VEGUETA

Vol 25 Issue 01, April 2023

ISSN NO: 1133-598X

Leveraging Machine Learning Classifiers for Obesity Risk Estimation

Mrs. VSLP Neelima M	Mrs. B. Bhargavi	Mrs. M. Shilpa
Assistant Professor Department of Information Technology, Lords Institute of Engineering and Technology, Hyderabad,India. mneelima@lords.ac.in	Assistant Professor Department of Information Technology, Lords Institute of Engineering and Technology, Hyderabad,India. bhargavi@lords.ac.in	Assistant Professor Department of Information Technology, Lords Institute of Engineering and Technology, Hyderabad,India. Shilpa.m@lords.ac.in

ABSTRACT

Obesity is defined as having too much fat, which raises the risk of having health issues. It's now a major concern that, if left untreated, can cause heart problems that could be fatal. Sleep patterns, frequent consumption of fatty foods, and inactivity are all influenced by daily lifestyle choices. Two billion people worldwide suffer from obesity. Diabetes has also been identified in individuals who have gained weight. Data management, surgery, drug development, and clinical operations have all advanced recently. In the healthcare industry, machine learning is one of the rapidly expanding disciplines that greatly contributes to innovation and progress. The suggested approach is used in this case with supervised machine learning techniques. Using labelled datasets, it teaches algorithms to correctly identify data or forecast results. utilizing the Obesity and Lifestyle dataset, which includes several metrics and attributes such as age, weight, height, and eating frequency. The data is trained and tested using machine learning techniques such as K-nearest Neighbour (K-NN), Support Vector Machine (SVM), Random Forest, Decision Tree, Adaptive Boosting (ADA boosting), and Gradient Boosting Classifier. Here, we compare the aforementioned models with the XGBoost (eXtreme Gradient Boosting) classifier, a gradient-boosting framework that is applicable to several programming languages, as a further implementation. The accuracy of the two XGBoost classifiers is contrasted with that of the other six classifiers.

INDEX TERMS: Machine Learning,K-NN(K-Nearest Neigh- bour,SVM Support Vector Machine,Random Forest,Decision Tree, Adaptive Boosting,Gradient Boosting,XGBoost.

INTRODUCTION

Without improvements in prevention and treatment methods, the yearly global economic cost of overweight and obesity might reach $4.32 trillion by 2035, according to the World Obesity Federation's 2023 World Obesity Atlas. By then, the majority of people on the planet are expected to be obese, with paediatric obesity rates more than doubling. Although obesity has always been thought to be an issue in high-income nations, it is currently rapidly increasing in low-income areas and affecting people of nearly all ages. A growing number of medical disorders, such as diabetes, cardiovascular disease, cancer, renal disease, stroke, and high blood pressure, are associated with obesity.

Poor eating habits and inactivity are the main causes of obesity. Excessive energy consumption, particularly from fats and carbs, combined with insufficient activity causes the body to store fat. Exercise and yoga are among the activities that are frequently overlooked. Many people ignore obesity's major health consequences and write it off as a cosmetic issue. Although the rise in childhood obesity in Bangladesh is less than that of adult obesity, both trends are alarming. Weight gain in women is influenced by psychological and social factors.

Adults are frequently categorized as overweight or obese using the Body Mass Index (BMI), which is computed as weight in kilograms divided by height in meters squared. Despite not measuring fat directly, BMI is a useful and popular tool. Because of growth differences, interpreting BMI in children and adolescents is more difficult.

Prevention of obesity depends on an understanding of the connections between lifestyle variables and obesity. Machine learning (ML)-based predictive algorithms can be quite successful in this situation. Based on lifestyle data, this study investigates the application of machine learning (ML) classification algorithms to classify people as overweight/obese or non-overweight/non-obese. Machine learning (ML) algorithms may automatically identify complicated, non-linear patterns, providing superior prediction, usability, and adaptability compared to classic statistical methods that rely on predetermined correlations.

By managing complicated data types like time series, text, photos, and social media, deep learning (DL) further extends potential. Using locally acquired data and important obesity-related characteristics, the study's main objective is to create a predictive machine learning model. Finding the best characteristics for predicting obesity, utilizing supervised machine learning approaches, and assessing model performance using metrics like accuracy, recall, precision, and F1-score are some of the specific goals.

