Utilizing AI to Tackle Child Overnutrition in Morocco

Mustapha Berrouyne, Hinde Hami, Fatine Hadrya and Youness Jouilil

September 17, 2024

Utilizing AI to Tackle Child Overnutrition in Morocco

1st Mustapha Berrouyne
Faculty of Sciences,
Ibn Tofail University of Kenitra
Kenitra, Morocco
mbmberrouyne@gmail.com
https://orcid.org/0009-0009-7290-3608

2nd Hinde Hami
Faculty of Sciences,
Ibn Tofail University of Kenitra
Kenitra, Morocco
hinde.hami@uit.ac.ma
https://orcid.org/0000-0002-1803-0115

3rd Fatine Hadrya
Higher Institute of Health Sciences,
University Hassan First of Settat
Settat, Morocco
fatine.hadrya@uhp.ac.ma
https://orcid.org/0000-0002-0394-8329

4rd Youness Jouilil
dept. of Economics,
Hassan II University of Casablanca
Casablanca, Morocco
y.jouilil@gmail.com
https://orcid.org/0000-0002-0394-8329

***Abstract*—This study investigates the predictive power of various machine learning classifiers in identifying overnutrition among children under five years old in Morocco. Using data from the 2018 national population and family health survey, the research employs ten machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), Gradient Boosting, Random Forest, XGBoost, k-Nearest Neighbors, Decision Trees, Naive Bayes, Artificial Neural Networks, and Deep Learning models. The performance of these models was assessed using accuracy, sensitivity, specificity, kappa statistic, and area under the curve (AUC). The results reveal that Logistic Regression and SVM were the most effective models, achieving nearly ninety percent accuracy for predicting overweight and approximately ninety-seven percent accuracy for predicting obesity. These models also demonstrated high sensitivity and specificity. Gradient Boosting and Random Forest also showed strong performance, while Naive Bayes, despite its lower accuracy, excelled in AUC, indicating its proficiency in overall class distinction. The findings highlight the significant roles of birth weight, socioeconomic status, and parental education in influencing childhood overnutrition. By focusing on the Moroccan context, this study addresses a gap in the existing literature and provides actionable insights for developing targeted public health interventions. The study underscores the effectiveness of Logistic Regression and SVM in handling complex datasets and the importance of early life factors, socioeconomic status, and parental education. The research suggests the need for incorporating more recent and longitudinal data, exploring a broader range of variables, and employing advanced machine learning techniques in future studies to enhance predictive accuracy into the determinants of childhood overnutrition.**

***Index Terms*—Machine learning algorithms, child Overnutrition, childhood overweight, childhood obesity, Determinants.**

I. Introduction

Childhood obesity and overweight are a significant global public health challenge, with their prevalence rising over recent decades. It is leading to a range of serious health consequences. The concepts of overweight and obesity often stem from overnutrition. This one is at an increased risk for type 2 diabetes, cardiovascular diseases, and certain cancers due to excessive fat accumulation [1]–[4]. Obesity also significantly impairs health by disrupting sleep and mobility, and it heightens the risk of bone health issues [5], [6]. Furthermore, childhood obesity is strongly associated with adult obesity.

The World Health Organization (WHO) reported that over 37 million children under 5 were classified as overweight or obese globally in 2022, up from 32 million in 1990. The prevalence of overweight in this age group has driven by poor dietary habits, reduced physical activity, and high-calorie, low-nutrient foods [7].

In Morocco, overweight and obesity are starting to grow, with boys being more affected than girls. According to the 2018 National Population and Family Health Survey (NPFHS), 12.3% of boys are overweight compared to 9.2% of girls, and 3.6% of boys are obese compared to 2.2% of girls. The survey also indicates that urban residents are more affected by overweight than those in rural areas, with 11.7% of urban children being overweight compared to 9.7% in rural areas [8]. Based on the data from the 2011 NPFHS, the prevalence of overweight in children aged 0-59 months stands at 10.7% nationwide. Boys have a higher overweight rate of 12.5% compared to 8.8% in girls [9]. Our goal throughout this study is to choose the best classifier to predict overnutrition among children under five years old in Morocco.

The first section, literature review, covers existing studies on machine learning applications in public health. The second section, materials and methods, details the data sources, sampling methods, and machine learning algorithms used. The third section, results, presents the performance metrics of various models and identifies key predictors. The fourth section, important determinants of child overweight and obesity, analyzes the key factors influencing childhood overnutrition. The fifth section, discussions, interprets the findings, compares them with existing literature, and suggests public health interventions. The sixth section, limitations of the study, discusses the constraints and potential biases of the research. The seventh and final section, conclusion, summarizes key insights, addresses limitations, and proposes future research directions.

