Predicting Obesity in Nutritional Patients using Decision Tree Modeling

Orlando Iparraguirre-Villanueva[1], Luis Mirano-Portilla[2], Manuel Gamarra-Mendoza[3], Wilmer Robles-Espiritu[4]
Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, Lima, Perú[1, 2, 3]
Facultad de Ingeniería y Arquitectura, Universidad César Vallejo, Lima, Perú[4]

***Abstract*—Obesity has become a widespread problem that affects not only physical well-being but also mental health. To address this problem and provide solutions, Machine Learning (ML) technology tools are being applied. Studies are currently being developed to improve the prediction of obesity. This study aimed to predict obesity levels in nutritional patients by analyzing their physical and dietary habits using the Decision Tree (DT) model. For the development of this work, we chose to use the CRISP-DM framework to follow the development in an organized way, thus achieving a better understanding of the data and describing, evaluating, and analyzing the results. The results of this work yielded metrics with significant values for predicting obesity: so much so that the accuracy rate was 92.89%, the sensitivity rate was 94% and the F1 score was 93%. Likewise, accuracy metrics above 88% were obtained for each level of obesity, demonstrating the effectiveness of the DT model in predicting this type of task. Finally, the results demonstrate that the DT model is effective in predicting obesity, with significant results that motivate further research to continue improving accuracy in this type of task.**

Keywords*—*Obesity; Machine Learning (ML); Decision Tree (DT); Prediction; CRISP-DM

I. Introduction

Today, obesity has become a potentially serious health problem worldwide. It is a condition characterized by an abnormal or excessive accumulation of fat in the body, which can have negative effects on a person's health. Obesity is associated with several health problems, including diabetes, heart disease, high blood pressure, and some forms of cancer [1]. It is also linked to psychological problems such as depression and anxiety.

Obesity is a major public health problem that causes physical and psychological health problems [2]. Surprisingly, this problem has tripled in the last four decades and, unfortunately, continues to increase [3]. This can pose a major public health challenge, especially for children and adults [4]. According to studies, global projections of adult obesity rates in 2010, 2025, and 2030 indicate an increase in obesity levels depending on the individual's degree of obesity [5], [6]. In addition, research on childhood obesity confirms a significant increase, which is now considered a global epidemic [7]. In 2010, UNICEF revealed that 40 million children had grade 1 obesity, with 81% of them coming from Asian countries [8]. Furthermore, it was predicted that by 2020, nearly one in ten children worldwide, and one in eight children in Africa, would be obese or undernourished [9]. The regions with the highest rates of childhood obesity are those in Asia-Pacific [10]. Worryingly, the rate of obesity has increased worldwide in the last decade. This is now considered a serious public health problem due to its strong connection with chronic diseases such as diabetes [11]. Obesity is a complex problem influenced by genetics, lifestyle, and environmental factors. Public health experts are using ML tools to predict and identify individuals at risk for obesity to provide personalized interventions [12]. Tools such as these make it easy to identify who needs help and create a personalized plan to meet their unique needs. With the potential provided by predictive analytics, proactive steps can be taken to combat obesity and help people live healthier, happier lives [13]. This trend is particularly concerning because childhood obesity is associated with an increased risk of chronic disease later in life. In addition, the incidence of cardiovascular disease in adults has increased, further emphasizing the need for effective prevention strategies [14]. Importantly, genetic factors may also influence short-term changes in body mass index (BMI), especially during the early years of development. However, due to the cross-sectional nature of the research, it is very complex to establish a causal relationship [15]. Nevertheless, these results highlight the importance of early intervention and prevention efforts to address the obesity epidemic and its associated health risks.

Technological approaches, especially ML models, can be excellent predictive tools that can help predict the level of obesity. For example, the DT model is a map that shows possible outcomes based on a series of related decisions, using algorithms to predict the degree of obesity of individuals [16]. BMI tests are a key indicator for predicting body fatness [17], so in this work, we seek to develop a model that can be used as a predictive index of obesity [18]. However, it is important to note that external validation is vital, as it represents the most optimal situation [19]. Therefore, simulated experiments based on BMI samples may not be as accurate [20]. Ultimately, the use of a tool that can optimize and streamline obesity prediction processes through an app provides users with a better experience based on the results obtained. This, in turn, ensures better treatment by healthcare professionals [21].

The objective of this study is to provide a technological solution capable of predicting the level of obesity in nutrition patients, thus improving the accuracy of obesity level prediction, and patient awareness and reducing the time required to make such predictions.

To achieve this objective, the following sections are presented: Section II presents the most relevant studies on obesity and the use of ML; Section III builds the methodology and develops the case study; Section IV presents the results of the study; Section V discusses the results with related works; and finally, Section VI presents the conclusions of the paper.

