An Obesity Risk Level (ORL) Based on Combination of K-Means and XGboost Algorithms to Predict Childhood Obesity

Ghaidaa Hamed Alharbi, Mohammed Abdulaziz Ikram
Department of Computer Science, Umm Al-Qura University, Makkah, KSA

***Abstract*—Childhood obesity is a common and serious public health problem that requires early prevention measures. Identifying children at risk of obesity is crucial for timely interventions that aim to mitigate these adverse health outcomes. Machine learning (ML) offers powerful tools to predict obesity and related complications using large and diverse data sources. The article uses machine learning (ML) techniques to analyze children's data, focusing on a newly developed variable, the Obesity Risk Level (ORL), which categorizes participants into high, medium, and low risk levels. Two primary models were utilized: the K-Means algorithm for clustering participants based on shared characteristics and XGBoost for predicting the risk level and obesity likelihood. The results showed an overall prediction precision of 88.04%, with high precision, recall, and F1 scores, demonstrating the robustness of the model in identifying obesity risks. This approach provides a data-driven framework to improve health interventions and prevent childhood obesity, providing information that could shape future preventive strategies.**

***Keywords*—Prediction system; Childhood obesity; K-Means; XGBoost; Machine learning**

I. Introduction

Childhood obesity is a global health issue that affects millions of children and adolescents. According to the World Health Organization, the percentage of overweight and obese children and adolescents aged 5 to 19 years increased from 4% in 1975 to 18% in 2016 [1]. Obesity prevention is an important component of public health efforts, and central to the discipline of obesity prevention is the identification of children at risk. Childhood obesity is an increasingly important issue in healthcare, as it can cause many health problems, such as heart disease, diabetes, metabolic syndrome, mental issues, and early death [2]. Therefore, it is essential to prevent and treat childhood obesity. To achieve this, it is necessary to identify which children are likely to become or are already obese, as well as how they are affected by obesity, including factors such as BMI, weight change, obesity complications, and treatment results. This information can help plan effective and timely interventions that address risk factors and promote healthy habits. However, these factors can vary between different groups, locations, and time periods, so the results of one study cannot be universally applied to another.

In recent years, many studies have used different ML methods to predict childhood obesity and related outcomes, using various data sources, characteristics, and targets [3], [7], [12], [15], [19]. However, the issue of personalized risk prediction has recently become more important. Despite these studies, little importance has been given to personalized obesity risk assessments. Although these studies predict obesity or related outcomes, none of them explicitly focus on predicting an individual's personalized risk of developing obesity, which could help with more targeted prevention strategies. This gap highlights the need for models that go beyond general classifications and provide individualized risk assessments.

To overcome this problem, machine learning (ML) methods have become useful tools that can use large and diverse data sources to build accurate and reliable prediction models for childhood obesity and related outcomes. Recent developments in the field of artificial intelligence and machine learning have led to renewed interest in using data-driven approaches to address these complex health challenges. This study builds on previous research on childhood obesity and contributes to a growing body of work using ML to predict obesity outcomes.

The goal of this study is to use ML algorithms to analyze children's data and to design an intervention that can help children who are at risk of obesity or are already obese to improve their health. In this research, an Obesity Risk Level (ORL) was proposed to classify participants into three levels: high, medium, and low. The goal is to analyze the relationship between individual characteristics and obesity, helping to identify those at higher risk. To achieve this, two models were used: the K-Means algorithm to group participants based on shared traits, and XGBoost to predict risk level and classify likelihood of obesity. The results demonstrated an overall precision of 88.04%, with strong performance in metrics such as precision, recall, and F1 score. These findings validate the effectiveness of the model in making precise predictions, helping to identify targeted preventive efforts and health interventions.

The remainder of the paper is organized into six sections, each focusing on a different aspect of the research. Section II presents a detailed overview of the research background and related work. Section III describes the research methodology, including the K-Means clustering process and the use of XGBoost for predictive modelling. Section IV presents the result of the experiment. The discussion part is presented in Section V. Finally, Section VI concludes with a summary.

II. Review of the Literature

A. Related Works

In this review of the literature, current research on childhood obesity and the various methods used to predict and analyze it is explored in depth. A considerable amount of literature has been published on factors that contribute to childhood obesity, including genetics, environmental influences, and lifestyle choices [14], [17], [22-23]. Additionally, advanced machine learning techniques have been employed to estimate obesity levels and predict risk using clinical or behavioral data [4-5].

However, relatively little literature has been published on the application of advanced machine learning techniques specifically tailored for the prediction of childhood obesity. In addition, different predictive modelling techniques, such as logistic regression, decision trees, and artificial neural networks, used by researchers to forecast obesity trends are also examined. By examining the existing research, the objective is to uncover knowledge gaps and deliver insights that may lead to improved strategies for preventing and managing childhood obesity.

Among recent efforts to predict childhood obesity using advanced machine learning techniques, Gupta et al. [6] developed a deep neural network architecture based on recurrent neural networks (RNNs). Their study specifically investigated how temporal patterns in pediatric health data could be leveraged to forecast obesity risk over one, two, and three years intervals. They used RNNs with LSTM (long-short-term memory) cells combined with a separate feedforward network for training their model. They trained the models using a large pediatric electronic health record (EHR) data set. They achieved an AUC of 0.80, 0.93, and 0.92 for a 3-year window at 5, 11, and 18 years. They also compared the recurrent model with other machine learning models for the task of predicting childhood obesity and found the LSTM-based model demonstrates better performance compared to traditional machine learning models that ignore the temporality of the data by aggregating the data.

Similarly, in a separate study by Robert Hammond et al. [10] conducted a study in which they used EHR data to predict obesity at five years of age with an AUC similar to cohort studies. They applied different machine learning algorithms for binary classification and regression tasks. They also created separate models for boys and girls. They discovered that the most important prediction characteristics were the weight of the length z score, the BMI from 19 to 24 months, and the last BMI value before age two in each of the models. The best models achieved an AUC of 81.7% for girls and 76.1% for boys in predicting obesity. Their findings indicate that machine learning methods can use EHR data to predict future childhood obesity and help clinicians and researchers design better interventions, policies, and clinical decisions.