II. Literature Review

Pang et al. (2021) compared several models, including XGBoost, decision trees, Gaussian naive bayes, Bernoulli naive bayes, logistic regression, neural networks, and support vector machines. XGBoost outperformed all other models with an AUC of 0.81, precision of 30.9%, accuracy of 66.14%, and specificity of 63.27%. Their findings identified important predictors such as weight, height, head circumference, body temperature, and respiratory rate [10].

Siddiqui et al. (2021) applied traditional ML models, including support vector machines, along with deep learning techniques like artificial neural networks and recurrent neural networks. Random forests and artificial neural networks showed high accuracy rates of 8% and 84%, respectively. Their findings underscored the importance of early predictors, such as diet and physical activity [6].

Lim et al. (2023) used logistic regression, decision trees, and random forests to predict childhood obesity. Logistic regression achieved the best performance with an AUC of 0.82, accuracy of 74%, sensitivity of 76%, and specificity of 73%. They identified key risk factors including gender, previous BMI, and maternal weight, emphasizing the importance of physical activity and maternal psychological status [11].

Gonzalo et al. (2020) utilized gradient boosting machines and deep learning models for predicting childhood obesity. XGBoost achieved an AUC of 0.86, while Deep Learning achieved an AUC of 0.84. They highlighted that XGBoost and deep learning methods outperform traditional models [12].

Lebron et al. (2020) explored naïve Bayesian classifiers along with classification and regression tree analysis. Classification and regression tree provided a visual representation of decision rules (AUC of 0.81, accuracy of 85%, sensitivity of 78%, and specificity of 80%) [13].

Hammond et al. (2019) combined regression models such as LASSO, random forests, and gradient boosting with logistic regression classifiers. Their study found that the LASSO regression model achieved the best performance with an AUC of 81.7% for girls and 76.1% for boys [14].

Gray et al. (2019) combined multiple regression models,including LASSO, ridge regression, and elastic net, with logistic regression classifiers. They utilized bootstrap cross-validation on 70% of the sample for model training and 30% for testing. Their study found that the LASSO regression model provided the best prediction of overweight/obesity [15].

Triantafyllidis et al. (2020) utilized decision trees, artificial neural networks, and logistic regression. Decision trees and artificial neural networks were most effective (AUC=0.83) [16].

Dugan et al. (2015) combined several machine learning techniques, including Random Tree, Random Forest, ID3, J48, Naïve Bayes, and Bayes Net, to predict early childhood obesity. Among the methods analyzed, the ID3 model provided the best overall performance, achieving an accuracy of 85% and a sensitivity of 89% [17].

Despite significant advancements in the study of childhood overnutrition, several critical gaps persist in the existing literature. This study aims to bridge these gaps by providing a novel and comprehensive analysis focused on the Moroccan context, utilizing advanced machine learning techniques.

Firstly, research on childhood overnutrition in Morocco using advanced machine learning techniques is notably scarce. Furthermore, this study compares ten different algorithms, identifying the best-performing models and analyzing their strengths and weaknesses.

Additionally, while many studies identify key predictors of childhood overnutrition, they often fall short in offering practical public health recommendations.

III. Materials and Methods

A. Data Source

The nutrition data from the 2018 National Population and Family Health Survey (NPFHS) conducted by the Moroccan Ministry of Health was utilized. The survey aimed to ensure that it accurately represented Moroccan society through a stratified sampling design executed in two stages. The study sample consisted of 5983 children under the age of five. The Ministry of Health preprocessed the data and documented the results in their 2019 report, providing a thoroughly cleaned and ready-to-analyze database.

We divided the data 70:30 after the preprocessing stage [18]. The testing dataset, comprising the remaining 30% of the data, was used to validate the predictive models and assess their ability to forecast undernutrition. Conversely, the training dataset, making up 70% of the data, was used to train the models, identify the best ones, and evaluate their effectiveness.

B. Outcome variables

Two dependent variables will be used to approach malnutrition: overweight and obesity. These variables were constructed using WHO standards [19]. The following code applies to each of these variables: 1 if the child has this characteristic, and 0 if not.

C. Explanatory variables

Detailed below are the characteristics of mothers, families, and children that were found to be associated with overnutrition as explanatory variables. We describe the child's variables as follows: Child age months quantifies the individual's age in months; genre identifies the gender of the child; twin child signifies if the child is a twin; and child rank sheds light on the position of the child within their sibling hierarchy. Finally, birthweight records the child's weight at the time of birth.