II. Literature Review

Multiple organizations, such as WHO/PHO, regularly publish reports or articles related to obesity. In addition, students, researchers, and independent groups have published papers that aim to address the problem of obesity using technology. For example, in study [22], a study analyzed six ML techniques to create a model capable of classifying obesity in individuals using a 3D scanner, X-ray equipment, and a body composition analyzer. The study obtained indicators above 75%, with the Random Forest technique presenting the best results. In the study [23], an algorithm was developed to analyze and predict whether infants are at risk for obesity based on their fourperiod BMI data. The algorithm was tested with 18818 infant samples and seven ML algorithms, and the Multilayer Perceptron algorithm provided the best results. It achieved an accuracy rate of 96%, with only 4% of cases classified as "At Risk", and a sensitivity value of 92%. This highlights the importance of being able to predict obesity in individuals. Also, in ref [24] they analyzed obesity in India using several algorithms such as Xero, EM, Apriori, and Best-First. They then evaluated better-known algorithms such as KNN, Linear Regression, and AdaBoost to predict and/or forecast obesity and gain new insights into the prediction of obesity in people. The study concluded that there are various levels of obesity in the population of the district where the research was conducted.

Similarly, in study [25], analysts developed an analysis of various ML algorithms such as K-NN, SVM, Logistic Regression (LR), Bayesian Networks, Random Forest, DT, AdaBoost, MLP, and Gradient Boosting to predict the risk of obesity. Two tests were performed, applying PCA in the second one, and the best result was an accuracy rate of 97.09% achieved by LR. The study concluded that obesity risk prediction was approached by evaluating nine ML techniques, and the most outstanding results were obtained with the Linear Regression technique. In study [26] conducted a study on single nucleotide polymorphisms related to eating habits that resulted in BMI readings equal to or greater than 25 kg/m² in 100 samples and BMI less than 25kg/m2 in 51 samples. The study also showed that individuals with allelic variants AgRP, Ala67Ala, ADRB2, Gln27Glu, Glu27Glu, INSIG2, Ala12Ala, and Pro 12 pro tend to develop obesity. Also, in [27], a predictive model was created to predict obesity in adult populations using ML techniques such as LR, Random Forest, Decision Tree, SVM, Gradient Boost, and Ada Boost. The study showed that LR and Decision Tree had the best performance in predicting obesity in adults based on accuracy. On the other hand, in study [28] a predictive model was developed using DT, LR, and KNN to estimate obesity levels from data related to dietary and physical habits, as well as other factors related to BMI. The study concluded that DT was the most effective technique for estimating obesity levels, with better accuracy than the other two techniques evaluated.

In study [29], a predictive model was created to forecast the level of obesity in high school students. The model employed four ML techniques: Binary LR, Enhanced DT, Weighted KNN, and Neural Networks. The results showed that the Binary LR technique had an accuracy rate of 56.02%, DT had an accuracy of 80.23%, KNN had an accuracy of 88.82%, and Neural Networks had an accuracy of 84.22%. The model with the highest accuracy was KNN, indicating that obesity is a major problem that needs to be addressed from various perspectives to reduce its prevalence among young people. Along the same lines, in study [30], they used ML algorithms such as SVM, DT, and Neural Networks, and applied Principal Component Analysis to determine the main factor of obesity in individuals using a dataset based on obesity-related patterns. The result was an accuracy level of 90% in both Neural Networks and DT algorithms while highlighting that a crucial factor in obesity is the presence of family members with obesity or overweight. Furthermore, in a study by [31], a predictive algorithm was developed to identify factors contributing to obesity and estimate obesity levels using unsupervised learning methods. The algorithm achieved an accuracy level of 97.8% using the cubic SVM technique.

III. Method

This section develops the theoretical basis of the DT model and the methodology used in the development of the case study.

A. Decision Tree

A DT is a nonparametric supervised learning algorithm that can be used for both classification and regression tasks. It has a hierarchical tree structure consisting of root nodes, branches, internal nodes, and leaf nodes [32], [33], [34]. Depending on the available features, both types of nodes perform evaluations by forming homogeneous subsets represented, by leaf nodes or end nodes. Leaf nodes represent all possible outcomes of a dataset [35], [36]. DT learning uses a divide-and-conquer strategy to determine the best-split point in the tree by greedy search. This partitioning process is recursively repeated from top to bottom until all or most of the records are classified under the given class label [37].

This approach provides a high degree of insight by determining the independent variable for each distribution in each branch of the tree. In addition, other algorithms or techniques belonging to the DT group, such as Random Forest or eXtreme Gradient Boosting, are based on decision trees [38], [39].

B. Understanding the Data

In this phase, we define the data set used for the development of the study, which comes from the Kaggle platform and comprises 17 variables relevant for predicting obesity levels based on dietary and physical patterns. To interpret the information, the data values are analyzed. Therefore, the logic of the data must be handled to accurately identify the functioning of the variables. This stage is crucial to understanding the behavior of the data, which helps to make informed decisions during the study.