In classification issues, feature selection is essential since it lowers computational burden and noise. It may not be beneficial to use every variable. To choose pertinent features and increase accuracy, strategies like the wrapper approach Recursive Feature Elimination with Cross-Validation (RFECV) are employed. In medical datasets, genetic algorithms are also used to optimize features, frequently in conjunction with Support Vector Machine (SVM) parameter tweaking.

The goal of embedded methods is to minimize computing cost while combining the advantages of filter and wrapper techniques by including feature selection into the model training process. These techniques work well for feature selection without sacrificing model performance.

Following data preprocessing, this study employed six supervised machine learning algorithms:

K-Nearest Neighbors (KNN), SVM, Random Forest, Decision Tree, AdaBoost, and Gradient Boosting. Reading raw data, determining factors linked to obesity, and forecasting results were all part of the investigation. The XGBoost algorithm, which uses a gradient boosting decision tree approach, performed the best among them. In order to fix the mistakes of earlier models, boosting involves adding models one after the other. An ablation investigation on model components demonstrated that XGBoost outperformed previous models and considerably increased prediction accuracy.

RELATED WORKS

The healthcare industry has made extensive use of machine learning (ML) approaches to forecast particular ailments based on a variety of variables. Artificial Neural Networks (ANNs), particularly Multilayer Perceptrons trained using the Levenberg-Marquardt approach in conjunction with Random Forests, have demonstrated remarkable predictive performance, as mentioned in article [4]. Regression analysis was also used in earlier studies to forecast BMI at age 14 using data from early infancy. One well-known system, OB-CITY, uses consumer demand analysis, expert advice, implementation techniques including the Analytic Hierarchy Process, and education to empower parents and reduce childhood obesity.

It is known that simple indicators like BMI, Body Adiposity Index (BAI), waist circumference (WC), and waist-to-height ratio (WHtR) can be used to diagnose disorders linked to obesity, such as metabolic syndrome, hypertension, and cardiovascular diseases. These are favored in clinical settings due to their ease of use, affordability, and non-invasiveness. By eliminating the requirement for specialist equipment and staff, such metric inspections reduce healthcare costs when compared to lab-based techniques [31].

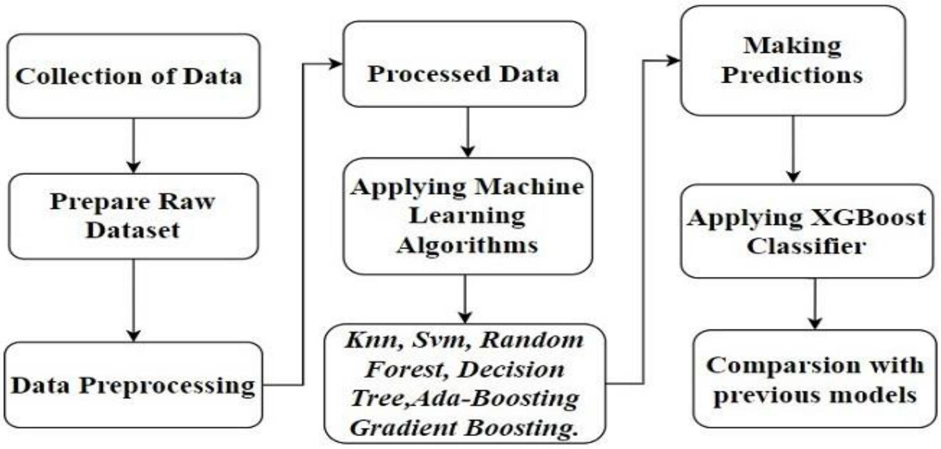
Colmenarejo's review [5] assessed machine learning (ML) models for predicting childhood obesity and suggested ML as an alternative to conventional statistical techniques because of its capacity to handle big datasets and analyze nonlinear relationships. The study did observe, however, that feature selection and model interpretability were not given enough attention. According to a narrative evaluation of 15 peer-reviewed research, predictors differed by age group: parental obesity and antecedent weight were substantially associated with early childhood obesity, but social environment and inactivity were important determinants in later childhood. Ethnicity-specific models may increase predicted accuracy, according to the study.