In the study by Xueqin Pang et al. [12] seven machine learning models have been developed to predict childhood obesity between the ages of 2 and 7 years using data from the Electronic Healthcare Record (EHR). Furthermore, these studies relied on machine learning algorithms to assess and predict obesity. Furthermore, Cheong Kim et al. [8] applied GBN-MB along with several other algorithms, such as GBN, LR, DT, SVM, and NN, as part of a proof of concept to analyze public health and simulate future outcomes through What-If analysis. Xiaolu Cheng et al. [9] focused on understanding the relationship between physical activity and weight status, using data from NHANES (2003–2006) and evaluating 11 classification algorithms, including logistic regression, k-NN, RBF, and J48.

Faria Ferdowsy et al. [11] applied nine different ML algorithms, including k-NN, random forest, and logistic regression, to classify obesity levels (high, medium, and low) in a dataset of over 1100 individuals. Each of these studies explored different ML models tailored to their research objectives. Cheong Kim et al. [8] demonstrated that the GBN-MB model produced the best results in simulating health outcomes and guiding public health professionals. Similarly, Xiaolu Cheng et al. [9] found that physical activity was a key predictor of weight status, with machine learning models offering valuable insights into demographic factors such as age, gender, and race/ethnicity.

In summary, the reviewed studies illustrate the diversity of approaches for predicting and estimating obesity, ranging from Bayesian-optimized neural networks to decision trees, Bayesian networks, and deep learning models. Ultimately, while each method has its strengths and limitations, the growing trend is towards utilizing electronic health records, integrating multiple data sources, and developing interpretable models for clinical and public health applications. Looking ahead, the future of obesity prediction research will likely focus on improving the balance between precision, interpretability, and practical applicability, particularly in pediatric and adolescent populations where early intervention is critical.

B. Machine Learning Models

Machine learning methods, such as Decision Trees, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVC), Logistic Regression, Random Forest, AdaBoost, XGBoost, and K-Means, are powerful tools for developing predictive models that help identify factors that contribute to childhood obesity. Although many studies have demonstrated the effectiveness of these models in various healthcare applications, relatively few have focused on their specific use in the prediction of childhood obesity. This gap highlights a critical area for future research.

These models are capable of evaluating complex relationships between variables such as genetics, lifestyle, environmental influences, and diet habits. Numerous studies have demonstrated that machine learning methods are highly efficient in handling large datasets, revealing hidden patterns that traditional statistical methods may overlook. For example, models such as Random Forest and XGBoost have been shown to significantly improve the accuracy of obesity-related predictions. Using these advanced algorithms, researchers aim to identify high-risk groups for childhood obesity, allowing the development of targeted and effective preventive strategies. Furthermore, these models provide deeper insights into how specific factors influence obesity, allowing for more personalized interventions customized to individual needs.

Although much of the current research has focused on improving predictive accuracy, there is a growing emphasis on

ensuring that these models are interpretable and actionable. Making models more understandable will allow healthcare professionals to apply findings in practical ways, improving resource allocation for the prevention and management of obesity in communities. Early identification of risk factors through machine learning has been shown to improve outcomes by enabling timely interventions.

Incorporating data from various sources, such as electronic health records (EHR) and lifestyle surveys, is expected to further improve the ability of these models to predict obesity risks. By analyzing a wide range of factors and sources, machine learning can support the design of more comprehensive and effective strategies to mitigate the increasing rates of childhood obesity.

C. Obesity Childhood

Childhood obesity is a global health problem characterized by excess body fat that significantly affects both physical and mental well-being. The following are key points regarding childhood obesity.

1) Definition and measurement: Childhood obesity is commonly assessed using the body mass index (BMI), a tool used to determine whether a child's weight is appropriate for their age and height [1].

2) Health risks: Being obese as a child can cause many health problems, such as heart disease, diabetes, metabolic syndrome, mental problems, and early death [2].

3) Causes: Multiple factors contribute to childhood obesity, including poor nutrition, lack of physical activity, genetic predisposition, and environmental influences such as access to healthy food and safe spaces for exercise [20].

4) Prevention and management: Addressing childhood obesity requires a holistic approach that incorporates healthy diet changes, increased physical activity, and behavioral modifications. Family involvement and community support are essential to create environments that promote healthy living [21].

D. Obesity Factors

1) Genetic: Genetics plays a crucial role in shaping children's likelihood of developing obesity. Genes that regulate appetite, metabolism, and fat storage increase the risk of obesity, particularly in children with a family history of the disease. Studies have found that somewhere between 25% and 40% of your BMI is actually inherited [17].

2) Physical activity: A sedentary lifestyle is widely recognized as one of the key factors strongly associated with the increase in obesity rates. This lifestyle is characterized by prolonged periods of inactivity, such as sitting for long hours while watching television or participating in other screen-based activities. Research has shown that each additional hour spent watching television per day can increase the likelihood of developing obesity by 2% [17]. Over time, these seemingly small increases can accumulate, contributing to significant health risks, as reduced physical activity leads to lower energy expenditure and, consequently, weight gain.

3) Dietary habits: There is a strong connection between childhood obesity and eating habits [22]. Choosing nutrients dense foods and maintaining balanced eating patterns are essential to prevent obesity. Proper portion control and reducing high-calorie and low-nutrient foods can significantly reduce the risk of childhood obesity.

4) Environmental factors: Children who live in unsafe areas or who do not have access to well lit, safe walking routes have fewer opportunities to be physically active [17]. In interviews conducted by Jenny Veitch et al. [23], the most commonly reported factor affecting where children played was parents' concerns about their child's safety, with 94% of parents expressing this concern. Safety was a crucial factor in parents' decisions about where their children could play. Therefore, the availability of a safe neighborhood was directly related to increased opportunities for children to engage in active free play.

5) Sleep hours: Poor sleep and sleep disturbances are associated with weight gain in children. A study by Christopher Magee et al. [24] suggests that poor sleep may be a contributing factor to childhood obesity.

6) Socioeconomic Status (SES): Robert Rogers [25] stated that their findings suggest that low SES plays a more significant role in the nation's childhood obesity epidemic than any other demographic factors.

Childhood obesity is a multifaceted problem influenced by a combination of genetic factors, environmental conditions, and lifestyle choices. Genetic factors may contribute to childhood obesity, but environmental factors, such as access to healthy foods and safe recreational spaces, and lifestyle habits, such as eating patterns and levels of physical activity, are equally important. To effectively address this complex problem, parents must provide ongoing emotional support to their children, regardless of their weight. Parents must instead focus on creating a supportive and positive home environment that encourages everyone to eat healthy and be physically active regularly. By promoting these habits and spreading them throughout the family, parents can help reduce the risk of obesity and support children in achieving and maintaining a healthy weight.