Here are some characteristics of the mother: education in a couple, education for partners and mothers, mother's age, the status of breastfeeding and the place of the child's birth.

Here are some characteristics of the Households: the size of households, the wealth of a household the residential areas and the region.

D. Statistical methods

The researcher utilized Python tool to examine the data. This software offers a more flexible environment for data analysis [20], [21]. This programming language is used in the analysis process so that advanced predictive models can be used, and feature importance evaluations can be run to find the dataset's most important predictors.

E. Training Models

We split the dataset into two halves: we set aside 30% for testing and 70% for training. Based on our literature review, we explored the potential for overnutrition prediction using ten widely used machine learning classifiers: gradient boosting (GB), random forests (RF), XGBoost, k-nearest neighbors (KNN), support vector machines (SVM), artificial neural networks (ANN), logistic regression (LR), decision trees (DT), Naive Bayes (NB), and deep learning models (DL). These methods have been shown to effectively identify key predictors and provide robust, accurate predictions in various studies on childhood obesity and related health conditions.

1) Random Forest: One approach to supervised ensemble learning is random forest, which uses decision trees [19], [23]. The RF algorithm iteratively produces a regression classification tree by sampling the training data set variables with new sets of predictor variables. Because of its speed and versatility, it finds dual use in regression and classification.

F. Evaluation of the Model's Performance

The following four metrics will be used to evaluate the model's performance [20], [22], [23].

1) Accuracy: A predictive algorithm's efficiency is proportional to its accuracy. Out of all the examined data points, the predictive algorithm determines the percentage of accurately predicted points. In this study, we combined the top accuracy results from various machine learning algorithms that used k-fold techniques and feature selection.

$$Accuracy = \frac{T.Pos + T.Neg}{T.Pos + F.Neg + F.Pos + T.Neg} \tag{1}$$

2) Sensitivity: Sensitivity, often referred to as true positive rate, measures the proportion of actual positive cases that are accurately identified as positive. It is also known as recall. This indicates that a certain percentage of real positive cases might be incorrectly classified as negative, which is termed as the false negative rate.

$$Sensitivity = \frac{T.Pos}{T.Pos + F.Neg} \tag{2}$$

3) Specificity: Specificity, or the proportion of true negatives, measures the cases accurately predicted as negative. Consequently, it is possible to incorrectly classify some actual negative cases as positive, an occurrence known as false positives.

$$Specificity = \frac{T.Neg}{T.Neg + F.Pos} \tag{3}$$

4) Cohen's K: The statistic compares different classifiers, with accuracy represented by a confusion matrix.

$$K = \frac{total\_Accuracy - random\_Accuracy}{1 - random\_Accuracy} \tag{4}$$

Another way to look at it is:

$$total\_Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Accuracy} = \frac{(TN+FP)(TN+FN)+(FN+TP)(FP+TP)}{\text{Total} \times \text{Total}} \tag{6}$$

These metrics were chosen for a comprehensive evaluation of the model, addressing overall correctness, positive case identification, negative case identification, and agreement beyond chance.

IV. RESULTS

A. Prediction Models

Tables 1 and 2 presents the performance indicators of ten machine learning algorithms utilized for predicting overweight and obesity, respectively, among Moroccan children under five. The tables is structured to compare the models across several key metrics: Accuracy, Specificity, Sensitivity, Kappa Statistic (κ), and Area Under Curve (AUC). Each metric offers insight into the models' ability to correctly identify cases of overweight and obesity.

1) Overweight: In terms of accuracy, Logistic Regression and SVM exhibit the highest performance with an accuracy of 0.887, demonstrating strong reliability in correct classifications. Gradient Boosting and Random Forest also show high accuracy, with values of 0.886 and 0.883, respectively. Naive Bayes, with an accuracy of 0.758, is the least effective in comparison.

Regarding specificity, Logistic Regression shows the highest specificity at 0.563, indicating better performance in identifying non-overweight individuals. Gradient Boosting and Random Forest follow with specificities of 0.506 and 0.501, respectively. Naive Bayes has a low specificity of 0.117, indicating challenges in this aspect.

Considering sensitivity, Logistic Regression, SVM, and Gradient Boosting exhibit near-perfect sensitivity (1.000 and 0.999, respectively), indicating their effectiveness in identifying overweight cases. Naive Bayes, with a sensitivity of 0.810, is the lowest in this aspect.

Evaluating the kappa statistic (κ) shows that Logistic Regression (0.450) has a relatively high kappa value, indicating better agreement beyond chance. Naive Bayes also has a moderate kappa value of 0.350, while SVM and Gradient Boosting have lower kappa values, indicating poorer agreement.