Adolescent obesity is frequently caused by an imbalance between caloric intake and expenditure. High calorie diets, sugary drinks, and sedentary habits like spending a lot of time in front of a screen are risk factors. On the other hand, moderate exercise promotes bone strength, blood pressure, and a healthy weight. The health effects of adolescent obesity, such as high cholesterol, cardiovascular disease, asthma, and Type 2 diabetes, are highlighted in articles [2][15], underscoring the significance of early detection and intervention.

Finding patterns and connections in big datasets has been attained through the use of structural equation modeling and machine learning. Using data gathered from infancy to age two, some models forecasted childhood obesity by highlighting physical growth measurements while frequently ignoring behavioral patterns. Using logistic regression on YRBSS data from 2007 and 2009, Article [32] investigated adolescent obesity and sleep behavior; nevertheless, no significant relationships were discovered, urging the use of more sophisticated modeling and larger datasets. Although some ML-based studies predicted obesity using validated health behavior surveys, their small sample sizes indicate that more extensive study is required to increase dependability.

PROPOSED METHOD

The proposed method predicts individual obesity levels using machine learning by analyzing features such as weight, height, diet, physical activity, and alcohol consumption. It addresses challenges like childhood and adult obesity by classifying data into categories (e.g., low, medium, high obesity) and comparing algorithm performance. Further implementation details are discussed in later sections.


Collection of Data
Prepare Raw Dataset
Data Preprocessing
Processed Data
Applying Machine Learning Algorithms
Knn, Svm, Random Forest, Decision Tree,Ada-Boosting Gradient Boosting.
Making Predictions
Applying XGBoost Classifier
Comparsion with previous models

A. Methodology

From the dataset of 2,111 entries with 17 attributes, 80% was used for training and 20% for testing. As per article [1], supervised ML algorithms like KNN, SVM, Random Forest, Decision Tree, AdaBoost, and GBM were applied. Additionally, XGBoost was used for comparison, evaluating accuracy, sensitivity, specificity, recall, and F1 score.

B. KNN

K-NN was the first algorithm used, known for its simplicity and effectiveness in classification tasks. As a non-parametric, supervised learning method, it handles both classification and regression without assuming underlying data distribution, using similarity metrics for predictions [1].

C. SVM

SVM is a versatile supervised learning algorithm used for classification, regression, and outlier detection. It constructs hyperplanes in high-dimensional space to separate classes with maximum margin. As noted in article [1], the decision boundary is defined by the equation: $W \cdot X + b = 0$.

D. Random Forest

Random Forest is a fast, adaptable supervised learning method that uses multiple decision trees on data subsets to improve prediction accuracy. It aggregates the results by majority voting for final output. Despite some limitations, it's effective and widely used [1].

E. Decision Tree

As noted in paper [1], Decision Trees are a supervised ML technique that splits data based on specific criteria using decision nodes and leaves. They handle categorical features well with minimal preprocessing.

F. Ada-Boosting

Boosting algorithms combine weak classifiers to form a strong one, improving accuracy and reducing overfitting. As per paper [1], AdaBoost, introduced by Freund and Schapire in 1996, adjusts classifier weights with each new data sample to better predict challenging cases.

G. Gradient Boosting

Gradient Boosting Machines (GBMs) are flexible, customizable models that build base learners aligned with the loss function's negative gradient. They effectively capture complex non-linear relationships and perform well across various real-world applications.

H. XGBoost

XGBoost is a decision tree-based ensemble learning method using Gradient Descent, offering flexibility and efficient computation. Bagging, or Bootstrap Aggregation, creates parallel models from random data samples, with Random Forest being a popular example. Boosting, a sequential process, builds models that correct errors from previous ones. In article [1], accuracy, precision, recall, F1-score, sensitivity, specificity, and ROC curve were calculated for each algorithm to evaluate performance. Accuracy measures the percentage of correctly classified samples.

I. Project Design

Data collection involved gathering information from social media and Google Forms. Afterward, data processing was performed to handle missing values, categorize, and encode text into numerical data. Data cleaning was done by removing noisy values using outlier detection and evaluating the correlation matrix. The obesity results column was removed, and correlation analysis identified the traits impacting the outcomes. Noisy values were also addressed during data preprocessing.

