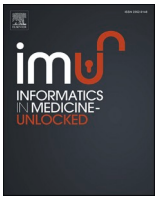
Informatics in Medicine Unlocked 29 (2022) 100901

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Estimation of obesity levels based on dietary habits and condition physical using computational intelligence

Juan Piero Santisteban Quiroz

Universidad Nacional Mayor de San Marcos, Lima, Peru

ARTICLE INFO

Index Terms:
Obesity
XG Boost
Light GBM
Random forest
Decision tree
Extremely randomized trees
Logistic regression
Data mining
AUC
ROC

ABSTRACT

Obesity is a disease that affects the health of men and women, and in recent decades it had an increasing trend, the WHO estimates that by 2030 more than 40% of the world's population will be overweight and more than a fifth will be obese Consequently, researchers have made great efforts to identify early the factors that influence the generation of obesity. There are tools limited to the calculation of BMI, omitting other relevant factors such as: if the individual has a family history of obesity, time spent on exercise routines, genetic expression profiles and other factors. In this study, a computational intelligence model is created, based on supervised and unsupervised data mining techniques such as Light Gradient Boosting Machine (Light GBM) classifier, random forest (RF), decision tree (DT), Extremely Randomized Trees (ET) and the and logistic regression (LR), to identify obesity levels based on lifestyle. In this research, the main source of data was a study of 2,111 people from the countries Colombia, Mexico and Peru, aged between 14 and 61 years. The study takes a set of data related to the main causes of obesity, based on the objective of referring to the high caloric intake, the decrease in energy expenditure due to lack of physical activity, eating disorders, genetics and socioeconomic factors. The results show that the LightGBM classification model has the highest weighted value of AUC (0.9990), improving the results of previous studies with similar antecedents.

1. Introduction

Obesity is a disease that affects the health of men and women, and in recent decades it has tended to increase; such is the case that "worldwide between 1980 and 2008, the average body mass index (BMI) increased by 0.4 $\frac{kg}{m^2}$ in men and 0.5 $\frac{kg}{m^2}$ in women per decade. In Latin America, the increase per decade was 0.6 $\frac{kg}{m^2}$ in men and 1.4 $\frac{kg}{m^2}$ in women. According to projections based on information from the World Health Organization (WHO), and if the trend continues, it is estimated that by 2030 more than 40% of the world's population will be overweight and more than one fifth will be obese" [1].

The obesity, identified by body mass index (BMI), and abdominal obesity (OA) by waist circumference (WC), are considered risk factors for the development of cardiovascular diseases, diabetes mellitus, dyslipidemia, among others, through metabolic disorders such as insulin resistance (IR [2].

Several researchers have made great efforts to identify early the factors that influence the generation of obesity, even creating web tools such as the calculation of the body mass index (BMI) [3,4], here you can calculate the level of obesity of a person; however, these tools are limited to calculating BMI, omitting other relevant factors such as whether the individual has a family history of obesity, time spent on exercise routines, gene expression profiles, and other factors. The high cost of living caused by obesity can cause significant financial hardship for individuals and families with this disease. Therefore, an intelligent tool is needed, capable of efficiently detecting the risk of obesity.

The purpose of this study is to make people aware of the risk of obesity based on some key factors of their lifestyle, for which a computational intelligence model is created, comparing supervised data mining techniques such as Light GBM, XG Boost, random forest (RF), decision tree (DT), Extreme Random Trees (ET) and logistic regression, to identify levels of obesity based on 16 factors of eating habits and physical condition, using as main data source a study of 2,111 people from the countries Colombia, Mexico and Peru, aged between 14 and 61 years. This study takes a set of data related to the main causes of obesity, based on the objective of referring to the high caloric intake, the decrease in energy expenditure due to lack of physical activity, eating disorders, genetics and socioeconomic factors [5]. The results show that the LightGBM classification model has the highest weighted value of AUC (0.9990), improving the results of previous studies with similar

E-mail address: juan.santisteban1@unmsm.edu.pe.

https://doi.org/10.1016/j.imu.2022.100901
Received 30 July 2021; Received in revised form 26 February 2022; Accepted 27 February 2022
Available online 6 March 2022

Table 1
Description of the variables.

Attribute	Description
Gender	Sex
Age	Age
Height	Height
Weight	Weight
FHWO	Overweight family members
FAVC	Consume high-calorie foods frequently
FCVC	Number of meals where you usually eat vegetables
NCP	Number of main meals a day
CAEC	Eat food between meals
SMOKE	How often you smoke
CH2O	Liters of water you drink a day
SCC	Monitor the calories you consume daily
FAF	Frequency of days per week that you often have physical activity
TUE	Time of use of technological devices on a daily basis
CALC	Frequency of alcohol intake.
MTRANS	Means of transportation that you use regularly
NObeyesdad	Body mass index

