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

Introduction

Obesity and other anthropometric factors such as body mass index, waist circumference, and weight are leading risk factors in diseases in almost every country in the world, and non-communicable diseases are the leading cause of death, accounting for more than 80% of deaths in most countries in the world. The diseases present a major economic and social burden, especially in low- and middle-income countries like Sri Lanka [1], the burden of which increases due to lifestyle habits associated with poor dietary habits, lack of physical exercise, and tobacco-related habits [1].

The proposed study employs machine learning (ML) with the objective of finding anthropometric measurements that may predict the presence of NCD in order to implement early diagnosis without invasive procedures [2]. The main parameters are age, weight, height, gender, body fat mass (MBF), total body water (TBW), percentage body fat (PBF), body mass index (BMI), visceral fat area (VFA), and waist-to-hip ratio (WHR). The primary goal is to come up with a predictive model that may determine NCD patients or risk factors using these measurements [3]. Sub-objectives include the choice of the most appropriate ML methodologies, the chosen specific learning technique, and the establishment of significant anthropometric parameters [4]. The study covers the issue of manual

identification of NCD by non-proficient practitioners by suggesting the automated system to be used by medical workers and encourages regular evaluation of body composition [5].

Materials and Methods

Dataset

At Jaffna Teaching Hospital (JTH) and the Sabaragamuwa University of Sri Lanka (SUSL), 300 adults aged over 18 were taken, and children under 18 years and pregnant women were excluded. The data include 164 men and 136 women, and the measurements of the body were taken with the help of a body composition analyzer (e.g., TANITA model). The height and the weight were recorded independently and entered into the analyzer. The ten characteristics were logged and measured as age, weight, height, gender (coded), MBF, TBW, PBF, BMI, VFA, and WHR in paper form, transduced to an Excel, and into CSV form. Data preprocessing consisted of cleaning (null/duplicates/outliers), normalization (Min-Max scaling), and standardization with the use of the pandas and scikit-learn packages in Python. The target variable was binary (0: no NCD; 1: NCD present) manually categorized with the help of two nutritionists based on the WHO values (e.g., BMI 18.525 normal, >27.5 at risk; VFA >90cm among men at risk).

Machine Learning Models

Binary classification ML algorithms were used and supervised: Random Forest (ensemble of decision trees to reduce overfitting), Extreme Gradient Boosting (XGBoost; tree boosting sequential with regularization), Artificial Neurons Network (ANN; multi-layer case maker of non-linear patterns), Decision Tree algorithm (hierarchical splitting in rules with interpretability), AdaBoost (adaptive boosting of weak learners), Logistic Regression (probabilistic modeling using logit), Catboost (gradient boosting optimized to deal with categorical data), Python libraries, e.g., XGBoost, CatBoost packages, were used to construct the models using scikit-learn and other packages. The Grid Search Cross-Validation was used to tune hyperparameters.

Training and Evaluation

The data was divided into an 80:20 training and test dataset (240 training, 60 test samples). The data were preprocessed, and models were trained on the results and then measured against accuracy, mean squared error (MSE), absolute squared error (ASE), precision, recall, F1-score, and confusion matrices. It was assessed only based on binary types of outcomes (presence/absence of NCD), where 0 meant healthy and 1 meant risk. Overfitting has been addressed by the use of methods such as cross-validation and ensembles.

Results and Discussion

The measure of accuracy was over 85% in all models, with Random Forest the most accurate 98.90% (MSE/ASE: 1.09 %), followed by XGBoost and ANN 97.80% (MSE/ASE: 2.19%), Decision Tree

96.05% (MSE/ASE: 3.94%), AdaBoost 93.40% (MSE/ASE: 6.59%), Logistic Regression 88.52% (The precision, recall, and F1-scores were consistently high (0.85; 0.85; 0.90+) and Random Forest had a nearly perfect balance (F1 0.99 in both classes). Confusion matrices have shown that there are a few mistakes in misclassification, e.g., Random Forest produced 1 false negative. Analysis of significance on features estimated the aspect of VFA to be the most important feature, followed by PBF and BMI, which were consistent with obesity being a major NCD risk. The high rate of ensemble methods, such as Random Forest and XGBoost, is associated with non-linear relationships and a decrease in variance in anthropometric data. These outcomes have even exceeded the conventional manual evaluation, allowing affordable, non-invasive NCD screening. A weakness is that it only consists of a small sample size (300 samples), and the study is limited to the northern and southern parts of Sri Lanka, which would degrade generalizability. The possible widespread global setting application in resource-poor environments is described as currently by focusing on the VFA application in preventing the mortality of NCDs through early intervention.

