Predictive Performance of Machine Learning Algorithms Regarding Obesity Levels Based on Physical Activity and Nutritional Habits: A Comprehensive Analysis

Paulo Henrique Ponte de Lucena, Lídio Mauro Lima de Campos, and Jonathan Cris Pinheiro Garcia

***Abstract*—Obesity is a complex chronic disease resulting from the interaction of multiple behavioral factors. This paper presents the application of Machine Learning to identify the primary groups of behaviors contributing to the development of obesity. Supervised machine learning emphasizes decision trees and deep artificial neural networks from datasets. The study also references related work that utilizes predictive methods to estimate obesity levels based on physical activity and dietary habits. Furthermore, it compares the performance of classification algorithms such as J48, Naive Bayes, Multiclass Classification, Multilayer Perceptron, KNN, and decision trees when predicting diabetes cases. The objective is to analyze different tools in the assessment based on physical activity and dietary habits, contributing to the improvement of obesity risk diagnosis. In addition, MLP and J48 demonstrated strong performance among all the algorithms, but BPTT achieved the highest overall performance.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8829

***Index Terms*—Artificial Neural Networks, Obesity, Machine Learning.**

I. Introduction

Obesity is a complex disease marked by excessive body fat that can harm health, defined by a BMI (Body Mass Index) of 30% or higher [1]. Obesity increases the risk of chronic and cardiovascular diseases, numerous types of cancer, musculoskeletal disorders, metabolic syndrome, diabetes and kidney disease. Furthermore, it triggers inflammatory processes and produces inverse vascular changes, such as arterial stiffness [2], [3].

According to data from the World Health Organization (WHO), many people over 18 suffer from obesity caused by several factors associated to food intake with high caloric content, sedentary lifestyle, and transportation habits [4].

Currently, scientists and health professionals are increasingly interested in recording and analyzing extensive datasets order to obtain more in-depth knowledge and understanding of this problem, with the goal of preventing, diagnosing, monitoring, and curing obesity more effectively [5].

P. H. Ponte de Lucena, L. M. Lima de Campos, and J. C. Pinheiro Garcia are with Federal University of Pará, Belém - PA, Pará, Brazil (emails: paulohlucena11@gmail.com, limadecampos@gmail.com, and jonathancris00@gmail.com).

Approaches that use Machine Learning (ML) techniques can gather data from diverse groups of individuals to provide personalized predictions. These techniques can be used to develop obesity risk categories and guidelines for directing public policies. More specifically, ML can offer a personalized treatment method for each patient, according to their specific characteristics. By using ML-based methods, it is possible to estimate future obesity risk and relevant information for governments. Furthermore, it can be very challenging for a nutritionist to analyze 19 variables and rank a person's obesity level. However, by using deep learning, the ability to make accurate predictions is greatly enhanced, thus assisting professionals in their daily tasks who utilize this technique.

This study aims at implementing ML methods to determine whether a person suffers from obesity. The performance of ML methods largely depends on the dataset and training algorithms, whereas choosing the right training algorithm can improve models performance, although, some algorithms that perform well in some datasets, might fail in others.

Approaches that use ML techniques, can combine data from diverse groups of individuals in order to provide personalized predictions [6] [7].

This article is structured as follows. Section II discusses related work, section III presents the background of the study, Section IV presents Material and Methods. Lastly, simulation results and conclusions are presented in Sections V and VI.

II. Related Work

In the study [8] , the authors employed ML methods to predict obesity levels in children. They utilized the SEMMA data mining approach for dataset selection and modeling, whereas three techniques were applied: Decision Trees (J48), Bayesian Networks (Naïve Bayes), and Logistic Regression (Simple Logistics). The J48 method yielded the highest accuracy rate (97.4%).

The purpose [9] of this article is to move towards a machine-learning-based pathway for predicting the risk of obesity using machine-learning algorithms. It collects more than 1100 data from various types of people of different ages and collects information both from the obese and the non-obese. For that, the authors used the algorithm of k-nearest neighbor (k-NN), random forest, logistic regression, multilayer perceptron

(MLP), support vector machine (SVM), naïve Bayes, adaptive boosting (ADA boosting), decision tree, and gradient boosting classifier. It establishes high, medium and low levels of obesity from experimental results. The accuracy came out from logistic regression with a value of 97.09%.

This work [10] introduced a hybrid approach for obesity prediction utilizing ensemble ML techniques that combined random forest, generalized linear model, and partial least squares methods whereas this hybrid model attained an accuracy of 89.68%. Later on, in a subsequent study, researchers suggested incorporating more than three algorithms for enhancing the ensemble-based hybrid approach.

This research [11] aimed to predict the level of obesity based on physical activities and eating habits using the trained neural network model whereas Chi-square, F-Classify, and classification algorithms were used to identify the most critical factors associated with obesity. The performance of the models was compared using a neural network trained with different resource sets. The results were able to predict the level of obesity with average accuracies of 93.06%, 89.04%, 90.32%, and 86.52% .

This study employed seven distinct ML algorithms on openly accessible datasets from the UCI ML repository. Accuracy levels of these algorithms were assessed before opting for hybrid approaches. The proposed hybrid model, which utilized a majority vote for obesity prediction and classification, achieved an accuracy of 97.16%, thus surpassing both individual models and other previously developed hybrid models [12].

The article [13] addresses a proposed predictive method for estimating obesity levels based on physical activity and nutritional habits. The study involves analyzing these factors as indicators to predict obesity in individuals whereas the authors explore the relationship between lifestyle and obesity aiming to provide an effective tool for assessing and addressing this health issue.

The research [14] aims to investigate how physical activity relates to weight status, which is assessed using Body Mass Index (BMI), and to evaluate the effectiveness and predictive capabilities of various machine learning and conventional statistical methods widely used. The dataset includes a total of 7162 participants who met the inclusion criteria (3682 males and 3480 females), with an average age ranging from 48.6 years (normal weight) to 52.1 years (overweight) whereas the classification algorithms employed in the study include logistic regression, Naïve Bayes, Radial Basis Function (RBF), local k-nearest neighbors (k-NN), classification via regression (CVR), random subspace, decision table, multiobjective evolutionary fuzzy classifier, random tree, J48, and multilayer perceptron. These algorithms were compared against estimates from a traditional logistic regression model and the random subspace algorithm showed a slight advantage over the other ten models, including logistic regression.

