1997.

[38] D. Magerman, «Statistical decision-tree models for parsing,» *Proceedings of the 33rd annual meeting on Association for Computational Linguistics,* vol. 6, pp. 276-283, 1995.

[39] G. Machado, M. Recamonde Mendoza y L. G. Corbellini, «What variables are important in predicting bovine viral diarrhea virus? A random forest approach,» *Veterinary Research,* 2015.

[40] S. Abirami y P. Chitra, «Chapter Fourteen - Energy-efficient edge based real-time healthcare support system,» de *Advances in Computers*, 2020, pp. 339-368.

[41] A. H. Fath, F. Madanifar y M. Abbasi, «Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict soltuion gas-oil ratio of crude oil systems,» *Petroleum,* vol. 6, 2020.

[42] S. Kumar Satapathy, S. Dehuri, A. Kumar Jagadev y S. Mishra, «Chapter 1 - Introduction,» de *EEG Brain Signal Classification for Epileptic Seizure Disorder Detection*, 2019, pp. 1-25.

[43] V. Vapnik, Statistical Learning Theory, Wiley and Sons, 1998.

[44] C. Papageorgiou, M. Oren y T. Poggio, «A general framework for object detection,» *Proceedings of the International Conference on Computer Vision,* 1998.

[45] T. Joachims, «Text categorization with support vector machines,» *Proceedings of the European Conference on Machine Learning,* vol. 4, pp. 137-142, 1998.

[46] Y. Kim y H. Ling, «Human activity classification based on micro-doppler signatures using a support vector machine,» *IEEE Transactions on Geoscience and Remote Sensing,* vol. 47, pp. 1328-1337, 2009.

[47] E. Garcia-Gonzalo, Z. Fernandez-Muñiz, P. J. García Nieto, A. B. Sanchez y M. Menéndez Fernandez, «Hard-Rock stability analysis for span design in entry-type excavations with learning classifiers,» *Materials,* vol. 9, 2016.

[48] J. Novakovic, «Toward optimal feature selection using ranking methods and classification algorithms,» *Yugoslav Journal of Operations Research,* vol. 21, 2016.

[49] S. H. Fang, T. Lin y P. Lin, «Location fingerprinting in a decorrelated space,» *IEEE Transactions on Knowledge and Data Engineering,* vol. 20, pp. 685-691, 2008.

[50] D. Gupta, S. Khare y A. Aggarwal, «A method to predict diagnostic codes for chronic diseases using machine learning techniques,» *2016 International Conference on Computing, Communication and Automation (ICCCA),* pp. 281-287, 2016.

[51] K. Kandaswamy, G. Pugalenthi, M. Hazrati, K. U. Kalies y T. Martinetz, «BLProt: prediction of bioluminiscent proteins based on support vector machines and relieff feature selection,» *BMC bioinformatics,* vol. 12, p. 345, 2011.

[52] H. Zhao y X. Li, «A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism».

[53] B. Xu, D. Lo, X. Xia, A. Sureka y S. Li, «EFSPredictor: Predicting configuration bugs with ensemble feature selection,» *Asia-Pacific Software Engineering Conference (APSEC),* pp. 206-213, 2015.

[54] S.-B. He, M.-M. Li, B. X. Zhang, X. T. Ye, R. F. Du, Y. Wang y Y. J. Qiao, «Construction of Metabolism prediction models for CYP450s, 3A4, 2D6 and 2C9 based on microsomal metabolic reaction system».