Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018

Sri Astuti Thamrin[1†], Dian Sidik Arsyad[2†], Hedi Kuswanto[1], Armin Lawi[3] and Sudirman Nasir [4]*

[1] Department of Statistics, Faculty of Mathematics and Natural Science, Hasanuddin University, Makassar, Indonesia,
[2] Department of Epidemiology, Faculty of Public Health, Hasanuddin University, Makassar, Indonesia, [3] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, Indonesia, [4] Department of Health Promotion, Faculty of Public Health, Hasanuddin University, Makassar, Indonesia

Frontiers in Nutrition 8:669155.(2021) doi: 10.3389/fnut.2021.669155

www.frontiersin.org

Abstract

Obesity is strongly associated with multiple risk factors. It is significantly contributing to an increased risk of chronic disease morbidity and mortality worldwide. There are various challenges to better understand the association between risk factors and the occurrence of obesity. The traditional regression approach limits analysis to a small number of predictors and imposes assumptions of independence and linearity. Machine Learning (ML) methods are an alternative that provide information with a unique approach to the application stage of data analysis on obesity. This study aims to assess the ability of ML methods, namely Logistic Regression, Classification and Regression Trees (CART), and Naïve Bayes to identify the presence of obesity using publicly available health data, using a novel approach with sophisticated ML methods to predict obesity as an attempt to go beyond traditional prediction models, and to compare the performance of three different methods. Meanwhile, the main objective of this study is to establish a set of risk factors for obesity in adults among the available study variables. Furthermore, we address data imbalance using Synthetic Minority Oversampling Technique (SMOTE) to predict obesity status based on risk factors available in the dataset. This study indicates that the Logistic Regression method shows the highest performance. Nevertheless, kappa coefficients show only moderate concordance between predicted and measured obesity. Location, marital status, age groups, education, sweet drinks, fatty/oily foods, grilled foods, preserved foods, seasoning powders, soft/carbonated drinks, alcoholic drinks, mental emotional disorders, diagnosed

hypertension, physical activity, smoking, and fruit and vegetables consumptions are significant in predicting obesity status in adults. Identifying these risk factors could inform health authorities in designing or modifying existing policies for better controlling chronic diseases especially in relation to risk factors associated with obesity. Moreover, applying ML methods on publicly available health data, such as Indonesian Basic Health Research (RISKESDAS) is a promising strategy to fill the gap for a more robust understanding of the associations of multiple risk factors in predicting health outcomes.

Keywords: classification, Logistic Regression, machine learning, Naive Bayes, obesity status