Machine learning techniques to predict overweight or obesity

Elias Rodríguez[a], Elen Rodríguez[a], Luiz Nascimento[a,b], Aneirson da Silva[a] and Fernando Marins[a]

[a] *São Paulo State University (UNESP), Av. Dr. Ariberto Pereira da Cunha, 333, Guaratinguetá/SP, 12516-410, Brazil*
[b] *University of Taubaté (UNITAU), Av. Professor Walter Taumaturgo, 739, Taubaté/SP, 12030-040, Brazil.*

Abstract
Overweight and obesity are considered a public health problem, as they are related to the risk of various diseases, and also to the risk of increased morbidity and mortality. The main objective of this work was to apply machine learning techniques for the development of a predictive model for the identification of people with obesity or overweight. The model developed was based on data related to the physical condition and eating habits. Furthermore, the machine learning classification algorithms that were tested were: decision tree,support vector machines, k-nearest neighbors, gaussian naive bayes, multilayer perceptron, random forest, gradient boosting, and extreme gradient boosting. Model hyperparameters were tuned to improve accuracy, resulting in that the model with the best performance was a random forest with 78% accuracy, 79% precision, 78% recall, and 78% F1-score. Finally, the potential of using machine learning models to identify people who are overweight or obese was demonstrated. The practical use of the model developed will allow specialists in the health area to use it as an advantage for decision-making.

Keywords 1
Overweight and obesity, machine learning, classification models, body mass index

1. Introduction

Overweight and obesity, according to the World Health Organization (WHO), can be defined as the excessive accumulation of fat in different parts of the body [1] , and is recognized as an important public health problem as it is related to various diseases, and even morbidity and mortality [2].

Some diseases associated with obesity are: type 2 diabetes mellitus, hypertension, stroke, osteoarthritis, depression, Alzheimer's, and some types of cancer such as breast, prostate, kidney, ovary, liver, and colon cancer, among others [1, 3].

In this context, an adult is overweight when he has a Body Mass Index (BMI) ≥ of 25 kg/m^2 and is obese when a BMI ≥ of 30 kg/m^2 [1, 4]. In addition, according to WHO analyzes, over the years, the prevalence of obesity worldwide has almost tripled [1], becoming not only a problem in developed countries but also in developing countries [5, 6].

Figure 1 illustrates the evolution of the prevalence of overweight (a) and obesity (b) globally and by region. Regarding overweight, the global prevalence in 1975 was 20.2%, and by 2016 the prevalence increased to 39.1% (Figure 1(A)). The regions with the highest prevalence of overweight were Europe and America, with values of 40.0% and 35.0% in 1975, and the prevalence values increased to 62.3% and 63.3% in 2016, respectively [7].

IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine, November 19–21, 2021, Valencia, Spain
EMAIL: elias.aguirre@unesp.br (E. Rodríguez); elen.aguirre@unesp.br (E. Rodríguez); fernando.nascimento@unesp.br (L. Nascimento); aneirson.silva@unesp.br (A. da Silva); fernando.marins@unesp.br (F. Marins)
ORCID: 0000-0003-1120-1708 (E. Rodríguez); 0000-0002-3829-4118 (E. Rodríguez); 0000-0001-9793-750X (L. Nascimento); 0000-0002-2215-0734 (A. da Silva); 0000-0001-6510-9187 (F. Marins)