C. Description of Data

In this phase, a general description of the data set is provided, including its variables, data types, and a brief description of what it represents. Table I presents the characteristics of the data set. The purpose of this section is to provide the reader with a clear understanding of the content of the data set and to interpret the results more accurately.

TABLE. I DATA SET DICTIONARY

Variable	Type	Description
Height	Float64	Person height in meters.
Weight		Person weight in kilograms.
FCVC		Frequent consumption of vegetables.
NCP		Number of main meals per day.
TUE		Use of technology devices in hours.
SMOKE		¿Does the person smoke?
CH2O		Dairy consumption of wáter.
FAF		Frequency of weekly physical activity.
Age		Age of the person.
Gender	Object	Gender of the person.
Oerweight_family_history		There are relatives with obesity.
FAVC		Frequent consumption of high-calorie foods.
CAEC		Food consumption between meals.
SCC		Monitoring of calorie consumption.
CALC		Frequency of alcohol consumption.
MTRANS		Means of transport usually used.
NSP		Obesity level (Target).

The dataset for this work is composed of more than 2000 entries and 17 variables, ranging from dietary patterns, physical habits, and general demographic data such as age and sex. The most relevant characteristic is the attribute "NSP", which reflects the patient's level of obesity and serves as the target variable.

D. Exploratory Data Analysis

In this phase of the data analysis process, the matplotlib-based Seaborn library was used to draw and explore the statistical data. This library is an excellent tool that integrates tightly with the panda’s data structures. Seaborn focuses on what the different elements of the graphs mean, rather than the details of how to draw them. Among its main functions is the plotting of data frames and matrices that are performed internally in semantic mapping and statistical aggregation. The graphs produced by this library provide valuable information about the dispersion of the data about the mean value of each variable or characteristic. It is very important to adjust the attributes of the axes, as shown in Fig. 1, where the degree of obesity of each patient is presented.

Fig. 1 shows the distribution of NSP variables among different levels of obesity in the data set. Fig. 1 clearly shows that there are more overweight patients than patients with any other degree of obesity. This distributional information can be used to determine the prevalence of obesity in a population and help design appropriate interventions to control obesity.

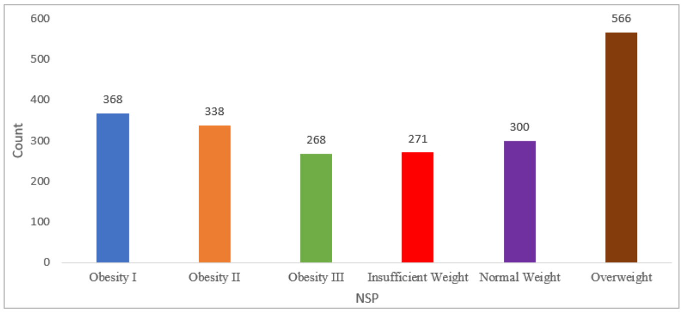


Fig. 1. NSP variable histogram.

In addition to distributional information, Fig. 2 shows the correlation between pairs of variables. There is a correlation between age and obesity level, which is useful for data set analysis. This information can be used to understand how different variables are related to each other and can help to better understand the data set.