Table 1: Model Performance of ML models

Algorithms	**Accuracy (%)**	**Mean Squared Error [MSE] (%)**	**Absolute Squared Error [ASE] (%)**
Random Forest	98.90	1.09	1.09
Extreme Gradient Boosting	97.80	2.19	2.19
ANN	97.80	2.19	2.19
Decision Tree	96.05	3.94	3.94
Ada Boost	93.40	6.59	6.59
Logistic Regression	88.52	11.47	11.47
Cat Boost	87.91	12.08	12.08
SVM	85.24	14.75	14.75

Table 2: Evaluation metric of each ML model

Model	**Class**	**Precision**	**Recall**	**F1-Score**	**Support**	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1-Score**	**Weighted Precision**	**Weighted Recall**	**Weighted F1-Score**	**MSE (%)**	**ASE (%)**
Random Forest	0	0.97	1.00	0.99	33	-	-	-	-	-	-	-	1.09	1.09
	1	1.00	0.98	0.99	58	0.99	0.99	0.99	0.99	0.99	0.99	0.99	-	-
Extreme Gradient Boosting	0	0.97	0.97	0.97	33	-	-	-	-	-	-	-	2.19	2.19
	1	0.98	0.98	0.98	58	0.98	0.98	0.98	0.98	0.98	0.98	0.98	-	-
Artificial Neural Network	0	1.00	0.94	0.97	35	-	-	-	-	-	-	-	2.19	2.19
	1	0.97	1.00	0.98	56	0.98	0.98	0.97	0.98	0.98	0.98	0.98	-	-
Decision Tree	0	0.87	0.96	0.92	28	-	-	-	-	-	-	-	3.94	3.94
	1	0.98	0.92	0.95	48	0.93	0.92	0.94	0.93	0.94	0.93	0.93	-	-

AdaBoost	0	0.89	0.97	0.93	40	-	-	-	-	-	-	-	6.59	6.59
	1	0.98	0.90	0.94	51	0.93	0.93	0.94	0.93	0.94	0.93	0.93	-	-
Logistic Regression	0	0.83	0.93	0.88	27	-	-	-	-	-	-	-	11.47	11.47
	1	0.94	0.85	0.89	34	0.89	0.88	0.89	0.88	0.89	0.89	0.89	-	-
CatBoost	0	0.84	0.90	0.87	40	-	-	-	-	-	-	-	12.08	12.08
	1	0.92	0.86	0.89	51	0.88	0.88	0.88	0.88	0.88	0.88	0.88	-	-
Support Vector Machine	0	0.76	0.79	0.78	24	-	-	-	-	-	-	-	14.75	14.75
	1	0.86	0.84	0.85	37	0.82	0.81	0.81	0.81	0.82	0.82	0.82	-	-

Conclusion

In this case, the research was able to construct ML models of the NCDs prediction based on the anthropometric data, reaching the maximum percentage of prediction equal to 98.90% accuracy with Random Forest as the most accurate algorithm. Having pinpointed the risk factor as VFA, the given approach presents an effective screening instrument to detect it in its early stages, which decreases the necessity of more invasive tests and aids health care in such low-resource regions as Sri Lanka. Additional validation by covering a larger dataset dealing with extended validation is planned in the future as well as the implementation of hybrid/deep learning in order to forecast nutritional levels and NCD risk.

References

[1] P. E. Petersen, "The World Oral Health Report 2003: continuous improvement of oral health in the 21st century–the approach of the WHO Global Oral Health Programme," *Community Dentistry and oral epidemiology,* vol. 31, pp. 3-24, 2003.

[2] D. Abegunde and A. Stanciole, "An estimation of the economic impact of chronic noncommunicable diseases in selected countries," *World Health Organization, Department of Chronic Diseases and Health Promotion,* vol. 2006, 2006.

[3] A. Parashar, M. Willeboordse, A. K. Gupta, and O. C. van Schayck, "Effect of brief interventions to promote behavior change on clinical outcomes of selected non-communicable diseases: The World Health Organization (WHO) Package of Essential Non-communicable disease (PEN) Interventions for primary health care settings–study protocol of a quasi-experimental study," *Contemporary Clinical Trials,* vol. 113, p. 106675, 2022.

[4] A. Bohr and K. Memarzadeh, *Artificial intelligence in healthcare*. Academic Press, 2020.

[5] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal,* vol. 15, pp. 104-116, 2017.