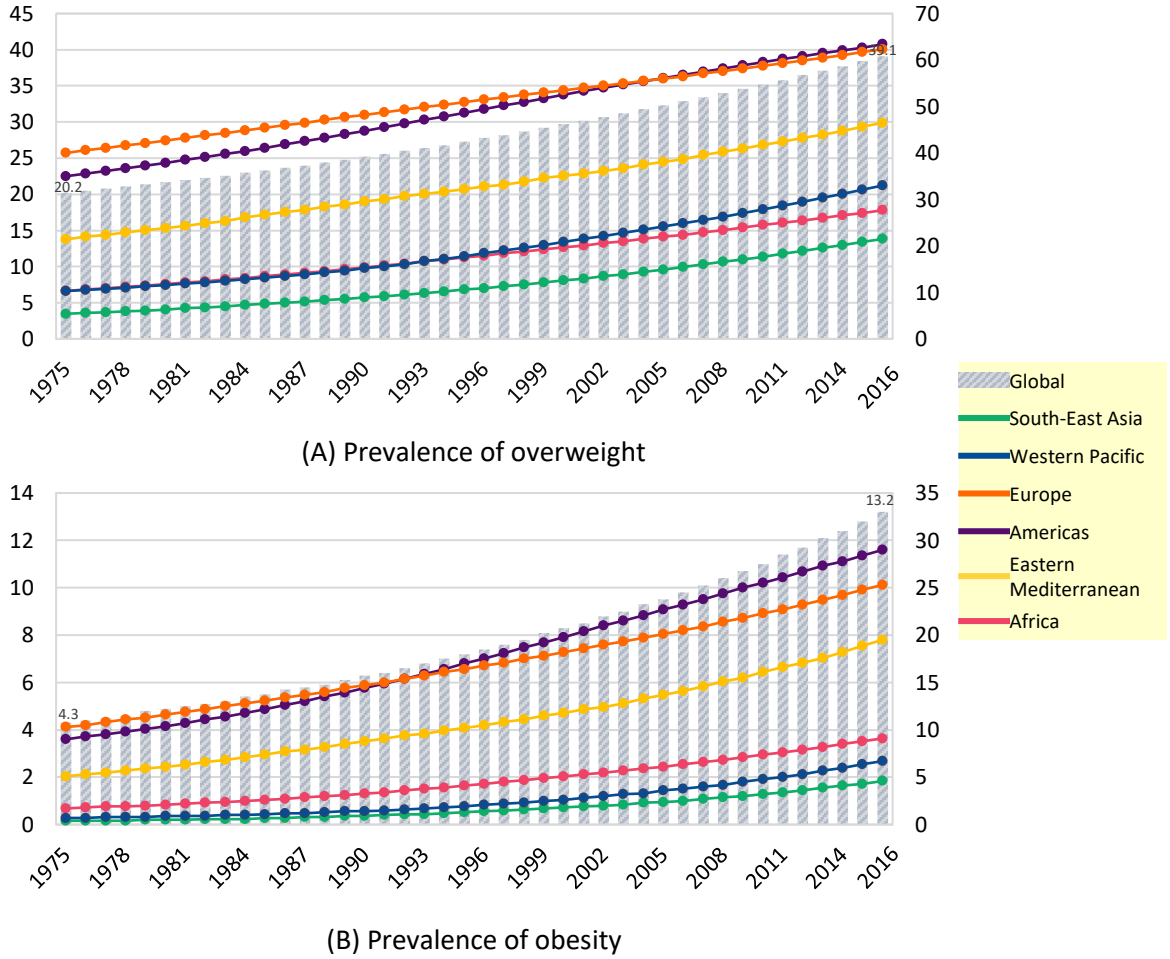


Figure 1: Prevalence of overweight and obesity globally and by region

Concerning obesity, the global prevalence in 1975 was 4.3%, and the increase in prevalence was greater for 2016, with a value equal to 13.2% (Figure 1(B)). Similarly, the regions with the highest prevalence of obesity were also Europe and America, with values of 10.1% and 9.0% in 1975, and in 2016 the prevalence of overweight values increased to 25.3% and 29.0%, respectively [7].

On the other hand, if the growing trend of the prevalence of obesity continues, it is estimated that by 2030 about 50% of the global population will be overweight or obese [8, 9].

Some studies affirm that the increase in the prevalence of obesity and overweight at a global level is due to complex changes in the population, concerning lifestyle, increased calorie consumption, decreased physical activity, and other factors such as urbanization, environmental changes, socioeconomic status, and genetic changes [8, 10].

Therefore, preventing obesity is complex, since it requires changes in the physical activity and eating habits of the population, in addition to the collective support from the government, industry, and the scientific and medical community [10], to minimize overweight by enabling the population to make sensible decisions regarding their lifestyle [11].

In this sense, in the literature, various studies can be found in which the development of technological tools is proposed, such as the application of machine learning [12, 13, 14], to support decision-making for specialists in the area, to reduce the prevalence of obesity and overweight.

Machine Learning can be defined as the study of computational methods for the identification of complex patterns in millions of data in order to build predictive models [15]. In addition, machine learning techniques in the health area have been gaining popularity in recent years [16].

This article aims to develop a predictive model using machine learning techniques with data collected through a survey to identify people with obesity and overweight and to make timely decisions.

The data collected for the development of the model are related to the physical condition and eating habits.

The work is organized as follows: Section 2 maps the publications in the literature on machine learning related to obesity or overweight; Section 3 describes the methodology and dataset used for the study; Section 4 presents the results and discussion; and finally, Section 5 brings the conclusions of this work, followed by the bibliographic references.

2. Machine learning in overweight and obesity

The number of publications where machine learning techniques were used in health problems is increasing, and the approach about the overweight and obesity is no exception. Thus, this section presents an analysis of the publications in SCOPUS about the application of machine learning techniques with data related to obesity or overweight.

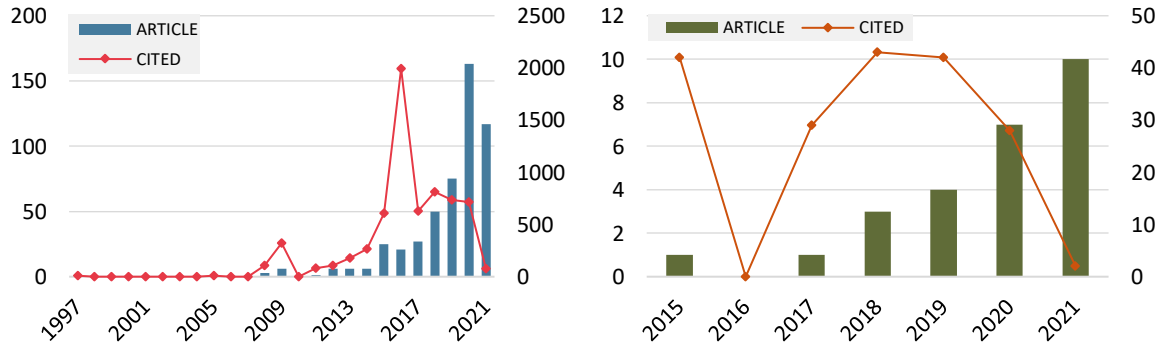
Table 1 presents the number of documents published with the keywords 'Obesity' AND/OR 'Overweight' AND 'Machine Learning'. The collection of the dataset for the analysis was refined by consulting published academic documents, including only articles and review articles, published from 1997 to 2021 (until the month of August).

Table 1
Results of the Scopus search

Keywords	Type	Publications	Citations	Impact fator	h-index
('Obesity' OR 'Overweight') AND 'Machine Learning'	Complete	508	6,649	13.09	36
('Obesity' OR 'Overweight') AND 'Machine Learning'	Title	26	186	7.15	7

From Table 1, it is evident that the combination of the keywords 'Obesity', 'Overweight' and 'Machine Learning' refined by 'Article title, Abstract, Keywords" results in a considerable number of publications (508) and citations (6,649), which is much greater than the number of publications (26) with the refined search just by 'Article title', which has 186 citations.

Figure 2 shows the time series of the distribution of articles published and the number of citations per year.



(A) Search result for 'Article title, Abstract, Keywords'

(B) Search result for 'Article title'

Figure 2: Number of publications and citations per year

In both cases (Figure 2 (A) and (B)), it is shown that there has been an increase in the number of publications, which indicates that there is a growing interest in the application of machine learning techniques in problems related to obesity. In addition, 70% and 81% of the documents found were published in the 2019-2021 period, as shown in Figure 2 (A) and (B) respectively.

