Diabetes Analysis Using Various Machine Learning Techniques

Sangeeta Bairagi[1], Ankur Taneja[2]

[1,2] Department of Computer Science and Engineering, Sam College of Engineering and Technology, Bhopal (M.P)

---***---

Abstract - *Diabetes is a high-risk medical condition characterized by abnormally high blood sugar levels. It is a leading cause of death worldwide. According to increasing morbidity in recent years, the number of diabetic patients worldwide will reach 642 million by 2040, implying that one out of every ten adults will have diabetes. It is undeniable that this needs a great deal of attention. A variety of data mining and machine learning techniques have been used on the diabetes dataset to predict disease risk. The aim of this paper is to examine these machine learning techniques using output metrics and method features. The research includes the Pima Indian diabetes dataset, which includes 768 patients, 268 of whom are diabetic and 500 of whom are not.*

Key Words: Diabetes, Machine Learning, Logistic Regression, Decision, Pima

1.INTRODUCTION

Diabetes mellitus (DM) is commonly called diabetes. It is a medical problem that is severe and complex. The pancreas does not produce enough insulin, so blood sugar rises, and it affects various organs, in particular the eyes, kidneys, nerves [1]. It is for this reason that diabetes is referred to as the silent killer. Three kinds of diabetes exist: type I diabetes, type II diabetes, and gestational diabetes [2]. The pancreas produces very little insulin in the case of type I diabetes or even no insulin. Roughly 5 to 10% of all diabetes is type I and occurs not only in puberty or infancy, but also in adulthood as well [3]. Type II diabetes occurs if insulin is not adequately released by the body. Approximately 90% of diabetic patients are of type II diabetes in the world. Form II is like the third type of diabetes, gestational diabetes mellitus (GDM). In many ways since it requires a mixture of comparatively inadequate secretion of insulin. Approximately 2-10% of all pregnant women are affected by gestational diabetes, after delivery, it can progress or disappear.

Diabetes disease diagnosis and interpreting diabetes data is the difficult problem. Various machine learning methods are used for dealing with healthcare problems which are typical in nature. Most of the medical data contains non-linearity, non-normality and an inherent correlation structure. Therefore, the conventional and extensively used classification techniques like naive bayes, random forest and decision tree etc. but cannot classify the data properly. In this paper review of various machine learning methods is presented and compared their accuracy on pima Indian dataset.

2. LITERATURE REVIEW

Now a day, Diabetes is a general chronic disease which poses a great risk to individual's physical condition. The blood glucose is a main property of diabetes which is higher than the normal level, because of defective insulin secretion with special biological effects, [1]. Diabetes can direct to persistent damage and dysfunction of different tissues, especially kidneys, eyes, heart, blood vessels and nerves [2]. The distinctive medical symptoms are increased thirst and regular urination, high blood glucose levels [3]. Diabetes cannot be treated successfully with medications alone and the patients are requisite insulin therapy.

With the advancement of living standards, diabetes is becoming more and more prevalent in the everyday lives of people. Therefore, a subject worth researching is how to easily and reliably diagnose and evaluate diabetes. In medicine, diabetes diagnosis is based on fasting blood glucose, glucose tolerance, and spontaneous levels of blood glucose [3] [4]. The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors [5]. The most important problems are how to pick the correct fea tures and the right classifier for the machine learning proces ses.

In recent times, several algorithms are used to forecast diabetes, including the conventional machine learning method [6], such as support vector machine (SVM), decision tree (DT), logistic regression etc. [7] constructed prediction models based on logistic regression for different onsets of type 2 diabetes prediction in order to deal with the high dimensional datasets. In [8], authors concentrated on glucose and used diabetes, which is a multivariate regression problem, to predict support vector regression (SVR). In addition, more and more studies have used ensemble techniques to enhance the accuracy of [6]. A new ensemble method, Rotation Forest, which incorporates 30 machine learning techniques, was proposed in [9]. In [10], authors suggested a method of machine learning that modified the rules for the prediction of SVM. Machine learning approaches are commonly used to predict diabetes and produce preferred results. Decision tree is one of the common methods of machine learning in the medical field, which has the power to classify gratefully. Many decision trees are created by Random Forest. The neural network is a common method of machine learning that has improved performance in many aspects recently. So, we used decision tree, random forest (RF) and neural network to predict diabetes in this research.

2.1 Dataset

The review of machine learning methods is performed on the Pima Indian dataset for Indians [11]. All patients are females of Pima Indian heritage who are at least 21 years old. The dataset comprises 8 pregnancy features, plasma glucose concentration after a 2-h oral glucose tolerance test, diastolic blood pressure, skin fold thickness of triceps 2-h serum insulin, body mass index, pedigree feature and age of diabetes. This dataset contains 786 initial values of diabetic data including missing values which are removed, remaining dataset is 392.

