# Comparative analysis of data mining techniques to classify and predict overweight in children and adolescents


**Original Research Article**

**ISSN :** 2456-1045 (Online)
**(ICV-MDS/Impact Value):** 63.78
**(GIF) Impact Factor:** 4.126
**Publishing Copyright @** International Journal Foundation
**Journal Code:** ARJMD/MDS/V-32.0/I-1/C-11/DEC-2018
**Category :** MEDICAL SCIENCE
**Volume :** 32.0 /**Chapter**- XI / **Issue** -1(DECEMBER-2018)
**Journal Website:** www.journalresearchijf.com
**Paper Received:** 25.12.2018
**Paper Accepted:** 03.01.2019
**Date of Publication:** 17-01-2019
**Page:** 56-65

Name of the Author (s):

**José Sulla-Torres[1], Rossana Gómez-Campos[2], Marco Cossio-Bolaños[1,2], Luis Alfaro-Casas[1], Francisco Antonio Pereira Fialho[3]**

*[1]Universidad Nacional de San Agustín de Arequipa, Perú*

*[2]Universidad Católica de Maule, Talca-Chile,*

*[3]Federal University of Santa Catarina - Florianopolis - SC – Brazil*

**Citation of the Article**

**Torres J.S. ; Campos R.G. ; Bolaños M.C. ; Casas L.A.; Fialho F.A.P (2018)** Comparative analysis of data mining techniques to classify and predict overweight in children and adolescents.; *Advance Research Journal of Multidisciplinary Discoveries.32(11)pp. 52-65*

**ABSTRACT**

The objective of this study was to compare which of the five popular algorithms for the classification of data mining (i.e., Classifier K-NN, Decision Tree, Bayesian Web, Efficient Bayesian Multivariate Classifier, and Forest by Penalizing Attributes) comes closest to reality in terms of being able to classify and predict excess body weight in children and adolescents. Towards the objective of analyzing the data, a database of 5,962 primary and secondary school students was utilized. The subjects, whose ages range from six to seventeen years, are from the region of Itaupú-Paraná, Brazil. The program used for the data mining was WEKA (Waikato Environment for Knowledge Analysis). The data mining was applied following the methodology of Oracle Technology Network (OTN), and the data were analyzed in three groups: boys, girls, and both sexes together. The results indicate that the technique J-48 proved to be the most realistic classifier and can thus contribute to predicting the risk of excess body weight in children and adolescents in the region of Lago de Itaipú, Brazil. The results suggest the use of technique J-48 in order to identify excess body weight in children and adolescents on a grand scale. This technique can be used in clinical and epidemiological contexts with samples of Brazilian youth.

**KEYWORDS :**

Data mining, classification, prediction, decision tree, overweight, obesity.

## I. INTRODUCTION

Data mining is defined as a non-trivial process for the identification of valid, novel, and potentially-useful patterns, and, in the last instance, deducible from the data, it is a stage in the 'Knowledge Discovery in Databases' (KDD) process, whose objective is the automatic interpretation of large data sets (1). It is considered an emergent area within computational intelligence due to its use in the analysis of large databases (2).

In essence, it joins together the advantages of several areas, for example, statistics, artificial intelligence, graphical computation, databases, and massive processing. What is more, it arises as a technology which attempts to help extract useful knowledge of the data related to the data construction; however, this technology still has not been maximally exploited (3). In this sense, there have been several studies which have utilized this technology in order to analyze large volumes of information in diverse areas of human knowledge (4), given that unlike traditional statistics, it takes advantage of the extensive use of 'machine learning' methods in order to be able to handle the volume of data as well as the computational complexity of the KDD problems (5).

Currently, there are several diverse state and/or private organization databases related to physical growth and nutrition. These databases store anthropometric variables such as weight, height, stature, and waist circumference, among others. The most well-known are the Center for Disease Control and Prevention, CDC-2000 (6), CDC-2012 (7) and the World Health Organization (WHO) (8). They are also available online and can be utilized for diverse ends.

From this perspective, the different types of databases, such as those relating to the health sector, contain large volumes of compiled variables, which are increasing significantly every day. In this sense, it is essential to possess an adequate, analytical tool capable of managing and analyzing large quantities of data (9). In fact, the use of data-mining techniques for children and adolescents could play a fundamental role in the classification and prediction of excess body weight and obesity, given that according to Goebel *et al*. (10), data mining can be utilized in diverse foresight, regression, classification, grouping, and association activities respectively.

Consequently, studying excess body weight and obesity in children and adolescents is highly relevant, given that it is considered a global, public health problem and is increasingly affecting many developed and developing countries alike. In fact, it is very well-known that the number of children and adolescents with excess body-weight problems is increasing, hence entailing an increase in the risk of early onset of cardio-metabolic diseases (11). In this context, there are several studies conducted at the international level which document the elevated prevalence of excess body weight in school-age children (12), The results of these studies indicate that this situation can be attributed to the early adoption of modern lifestyles on the part of the children (13).

Therefore, using data-mining techniques, it is possible to uncover unknown patterns while analyzing databases of children and adolescents in which anthropometric variables have been registered, for example, age, sex, skin color, and socioeconomic status, respectively. In this sense, the objective of this study was to compare which of the five popular, data-mining algorithms (Classifier K-NN, Decision Tree, Bayesian Network, Efficient Bayesian Mulitvariate Classifier and Forest by Penalizing Attributes) approximates reality in terms of the classification and prediction of excess body weight in children and adolescents from the region of Lago de Itapú (Paraná, Brazil).

## II. MATERIALS AND METHODS

### 2.1. Database

In this study, an investigation of the descriptive, transversal type was developed. The data utilized come from a database gathered from the area around Lake Itapú (on the border of Brazil and Paraguay) in 2013. Eleven municipalities participated in the study (i.e., Foz do Iguaçu, Santa Terezinha de Itaipu, São Miguel do Iguaçu, Itaipulândia, Missal, Santa Helena, Entre Rios do Oeste, Pato Bragado, Marechal Cândido Rondon, Mercedes y Guaíra). This region belongs to the state of Paraná, Brazil. The database is compiled and maintained by the *Ibero-American Network of Human Biological Development*. The effectuated sample was of the stratified (probabilistic) type, in which 5,962 primary and secondary school students of both sexes (2,938 boys and 3,024 girls) were selected. A total of 34 schools neighboring the region were selected. The age range of the subjects varies between six and seventeen years. The whole data-gathering procedure was the task of ten professionals trained in anthropometric techniques, who carried out the data collection.

The data was collected with the consent of the parents and/or legal guardians of the minors by way of written signatures, so they were duly informed about the study in that they authorized the realization of the anthropometric measurements of the children and adolescents of both sexes. Moreover, the data collection also had the authorization of the ethics-in-research committee of the Faculty of Medicine of *Universidad Estadual de Campina*, Sao Paulo, Brazil.

