Forecasting Diabetes Using Feature Selection in Machine Learning Models to Identify Key Indicators

Suejit Pechprasarn[1]*, Nichapa Srisaranon[2], and Panpatchanan Yimluean[2]

[1]College of Biomedical Engineering, Rangsit University, Pathumthani 12000, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok 10200, Thailand
The last two authors have contributed equally.
*Corresponding author; E-mail: suejit.p@rsu.ac.th

Abstract

In this investigation, the authors delve into the impact of diabetes, a prevalent chronic condition in the United States and developed countries that poses significant economic burdens and affects millions of individuals. This study utilizes a publicly accessible clinical dataset from the Behavioral Risk Factor Surveillance System (BRFSS), encompassing 253,680 patients and featuring 21 predictors, excluding the label. These predictors include factors such as high blood pressure, cholesterol levels, body mass index, smoking habits, and various health-related aspects. The study focuses on training different supervised classification models, comparing their performance using classification performance metrics, and employing unsupervised machine learning approaches to identify essential features among the 21 predictors. The primary focus is to achieve a balanced training dataset. Using MATLAB2023a, 34 distinct machine-learning models were trained, and their classification performance metrics were evaluated. Notably, the Quadratic Support Vector Machine (SVM), Coarse Gaussian SVM, and Narrow Neural Networks exhibited the highest training accuracy at 76.3%. For unseen test datasets, the Bilayered Neural Network emerged with the highest accuracy at 74.7%. However, a comprehensive evaluation considering average accuracy, precision, recall, and F-1 scores highlighted the Quadratic SVM as the top performer. To gain insights into factors influencing diabetes diagnosis, feature selection methods were employed, identifying crucial components for constructing a more efficient model with fewer parameters. The selected predictors, including high blood pressure, cholesterol levels, body mass index, heart disease or heart attack, physical activities, general health condition, days with a bodily injury in the past 30 days, difficulties in walking or climbing stairs, and age, contributed to a streamlined model with an accuracy of 75.4%, showcasing the potential for a practical and simplified diagnostic approach. The authors also discuss and propose a systematic methodology to lower the complexity of the trained model without compromising the performance of the classification model.

Keywords: *Diabetes Diagnosis, Machine Learning, Principal Component Analysis, Complexity-reduced Model, Intelligent Diagnostic Software*