Abstract

With progression in technology specifically in last three decades or so, an enormous magnitude of information has been transitioned into a digital form, which resulted in formation of enormous data repositories. With accrual of information in these repositories a challenge persisted as how to extract meaningful knowledge from it. Data mining as a tool was used to tackle the situation. Data mining is one of the young and promising fields that have gotten consideration of the scientists in both industry and scholastics. “Knowledge discovery in databases” a synonym used by many for term data mining is a technique aimed at revealing significant and interesting patterns from data. The advancement in both Information Technology and Medical sciences has led to accumulation of huge amount of clinical data that gets generated from electronic health records. These emerging technologies has led to amalgamation of healthcare and computing ensuing in formation of health informatics.

Diabetes is one of the most appalling disease that mankind is facing currently. The disease occurs because of body’s improper response to insulin: which is an important hormone in our body that converts sugar into energy needed for proper functioning of regular life. The challenge posed by the disease has felt the need to develop the medical prognostic DSS Decision Support System which can aid practitioners in diagnosis process. For effective medical diagnosis, it is important to extract the knowledge from medical databases. The Endeavour of knowledge extraction from database is to generate clear and understandable narrative of patterns.

The main objective of this research is designing a predictive model using data mining tools and technique for diabetes (Type II Diabetes) prognosis which should be more accurate and generalized and can predict diabetes based on both clinical and non clinical parameters. The complete work has been categorized into three sections; in section one an extensive review of literature has been carried out inorder to find the research gaps and to abstract the knowledge regarding the diabetes prediction frameworks that have been employed so far. Here it was found that dataset being used for development purposes had many flaws, we designed
models on our freshly obtained dataset using all major machine learning algorithms to obtain the best resulting classifiers. A highest accuracy rate of 94.14% was obtained using bagging classifier. In section two we enhanced the predictive accuracy of our models by reducing the class imbalance, which is one of the hottest research topics in machine learning and is the major problem with datasets. The results achieved are satisfactory here. The highest accuracy reached here was on decision tree classifier with an accuracy rate of 94.70%. In section three we conducted a novel experiment to develop the model that can predict diabetes using only non-clinical parameters. The aim here is to design a predictive model which does not require a patient to undergo blood tests and diabetes is predicted using only non-clinical parameters. This experiment is done keeping in view that all patients are not well off to perform the costly blood tests especially in developing and poor nations. The model developed here had an acceptable accuracy of 85.74%.

We have used 734 patients records obtained from superb diagnostic center in Srinagar. The researchers may further add more instances to our dataset which might be purposeful as strong results could be achieved with larger dataset. Researchers may add some more attributes to dataset that also have some relevance to diabetes prediction take for example for our model with clinical parameters we can add lipid levels of blood which also has some relevance with diabetes, for model with non-clinical parameters we add some life style related parameters like intake of junk food or if a person takes daily exercises or not. The model can also be enhanced by applying other classifiers that can predict the vulnerability of diabetic patient to other allied diabetes triggered disease like heart disease, renal disease etc. and patient can be forewarned about the complications and can be asked to change the lifestyle habits to lessen the complications of disease.

Keywords: Diabetes, Data mining, Prognosis, Knowledge discovery, class imbalance.