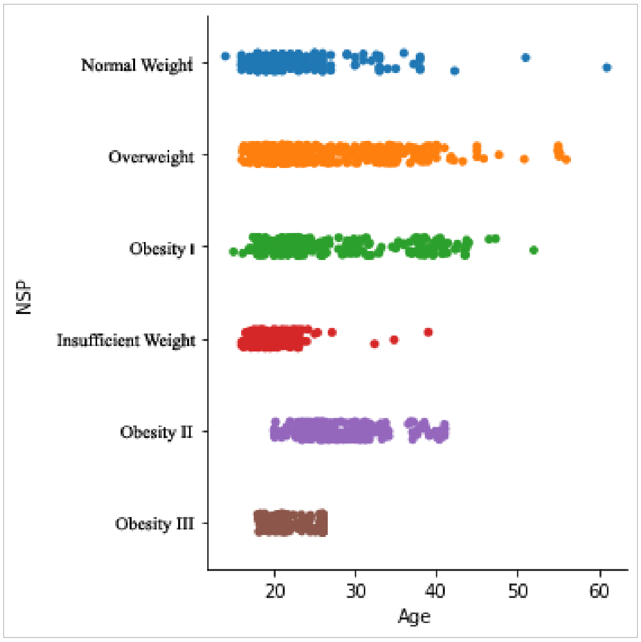
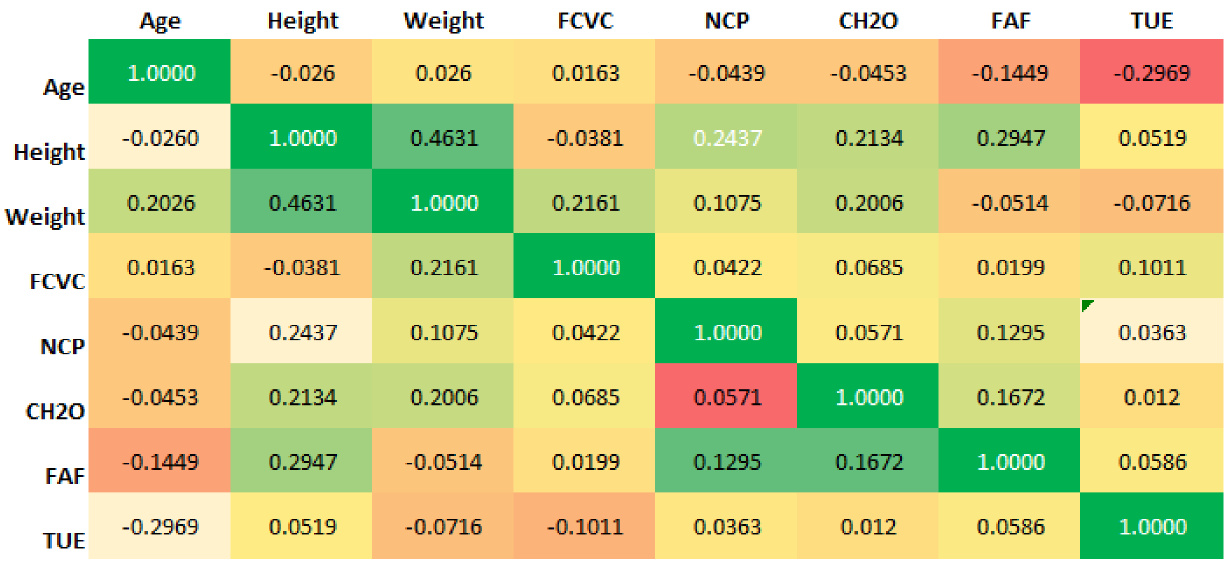


Fig. 2. Relationship between Age - NSP variables.

After the analysis of the data set, the degree of correlation between the variables was also examined. For this purpose, the correlation matrix of variables was used to analyze the relationship between different variables in the data set. The relationships between variables provide information on the strength and direction of the relationships between variables in the data set, and the correlation matrix is used to measure the correlation coefficient. For example, if two variables are highly correlated, this may indicate that they both measure the same phenomenon. Conversely, if two variables are negatively correlated, this may indicate that they are measuring opposite phenomena. As it can be seen in Fig. 3, there are two pairs of variables with high correlation coefficients: weight/height and age/truth.



	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
Age	1.0000	-0.026	0.026	0.0163	-0.0439	-0.0453	-0.1449	-0.2969
Height	-0.0260	1.0000	0.4631	-0.0381	0.2437	0.2134	0.2947	0.0519
Weight	0.2026	0.4631	1.0000	0.2161	0.1075	0.2006	-0.0514	-0.0716
FCVC	0.0163	-0.0381	0.2161	1.0000	0.0422	0.0685	0.0199	0.1011
NCP	-0.0439	0.2437	0.1075	0.0422	1.0000	0.0571	0.1295	0.0363
CH2O	-0.0453	0.2134	0.2006	0.0685	0.0571	1.0000	0.1672	0.012
FAF	-0.1449	0.2947	-0.0514	0.0199	0.1295	0.1672	1.0000	0.0586
TUE	-0.2969	0.0519	-0.0716	-0.1011	0.0363	0.012	0.0586	1.0000

Fig. 3. Variables correlation matrix.

The correlation matrix of variables presented in Fig. 3 shows correlation values ranging from -1 to 1, indicating the strength and direction of the relationship between the two variables. For example, age shows a weak negative correlation with most of the other variables, implying that as age increases, these variables tend to decrease. On the other hand, height shows a positive correlation with weight, suggesting that as height increases, weight also tends to increase. Similarly, weight shows a positive correlation with height. While there is a moderate correlation between FCVC and CH2O, there is a weak negative correlation with FAF and TUE. FCVC, or the frequency of eating raw vegetables, is directly related to body weight, meaning that people who eat more raw vegetables tend to lose weight. In addition, the number of main meals (PNC) was positively correlated with height and water intake.

In analyzing the data collected, we found that age plays a decisive role in determining the level of physical activity. Participation in physical activity appears to decrease as people age. In addition, there is a positive correlation between height and physical activity suggesting that taller people are likely to be physically active. In addition, the data indicated a weak positive correlation between PNC and CH2O and physical activity levels. However, it should be noted that these results may be influenced by other factors, so further studies may be needed to obtain better results.

Finally, the screen time variable showed a strong negative correlation with age. This finding means that as people age, their tendency to use electronic devices for longer period's decreases significantly. These results are critical to help us understand the factors that influence physical activity levels and screen time use patterns across all age groups.