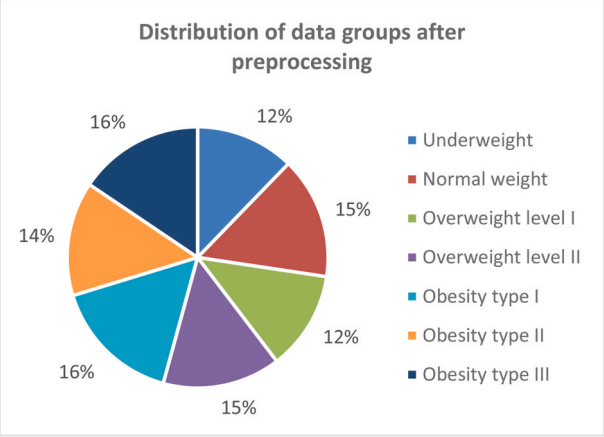


Fig. 1. Distribution of data groups after preprocessing.

background.

2. Related works

In this section I analyze the work of some researchers related to the prediction of obesity levels, trying to understand the method applied and the forms of validation.

The authors in Ref. [6] presented an intelligent method, based on supervised and unsupervised data mining techniques using machine learning techniques that included: decision tree (DT), Support vector machine (SVM) and Simple K-Means to detect obesity levels and help people and health professionals to have a healthier lifestyle in the face of this global epidemic. They used a data set of 178 students that included 18 variables that can determine if a person is obese. They distributed the data in 4 groups using Simple K-Means, the grouping result was not balanced, since 69% of the data were people who were considered not prone to overweight problems and the remaining 31% were people prone to overweight problems, and the authors did not mention whether they applied any data balancing technique. The results obtained by the DT + Simple K-Means hybrid method had a precision of 98,5% and an ROC area of 99,5%.

The authors in Ref. [7] presented a prediction system for obesity risk in Bangladeshi, using machine learning techniques. The data used was the product of surveys of 1,100 Bangladeshi people, which included 28 factors considered to cause obesity, based on previous research and expert judgment. They distributed the data in 3 groups: high (48%), medium (30%) and low (22%) risk of obesity, there is evidence of an imbalance in the data on the risk of high obesity with respect to the medium and low risk, the authors did not mention if they applied any data balancing technique. They performed the training with 80% of the data, applying supervised machine learning techniques such as: k-NN, logistic regression, SVM, CART, random forest, MLP, Ada boost and GBM. The results obtained show the logistic regression technique with the highest performance, obtaining a 97,09% precision, F1-score (97%) and the GBM technique with the least precision with 64.08% and F1-score (57%), additionally, the sensitivity, specificity and recall were evaluated.

The authors in Ref. [8] presented a software for obesity estimation based on the SEMMA data mining methodology, using machine learning techniques: decision trees (DT), Bayesian networks and logistic regression (LR). The data used were the product of a study in Colombia, Mexico and Peru, in which 712 university students aged between 18 and 25 participated., where a survey was applied to understand the behavior of obese people and identify through questions the level of obesity based on physical features, social factors and others. After the training of the models, these were validated through 3 metrics: Recall, true positive rate and false positive rate, the authors did not evaluate the ROC curve. The decision tree model was selected as the best technique because it obtained the best precision (97,4%). "The best technique was implemented in desktop software developed with the Java language supported by the Weka Toolkit, in which the user must complete a form with questions related to their physical characteristics, social habits, food, among others, and the model predicted the level of obesity that the individual had".

The reference authors [9] presented an intelligent model to predict the high risk of childhood obesity before the greatest increase in BMI occurs. The data used was 132.262 electronic medical health records of Israel's children from 2002 to 2018. The data included diagnoses, prescribed medications, child and family lab tests, and demographic data. The models were trained with data from the first 2 years of life and the risk of obesity was estimated at 5 and 6 years. "The gradient boosting trees technique was trained with a portion of the data, the quality of the model was evaluated by calculating the area under the ROC curve and the area under the Precision-Recovery curve. The results show that the gradient boostring trees technique obtained an auROC of 0,803 and auPR of 0,312".

The authors in Ref. [10] presented a spatial regression model to geographically determine the prevalence of obesity in adults in New York in 2013. The local scale data used was downloaded from the NYC Community Health Profiles and Atlas which contained the prevalence of obesity, socioeconomic status, behavioral factors, and surrounding building characteristics. The spatial regression model of the Matrix Exponential Spatial Specification (MESS) model was trained. To determine the performance of the model, the coefficient of determination R^2 was used. The results of the experimentation show that the MESS model obtained an R^2 of 0,8353, in addition the results indicated that the reduction of adult obesity is better benefited by the decrease in the consumption of sugary drinks than by the increase in physical activity.

The authors in Ref. [11] conducted a study that aimed to examine multiple intergenerational risk factors for obesity among children aged 24–80 months using data from national cohorts in Korea. The data used was extracted from the Korean National Health Insurance (KNHI) database, with a total of 1,001,775 family records. Socioeconomic status (SES) and factors related to parents and children were examined. The decision tree machine learning technique was trained, and descriptive statistics were applied to predict the prevalence of factors associated with obesity in families. The results show that the prevalence of obesity was 6.57% and that of overweight was 11.31% in the entire study population. The best factors predicted by the decision tree were the following: obese mothers before conception, obese fathers, recipients of non-medical help, and mothers with hypertension during pregnancy.