Table 2 shows the description of some articles found in this analysis, which are related to this work.

Table 2
Articles related to this research

Author	Title	Data	Machine learning methods	Final model
Thamrin et al., 2021 [17]	Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018.	Data from an Indonesian national scale survey, with information from 300,000 families.	Logistic Regression, Classification and Regression Trees (CART), and Naïve Bayes.	The Logistic Regression model with 72% accuracy, 71% specificity, and 69% precision.
Dunstan et al., 2020[18]	Predicting nationwide obesity from food sales using machine learning.	Data was collected from 79 countries, with variables related to food and beverage sales, and the prevalence of obesity in adults.	Support vector machine, random forest, and extreme gradient boosting.	The random forest model with 0.057 root mean square error.
Machorro-Cano et al., 2019 [19]	PISIoT: A machine learning and IoT-based smart health platform for overweight and obesity control.	The dataset was collected by monitoring 40 elderly, of which a wearable device was assigned to each to obtain the biomedical variables.	Random forest, and the CART and J48 decision tree algorithms.	The J48 model.
Wang et al., 2018 [20]	Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing.	Data were collected with the informed consent of 139 recruited, with 74 obese and 65 non-obese individuals.	Support vector machine, k-nearest neighbor, and decision tree.	The SVM model with 70.77% accuracy, 80.09% sensitivity, and 63.02% specificity.
Dugan et al., 2015 [12]	Machine learning techniques for prediction of early childhood obesity.	Data collected from a pediatric clinical decision support system.	RandomTree, RandomForest, J48, ID3, Naïve Bayes, and Bayes trained.	The ID3 model with 85% accuracy and 89% sensitivity.

3. Material and Methods

Based on the purpose of this work, Figure 3 shows the pipeline diagram of the prediction model for the identification of obese or overweight people, based on the data collected in the survey, with the development of the predictive model (training and testing) and the selection of the final model.

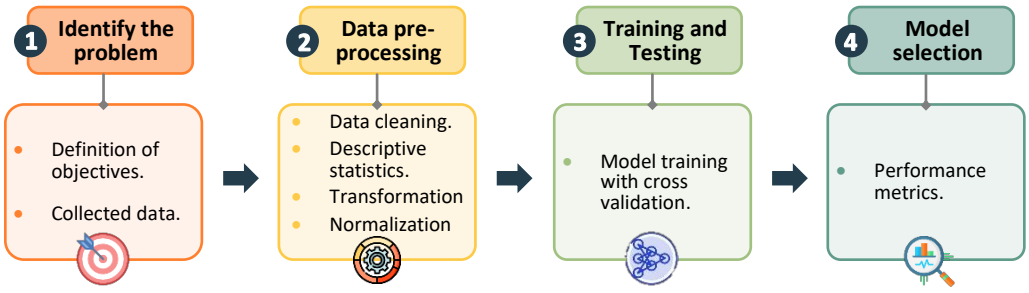


Figure 3: Pipeline diagram of the prediction model for the identification of obese or overweight people

3.1. Database Collection

The data set for this study was collected through a survey, in which 16 questions related to the interviewees' dietary habits and physical condition were applied. Table 3 presents the survey questions and dataset features.

Table 3
Survey questions and features

n°	Feature	Questions	Answers
Physical description features			
1	n_gender	What is your gender?	Female/ Male
2	n_age	What is your age?	Numeric
3	n_height	What is your height?	Numeric (m)
4	n_weight	What is your weight?	Numeric (kg)
5	c_FMOW	Has a family member suffered or suffers from overweight?	Yes/ No
Features of dietary habits			
6	c_ECFF	Do you eat high caloric food frequently?	Yes/ No
7	c_EVM	Do you usually eat vegetables in your meals?	Never/ Sometimes/ Always
8	c_MMHD	How many main meals do you have daily?	Between 1 or 2/ Three/ More than three
9	c_EFBM	Do you eat any food between meals?	Never/ Sometimes/ Frequently/ Always
10	c_SMOKE	Do you smoke?	Yes/ No
11	c_WDRD	How much water do you drink daily?	Less than a liter/Between 1 and 2 L/ More than 2 L
12	c_DRAL	How often do you drink alcohol?	I do not drink/ Sometimes/ Frequently/ Always
Physical condition features			
13	c_MCED	Do you monitor the calories you eat daily?	Yes/ No
14	c_HPHA	How often do you have physical activity?	Never/1 or 2 days/2 or 4 days/4 or 5 days
15	c_TTEC	How long do you use technological devices?	0–2 hours/ 3–5 hours/ More than 5 hours
16	c_TRANSP	What type of transportation do you usually use?	Automobile/ Motorbike/ Bike/ Public Transportation/ Walking

As shown in Table 3, the 16 features of the dataset can be grouped into three categories: (a) physical description features, (b) features of dietary habits, and (c) physical condition features.

Physical description Features. Includes features such as gender, age, height, weight and the existence of overweight relatives (c_FMOW).

Features of eating habits. Includes the features of consumer foods high in calorie (c_ECFF), vegetable consumption (c_EVM), number of main meals (c_MMHD), food between meals (c_EFBM), water consumption (c_WDRD), alcohol consumption (c_DRAL) and tobacco consumption (c_SMOKE).

Physical condition Features. Includes calorie consumption (c_MCED), physical activity (c_HPHA), use of technology devices (c_TTEC), and type of transport used (c_TRANSP).

The target feature was of the categorical type, and was calculated through the data labeling process, in which, for each instance of the dataset, the BMI was calculated using the weight (measured in kilograms) and height (measured in meters) information, according to Equation (1):

$$BMI = \frac{Weight(kg.)}{Height(m.) \times Height(m.)} \quad (1)$$

In the identification of classes for data labeling, the classification table of BMI values was used (Table 4), which is made available by the World Health Organization (WHO) [4].