The aim of the research outlined in [15] is to create a model that can accurately predict the likelihood of diabetes in patients. To achieve this goal, the study utilizes three different ML classification algorithms: Decision Tree, SVM, and Naive Bayes. These algorithms are employed to detect diabetes at its early stages whereas the performance of each one is assessed using a range of metrics including Precision, Accuracy, F-Measure, and Recall. Precision measures the instances correctly classified versus those classified incorrectly. The findings indicate that Naive Bayes demonstrates superior performance compared to the other algorithms.

III. Background

ML is a dynamic field of computational techniques engineered to replicate human intelligence through environmental learning. ML techniques have been proven effective in diverse fields, thus encompassing pattern recognition, computer vision, spacecraft engineering, finance, entertainment, computational biology, and notably in medical and nutritional contexts [16].

Therefore, health is essential for any human being, which is why there are currently numerous technological advances and several attempts to provide health to citizens. ML is a promising area of research, and as such it has attracted many researchers to use it to solve different types of multi-criteria decision problems. Therefore, many researchers are using ML to predict critical illnesses [17], [18].

More specifically, ML can provide personalized treatment plans by identifying the unique traits of each patient. Furthermore, assessing the effectiveness and potential side effects of various treatments on an individual basis can help guide the choice of therapy and ongoing patient monitoring. Predicting the likelihood of future obesity using ML-based techniques can yield valuable insights and data on numerous personal variables [13].

IV. Material and Methods

This section describes the methodology applied to a dataset sourced from [5], containing 17 variables. The objective is to determine the obesity level of individuals across seven categories and assess the predictive accuracy of the model's outcomes. The study employs various architectures including deep neural networks (multilayer perceptron), J48, Naive Bayes, Multiclass Classification, and KNN.

A. *Dataset Description*

The dataset utilized in this study, sourced from [5], encompasses information regarding obesity levels among individuals from Mexico, Peru, and Colombia. Participants range in age from 14 to 61 and exhibit diverse dietary habits and physical conditions.

Data collection was facilitated through a web platform where anonymous users responded to a survey comprising various questions. This dataset comprises 2111 records, featuring 17 variables, with 16 serving as inputs and 1 as the output. Detailed descriptions of these variables are provided below.

- **Gender:** categorical variable according to the biological sex of each person(male or female).
- **Age:** numeric variable that uses the age, in years, of a person.
- **Height:** numeric variable that uses, in meters, the height of a person.

- **Weight:** numeric variable that uses the weight of a person, in kilograms.
- **Historic of overwheight family:** categorical variable that analyzes if a person has some case of overweight in the family.Answers can be: yes or no.
- **Consumption of high-calories foods (FAVC):** categorical variable that analyzes if a person often ingests high-calorie foods.Answers can be: yes or no.
- **Consumption of vegetables (FCVC):** categorical variable that analyzes how regularly the person ingests vegetables in meals.Answers can be: never, sometimes or always.
- **Amount of main meals (NCP):** numeric variable that analyzes the number of main meals made in a day.Answers can be between 1 and 2, 3 or more than 3.
- **Consumption of food between meals (CAEC):** categorical variable that analyzes how often person ingests food between meals during the day.Answers can be: no,sometimes, frequently or always.
- **Smoke:** categorical variable that analyzes if a person smokes or not.Answers can be: yes or no.
- **Consumption of water (CH2O):** numeric variable that analyzes the amount, in liters, of water intake during the day.
- **Monitor calories (SCC):** categorical variable that analyzes if a person monitors the amount of calorie intake during the day.Answers can be: yes or no.
- **Physical activity (FAF):** numeric variable that analyzes how often a person does physical activities.
- **Time using electronic devices (TUE):** numeric variable that analyzes how many hours per day the person uses electronic devices, like videogames, computers, TVs, cell phones, and others.
- **Consumption of alcohol (CALC):** categorical variable that analyzes how often a person drinks alcoholic beverages.Answers can be: i do not drink, sometimes, frequently or always.
- **Type of transportation used (MTRANS):** categorical variable that analyzes what type of transportation is most used by the person.Answers can be: automobile, motorbike, bike, public transportation or walking.
- **Obesity levels according BMI method (NObesity):** categorical variable that shows the obesity level according to the Index Body Mass Index (BMI) calculation.Answers can be: insufficient weight, normal weight, overweight level I, overweight level II, obesity type I. obesity type II or obesity type III.

B. Machine Learning Models

In this section, we describe the Neural Network's, J48, Naive Bayes, Multiclass Classification and KNN architectures that were used in the experiments.

*1) **Deep Feedforward NN**:* This model, characterized by a three-layer or more network of processed units connected by acyclic links, Fig. 1, is commonly employed in systems and biology studies, including evolutionary studies. Information flows through the ANN in discrete time.

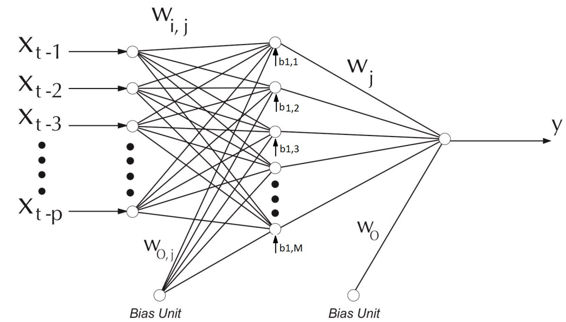

Fig. 1. Structure of a deep feedforward neural network.

The output o_j of node j is calculated by Eq. 1:

$$\vartheta(o_j) = \frac{1}{1 + exp^{-k.\sum_{i \epsilon I_j}(w_{ij}x_i + b_j)}} \tag{1}$$

Here, I_j is the set of nodes connected to node j, w_{ij} is the strength of the connection between node i and node j, o is the output value of node i, and b_j is the bias.