According to the Area Under the Curve (AUC), which evaluates the overall ability to distinguish between classes, Naive Bayes achieves the highest AUC at 0.63, indicating better overall performance. Logistic Regression and Gradient Boosting follow with AUC values of 0.62 and 0.61, respectively. K Nearest Neighbors (KNN) and Decision Trees, with AUC values of 0.51-0.52, indicate less reliability.

Given these observations, Logistic Regression and SVM, with their high accuracy and sensitivity, are excellent for identifying overweight cases. Logistic Regression also has the highest specificity, making it particularly balanced. Gradient Boosting and Random Forest, with relatively high specificity and accuracy, are robust alternatives.

Naive Bayes, despite its lower accuracy and sensitivity, excels in AUC and shows moderate specificity, offering a balanced performance suitable for applications needing both sensitivity and specificity. K Nearest Neighbors and Decision Trees perform poorly overall, suggesting they are less effective for this task. Artificial Neural Networks and Deep Learning models provide good accuracy and sensitivity with moderate AUC, indicating effectiveness but not top-tier performance.

TABLE I
THE PERFORMANCE INDICATORS OF TEN MACHINE LEARNING ALGORITHMS USED FOR PREDICTING OVERWEIGHT

Algorithms	Accuracy	Specificity	Sensitivity	K	AUC
Logistic Regression	0.887	0.563	1.000	0.450	0.62
K Nearest Neighbors	0.879	0.390	0.986	0.039	0.51
Decision Trees	0.785	0.410	0.862	0.184	0.52
Gradient Boosting	0.886	0.506	0.999	0.005	0.61
XGBOOST	0.865	0.119	0.970	0.044	0.56
Random Forest	0.883	0.501	0.996	0.005	0.57
Support Vector Machines	0.887	0.102	1.000	0.100	0.56
Naive Bayes	0.758	0.117	0.810	0.350	0.63
Artificial Neural Networks	0.823	0.055	0.910	0.141	0.57
Deep Learning Model	0.883	0.036	0.993	0.029	0.58

2) Obesity: When examining accuracy, Logistic Regression, Random Forest, and SVM stand out with the highest accuracy of 0.967, indicating they are highly reliable in making correct classifications. Gradient Boosting and K Nearest Neighbors (KNN) also show strong performance with accuracy values of 0.966 and 0.965, respectively. On the other hand, Naive Bayes has the lowest accuracy at 0.860, making it the least effective among the algorithms.

In terms of specificity Logistic Regression leads with a specificity of 0.350. Naive Bayes follows with a specificity of 0.317, showing a moderate ability to identify non-obese cases. However, Gradient Boosting and XGBOOST have very low specificities of 0.001 and 0.004 respectively, indicating significant difficulties in correctly identifying non-obese cases.

Evaluating sensitivity shows that Logistic Regression, SVM, and Random Forest all achieve perfect sensitivity (1.000), making them highly effective in detecting obese cases. Naive Bayes, with a sensitivity of 0.875, performs the lowest in this aspect but still shows strong performance.

Looking at the kappa statistic (κ), which adjusts for chance agreement, Logistic Regression (0.442) and Naive Bayes (0.417) show the highest kappa values, indicating better-than-chance agreement. In contrast, Gradient Boosting and SVM have lower kappa values, reflecting poorer agreement beyond chance.

Considering the Area Under the Curve (AUC), Logistic Regression achieves the highest AUC at 0.653, indicating strong overall performance. Naive Bayes follows closely with an AUC of 0.651. KNN and XGBOOST have lower AUC values, indicating they are less reliable in class discrimination.

Given these observations, Logistic Regression and SVM, with their high accuracy and sensitivity, are excellent for identifying obese cases. Logistic Regression also has relatively high specificity and AUC, making it a particularly balanced choice. Random Forest and Gradient Boosting, with their high accuracy and sensitivity, are also robust alternatives.

TABLE II
PERFORMANCE OF VARIOUS ALGORITHMS IN PREDICTING OBESITY.