E. Data Verification and Structuring

During this process, null or empty values are searched for in the data of each variable, to prepare them for training and verifying their post-processing behavior. These steps are essential to ensure data quality before proceeding with analysis and modeling.

During this stage, data scaling, balancing, and/or transformation are performed to structure the data set. These tasks are carried out to sort the data for each variable, which reduces the possible dispersion of values and improves the efficiency of the model. Previously, a data table was created for each variable showing the minimum, maximum, mean, median, and standard deviation values, to determine if scaling, balancing, or transformation techniques are required according to the model requirements, as shown in Table II. In addition, during model construction, data quality must be ensured and properly prepared, since scaling methods require that each variable be placed without repetitions to avoid negative impacts on the model due to the size of the variables.

TABLE. II VERIFICATION OF VARIABLES

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
Count	2111.00	2111.0	2111.0	2111.00	2111.00	2111.00	2111.00	2111.00
Mean	24.3126	1.7016	86.586	2.41904	2.68562	2.00801	1.01029	0.65786
Std	6.3459	0.0933	26.191	0.53392	0.77803	0.61295	0.85059	0.60892
Min	14.000	1.4500	39.000	1.00000	1.00000	1.00000	0.00000	0.00000
Max	61.0000	1.980	173.000	3.00000	4.0000	3.0000	3.0000	2.0000

F. Model Construction and Validation

In this phase, the model is used to make decisions based on the knowledge generated. The data set was divided into 20% for testing and 80% for training and validation. The Pandas, NumPy, and Scikit-learn libraries were used to implement the code in Python. This section allows the results of the model to be independently verified and evaluated. In addition, the command data_train.groupby('NSP') is used to display the number of people per NSP (obesity level).size() variable. This will allow us to see how people are distributed at each obesity level. It is also crucial to look at the type of variable in each column of the dataset. A graphical visualization can be performed within the cross-validation to achieve an 80% accuracy level.

G. Model Creation, Training, and Testing

Before creating a predictive model, it is important to verify the target variable (NSP) in the given data set. Table III shows the distribution of each NSP class in the training data set. Since imbalances in the data can affect the performance of the model, analyzing the distribution of the classes of the target variable is critical. Table III presents the number of instances of each NSP class in the training data set, indicating the proportion of each class in the data. This analysis is necessary to understand the distribution of classes, identify the need for balancing techniques such as oversampling or undersampling, and ensure that the model is trained equally on all classes. By addressing any imbalance, equitable model training and accurate predictions are ensured.

TABLE. III NSP VARIABLE DISTRIBUTION

	Obesity I	Obesity II	Obesity III	Insufficient Weight	Normal Weight	Overweight
Count	295	270	216	224	234	450

To ensure that the distribution of the data contained in each class of the target variable "NSP" have similar weights, or in case they do not, a scaling process must be performed. It can be observed that the class "Overweight" has more data compared to other classes with similar amounts according to the distribution of the classes of the target variable. Therefore, the weight of the "Overweight" class will be adjusted to match that of the other classes. Subsequently, a cross-validation of the Decision Tree will be performed. This process aims to estimate the optimal depth of the tree, which will increase the efficiency of the model. Then, the decision tree is generated using the training data and the previously calculated parameters (maximum depth, and weights for each class of the NSP variable).

IV. RESULTS

After training the DT model with the training data, we proceeded to predict the results with the test data, resulting in an impressive accuracy rate of 92.89%. This indicates a high level of model performance and reliability. In addition, the confusion matrix helped us to identify correct and incorrect predictions, giving us a more detailed view of the model's effectiveness, as shown in Fig. 4.

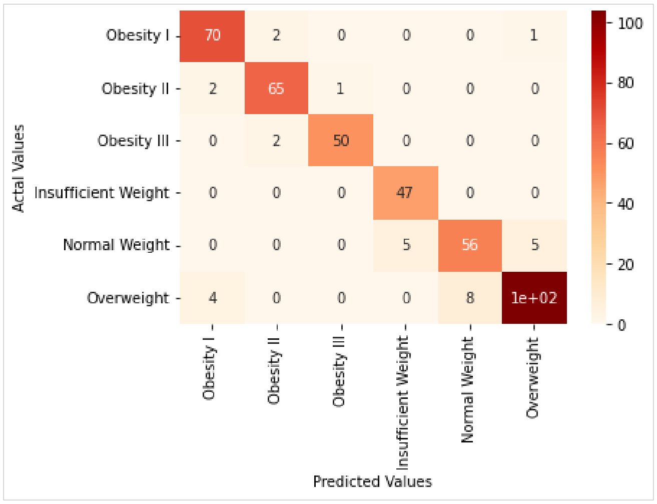


Fig. 4. Obesity levels correlation matrix.