The parameter k measures the steepness of the sigmoidal function Eq. 1). As k is positive, the sigmoidal function is monotonically increasing, continuous and differentiable over the whole domain. In our experiments, we assigned the typical value for the training of neural networks with BP (namely, k = 1).

The parameters of the neural network w_{ij} are changed by an amount Δw_{ij}, which is calculated by Eq. 2.

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{2}$$

Where the parameter η is the learning rate and *E* is the error in the output layer. The δ term in Eq.3) is a momentum term, introduced by [19] to accelerate the learning process while avoiding instability in the algorithm.

$$\Delta w_{ij}(t+1) = -\eta \frac{\partial E}{\partial w_{ij}} + \delta \Delta w_{ij}(t) \tag{3}$$

*2) **J48**:* The book [20] provides a comprehensive description of the C4.5 algorithm, widely used in building decision trees in ML. Quinlan addresses various aspects of decision trees, from the construction process to pruning techniques and problem-solving such as handling missing values. While not covering competing algorithms, the book is a valuable tool for researchers and students in ML, assuming no prior knowledge on the subject. Quinlan provides practical and enlightening examples from real datasets, thus making the content accessible and informative.

Algorithms for decision tree construction are among the most well-known and widely used, with C4.5 being probably the most popular in the ML community. Here is a simplified description of the C4.5 algorithm:

Entropy Calculation: Entropy is a measure of the impurity of a dataset. For a dataset **S** with **k** classes, entropy **(H(S)))** is calculated as:

$$H(S) = -\sum_{i=1}^{k} p_i \cdot \log_2(p_i) \tag{4}$$

where p_i is the proportion of instances in **S** belonging to class **i**.

Information Gain Calculation:

The information gain is employed to ascertain the most suitable attribute for dividing the tree at each node. If A is an attribute with v possible values, and S is the current dataset, the information gain **(Gain(A))** is given by:

$$Gain(A) = H(S) - \sum_{v} \frac{|S_v|}{|S|} \cdot H(S_v) \tag{5}$$

where S_v is the subset of **S** for which the attribute **A** equals **v**

Stop Criterion:

The algorithm for tree construction continues recursively until some stopping criterion is reached. This may include the maximum depth of the tree, the minimum number of instances in a node, or other criteria.

Tree Pruning:

Pruning is a technique to prevent overfitting of the tree to the training data. The C4.5 algorithm uses pruning techniques to remove parts of the tree that are not statistically significant.

The formula above represents a simplified version of the C4.5 algorithm. It is worth noting that the actual implementation may involve more details and considerations, especially when dealing with specific features such as handling missing values, treatment of continuous attributes, etc.

*3) **Naive Bayes**:* According to [21], the Naive Bayes classifier offers a simple, semantically clear approach to representing, using, and learning probabilistic knowledge. This approach is tailored for supervised induction tasks, aiming to precisely forecast the class of test instances. In this scenario, training instances contain class information, and the objective is to achieve accurate predictions. Such a classifier can be considered as a specialized form of a Bayesian network, termed "naive" because it relies on two important simplifications. More specifically, it operates under the assumption that predictive attributes are independent of each other given the class, and it suggests that there are no hidden or latent attributes affecting the prediction process. This forms the basis of the Naive Bayes algorithm:

Bayes Theorem:

The Naive Bayes classifier leverages Bayes' Theorem to compute the conditional probability of a class **C** given a set of attributes$X = \{x_1, x_2, \ldots, x_n\}$. The theorem is expressed as follows:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{6}$$

Assumption of Conditional Independence:

The term "naive", as in Naive Bayes, refers to the presumption of conditional independence among predictive attributes when considering the class.. This is expressed as:

$$P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C) \tag{7}$$

Prediction:

To predict the class **C** of an instance with attributes **X**, The Naive Bayes classifier chooses the class that maximizes the probability $P(C|X)$. This can be expressed as:

$$\hat{C} = \arg\max_{C} P(C) \cdot P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C) \tag{8}$$

where $\hat{C}$ is a scheduled class.

Parameter Estimation:

The parameters $P(C)$ and $P(x_i|C)$ are estimated from the training set.

This formula represents a simplified version of the Naive Bayes algorithm whereas actual implementation may involve additional techniques such as smoothing (to handle zero probabilities), handling missing data, and domain-specific considerations.

*4) **Multiclass Classification**:* The primary objective within the expansive domain of ML is often to forecast an outcome utilizing the data at hand. This predictive endeavor is commonly referred to as a "classification problem" when the outcome entails distinct classes, whereas it is termed a "regression problem" when the outcome is a numerical measurement. In the realm of classification, the typical scenario entails only two classes, although situations may arise with more than two classes. In such cases, the objective shifts to "multi-class classification." [22]

Let's consider a multiclass problem with K classes, where $K > 2$. Suppose you have binary classifiers $C_1, C_2, \ldots, C_{K-1}$, each trained to distinguish between a specific class and the rest.

The prediction for a new instance is made by selecting the class for which the corresponding binary classifier has the highest confidence.

Below is a general formula that can be used for prediction in a multiclass system using binary classifiers:

$$haty = \arg\max_{i} C_i(x) \tag{9}$$

where $\hat{y}$ is a scheduled class, and $C_i(x)$ it is the confidence assigned by the binary classifier i.

*5) **K-nearest neighbours classifier (KNN)**:* The study [23] explores instance-based learning algorithms in their pioneering research published under the title "Instance-based learning algorithms" in the Machine Learning journal. Among these techniques is the K-nearest neighbors (KNN) method, which falls within the category of supervised ML algorithms within the instance-based learning framework.

KNN is an example of a classifier that uses the similarity between instances to make predictions, employing a nearest neighbors method, as discussed in sections X and Y of the article. Instances are compared by measuring the distance between vectors or data points arranged within Euclidean space. In classification problems, the prediction in KNN is determined by the class with the highest representation (mode). In regression tasks, the prediction outcome is determined by calculating the average of the K closest values.

