WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Diabetes Prediction using Machine Learning

R. Ramcharan[1], J. Prathyusha[2], S. Ramesh[3], P. Ranadheer[4], D. Ramya[5], V. Ranjith[6], Prof. Sameera Sultana[7]

[1,2,3,4,5,6]B. Tech, CSE- (AI & ML), Hyderabad, India
[7]CSE- (AI&ML), Malla Reddy University (MRUH), Hyderabad, India
[1]2111CS020388@mallareddyuniversity.ac.in, [2]2111cs020391@mallareddyuniversity.ac.in, [3]2111cs020389@mallareddyuniversity.ac.in,
[4]2111cs020392@mallareddyuniversity.ac.in, [5]2111cs020390@mallareddyuniversity.ac.in, [6]2111cs020393@mallareddyuniversity.ac.in

ABSTRACT :

One of the most serious illnesses that many people have is diabetes mellitus. Diabetes mellitus can be brought on by a number of factors, including advanced age, obesity, a poor diet, genetics, high blood pressure, and lack of exercise.

Diabetes increases a person's risk of developing heart disease, renal problems, stroke, eye issues, nerve damage, and other illnesses.

At the moment, hospitals gather the data needed for a diabetes diagnosis using a variety of tests, and then treat patients according to the diagnosis.

In order to better classify diabetes, we have suggested a diabetes prediction model that takes into account both regular factors like age, insulin, BMI, glucose, and so on, as well as a few exogenous elements that cause diabetes.

The new dataset improves classification accuracy when compared to the previous dataset. Additionally, a pipeline model for diabetes prediction was implemented with the goal of increasing classification accuracy.

INTRODUCTION

Diabetes is the fast growing disease among the people even among the youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

Symptoms of Diabetes

- Frequent Urination
- Increased thirst
- Tired/Sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating
- frequent infections

LITERATURE REVIEW

Machine learning techniques are increasingly useful in the field of medicine. Many researchers have used various machine learning and deep learning techniques and algorithms to predict diabetes. Aishwarya and Vaidehi used several machine learning algorithms like support vector machines, random forest classifier, decision tree classifier, extra tree classifier, Ada boost algorithm, perceptron, linear discriminant analysis algorithm, logistic regression, K-NN, Gaussian naive gulf, Baging Algorithm. Bagging and gradient boosting classifier. To test the different models, they used two different datasets - PIMA India and another Diabetes dataset. Logistic regression gave them 96% accuracy. On the other hand, Tejas and Pramila chose two algorithms - Logistic Regression and SVM - to build a predictive model for diabetes. Data processing was done to get better results. They found that SVM performed better with 79 percent accuracy. Three different machine learning algorithms – Random Forest, Decision Tree and Naive Bayes are used to built ML model. Material pre-processing techniques are used. The results showed that the highest accuracy of 94% was obtained by the Random Forest algorithm. Deepti and Dilip used decision tree, SVM and Naive Bayes algorithms.Both articles used the Pima Indian Diabetes database. Deep learning methods for diabetes prediction. The first used a multilayer Feed-Forward neural network. A backpropagation algorithm was used to train the model. They also used the PIMA-India dataset and normalized it before preprocessing to obtain numerical stability. Their accuracy was 96%. All the above studies provided a comparative performance analysis of different machine learning algorithms. Some of them used data preprocessing and cross-validation techniques to improve accuracy, but they all focused more on comparing performance different models instead of improving one model. In this article, I focused on one model and explored techniques that not only improve accuracy, but also improve execution speed, thus increasing performance. This article shows that, in addition to the choice of algorithms, data pre- and post-processing play an important role. general improvement of the model

PROBLEM STATEMENT:

To design a machine learning model that delivers precise and interpretable predictions of diabetes risk by integrating diverse patient data. The aim is to enable early identification of individuals prone to diabetes, empowering healthcare professionals to implement timely and tailored preventive strategies for improved patient outcomes.

METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction.

Data Description: This module includes data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results.

Dataset description is given below. This Diabetes dataset contains 800 records and 10 attributes.

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Outcome	C

METHODS AND ALGORITHMS

The first steps are the selection of the model dataset and the evaluation of its characteristics. The first dataset selected for this paper is the PIMA India dataset. There are a total of nine characteristics/variables, eight of which are predictor variables and one target variable. Features are as follows: ▪ Pregnancy: several times the patient has been pregnant. ▪ Glucose: plasma glucose concentration over two hours orally glucose tolerance test. ▪ Blood pressure: diastolic blood pressure (mm Hg). ▪ Skin thickness: Triceps skinfold thickness (mm). ▪ Insulin: two-hour serum insulin (mu U/ml). ▪ BMI: body mass index (weight in kg/(height in meters. ▪ Diabetes Pedigree Function/DPF: A function that calculates the probability diabetes based on family history. ▪ Age: in years. ▪ Outcome: Categorical variable (0 if not diabetic, 1 if diabetic). it is target variable. ▪ Another dataset used is the Vanderbilt dataset. It consists of 16 characteristics, one of which is the target variable, i.e., diabetes: ▪ Patient Number: Identifies patients by number ▪ Cholesterol: total cholesterol ▪ Glucose: fasting blood sugar ▪ HDL: HDL or good cholesterol ▪ Age: Age of the patient

2. Data Pre-processing -This phase model handles inconsistent data, missing values and other impurities that could cause effectiveness of data. Data Pre-processing is done to improve the quality and to obtain accurate results. A. Missing values removal - Instances with zero as worth are removed. Through eliminating irrelevant instances, we make feature subset and this process is called features subset selection, which help to work faster. B. Splitting of data - After removal of irrelevant instances, data is normalized in training and testing the model. When data is splitted then we train the efficient algorithm on the training data set and keep test data set aside. C. Apply Machine Learning – After pre-processing of the data we will split the data into training and testing parts, 80% of the data for training part and 20% of the data for the testing part and now we will train the data using machine learning classification algorithms. These algorithms include Random Forest, Decision Trees, Naïve Bayes. We will train the data using these algorithms and after training the data we will measure the accuracy using test data.

EXPERIMENTAL RESULTS

These are the results of our project is:

CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 96%.

FUTURE ENHANCEMENT

In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

The above model is used to predict whether a person has diabetes or not using their health records and in future we can build a perfect model using deep learning techniques and providing best accuracy and further we can also build a Web application using flask so that users can give the parameters and based on those attributes the model will predict

REFERENCES

[1] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[2]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[3] Desmond Bala Bisandu, Godwin Thomas "Diabetes Prediction using Data mining Techniques," International journal of research and Innovation in Applied Sciences, volume 4, pp. 103-111, 2019.

[4] B.Suvarnamukhi , M. Seshashayee, "Big Data Processing System for Diabetes Prediction using Machine Learning Techniques," in International Journal of Innovative Technology and Exploring Engineering, volume 8, pp. 4478–4483, 2019.

[5] Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology, Volume 9, pp. 921-925, 2020. [6] N.Sneha , Tarun Gangil, "Analysis of diabetes meelitus for early prediction using optimal features selection", Journal of Big data, pp. 1-19, 2019.

[6]. World health organization,15 may 2020, Diabetes.

[7]. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 554– 559doi:10.1109/CICN.2010.109.

[8]. https://www.kaggle.com/johndasilva/diabetes