III. Methodology

In this part of the article, details of the methods used to analyze and model childhood obesity are provided. The explanation begins with the data collection process, including the sources and characteristics of the dataset used. Next, the applied methodologies are described, starting with K-Means clustering to group data based on similarities, followed by XGBoost for predictive modelling. The section also covers the tools and software utilized and addresses any technical challenges encountered during implementation. Finally, the evaluation metrics used to assess model performance are discussed. This methodology provides a comprehensive approach to understanding and predicting childhood obesity based on the data analyzed.

A. Dataset

The study is based on a comprehensive dataset gathered by a university [13], reflecting a wide range of student profiles from various schools. In collaboration with school administrations, surveys collected data on demographic characteristics and anthropometric measurements. It is noteworthy that a subset of these data, specifically involving 411 identified students with 15 variables, will be utilized in the study. This carefully selected dataset allows for a detailed analysis, contributing to a comprehensive understanding of the educational context and the relationships between various factors. Table I outlines the features used in this study along with their detailed descriptions. Each feature was selected based on its relevance and contribution to the objectives of the research. The descriptions provided help to clarify the significance of these variables, offering a clearer understanding of how they interact and influence the overall analysis. When exploring these characteristics, deeper insights can be gained into the factors that shape the outcomes observed in the data, making them essential for drawing informed conclusions from the study.

TABLE I. DATASET FEATURES

Parameter	Description
St_Height	Height of the student in centimeters
St_Weight	Weight of the student in kilograms
M_Weight	Mother's weight in kilograms
M_Height	Mother's height in centimeters
F_Weight	Father's weight in kilograms
F_Height	Father's height in Centimeters
Appearance_self	Body image perception, rated 1 to 5
Body_pride	Body self-esteem, rated 1 to 5
Neighborhood_jog_safe	Safety of jogging in the neighborhood, rated 1 to 5
Neighborhood_bike_safe	Safety of cycling in the neighborhood, rated 1 to 5
Family_income	Family income level, rated 1 to 8
Gender	Student's gender
Birthday	Student's date of birth
Obese_Student	Obesity classification (binary)
S_BMI	Student's BMI category

These variables, which cover a diverse range of information and survey responses, serve as the foundation for the study's analysis. They offer a comprehensive snapshot of the students' profiles, facilitating a nuanced examination of the factors under scrutiny.

B. Prediction System

First: K-means algorithm

The K-Means algorithm is a common technique in machine learning to group data into groups (clusters) so that the data within each group are as similar as possible [16].

K-Means is a clustering algorithm that partitions a dataset into K distinct clusters. Here is a simplified explanation of the steps.

1) Specify the number of clusters(k): Determine the number of clusters into which you want to divide the data. This number is called k.

2) Initialise the cluster centre: Randomly select k points from the data as the initial cluster center.

3) Assign data points to clusters: For each data point, calculate the distance between it and the center of each cluster. Assign points to the nearest cluster center.

4) Update cluster centres: After assigning points to clusters, recalculate the centers of each cluster based on the points assigned to each cluster.

5) Repeat: Repeat steps 3 and 4 until the center is stable and does not change significantly.

6) Complete: When the center is stable, the algorithm is complete and the clusters are formed.

Objective Function:

$$Minimize \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \quad (1)$$

where,

- k is the number of clusters.
- C_i represents the i-it cluster.
- x is a data point.
- μi is the centroid of the cluster C_i

As shown in Fig. 1, the plot on the left shows the data before applying K-Means, where the points are ungrouped and displayed in uniform color, indicating that no clustering has been applied. In contrast, the plot on the right shows the data after K-Means clustering, where the data points are grouped into distinct clusters, each represented by a different color. The red stars indicate the centroids of each cluster, which are the central points of each group.

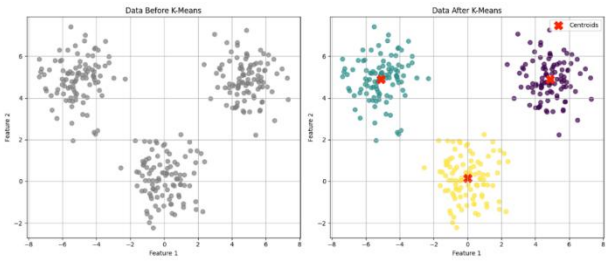

Fig. 1. Data visualisation before and after applying the k-means clustering algorithm.

C. Implementation

The K-Means algorithm has been successfully applied to the data with K=80. A new column named "Cluster" has been added, showing which cluster each student belongs to. Defining K=80; this means that we want to divide the students into 80 groups. Each group represents students with similar characteristics according to the specified features.

1) Choose the columns for clustering: To perform K-Means, we have used the columns of the data because they

reflect important factors such as parents' weight and height, personal behavior, environment, and family income.

2) Scaling: Since the columns contain values of different scales (such as family income vs. height), we scaled them using Standard Scaler to ensure that each feature contributes equally. Scaling means converting the values to a uniform range so that each column has a mean of 0 and a standard deviation of 1.

3) Initialise clusters: The initial centers are chosen randomly for 80 clusters.

Assign points to clusters: For each student, the distance between their data (such as mother's weight, father's height, family income, etc.) and each of the cluster centers is calculated.

The student is assigned to the nearest cluster.

1) Update cluster centres: A new center is calculated for each cluster based on the average values of all students within it.

Iteration: The previous two steps are repeated until the clusters stabilize (i.e., there is no significant change in the assignment of students).

a) Output (group column)

A new column named "Cluster" has been added to the data. This column contains the cluster number to which each student belongs (from 0 to 79, since we set K=80).

b) Example of results

- Student 1: Data: Mother's weight 52.4, Mother's height 157, Father's weight 82... etc.
- Assigned to group 78.
- Student 2: Data: Mother's weight 55, Mother's height 172, Father's weight 90, etc.

Assigned to group 3.