The K-nearest neighbors (KNN) algorithm presents a simplified description of the mathematical formula of the KNN algorithm:

Distance Calculation

The distance between two instances A and B in Euclidean space can be calculated using different metrics, such as Euclidean distance. The general formula for the distance between two instances A and B is:

$$\text{Distance}(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{10}$$

where A_i and B_i are the Coordinates of the instances in the different attributes, and n is the number of attributes.

Classification Voting

In the case of classification problems, the prediction in KNN is determined by the class with the highest representation among the K nearest neighbors. If C_i represents the class of instance i, the prediction $\hat{y}$ for a new instance is given by:

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{K} \delta(C_i, C_j) \tag{11}$$

where δ is a Kronecker delta function that is 1 when $C_i = C_j$ and 0 otherwise.

Average for Regression:

In regression scenarios, the prediction outcome is established by computing the mean of the K nearest values. If Y_i represents the value of instance i, the prediction $\hat{y}$ for a new instance is given by:

$$\hat{y} = \frac{1}{K} \sum_{j=1}^{K} Y_j \tag{12}$$

These formulas represent a simplified version of the KNN algorithm. It's worth noting that the selection of the distance metric and the choice of the value for K are crucial parameters that significantly influence the algorithm's performance in real-world scenarios.

C. Advantages and Disadvantages of Algorithm.

The J48 algorithm has the merit of highly readable classification rules and high accuracy although is highly time-consuming. Naive Bayes is computationally efficient, because it handles high-dimensional data well, but it has a few limitations: it assumes independence between features, which may not hold in all scenarios. KNN on the other hand is easy to implement, as there is no training period and new data can be added at any time since it won't affect the model. One advantage of MLPs is their compatibility with all training software. However, MLPs have limitations regarding their architecture. They are less powerful than other topologies, such as BPTT. Another disadvantage of MLPs is the same required time to train plus a large number of iterations. BPTT has several advantages, such as being significantly faster for training RNNs, although it also has drawbacks, including the difficulty with local optima and the potential for vanishing or extrapolating gradients.

D. Methodology

The first phase involved gaining a thorough understanding of the domain, specifically the prediction of obesity levels based on physical activity and nutritional habits. This included defining the research problem, setting clear objectives, identifying relevant variables, and understanding the key factors that influence obesity. The aim was to establish a strong foundation for the subsequent phases by ensuring a comprehensive grasp of the context and requirements of the study.

The second phase, regards creating a target dataset: Focus on creating the target dataset and subsetting data samples or variables. It was described in the previous section

The third phase was data cleaning, pre-processing and Data transformation. An initial examination of the dataset was conducted to identify and handle any missing or null values. Fortunately, no missing data entries were found, thus ensuring the dataset's completeness. Categorical values "yes," "no,", "always", "never", "sometimes", "frequently", "i do not drink", "male", "female", "automobile", "motorbike", "bike", "public transportation" and "walking" were systematically converted into numerical values to facilitate the application of ML methods - for example: from "female" to 0 and "male" to 1, and the same logic to the other sequences presented previously-; the variables that underwent these changes were: Historic of overweight family, Consumption of high-calories foods (FAVC), Consumption of vegetables (FCVC), Consumption of food between meals (CAEC), Smoke, Monitor calories (SCC), Consumption of alcohol (CALC), Gender and Type of transportation used (MTRANS). Another treatment was in variables such as "Height", "Weight", "Consumption of water (CH2O)" and "Time using eletronic devices (TUE)" for using only two floating points, and in variable "Age" to use only the integer part. These transformations were crucial for enabling the algorithms to process the data effectively. The "Obesity levels according BMI method (NObesity)" categorical values were then normalized to a range between 0 and 1 - from "Insufficient_Weight" to 0.14, from "Normal_Weight" to 0.28, from "Overweight_Level_I" to 0.42, from "Overweight_Level_II" to 0.56, from "Obesity_Type_I" to 0.70, from "Obesity_Type_II" to 0.84, from "Obesity_Type_III" to 0.98 -. This step was essential to ensure compatibility with the sigmoid activation function used by the neural network, which operates within this range. As a result of these pre-processing steps, the dataset was refined to include 17 distinct attributes and a total of 2111 individual records. Data balancing techniques, as outlined by the researcher [5], were applied to ensure the dataset was evenly distributed across different categories. Figures 2 and 3 illustrate the impact of data balancing, comparing the unbalanced and balanced datasets visually.

The fourth phase involved selecting the most appropriate data mining method based on the specificities of the research problem. Classification was chosen as the optimal method due to its suitability for categorizing data into predefined classes. This decision was based on the need to classify individuals into different obesity levels based on their physical activity and nutritional habits.

The fifth phase was exploratory analysis and model selection: during this step, various machine learning algorithms were evaluated and selected for model building. The chosen algorithms included J48, Naive Bayes, Multiclass Classification, and K-nearest neighbors classifier. These algorithms were selected for their proven effectiveness in pattern recognition and classification tasks. The exploratory analysis involved testing these algorithms on the dataset to determine their performance and suitability for the research objectives.

The sixth phase was Data mining: in this phase, the selected machine learning algorithms were applied to the dataset to search for patterns and relationships. This involved training the models on the dataset and evaluating their performance in classifying individuals into different obesity levels. The goal was to identify significant patterns and insights that could be used to predict obesity based on physical activities and nutritional habits.

The final step was to interpret the patterns and models extracted during the data mining phase. This involved analyzing the results to understand the implications of the findings, such as identifying key predictors of obesity and understanding how different variables interact. The insights gained from this analysis were used to draw conclusions and make recommendations for predicting obesity levels and informing intervention strategies.

It's important to highlight that the performance evaluation technique adopted was K-Fold Cross Validation. This robust method estimates the error of the learning method in observations not used in training, meaning it assesses how the constructed model will perform on new data. This is valid only if we maintain the same joint probability of the explanatory variables and the response variable used during training. The k-fold Cross Validation involves dividing the dataset into k folds. For each fold, we estimate the method without including that specific fold and assessing the average error on the fold not used during training.

