ICAS 2025

Faculty of Applied Science
University of Vavuniya

Comparative Study of Machine Learning Models for Non-Communicable Disease Prediction

Abstract

Morbidity and mortality matters Cardiovascular and other non-communicable diseases (NCDs), including diabetes and cancers, are global leading causes of morbidity and mortality globally. This analysis draws on the techniques of supervised machine learning (ML) models to predict the risks of NCDs based on anthropometric data drawn from 300 participants in the Jaffna Teaching Hospital and Sabaragamuwa University in Sri Lanka. Preprocessing of data was done through eliminating noise, normalization, and correction of outliers. Such models as Random Forest, XGBoost, ANN, Decision Tree, AdaBoost, Logistic Regression, CatBoost, and SVM were trained and optimized over a grid search cross-validation. The findings present that Random Forest reported the best accuracy (98.9), whereas ensemble models were superior to other conventional classifiers. The method offers a suitable instrument in the timely detection of NCDs that facilitate preventive care where resources are scarce.

Introduction

Non-communicable diseases (NCDs) cause more than 70% of deaths worldwide. Early detection and prediction are crucial for prevention. Machine learning has shown promise in predictive healthcare, offering cost-effective, data-driven solutions. This study investigates ML models applied to anthropometric data to predict NCD risk in Sri Lanka.

Methodology

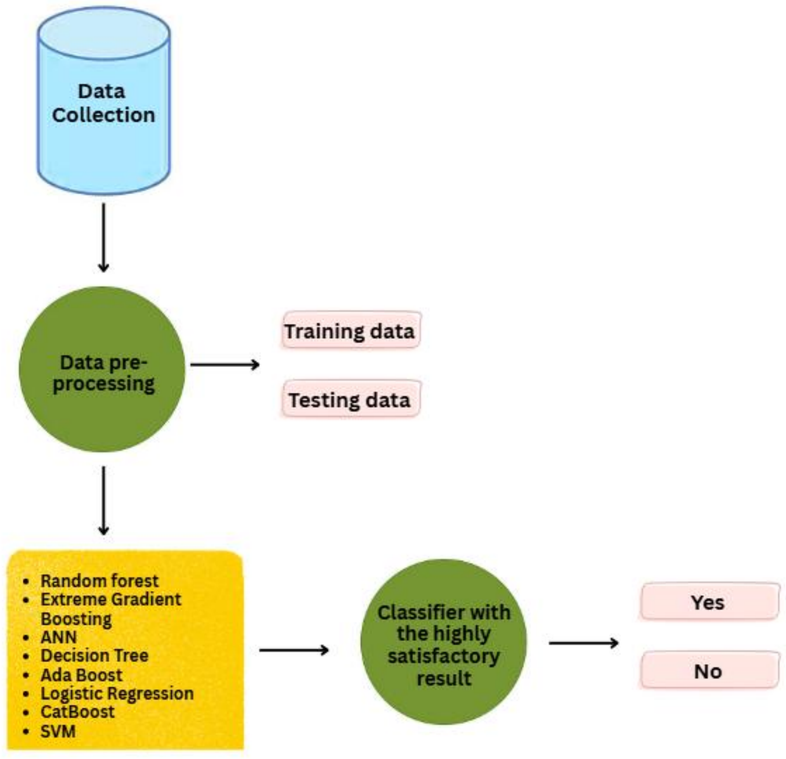

Figure 1: High-level architecture

Objective

1. To preprocess and analyze anthropometric data for NCD risk prediction.
2. To evaluate and compare the performance of multiple supervised ML algorithms.
3. To identify the most effective model for early detection of NCDs.

Results

Table 1: Algorithm's accuracy, mean squared error and absolute squared error

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

Conclusion

Among the tested models, **Random Forest** emerged as the most accurate (98.9%) and reliable, followed closely by **XGBoost** and **ANN**. Traditional models such as Logistic Regression and SVM performed less effectively. These findings confirm that ensemble learning techniques can significantly enhance NCD risk prediction using simple anthropometric data. Such models provide a low-cost, scalable, and effective tool for preventive healthcare.

References

- K. Tripathi and H. Garg, "Machine Learning techniques for Cardiovascular Disease" in IOP Conference Series: Materials Science and Engineering, 2021, p. 012140. R. E. Ali, H. El
- Kadi, S. S. Labib, and Y. I. Saad, "Prediction of potential-diabetic obese-patients using machine learning techniques" 2019.