Table 4
Body mass index classification

Classification	BMI
Underweight	Below 18.5
Normal weight	18.5 – 24.9
Pre-obesity	25.0 – 29.9
Obesity class I	30.0 – 34.9
Obesity class II	35.0 – 39.9
Obesity class III	Above 40

3.2. Data transformation and normalization

3.2.1. Dataset balancing

Dataset balancing: Classification models trained with unbalanced data tend to make biased predictions with wrong results, so in many cases, the classes with fewer instances are not enough for the model, and it is necessary to apply a sub-process so that the data is balanced [21]. In this case, the classes in the collected dataset were unbalanced, so the *"oversampling"* technique was used to balance the minority classes in the training dataset.

3.2.2. Categorical data encoding

In the database used, about 80% of the variables were categorical, so data transformation techniques were used since some of the machine learning algorithms do not allow the use of non-numerical data [21]. The features of c_MMHD, c_WDRD, c_HPHA, c_TTEC and c_TRANSP were transformed with the one hot encoding technique, and the features of c_gender, c_FMOW, c_ECFF, c_SMOKE, c_MCED, c_EVM, c_DRAL, and c_EFBM with the ordinal encoding technique. The label encoding technique was used to transform the classes of the target feature.

3.2.3. Data normalization

Data normalization helps to ensure that features with different ranges do not affect the trained model. In *MIN-MAX* normalization, the data is scaled in the range of $[0 - 1]$ or $[0.0 - 1.0]$ [22], and is calculated according to Equation (2):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2)$$

where x' is the normalized value, x is the original value, and x_{max} and x_{min} are the minimum and maximum values, respectively [22].

3.3. Classification in supervised machine learning

The main task of supervised learning is to learn the behavior of a function $\mathcal{Y} = f(\mathcal{X}, \beta)$ based on the features $\mathcal{X} \times \mathcal{Y}$, belonging to the training dataset $\mathcal{T}$, where $\mathcal{X} = (x_1, x_2, \dots x_d) \in \mathbb{R}^d$, each x_d is a feature, and $\mathcal{Y} \in C$ with $C = (c_1, c_2, \dots, c_j)$, j labels [23]. In classification problems, the objective of the learned categorical function $f(\mathcal{X})$ is the mapping of the input variables, to predict a new label $\mathcal{Y}$ [24].

The classification models used in this work are described below:

- ***Decision Tree (DT).*** DT is very popular for its simple structure, ease of interpretation, and for its efficiency [26]. The construction of the tree presents an iterative process, starting from a training dataset ($\mathcal{T}$) with n observations, recursively partitioned, thus dividing into increasingly homogeneous data subsets [24, 25].
- ***Support Vector Machines (SVM).*** SVM construct optimal separation limits between variables, applying the input data in a larger nonlinear space, called characteristic space [27]. Furthermore, the algorithm uses different kernel functions to model different degrees of nonlinearity and efficiency [25].
- ***K-Nearest Neighbors (KNN).*** The main objective of the KNN classifier is to predict the closest value using distance as a basis is a widely used technique the euclidean distance [23]. The classification of the input data is based mainly on the selection of the majority class among its nearest neighbors [28].
- ***Gaussian Naive Bayes (GNB).*** GNB classifier is considered a powerful probabilistic algorithm, based on Bayes Theorem, in which it ignores the possible dependencies between characteristics, reducing the multivariate problem to a univariate problem [23, 24, 28].
- ***Multilayer Perceptron (MLP).*** MLP is a nonparametric estimator, which has the structure of an artificial neural network and is formed by one or more interconnected layers [24]. This algorithm uses the backpropagation technique to improve the prediction of the model, in which the gradient is calculated using the error function and the neural network weights, and each node in the model uses a non-linear activation function [25].
- ***Random Forest (RF).*** RF is a flexible algorithm and is an expansion of the Decision Tree. This algorithm creates randomness from a dataset and trains each of its trees with different random data, then the trees are grouped, and by combining their results the errors are calculated to have a more accurate prediction [23].
- ***Gradient Boosting (GB).*** GB is a robust classifier that combines several sequential classifiers, each classifier has a different weight, and the error calculated by each classifier is used to improve the prediction value [23, 27].
- ***Extreme Gradient Boosting (XGB).*** XGB method is an improved version of Gradient Boosting, in which this algorithm uses a split search approach for existing sparse patterns in the data to make the training of the model more efficient, and the risk of overfitting the model is controlled [23, 24].

3.4. Cross Validation and performance measures

Cross validation is used for the evaluation of machine learning models, where the dataset $\mathcal{X}$, used for the development of the model, is partitioned into k subsets (k-folds) of data of equal size ($\mathcal{X}_i$, where $i = 1, 2, \dots, k$) [23,24]. In the division of the dataset, $k - 1$ is used for model training $\mathcal{T}$, and the remainder is used for model validation $\mathcal{V}$, as shown in Equation (3):

$$\begin{array}{ll} \mathcal{V}_1 = \mathcal{X}_1 & \mathcal{T}_1 = \mathcal{X}_2 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_k \\ \mathcal{V}_2 = \mathcal{X}_2 & \mathcal{T}_2 = \mathcal{X}_1 \cup \mathcal{X}_3 \cup \cdots \cup \mathcal{X}_k \\ \vdots & \vdots \\ \mathcal{V}_k = \mathcal{X}_k & \mathcal{T}_k = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_{k-1} \end{array} \quad (3)$$

The main advantage of the k-fold method is that each subset is used in the testing process only once, reducing the possibility of biased training [23].

On the other hand, the models are evaluated, assuming that the validation data ($\mathcal{V}$) follow the same distribution as the training dataset ($\mathcal{T}$). In the literature, there are several evaluation metrics, however, a standard metric to measure the efficiency of prediction models has not yet been established [17]. In

this work, the performance of the tested models is evaluated based on the comparison of different metrics.