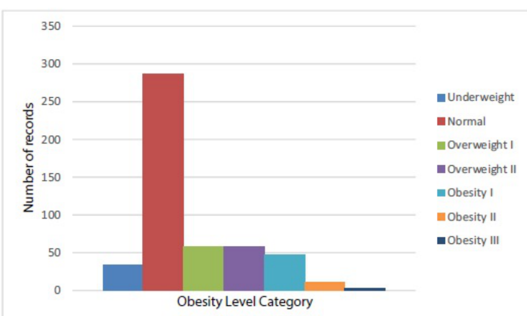

Fig. 2. Graphic with unbalanced dataset, Font: Palechor and Manotas [5].

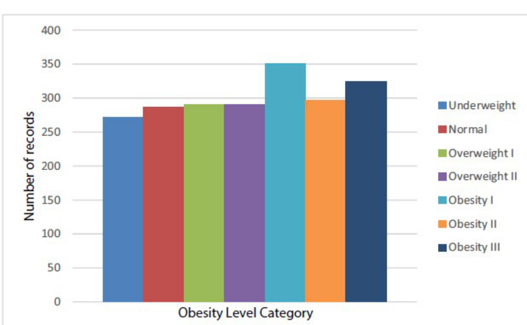

Fig. 3. Graphic with balanced dataset, Font: Palechor and Manotas [5].

The correlation matrix, depicted in Fig. 4, illustrates the relationships between all possible pairs of input variables.

So, it is possible to notice that the highest correlations are between Height and Gender (0.62), Family History of Overweight and Weight (0.5), and Weights and Height (0.46). Therefore, it is understood that, for the most part, the tallest heights are close to 1 (male) and the opposite close to 0 (female). Similarly, it is understood that for the majority of higher weights, they are close to 1 (family history), while for lower weights, they are close to 0 (no family history). There is also a relationship indicating that the taller the person, the higher the weights presented, and the opposite is true.

It is also noticeable that there are no very strong correlations (above 0.8), and most of them are in the blue range (below 0.2), indicating that the chance of overfitting is lower. There is no need to remove input variables for testing, confirming that the tests were conducted with all the variables mentioned in the dataset description subsection.

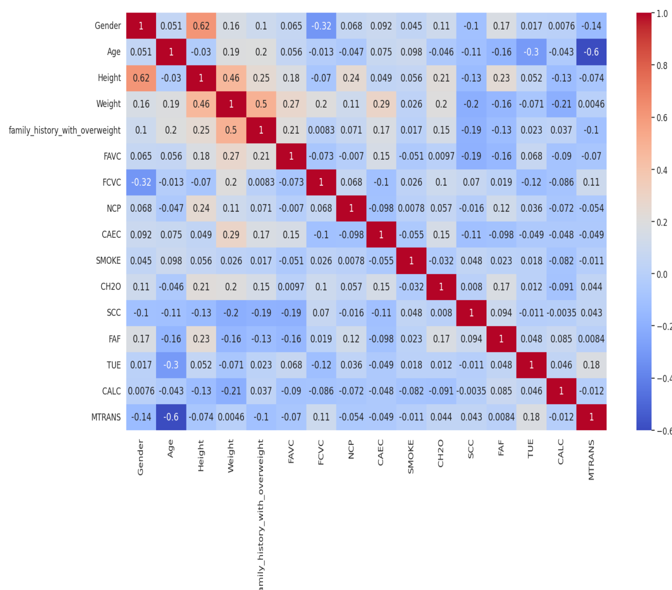

Fig. 4. Correlation matrix of all inputs in pair.

E. Metrics

This section describes others metrics, in addition to presented in section IV used in the experiments: Root mean squared error (RMSE) calculated by Eq. 13 and and Body mass index(BMI) calculated by Eq. 14.

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}\Big(p_i true - p_i forecast\Big)^2} \tag{13}$$

where: N is the number of samples, $p_i true$ is the actual price and $p_i forecast$ is the forecasted price.

$$BMI = \frac{Weight}{Height * Height} \tag{14}$$

where: Weight is the weight,in kg, and Height is the height, in meters, of person.

V. SIMULATION RESULTS

This section presents the comparison of different data mining algorithms, such as J48, Naive Bayes, K-nearest neighbors classifier (KNN), Multilayer Perceptron, and Multiclass classification, in classification obesity level.

A. *Multilayer Perceptron Neural Networks*

All tests used Sigmoid such intermediate activate function and output activate function. Furthermore, the optimizer used is SGD - Stochastic Gradient Descent -.

1) ***Multilayer perceptron neural networks with two intermediate layers****:* This section presents simulation results for multilayer perceptron neural networks with two intermediate layers using K-FOLD CROSS VALIDATION.

The Table I illustrates five favorable outcomes, each one surpassing a threshold of 90% accuracy. Nevertheless, only two instances exhibit accuracies of 95% or higher. In essence, merely 20% of the conducted tests yield such notably high results: specifically, ID 3 with a 99.96% accuracy and ID 9 with a 98.93% accuracy, both achieved after 90,000 training epochs. Notably, the network architecture comprises 10 neurons in the first intermediate layer, 6 in the second intermediate layer, learning rate 1 and learning rate 2 being 0.6 and 0.4, respectively. However, k = 2 is applied to ID 3 and ID 9 utilizes k = 5. The accuracy range spans from a minimum of 71.77% to a maximum of 99.96%, indicating a lack of regularity, with results fluctuating between 70% and 100%.