J. Hardware and Software Specifications

Python is a versatile and simple programming language widely used in machine learning due to its consistency and powerful tools. It's ideal for building modern software, particularly machine learning models. Google Collaboratory (Colab), a cloud-based Jupyter notebook service by Google, is commonly used for Python programming. Colab offers free access to computing resources, including GPUs, and allows users to run Python code online. The code typically starts by mounting Google Drive to access datasets, then imports libraries like pandas, seaborn, matplotlib, numpy, and scikit-learn for feature classification, regression, and clustering. Data is read from CSV files, containing around 17 features such as gender, age, height, weight, and family history of obesity.

RESULTS

A list of classifiers is selected and stored; each classifier in the list is looped through before the pre-processor is applied. The accuracy score for each classifier will be printed out. The model implementation makes use of models like KNN, SVM, Random Forest, Decision Tree, Ada-Boosting, Gradient Boosting, and others. We may observe that the aforementioned classifiers perform at varying levels by comparing their accuracy rates: The accuracy of KNN is approximately 0.86 percent, that of SVM is approximately 0.43 percent, that of Decision Trees is 0.96 percent, that of Random Forest is 0.94 percent, that of Ada Boosting is approximately 0.34 percent, and that of Gradient Boosting is the highest of all classifiers at 0.97 percent. Here, the accuracy is obtaining a percentage of 0.84 following the application of the XGBoost model. When compared to other models, Gradient Boosting has a good accuracy score.

DIFFERENT MODEL SCORES OF EACH CLASSIERS:

Classifiers	Model-score
KNN	0.863
SVM	0.434
Random Forest	0.948
Decision Tree	0.967
Ada-Boosting	0.331

Gradient Boosting	0.97
XGBoost	0.84

CONCLUSION

Developing health education programs for primary care professionals is crucial in preventing and managing obesity. Obesity rates vary due to local environmental, socioeconomic, healthcare, and demographic factors. Machine learning (ML) models outperform traditional methods in capturing these variations. Despite many treatment efforts, a guaranteed effective therapy for obesity remains elusive. Robust ML models can aid decision-making based on population factors, though results are not always definitive. Future work aims to expand the dataset to include a wider obesity range for real-time analysis. IoT integration can enhance monitoring, diagnosis, and prevention of obesity-related issues. ML, a subset of AI, uses statistical models to identify patterns and make predictions. Future goals include detecting other obesity-linked diseases and enhancing the ML framework using deep learning models like LSTM and CNN for improved accuracy.

REFERENCES

[1] Ferdowsy, Faria, et al,”A machine learning approach for obesity risk prediction.” Current Research in Behavioral Sciences 2 (2021).

[2] Singh, Balbir, and Hissam Tawfik, ”Machine learning approach for the early prediction of the risk of overweight and obesity in young people.” International Conference on Computational Science. Springer, Cham, 2020.

[3] Musa, Fati, Federick Basaky, and E. O. Osaghae, ”Obesity prediction using machine learning techniques.” Journal of Applied Artificial Intel- ligence 3.1 (2022): 24-33.

[4] R. Hu et al, ”OB CITY–Definition of a Family-Based Intervention for Childhood Obesity Supported by Information and Communica- tion Technologies,” in IEEE Journal of Translational Engineering in Health and Medicine, vol. 4, pp. 1-14, 2016, Art no. xxxxxx, doi: 10.1109/JTEHM.2016.2526739.

[5] H. Siddiqui et al, ”A Survey on Machine and Deep Learning Models for Childhood and Adolescent Obesity,” in IEEE Access, vol. 9, pp. 157337-157360, 2021, doi: 10.1109/ACCESS.2021.3131128.

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

[7] Safaei, Mahmood,et al. ”A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various

machine learning approaches used to predict obesity.” Computers in biology and medicine 136 (2021): 104754

[8] 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.

[9] Chatterjee, Ayan, Martin W. Gerdes, and Santiago G. Martinez, ”Identifi- cation of risk factors associated with obesity and overweight—a machine learning overview.” Sensors 20.9 (2020): 2734.

[10] Lee, Yu-Chi, et al. ”Using machine learning to predict obesity based on genome-wide and epigenome-wide gene–gene and gene–diet inter- actions.” Frontiers in Genetics 12 (2022): 783845.