PREDICTIVE OBESITY AMONG CHILDREN AND ADOLESCENT USING LOGISTIC REGRESSION AND RANDOM FOREST MODEL

Siti Norain Mohd Yusoff, Nor Azriani Mohamad Nor
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis Branch, Malaysia

e-mail: aieynyusoff5063@gmail.com, norazriani@uitm.edu.my

Abstract

Childhood obesity has become a serious public health concern in Malaysia and worldwide. This study aims to predict obesity among children and adolescents using two classification techniques: logistic regression and random forest. The dataset was obtained from Kaggle and includes 2,111 records with 16 demographic, dietary, and lifestyle variables. Logistic regression was used to identify significant predictors, while the random forest model captured complex interactions among variables. The performance of both models was evaluated using accuracy, precision, recall, specificity, F1-score, and AUC. The logistic regression model achieved high recall (98.9%) but very low precision (2.0%) and F1-score (3.8%), indicating a high false-positive rate. The random forest model showed more balanced metrics with a precision of 48.9% and F1-score of 46.1%, although the accuracy was lower at 49.9%. The findings suggest that while logistic regression can identify at-risk individuals, random forest provides more reliable predictions. Both models, however, showed weak performance overall (AUC ≈ 0.5), highlighting the need for improved data and modeling approaches.

Keywords: *Childhood Obesity, Logistic Regression, Random Forest, Prediction Model, Malaysia*

1. Introduction

The rising prevalence of obesity among Malaysian children and adolescents poses serious health risks. Factors such as high-calorie diets, low physical activity, and technology overuse contribute significantly. This study focuses on predicting childhood obesity using logistic regression and random forest models. The objective is to identify key risk factors and evaluate the predictive performance of both models. The outcome may assist healthcare providers in early detection and prevention strategies tailored to local populations.

2. Methodology

The dataset used for this study was sourced from Kaggle, consisting of 2,111 records with 16 attributes. It includes variables like age, gender, family history, physical activity, eating patterns, and screen time. Data was split into training (80%) and testing (20%) sets. Logistic regression was applied to identify significant predictors based on p-values, while random forest focused on ranking feature importance. Both models were implemented using RStudio. Performance evaluation included standard classification metrics: accuracy, precision, sensitivity, specificity, F1-score, and AUC. Logistic regression emphasized interpretability, while random forest offered robustness in handling non-linear relationships.

3. Results / Findings

The logistic regression model showed high recall (98.9%) but very low precision (2.0%), suggesting over-prediction of obesity. It identified “eating between meals” as the only significant predictor. In contrast, random forest identified five key predictors: family history with overweight, frequency of vegetable intake, age, physical activity, and number of meals. It showed better balance in metrics, with 48.9% precision and 46.1% F1-score, although the overall accuracy was lower (49.9%). Both models performed poorly in terms of AUC (≈ 0.5), suggesting weak overall discrimination between obese and non-obese individuals.

4. Novelty of Research / Product / Project

This study applies and compares traditional statistical (logistic regression) and machine learning (random forest) methods to predict obesity in children a growing issue in Malaysia. The novelty lies in analyzing variable importance from both models and highlighting their strengths and limitations. This comparison provides insight into which model is better suited for practical application in health settings, especially under data imbalance conditions. The project sets the groundwork for future development of more advanced prediction tools.

5. Conclusion or Future Work

For future work, firstly, include more important variables such as sleep duration, stress level, screen time during weekend, and family income. This is to ensure that the model can detect obesity more accurately by considering other possible risk factors. Next, use advanced machine learning models like gradient boosting or neural networks. This is because these models can handle complex data better and give more reliable prediction results.

REFERENCE LIST

Devetyarov, D., & Nouretdinov, I. (2010). Prediction with confidence based on a random forest classifier. IFIP Advances in Information and Communication Technology, 37–44. https://doi.org/10.1007/978-3-642-16239-8_8

Ministry of Health Malaysia. (2019). National Health and Morbidity Survey 2019. https://www.iku.gov.my/nhms

Muhramasaputra. (2023). *Obesity based on eating habits and physical condition* [Dataset]. Kaggle. https://www.kaggle.com/datasets/muhramasaputra/obesity-based-on-eating-habits-and-physical-cond

UNICEF. (2025). Overweight and Obesity Among Children and Adolescents. https://www.unicef.org/rosa/reports/overweight-and-obesity-among-children-and-adolescents

UNICEF. (2025). Overweight and Obesity Among Children and Adolescents. https://www.unicef.org/rosa/reports/overweight-and-obesity-among-children-and-adolescents