TABLE I
RESULTS WITH BALANCED DATA USING A MULTILAYER PERCEPTRON NEURAL NETWORKS WITH TWO INTERMEDIATE LAYER USING K-FOLD CROSS VALIDATION METHOD

ID	K	Depth	Neurons	Epoch	Learning Rate	Accuracy
1	2	2	5,3,1	100000	0.5, 0.2	82.43%
2	2	2	7,4,1	70000	0.4, 0.3	71.77%
3	2	2	10,6,1	90000	0.6, 0.4	99.96%
4	3	2	10,6,1	90000	0.6, 0.4	94.45%
5	3	2	8,5,1	50000	0.9, 0.7	87.26%
6	4	2	5,3,1	100000	0.5, 0.2	84.99%
7	4	2	10,6,1	90000	0.6, 0.4	90.96%
8	4	2	8,5,1	50000	0.9, 0.7	88.37%
9	5	2	10,6,1	90000	0.6, 0.4	98.93%
10	5	2	8,5,1	50000	0.9, 0.7	93.76%

2) ***Multilayer perceptron neural networks with one intermediate layer****:* This section presents simulation results for multilayer perceptron neural networks with one intermediate layer using K-FOLD CROSS VALIDATION.

The subsequent table (Table II) reveals highly favorable outcomes, with 9 out of 10 results exceeding the threshold of 95% accuracy. Put differently, 90% of all conducted tests yield outstanding results. The test falling below the 95% accuracy mark is represented by ID 6, attaining an accuracy of 92.72% after 50,000 training epochs, featuring 2 neurons in the intermediate layer, a learning rate set at 0.15, and utilizing k = 4. Despite not reaching the classification of an exceptional result, this test maintains a satisfactory accuracy, surpassing the 90% threshold. The range spans from a minimum of 92.72% to a maximum of 99.96%, revealing a noteworthy consistency in results, all of which surpass the 90% accuracy.

TABLE II
RESULTS WITH BALANCED DATA USING A MULTILAYER PERCEPTRON NEURAL NETWORKS WITH ONE INTERMEDIATE LAYER USING K-FOLD CROSS VALIDATION METHOD

ID	K	Depth	Neurons	Epoch	Learning Rate	Accuracy
1	2	1	2,1	50000	0.15	99.96%
2	2	1	6,1	20000	0.5	98.35%
3	2	1	4,1	40000	0.2	95.08%
4	3	1	2,1	50000	0.15	98.70%
5	3	1	4,1	40000	0.2	97.59%
6	4	1	2,1	50000	0.15	92.72%
7	4	1	8,1	30000	0.3	96.92%
8	4	1	6,1	20000	0.5	99.38%
9	5	1	2,1	50000	0.15	98.80%
10	5	1	8,1	30000	0.3	99.54%

3) ***Multilayer perceptron neural networks BPTT with one intermediate layer****:* This section presents simulation results for multilayer perceptron neural networks BPTT with one intermediate layer using K-FOLD CROSS VALIDATION.

The ensuing table (Table III) delineates exceptional results, where in all tests surpass the 95% accuracy threshold. The minimum accuracy observed is 98.30% in the case of ID 1, employing 8 neurons in the intermediate layer, a learning rate of 0.9, 30000 epochs, and k = 2. This test, while having a slightly lower accuracy, still hovers around 98%, indicating a very good regularity within the range of 98% to 100% accuracy. This consistency underscores the infrequent occurrence of suboptimal results in the deep neural network model under consideration.

Another noteworthy aspect of these tests is their ability to achieve high accuracy with a reduced number of epochs. For instance, IDs 2, 4, 7, and 10 exhibit accuracy levels above 99% while utilizing only 10,000 training epochs. This highlights the efficiency of the model in converging to highly accurate results with relatively less training epochs.

TABLE III
RESULTS WITH BALANCED DATA USING A MULTILAYER PERCEPTRON NEURAL NETWORKS BPTT WITH ONE INTERMEDIATE LAYER USING K-FOLD CROSS VALIDATION METHOD

ID	K	Depth	Neurons	Epoch	Learning Rate	Accuracy
1	2	1	8,1	30000	0.9	98.30%
2	2	1	10,1	10000	0.7	99.96%
3	2	1	12,1	90000	0.4	99.96%
4	3	1	10,1	10000	0.7	99.99%
5	3	1	12,1	90000	0.4	99.95%
6	4	1	8,1	30000	0.9	99.97%
7	4	1	10,1	10000	0.7	99.79%
8	4	1	12,1	90000	0.4	99.91%
9	5	1	8,1	30000	0.9	99.99%
10	5	1	10,1	10000	0.7	99.99%

4) ***General analysis of the results****:* The analysis reveals that the deep neural network employing two intermediate layers manifests inferior performance, characterized by relatively lower and moderate accuracies, the majority of which fall within the range of 70% to 90%. Only two instances surpass the 95% accuracy threshold. In contrast, the deep neural network with one intermediate layer exhibits significant improvement compared with the architecture of two intermediate layers. A noteworthy 90% of the tests yield remarkable results, with merely one instance falling slightly below 95% accuracy but still exceeding 90%. The values demonstrate a limited variance, residing between 90% and 99%, highlighting a notable enhancement over the two intermediate layer artificial neural network (ANN).

Furthermore, the superior performance is observed in the deep neural network employing Backpropagation Through Time (BPTT) with one intermediate layer. This configuration attains exceptional outcomes, with all results surpassing the 95% accuracy mark. The accuracy spectrum extends impressively from 98% to 99%, showcasing a remarkable consistency in the achieved results. Moreover, this ANN demonstrates efficacy in operating with a reduced number of epochs and exhibits satisfactory outcomes with well-tuned hyperparameters.