The anthropometric variables evaluated were weight (kg), height (cm), trunk cephalic height (cm), and tricipital and subscapular skin folds. The standardized protocol of Ross & Marfell-Jones was adopted.

Body-mass index (BMI) was calculated [BMI = weight (kg)/height $(m)^2$], and the cut-off points were utilized in order to classify the subjects into categories of normal, overweight, obese, and excess body weight (overweight+obesity) according to IOTF (14). The percentage of fat was determined utilizing the formula proposed by Boileau, Lohman, and Slaughter (15), and biological maturation was calculated by way of a regression equation specific to each sex, which was proposed by Mirwald *et al*. (16). This technique allows for the determination of biological age in a transversal manner.

Chronological age was likewise determined in a transversal manner, taking into considering the dates of birth of the subjects and those of the evaluation. In order to identify skin color, simple observation was adopted as a technique, where it was registered according to the lightness or darkness of the subjects' skin. Socio-economic level was evaluated via a survey proposed by the *Brazilian Association of Research Businesses* BARB (17). In general, the variables which characterize the studied sample are observed in Table 1.

Starting from the variables evaluated, a descriptive, statistical analysis was conducted (i.e., frequencies, percentages, arithmetic mean, and standard deviation) for both sexes. The differences between the databases (trial run and experiment) were verified by way of a t-test for related samples. The differences between sexes were determined through a t-test for independent samples. In all cases $p<0.001$ was adopted. Normality was verified by the 'goodness of fit test' of Kolmogorov-Smirnov (KS). The whole processing was conducted in SPSS 18.0 (SPSS Inc., IL, USA).

**2.2. Experimental design**

Currently, several types of data-mining tools are available, one of them being Waikato Environment Knowledge Analysis (WEKA). This tool is a software platform which is used for automatic learning and data mining. It is written in Java and contains a collection of visualization tools and algorithms for data analysis and predictive modelling. Working together with a graphic interface, it permits the user to access its functions easily. It supports several standard, data-mining tasks, especially data pre-processing, clustering, classification, regression, visualization, and selection, respectively.

WEKA was utilized for the experiments for three specific reasons: a) it is a user-friendly tool for health-care professionals; b) it is a free-access software; and c) it is fast and efficient (18).

In essence, all the WEKA techniques are grounded in the assumption that the data are available in a flat file or a relation in which each data registry is described by a fixed number of attributes (normally numerical or nominal, although other types are also supported).

For the experiments, the Oracle Technology Network (OTN) methodology was utilized. This graphic interface helps analyze databases whose extraction process can be observed in Figure 1. The architecture implies five steps: data selection, data preparation, data analysis, database results, and implementation.

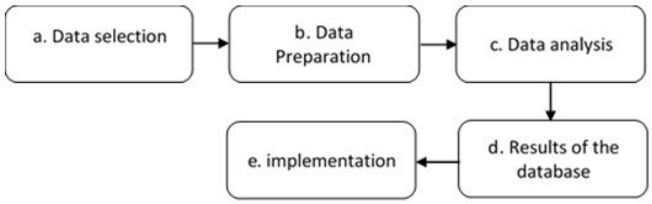


**Fig. 1. Architecture for database analysis regarding the excess body weight of schoolchildren**

**a) Data Selection:** This is the first step where the database is identified (Ibero-American Network of Research in Human Biological Development) and frequent errors and some inconsistencies are also identified.

**b) Data preparation**: This second step is crucial because the data are stored in archives CSV in WEKA. In order to conduct trial runs, 66% was adopted in a random manner. The literal fields (Sex, Color_Skin, Socioeconomic_Liste, classify_BMI_IOTF) were presented in TEXT type, field format, while all the others were of NUMERICAL type (BMI, PGR, Percentage_fat, Mass_fat, Mass_lean), given that they are products of the previous calculations.

The transformation of the fields into their corresponding real type was effectuated in the following way:

- Sex (0- Female, 1-Male).

- Skin color (1-White, 2-Black).

- Socioeconomic level (1-A1, 2-A2, 3-B1, 4-B2, 5-C1, 6-C2, 7-D, 8-E)

The equations employed in the fields generated for body-mass index (BMI), Peak Growth Rate (PGR), Fat Percentage (%fat), Fat Free Mass (FFM) and Fat Mass (mass_fat) are as follows:

$$BMI = weight(kg)/height^2 \text{ ....}\ldots\ldots\ldots\ldots\ldots(1)$$

Boys:

PGR=-9.232+0.0002708(LMI*TCH)-0.001663(Age*LMI)
+0.007216(Age*TCH)+0.02296(Weight/Height)……(2)

Girls :

PGR=-9.37+0.0001882(LMI*TCH)+0.0022(Age*LMI)
+0.005841(Age*TCH)-0.02658(Weight/Height)……(3)

%fat=1.2*BMI+0.23*Age-10.8*Sex-5.4……(4)

mass_fat=(%fat*weight/100)……(5)

mass_lean=weight-mass_fat……(6)

The data construction has allowed for the generation of the data-registry set, which will be utilized in the data-mining tool WEKA. Towards this end, the corresponding archive CSV (comma-separated values) has been generated with the following fields:

| | |
|---|---|
| - Age | entire |
| - Sex | entire (0-female, 1-male) |
| - Skin color | entire (1-white, 2-black) |
| - Body weight | decimal (2,1) |
| - Height | decimal (1,2) |
| - MC (2,2) | decimal (2,2) |
| - TCH | decimal (2,1) |
| - CP | decimal (2,1) |
| - PGR | entire |
| - Fat percentage | decimal (2,1) |
| - Mass_fat | decimal (1,2) |
| - Mass_lean | decimal (2,2) |
| - Socio-economic level | entire |
| - classify_BMI_IOTF | varchar (9) |

**c) Data analysis:** In this stage the appropriate values of the master table are selected. As prediction variables, weight excess (WE) and normality without any weight excess (WWE) were utilized, and a model will be obtained as a result. This was obtained through five types of algorithms: Decision Trees J48, K-NN, Bayesian Networks, Efficient Bayesian Multivariate Classifier, and Forest by Penalizing Attribute.

**Decision trees J48**

This is a version of the tree-decision algorithm C4.5 (19). Decision trees are located within supervised, classification methods, in which there is a dependent variable or class, and the objective of the classifier is to determine the value of excess body weight and normal body weight. The tree is generated according to different measurement variants, such as 'information gain', 'gain ratio' (20) and 'based on distance' (21). The algorithm is designed in such a way that it functions with the 'divide and conquer' strategy, which is an approach for developing a decision tree that is implemented in algorithms such as C4.5 (22). The procedure is described below:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \text{.....}(7)$$

Where pi is the probability that an arbitrary sample belongs to $C_i$.