According to Marsland [23], the following lines describe the performance metrics used in this work:

- ***Confusion matrix.*** It helps to evaluate the quality of the classification model and is represented by a matrix, which allows visualizing the performance of each class of the prediction model, as illustrated in the following matrix:

		Predicted		
		Positive	Negative	Total
Actual	Positive	TP	FP	TP + FP
	Negative	FN	TN	FN + TN
	Total	TP + FN	FP + TN	

- ***Accuracy.*** Defined as the sum of the total of true positives and true negatives divided by the total number of results, Equation (4):

$$Accuracy = \frac{n° TP + n° FP}{n° TP + n° FP + n° FN + n° TN} \tag{4}$$

- ***Precision.*** Defined as the ratio of the number of correctly predicted true positives to the total number of predicted positives, according to Equation (5):

$$Precision = \frac{n° TP}{n° TP + n° FP} \tag{5}$$

- ***Recall.*** It is also known as the true positives rate. Defined as the ratio of the number of correct positives to the total predictions classified as positives, as was shown in Equation (6):

$$Recall = \frac{n° TP}{n° TP + n° FN} \tag{6}$$

- ***F1-score.*** A single measure that combines the sensitivity and precision value of the prediction model, according to Equation (7):

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

4. Results and discussion

This section presents the exploratory analysis of the collected data and the main results of the experiments performed in the model development process.

4.1. Exploratory Analysis

The age, weight, and height features are quantitative of the rational type. Table 5 presents the descriptive statistics of the features.

Table 5
Descriptive statictics

Statictics	Mean	S.D.[a]	Min[b]	Q1[c]	Q2[d]	Q3[e]	Max[f]
Age	24.31	6.35	14.00	19.95	22.78	26.00	61.00
Height	1.70	0.09	1.45	1.63	1.70	1.77	1.98
Weight	86.59	26.19	39.00	65.47	83.00	107.43	173.00

a) S.D.: Standard deviation, b) MIN: Minimun, c) Q1: First quartile, d) Q2: Second quartile (median), e) Q3: Third quartile, f) MAX: Maximun.

Additionally, Figure 4 shows the data distribution of the numerical features.

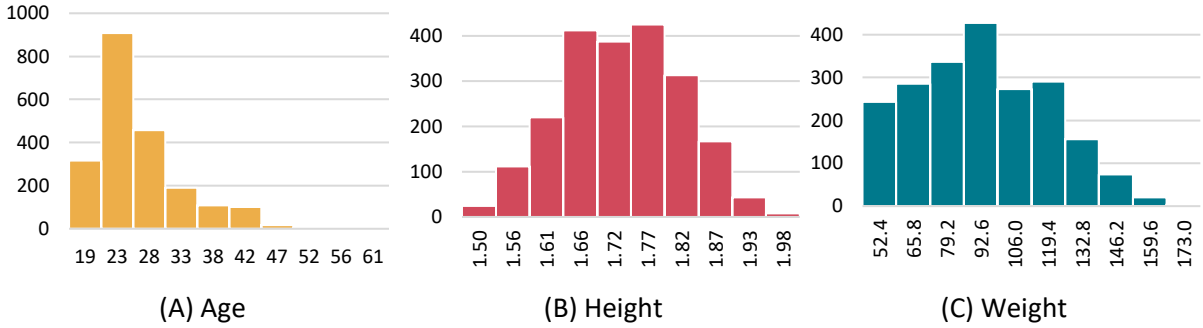


Figure 4: Distribution of the numerical features

The age data used for this work ranged from 14 to 61 years of age, with a mean age of 24 years. The mean height of the data collected was 1.7 (m.), the range was 1.45 (m.) as a minimum and 1.98 (m.) as a maximum. The data collected related to weight ranged from 39.00 (kg.) to 173.00 (kg.), and the mean weight was 86.59 (kg.), Table 5.

The target feature of the classes was calculated using the weight and height of each instance with Equation (1). Figure 5 shows the distribution of the calculated BMI values.

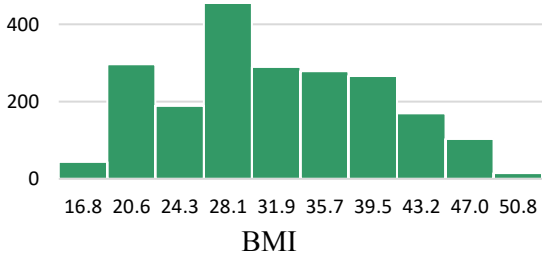


Figure 5: Distribution of the BMI

The calculated BMI values ranged from 13.00 to 50.81, with a mean equal to 29.70 and a standard deviation of 8.01. In Figure 6 of the number of instances per class of the target feature, it is observed that the classes of the collected dataset were unbalanced, as the pre-obesity class (26.8%) has the largest number of instances.

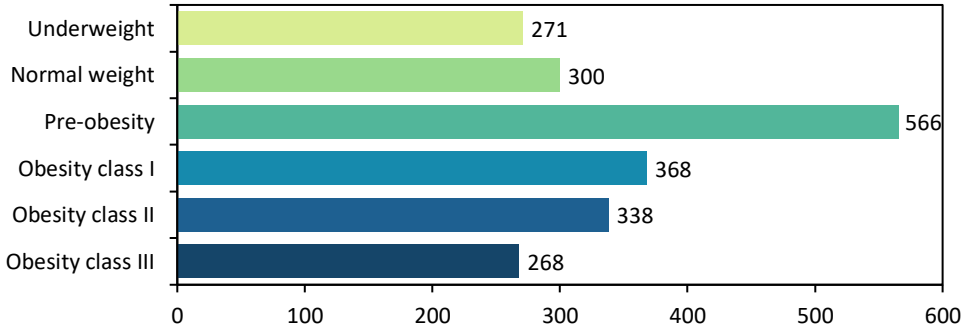


Figure 6: Target feature classes

4.2. Experiments and evaluation

The experiments were carried out with 14 features of the data set, variables such as: n_gender, n_age, c_FMOW, c_ECFF, c_EVM, c_MMHD, c_EFBM, c_SMOKE, c_WDRD, c_DRAL, c_MCED, c_HPHA, c_TTEC, and c_TRANSP (Table 3). The weight and height features were used to calculate the BMI, and also to classify each instance.