TABLE IV
RESULTS WITH BALANCED DATA USING J48, NAIVE BAYES, KNN AND MULTICLASS CLASSIFICATION ALGORITHMS USING K-FOLD CROSS VALIDATION METHOD

Algorithm	K	Accuracy	Error Rate	Mean Accuracy	Mean Error Rate
	4	93%	7.01%		
	5	93.23%	6.77%		
	6	94.08%	5.92%		
J48	7	93%	6.53%	93%	7.01%
	8	93.94%	6.06%		
	9	94.60%	5.40%		
	10	93.75%	6.25%		
	4	67%	33.11%		
	5	68.02%	31.98%		
	6	67.17%	32.83%		
NaiveBayes	7	68%	32.31%	67%	33.11%
	8	67.46%	32.54%		
	9	67.41%	32.59%		
	10	67.41%	32.59%		
	4	81%	18.85%		
	5	81.05%	18.95%		
	6	82.19%	17.81%		
KNN	7	83%	16.63%	81%	18.85%
	8	82%	18.05%		
	9	82%	18.10%		
	10	82%	17.95%		
	4	78%	22%		
	5	78%	22%		
	6	78%	22%		
Multiclass Classification	7	78%	22%	78%	22.03%
	8	78%	22%		
	9	78%	22%		
	10	78%	22%		

B. J48, Naive Bayes, K-nearest neighbors classifier and Multiclass classification.

The results presented in Table IV reveal a comparison among different prediction algorithms, focusing on percentage accuracy and error. The J48 algorithm stands out by displaying a notable difference in accuracy compared to others. J48 achieved an accuracy rate of 93%, while other algorithms recorded a maximum of 81%, 78%, and 67%. In comparison, the work [13] addresses the estimation of obesity levels based on physical activity and nutritional habits. The study, which is focused on the relationship between lifestyle and obesity, proposes an effective tool to assess and deal with this health issue. In the study [15], the aim is to develop an accurate model for predicting the probability of diabetes in early stages. Three ML algorithms: Decision Tree, SVM, and Naive Bayes, are compared in terms of metrics such as Precision, Accuracy, F-measure, and Recall. Naive Bayes was the best - with 76.30% compared to other algorithms.

C. Statistical Comparison of the Algorithms.

In addition, in order to evaluate the significance of the results obtained by the algorithms, we statistically compared the performance of the algorithms (considering the accuracy of each one), by using Welch's t-test [24] as the standard strategy, with 95% confidence, and some assumptions were adopted: continuous measurement, independence of results, homoscedasticity and normal distribution. Additionally, Bootstrapping was used to perform the test when the data did not have a normal distribution. When it comes to deciding whether the two performances differ from each other, we test the significance of the difference between u1 and u2 ($p < 0.05$). When there were no significant statistical differences between the accuracy values of two algorithms on a given dataset, we considered that both algorithms performed equally well (Ho: u1=u2) and awarded 1 point to each algorithm, in this case, we accepted the Hypothesis Ho. On the other hand, if two algorithms obtain significantly different accuracy (H1: u1 $\neq$ u2), the better performing algorithm is awarded 2 points and the other zero points. Consequently, we reject the null hypothesis Ho and accept the alternative hypothesis H1. The overall performance of each algorithm is then calculated by adding all points achieved in the pairwise comparisons. For example, considering the accuracy values of the BPTT algorithm, illustrated in Table III, and the accuracies of the KNN algorithm, detailed in Table IV. The two algorithms obtain significantly different accuracy, u1=99.70286 $\neq$ u2=81.89143, and p-value = 2.997e-15 < 0.05 (test-t with 95% confidence). Considering the normality criterion, the Shapiro-Wilk test indicated, W = 0.82545, p-value = 0.09832 > 0.05, and normal distribution. Thus the BPTT is better than the KNN algorithm. Table V illustrates the statistical comparison between the algorithms and the sum of points obtained by each algorithm.

TABLE V
SCORING EACH ALGORITHM BY COMPARING THE STATISTICAL METHOD WELCH'S T-TEST

	MLP	BPTT	J48	NB	KNN	MC	Total
MLP	-	0	1	2	2	2	7
BPTT	2	-	2	2	2	2	10
J48	1	0	-	2	2	2	7
NB	0	0	0	-	0	0	0
KNN	0	0	0	2	-	2	4
MC	0	0	0	2	0	-	2

VI. CONCLUSIONS

This research sought to compare the performance of classification algorithms such as BPTT, J48, Naive Bayes, Multiclass Classification, Multilayer Perceptron, KNN, and decision trees regarding predicting obesity levels cases. Table V statistically compares the algorithms' performance by using the approach mentioned in Section V.C. BPTT obtained the best overall performance. The demonstrated average accuracies surpassed 98%, showing excellent results. The MLP and J48 algorithms showed good performance compared to the others. However, the j48 has the disadvantage of relatively higher time consumption, compared to MLP, which is the best. NB was the algorithm that presented the worst performance, It assumes independence between features, which may not hold up in all scenarios in predicting obesity levels cases.

ACKNOWLEDGMENTS

We would like to thank the Federal University of Pará for having supported this study.

REFERENCES

[1] J. F. López-Gil, S. M. Wu, T.-L. I. Lee, C.-W. Shih, S. Tausi, V. Sosene, P. P. Maani, M. Tupulaga, Y.-T. Hsu, C.-R. Chang, S.-C. Shiau, Y.-H. Lo, C.-F. Wei, P.-J. Lin, and M. S. Hershey, “Higher imported food patterns are associated with obesity and severe obesity in tuvalu: A latent class analysis,” *Current Developments in Nutrition*, p. 102080, 1 2024, doi: 10.1016/j.cdnut.2024.102080.

[2] A. S. Kelly, S. E. Barlow, G. Rao, T. H. Inge, L. L. Hayman, J. Steinberger, E. M. Urbina, L. J. Ewing, and S. R. Daniels, “Severe obesity in children and adolescents: Identification, associated health risks, and treatment approaches: A scientific statement from the american heart association.,” *Circulation*, vol. 128, pp. 1689–1712, 10 2013, doi: 10.1161/CIR.0b013e3182a5cfb3.

[3] Collaborators, “Health effects of overweight and obesity in 195 countries over 25 years,” *New England Journal of Medicine*, vol. 377, pp. 13–27, 7 2017, doi: 10.1056/NEJMoa1614362.

[4] R. C. Cervantes and U. M. Palacio, “Estimation of obesity levels based on computational intelligence,” *Informatics in Medicine Unlocked*, vol. 21, p. 100472, 2020, doi: 10.1016/j.imu.2020.100472.

[5] F. M. Palechor and A. de la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico,” *Data in Brief*, vol. 25, p. 104344, 8 2019, doi: 10.1016/j.dib.2019.104344.