Second: XGBoost

XGBoost is a powerful and efficient tool to boost tree-based models. Moreover, it has been improved to become highly scalable, which is why it is widely used in various machine learning applications [18]. In fact, XGBoost stands out as a highly effective and popular algorithm for this purpose [18]. Specifically, it improves on traditional tree boosting by building multiple models sequentially, where each model corrects the errors of its predecessors to enhance overall performance. XGBoost addresses these issues with several key innovations. First, it incorporates a regularization term into the objective function to reduce model complexity and prevent overfitting. Additionally, it efficiently handles missing values and sparse data, which is crucial for working with large and incomplete datasets. In terms of speed and scalability, XGBoost employs advanced parallelization techniques, processing feature blocks in parallel and distributing tasks across multiple processors. Moreover, it optimizes memory access patterns to reduce unnecessary operations, thereby enhancing performance with massive datasets. Furthermore, XGBoost can handle datasets larger than memory capacity by leveraging external storage efficiently. As a result, the strengths of XGBoost include its impressive speed and efficiency, achieved through effective parallelization and optimized memory usage. In addition, its accuracy and effectiveness are enhanced by regularization and second-order optimization techniques.

Consequently, XGBoost is widely used across various fields due to its ability to manage large and sparse data effectively. In summary, XGBoost is a powerful and scalable tool for boosting tree-based models, offering significant improvements in performance and applicability in machine learning. To conclude, XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm designed for efficient and scalable decision tree boosting. Specifically, it improves traditional gradient boosting by incorporating advanced techniques such as regularization to prevent overfitting, parallel computation for speed, and handling of sparse data for improved performance on large and complex datasets.

c) XGBoost Equation and Explanation: The XGBoost equation is based on the gradient boosting technique with several optimizations to make it more efficient and effective. The main idea is to gradually create decision trees that correct the errors made by previous trees, and these corrections are combined to form the final prediction.

d) XGBoost Equation: Let us assume that there are K decision trees. The prediction of the model for any sample x_i is calculated as follows:

$$\hat{y}_i = \sum f_k(x_i) \quad (2)$$

where,

- $\hat{y}_i$ is the final prediction for the sample x_i.
- $f_k(x_i)$ is the function representing the k -th decision tree, which produces a prediction for sample x_i.
- K is the number of trees.

e) Objective Function: The model is optimized by minimizing the objective function, which consists of two parts:

Loss function: Measures the difference between the predictions and the actual values. The commonly used loss function is the squared error for regression problems or the logistic loss for classification problems.

$$L(\hat{y}, y) = \sum l(y_i, \hat{y}_i) \quad (3)$$

where $l(y_i, \hat{y}_i)$ is the loss function that measures the difference between the prediction $\hat{y}\_i$ and the actual value y_i for sample i.

Regularization term: Helps to control the complexity of the model and reduce overfitting by penalizing the number of trees and their complexity.

$$\Omega(f) = \gamma T + (1/2)\,\lambda \sum w_j \quad (4)$$

where,

γ is the parameter that penalizes adding new nodes to the tree.

T is the number of nodes in the tree.

λ is the regularization parameter that controls the size of the weights associated with the nodes.

w_j is the weight associated with each node in the tree.

f) Final Objective Function:

$$Objective = \sum l(y_i, \hat{y}_i) + \Sigma \Omega(f_k) \quad (5)$$

where,

The first part $\Sigma\, l(y_i, \hat{y}_i)$ represents the total loss over all the samples. The second part $\Sigma\, \Omega(f_k)$ is the sum of the regularization terms for each tree, limiting the model's complexity.

g) How XGBoost Optimizes: XGBoost builds trees progressively, updating predictions at each step. This is done using Gradient Descent, where the first and second derivatives of the loss function are computed to optimize the model gradually:

First Derivative (Gradient): Represents the rate of change of the loss with respect to the prediction.

$$g\_i = \partial L(y_i, \hat{y}_i) / \partial \hat{y}_i \quad (6)$$

Second derivative (Hessian): Represents the rate of change of the first derivative (used to speed up the gradient calculation).

$$h\_i = \partial^2 L(y_i, \hat{y}_i) / \partial \hat{y}_{i^2} \quad (7)$$

h) Problem Setup: Aim to classify whether a student is obese ("obese student" = 1) using various features.

1. Input data:

The sample contains the following columns:

Features: q5, q6, q11, q12, s_q7_1, s_q7_3, s_q8_1, s_q8_2, q16, gender, cluster.

Target: obese_student

2. Objective of prediction:

The goal is to determine whether a student is obese (obese_student = 1) or not (obese_student = 0) based on the input features.

3. How the prediction works:

The XGBoost model predicts each feature taking it as input to the model. Based on the relationships discovered during training, the model calculates a probability score to classify the student as "obese" or "non-obese."

4. Example:

Features:

q5 = 80.0, q6 = 165.0, q11 = 55.0, q12 = 175.0, s_q7_1 = 4, s_q7_3 = 3, s_q8_1 = 4, s_q8_2 = 4, q16 = 3, gender = 0, Cluster = 9

Steps:

After combining tree contributions, the final score might be $\hat{y} = -0.3$.

Applying the Sigmoid function:

$$P(obese) = \frac{1}{1 + e^{-(-0.3)}} \approx 0.43$$

Since $P(obese) < 0.5$, the model predicts obese_student = 0 (non-obese).

D. Integrating K-Means and XGBoost

Combining K-Means with XGBoost presents a robust framework for anomaly detection in logarithmic data sets, leading to accurate and fast results with fewer errors [26]. It leverages the strengths of both algorithms to improve prediction accuracy and model performance.

1) Purpose of Combining K-Means and XGBoost: One effective method of combining K-Means and XGBoost is to use K-Means as a preprocessing step. In this approach, K-Means clusters the data into groups based on shared characteristics, allowing for the identification of patterns in the data. These cluster labels are then used as additional features for the XGBoost model, improving its ability to make predictions. By integrating clustering before applying XGBoost, the model can capture more nuanced relationships between variables, leading to more accurate predictions, particularly in complex datasets like those related to childhood obesity.

2) Methods to Combine K-Means and XGBoost: K-Means Clustering: Initially, K-Means is applied to group data into distinct clusters based on the similarity of data points. The primary objective of K-Means is to partition the data set into cohesive clusters where the data points within each cluster are similar and close to each other.

Utilizing Cluster Information as a Feature: Once K-Means has assigned clusters, the cluster label (the number assigned to each data point's cluster) is incorporated as an additional feature in the dataset. This enhanced data set, which now contains cluster information, is then fed into the XGBoost model. The added cluster labels provide the model with valuable insights about the underlying structure of the data, potentially improving the model's predictive accuracy.

