



Article Title: Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

Aswathy G Ashok¹

¹Department of Computer Science and Engineering, Sivaji College of Engineering and Technology, Manivila, Anna University

ABSTRACT

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. In this work, Boosting decision tree model (BDT) is used to predict the diabetes. The proposed model is used boosting algorithm to enhance the perform ace of the model. The accuracy is different for every model when compared to other models. The Result shows that the proposed model achieved higher accuracy compared to other machine learning techniques.

Keywords: Artificial Intelligence, MATLAB PYTHON

1 Introduction

In medicine, early illness prediction is crucial. Diabetes is become the most hazardous illness worldwide. Due to our fast-food diets, diabetes is now considered a lifestyle circumstance. Recognising its symptoms is key to predicting this condition. If the person can identify diabetes illness signs, managing diet and avoiding sugar makes it easier to manage. Lack of exercise contributes to diabetes in modern lifestyles. Diabetics' blood glucose levels remain high. Diabetes is caused by the pancreas not producing enough insulin or cells not responding to insulin. Diabetic ketoacidosis (DKA), a metabolic disorder that causes nausea, vomiting, stomach discomfort, acetone breath, deep breathing, and in extreme cases, loss of consciousness, can occur in people with diabetes. Hospitals must address DKA. Hyperosmolar hyper glycaemic state (HHS) is common in type 2 diabetes and causes thirst due to high blood sugar. Type 1 and type 2 diabetics often experience shortened blood sugar because to medication. Most situations are unimportant. Effects might range from uneasiness, sating, shivering, and increased hunger to misunderstanding, changes in behaviour including



Article Title: Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

aggressiveness, convulsions, unconsciousness, and brain damage or death in severe situations. Low blood sugar causes quick breathing, sweating, and cool, soft skin, although not always. Self-treat minor to moderate instances using fast-absorbing carbs. Glucagon injections prevent unconsciousness in severe situations.

2 Recent Works

2.1 Predicting diabetes with a group of several machine learning classifiers

In their paper from 2020, Kamron Hassan Md., Ashraf Alan Md., Das D., Husain E., and Hassan M. provide a process that combines feature selection, data pre-processing, and hyper parameter optimisation. Due to the small amount of labelled data and the presence of missing values in the diabetes datasets, a firm and accurate computation of diabetes is quite challenging. The missing values rejection, filling the lost values, data standardisation, feature selection, K-fold cross-validation, and various Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, Adobos, Naive Bayes, and Boost) and Multilayer Perceptron (MLP) were used in this work to propose a robust framework for diabetes estimation.

2.2 Diabetes mellitus prediction using machine learning techniques

A method was created by Zhou Q., Quad K., Lou Y., Yin D., Jud Y., and Tang H. [2018] in which the writers were directed by two sets of data. By 2040, there will be 642 million diabetics globally, or one in every 10 persons, as a result of the disease's rise in prevalence during the past several centuries. Without a doubt, this threatening figure has to be given more attention. With machine learning's quick progress, it has been used in several medical and health domains. In order to predict diabetes, a decision tree, a random forest, and a neural network are employed in this study. The hospital physical examination statistics in Luzon, China, make up the dataset. It has 14 different qualities. In the study, five-fold cross validation was utilised to test the models. Creator chose a few techniques with the better performance to carry out independent test trials in order to confirm the approaches' universal applicability.

2.3 Diabetes illness classification and prediction utilising a machine learning paradigm

In order to choose the suitable characteristics from a collection of data, Maniruzzaman M., Raman M. J., Al-Mehedi Hasan M., et al. [2019] developed a method using logistic regression, a probability value, and an odds ratio. NHANES (National Health and Nutrition Examination Survey) is a freely accessible dataset. Using this method, the authors sank 9858 values into 6561 values by dropping the missing data. The dataset that was captured is not very accurate. From the medical laboratories, a data record is chosen, and it is only then processed individually. Using odds ratios (OR) and p-values, logistic regression (LR) is used



Article Title: Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

to find the risk variables for diabetes. To predict diabetes, validated four classifiers, including Naive Bayes (NB), Decision Tree (DT), AdaBoost (AB), and Random Forest (RF).

