



USE OF MACHINE LEARNING FOR DIABETIC PREDICTION

De Silva L.H.S.¹, Nandana Pathirage¹ and Jinasena T.M.K.K.^{2*}

¹Department of Computer Science, General Sir John Kotelawala Defence University, Sri Lanka

²Department of Computer Science, University of Jayewardenepura, Sri Lanka

kasunkosala@yahoo.com

ABSTRACT

Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases etc. As per the current system in Sri Lanka, patients need to visit diagnostic center, consult their doctor and sit tight a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. But with the rise of Machine Learning approaches we have been able to find a solution to this issue, we have developed a system using data mining which have the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treat the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes related data. Because of that, it has a significant role in diabetes research more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification methods namely, Decision Tree, Naïve Bayes and Support Vector Machine algorithms. Currently the models give accuracies of 84.6667%, 76.6667%, and 77.3333% for Decision Tree, Naïve Bayes, and SMO Support Vector Machine respectively. These results have been verified using Receiver Operating Characteristic curves in a cost sensitive manner. The developed ensemble method uses votes given by the other algorithms to produce the final result. This voting mechanism eliminates the algorithm dependent misclassifications. It also helps to get more accurate prediction for the disease. Weka data mining extension was used for data preprocessing and experimental analysis. Results shows a significant improvement of accuracy of the ensemble method compares to other existing methods.

Keywords: Data Mining, Diabetes, Machine Learning, Decision Tree, Support Vector Machine