The authors in Ref. [12] presented a prediction model of overweight and obesity in employees of a medical company in the Kingdom of Saudi

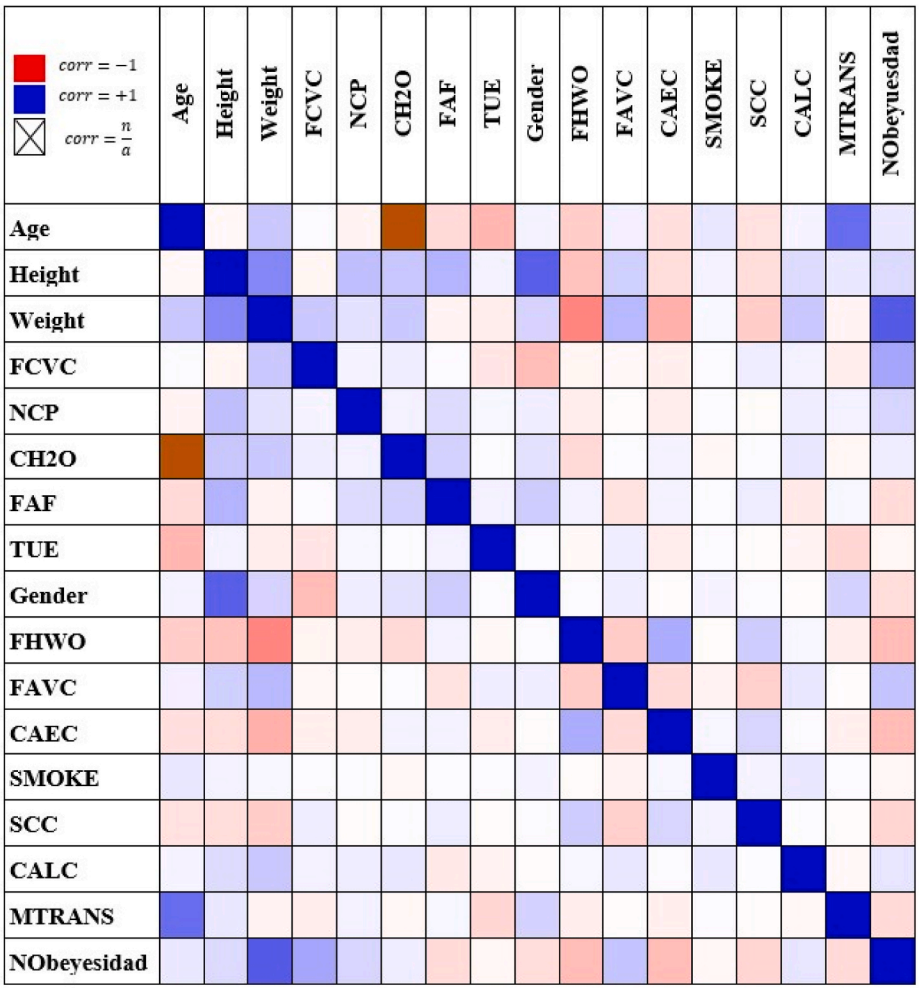


Fig. 2. Correlation matrix of attributes.

Table 2
Average performance evaluation of the techniques used.

Technique	Accuracy	AUC	F_1-score	Precision	Recall
Light GBM	97.45%	99.90%	97.43%	97.50%	97.44%
XG Boost	96.85%	99.86%	96.83%	96.86%	96.86%
Random forest	95.81%	99.78%	95.81%	95.67%	95.81%
Decision tree	93.00%	95.92%	92.95%	93.01%	93.00%
extremely random trees	94.62%	99.62%	94.62%	94.83%	94.61%
Logistic regression	83.02%	97.28%	82.63%	82.77%	82.96%

Arabia using logistic regression (LR). The data used was the product of a survey of 505 employees of the chemical company Sadara in 2019, the participants were between 18 and 62 years old. The data that was collected is related to sociodemographic information, lifestyle habits and physical characteristics. The reason for the study on obesity in employees is because workers with obesity, overweight or some non-communicable disease (NCD) can reduce productivity by up to 60% [12]. The logistic regression (LR) classification technique was trained. The authors do not mention the steps of data preprocessing, the grouping of the data by type of obesity shows an imbalance, since only 3.4% of the data were from people with insufficient weight and the authors did not mention if they applied any balancing technique of data.