XGBoost for Prediction: After enriching the dataset with cluster information, XGBoost is used for either classification or regression tasks. As a powerful and widely adopted tree-boosting algorithm, XGBoost can utilize the cluster feature to better capture patterns and relationships in the data, leading to more precise predictions.

This approach of combining K-Means clustering with XGBoost preprocessing helps improve model performance by adding a layer of cluster-based structure to the data.

As shown in Fig. 2, the process begins with applying K- Means clustering to the raw data, generating cluster labels. These labels are then combined with the original data, creating an enhanced dataset. Finally, the enhanced dataset is used as input for the XGBoost model to produce the prediction results.

Flow diagram:

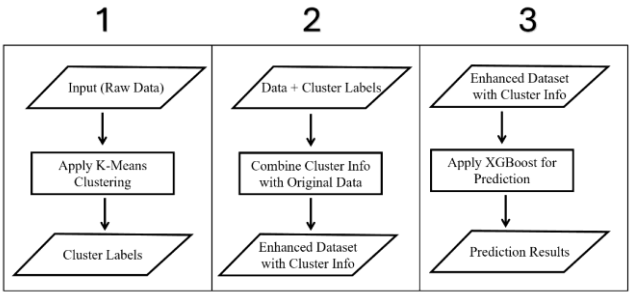

Fig. 2. Integration of K-Means clustering and XGBoost for predictive modelling.

3) Benefits and Challenges

a) Benefits: Improved accuracy: By clustering data first, XGBoost can make more accurate predictions within each group.

Discovering Hidden Patterns: Clustering can reveal hidden patterns that were not visible in the raw data, allowing XGBoost to make better-informed predictions.

b) Challenges: Choosing the right number of clusters: Deciding how many clusters to use in K-Means can be difficult and may require testing different values to find the best result.

Increased computational complexity: Combining K-Means and XGBoost can be computationally expensive, especially for large datasets, as it requires running both clustering and multiple model training processes.

Combining K-Means with XGBoost is an effective strategy to improve predictions in complex and heterogeneous data sets. K-Means helps to organise the data and uncover hidden patterns, while XGBoost uses this information to make more accurate predictions. This approach is particularly useful in domains such as marketing, healthcare, and finance, where segmenting data can lead to more targeted and effective models.

E. Clustering with K-Means

K-Means clustering was employed to group the data into 80 clusters. During the experimentation phase, several different values for the number of clusters to determine the optimal configuration for the data. Through this process, it became clear that 80 clusters offered the best balance between maintaining a manageable level of complexity and achieving a high level of classification accuracy. Therefore, after careful consideration and testing, 80 clusters were chosen as the ideal number, as they effectively captured the nuances in the data while optimizing the overall performance of the model. This thorough approach is one of the main reasons for settling on 80 clusters. Each cluster was analyzed for the percentage of obese students, leading to the classification of clusters into risk levels (high, medium, low). The data was then annotated with these risk levels.

Risk Level Classification (Obesity Risk Level (ORL)):

$$ORL \begin{cases} \text{High Risk} & \text{if Percentage of Obese Students} \geq 70\% \\ \text{Medium Risk} & \text{if } 40\% \leq \text{Percentage of Obese Students} < 70\% \\ \text{Low Risk} & \text{if Percentage of Obese Students} < 40\% \end{cases}$$

The classification of risk level based on the percentage of obese students is designed to reflect the severity of the health issue and highlight the influence of surrounding factors on obesity rates. Here is why it is divided in this way:

High Risk (70%): When the percentage of obese students is very high (greater than or equal to 70%), it indicates a severe issue, suggesting that the surrounding factors—such as those studied in this research, including parental weight and height, neighborhood safety, monthly income, and psychological factors—have a significant impact on individuals, potentially leading to a higher likelihood of obesity. In this case, these factors seem to play a major role in shaping the health outcomes of students.

Medium Risk ($40\% \leq x < 70\%$): At this level, the obesity rate is moderate, indicating that surrounding factors still have a considerable influence on students. This means that while not as severe as in the "high-risk" category, these environmental, social, and psychological factors still contribute to obesity, but perhaps to a lesser extent. Therefore, it is essential to address these factors to prevent further escalation.

Low Risk (<40%): When the percentage of obese students is below 40%, it suggests that the impact of surrounding factors is relatively less significant in leading to obesity. However, preventive action is still crucial, as factors such as parental characteristics, neighborhood conditions, and psychological well-being can still play a role in maintaining or reducing obesity rates over time.

To conclude, the higher the percentage of obese children in each risk category, the greater the influence of surrounding factors, such as those explored in this study, on individuals, potentially leading to obesity. This classification underscores the need for targeted interventions addressing these factors to mitigate obesity rates at different risk levels.

F. Modelling with XGBoost

After that, the XGBoost algorithm is applied to train the model in the dataset. XGBoost is one of the most powerful machine learning algorithms and is based on the Gradient Boosting technique to progressively build multiple trees. Each tree corrects the errors made by the previous ones and, through this iterative process, the accuracy of the predictions improves with every new tree added to the model.

G. Steps to Analyse and Classify Student Obesity Risk

This section provides a structured overview of the process used to analyze student data to identify and classify risk levels for obesity. The workflow includes data loading, preprocessing, exploratory analysis, data balancing, and the application of the K-Means clustering algorithm, culminating in user-specific obesity risk assessment.

1) These steps cover the entire process from data loading to user-specific risk assessment.

a) Data Loading and Exploration:

- Load Data: Import data from a CSV file and examine the structure, including data types and missing values.