3 Proposed Work Explanation

To predict diabetes, a variety of models were utilised. The key concern is the accuracy of the presented model. To achieve more precision, a new model is necessary. Diabetic forecasting utilising a machine learning-based diabetic decision support system with decision level fusion. Using fuzzy logic, two machine learning approaches are combined in the proposed model. The fuzzy decision system has a higher accuracy of 94.87 than the other prevalent methods. Creator can save many lives by using this diagnosing methodology. Furthermore, the fatality rate from diabetes can be lowered if the disease is detected early and preventive interventions are used.

3.1 Data Collection

The investigation is carried out using a freely available dataset known as the Pima Indians Diabetes Database.

The aforementioned dataset includes eight characteristics.

Pregnancies

It is most likely that an increase in the number of pregnancies would increase the level of insulin in the blood.

Glucose

It is a form of sugar that is commonly consumed in everyday diets. Blood pressure: the force with which blood circulates throughout the body.

Skin thickness

Skin covers the majority of the body and varies in thickness, with the thinnest skin covering the eyelids. The soles of the feet and the palm of the hand have thick skin.

Insulin

Insulin is a hormone that lowers blood glucose levels.

Body mass index (BMI)

BMI is an individual's weight in kilograms divided by the square of their height in metres.

Diabetes pedigree function

The function that determines the likelihood of diabetes based on family history.



Article Title: Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

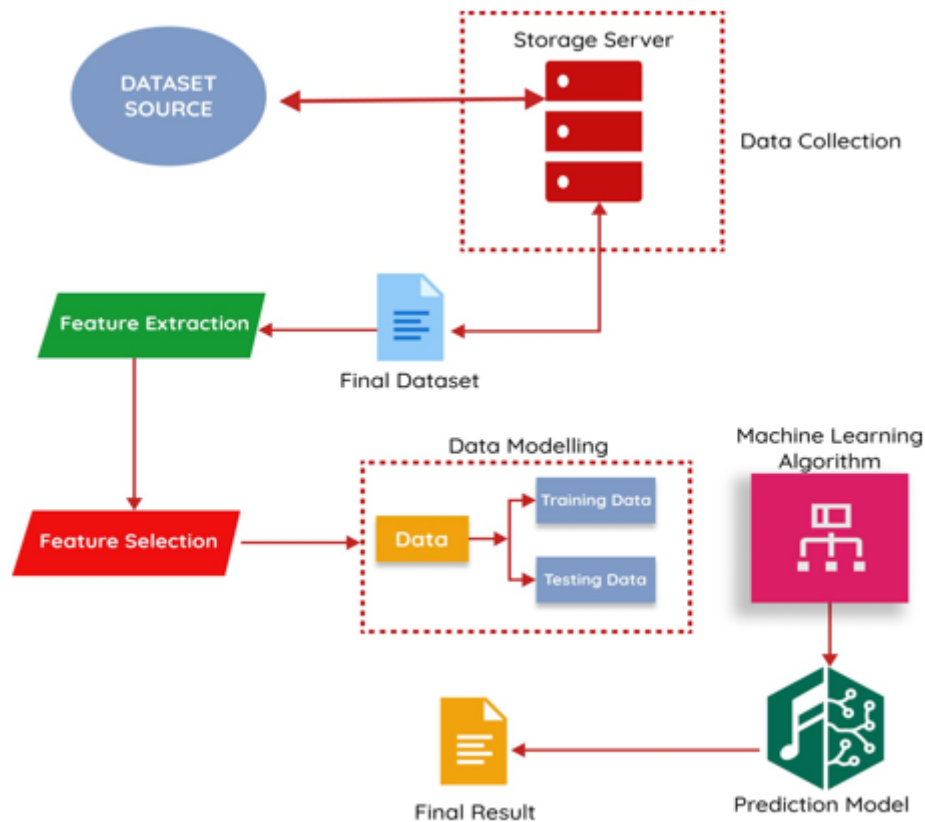


Figure 1: System Architecture

3.2 Mathematical Expression & Symbols

Each sample travels through the freshly created tree's decision nodes until it hits a specific lead. A prediction is made using the residual in that leaf. Calculate a fresh set of residuals based on the predictions produced in the previous stage. The leaves of the following decision tree will then be constructed using the residuals. Repeat the steps until the number of iterations (i.e., the number of estimators) meets the number indicated by the hyper parameter. Make a final prediction about the value of the target variable using all of the trained trees in the ensemble. All of the residuals predicted by the trees will be added to the mean computed in the first phase to get the final forecast.

