Oral presentation: 213

Prediction of type 2 diabetes risk factor using machine learning in Sri Lanka

R. M. S. D. Menike*, S. G. V. S. Jayalal and N. Algiriyage

Department of Industrial Management, Faculty of Science, University of Kelaniya, Sri Lanka.
*sddamayanthi93@gmail.com

Diabetes mellitus is in third place in the index of 20 major diseases affecting deaths in Sri Lanka. Diagnosis of diabetes is a key and insipid task. A successful, easy and correct method has not been identified to identify the diabetes mellitus in the early stage. Currently, the Diabetes detection is done using blood tests, such as Glycated hemoglobin (A1C) test, Random blood sugar test**,** Fasting Plasma Glucose test, Oral Glucose Tolerance Test, and Blood Sugar Series. People who do not have a special condition are generally unwilling to go for a blood test, which is a process that costs them time and money. Diabetes mellitus cannot be fully cured, but if identified in prediabetes, it is possible to prevent prediabetes from developing into type II by the actions such as eating healthy foods, losing weight, being physically active. As there are no regular medical checkups to diagnose pre-diabetes among the general public, identification of pre-diabetes is problematic in Sri Lanka. Machine learning techniques have been successfully applied to predict the risk factor for diabetes mellitus in other countries. Due to the high variance of economic and cultural factors, it is very difficult to come up with a common model to all countries. The detection of diabetes from some important risk factors is a multi-layered process. This research is primarily aimed at identifying factors that contribute to the prevalence of diabetes in Sri Lanka and finding a mechanism to predict the risk of diabetes through the use of machine learning algorithms over the identified factors. The gathered dataset consists of anthropometric and behavioral data of a set of people who have diabetes and don't have, such as age, BMI, gender, heredity, and Hypertension etc. Wrapper methods are used to identify the most influential factors of these factors that affect diabetes mellitus. Since earlier studies have shown better performances, Support Vector Machine, J48, Random Forest algorithms are used for classification of selected dataset. As a result, three models are generated, and the performance of each model is measured analyzing measurements such as accuracy, specificity sensitivity. The outcome model of the study is the one that shows the best performance. That model presented by this scrutiny as the final output can be used by the public without specific domain knowledge, provide a more accurate clue of the diabetes risk of themselves by giving the data related to the identified factors as inputs.

Keywords: Data Mining, J48, random forest, support vector machines, Type II Diabetes