The authors in Ref. [13] presented a prediction model for obesity in early childhood using machine learning. The data used were the electronic medical records of 860,510 patients under 2 years of age with 11, 194,579 medical consultations at the Children's Hospital of Philadelphia. Seven machine learning techniques were trained: XG Boost, Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Neural Networks (NN), Gaussian Naive Bayes (GNB) and Bernoulli Naive Bayes (BNB); to predict the incidence of obesity as defined by the Philadelphia Centers for Disease Control and Prevention. The performance of the models was evaluated using multiple classification metrics: AUC, precision, F1-score, accuracy, and specificity, and the differences between seven models were compared using Cochran's Q test and post-hoc pairwise test. The results obtained reflect that the XG Boost technique classifies the research problem better, since it reached an AUC of 0.81 and an accuracy of 64.14%. The XG Boost technique proved to have a good performance when classifying the incidence of obesity, therefore, it will be one of the classification techniques that will be trained in this research.

This research tried to cover some gaps that were evidenced in previous works by other authors, focusing on applying good practices in data preprocessing, training new learning techniques and comparing the results with the metrics of some classification techniques that are traditionally used in machine learning. The results of this research could support public health management, evaluating groups at high risk of obesity and developing or providing effective interventions in the early stages of the disease.

3. Methodology

The author used 80% of the data set for training, and the remaining

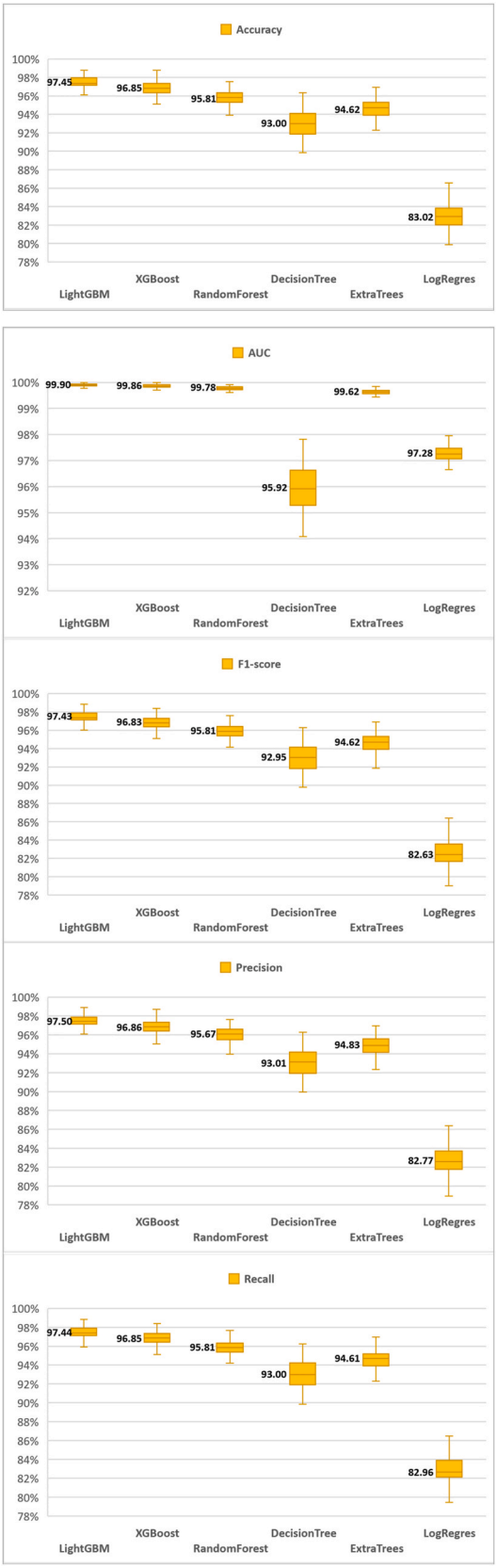


Fig. 3. Comparative graphs of the results.

20% was used for testing the algorithms. Section IV shows the techniques applied in data preprocessing. The following data mining techniques were used: Light Gradient Boosting Machine (Light GBM), Extreme Gradient Booting (XG Boost), Random Forest (RF), Decision Tree (DT), Extremely Randomized Trees (ET) and Logistic Regression (LR). The following metrics were evaluated: accuracy, AUC, F_1-score, precision and recall, for each algorithm used. Based on these metrics, the algorithm that had the best performance when classifying the data was determined. The training was performed 100 times. And the results of the 5 metrics were averaged to determine the model with the best performance.

3.1. Dataset

We used the dataset published by Mendoza et al. [14], which has 16 factors that can determine whether a person is obese. To collect the information, they presented the variables as questions through a survey, which they applied to a group of people from the countries of Colombia, Mexico and Peru. In the selected dataset, it was observed that 2,111 people participated in the study, with ages between 14 and 61 years.

3.2. Light Gradient Boosting Machine (light GBM)

"The Light GBM algorithm stands for Light Gradient Boosting Machine. Gradient boosting machines construct sequential decision trees. Each tree will be built based on the error of the previous tree. Finally, the predictions will be made by the sum of all those trees. Light GBM applies tree growth in terms of leaves or nodes. Light GBM was created by Microsoft and has been adopted by many machine learning studios due to its speed and performance, currently it runs 6 times faster than the XGBoost algorithm." [15].