In the data transformation process, the characteristics of c_MMHD, c_WDRD, c_HPHA, c_TTEC and c_TRANSP were decomposed into dummy variables. The age feature was normalized so that the

difference between the ranges of each feature does not affect the performance of the model. Furthermore, class balancing was only applied in the training set, since the target feature has 6 unbalanced classes (see Table 4 and Figure 6).

In the training and evaluation process, eight machine learning methods were tested (subsection 3.3). The prediction of the tested models were internally validated by 10-fold cross-validation. On the other hand, the Random Search method was used to find the best combinations of hyperparameters for each of the tested models. Table 6 presents the optimal hyperparameters for each model.

Table 6
Hyperparameters of the models

Models	Hyperparameters
Decision tree	splitter: 'best', min_samples_split: 2, min_samples_leaf: 1, criterion: 'entropy'
Support vector machines	C: 90.11, break_ties: True, kernel: 'linear', probability: False, shrinking: True, tol: 0.51
K-Nearest Neighbors	weights: 'distance', p: 1, n_neighbors: 2, leaf_size: 8, algorithm: 'kd_tree'
Gaussian Naive Bayes	var_smoothing: 0.0009
Multilayer Perceptron	tol: 0.01, solver: 'lbfgs', shuffle: True, max_iter: 530, learning_rate_init: 0.30, learning_rate: 'invscaling', alpha': 0.69, activation: 'relu'
Random forest	n_estimators: 390, min_samples_split: 8, min_samples_leaf: 1, criterion: 'gini', bootstrap: False
Gradient Boosting	loss: 'deviance', criterion: 'mse'
Extreme Gradient Boosting	'booster': 'gbtree', 'eta': 0.0033752879456183816, gamma: 1, predictor: 'cpu_predictor', tree_method: 'approx'

Figure 7 shows the performance behavior in each fold, according to the metrics of accuracy, precision, recovery, and F1-score.

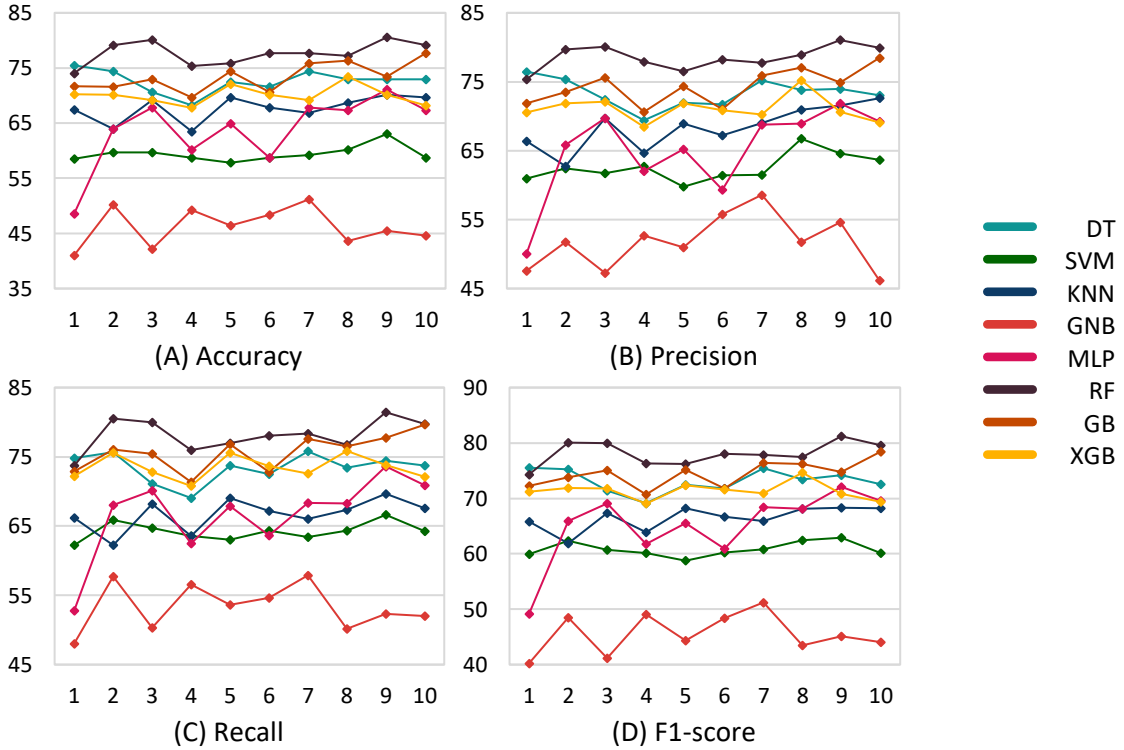


Figure 7: Performance of models tested with cross-validation (10-folds)

Figure 7 shows that each model generated with cross-validation showed a different behavior, according to each subset of data from each fold. The models that presented a greater variation in their performance values in each fold, according to the metrics considered, were the gaussian naive bayes (GNB) and multilayer perceptron (MLP) models.

Additionally, Table 7 present the mean performance values of the trained models.

