Comparison of machine learning techniques used in type II diabetes risk prediction

Kaluarachchi K.N.[1*], Premachandra K.P.[1], Dissanayake R.B.N.[1]

[1] *Department of Physical Science, Faculty of Applied Sciences, Rajarata University of Sri Lanka*

Native Americans living in Arizona called PIMA have several medical problems such as diabetes. Diabetes prediction of females in this population has been done using various machine learning techniques. The objective of this study is to compare eight supervised machine learning models to identify the best algorithm with a low bias-variance trade-off for the diagnosis of type II diabetes among female PIMA population. Eight prediction models namely logistic regression, decision tree, random forest, naïve bayes, k-nearest neighbor, support vector machine, gradient boosting, and artificial neural network (ANN) were developed for type II diabetes using the data published by the National Institute of Diabetes, Digestive and Kidney Diseases in the USA (PIMA Indian Diabetes Dataset). Among the 768 patient records, 430 (50% with diabetes and 50% without diabetes) were used to train the models to reduce data biasness, and the remaining 338 records were used for testing. The performance of each model was evaluated and compared using testing accuracy, mean squared error (MSE), sensitivity, precision, and F1-score. The results showed that the random forest model has the highest testing accuracy of 83.12% and the lowest MSE. This result shows that most significant predictor variables are number of pregnancies, insulin level, BMI level, and age. The ANN model achieved the highest MSE, due to the limited number of training data. Therefore, the random forest model with number of 50 subtrees is the most accurate machine learning model that can be used to diagnose type II diabetes in the PIMA Indian Diabetes Dataset.

Keywords: Machine Learning, Testing Accuracy, Types II Diabetes Prediction, PIMA Indian

*Corresponding author: knk.nisansala2012@gmail.com