3.3. Extreme gradient boosting (XG boost)

Extreme gradient boosting or also known as XGBoost "is an algorithm used in machine learning to solve classification problems, it is an implementation of gradient-driven decision trees designed for speed and performance. XGBoost is a software library created by Tianqi Chen which offers the implementation of the model and supports three main forms of gradient augmentation: gradient augmentation algorithm, stochastic gradient augmentation, and regularized gradient augmentation. XGBoost applies tree growth level-wise." [16].

3.4. Decision tree (DT)

"Decision trees are related to the theory of probability, specifically to the Bayesian decision process. A decision tree is so named because it resembles a tree, although for convenience it is horizontal. The base of the tree is the initial decision point. The actions available to the decision maker are expressed in chronological order and can be of two types: Decision points or decision points with uncertainty" [17]. "In each division we want to find a characteristic that divides the labeled data in such a way that the secondary nodes are more homogeneous than the parent node they came from. In other words, the nodes become purer as we move through the tree" [18].

3.5. Random forest (RF)

Random Forest is a type of tree-based prediction model. A type of tree-based prediction model is known as Random Forest. He does not use a single tree but a group of them. These can be CART or CI type. Its development was driven by the search for a decrease in variability and the consequent increase in the reliability of the model and its robustness" [19]. "The term random is derived from the fact that we take random samples from the training set, and since we have a collection of trees, it is natural to call it forest, hence random forest. To create the root

Table 3
Comparison of the results with previous research.

Autors	Best technique	Accuracy	AUC	F_1-score	Precision	Recall
This work	LightGBM	97.45%	99.89%	97.43%	97.46%	97.44%
[6]	DT + Simple k-Means	–	99.5%	98.5%	98.5%	98.5%
[7]	Logistic regression	97.09%	100%	97%	97%	97%
[8]	Decision tree	–	–	–	97.4%	97.8%
[9]	Gradient Boosting trees	–	80.3%	–	.	31.2%
[22]	Decision tree	85%	–	–	–	–
[23]	Naïve Bayes	75%	–	–	–	–

node or any node in the tree, a random subset of entities is selected. For each of these selected features, the algorithm searches for the optimal cut-off point to determine the division of the given function. The characteristic of the randomly selected subset that produces the purest division is used to create the root node". [20].

3.6. Extremely Randomized Trees (ET)

"Extremely random trees or also known as extra trees is a machine learning algorithm used to solve classification problems, based on nodes and associations, seeking to determine the division of a main set of data into subsets of similar characteristics randomly selected. Since the divisions are chosen randomly for each feature in the additional tree classifier, it is less computationally expensive than a Random Forest" [20].

3.7. Regresión logística (LR)

"Logistic Regression is a machine learning predictive analytics algorithm used for classification problems based on the concept of probability. It uses a cost function known as a sigmoid function or logistic function (Equation (1)) to map predictions to probabilities. The logistic regression hypothesis tends to limit the cost function between the values of 0 and 1" [21].

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{1}$$

Equation (1). Formula of a sigmoid function.

3.8. Accuracy

Accuracy is defined as the percentage of the total sample that was correctly classified. It can be measured by the following equation:

$$Accuracy = \frac{MCC}{ME} * 100\% \tag{2}$$

Equation (2). Formula of accuracy.Where:

MCC, it is the total of the sample that was correctly classified.
ME, it is the total of the sample that was evaluated.

3.9. Area under the curve ROC (AUC)

The area under the ROC curve (AUC), is the measure of the area that forms a probability curve that plots the ratio of true positives (TPR) or also called sensitivity against the ratio of false positives (FPR) at various threshold values, this curve is known as the ROC. Area values range from 0 to 1, the higher the AUC, the better the model will perform in distinguishing between positive and negative classes.

$$Sensitivity\ (TPR) = \frac{VP}{VP + FN} * 100\% \tag{3}$$

Equation (3). Formula of sensitivity

$$Specificity = \frac{VN}{FP + FN} * 100\% \tag{4}$$

Equation (4). Formula of specificity

$$FPR = 1 - Specificity \tag{5}$$

Equation (5). Formula of false positive ratio.Where:

VP, is the total of true positives
VN, is the total of true negatives
FP, is the total false positives
FN, is the total of false negatives

3.10. Precision

Precision is defined as the percentage of the total positive sample that was classified positively, it can be measured according to the following equation:

$$Precision = \frac{VP}{VP + FP} * 100\% \tag{6}$$

Equation (6). Formula of precision.Where:

VP, is the total of true positives
FP, is the total false positives

3.11. Recall

Recall is understood as the percentage of the total positive sample that was correctly classified as positive.

$$Recall = \frac{VP}{VP + FN} * 100\% \tag{7}$$

Equation (7). Formula of recall.