[6] W. Osama, K. R. N, and N. Ghada, “Machine learning for predicting deposits bank market shares in emerging markets: Evidence from egypt,” *REVISTA GESTÃO & TECNOLOGIA*, pp. 473–492, 2023, doi: 10.20397/2177-6652/2023.v23i4.2729.

[7] M. H. L. Cunha and H. S. Lara, “Machine learning as a way to have more accuracy when defining milk quality classification,” Seven Congress, 8 2023, doi: 10.56238/homeinternationalanais-069.

[8] E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. De-La-Hoz-Manotas, R. C. Morales-Ortega, and S. H. B. Adriana, “Obesity level estimation software based on decision trees,” *Journal of Computer Science*, vol. 15, pp. 67–77, 1 2019, doi: 10.3844/jcssp.2019.67.77.

[9] F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah, and M. T. Habib, “A machine learning approach for obesity risk prediction,” *Current Research in Behavioral Sciences*, vol. 2, p. 100053, 11 2021, doi: 10.1016/j.crbeha.2021.100053.

[10] K. Jindal, N. Baliyan, and P. S. Rana, *Obesity Prediction Using Ensemble Machine Learning Approaches*, pp. 355–362. 2018, doi: 10.1007/978-981-10-8636-6_37.

[11] F. H. Yagin, M. Gülü, Y. Gormez, A. Castañeda-Babarro, C. Colak, G. Greco, F. Fischetti, and S. Cataldi, “Estimation of obesity levels with a trained neural network approach optimized by the bayesian technique,” *Applied Sciences*, vol. 13, p. 3875, 3 2023, doi: 10.3390/app13063875.

[12] D. D. Solomon, S. Khan, S. Garg, G. Gupta, A. Almjally, B. I. Alabduallah, H. S. Alsagri, M. M. Ibrahim, and A. M. A. Abdallah, “Hybrid majority voting: Prediction and classification model for obesity,” *Diagnostics*, vol. 13, p. 2610, 8 2023, doi: 10.3390/diagnostics13152610.

[13] H. G. G. Bag, F. H. Yagin, Y. Gormez, P. P. González, C. Colak, M. Gülü, G. Badicu, and L. P. Ardigò, “Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits,” *Diagnostics*, vol. 13, p. 2949, 9 2023, doi: 10.3390/diagnostics13182949.

[14] X. Cheng, S. yu Lin, J. Liu, S. Liu, J. Zhang, P. Nie, B. F. Fuemmeler, Y. Wang, and H. Xue, “Does physical activity predict obesity—a machine learning and statistical method-based analysis,” *International Journal of Environmental Research and Public Health*, vol. 18, p. 3966, 4 2021, doi: 10.3390/ijerph18083966.

[15] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[16] I. E. Naqa and M. J. Murphy, *What Is Machine Learning?*, pp. 3–11. Springer International Publishing, 2015, doi: 10.1007/978-3-319-18305-3_1.

[17] S. Tajjour, S. Garg, S. S. Chandel, and D. Sharma, “A novel hybrid artificial neural network technique for the early skin cancer diagnosis using color space conversions of original images,” *International Journal of Imaging Systems and Technology*, vol. 33, pp. 276–286, 1 2023, doi: 10.1002/ima.22784.

[18] M. Arora and I. Bhardwaj, *Advances in Information Communication Technology and Computing*, vol. 392. Springer Nature Singapore, 2022, doi: 10.1007/978-981-19-0619-0.

[19] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*. The MIT Press, 1986, doi: 10.7551/mitpress/5236.001.0001.

[20] S. L. Salzberg, “C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993,” *Machine Learning*, vol. 16, pp. 235–240, 9 1994, doi: 10.1007/BF00993309.

[21] G. H. John and P. Langley, “Estimating continuous distributions in bayesian classifiers,” 2 2013, doi: 10.48550/arXiv.1302.4964.

[22] M. Grandini, E. Bagli, and G. Visani, “Metrics for multi-class classification: an overview,” *arXiv preprint arXiv:2008.05756*, 2020, doi: 10.48550/arXiv.2008.05756.

[23] D. W. Aha, D. Kibler, and M. K. Albert, “Instance-based learning algorithms,” *Machine Learning*, vol. 6, pp. 37–66, 1 1991, doi: 10.1007/BF00153759.

[24] M. Delacre, D. Lakens, and C. Leys, “Why psychologists should by default use welch’s t-test instead of student’s t-test,” *International Review of Social Psychology*, vol. 30, pp. 92–101, 4 2017, doi: 10.5334/irsp.82.

Paulo Henrique Ponte de Lucena is graduating in the Information Systems at the Federal University of Pará (UFPA) and he is a research scholarship recipient under the Institutional Program for Scientific Initiation Scholarships (PIBIC - UFPA). Engaged in research focused on Artificial Intelligence, specifically Deep Learning.

Lídio Mauro Lima de Campos holds a degree in ELECTRICAL ENGINEERING from the Federal University of Pará (1998), a degree in DATA PROCESSING TECHNOLOGY from the University of the Amazon (1990), a master’s degree in Computer Science from the Federal University of Santa Catarina (2001), and a Ph.D. in Electrical Engineering from the Federal University of Pará (2016). Currently, he is an Associate Professor II at the Faculty of Computing / Institute of Exact and Natural Sciences at UFPA. He has experience in the field of Computer Science, with a focus on the following topics: Artificial Intelligence, Machine Learning, Deep Learning, Time Series, and Neuroevolutionary Algorithms.

Jonathan Cris Pinheiro Garcia is graduated in Geophysics (2013) and Information Systems (2024) from the Federal University of Pará participated in the Institutional Program for Scientific Initiation Scholarships (PIBIC) during undergraduate studies. In 2017, he was part of the project "Implementation and Testing of Direct RNAs in Classification Tasks and Reinforcement Learning." And in 2022, participated in the project "Development of a Physical and Dynamic Bar Chart." Currently, he works as a Senior Development Analyst, excelling in the use of technologies such as Java, Spring, JSF, and Angular.