Machine Learning–Driven Prediction of Non-Communicable Diseases Using Anthropometric Data

Abstract

This dissertation examines how machine learning can be used to forecast non-communicable diseases (NCDs), including heart disease, diabetes, and cancer, and explores how their prevalence is largely due to obesity. Based on a sample of anthropometric measurements from adults (Jaffna Teaching Hospital (JTH) and Sabaragamuwa University of Sri Lanka (SUSL)), excluding children and pregnant women, the study uses ten variables: age, weight, height, gender, body fat mass (MBF), total body water (TBW), body fat percentage (PBF), body mass index (BMI), visceral fat area (VFA), and waist-to-hip ratio (WHR). The most significant characteristic turned out to be the visceral fat area (VFA). Machine learning algorithms were employed in binary classification: Random Forest, Extreme Gradient Boost (XGBoost), Artificial Neural Networks (ANN), Decision Tree, AdaBoost, Logistic Regression, CatBoost, and Support Vector Machine (SVM). The models achieved accuracy rates above 85%, with Random Forest reaching the highest at 98.90%, and outputs indicating Yes (NCD patient) or No. The study utilizes data mining to facilitate knowledge discovery in healthcare systems, which could be beneficial for early detection and enhancing patient well-being through a knowledge discovery in databases (KDD) approach.

Keywords: obesity; machine learning; non-communicable diseases; prediction; binary classification