Where:
VP, is the total of true positives
FP, is the total false positives
FN, is the total of false negatives

3.12. F_1-score

F_1-score is the measure of the harmonic mean of recall and precision.

$$F_1 score = \frac{2 * P * R}{P + R} * 100\% \tag{8}$$

Equation (8). Formula of F_1-score.Where:

P, is the precision
R, is the recall

4. Experimentation

This study proposes a method of artificial intelligence applying machine learning techniques to estimate the level of obesity of people. To achieve this, the following steps were followed.

4.1. Data preprocessing

Table 1 shows the description of the data used for the determination of the mathematical model.

To balance the data, the MaxAbsScaler function was applied to scale each characteristic by its maximum absolute value and the MeanImputer imputation method to replace the missing values with the average value of the characteristic. After applying the data preprocessing, the distribution of the records is shown in Fig. 1.

4.2. Correlation matrix

Fig. 2 shows the correlation matrix of the data used for the classification; the categorical data were transformed to numerical, with an initial value of 0, increment of 1 and a maximum value of 100 categories; then the linear correlation technique was applied including only pairs of compatible columns, with an upper limit of possible values of 50 and a p-value corresponding to the probability of obtaining the correlation at both ends of the Pearson correlation coefficient. The correlation range is from −1 (red color) to +1 (blue color) and it degrades based on the correlation level of each attribute.

4.3. Training

Microsoft Azure Machine Learning and Kaggle was used for training and testing, the classification models Light GBM, XG Boost, random forest (RF), decision tree (DT), Extreme Random Trees (ET) and logistic regression were tested, the hyperparameters used by each machine learning technique used are found in the appendix of the document Tabla 5, using the data to estimate obesity levels based on eating habits and physical condition in people from Colombia, Peru and Mexico [14]. 80% of the data will be used for model training and the remaining 20% will be used for testing. The models must determine the level of obesity based on 7 levels indicated by the WHO: underweight, normal weight, level I overweight, level II overweight, type I obesity, type II obesity and type III obesity. The training was performed 100 times. The results of the 5 metrics were average. In this way, a comparison will be made between the results, defining the model that best fits the data set.

5. Results

Table 2 shows the performance of each machine learning technique used in training, the performance was determined based on the average of the following metrics: accuracy, AUC, F1-score, precision and recall. The algorithm that best fits our problem was determined according to its performance.

As shown in Table 2 and Fig. 3, the LightGBM technique the highest accuracy, AUC, F_1-score, precision y recall. The performance results of the Logistic regression technique were the lowest of the 6 techniques evaluated. Considering the results, the best performance for the data model was obtained using the Light Gradient Boosting Machine technique.

6. Discussion

It was necessary to compare this work with similar investigations, after that, it was possible to evaluate the goodness of the proposed obesity risk prediction model. After following a good amount of research articles, we concluded that many works had been done to predict obesity, but it was difficult to find works that were relative to ours.

Table 3 compares our results with other investigations that address the same problem; we can say that the performance of our model is better than those proposed in other investigations. We collect data from people of different classes and ages. But it is true that our dataset is not a larger dataset like others, but we tried our best to find the best performance by applying various machine learning algorithms and got a satisfactory result.

7. Conclusion

Data mining is a discipline responsible for performing exploratory data analysis to identify patterns or behaviors in the information. In this research, a method based on computational intelligence is created, using supervised classification techniques that include Light gradient boosting machine (Light GBM), extreme gradient boosting (XG Boost), decision tree (DT), random forest (RF) and extremely random three (ET) and logistic regression (LR). Methods were compared across performance metrics: accuracy, AUC, F1-score, precision, and recall. The processes of preparation, data transformation to identify missing data, outliers, and correlation analysis; training and classification were performed using the Kaggle and Microsoft Azure Machine Learning data mining platform. Finally, the performance results obtained by the LightGBM technique (AUC 99.90%, accuracy 97.45%), surpassed the results obtained in previous studies such as [6] that obtained a 99.5% of AUC [7], that obtained a 97.09% of precision [8], that obtained a 97.4% of precision [22], that obtained a 85% of precision [23], that obtained a 75% de precision.

The use of the specialized Microsoft Azure Machine Learning and Kaggle platforms within this study allowed us to develop highly efficient, scalable and reusable classification models for estimating obesity.

The analysis of the level of obesity is a current need in society since it is present throughout the world and can affect people regardless of their age or sex. The results presented may be useful to analyze the relevance of methods based on computational intelligence to study different diseases, detect them adequately and minimize the impact on society.

For the present research the author used two of the most popular gradient magnification frames currently Light GBM and XG Boost. The model with the highest weighted value of AUC (0,9990 and 0,9986) was constructed using the XG Boost classifier and the Max absolute scaler standardization algorithm. In addition to the AUC, other metrics were calculated: accuracy, F1-score, precision, recall.

8. Limitations

The data set used in this research consists of 2,111 records, which was not balanced in the same proportion of data for the seven levels of obesity, had a smaller number of records of people with underweight and overweight level I; statistical tests typically require a larger sample size to ensure a representative distribution of groups of people to whom the results will generalize, In machine learning models, a greater amount of data leads to more reliable models, but as long as they are representative and of good quality; the MaxAbsScaler data transformation and scaling technique was used to minimize the impact of outliers or very large data on our data.