In addition, to measure the performance of the model, several metrics were obtained for all classes of the "NSP" variable, as presented in Table IV. These metrics allow us to evaluate the percentages of accuracy, recall, and F1 scores obtained from the prediction of the test data. This provides a comprehensive assessment of the accuracy of the model, allowing us to determine the effectiveness of the prediction methodology. Using these metrics, it is possible to further refine and improve the results of the model to achieve better predictions in the future.

TABLE. IV CLASSIFICATION REPORT

	Precision	**Recall**	**F1-Score**	**Support**
Insufficient Weight	0.92	0.96	0.94	73
Normal Weight	0.94	0.96	0.95	68
Overweight	0.98	0.96	0.97	52
Obesity I	0.90	1.00	0.95	47
Obesity II	0.88	0.85	0.86	66
Obesity III	0.95	0.90	0.92	116
Accuracy			0.93	422
Macro AVG	0.93	0.94	0.93	422
Weighted AVG	0.93	0.93	0.93	422

To evaluate the classification of variables for each level of obesity obtained from the NSP, precision measures were analyzed for each variable in each type of obesity. The results showed an accuracy of 92% for underweight classes, 94% for normal weight, 98% for overweight, 90% for obesity I, 88% for obesity II and 95% for obesity III. These precision values indicate how accurately the model classified each level of obesity. Higher accuracy values mean that the model can correctly classify a higher proportion of cases at that level of obesity.

These results are obtained from the sample size and are based on an accuracy of 80% during the creation of the DT. The results reveal that the model is very effective in identifying different levels of obesity. The accuracy values further demonstrate the accuracy of the model's classifications, with the highest accuracy rate recorded for the Overweight category. The results are based on a consistent sample size, which lends credibility to the model's ability to accurately classify cases.

V. Discussion

The results of this research validate the main objective, but a comparative analysis with previous studies is necessary to demonstrate the relevance of the obesity prediction study. After evaluating several studies related to the topic presented in this research, we refer to the results found in the studies most like this work, which aim to predict obesity. For example, a study [40] predicted obesity using 3D scanner data with an accuracy of 80% and an accuracy of 84%, which are lower than the results obtained in this work. However, in the study [23], which predicted the risk of childhood obesity using a dataset of 18,818 infants in four age periods and applying seven ML algorithms, the Multilayer Perceptron algorithm obtained the best indicator with an accuracy of 96%, which is higher than the results of this work.

Several ML techniques were applied to predict obesity, with logistic regression and DTs being the most relevant models.

Furthermore, in study [27] predicted obesity in adults using their dietary patterns and various ML techniques, with logistic regression and DT models being the most relevant for predicting obesity. Finally, the results obtained in study [30] show similar levels of accuracy, where models such as SVM, DT, neural networks, and PCA were used to find that a decisive factor in obesity is family history, reaching an accuracy rate of 90% for DT and neural networks. Based on the above, we state that this work is very relevant in terms of obesity prediction since it has reached an accuracy rate of 92.89%, higher than those obtained in the studies. In conclusion, this work has managed to predict obesity with a very effective level of accuracy and contributes knowledge to a global problem that affects everyone.

VI. Conclusion

Several researchers and institutions are looking for technological solutions that allow a better prediction of obesity levels either at early ages, young people, or adults; in the present work, we sought, using decision trees, to predict the level of obesity in nutrition patients. The data set used for the present investigation consists of 2111 records and 17 variables that encompass both physical and dietary habits. The result achieved showed that DTs are very efficient in the prediction of the level of obesity, with the support of previous studies and the results obtained with the prediction of the test data, which were 92.89%, and it can be stated that the study was successful. The results obtained support the position that the application of DT to predict obesity levels does achieve its purpose, although, for future research or proposals for improvement, it is recommended to apply ML, more specifically DT, too much larger groups or to apply it to new scenarios or patterns that offer another point of view on obesity prediction. Regarding the limitations of the study:

1) The results of the study may not generalize to other populations or contexts because it is based on a specific dataset collected through the Kaggle platform. The accuracy of the model may also be affected by the representativeness and quality of the data.

2) Although class imbalance in the dataset has been addressed during model building, it remains an issue for predicting obesity. Obesity classes may not be equally represented in the population, which could bias the model results.

3) Several variables have been used to predict obesity, but the importance of each of these variables may be limited. Prediction may be more significantly affected by some characteristics than others, which may not be fully reflected in the results presented.

For future work, it is recommended to apply ML models to predict the level of obesity in demographic populations and to work with data covering different ethnic groups, ages, genders, and geographic locations. In addition, it is recommended to compare several ML models to predict obesity, such as LR, Support Vector Machines, Neural Networks, and Random Forest, among others. This work will allow us to determine if ML models such as DT are still the best option or if other ML algorithms offer equal or superior performance.

References

[1] M. Calderón-Díaz, L. J. Serey-Castillo, E. A. Vallejos-Cuevas, A. Espinoza, R. Salas, and M. A. Macías-Jiménez, "Detection of variables for the diagnosis of overweight and obesity in young Chileans using machine learning techniques.," *Procedia Comput Sci*, vol. 220, pp. 978–983, 2023, doi: 10.1016/j.procs.2023.03.135.