Entropy calculation:

$$E(A_i) = \sum_{j=1}^{q} \frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \ldots\ldots\ldots\ldots(8)$$

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \ldots\ldots\ldots\ldots\ldots(9)$$

Where $p_{ij} = s_{ij} / |s_j|$, and $|s_j|$ is the number of samples in the subset. Then, the information codification that will be gained in $A$ is:

$$Gain(A_i) = I(s_1, s_2, \ldots, s_m) - E(A_i) \ldots\ldots\ldots\ldots\ldots(10)$$

The attribute $A_i$ with the highest information gain is selected as the node root. The divisions of the node root are formed according to different values of $a_{ij}$, $j$=1,…,$q$. The tree grows in this way until all the samples are of the same class, and then the node becomes a sheet and is labeled with that class.

**Forest by Penalizing Attributes (Forest PA)**

The algorithm Forest PA constructs a set of highly-precise decision trees through the exploitation of the force of all the non-class attributes available in a dataset. At the same time, in order to promote a great diversity, the algorithm imposes sanctions (disadvantageous weights) to the attributes which participated in the last tree in order to generate subsequent trees. What is more, other worries related to weight are taken into consideration so that the trees generated by the algorithm remain individually precise and conserve their great diversity (23).

**k-NN (k nearest neighbors)**

This is a non-parametric, supervised classification method. It is based on trial runs working with examples. It is a 'lazy-learning' type, where the function approximates only locally, and the whole calculation is directed towards classification. In WEKA, this algorithm is referred to as IBk (instance-based learning with fixed neighbors).

The functioning is simple. When a new sample arrives, it verifies the distance found between all the elements and classifies the group of least distance. In order for a data registry 't' to be classified, its closest 'k' neighbors are retrieved, and a neighbor of 't' is formed. A vote between the data registries in the neighborhood is realized. It is usually used in order to decide the classification of 't', with or without weight consideration based on distances. However, in order to apply the algorithm k-NN, an appropriate value for 'k' is needed, and the success of the classification is very much dependent on this value. In this sense, the k-NN method is based on 'k'. There are many ways to choose the value of 'k'. One simple way is to execute the algorithm many times with different 'k' values, and the value of the best execution is chosen (24). Three key elements must be considered:

- A set of labeled elements.
- A distance or metric of similarity in order to calculate the distance between the objects.
- The value of 'k', the number of close neighbors.

**Bayesian networks**

These are non-cyclical graphics, where the nodes represent the variables and the arcs represent the dependencies between the variables. The structure of a Bayesian network supplies information regarding the relationships of conditional dependence and independence existent among the variables. The estimation process of a Bayesian network consists of a structural-learning stage and a parametric-learning stage. The artificial, Bayesian intelligence or BNs is a mathematical frame utilized to deduce probabilistic, cause-and-effect relationships directly starting from the data (25).

**Efficient Bayesian Multivariate Classifier (EBMC)**

This algorithm conducts a voracious search in a subspace of Bayesian networks in order to find what best predicts a target node. The EBMC lessens the difficulty of making strong, conditional-independence suppositions. This is to say that it supposes that the predictors are conditionally independent given the result, while, in reality, they are also often conditionally dependent given the result (26).

**d) Database results**: In this penultimate stage, the desired algorithm is chosen. New patterns from the database are extracted by way of predictions, probabilities, and visualization.

**e) Implementation**: In this last stage, the selected model is applied to new predictions.

## III. RESULTS

Table 1 shows the variables that characterize the studied sample. The boys present greater height, lower percentage of body fat and fat mass, and greater fat-free mass compared to the girls. The prevalence of an excess body weight in both sexes is similar (boys 20.66% and girls 19.68%).

**Table 1. Characteristics of the studied sample**

| Attributes (codification) | Boys | | Girls | | Both Sexes | |
|---|---|---|---|---|---|---|
| | *X* | *SD* | *X* | *SD* | *X* | *SD* |
| N | 2938 | | 3024 | | 5962 | |
| C.A. (years) | 11,30 | 3,23 | 11,60 | 3,20 | 22,90 | 6,43 |
| ***Anthropometry*** | | | | | | |
| Weight (kg) | 45,40 | 17,80 | 43,90 | 14,70 | 89,30 | 32,5 |
| Height (cm) | 1,51 | 0,19 | 1,49 | 0,15 | 3,00 | 0,34 |
| T.C.H. (cm) | 77,94 | 9,35 | 78,26 | 8,13 | 156,20 | 17,48 |
| ***Body Composition*** | | | | | | |
| Fat Percentage (%G) | 15,85 | 7,47 | 22,60 | 6,81 | 38,45 | 14,28 |
| Fat Mass (kg) | 7,70 | 5,96 | 10,58 | 5,95 | 18,28 | 11,91 |
| Fat-Free Mass (Kg) | 37,69 | 13,84 | 33,38 | 9,65 | 71,07 | 23,49 |
| ***Skin Color*** | ***fi*** | ***%*** | ***fi*** | ***%*** | ***fi*** | ***%*** |
| 1. White | 2628 | 89.40 | 2717 | 89.80 | 5345 | 89,65 |
| 2. Black | 310 | 10.60 | 307 | 10.20 | 617 | 10,35 |
| Total | 2938 | 100.00 | 3024 | 100.00 | 5962 | 100,00 |
| ***Socio-economic Level*** | | | | | | |
| 1. A1 | 5 | 0.17 | 1 | 0.03 | 6 | 0,10 |
| 2. A2 | 117 | 3.99 | 80 | 2.65 | 197 | 3,30 |
| 3. B1 | 399 | 13.58 | 380 | 12.57 | 779 | 13,07 |
| 4. B2 | 1037 | 35.30 | 1024 | 33.86 | 2061 | 34,57 |
| 5. C1 | 895 | 30.46 | 976 | 32.28 | 1871 | 31,38 |
| 6. C2 | 407 | 13.85 | 481 | 15.91 | 888 | 14,89 |
| 7. D | 75 | 2.55 | 81 | 2.67 | 156 | 2,62 |
| 8. E | 3 | 0.10 | 1 | 0.03 | 4 | 0,07 |
| Total | 2938 | 100.00 | 3024 | 100.00 | 5962 | 100,00 |
| ***BMI Classification (kg/m²)[a]*** | | | | | | |
| Normal | 2331 | 79.34 | 2429 | 80.32 | 4760 | 79,83 |
| Overweight | 607 | 20.66 | 595 | 19.68 | 1202 | 20,16 |
| Total | 2938 | 100.00 | 3024 | 100.00 | 5962 | 100,00 |

Legend: X = Average, SD = Standar Desviation, N = Quantity, C.A. = chronological age, TCH = trunk-cephalic height, BMI = Body Mass Index, a = Determined by IOTF (criterion).