Table 7
Mean performance values of the models

Models	Accuracy ($CI_{95\%}$[a])	Precision ($CI_{95\%}$)	Recall ($CI_{95\%}$)	F1-score ($CI_{95\%}$)
Decision Tree	72.62 (±1.49)	73.33 (±1.49)	73.43 (±1.49)	73.11 (±1.47)
Support Vector Machines	59.45 (±1.03)	62.55 (±1.44)	64.24 (±0.94)	60.83 (±0.94)
K-Nearest Neighbors	67.69 (±1.67)	68.37 (±2.23)	66.70 (±1.66)	66.44 (±1.53)
Gaussian Naive Bayes	46.24 (±2.48)	51.70 (±2.84)	53.33 (±2.41)	45.55 (±2.58)
Multilayer Perceptron	63.77 (±4.65)	65.11 (±4.66)	66.60 (±4.17)	65.05 (±4.69)
Random Forest	77.69 (±1.52)	78.53 (±1.25)	78.15 (±1.69)	78.09 (±1.52)
Gradient Boosting	73.43 (±1.88)	74.34 (±1.85)	75.67 (±1.85)	74.45 (±1.69)
Extreme Gradient Boosting	70.06 (±1.21)	71.10 (±1.34)	73.51 (±1.23)	71.37 (±1.11)

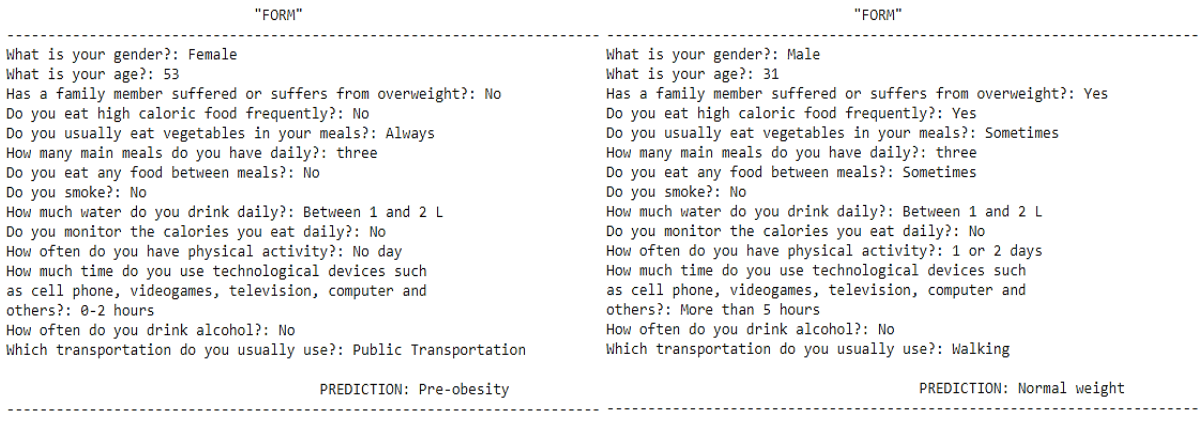
(a) $CI_{95\%}$: Confidence interval

The results of Table 7 show that the gaussian naive bayes model presented the lowest performance values, compared to the other models. On the other hand, the extreme gradient boosting, decision tree, gradient boosting, and gradient boosting models showed better performance, with values higher than 70% in the evaluated metrics.

The model with the best performance in all the metrics evaluated was the random forest (final model selected), with 77.69% (±1.52) accuracy, 78.53% (±1.25) precision, 78.15% (±1.69) recall, and 78.09% (±1.52) F1-score.

The random forest model selected even obtained values that surpass the results obtained in other similar researches where obesity was analysed with features associated with eating habits and physical condition, such as the study by Tharmin et al. [17] who selected the logistic regression model as the best, generating 72% accuracy and 69% precision, and Wang et al. [20] with 70.77% accuracy in the SVM model selected.

Finally, the final model was tested with real data, to identify people with obesity or overweight. Figure 8 shows each of the data entered in the final model and its respective prediction for each query made.



(A) First query with real data (B) Second query with real data

Figure 8: Model predictions with real data.

The BMI for each interviewee was calculated manually, to check the classes predicted by the classification model. From Figure 8(A), the weight and height of the 53-year-old interviewee were 65 kg and 1.57 m, respectively. The BMI of the first interviewee was 26.37 kg/m^2, which according to

Table 4 is classified as pre-obesity. The weight and height of the 31-year-old interviewee were 75 kg and 1.74 m respectively, and the BMI = 24.77 kg/m^2, which is classified as normal weight (Figure 8(B)). Therefore, these results show that the values obtained with the model and those calculated manually were the same.

5. Conclusion

Overweight and obesity are considered an epidemiological problem since it is related to various diseases such as hypertension, type 2 diabetes mellitus, osteoarthritis, stroke, some types of cancer, and even death. At the global level, the prevalence of overweight and obesity has increased considerably, and it is estimated that by 2030 the global prevalence of overweight will increase to around 50%.

Preventing obesity is not an easy task as it requires the intervention of government, industry, and specialists in the area. Thus, curbing overweight and its effects requires important changes in the population's lifestyle, mainly in physical activity and eating habits.

In this context, thanks to technological advances, there are currently tools such as artificial intelligence and machine learning that can be used as support for the prevention and identification of people with overweight and obesity. Machine learning models can be used as an advantage in the medical field since specialists in the area can make use of these tools as a support tool for decision-making.

In this work, eight machine learning models were tested, such as decision tree, support vector machines, k-nearest neighbors, gaussian naive bayes, multilayer perceptron, random forest, gradient boosting, and extreme gradient boosting, to develop an intelligent model for the identification of people with obesity or overweight, which will serve as support in the decision-making to specialists in the area.

On the other hand, the data used for the development of the model were collected through a survey, in which information related to the eating habits and physical activity of the interviewees was collected. The database labeling was performed based on the BMI classification table, in which the weight and height of each instance were used.

Furthermore, 10-fold cross-validation was used and the hyperparameters of the tested models were optimized. The results showed that the random forest model obtained the best results in the performance metrics, with 77.69% accuracy, 78.53% precision, 78.15% recall, and 78.09% F1-score. Other tested models that obtained values in the performance metrics higher than 70% were extreme gradient boosting (70.06% accuracy), decision tree (72.62% accuracy), and gradient boosting (73.43% accuracy).

Finally, machine learning models showed good performance for BMI classification, even when data related to diet and eating habits were used. The models developed demonstrated that machine learning is a powerful tool that can be used in the medical field to make decisions for timely treatment for people at risk of obesity.

6. Acknowledgements

This research was supported by the Coordination for the Improvement of Higher Education Personnel [CAPES - 001].