[2] D. Ryan, S. Barquera, O. Barata Cavalcanti, and J. Ralston, "The Global Pandemic of Overweight and Obesity," *Handbook of Global Health*, pp. 1–35, 2020, doi: 10.1007/978-3-030-05325-3_39-1.

[3] 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, no. 1, pp. 67–77, 2019, doi: 10.3844/jcssp.2019.67.77.

[4] "Obesidad y sobrepeso." https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight (accessed Jun. 20, 2023).

[5] R. N. Hiremath, M. Kumar, R. Huchchannavar, and S. Ghodke, "Obesity and visceral fat: Indicators for anemia among household women visiting a health camp on world obesity day," *Clin Epidemiol Glob Health*, vol. 20, Mar. 2023, doi: 10.1016/j.cegh.2023.101255.

[6] H. M. Salihu, S. M. Bonnema, and A. P. Alio, "Obesity: What is an elderly population growing into?," *Maturitas*, vol. 63, no. 1. pp. 7–12, May 20, 2019. doi: 10.1016/j.maturitas.2019.02.010.

[7] J. L. Díaz-Ortega, A. Q. Tácunan, M. G. Ancajima, L. C. Caracholi, and I. Y. Azabache, "Atherogenicity indicators in the prediction of metabolic syndrome among adults in trujillo-peru," *Revista Chilena de Nutricion*, vol. 48, no. 4, pp. 586–594, Aug. 2021, doi: 10.4067/S0717-75182021000400586.

[8] "GUÍA PROGRAMÁTICA DE UNICEF," 2020, Accessed: Jun. 20, 2023. [Online]. Available: https://www.unicef.org/media/96096/file/Overweight-Guidance-2020-ES.pdf

[9] O. Otitoola, W. Oldewage-Theron, and A. Egal, "Prevalence of overweight and obesity among selected schoolchildren and adolescents in Cofimvaba, South Africa," *South African Journal of Clinical Nutrition*, vol. 34, no. 3, pp. 97–102, 2021, doi: 10.1080/16070658.2020.1733305.

[10] "Overweight and obesity among adults | Health at a Glance 2021 : OECD Indicators | OECD iLibrary." https://www.oecd-ilibrary.org/sites/

0f705cf8-en/index.html?itemId=/content/component/0f705cf8-en (accessed Jun. 20, 2023).

[11] M. J. Duncan, C. Hall, E. Eyre, L. M. Barnett, and R. S. James, "Pre-schoolers fundamental movement skills predict BMI, physical activity, and sedentary behavior: A longitudinal study," *Scand J Med Sci Sports*, vol. 31, no. S1, pp. 8–14, Apr. 2021, doi: 10.1111/sms.13746.

[12] Z. Ren *et al.*, "Status and transition of normal-weight central obesity and the risk of cardiovascular diseases: A population-based cohort study in China," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 32, no. 12, pp. 2794–2802, Dec. 2022, doi: 10.1016/j.numecd.2022.07.023.

[13] J. E. Selem-Solís, A. Alcocer-Gamboa, M. Hattori-Hara, J. Esteve-Lanao, and E. Larumbe-Zabala, "Nutrimetry: BMI assessment as a function of development," *Endocrinología, Diabetes y Nutrición (English ed.)*, vol. 65, no. 2, pp. 84–91, Feb. 2018, doi: 10.1016/j.endien.2018.03.004.

[14] Q. Su, Y. Wu, B. Yun, H. Zhang, D. She, and L. Han, "The mediating effect of clinical teaching behavior on transition shock and career identity among new nurses: A cross-sectional study," *Nurse Educ Today*, p. 105780, Jun. 2023, doi: 10.1016/j.nedt.2023.105780.

[15] K. Silventoinen and H. Konttinen, "Obesity and eating behavior from the perspective of twin and genetic research," *Neuroscience and Biobehavioral Reviews*, vol. 109. Elsevier Ltd, pp. 150–165, Feb. 01, 2020. doi: 10.1016/j.neubiorev.2019.12.012.

[16] F. Bollwein and S. Westphal, "Oblique decision tree induction by cross-entropy optimization based on the von Mises–Fisher distribution," *Comput Stat*, vol. 37, no. 5, pp. 2203–2229, Nov. 2022, doi: 10.1007/s00180-022-01195-7.

[17] S. Tanaka *et al.*, "A clinical prediction rule for predicting a delay in quality of life recovery at 1 month after total knee arthroplasty: A decision tree model," *Journal of Orthopaedic Science*, vol. 26, no. 3, pp. 415–420, May 2021, doi: 10.1016/j.jos.2020.04.010.