Algorithms	Accuracy	Specificity	Sensitivity	κ	AUC
Logistic Regression	0.967	0.350	1.000	0.442	0.653
K Nearest Neighbors	0.965	0.003	0.998	0.050	0.555
Decision Trees	0.929	0.036	0.958	0.083	0.529
Gradient Boosting	0.966	0.001	0.999	0.060	0.621
XGBOOST	0.965	0.004	0.998	0.040	0.534
Random Forest	0.967	0.263	1.000	0.019	0.588
Support Vector Machines	0.967	0.303	1.000	0.430	0.667
Naive Bayes	0.860	0.317	0.875	0.417	0. 651
Artificial Neural Networks	0.959	0.012	0.991	0.017	0.573
Deep Learning Model	0.965	0.004	0.998	0.000	0.582

B. Important determinants of child Overnutrition

1) Overweight: Figure 1 shows the feature importance from the logistic regression model for childhood overweight predictors, expressed as odds ratios. The results highlight the significant roles of birth characteristics, socioeconomic factors, parental education, geographic location, and lifestyle in influencing childhood overweight. Birth weight is the most critical predictor, underscoring early life factors. Socioeconomic status and parental education are crucial, likely affecting lifestyle choices and resource access. Geographic disparities suggest the need for region-specific interventions. Additionally, family dynamics and maternal characteristics add further layers of influence for consideration in public health strategies.

Furthermore, children with normal birth weight (birth_weight_3) are 1.75 times more likely to be overweight than those with low birth weight. Those with elevated birth weight (birth_weight_4) have a 1.70 times greater chance, and children with normal birth weight (birth_weight_2) are 1.65 times more prone to overweight.

Geographic location also matters. Children in ”Laayoune-Sakia El Hamra” (region_11) and ”Beni Mellal-Khenifra” (region_5) have a 1.50 and 1.45 times higher probability, respectively, of being overweight than those in ”Tanger-Tetouan-Al Hoceima” (region_1).

Higher household wealth is crucial. Children from wealthier households, ”Rich” (household_wealth_4) and ”Richer” (household_wealth_5), are 1.40 and 1.35 times more likely to be overweight than those from the poorest households.

Education levels of mothers and partners are significant. Children whose mothers have preparatory/middle education (maternal_educ_level_3) are 1.30 times more likely to be overweight. Similarly, children whose partners have secondary education (partner_educ_level_4) face a 1.25 times higher risk.

Child and family characteristics influence overweight risk. Fourth or later-born children (child_rank_3) have a

1.20 times greater likelihood. Children of older mothers (maternal_age_5) and (maternal_age_4) are 1.15 and 1.10 times more likely to be overweight compared to those of mothers under 25 years.

Diet and lifestyle are important. Children with lower quality food (quality_food_2) have a 1.25 times higher likelihood of being overweight. Non-breastfed children (breastfeeding_2) show a 1.20 times increased risk.

Other factors include environment and context. Rural children (area_2) are 1.10 times more likely to be overweight.

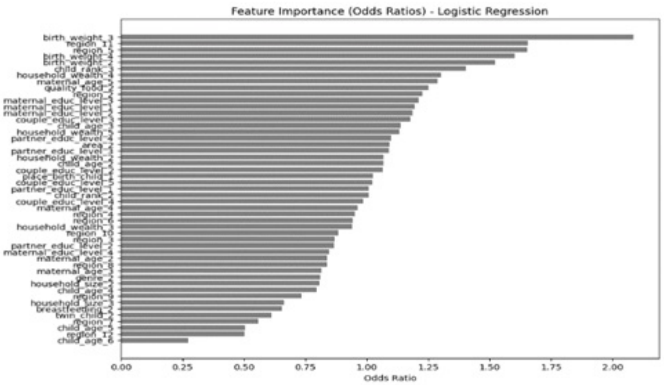

Fig. 1. Important determinants of child Overweight

2) Obesity: Figure 2 indicates that a complex interplay of birth characteristics, socioeconomic factors, parental education, geographic location, and lifestyle influences the risk of childhood obesity. Regions such as ”Laayoune-Sakia El Hamra” and ”Beni Mellal-Khenifra” highlight significant geographic disparities in obesity risk. Children in ”Laayoune-Sakia El Hamra” (region_11) are nearly 2.7 times more likely to be obese than those in ”Tanger-Tetouan - Al Hoceima” (region_1), while those in ”Beni Mellal-Khenifra” (region_5) have almost 2.4 times greater risk.

Socioeconomic status and parental education crucially impact lifestyle choices and resource access. Birth weight and maternal age are significant early life factors. Children whose partners have preparatory/middle (partner_educ_level_3) or secondary (partner_educ_level_4) education levels are 2.3 and 2.2 times more likely to be obese, respectively. Children of mothers aged 40+ years (maternal_age_5) are 2.1 times more likely to be obese, while those of mothers aged 35-39 years (maternal_age_4) are 2.0 times more likely.

Household wealth significantly influences obesity risk. Children from the wealthiest households (household_wealth_5) are 2.0 times more likely to be obese, and those from the ”Rich” category (household_wealth_4) are 1.9 times more likely. Geographic disparities are also notable in regions like ”Oriental” (region_2) and ”Casablanca-Settat” (region_10), where children are 1.8 times more likely to be obese.

Other influential factors include birth weight, with children of elevated birth weight (birth_weight_4) being 1.7 times more likely to be obese. Lower quality food (quality_food_2) increases obesity risk by 1.6 times. Maternal and partner education levels, such as ”Préparatoire/moyen” (partner_educ_level_2) and ”Secondaire” (partner_educ_level_3), show higher obesity likelihoods with odds ratios of 1.5 and 1.4, respectively. Children with normal birth weight (birth_weight_2) have a 1.3 times higher likelihood of obesity.

Additional child and family characteristics, such as being the fourth or later in birth order (child_rank_3) and living in rural areas (area_2), also indicate higher obesity risk, with odds ratios of 1.2 and 1.1, respectively. Being born at home (place_birth_child_1) also suggests an increased risk of obesity.

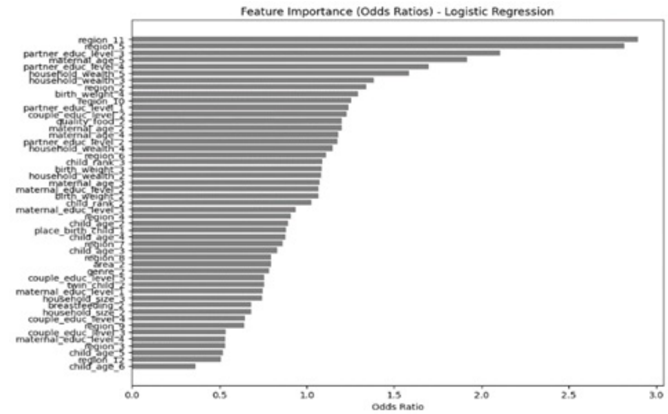

Fig. 2. Important determinants of child Obesity

In short, childhood overweight and obesity are influenced by birth characteristics, socioeconomic factors, parental education, geographic location, and lifestyle. Birth weight is a critical predictor, highlighting early life factors. Geographic disparities show certain regions with higher risks. Higher household wealth and parental education significantly increase the likelihood of both conditions. Additionally, child and family characteristics, such as birth order and maternal age, play significant roles. The odds ratios for obesity are generally higher than those for overweight, indicating a more pronounced impact on obesity risk.

Our study reveals significant insights into the determinants of childhood overweight and obesity, suggesting potential causal relationships that can inform public health strategies. Key factors include birth characteristics, socioeconomic status, parental education, geographic location, and lifestyle.

High birth weight is a strong predictor, emphasizing the need for optimal maternal health and nutrition during pregnancy to prevent early-life overnutrition. Socioeconomic factors also play a crucial role, with higher household wealth linked to increased risk, likely due to greater access to high-calorie foods and sedentary lifestyles. Public health messages should thus target all socioeconomic groups.

Parental education, particularly maternal education, is another critical determinant. Even among educated families, higher risks persist, suggesting that knowledge alone is insufficient without behavioral changes and environmental support. Enhancing parental education programs with practical strategies for healthy living is essential.

Geographic disparities indicate that regional factors such as cultural practices, the availability of healthy foods, and local health policies significantly impact childhood overweight and obesity. Regions like ”Laayoune-Sakia El Hamra” and ”Beni Mellal-Khenifra” show higher risks, necessitating region-specific interventions.

Lifestyle factors, including diet and physical activity, are vital. Lower-quality food and a lack of breastfeeding lead to higher risks. Promoting breastfeeding and nutritious food access, reducing processed food consumption, and encouraging physical activity are key strategies.

These findings underscore the need for comprehensive policies and a holistic approach to address childhood overnutrition. Educating parents about healthy lifestyle choices is crucial to breaking the cycle of overnutrition. Policies should create supportive environments, including urban planning for physical activity, marketing regulations for unhealthy foods, and subsidies for healthy foods.

V. Discussions

The results of the study indicate that Logistic Regression and Support Vector Machines (SVM) are the most reliable classifiers for predicting both overweight and obesity among children. These models showed the highest accuracy, with Logistic Regression and SVM each achieving an accuracy of 0.887 for overweight prediction and 0.967 for obesity prediction. This finding aligns with the results from Lim et al. (2023), who found that Logistic Regression performed best in predicting childhood obesity with an AUC of 0.82, an accuracy of 74%, and a specificity of 73%. Moreover, the study's findings on the high performance of Gradient Boosting and Random Forest are consistent with Gonzalo

et al. (2020), who reported that XGBoost achieved an AUC of 0.86 while Deep Learning models achieved an AUC of 0.84. Both studies emphasize the effectiveness of these models in accurately predicting childhood obesity. Similarly, Siddiqui et al. (2021) reported high accuracy rates for Random Forests and Artificial Neural Networks, which also showed good performance in the present study. Additionally, the emphasis on birth weight, socioeconomic status, and parental education as significant predictors of childhood overweight and obesity is supported by the findings of previous studies. For instance, the importance of early life factors such as birth weight is echoed in the work of Hammond et al. (2019), who identified maternal weight as key risk factor.