- Exploration: Review the dataset to understand its composition and identify potential issues such as missing or incorrect values.

b) Data Processing and Transformation:

- Text to Numeric Conversion: Convert text data into a numerical format for easier processing.
- Data cleaning: Remove unnecessary data points and create new relevant features.

c) Exploratory Analysis:

- Correlation Analysis: Compute and display the correlation matrix between features to identify relationships between variables.
- Outlier Detection: Use the interquartile range (IQR) method to detect outliers in numerical features.
- Box plot: Visualize the distribution of numerical data and identify outliers.
- Bar plot: Plot categorical data such as gender and the presence of obesity to understand distributions.

d) Data Balancing:

- Apply SMOTE (Synthetic Minority Over-sampling Technique): Balance the dataset by increasing the number of samples in underrepresented categories.
- Post-SMOTE Analysis: Display the dataset distribution after applying SMOTE to ensure balance.

e) Clustering of K-Means:

- Cluster Formation: Apply the K-Means algorithm with 80 clusters to the weighted data of 694 samples. Through experimentation, 80 was determined to be the optimal number of clusters.
- Obesity Percentage Calculation: Calculate the percentage of obese students within each cluster and determine the corresponding risk levels.

f) Report and Analyze Results:

- Summary: Provide a detailed summary of the number of obese students and their percentages at different risk levels.
- Cluster Classification: Classify the clusters into predefined risk levels (High, Medium, Low) and display the data accordingly.

g) User-Specific Risk Assessment:

- Data input: Collect input data from the user.
- Risk Level Determination: Based on user input, classify your risk level of obesity using the K-Means model.

The objective of this process is to thoroughly analyze student data to uncover obesity patterns and identify risk levels. This involves examining various factors, such as birth date, sex, and other characteristics. By applying K-Means clustering, students are grouped based on similar characteristics, allowing us to better understand obesity trends and provide actionable insights for intervention.

2) Apply XGBoost: Once the data have been clustered, the next step is to develop a predictive model using the XGBoost algorithm. XGBoost is known for its high performance, making it ideal for handling large datasets and complex relationships.

a) Data Preparation:

Input data: Use the clustered data from the K-Means algorithm. The data, now organised into clusters, helps to reveal deeper relationships and structures.

b) Model Training:

Train XGBoost: Train the XGBoost model on the clustered dataset, feeding the input features and their corresponding target labels to learn patterns and predict results.

c) Hyperparameter Tuning:

Optimise model: Improve the performance of the model by fine-tuning key hyperparameters such as the learning rate, maximum depth, and number of estimators. This can be done using methods such as grid search or random search.

d) Model Evaluation:

Performance metrics: Evaluate the model using metrics such as precision, precision, recall, and F1 score to determine how well the model predicts obesity risk levels.

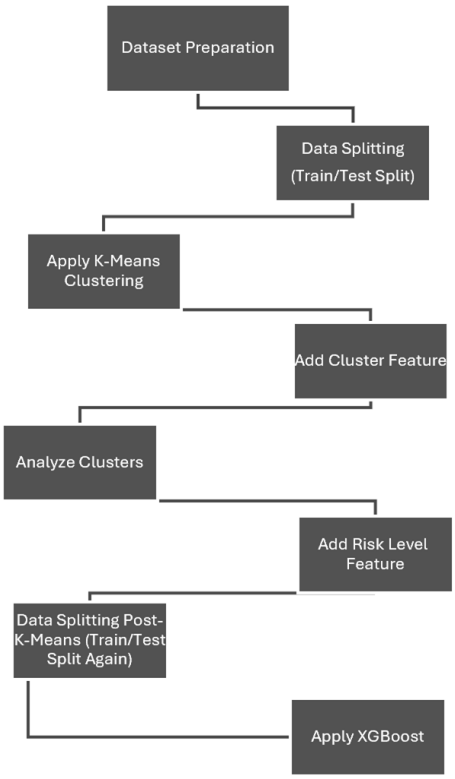

Fig. 3. Flowchart of the analysis process using K-Means and XGBoost.

As shown in Fig. 3, the process of applying K-Means clustering followed by XGBoost in the analysis of obesity risk. The flowchart highlights key steps such as data preparation, clustering, risk level classification, and finally, the use of XGBoost to build a robust predictive model. The diagram provides a clear view of how clustering is integrated with predictive modelling to assess the risks of obesity in students.

IV. Experiments and Results

The primary objective of this analysis was to determine whether it is possible to predict childhood obesity based on environmental factors. The study aims to explore the relationship between a child's environment and the probability of obesity by applying machine learning models such as K- Means for clustering and XGBoost for predictive modelling.

A. Overview of the Dataset

The dataset used for this study comprised 411 identified students, including 15 key variables related to both the child's physical characteristics and environmental factors. After applying SMOTE to balance the dataset, the number of samples increased to 694. The dataset includes the following key variables:

Mother's Weight (M_Weight), Mother's Height (M_Height), Father's Weight (F_Weight), Father's Height (F_Height), Appearance self-assessment (appearance_self), Body pride (body_pride), Neighborhood safety for jogging (neighborhood_jog_safe), Neighborhood safety for biking (neighborhood_bike_safe), Family's monthly income (family_income), Gender, Birthday, Obese Student (Obese_Student)

B. Applied Models

The analysis involved the use of two main models:

1) K-Means for clustering students based on similarities in their features.

2) XGBoost for building predictive models based on the generated clusters.

C. Performance Metrics

The XGBoost model was evaluated based on several performance metrics that provided insight into its effectiveness in predicting obesity risk. The model achieved the following metrics: As seen in Table II, the performance metrics for the obesity classification model are outlined, including Precision, Recall, F1-score, and Support for both non-obese (Class 0.0) and obese (Class 1.0) categories. These metrics provide a detailed evaluation of the model's ability to accurately classify obesity.

TABLE II. Performance Metrics for Obesity Classification Model

Class	Precision	Recall	F1-score	Support
0 (Not Obese)	0.93	0.85	0.89	118
1 (Obese)	0.82	0.92	0.87	91

As seen in Table III, the macro and weighted average performance metrics of the obesity classification model are presented. The model achieves a macro average precision of 0.88, recall of 0.89, and an F1-score of 0.88. Similarly, the weighted averages show strong performance, with precision at 0.89, recall at 0.88, and an F1-score of 0.88. The overall accuracy of the model is 88.04%, reflecting its reliability in predicting obesity across the data set.

These results indicate strong performance across all metrics, validating the model's ability to accurately predict childhood obesity based on environmental factors.

TABLE III. Overall Performance Metrics for Obesity Classification Model

Metric	Value
Macro Average Precision	0.88
Macro Average Recall	0.89
Macro Average F1-score	0.88
Weighted Average Precision	0.89
Weighted Average Recall	0.88
Weighted Average F1-score	0.88
Overall Accuracy	88.04%

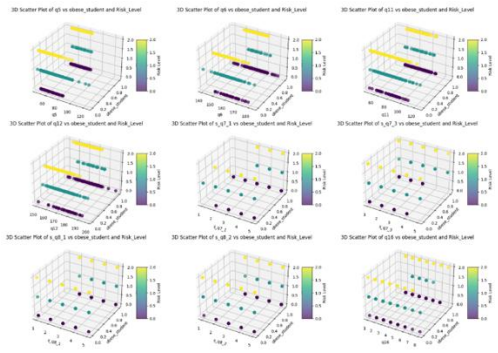

Fig. 4. 3D scatter plots showing the relationship between various characteristics, obesity, and risk levels.

