Ensemble Classification Approach For Machine Learning Using Diabetes Mellitus Data

Dr. P.Chandra Kanth[1],
Department of Computer Science & Engineering,
Audisankara College of Engineering and Technology, Gudur, AP

Sk.Naveed[2]
PG scholar, Department of MCA,
Audisankara College of Engineering and Technology, Gudur, AP

Abstract - Diabetes is a dreadful disease identified by escalated levels of glucose in the blood. Machine learning algorithms help in identification and prediction of diabetes at an early stage. The main objective of this study is to predict diabetes mellitus with better accuracy using an ensemble of machine learning algorithms.The proposed ensemble soft voting classifier gives binary classification and uses the ensemble of three machine learning algorithms viz. random forest, logistic regression for the classification.The prediction analysis techniques are based on the clustering and classification. The diabetes mellitus is the diabetes disorder which is caused due to increase in sugar level in the blood.

Keywords – Ensemble classification, Prediction, Machine learning,Classifiers

INTRODUCTION

The process through which important information can be extracted from raw data is called data mining. Extraction of huge amount of data is necessary such that important information can be acquired. Large amount of data is available in every field and huge amount of time is consumed when analyzing this complete data [1]. For extracting the knowledge, data mining process is used such that no extra raw material is included. Mining is defined as the process through which important data is extracted. A prediction technique is used to discover the relationship which exists among the independent and dependent variables [2]. For predicting the future profits, prediction analysis techniques have been used in several fields. A majorchronic disease that is being of major health concern all across the world is called diabetes.When minimal amount of insulin required to maintain the rate of glucose is not outgrown, diabetes can arise [3]. A healthy diet, regular exercise as well as insulin injections are some of the basic methods to control diabetes. Several other problems like heart disease, kidney disease, blindness or blood pressure related problems can arise due to

diabetes [4] A group of metabolic disorders which result in causing abnormal insulin secretion or action is called Diabetes Mellitus (DM).Around 200 million people all across the globe are affected by DM which is thus known as the most common endocrine disorders. Over the years, it is assumed that the rate of growth of this disease will rise. Diabetes mellitus is broadly categorized into three types [5]. When a body fails to provide insulin, Type 1 Diabetes Mellitus is caused. An insulin resistance that is caused when the cells fail to use insulin properly results in causing Type 2 Diabetes Mellitus. When a pregnant women which was previously diagnosed on diabetes develops a high blood glucose level, Type 3 diabetes called Gestational diabetes occurs [6]. The most commonly found diabetes form which is also known as insulin resistance is T2D. The type of lifestyle, dietary habits and heredity are certain causes of T2d. Weight loss, polyurea [7], and polydipsia are some of the symptoms of DM. Further, depending upon the blood glucose levels, the diagnosis can vary. Mostly because of the chronic hyperglycemia [8], there arevarious complications seen in case of DM progression. There are several heterogeneous patho-physiological conditions covered by DM [9]. There are micro-and macro-vascular disorders within which mostly all the common complications have been categorized. It is very important to provide prevention and treatment becauseof the high DM mortality [10].