Final Classification Equation is

$$F(x) = \text{sign}\left(\sum_{m=1}^M \Theta_m f_m(x)\right)$$

Where, $f_m = m^{\text{th}}$ weak classifier

Θ_m = corresponding weight



Article Title: Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

3.3 Advantages

The maximum accurate diabetes prediction of the patients is obtained comparing to other methods

4 Features of Diabetes Pedigree Function

Table 1: Diabetes Pedigree Function

Features	Description
Pregnancies	Number of prior pregnancies
Glucose	Existing glucose level of patients.
Blood pressure	Blood pressure level recorded at the specific time
Skin thickness	Patient's skin thickness level
Insulin	Amount of active insulin in the body
BMI	Body size index of the person
Diabetes Pedigree Function	Diabetes disease family history
Age	Age of an individual

5 Results and Discussion

The suggested approach omits the boosting strategy to enhance the model's performance. When compared to other models, each model's accuracy varies. The outcome shows that, in comparison to previous machine learning techniques, the suggested model reached improved accuracy.

6 Conclusion

Numerous models have been abandoned for the assessment of diabetes, but researchers' top concern has always been the accuracy of the projected models used for disease estimation. As a result, a novel technique with the maximum degree of accuracy is needed for illness prediction. Machine learning is a rapidly expanding topic of research that deals with how machines learn via engagement. In this research, a system is created that, by combining the findings of several machine learning approaches, can calculate a patient's first diagnosis of diabetes with a better degree of accuracy. In this study, a boosting decision tree model (BDT) is employed to forecast diabetes. The suggested approach omits the boosting strategy to enhance the model's performance. When compared to other models, each model's accuracy



Article Title: Early Stage Diabetes Prediction of Patients in Big Data Healthcare Using Sophisticated Machine Learning Approach

varies. The outcome shows that, in comparison to previous machine learning techniques, the suggested model reached improved accuracy.

References

1. F. Islam; R. Ferdousi; S. Rahman; H. Y. Bushra, Year: 2019, "Computer Vision and Machine Intelligence in Medical Image Analysis", London, U.K.: Springer, 2019.
2. A. Frank; A. Asuncion, Year: 2010, "UCI Machine Learning Repository", Accessed: Oct.22, 2021[Online].Available:<http://archive.ics.uci.edu/ml>.
3. Gaurav Pradhan; Ratika Pradhan; Bidita Khandelwal, Year: 2021, "A study on various machine learning algorithms used for prediction of diabetes mellitus in Soft Computing technique and applications(Advances in Intelligent Systems and Computing)", Vol: 1248, London, UK:Springer, 2021, pp. 553 – 561.
4. S. Kumari; D. Kumar; M. Mittal, Year: 2021, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," Int. J. Cogn. Comput. Eng., Vol: 2, pp. 40 – 46.
5. M. A. Sarwar; N. Kamal; W. Hamid; M. A. Shah, Year: 2018, "Prediction of diabetes using machine learning algorithms in healthcare," in Proc. 24th Int. Conf. Autom. Comput. (ICAC), Sep. 2018, pp. 6 – 7.
6. S. K. Dey; A. Hossain; M. M. Rahman, Year: 2018, "Implementation of a web application to predict diabetes disease: An approach using machine learningalgorithm," in Proc.21stInt.Conf.Comput. Inf.Technol. (ICCIT), Dec. 2018, pp. 21 – 23.
7. A. Mir; S. N. Dhage, Year: 2018, "Diabetes disease prediction using machine learning on big data of healthcare," in Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA), Aug. 2018, pp. 1 – 6.