The results of the analysis are supported by 3D scatter plots as seen in the Fig. 4. These graphs visually demonstrate the relationship between different characteristics such as parental weight, neighborhood safety, and child's risk of obesity. The color-coded risk levels (purple, teal, yellow) highlight distinct clusters where, environmental factors align with the likelihood of obesity. This representation makes it easier to identify how certain characteristics influence the risk levels.

D. Risk Level Categorisation

As seen in Table IV, the clusters are categorized by their risk levels based on the proportion of obese individuals in each group. The table includes the cluster number, the number of obese individuals, the percentage of obese individuals, and the associated risk level. Clusters with a higher percentage of obesity are classified as high-risk, while those with lower percentages are categorized as 'medium risk' or 'low risk.' This classification helps identify obesity prevalence within different clusters, providing insight into the varying levels of risk across the groups.

This section provides a detailed breakdown of obesity rates among children, classified into high-, medium-, and low-risk clusters. Each row in the table represents a cluster, detailing the number of obese children, the percentage of obese children within that cluster, and the associated risk level.

1) High risk: Clusters in the high-risk category exhibit a significantly high percentage of obese children. Key points include:

Clusters 18, 62, and 41 show a 100% obesity rate, indicating that all children in these clusters are classified as obese. Clusters 14 and 25 demonstrate high obesity rates of 92.86% and 84.62%,

respectively. The lowest percentage within this category is 70% (Cluster 40), which still indicates a substantial proportion of obese children. These findings suggest that in high-risk groups, most children are obese, with some groups having all children classified as obese.

TABLE IV. ALL RISK LEVELS CLUSTERS

Cluster	Number_of_Obese	Percentage_of_Obese	Risk_Level
18	12	100	high risk
62	3	100	high risk
41	5	100	high risk
35	6	100	high risk
14	13	92.86	high risk
…	…	…	…
22	10	66.67	medium risk
15	6	66.67	medium risk
78	4	66.67	medium risk
7	8	66.67	medium risk
…	…	…	…
58	1	12.5	low risk
21	0	0	low risk
67	0	0	low risk
24	0	0	low risk
…	…	…	…

2) Medium risk: Clusters in the medium-risk category have lower obesity rates compared to high-risk clusters, but the rates are still significant:

Clusters 22, 15, and 78 show an obesity rate of 66.67%, which means that approximately two-thirds of the children in these clusters are obese. Clusters such as 68 and 65 show obesity rates of around 64.29% and 62.50%, respectively. The lowest obesity rate within this category is 42.86% (Clusters 13, 77, and 27).

This category indicates a moderate level of obesity, with a substantial proportion of children classified as obese, though less prevalent than in the high-risk clusters.

3) Low Risk: In the low-risk category, obesity rates are significantly lower:

Clusters such as 45, 16, and 73 show a 33.33% obesity rate, which means that only one-third of the children in these clusters are obese. Some clusters, such as Cluster 79, have an obesity rate as low as 15.38%. Several clusters (eg, clusters 21, 67, and 24) show a 0% obesity rate, indicating that there are no obese children in these clusters. These clusters display minimal levels of obesity, and many clusters having no children classified as obese at all.

E. Obesity by Risk Levels

As seen in Table V, the summary of obesity by risk levels provides an overview of the number of obese individuals, total individuals, and the percentage of obese individuals within each risk category. The high-risk group exhibits the highest percentage of obese individuals, at 80.63%, followed by the medium-risk group at 56.16%, and the low-risk group at 12.18%.

TABLE V. OBESITY SUMMARY BY RISK LEVEL

Risk_Level	Number_of_Obese	Total_Number	Percentage_of_Obese
high-risk	204	253	80.63
medium risk	114	203	56.16
low risk	29	238	12.18

This summary highlights the varying prevalence of obesity at different risk levels, with obesity rates much higher in high-risk groups compared to medium and low-risk groups.

The results confirmed the initial hypothesis that a child's surrounding environment plays a significant role in their likelihood of becoming obese. Factors such as parental weight, neighborhood safety, and socioeconomic status (as indicated by family income) were found to be closely related to obesity risk levels. This finding is consistent with previous research, confirming the importance of these environmental influences on childhood obesity.

The results were generally consistent with previous studies, which also emphasized the role of environmental and familial factors in determining the risk of obesity. However, the integration of machine learning techniques, such as K-Means and XGBoost, provided a more refined predictive model with high precision in this study.

V. DISCUSSION

The purpose of this chapter is to interpret and discuss the results obtained from the analysis of childhood obesity based on environmental factors. By applying machine learning techniques such as K-Means for clustering and XGBoost for predictive modelling, this study aimed to uncover patterns and relationships between various factors that influence childhood obesity. The discussion section will evaluate the performance of these models, explore the significance of key findings, and highlight areas for further improvement and future research.

A. Model Performance

The models showed strong performance, especially in terms of precision, recall, and F1 scores. The XGBoost model, in particular, was highly effective in predicting obesity risks, achieving an overall accuracy of 88.04%. These findings align with existing research, reinforcing the efficacy of machine learning techniques in predicting childhood obesity based on environmental factors.

B. Insights from the Study

The experiment confirmed that various factors, such as parental weight, neighborhood safety, and family income, significantly impact childhood obesity. The models provided valuable information on these relationships. However, there are other important factors, such as duration of sleep, medical conditions, and diet patterns that could further refine the predictions of obesity if included in future research. Despite

promising results, including an accuracy of 88.04%, it is important to consider that there are other factors that influence obesity that were not included in this study. For example, factors such as sleep duration, medical conditions, and eating patterns may play a significant role in determining the risk of obesity among children. These factors could have substantial impacts on the results we obtained and may contribute to improving accuracy if considered in future models. Therefore, it is recommended that future research incorporates a broader range of relevant factors and variables, including dietary patterns, to provide a more comprehensive and accurate assessment of the risk of obesity. Additionally, including more diverse data can help improve model performance and offer deeper insights into the factors that influence obesity.