While the study found Logistic Regression and SVM to be the most effective models, it contrasts with Pang et al. (2021), who identified XGBoost as the top-performing model with an AUC of 0.81. This discrepancy could be attributed to differences in the datasets and the specific context of Moroccan children, which may not be fully captured by models trained on different populations. Moreover, the study's results diverge from Triantafyllidis et al. (2020), who found Decision Trees and Artificial Neural Networks to be the most effective with an AUC of 0.83. In the present study, Decision Trees and K-Nearest Neighbors showed lower reliability, suggesting that these models might be less suitable for the Moroccan dataset used. Another notable divergence is observed with Naive Bayes. While Gonzalo et al. (2020) and Gray et al. (2019) highlighted the strong performance of gradient boosting and regression models like LASSO, the current study found Naive Bayes to excel in AUC (0.63) despite its lower accuracy (0.758) and specificity (0.117) for overweight prediction. This indicates that while Naive Bayes can effectively distinguish between classes, it may struggle with accurate classification in this context.

The high performance of Logistic Regression and SVM can be attributed to their robustness in handling large datasets and their capability to manage multicollinearity among predictors. This is particularly important given the complex interplay of factors influencing childhood overnutrition, such as birth weight, socioeconomic status, and parental education. The study underscores the significant roles of these factors, consistent with the broader literature on childhood obesity and overweight.

Looking ahead, the study suggests several prospects. Public health interventions should target high-risk regions and specific demographic groups identified in the study. For instance, regions like ”Laayoune-Sakia El Hamra” and ”Beni Mellal-Khenifra” exhibit higher obesity risks, necessitating focused intervention strategies. Additionally, policies aimed at enhancing parental education and improving socioeconomic conditions could significantly mitigate childhood overnutrition. Further research could explore the integration of more advanced machine learning models and hybrid approaches to enhance prediction accuracy. Longitudinal studies are also recommended to assess the long-term effectiveness of public health interventions.

This study offers unique contributions by presenting a comprehensive, context-specific analysis of childhood overnutrition in Morocco. Firstly, the study focuses on Morocco, employing advanced machine learning techniques to provide tailored insights relevant to the local population. Secondly, the study compares ten different machine learning algorithms to identify the best-performing models for predicting childhood overweight and obesity.

However, the dataset from the 2018 national population and family health survey may not represent current population dynamics due to changes in demographics, socioeconomic conditions, and public health policies, introducing temporal bias. Selection biases may also exist if certain regions or demographic groups are underrepresented, leading to skewed results. These biases can significantly impact the study's results. To mitigate these biases, future studies should use more recent and comprehensive datasets reflecting current population dynamics.

We should interpret the study's findings cautiously when formulating public health policies, given these limitations. Policymakers should consider these potential biases and complement the findings with additional research. Addressing these limitations in future research can provide more robust insights, leading to more effective strategies to combat childhood overnutrition in Morocco.

VI. Conclusions

This study provides a comprehensive analysis of various machine learning classifiers to predict overnutrition among children under five years old in Morocco, using data from the national population and family health survey. The study evaluated ten machine learning algorithms. Performance metrics such as accuracy, sensitivity, specificity, kappa statistic (κ), and area under the curve (AUC) were used to assess these models. Logistic Regression and SVM were found to be the most effective for predicting both overweight and obesity, with high accuracy rates of nearly ninety percent for overweight and approximately ninety-seven percent for obesity. These models also showed high sensitivity and specificity. Gradient Boosting and Random Forest performed well, while Naive Bayes, despite lower accuracy, excelled in AUC, indicating its proficiency in class distinction.

This research makes several unique contributions to public health and machine learning. By focusing on Morocco, it addresses a gap in literature dominated by other regions. The study's context-specific analysis is vital for developing targeted interventions. It provides a detailed comparison of ten machine learning models, highlighting their strengths and weaknesses in predicting childhood overnutrition. It also emphasizes the significant roles of birth weight, socioeconomic status, and parental education, which can inform targeted public health strategies.