The instances classified, beginning from the data-confusion matrix, with a 66% percentage split of test options and the evaluation time required in each technique, are observed in Table 2. The results indicate that the J-48 decision-tree algorithm is that which most approximates the utilized criterion (IOFT) with both sexes (all). The percentage of instances classified as normal and excess body weight (EW) reaches 97.6% in both sexes. In relation to the time utilized in order to construct the model, the BayesNet technique proved to be the fastest in relation to all the other techniques, although the other techniques were also very quick, above all in terms of constructing the model for each sex.

**Table 2. Classification of excess body weight through the data-confusion matrix and evaluation time in order to construct the model**

| Techniques | Boys | | | Girls | | | All | | |
|---|---|---|---|---|---|---|---|---|---|
| | fi | % | Time | fi | % | Time | fi | % | Time |
| ***J48*** | | | | | | | | | |
| Instances classified as normal | 778 | 77.9 | 0 | 817 | 79.5 | 0 | 1596 | 78.7 | 0.07 |
| Instances classified as EW | 198 | 19.8 | | 187 | 18.2 | | 383 | 18.9 | |
| Instances classified as incorrect | 23 | 2.3 | | 24 | 2.3 | | 48 | 2.4 | |
| Total | 999 | 100.0 | | 1028 | 100.0 | | 2027 | 100.0 | |
| ***IBK*** | | | | | | | | | |
| Instances classified as normal | 768 | 76.9 | 0 | 800 | 77.8 | | 1561 | 77.0 | 0.5 |
| Instances classified as EW | 166 | 16.6 | | 164 | 16.0 | | 326 | 16.0 | |
| Instances classified as incorrect | 65 | 6.5 | | 64 | 6.2 | | 140 | 7.0 | |
| Total | 999 | 100.0 | | 1028 | 100.0 | 0.12 | 2027 | 100.0 | |
| ***Bayes Net*** | | | | | | | | | |
| Instances classified as normal | 733 | 73.4 | 0 | 744 | 72.4 | | 1471 | 72.6 | 0.02 |
| Instances classified as EW | 169 | 16.9 | | 143 | 13.9 | | 294 | 14.5 | |
| Instances classified as incorrect | 97 | 9.7 | | 141 | 13.7 | | 262 | 12.9 | |
| Total | 999 | 100.0 | | 1028 | 100.0 | 0 | 2027 | 100.0 | |
| ***EBMC (Efficient Bayesian Multivariate Classification)*** | | | | | | | | | |
| Instances classified as normal | 757 | 75.8 | 0.24 | 807 | 78.5 | | 1571 | 77.5 | 0.34 |
| Instances classified as EW | 65 | 6.5 | | 144 | 14.0 | | 315 | 15.5 | |
| Instances classified as incorrect | 177 | 17.7 | | 78 | 7.5 | | 141 | 7.0 | |
| Total | 999 | 100.0 | | 1028 | 100.0 | 0.21 | 2027 | 100.0 | |
| ***Forest PA*** | | | | | | | | | |
| Instances classified as normal | 785 | 78.6 | 0.63 | 816 | 79.4 | | 1600 | 78.9 | 0.36 |
| Instances classified as EW | 188 | 18.8 | | 183 | 17.8 | | 375 | 18.5 | |
| Instances classified as incorrect | 26 | 2.6 | | 29 | 2.8 | | 52 | 2.6 | |
| **Total** | 999 | 100.00 | | 1028 | 100.0 | 0.64 | 2027 | 100.0 | |

Legend: fi = frecuency, % = Percentage.

The cross-validation values are observed in Tables 3 and 4. The J-48 decision-tree technique showed superior, concordance values (Kappa) in relation to the other techniques (0.9262-0.9305). The absolute-error percentages are likewise inferior (8.29%-8.55%) for both sexes in relation to the other techniques. The coverage values are found to be approaching 100%. All the values observed in the table are derived from the evaluations and are indicated in the percentage-split parameter, which reaches 66%. The instances are divided, as this parameter indicates, and in each evaluation the instances are taken as test data and the rest as trial-run data in order to construct the model. The calculated errors are the average of all the executions.

**Table 3. Detail of the cross-validation values for the data-mining techniques broken down for boys and girls**

| Detail of results | Boys | | | | | Girls | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | J48 | IBK | Bayes Net | EBMC | Forest PA | J48 | IBK | Bayes Net | EBMC | Forest PA |
| ***Cross validation*** | | | | | | | | | | |
| Instances classified correctly | 976 | 934 | 902 | 924 | 973 | 1004 | 964 | 887 | 951 | 999 |
| Instances classified incorrectly | 23 | 65 | 97 | 72 | 26 | 24 | 64 | 141 | 78 | 30 |
| Kappa statistic | **0.9305** | 0.7959 | 0.715 | 0.7883 | 0.9190 | **0.9252** | 0.7984 | 0.7804 | 0.7417 | 0.9062 |
| Absolute mean error | 0.0274 | 0.0655 | 0.109 | 0.098 | 0.0550 | 0.0265 | 0.0627 | 0.0693 | 0.1160 | 0.0597 |
| Square root mean error | 0.1429 | 0.2549 | 0.287 | 0.276 | 0.1474 | 0.1505 | 0.2494 | 0.2627 | 0.2445 | 0.1509 |
| Relative absolute error | **8.29%** | 19.87% | 33.07% | 30.53% | 16.78% | **8.34%** | 19.77% | 21.47% | 36.56% | 18.81% |
| Relative square root error | 34.94% | 62.3% | 70.17% | 71.43% | 36.49% | 37.65% | 62.41% | 65.22% | 61.09% | 37.70% |
| Coverage of cases (0.95 level) | 98.59% | 93.49% | 94.9% | 98.19% | 99.70% | 97.95% | 92.77% | 93.09% | 98.93% | 99.90% |
| Mean rel. region size (0.95 l.) | 51.05% | 50% | 57.80% | 57.15% | 58.36% | 51.02% | 50.00% | 50.00% | 65.40% | 60.01% |
| Total number of instances | 999 | 99 | 999 | 999 | 999 | 1028 | 1028 | 1028 | 1028 | 1028 |