C. Recommendations for Future Research

Future research should incorporate a broader range of variables, including diet patterns, sleep habits, and medical conditions, to provide a more comprehensive assessment of the risk of obesity. Furthermore, collecting more diverse data could enhance model performance and provide deeper insights into the factors influencing childhood obesity.

VI. Summary and Key Findings

A. Study Approach and Methodology

The research addressed the challenges of childhood obesity using machine learning models, specifically XGBoost for prediction and K-Means Clustering to categorize students based on risk levels. The study began with an extensive review of the literature to understand the current state of research in this field, followed by a comparison of different machine learning approaches. This provided the foundation for selecting and implementing the models that were ultimately used.

B. Results

Through the application of K-Means Clustering, students were classified into different groups based on their risk of obesity. This clustering method offered valuable information on how various factors, such as socioeconomic status, parental characteristics, and neighborhood conditions, influence a child's likelihood of developing obesity. The clusters created a framework that allowed for a deeper analysis of the risk levels.

Once the clusters were established, XGBoost was applied to predict the risk of obesity with high precision. The model achieved an accuracy of 88.04%, demonstrating its effectiveness in handling the dataset and providing reliable predictions for childhood obesity based on the characteristics extracted from the clustering process. The combination of these techniques proved to be a powerful approach for predictive modelling in this domain.

C. Broader Implications

These findings highlight the potential of machine learning techniques to advance our understanding of childhood obesity. The success of XGBoost in predicting risk of obesity offers a strong foundation for further research and development. In particular, these methods could be refined to support early interventions, potentially reducing the prevalence of childhood obesity by targeting high-risk groups more effectively.

D. Conclusion

In summary, this study provides a compelling case for the application of machine learning in health-related fields, specifically in understanding and predicting childhood obesity. The results achieved, especially the accuracy of classification and prediction, suggest that machine learning can play a vital role in future research and public health interventions aimed at combating childhood obesity. Future studies can build on this by incorporating more variables or experimenting with alternative algorithms to improve prediction accuracy and develop targeted intervention solutions.

Acknowledgment

The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia, for funding this research work through grant number: 25UQU44280247GSSR01G.

Funding

This research work was funded by Umm Al-Qura University, Saudi Arabia, under grant number: 25UQU44280247GSSR01G.

References

[1] World Health Organization. Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[2] Reilly, John J., and Joanna Kelly. "Long-term impact of overweight and obesity in childhood and adolescence on morbidity and premature mortality in adulthood: systematic review." *International journal of obesity* 35.7 (2011): 891-898.

[3] Yagin, Fatma Hilal, et al. "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique." *Applied Sciences* 13.6 (2023): 3875.

[4] De-La-Hoz-Correa, Eduardo, et al. "Obesity level estimation software based on decision trees." (2019).

[5] Mondal, Pritom Kumar, et al. "Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers." *Sensors* 23.2 (2023): 759.

[6] Gupta, Mehak, et al. "Obesity Prediction with EHR Data: A deep learning approach with interpretable elements." *ACM Transactions on Computing for Healthcare (HEALTH)* 3.3 (2022): 1-19.

[7] Lingren, Todd, et al. "Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers." *Applied clinical informatics* 7.03 (2016): 693-706.

[8] Kim, Cheong, et al. "Predicting factors affecting adolescent obesity using general bayesian network and what-if analysis." *International journal of environmental research and public health* 16.23 (2019): 4684.

[9] Cheng, Xiaolu, et al. "Does physical activity predict obesity—a machine learning and statistical method-based analysis." *International Journal of environmental research and public Health* 18.8 (2021): 3966.

[10] Hammond, Robert, et al. "Predicting childhood obesity using electronic health records and publicly available data." *PloS one* 14.4 (2019): e0215571.

[11] Ferdowsy, Faria, et al. "A machine learning approach for obesity risk prediction." *Current Research in Behavioral Sciences* 2 (2021): 100053.

[12] Pang, Xueqin, et al. "Prediction of early childhood obesity with machine learning and electronic health record data." *International Journal of Medical Informatics* 150 (2021): 104454.

[13] Dataset:(GitHub - fanwenxiaoyu/ChildhoodObesity: Obesity analysis of a questionnaire dataset from Turkey)

[14] Yardim, Mahmut, et al. "Prevalence of childhood obesity and related parental factors across socioeconomic strata in Ankara, Turkey." *Eastern Mediterranean Health Journal* 25.6 (2019).

[15] Colmenarejo, Gonzalo. "Machine learning models to predict childhood and adolescent obesity: a review." *Nutrients* 12.8 (2020): 2466.

[16] Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." *IEEE access* 8 (2020): 80716-80727.

[17] Anderson, Patricia M., and Kristin F. Butcher. "Childhood obesity: trends and potential causes." *The Future of children* (2006): 19-45.

[18] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[19] Badawy, Mohammed, Nagy Ramadan, and Hesham Ahmed Hefny. "Healthcare predictive analytics using machine learning and deep learning techniques: a survey." *Journal of Electrical Systems and Information Technology* 10.1 (2023): 40.

[20] Ebbeling, Cara B., Dorota B. Pawlak, and David S. Ludwig. "Childhood obesity: public-health crisis, common sense cure." *The lancet* 360.9331 (2002): 473-482.

[21] Sahoo, Krushnapriya, et al. "Childhood obesity: causes and consequences." *Journal of family medicine and primary care* 4.2 (2015): 187-192.

[22] Huang, Jia-Yi, and Sui-Jian Qi. "Childhood obesity and food intake." *World Journal of Pediatrics* 11 (2015): 101-107.

[23] Veitch, Jenny, et al. "Where do children usually play? A qualitative study of parents' perceptions of influences on children's active free-play." *Health & place* 12.4 (2006): 383-393.

[24] Magee, Christopher, Peter Caputi, and Don Iverson. "Lack of sleep could increase obesity in children and too much television could be partly to blame." *Acta paediatrica* 103.1 (2014): e27-e31.

[25] Rogers, Robert, et al. "The relationship between childhood obesity, low socioeconomic status, and race/ethnicity: lessons from Massachusetts." *Childhood obesity* 11.6 (2015): 691-695.

[26] Henriques, João, et al. "Combining k-means and xgboost models for anomaly detection using log datasets." *Electronics* 9.7 (2020): 1164.