The findings have important theoretical and managerial implications. Theoretically, it underscores the effectiveness of Logistic Regression and SVM in handling complex datasets and highlights the importance of early life factors, socioeconomic status, and parental education in the framework of childhood overnutrition. For policymakers and public health officials, the study offers actionable insights into effective predictive models and key risk factors, aiding the design of targeted interventions and policies to mitigate childhood overnutrition in Morocco.

Future research should incorporate more recent and longitudinal data to identify trends and causality more accurately. It should also include a broader range of variables, such as genetic factors, dietary habits, physical activity levels, and psychological factors, to enhance the comprehensiveness of predictive models. Exploring advanced machine learning techniques and hybrid approaches could improve predictive accuracy and provide deeper insights into the determinants of childhood overnutrition. Evaluating the impact of specific public health interventions and policies on childhood overnutrition can offer valuable feedback for policymakers.

References

[1] World Health Organization. (2023). Malnutrition: Key facts. https://www.who.int/news-room/fact-sheets/detail/malnutrition

[2] Vucenik Ivana, Stains, Joseph P. (2012). Obesity and cancer risk: evidence, mechanisms, and recommendations. Annals of the New York Academy of Sciences, 1271(1), 37–43. https://doi.org/10.1111/j.1749-6632.2012.06750.x

[3] Agras WS, Hammer LD, McNicholas F, Kraemer HC. (2004). Risk factors for childhood overweight: A prospective study from birth to 9.5 years. J Pediatr 145:20–25. https://doi.org/10.1016/j.jpeds.2004.05.038

[4] Centers for Disease Control and Prevention. (2023). Childhood Obesity Causes and Consequences. https://www.cdc.gov/obesity/childhood/causes.html

[5] Shi X, Nikolic G, Epelde G, et al. (2021). An ensemble-based feature selection framework to select risk factors of childhood obesity for policy decision making. https://doi.org/10.1186/s12911-021-01517-5

[6] Siddiqui JK, Wu S, Blanchard H. (2021). Survey on Machine and Deep Learning Models for Childhood and Adolescent Obesity. https://doi.org/10.1109/ACCESS.2021.3071659

[7] UNICEF, WHO and World Bank. (2021). Levels and Trends in Child Malnutrition: Child Malnutrition Estimates - Key findings of the 2021 edition.

[8] Ministère de la Santé. (2019). Rapport de l'enquête nationale sur la population et la santé familiale de 2018.

[9] Ministère de la Santé. (2011). Rapport de l'enquête nationale sur la population et la santé familiale de 2010-11.

[10] Pang X, Forrest CB, Lê-Scherban F, Masino AJ. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. Int J Obes (Lond). 45:1189–1232.

[11] Lim H, Lee H, Kim J. (2023). A Prediction Model for Childhood Obesity Risk Using Machine Learning: A Panel Study on Korean Children. Scientific Reports 13:10122.

[12] Gonzalo C, Lee D, Kim J. (2020). Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. Nutrients, 12(8):2466.

[13] Gray JC, Schvey NA, Tanofsky-Kraff M. (2019). Demographic, psychological, behavioral, and cognitive correlates of BMI in youth: Findings from the Adolescent Brain Cognitive Development (ABCD) study. Psychol Med 2019;50:1539–1547.

[14] Lebron C, Lee H, Kim J. (2020). Predicting Childhood Obesity Using Machine Learning Techniques. J Pediatr 2020;49:26–30.

[15] Hammond R, Athanasiadou R, Curado S, et al. (2019). Predicting childhood obesity using electronic health records and publicly available data. PLoS One 2019;14:e0215571.

[16] Triantafyllidis AK, Kourou KD, Karatzanis I, et al. (2020). Computerized Decision Support and Machine Learning Applications for Childhood Obesity. Artif Intell Med 2020;104:101844.

[17] Dugan TM, Mukhopadhyay S, Carroll A, Downs S. (2015). Machine learning techniques for prediction of early childhood obesity. Appl Clin Inform 2015;6:506–520.

[18] Fikrewold H Bitew, Corey S Sparks and Samuel H Nyarko. (2021). Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia.

[19] World Health Organization. (2006). Child growth standards.Multi Center Growth Reference Study Group, Genève: OMS, 2006, 336p.

[20] Brownlee, J. (2016). Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End.

[21] Pedregosa F. et al,. (2011). Scikit-learn: Machine Learning in Python.

[22] Hastie T, Tibshirani R and Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer.

[23] Berrouyne M., Hami H. and Jouilil Y. (2024). Predictive Power of AI: Tackling Child Undernutrition in Morocco. The Sixth IEEE International Conference on intelligent Systems and Computer Vision, Fez, Morocco, 2024.