**Table 4. Detail of the cross validation values for the data-mining techniques broken down for both sexes.**

| Detail of results | All | | | | |
|---|---|---|---|---|---|
| | J48 | IBK | BayesNet | EBMC | Forest PA |
| ***Cross validation*** | | | | | |
| Instances classified correctly | 1979 | 1887 | 1765 | 1877 | 1975 |
| Instances classified incorrectly | 48 | 140 | 262 | 152 | 52 |
| Kappa statistic | **0.9262** | 0.7804 | 0.6101 | 0.7445 | 0.9192 |
| Absolute mean error | 0.0276 | 0.0693 | 0.1395 | 0.1188 | 0.0478 |
| Square root mean error | 0.1491 | 0.2627 | 0.3285 | 0.2421 | 0.1345 |
| Relative absolute error | **8.55%** | 21.47% | 43.22% | 36.51% | 14.80 % |
| Relative square root error | 37.00% | 65.22% | 81.56% | 59.08% | 33.39 % |
| Coverage of cases (0.95 level) | 98.02% | 93.09% | 94.27% | 98.57% | 99.90 % |
| Mean rel. region size (0.95 level) | 50.93% | 50.00% | 59.96% | 58.86% | 58.07 % |
| Total number of cases | 2027 | 2027 | 2027 | 2027 | 2027 |

In Tables 5 and 6, the values in relation to precision capacity and prediction regarding the techniques utilized in the study are observed. With boys, girls, and both sexes, the J-48 decision-tree technique evidenced values of superior precision, for example, Measurement-F and the area below the curve (ROC), in relation to the other techniques, as well as for classified metrics such as normal body weight and excess body weight.

**Table 5. Precision capacity by class of the data-mining techniques J48, IBK, and BayesNet**

| Precision by class | J48 | | | IBK | | | BayesNet | | |
|---|---|---|---|---|---|---|---|---|---|
| | Normal | EP | PM | Normal | EP | PM | Normal | EP | PM |
| ***Boys*** | | | | | | | | | |
| TP Rate (True Positive) | 0.989 | 0.934 | 0.977 | 0.976 | 0.783 | 0.935 | 0.931 | 0.797 | 0.903 |
| FP Rate (False Positive) | 0.066 | 0.011 | 0.054 | 0.217 | 0.024 | 0.176 | 0.203 | 0.069 | 0.174 |
| Precision | **0.982** | 0.957 | 0.977 | 0.943 | 0.897 | 0.934 | 0.945 | 0.758 | 0.905 |
| Recall | 0.989 | 0.934 | 0.977 | 0.976 | 0.783 | 0.935 | 0.931 | 0.797 | 0.903 |
| F-Measure | **0.985** | 0.945 | 0.977 | 0.959 | 0.836 | 0.933 | 0.938 | 0.777 | 0.904 |
| MCC | 0.931 | 0.931 | 0.931 | 0.799 | 0.799 | 0.799 | 0.715 | 0.715 | 0.715 |
| ROC Area | **0.979** | 0.979 | 0.979 | 0.879 | 0.879 | 0.879 | 0.944 | 0.944 | 0.944 |
| PRC Area | 0.990 | 0.946 | 0.980 | 0.940 | 0.749 | 0.899 | 0.982 | 0.873 | 0.959 |
| ***Girls*** | | | | | | | | | |
| TP Rate (True Positive) | 0.993 | 0.912 | 0.977 | 0.972 | 0.800 | 0.938 | 0.904 | 0.698 | 0.863 |
| FP Rate (False Positive) | 0.088 | 0.007 | 0.072 | 0.200 | 0.028 | 0.166 | 0.302 | 0.096 | 0.261 |
| Precision | **0.978** | 0.969 | 0.977 | 0.951 | 0.877 | 0.936 | 0.923 | 0.644 | 0.867 |
| Recall | 0.993 | 0.912 | 0.977 | 0.972 | 0.800 | 0.938 | 0.904 | 0.698 | 0.863 |
| F- Measure | **0.986** | 0.940 | 0.976 | 0.962 | 0.837 | 0.937 | 0.913 | 0.670 | 0.865 |
| MCC | 0.926 | 0.926 | 0.926 | 0.800 | 0.800 | 0.800 | 0.584 | 0.584 | 0.584 |
| ROC Area | **0.965** | 0.965 | 0.965 | 0.886 | 0.886 | 0.886 | 0.921 | 0.921 | 0.921 |
| PRC Area | 0.983 | 0.934 | 0.973 | 0.947 | 0.741 | 0.906 | 0.977 | 0.793 | 0.940 |
| ***All*** | | | | | | | | | |
| TP Rate (True Positive) | 0.989 | 0.927 | 0.976 | 0.967 | 0.789 | 0.931 | 0.911 | 0.712 | 0.871 |
| FP Rate (False Positive) | 0.073 | 0.011 | 0.060 | 0.211 | 0.033 | 0.174 | 0.288 | 0.089 | 0.247 |
| Precision | **0.982** | 0.955 | 0.976 | 0.947 | 0.860 | 0.929 | 0.925 | 0.673 | 0.874 |
| Recall | 0.989 | 0.927 | 0.976 | 0.967 | 0.789 | 0.931 | 0.911 | 0.712 | 0.871 |
| F- Measure | **0.985** | 0.941 | 0.976 | 0.957 | 0.823 | 0.930 | 0.918 | 0.692 | 0.872 |
| MCC | 0.926 | 0.926 | 0.926 | 0.782 | 0.782 | 0.782 | 0.610 | 0.610 | 0.610 |
| ROC Area | **0.980** | 0.980 | 0.980 | 0.878 | 0.878 | 0.878 | 0.914 | 0.914 | 0.914 |
| PRC Area | 0.991 | 0.939 | 0.981 | 0.942 | 0.722 | 0.897 | 0.975 | 0.800 | 0.939 |