2.2 Comparative analysis of machine learning methods

In this paper, comparison of three main machine learning methods named decision tree, logistic regression and random forest are compared using accuracy measure. The existing results taken from the paper [12] in which authors have compared the performance of various machine learning methods. In this paper, performance analysis of decision tree, random forest and logistic regression is done.

Decision tree is one of the most useful and efficient method in machine learning. It can handle the input data such as nominal, numerical, and alphabetical. The method can process the data with the missing errors and values. It is possible for this form of content to the number of channels and different bundles of services varies. Decision tree generates some decision rules which used to extract a significant quantity of available data. Decision tree node is generated by using information gain approach to resolve the suitable attributes in decision tree. From the highest information gain the main attribute can be selected. The random forest is also the part of machine learning which consists of aggregation of many decision trees, which results the low variance compared to the single decision tree.

The main difficulty of decision trees is that, in each step, the combination of single best variable and optimal split-point is selected, which may give different results. To overcome the weak points another method named logistic regression is applied on Pima Indian dataset. A logistic regression model computes the class membership probability for one of the two categories in the data set. A classification algorithm used to assign observations to a discrete group. Some of the examples of classification are email spam or not spam, online payment fraud or not fraud, Tumor Malignant or Benign. Using the logistic sigmoid function to return a probability value, logistic regression transforms its output.

3. RESULTS AND DISCUSSIONS

Fig 1 presented the comparison of accuracy predicted by three algorithms. It is observed that the logistic regression performed best on Pima Indian dataset. But accuracy is 76 % which has scope for improvement. In the future work it can be improved using change in the setup of hyper-parameters.

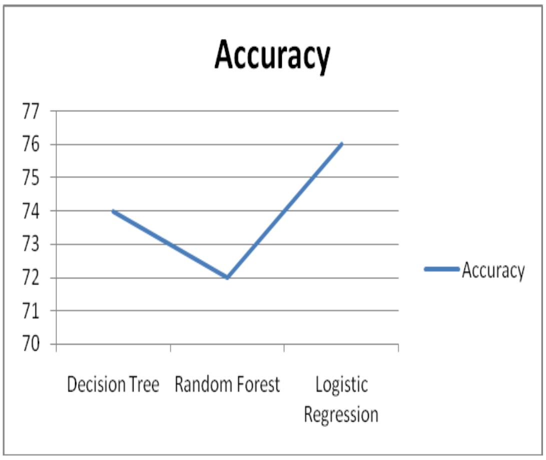

Fig. 1 Performance Evaluation

4. CONCLUSION

A review of various machine learning methods is presented in this paper with their evaluation at single place. Through this paper, authors can understand machine learning methods and usage of these methods with domain dependent or non-domain dependent features for diabetes dataset. Evaluations of these methods are also shown, from which methods of best result can be used in future for other applications. According to this evaluation, logistic regression performed best compared to decision tree and random forest for Pima Indian dataset. But the accuracy of logistic regression is 76% for Pima Indian dataset which still has some scope for improvement.

References:

[1] A. Lonappan, G. Bindu, V. Thomas, J. Jacob, C. Rajasekaran, and K. Mathew, "Diagnosis of diabetes mellitus using microwaves," Journal of Electromagnetic Waves and Applications, vol. 21, pp. 1393-1401, 2007.

[2] A. Krasteva, V. Panov, A. Krasteva, A. Kisselova, and Z. Krastev, "Oral cavity and systemic diseases—diabetes mellitus," Biotechnology & Biotechnological Equipment, vol. 25, pp. 2183-2186, 2011.

[3] M. I. N. Logical, S. BUZURA, V. DADARLAT, B. IANCU, A. PECULEA, E. CEBUC, et al., "2020 IEEE International Conference on Automation, Quality and Testing, Robotics."

[4] M. E. Cox and D. Edelman, "Tests for screening and diagnosis of type 2 diabetes," Clinical diabetes, vol. 27, pp. 132-138, 2009.

[5] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive

neuro-fuzzy inference system to diagnosis of diabetes disease," Digital Signal Processing, vol. 17, pp. 702-710, 2007.

[6] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," Expert Systems with Applications, vol. 38, pp. 8311-8315, 2011.

[7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and structural biotechnology journal, vol. 15, pp. 104-116, 2017.

[8] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," IEEE journal of biomedical and health informatics, vol. 20, pp. 39-46, 2015.

[9] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," Computer methods and programs in biomedicine, vol. 104, pp. 443-451, 2011.

[10] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes," IEEE journal of biomedical and health informatics, vol. 19, pp. 728-734, 2014.

[11] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," International Journal of Engineering Research and Applications, vol. 3, pp. 1797-1801, 2013.

[12] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019/01/01/ 2019.