7. References

[1] World Health Organization (WHO), Obesity and overweight, 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] H. Rosen, “Is Obesity A Disease or A Behavior Abnormality? Did the AMA Get It Right?”. Missouri medicine, 111(2014): 104–108.

[3] M. Blüher, “Obesity: global epidemiology and pathogenesis”. Nature Reviews Endocrinology, 15(2019): 288–298. doi: 10.1038/s41574-019-0176-8.

[4] World Health Organization (WHO), Body mass index–BMI, 2020. URL: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

[5] T. Bhurosy, R. Jeewon, “Overweight and obesity epidemic in developing countries: a problem with diet, physical activity, or socioeconomic status?”. The Scientific World Journal, (2014). doi:10.1155/2014/964236.
[6] A. Alami, A. Jafari, Z. Hosseini, “Differences in overweight/obesity prevalence by demographic characteristics and self-weight misperception status”. Clinical Nutrition ESPEN, 41(2021): 249–253. doi:10.1016/j.clnesp.2020.12.005.
[7] World Health Organization (WHO), NCD risk factors: Overweight / Obesity, 2021. URL:https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/overweight-obesity.
[8] T. Kelly, W. Yang, C.-S. Chen, K. Reynolds, J. He, “Global burden of obesity in 2005 and projections to 2030”. International Journal of Obesity, 32(2008): 1431–1437.
[9] R. Mehrzad, “Definition and introduction to epidemiology of obesity”. Obesity, (2020): 1–6. doi:10.1016/b978-0-12-818839-2.00001-6.
[10] Y.C. Chooi, C. Ding, F. Magkos, “The epidemiology of obesity”. Metabolism, 96(2018):6–10. doi: 10.1016/j.metabol.2018.09.005
[11] The Lancet Diabetes and Endocrinology, “ Tackling obesity in 2020—with a great resolution comes shared responsibility”. The Lancet Diabetes & Endocrinology, (2020). doi:10.1016/s2213-8587(20)30001-2.
[12] T.M. Dugan, S. Mukhopadhyay, A. Carroll, S. Downs, “Machine learning techniques for prediction of early childhood obesity”. A pplied Clinical Informatics, 6(2015):506–520. doi: 10.4338/ACI-2015-03-RA-0036.
[13] K.W. DeGregory et al., “A review of machine learning in obesity”. Obesity Reviews, 19(2018):668–685. doi: 10.1111/obr.12667.
[14] E. De-La-Hoz-Correa et al., “Obesity Level Estimation Software based on Decision Trees”. Journal of Computer Science, 15 (2019): 67–77. doi:10.3844/jcssp.2019.67.77.
[15] M.I. Jordan, T.M., “Mitchell, Machine learning: Trends, perspectives, and prospects”. Science, 349(2015):255–260. doi: 10.1126/science.aaa8415.
[16] S.N. Kumar, P. Saxena, R. Patel, A. Sharma, D. Pradhan, H. Singh et al., “Predicting risk of low birth weight offspring from maternal features and blood polycyclic aromatic hydrocarbon concentration”. Reproductive Toxicology, 94(2020):92–100. doi: 10.1016/j.reprotox.2020.03.009.
[17] S.A. Thamrin, D.S. Arsyad, H. Kuswanto, A. Lawi, S. Nasir, “Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018”. Frontiers in Nutrition, 8(2021). doi:10.3389/fnut.2021.669155.
[18] J. Dunstan, M. Aguirre, M. Bastías, C. Nau, T.A. Glass, F. Tobar, “Predicting nationwide obesity from food sales using machine learning”. Health Informatics Journal, 26(2020):652–663. doi:10.1177/1460458219845959.
[19] I. Machorro-Cano, G. Alor-Hernández, M.A. Paredes-Valverde, U. Ramos-Deonati, J.L. Sánchez-C ervantes, L. Rodríguez-Mazahua, “PISIoT: A machine learning and IoT-based smart health platform for overweight and obesity control”. Applied Sciences, 9(2019). doi: 10.3390/app9153037.
[20] H.-Y. Wang et al. “Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing”. Journal of Computational Biology, 25(2018):1347–1360. doi:10.1089/cmb.2018.0002.
[21] V. Kotu and B. Deshpande, “Chapter 3 - data exploration,” in Data Science (Second Edition), second edition ed., V. Kotu and B. Deshpande, Eds. Morgan Kaufmann, 2019:39–64.
[22] S. Jain, S. Shukla, and R. Wadhvani, “Dynamic selection of normalization techniques using data complexity measures”. Expert Systems with Applications, 106(2018):252–262. doi:10.1016/j.eswa.2018.04.008.
[23] S. Marsland, Machine Learning: An Algorithmic Perspective, 2015. 6000 Broken Sound Parkway NW, Suite 300, Boca Raton: Taylor and Francis Group, LLC.
[24] E. Alpaydim, Introduction to Machine Learning, 2004. Cambridge, Massachusetts, London, England: The MIT Press.
[25] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review”. Journal of Biomedical Informatics, 25(2002), 352–359. doi:10.1016/S1532-0464(03)00034-0.

[26] M. Bublyka, V. Lytvyna, V. Vysotskaa, L. Chyrunb, Y. Matseliukha, N. Sokulskac , “The Decision Tree Usage for the Results Analysis of the Psychophysiological Testing”. The 3rd Conference on Informatics and Data-Driven Medicine (IDDM 2020), 2753(2020): 458–472.
[27] J. Wu, J. Roy, and W. F. Stewart, “Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches”. Medical Care, 48(2010):S106–S113. doi: 10.1097/MLR.0b013e3181de9e17.
[28] N. Boyko, K. Boksho, “Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data”. The 3rd Conference on Informatics and Data-Driven Medicine (IDDM 2020), 2753(2020): 230–239.
[29] G. Li and J. Shi, “On comparing three artificial neural networks for wind speed forecasting”. Applied Energy, 87(2010): 2313–2320. doi: 10.1016/j.apenergy.2009.12.013.