Legend**:** EW = Excess body weight, WA = Weighed average

**Table 6. Precision capacity by class of the data-mining techniques EBMC and Forest PA**

| Precision by class | EBMC | | | Forest PA | | |
|---|---|---|---|---|---|---|
| | Normal | EW | WA | Normal | EW | WA |
| ***Boys*** | | | | | | |
| TP Rate (True Positive) | 0.953 | 0.863 | 0.935 | 0.989 | 0.917 | 0.974 |
| FP Rate (False Positive) | 0.137 | 0.047 | 0.118 | 0.083 | 0.011 | 0.068 |
| Precision | 0.964 | 0.827 | 0.936 | 0.979 | 0.954 | 0.974 |
| Recall | 0.953 | 0.863 | 0.935 | 0.989 | 0.917 | 0.974 |
| F-Measure | 0.959 | 0.845 | 0.935 | 0.984 | 0.935 | 0.974 |
| MCC | 0.804 | 0.804 | 0.804 | 0.919 | 0.919 | 0.919 |
| ROC Area | 0.970 | 0.973 | 0.971 | 0.987 | 0.987 | 0.987 |
| PRC Area | 0.990 | 0.920 | 0.976 | 0.992 | 0.979 | 0.990 |
| ***Girls*** | | | | | | |
| TP Rate (True Positive) | 0.981 | 0.699 | 0.924 | 0.991 | 0.888 | 0.971 |
| FP Rate (False Positive) | 0.301 | 0.019 | 0.245 | 0.112 | 0.009 | 0.091 |
| Precision | 0.929 | 0.900 | 0.923 | 0.973 | 0.963 | 0.971 |
| Recall | 0.981 | 0.699 | 0.924 | 0.991 | 0.888 | 0.971 |
| F- Measure | 0.954 | 0.787 | 0.920 | 0.982 | 0.924 | 0.970 |
| MCC | 0.750 | 0.750 | 0.750 | 0.907 | 0.907 | 0.907 |
| ROC Area | 0.956 | 0.958 | 0.956 | 0.995 | 0.995 | 0.995 |
| PRC Area | 0.986 | 0.879 | 0.965 | 0.999 | 0.984 | 0.996 |
| ***All*** | | | | | | |
| TP Rate (True Positive) | 0.973 | 0.763 | 0.930 | 0.991 | 0.908 | 0.974 |
| FP Rate (False Positive) | 0.237 | 0.027 | 0.194 | 0.092 | 0.009 | 0.174 |
| Precision | 0.941 | 0.880 | 0.929 | 0.977 | 0.955 | 0.966 |
| Recall | 0.973 | 0.763 | 0.930 | 0.991 | 0.964 | 0.974 |
| F- Measure | 0.957 | 0.817 | 0.929 | 0.982 | 0.908 | 0.974 |
| MCC | 0.777 | 0.777 | 0.777 | 0.920 | 0.935 | 0.920 |
| ROC Area | 0.970 | 0.970 | 0.970 | 0.976 | 0.976 | 0.976 |
| PRC Area | 0.990 | 0.908 | 0.973 | 0.999 | 0.989 | 0.997 |

**Legend:** EW = Excess body weight, WA = Weighed average

Figure 2 show the J-48 decision tree of both sexes, it can be appreciated that the percentage of body fat, age, sex, peak growth rate, and socioeconomic level are predictors of BMI and excess body weight. When analyzed separately, excess body weight in boys is present in subjects with both white and black skin and from middle, socioeconomic levels (level 3). On the contrary, in girls there is more of a risk for excess body weight in girls with white skin, and it is present in lower socioeconomic conditions.

```
J48 pruned tree
------------------
BMI <= 22.33
|  Fat_percentage <= 20.5
|  |  BMI <= 17.61: normal (2158.0)
|  |  BMI > 17.61
|  |  |  PGR classification <= -4
|  |  |  |  Age <= 6
|  |  |  |  |  BMI <= 17.9
|  |  |  |  |  |  sex <= 0: excess (2.0)
|  |  |  |  |  |  sex > 0
|  |  |  |  |  |  |  Socioeconomic_level <= 4
|  |  |  |  |  |  |  |  Height_m <= 1.22: excess (2.0)
|  |  |  |  |  |  |  |  Height_m > 1.22: normal (3.0/1.0)
|  |  |  |  |  |  |  Socioeconomic_level > 4: normal (4.0)
|  |  |  |  |  BMI > 17.9: excess (20.0)
|  |  |  |  Age > 6
|  |  |  |  |  BMI <= 19.46
|  |  |  |  |  |  Age <= 7
|  |  |  |  |  |  |  BMI <= 18.16: normal (11.0)
|  |  |  |  |  |  |  BMI > 18.16: excess (10.0/2.0)
|  |  |  |  |  |  Age > 7: normal (61.0/2.0)
|  |  |  |  |  BMI > 19.46: excess (11.0)
|  |  |  PGR classification > -4
|  |  |  |  BMI <= 20.84: normal (951.0)
|  |  |  |  BMI > 20.84
|  |  |  |  |  Age <= 12
|  |  |  |  |  |  Age <= 10: excess (4.0)
|  |  |  |  |  |  Age > 10
|  |  |  |  |  |  |  BMI <= 21.66: normal (6.0)
|  |  |  |  |  |  |  BMI > 21.66: excess (6.0/1.0)
|  |  |  |  |  Age > 12: normal (170.0)
|  Fat_percentage > 20.5
|  |  Age <= 11
|  |  |  BMI <= 18.84
|  |  |  |  Age <= 7
|  |  |  |  |  BMI <= 17.42: normal (56.0)
|  |  |  |  |  BMI > 17.42
|  |  |  |  |  |  BMI <= 18.1
|  |  |  |  |  |  |  Age <= 6: excess (9.0/1.0)
|  |  |  |  |  |  |  Age > 6: normal (14.0/1.0)
|  |  |  |  |  |  BMI > 18.1: excess (19.0)
|  |  |  |  Age > 7: normal (259.0/1.0)
|  |  |  BMI > 18.84
|  |  |  |  Age <= 8: excess (136.0/3.0)
|  |  |  |  Age > 8
|  |  |  |  |  BMI <= 20.28
|  |  |  |  |  |  Age <= 9
|  |  |  |  |  |  |  BMI <= 19.79
|  |  |  |  |  |  |  |  Fat_percentage <= 32: normal (17.0/1.0)
|  |  |  |  |  |  |  |  Fat_percentage > 32: excess (2.0)
|  |  |  |  |  |  |  BMI > 19.79: excess (9.0/1.0)
|  |  |  |  |  |  Age > 9: normal (77.0)
|  |  |  |  |  BMI > 20.28
|  |  |  |  |  |  Age <= 10: excess (82.0/5.0)
|  |  |  |  |  |  Age > 10
|  |  |  |  |  |  |  BMI <= 21.18: normal (24.0/1.0)
|  |  |  |  |  |  |  BMI > 21.18
|  |  |  |  |  |  |  |  sex <= 0
|  |  |  |  |  |  |  |  |  BMI <= 21.65
|  |  |  |  |  |  |  |  |  |  PGR classification <= 0
|  |  |  |  |  |  |  |  |  |  |  Height_m <= 1.48: normal (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  Height_m > 1.48: excess (2.0)
|  |  |  |  |  |  |  |  |  |  PGR classification > 0: normal (4.0)
|  |  |  |  |  |  |  |  |  BMI > 21.65: excess (10.0)
|  |  |  |  |  |  |  |  sex > 0: excess (7.0)
|  |  |  Age > 11
|  |  |  |  sex <= 0: normal (647.0/1.0)
|  |  |  |  sex > 0
|  |  |  |  |  BMI <= 21.53: normal (62.0)
|  |  |  |  |  BMI > 21.53
|  |  |  |  |  |  Age <= 12
|  |  |  |  |  |  |  IMC <= 21.88
|  |  |  |  |  |  |  |  Fat_percentage <= 24.8: normal (2.0)
|  |  |  |  |  |  |  |  Fat_percentage > 24.8: excess (4.0/1.0)
|  |  |  |  |  |  BMI > 21.88: excess (5.0)
|  |  |  |  |  Age > 12: normal (15.0/1.0)
BMI > 22.33
|  BMI <= 24.23
|  |  Age <= 13
|  |  |  PGR classification <= 1
|  |  |  |  Age <= 12: excess (158.0)
|  |  |  |  Age > 12
|  |  |  |  |  BMI <= 22.55
|  |  |  |  |  |  PGR classification <= -1: excess (4.0/1.0)
|  |  |  |  |  |  PGR classification > -1: normal (3.0)
|  |  |  |  |  BMI > 22.55: excess (30.0)
|  |  |  PGR classification > 1
|  |  |  |  BMI <= 23.31
|  |  |  |  |  Age <= 12
|  |  |  |  |  |  BMI <= 22.63: normal (3.0/1.0)
|  |  |  |  |  |  BMI > 22.63: excess (7.0)
|  |  |  |  |  Age > 12
|  |  |  |  |  |  Fat_percentage <= 33.2: normal (13.0)
|  |  |  |  |  |  Fat_percentage > 33.2: excess (3.0/1.0)
|  |  |  |  BMI > 23.31: excess (23.0)
|  |  Age > 13
|  |  |  Age <= 14
|  |  |  |  BMI <= 23.5
|  |  |  |  |  sex <= 0: normal (23.0)
|  |  |  |  |  sex > 0
|  |  |  |  |  |  BMI <= 22.91: normal (10.0)
|  |  |  |  |  |  BMI > 22.91: excess (6.0)
|  |  |  |  BMI > 23.5: excess (23.0/4.0)
|  |  |  Age > 14: normal (151.0/6.0)
|  BMI > 24.23
|  |  Age <= 15: excess (496.0)
|  |  Age > 15
|  |  |  BMI <= 24.8
|  |  |  |  Age <= 16
|  |  |  |  |  Fat_mass <= 19.2: excess (7.0)
|  |  |  |  |  Fat_mass > 19.2: normal (3.0/1.0)
|  |  |  |  Age > 16: normal (8.0)
|  |  |  BMI > 24.8: excess (107.0)

Number of sheets:  57
Size of the tree:  113
```



**Fig. 2. J-48 decision tree for children and adolescents of both sexes**

## IV. DISCUSSION

The results of the experiments conducted show that the J-48 algorithm is the technique which most approximates the real criterion with respect to the classification of schoolchildren with excess body weight and those with normal body weight. What is more, in the cross validation the concordance values, as much for boys, girls, and both sexes, are found to be close to 1 (0.9252-0.9305). The coverage values also evidence percentages very close to 100% (98.59% to 98.02%), making this algorithm really stand out in terms of its high, classification capacity, specifically with children and adolescents from the region of Lago de Itaipú, Brazil.

The criterion utilized (real class) in order to classify schoolchildren with normal and excess body weights was from the *International Group of Work about Obesity* (19). In fact, in the literature there are other international criteria, such as that of the CDC-2000, WHO (27), which allows for the classification of children and adolescents according to several health-and-nutrition-related categories (e.g., underweight, normal, overweight, and obese) and which are currently, widely utilized in order to diagnose health levels in growth ages in different parts of the world, but above all in those countries where reference criteria do not exist for the valuation of health levels and physical growth.

In fact, the J-48 technique proved to be a valuable tool for the classification of excess body weight, at least for Brazilian schoolchildren under study. This demonstrates that data mining provides valuable tools for the selection of adequate models, but above all when it is a question of analyzing clinical repositories which contain great quantities of biological, clinical, and administrative data (28). Lee *et al.* (29) even warn that the Bayesian classifier model presents less bias in relation to the traditional statistics of logistical regression, for which it is necessary that medical professionals and associations take advantage of the novel and powerful techniques which data mining offers (30).

With respect to precision and the prediction of excess body weight, the results show that the J-48 algorithm demonstrated a high capacity with respect to the three analyzed data sets (i.e., boys, girls, and both sexes). The values of Measurement-F, precision by class, sensitivity and specificity, and the area below the curve yielded values close to 1. However, the BayesNet and IBK techniques presented relatively inferior values with the three groups of analyzed data.

Starting from these discoveries, the J-48 decision tree was adopted as a predictive model, whose principal objective is inductive learning effectuated from observations and logical constructions. This algorithm creates standards for the prediction of the target variable (31). Basically, according to Han, Kamber (32) the trees, unlike other techniques, facilitate the interpretation of other data, provide a high grade of understanding of the knowledge utilized in decision making, reduce the number of independent variables, and allow for the establishment of the data-mining algorithm selection.

In this sense, the decision trees obtained (boys, girls, and both sexes) show an association between BMI and the percentage of body fat. Furthermore, attributes such as peak growth rate (PGR), age, and socioeconomic status are clear predictors of an excess in body weight. When the data were analyzed by sex, attributes were even evidenced with the girls which were not observed in the boys, for example, skin color and socioeconomic level (category five). What is being implied here is that with white girls, there is a greater risk of excess body weight, as there also is with girls from lower, socioeconomic levels. On the contrary, with the boy's excess body weight is present in subjects with white and black skin, and especially those from middle, socioeconomic levels (category 3).

In essence, these discoveries must be analyzed by health-science professionals, familiar with health and nutritional studies, who are trained to be able to interpret the results properly. Nevertheless, it is necessary to highlight that data mining basically depends on the quality of the collected data (33), given that this allows for the reduction of bias during the data collection process, and consequently, the classifications and/or predictions obtained by way of data mining can be validated. In fact, the database utilized in this study was collected by a group of well-trained professionals utilizing anthropometric assessment techniques, which guarantees the quality of the data.

Consequently, health-science professionals may adopt techniques which offer data mining once they are convinced of the utility (34) and thus begin the process of familiarization which must happen in order to be able to master the rigorous, technological task which is needed. Therefore, the databases stored digitally in state and private organizations must be taken advantage of in order to produce useful knowledge and, in this manner, help make decisions in organizations through the determination of patterns and models.

For future studies, it is suggested to utilize other real classifiers of excess body weight (national and international criteria) and compare them with data-mining techniques. What is more, the uses of traditional, statistical methods must be expanded and developed in order to investigate their predictive capacity versus that of data mining.

## V. CONCLUSION

Through the results obtained, it is concluded that data mining, the J-48 technique, proved to be the most realistic classifier and can contribute to the prediction of excess body-weight risk in children and adolescents from the region of Lago de Itaipú, Brazil. The results suggest the use of the J-48 technique in order to identify excess body weight in children and adolescents on a grand scale; however, more studies in other contextual realities are necessary in order to confirm these discoveries. This technique can be utilized in clinical and epidemiological contexts with samples of Brazilian youth.

## REFERENCES


**[1].** **Kriegel H-P, Borgwardt KM, Kröger P, Pryakhin A, Schubert M, Zimek A.** Future trends in data mining. Data Min Knowl Discov. 2007;

**[2].** **Kusiak A, Kern JA, Kernstine KH, Tseng BTL.** Autonomous decision-making: a data mining approach. IEEE Trans Inf Technol Biomed. 2000;4(4):274–84.

**[3].** **Yu Z (Jerry), Haghighat F, Fung BCM.** Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustain Cities Soc [Internet]. 2016;25:33–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2210670715001067

**[4].** **Vianna RCXF, de Barra Moro CMC, Moysés SJ, Carvalho D, Nievola JC.** Mineração de dados e características da mortalidade infantil Data mining and characteristics of infant mortality. Cad saúde pública. 2010;26(3):535–42.

**[5].** **Mannila H.** Data mining: machine learning, statistics, and databases. In: Proceedings of 8th International Conference on Scientific and Statistical Data Base Management. 1996.

**[7].** **Fryar CD, Gu Q, Ogden CL.** Anthropometric reference data for children and adults: United States, 2007-2010. Vital Health Stat 11. 2012;(252):1–48.

**[8].** **Mahbubani K.** The World Health Organization (WHO). Glob Public Health [Internet]. 2012;7(3):312–4. Available from: https://doi.org/10.1080/17441692.2011.652972

**[9].** **Tekieh MH, Raahemi B.** Importance of Data Mining in Healthcare. Proc 2015 IEEE/ACM Int Conf Adv Soc Networks Anal Min 2015 - ASONAM '15. 2015;

**[10].** **Goebel M, Gruenwald L.** A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explor Newsl. 1999;1(1):20–33.

**[11].** **Schmidt MD, Dwyer T, Magnussen CG, Venn AJ.** Predictive associations between alternative measures of childhood adiposity and adult cardio-metabolic health. Int J Obes. 2011;35(1):38–45.

**[12].** **Sun SS, Deng X, Sabo R, Carrico R, Schubert CM, Wan W, et al.** Secular trends in body composition for children and young adults: The fels longitudinal study. Am J Hum Biol. 2012;24(4):506–14.

**[13].** **Wardle J, Boniface D.** Changes in the distributions of body mass index and waist circumference in English adults, 1993/1994 to 2002/2003. Int J Obes. 2008;32(3):527–32.

**[14].** **Cole TJ, Bellizzi MC, Flegal KM, Dietz WH.** Establishing a standard definition for child overweight and obesity worldwide: international survey. Bmj. 2000;320(7244):1240.

**[15].** **Boileau RA, Lohman TG, Slaughter MH.** Exercise and body composition of children and youth. Scand J Sport Sci [Internet]. 1985;7(1):17–27. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0022263239&partnerID=40&md5=81a6266e19d5529bb2af4fffa0e16483

**[16].** **Mirwald RL, G. Baxter-Jones AD, Bailey DA, Beunen GP.** An assessment of maturity from anthropometric measurements. Med Sci Sport Exerc [Internet]. 2002;34(4):689–94. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005768-200204000-00020

**[17].** Associação Brasileira de Empresas de Pesquisa. Critério Brasil - ABEP [Internet]. CRITÉRIO BRASIL 201 E ATUALIZAÇÃO DA DISTRIBUIÇÃO DE CLASSES PARA 2016. 2016. Available from: http://www.abep.org/criterio-brasil

**[18].** **Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH.** The WEKA data mining software: an update. SIGKDD Explor [Internet]. 2009;11(1):10–8. Available from: http://portal.acm.org/citation.cfm?id=1656274.1656278

**[19].** **Quinlan J. C4.5:** programs for machine learning. Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1993. 235–240 p.

**[20].** **Wang Y, Makedon FS, Ford JC, Pearlman J. HykGene:** A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics. 2005;21(8):1530–7.

**[21].** **De Mántaras RL.** A Distance-Based Attribute Selection Measure for Decision Tree Induction. Mach Learn. 1991;6(1):81–92.

**[22].** **Hidayah I, Erna P. A, Kristy MA.** Application of J48 and bagging for classification of vertebral column pathologies. In: Conference Proceedings - 6th International Conference on Information Technology and Multimedia at UNITEN: Cultivating Creativity and Enabling Technology Through the Internet of Things, ICIMU 2014. 2015.

**[23].** **Adnan MN, Islam MZ. Forest PA:** Constructing a decision forest by penalizing attributes used in previous trees. Expert Syst Appl. 2017;

**[24].** **Guo G, Wang H, Bell D, Bi Y, Greer K.** On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE [Internet]. Meersman R, Tari Z, Schmidt DC, editors. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. 986-996 p. (Lecture Notes in Computer Science; vol. 2888). Available from: http://link.springer.com/10.1007/b94348

**[25].** **Vemulapalli V, Qu J, Garren JM, Rodrigues LO, Kiebish MA, Sarangarajan R, et al.** Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data. Artif Intell Med. 2016;

**[26].** **Jiang X, Cai B, Xue D, Lu X, Cooper GF, Neapolitan RE.** A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. J Am Med Inform Assoc. 2014;

**[27].** **Varghese C, Oyere O, Cowan M, Davis S, Norrving B.** World Health Organization. Stroke [Internet]. 2016 Aug [cited 2018 Jul 12];47(8):e210–e210. Available from: http://stroke.ahajournals.org/lookup/doi/10.1161/STROKEAHA.116.014233

**[28].** **Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al.** Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med. 2006;36(12):1351–77.

**[29].** **Lee BJ, Ku B, Nam J, Pham DD, Kim JY.** Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE J Biomed Heal informatics. 2014;18(2):555–61.

**[30].** **Holmes JH, Durbin DR, Winston FK.** Discovery of predictive models in an injury surveillance database: an application of data mining in clinical research. In: Proceedings of the AMIA Symposium. 2000. p. 359.

**[31].** **Saravanan N, Gayathri V.** Classification of dengue dataset using J48 algorithm and ant colony based AJ48 algorithm. In: 2017 International Conference on Inventive Computing and Informatics (ICICI) [Internet]. IEEE; 2017 [cited 2018 Jul 12]. p. 1062–7. Available from: https://ieeexplore.ieee.org/document/8365302/

**[32].** **Jiawei H, Kamber M, Han J, Kamber M, Pei J.** Data Mining: Concepts and Techniques [Internet]. San Francisco, CA, itd: Morgan Kaufmann. 2006. 745 p. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Data+Mining+Concepts+and+Techniques#1%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Data+mining+concepts+and+techniques%231%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Se

**[33].** **Gal A.** Challenge Paper: Data Quality Issues in Queue Mining 1 Queueing Networks and Queue Mining. J Data Inf Qual Artic ACM J Data Inf Qual. 2018;

**[34].** **Koh HC, Tan G,** others. Data mining applications in healthcare. J Healthc Inf Manag. 2011;19(2):65.


*****