[18] G. Radetti, A. Fanolla, G. Grugni, F. Lupi, and A. Sartorio, "Indexes of adiposity and body composition in the prediction of metabolic syndrome in obese children and adolescents: Which is the best?," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 29, no. 11, pp. 1189–1196, Nov. 2019, doi: 10.1016/J.NUMECD.2019.06.011.

[19] J. P. Santisteban Quiroz, "Estimation of obesity levels based on dietary habits and condition physical using computational intelligence," *Inform Med Unlocked*, vol. 29, Jan. 2022, doi: 10.1016/j.imu.2022.100901.

[20] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[21] A. S. Mohd Faizal, T. M. Thevarajah, S. M. Khor, and S. W. Chang, "A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach," *Comput Methods Programs Biomed*, vol. 207, Aug. 2021, doi: 10.1016/j.cmpb.2021.106190.

[22] S. Jeon, M. Kim, J. Yoon, S. Lee, and S. Youm, "Machine learning-based obesity classification considering 3D body scanner measurements," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–10, Feb. 2023, doi: 10.1038/s41598-023-30434-0.

[23] B. Singh and H. Tawfik, "Machine learning approach for the early prediction of the risk of overweight and obesity in young people," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2020, pp. 523–535. doi: 10.1007/978-3-030-50423-6_39.

[24] M. Mahapatra and K. K. Singh, "Prediction of causes and effects of obesity in India by supervise learning approaches," *Obes Med*, vol. 34, p. 100436, Sep. 2022, doi: 10.1016/J.OBMED.2022.100436.

[25] 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, Nov. 2021, doi: 10.1016/j.crbeha.2021.100053.

[26] C. Rodríguez-Pardo *et al.*, "Decision tree learning to predict overweight/obesity based on body mass index and gene polymporphisms," *Gene*, vol. 699, pp. 88–93, May 2019, doi: 10.1016/J.GENE.2019.03.011.

[27] K. N. Devi, N. Krishnamoorthy, P. Jayanthi, S. Karthi, T. Karthik, and K. Kiranbharath, "Machine Learning Based Adult Obesity Prediction," *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*, 2022, doi: 10.1109/ICCCI54379.2022.9740995.

[28] T. Cui, Y. Chen, J. Wang, H. Deng, and Y. Huang, "Estimation of obesity levels based on decision trees," *Proceedings - 2021 International Symposium on Artificial Intelligence and its Application on Media, ISAIAM 2021*, pp. 160–165, May 2021, doi: 10.1109/ISAIAM53259.2021.00041.

[29] Z. Zheng and K. Ruggiero, "Using machine learning to predict obesity in high school students," *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, vol. 2017-January, pp. 2132–2138, Dec. 2017, doi: 10.1109/BIBM.2017.8217988.

[30] Z. He, "Comparison Of Different Machine Learning Methods Applied To Obesity Classification," *Proceedings - 2022 International Conference on Machine Learning and Intelligent Systems Engineering, MLISE 2022*, pp. 467–472, 2022, doi: 10.1109/MLISE57402.2022.00099.

[31] Y. Celik, S. Guney, and B. Dengiz, "Obesity Level Estimation based on Machine Learning Methods and Artificial Neural Networks," *2021 44th International Conference on Telecommunications and Signal Processing, TSP 2021*, pp. 329–332, Jul. 2021, doi: 10.1109/TSP52935.2021.9522628.

[32] Q. Li, X. Wang, Q. Pei, X. Chen, and K.-Y. Lam, "Consistency preserving database watermarking algorithm for decision trees," *Digital Communications and Networks*, Jan. 2023, doi: 10.1016/j.dcan.2022.12.015.

[33] K. Ramya, Y. Teekaraman, and K. A. Ramesh Kumar, "Fuzzy-based energy management system with decision tree algorithm for power security system," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1173–1178, 2019, doi: 10.2991/ijcis.d.191016.001.

[34] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. F. Castañeda, and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," *Diagnostics 2023, Vol. 13, Page 2383*, vol. 13, no. 14, p. 2383, Jul. 2023, doi: 10.3390/DIAGNOSTICS13142383.

[35] S. Garg and P. Pundir, "MOFit: A Framework to reduce Obesity using Machine learning and IoT," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.08868

[36] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[37] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: A review," in *Procedia Manufacturing*, Elsevier B.V., 2019, pp. 698–703. doi: 10.1016/j.promfg.2019.06.011.

[38] H. Rao *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing Journal*, vol. 74, pp. 634–642, Jan. 2019, doi: 10.1016/j.asoc.2018.10.036.

[39] O. Iparraguirre-Villanueva *et al.*, "Comparison of Predictive Machine Learning Models to Predict the Level of Adaptability of Students in Online Education," 2023. doi: http://dx.doi.org/10.14569/IJACSA.2023.0140455.

[40] S. Jeon, M. Kim, J. Yoon, S. Lee, and S. Youm, "Machine learning-based obesity classification considering 3D body scanner measurements," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-30434-0.