The results of this research could support public health management, evaluating groups at high risk of obesity and developing or providing effective interventions in the early stages of this disease.

9. Recommendations

It is recommended to align the obesity levels analyzed with the Body Mass Index Classification proposed by the World Health Organization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

I'm grateful for the collaboration of Dr. José Alfredo Herrera Quispe,

his effort and advice were key to achieving the expected results and the dissemination of these, through this work and to the Universidad Nacional Mayor de San Marcos for supporting this study.

Table 4
Dataset used in experimentation

N°	Gender	Age	Height	Weight	FHWO	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	Nobeyesdad
1	Female	21	1.62	64	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportation	Normal_Weight
2	Female	21	1.52	56	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public_Transportation	Normal_Weight
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
2110	Female	24.361936	1.73945	133.346641	yes	yes	3	3	Sometimes	no	2.852339	no	1.139107	0.586035	Sometimes	Public_Transportation	Obesity_Type_III

APPENDIX

An extract of the data used in the experimentation published by Mendoza et al. [14] is presented in Table 4 below.
Next, Tabla 5 presents the hyperparameters used by machine Learning techniques to train the models.

Table 5
Hyperparameters of the different machine learning techniques used in training.

Machine learning technique	Data transformation	Training algorithm

Light Gradient Boosting Machine	{
	“spec_class”: “preproc",
	“class_name”: “MaxAbsScaler",
	“module”: “sklearn.preprocessing",
	“param_args”: [],
	“param_kwargs”: {},
	“prepared_kwargs”: {}
	}
		{
		“spec_class”: “sklearn",
		“class_name”: “LightGBMClassifier",
		“module”: “automl.client.core.common.model_wrappers",
		“param_args”: [],
		“param_kwargs”: {
		“min_data_in_leaf”: 20
		},
		“prepared_kwargs”: {}
		}

Extreme Gradient Booting	{
	“class_name”: “StandardScaler",
	“module”: “sklearn.preprocessing",
	“param_args”: [],
	“param_kwargs”: {
	“with_mean”: false,
	“with_std”: false
	},
	“prepared_kwargs”: {},
	“spec_class”: “preproc"
	}
		{
		“class_name”: “XGBoostClassifier",
		“module”: “automl.client.core.common.model_wrappers",
		“param_args”: [],
		“param_kwargs”: {
		“booster”: “gbtree",
		“colsample_bytree”: 0.8,
		“eta”: 0.5,
		“gamma”: 0,
		“max_depth”: 5,
		“max_leaves”: 0,
		“n_estimators”: 100,
		“objective”: “reg:logistic",
		“reg_alpha”: 0,
		“reg_lambda”: 1.4583333333333335,
		“subsample”: 0.8,
		“tree_method”: “auto"
		},
		“prepared_kwargs”: {},
		“spec_class”: “sklearn"
		}

Decision Tree	{
	“class_name”: “StandardScaler",
	“module”: “sklearn.preprocessing",
	“param_args”: [],
	“param_kwargs”: {
	“with_mean”: false,
	“with_std”: false
	},
	“prepared_kwargs”: {},
	“spec_class”: “preproc"
	}
		{
		“class_name”: “DecisionTreeClassifier",
		“module”: “sklearn.tree",
		“param_args”: [],
		“param_kwargs”: {
		“class_weight”: “balanced",
		“criterion”: “gini",
		“max_features”: null,
		“min_samples_leaf”: 0.01,
		“min_samples_split”: 0.01,
		“splitter”: “best"
		},
		“prepared_kwargs”: {},
		“spec_class”: “sklearn"
		}

Random Forest	{ “class_name”: “StandardScaler", “module”: “sklearn.preprocessing",	{ “class_name”: “RandomForestClassifier", “module”: “sklearn.ensemble",

(*continued on next page*)

Table 5 (*continued*)

Machine learning technique	Data transformation	Training algorithm

	"param_args": [],
	"param_kwargs": {
	"with_mean": false,
	"with_std": false
	},
	"prepared_kwargs": {},
	"spec_class": "preproc"
	}
		"param_args": [],
		"param_kwargs": {},
		"prepared_kwargs": {},
		"spec_class": "sklearn"
		}

Logistic Regression	{
	"class_name": "MaxAbsScaler",
	"module": "sklearn.preprocessing",
	"param_args": [],
	"param_kwargs": {},
	"prepared_kwargs": {},
	"spec_class": "preproc"
	}
		{
		"class_name": "LogisticRegression",
		"module": "sklearn.linear_model",
		"param_args": [],
		"param_kwargs": {
		"C": 4714.8663634573895,
		"class_weight": "balanced",
		"multi_class": "multinomial",
		"penalty": "l2″,
		"solver": "lbfgs"
		},
		"prepared_kwargs": {},
		"spec_class": "sklearn"
		}

Extreme Random Trees	{
	"class_name": "StandardScaler",
	"module": "sklearn.preprocessing",
	"param_args": [],
	"param_kwargs": {
	"with_mean": false,
	"with_std": true
	},
	"prepared_kwargs": {},
	"spec_class": "preproc"
	}
		{
		"class_name": "ExtraTreesClassifier",
		"module": "sklearn.ensemble",
		"param_args": [],
		"param_kwargs": {
		"bootstrap": false,
		"class_weight": "balanced",
		"criterion": "gini",
		"max_features": null,
		"min_samples_leaf": 0.01,
		"min_samples_split": 0.056842105263157895,
		"n_estimators": 200,
		"oob_score": false
		},
		"prepared_kwargs": {},
		"spec_class": "sklearn"
		}

Referencias

[1] Pajuelo Ramírez J, Torres Aparcana L, Agüero Zamora R, Bernui Leo y I. El sobrepeso, la obesidad y la obesidad abdominal en la población adulta del Perú. An Fac Med 2019;80(1):21–7.

[2] W Elffers T, de Mutsert R, J Lamb H, de Roo A, van Dijk KW, R Rosendaal F, Wouter Jukema J, Trompet y S. Body fat distribution, in particular visceral fat, is associated with cardiometabolic risk factors in obese women. *PLoS One*; 2017.

[3] World health Organization, «Body mass index calculator,» *WHO*, [En línea]. Available: http://www.emro.who.int/nutrition/information-resources/bmi-calculator.html. [Último acceso: 8 Octubre 2021].

[4] Centros para el Control y la, de Enfermedades Prevención. «Calculadora del IMC para adultos: sistema inglés,» CDC, 16 Febrero [En línea]. Available: https://www.cdc.gov/healthyweight/spanish/assessing/bmi/adult_bmi/english_bmi_calculator/bmi_calculator.html. [Accessed 8 October 2021].

[5] Safaei M, Sundararajan EA, Boulila W, Shapi'i y A. A systematic literature review on obesity: understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. Comput Biol Med 2021;136(104754):1–17.

[6] Martinez Palacio R Cañas Cervantes y U. Estimation of obesity levels based on computational intelligence. Informatics in Medicine Unlocked 2020;21(100472).

[7] Ferdowsy F, Samsul K, Jabiull y I. A machine learning approach for obesity risk prediction. Current Research in Behavioral Sciences 2021;2(100053).

[8] De la Hoz Correa E, Morales Ortega R, Mendoza Palechor F, De la Hoz Manotas A, Sánchez Hernandez y B. Obesity level estimation software based on decision trees. J Comput Sci 2019;15(10):67–77.

[9] Rossman H, Shilo S, Barbash-Hazan S, Shalom Artzi N, Hadar E, Balicer RD, Feldman B, Segal A Wiznitzer y E. Prediction of childhood obesity from nationwide health records. J Pediatr 2021;233:132–40.

[10] Sun Y, Wang S, Sun y X. Estimating neighbourhood-level prevalence of adult obesity by socio-economic, behavioural and built environment factors in New York City. *Public Health*; 2020. p. 57–62.

[11] Lee I, Bang K-S, Joint HM, Kim y J. Risk factors for obesity among children aged 24 to 80 months in korea: a decision tree analysis. J Pediatr Nurs 2019;46:15–23.

[12] Jaoua N, Woodman A, Withers y M. Predictors of overweight and obesity among employees of Sadara chemical company in the kingdom of Saudi Arabia. Obesity Medicine 2020;18:1–6.

[13] Pang X, Forrest C, Lê-Scherban F, Masino y A. Prediction of early childhood obesity with machine learning and electronic health record data. Int J Med Inf 2021;150:1–8.

[14] Mendoza Palechor y F, De la Hoz Manotas A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, vol. 25. *ELSEVIER*; 2019. 104344.

[15] Likin S. Sefiks [En línea]. Available: https://sefiks.com/2018/10/13/a-gentle-introduction-to-lightgbm-for-applied-machine-learning/. [Accessed 2 October 2021].

[16] Browniee J. Machine learning mastery,. 17 Agosto 2016 [En línea]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/. [Accessed 4 November 2021].

[17] Fabela Rodriguez Md lL. Toma de decisiones en Administración. Monterrey: UANL; 1995.

[18] Jeffares A. Towards data science [En línea]. Available: https://towardsdatascience.com/decision-trees-60707f06e836. [Accessed 2 November 2021].

[19] Martín BB. Predicción semanal de precios de la energía eléctrica utilizando bosques aleatorios. Madrid: UPM; 2017.

[20] Ceballos F. Towards data science [En línea]. Available: https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b. [Accessed 9 October 2021].

[21] Pant A. Towards data science [En línea]. Available: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148. [Accessed 12 October 2021].

[22] Dugan TM, Carroll y A, Downs SM. Machine learning techniques for prediction of early childhood obesity. Appl Clin Inf 2015;6:506–20.

[23] Bin Muhamad Adnan MH, Rashid W Husain y NA. A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction. ICCIS 2012:281–5.