Authors:

Keywords:

Title: Prediction of Type 2 Diabetes Occurrence Using Machine Learning Model

Early detection of type 2 diabetes (T2D) occurrence allows a person at risk to take preventive actions that can prevent or delay the progression of the disease. In this study we aim to develop a machine learning (ML) model to predict T2D occurrence in year Y+1 using the variables in year Y.

The dataset was collected from electronic health records between 2013 and 2018 at a medical agency located in Seoul, Korea. We utilized 169,024 instances from 80,692 patients with longitudinal data to build the ML model. Each instance has 1,444 variables.

To construct a prediction model, key variables or features were first selected using ANOVA and chi-square tests, and the recursive feature elimination methods. Then we employed random forest (RF) and XGBoost algorithms based on these variables to predict the outcome as normal, prediabetes or diabetes.

Selected variables were fasting plasma glucose (FPG), HbA1c, triglycerides, body mass index (BMI), r-GTP, gender, age, uric acid, smoking, drinking, physical activity and family history. The accuracy of the RF classifier in predicting the occurrence in year Y+1 was 70.7% while that of the XGBoost classifier was 69.6%.

The proposed prediction model using ML can provide both clinicians and patients with valuable information on the immediate incident of T2D. In addition to traditional predictors of T2D (FPG, HbA1c, BMI, family history), it should also be noted that variables such as r-GTP, uric acid, triglycerides and life style factors in year Y are influential variables in predicting T2D occurrence in year Y+1.