Chapter 5
Comparative Analysis of Machine Learning-Based Diabetes Prediction Approaches

Harsh Vardhan
National Institute of Technology, Hamirpur, India

Vijay Kumar
Dr B R Ambedkar National Institute of Technology, Jalandhar, India

ABSTRACT

The early prediction of diabetes mellitus may help improve the health of patients and cure them of this disease. In recent years, machine learning techniques have been widely used to predict diabetes in its early stages. In this chapter, an attempt has been made to analyse the performance of different machine learning techniques for diabetic prediction. Four well-known machine learning techniques, named as random forest, support vector machine, decision tree, and XGBoost are used. These techniques are evaluated on the Indian Diabetes dataset. Experimental results reveal that random forest algorithm achieved highest accuracy than the other techniques in terms of performance measures. These techniques will help to reduce diabetes incidence and health care costs. This work can be used to envisage diabetes in its early stages.

1. INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is responsible for regulating blood sugar levels in the human body. In the condition of diabetes, the body fails to regulate the amount of glucose in the blood (Saydah et al. 2004; Chhabra 2023). Hyperglycaemia, or raised blood sugar, is a common effect of diabetes and, over time, leads to serious damage to many of the body's systems, especially the nerves and blood vessels. If the diabetes is left untreated, then it can cause many health complications (Zimmet et al. 2016).

DOI: 10.4018/979-8-3693-2359-5.ch005

In recent years, diabetes mellitus has more than tripled. In 2021, about 537 million people were diabetic that is roughly about 7% of the entire world population. According to WHO, it is expected that about 800 million people will be diabetic by the end of 2045. The situation of diabetes in the world is in alarming condition. Hence, there is a need to take step before it gets out of control (Sartorius, 2022). Diabetes is still uncurable with the help of medical science (Mitratza et al. 2020). The treatment of diabetes is limited to prescribe the lifelong medicine that will help to halt the rate of damage done by diabetes. Therefore, the early detection of diabetes is even more useful for patients (Mir and Dhage 2018).

Machine learning techniques are used to detect the diabetes in a patient. In this paper, an attempt has been made to analyse the performance of machine learning techniques for the prediction of diabetes. The main contributions of this paper are

1. Machine learning techniques are used to predict diabetes from the patient data.
2. The techniques are tested on the Indian Diabetes dataset.
3. The performance of machine learning techniques is evaluated in terms of accuracy and F1-score.

The remaining structure of this paper is as follows. Section 2 discusses the brief description of machine learning models followed by related work done in the direction of diabetes prediction. The experimental results are presented in Section 3. Section 4 presents the discussion, followed by conclusions in Section 5.

2. METHODS

In this section, the machine learning models are briefly described followed by related work done in the direction of diabetes prediction.

2.1 Machine Learning Models

Four well-known machine learning models namely decision tree, support vector machines, random forest, and XG boost are briefly discussed in the preceding subsections.

2.1.1 Decision Tree

Decision tree comes under the category of supervised learning (Kamiński et al. 2018). In decision tree, decision is made by splitting the data into feature based splits. The decision tree works by forming a tree like structure that has multiple branches. This technique works for both regression and classification task. However, it is preferred for classification as compared to the regression. The formation of decision tree is complex for complex data. In decision tree, the splitting of data greatly affects the formulation of tree. The golden rule for splitting is to split tree in such a way that it will reduces the randomness present in the data.

2.1.2 Support Vector Machines

Support Vector Machines (SVM) is one of the most popular and powerful machine learning technique (Cortes and Vapnik 1995). SVM works for both classification and regression tasks. In SVM, the data