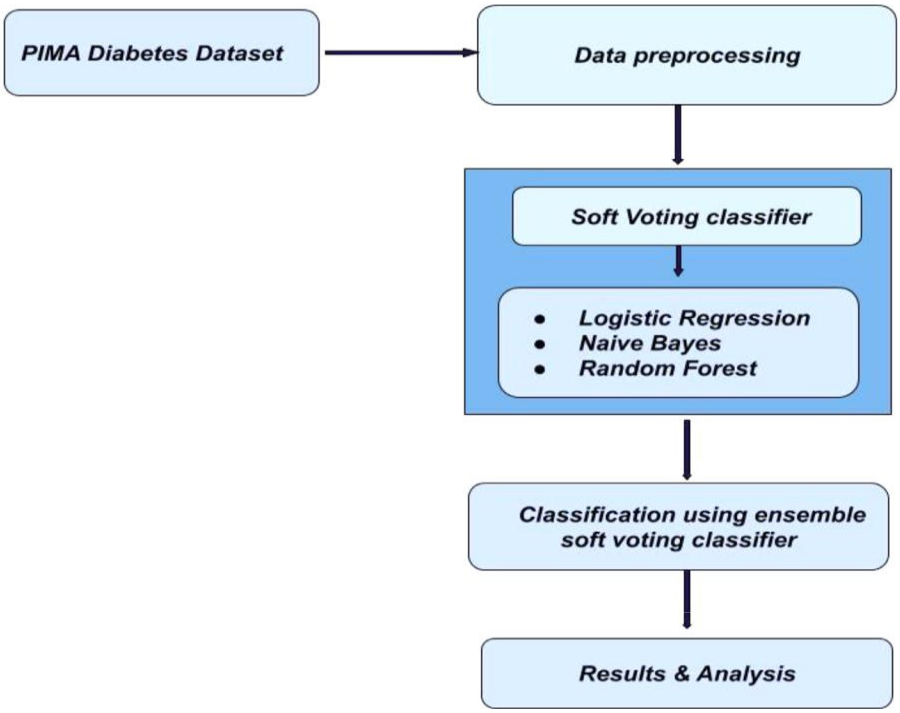
PROPOSED SYSTEM

An ensemble of machine learning algorithms viz. random forest, logistic regression, and Naïve Bayes with soft voting classifier have been proposed. The proposed methodology binary classifies the diabetes mellitus disease data into positive and negative classes.Experiments have been conducted on two different datasets viz. PIMA diabetes and breast cancer dataset.A comparison with existing techniques of the proposed methodology reveals superior results with the same number of defined parameters.Empirical evaluation of the proposed methodology has been conducted with conventional base classifiers such as AdaBoost, Logistic Regression, Support Vector Machine, Random forest, Naïve Bayes, Bagging, GradientBoost, XGBoost, CatBoost.

The naïve bayes classifier that is based on the assumption of class conditional independence which states that there are no dependencies among the attributes is called Naïve bayes classifier. It is called “Naïve” since it simplifies the computations involved. The three classifier which are applied individually and prediction result of each classifier is given

as input to voting method. The soft voting based method is applied which can generate the final result whichmaximum accuracy for the diabetes prediction.

SYSTEM ARCHITECTURE



The model building is the third phase, in which whole dataset will be divided into training set and test set. The training set will be more as compared to test set. The model of classification isapplied which take input training and test set. The method of voting based classification is applied which is combination of decision tree, support vector machine and naïve Bayes classifier for the final prediction. The decision tree is flow-chart-like tree structure in which a test on an attribute is represented by internal node, outcome of test by branch and classes by leaf nodes is called a decisiontree classifier. The SVM is statistical learning based classifier in which the decision boundaries are represented by support vectors is called SVM. The training data is represented by identifying the number of support vectors. The model is trained using the only portion ofdata. Binary classification is used to design the SVM originally.

CONCLUSION

Diabetes mellitus is an illness that is commonly found in adults now a-days. Hence the early recognition of this disease is the need of an hour. The main objective of this research work has to get the best accuracy and algorithm for predicting diabetes patients. Machine learning algorithms that have been applied in the previous five years were examined regarding their accuracy.

REFERENCES

[1] Alexis Marcano-Cede˜no, Diego Andina, "Data mining for the diagnosisof type 2 diabetes", IEEE, Vol. 11, issue 3, pp. 9-19, 2016.

[2]B. M. Patil, R. C. Joshi, DurgaToshniwal, "Association rule for classification of type -2 diabetic patients", 2010 Second International Conference on Machine Learning and Computing, Vol. 8,issue 3, pp. 7-23, 2010.

[3]Prova Biswas1,2, Ashoke Sutradhar3, PallabDatta, "Estimation of parameters for plasma glucose regulation in type-2 diabetics in presence of meal", IET Syst. Biol., 2018, Vol. 12 Iss. 1, pp. 18-25, 2018.

[4]MS.Tejashri n. Giri, prof. S.r.todamal, "data mining approach for diagnosing type 2 diabetes", international journal of science, engineering and technology, vol. 2 issue 8, 2014.

[5]P. Suresh Kumar and V. Umatejaswi, " Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017.

[6]M. Sharma, G. Singh, R. Singh, "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques", Elsevier, vol. 5, pp. 202-222, 2017.

[7]Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.

[8]Yan Luo, Charles Ling, Ph.D., Jody Schuurman, Robert Petrella, MD, "GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing", 2014 IEEE International Conference on Data Mining, Vol. 9, issue 8, pp. 12-23, 2014.

[9]AbdelghaniBellaachia and ErhanGuven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, 2010, pp. 234-239.

[10]Oyelade, O. J, Oladipupo, O. O andObagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance