International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2020): 6.733

Diabetes Prediction Using Machine Learning

Ritik

Student, School of Engineering and Technology, Sharda University, Greater Noida, India
thakurritik870[at]gmail.com

Abstract: ***Diabetes is one of most common diseases worldwide that occurs when your blood glucose is too high. Blood glucose is your main source of energy and comes from the food you eat. Annually it costs a lot of money to take care people with diabetes. Thus the most important issue arises to predict diabetes to be very accurately. One of these methods is using machine learning model. So in this paper, we used a machine learning to predict whether a person is diabetic or not. To build the onset Diabetes predictor by machine learning we used PIMA (Indian Diabetes Datasets). After analyzing dataset and training the machine learning model using Random Forest algorithm, the average probability was 0.75 and the average accuracy was coming to predict if a person is diabetic or not was 77.3%. but to get more accurate result, we used XgBoost (machine learning library) which in result able to give more accurate details with high accuracy of 85.24%.***

Keywords: Diabetes, Machine Learning, Algorithm, prediction

1.Introduction

Diabetes is a long-lasting disease that happens when the pancreas fails to create enough insulin, or when the body cannot use the insulin produced efficiently. Blood sugar in blood is control by a hormone called insulin. Hyperglycemia is a common result of uncontrolled diabetes and, over time, causes severe damage to many organs, particularly nerves and blood vessels.

In 2015, 8.5% of adults aged 17 years or older had diabetes. In 2013, diabetes was the cause of 1.5 million deaths, and high blood glucose caused 2.3 million deaths. Diabetes patients have doubled in the last ten years worldwide. More than 200 million people are infected and about a seven percent increase in the annual predominance of diabetes in the world. People for a long time suffered from different diseases that in some cases have been able to diagnose diseases and offer them the solution in order to enhance it, but unfortunately, sometimes, due to the lack of diagnosis of symptoms in patients for a long time may even threaten the life of the patient. Therefore, many studies have been done in the field of predicting for several diseases to the extent that today's human take advantage of decision supports models and smart method to predict [1, 2]. Deferment in the diagnosis and prediction of diabetes due to insufficient control of blood glucose increases macro vascular and Capillaries difficulties risk, ocular diseases and kidney failure [1, 2]. So we proposed an machine learning model to predict diabetes that can be useful and helpful for doctors and practitioners. In this research, we used the following attributes: Number of pregnancies (describes the number of times the person has been pregnant.), Glucose (describes the blood glucose level on testing), BMI (Body Mass Index: (weight in kg/ (height in m) ^2)), Blood Pressure (describes the diastolic blood pressure), Age (years), Diabetes (Whether or not the person has diabetes/ or any family history with diabetes), Insulin (describes the amount of insulin in a 2 hour serum test), Skin (Thickness describes the skinfold thickness of the triceps), Outcome (describes if the person is predicted to have diabetes or not) [15]. It should also be noted that the dataset has no missing values and thus, filling up the dataset using algorithms will not be necessary.

2.Literature Review

Diabetes is a disease that occurs when yours blood sugar is too high. Blood sugar is the main source of energy which comes from the food you eat. Thus only remedy or a way to control diabetes is Insulin, which balanced out the blood sugar level in our body. Insulin is a hormone made by the pancreas that helps glucose from food get into your cells that used for energy for your body. But sometimes your body doesn't make enough insulin. So glucose then stays in your blood and doesn't reach cells. To provide energy to cells we give artificial insulin [2].

Blood sugar in normal people after meal hover around 70 to 80mg/dl. Whereas, for diabetic person blood sugar goes 170mg/dl and above shown Figure 1 [2].

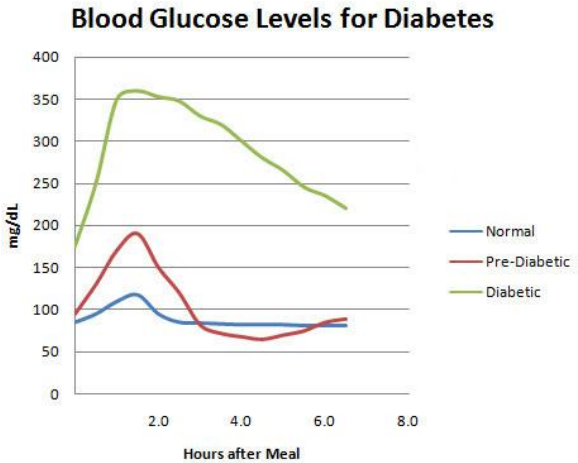

Figure 1: Blood sugar chart

There are two types of diabetes. In Type I diabetes, obliteration of beta pancreatic cells damage insulin construction and in type II, there is a progressive insulin confrontation in the body and ultimately may yield to the obliteration of pancreatic beta cells and faults in insulin production. In type II diabetes, it is known that genetic issues, obesity and lack of physical activity have a vital part in a person [1].

Even though cause of type I is unidentified, but according to

researchers it is caused by lack of insulin or no insulin produced. In type I beta cells have been destroyed. Number of factor that may indicate a greater risk of type I diabetes [2]:

- It can occur at any age.
- If a person has any family history of type I diabetes, it upsurges high risk.
- Lack of insulin production due to auto immune disease.

Type II diabetes occurs when the body either doesn't produce enough insulin or the insulin produced doesn't work properly (known as insulin resistance). In type II beta cells destroyed over time. Number of factor that may indicate a greater risk of type II diabetes [2]:

- Mostly occurs in adults.
- Genetic-a family history of type II diabetes increase the risk to young person.
- Weight-Excess body fat can cause insulin resistance.
- Poor diet-Diet with high cholesterol, calories and fat increase body's resistance to insulin.
- Lack of exercise-being inactive can cause insulin resistance and increase high risk of type II.

A practical approach to this problem is the application of regression analysis where past data is combined into some functions. The result is an equation in which both x_j inputs are multiplied by w_j; the sum of all product is constant, and then output $y= \sum x_j w_j +$, where $j = 0\ldots. n$.

2.1 Previous studies

- The author in [16] used data mining procedure for predicting the diabetes from medicals records. The author was inspired by death of people caused by diabetes in the world which involves avoiding the complication of disease. He came up with idea to develop a model using data mining which could classify diabetic patient control level based on their past medical data. The author used three data mining techniques which are Naïve Bayes, Logistic and J48. The researcher was executed using WEKA application which is data mining tool. The results showed that Logistic algorithm gave accuracy of 74.4%. Naïve Bayes gave accuracy of 74.2% and J48 gave accuracy of 73.5%. This proved that Logistic algorithm was more accurate than Naïve Bayes and J48. The research limited only to type II diabetes.
- The authors in [18] were intended to predict the diabetes types of patients based on their physical health and medical records using boosting ensemble technique, which internally use random committee classifier. The evaluation result of the technique showed accuracy gave a F-measured of 0.82 and ROC area of 0.82 for diabetes type I and II.

3.Methodology

By studying thoroughly the history of Diabetes, a number of factors have been recognized that have an impact on determining patients' cases in the subsequent period. These factors were prudently studied and coordinated with an appropriate number for coding the computer within the modeling environment. These factors were categorized as input variables and output variables that reflect some possible levels of disease status in terms of the assessment system. We will follow given data flow graph for prediction model evaluation shown in Figure 2.

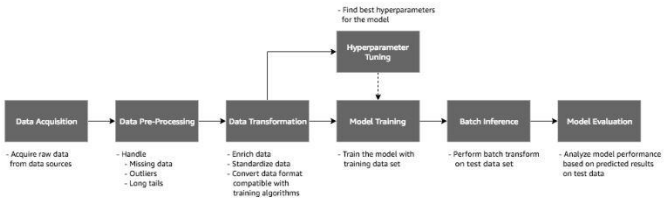

Figure 2: Flow Chart

3.1 Input variable

The specified input variables are those that can be obtained simply from the file system and the registry of diseases. Input variable are:

Table 1: attributes in the Data set

No.	Attribute name
1	Pregnancies: numbers of pregnancies
2	Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Blood Pressure: Diastolic blood pressure (mm Hg)
4	Skin Thickness: Triceps skin fold thickness (mm)
5	Insulin:-Hour serum insulin (mu U/ml)
6	BMI: Body mass index (weight in kg/ (height in m) ^2)
7	DiabetesPedigreeFunction: Diabetes pedigree function
8	Age: Age (years)
9	Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

These given factors were converted into suitable format for data analysis using pandas and numpy shown in Figure 3.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 3: Input Data Transformation

3.2 Correlation

When two sets of data are strongly linked together we say they have a high Correlation. It is Positive when the value increase together, and It is Negative when one value decrease as the other increases.

To find whether if there any correlation in PIMA dataset attributes, we plot a heat-map using matplotlib (plotting library in python) and seaborn (to get statistical data), after executing analyzed data via matplotlib we get a graph shown in Figure 4.

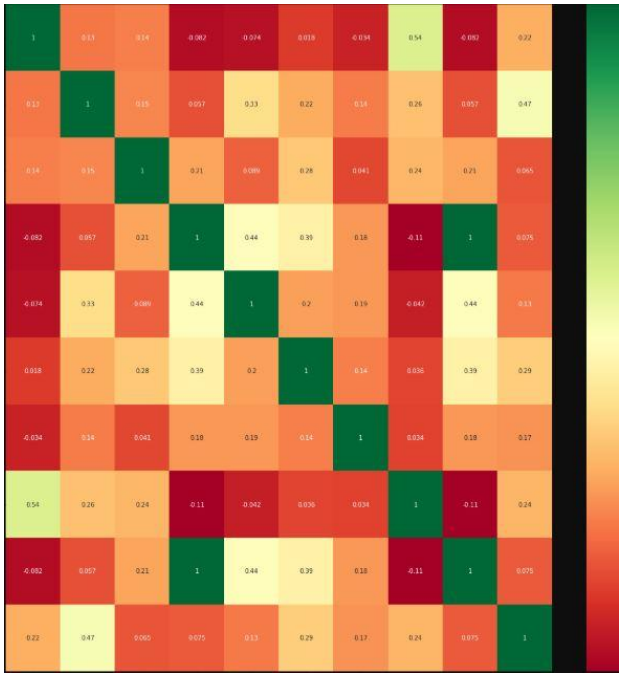

Figure 4: Heat-map

We can see all the correlated data in Figure 5. This particular correlation lies between-1 to +1.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Figure 5: correlation in Data set

3.3 Model Evaluation

As we mentioned above, the purpose of this experiment was to identify whether or not the person has diabetes. We used sklearn Regression algorithm, which predict a continuous-valued attribute associated with an object.

To predict a dependable variable value (y) based on the given independent variable (x). So, this regression technique find out relationship between x (input) and y (output). It follow **Y = mX+c** equation. We use same concept where there are more than two variable. This is known as multiple regression. For multiple regression equation of hyperlane is used, that is:

y =b0 + m1b1 + m2b2 + m3b3 + … … mnbn shown in Figure 6.

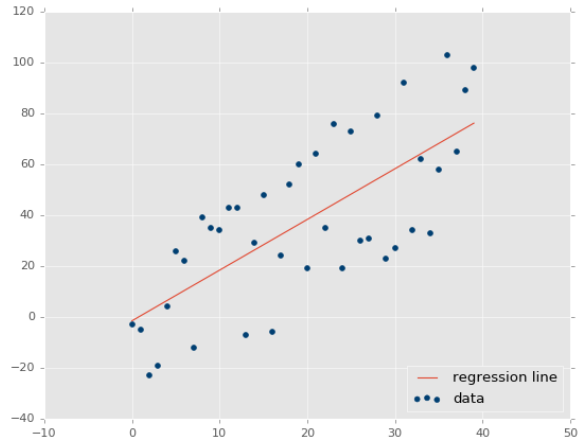

Figure 6: Regression

The proposed model is implemented in sklearn regression environment. The dataset for the diagnoses of diabetes were gathered from Pima Indians Diabetes Database which contain 768 samples and 9 attributes (as seen in Figure 1). After training and validating, the network, the following results were obtained.

The average probability was 0.75 and the accuracy of model to decide whether or not the person is diabetic was 77.3%.

- But we need more accurate outcome for prediction, so we use Xgboost algorithm.

3.4 Xgboost

Xgboost is short for eXtreme Gradient Boosting package. It is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework by (Friedman, 2001) (Friedman et al., 2000). The package includes efficient linear model solver and tree learning algorithm. The package is made to be extendible, so that users are allowed to define their own objectives easily. It has few features:

- Speed-xgboost is 10 times faster than gbm (gradient boosting classifier).
- Input Types-Xgboost has several types of input data such as Dense Matrix, Sparse Matrix, Data File, xgb. DMatrix (own class).
- Sparsity-xgboost accepts sparse input for both tree booster and linear booster.
- Customization-xgboost can customize objective function.
- Performance-xgboost can give better performance on several dataset.

After training and executing the model, the accuracy of the diabetes prediction was (85.24%).

4.Conclusion

In this paper, machine learning model was used to predict diabetes. You can design and implement complex medical processes using software. The software systems are more effective and efficient in various medical fields including predicting, diagnosing, treating and helping the surgeons,

physicians, and the general population. These systems can be implemented in a parallel way and are distributed in different measures. The diabetes dataset contain 768 samples with 9 attributes. After training, validating, and testing the dataset, we got (77.3%) accuracy with sklearn Regression but with XgBoost we got (85.24%) accuracy.

References

[1] World Health Organization (WHO), "Definition, Diagnosis and classification of diabetes mettlius and its complication", part 1. WHO/NCD/NCS/2016.2, (**2016**).
[2] H. Temurtas, N. Yumusak and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", Expert System, vol.36, (2009), pp.8610–15.
[3] A. Chavey, M. Kioon and D. Bailbé, "programming of beta-cell disorders and intergenerational risk of type 2 diabetes Diabetes", Maternal Diabetes, vol.40, no.5, (2014), pp.323-30D.
[4] D. Manzella, R. Grella, A. M. Abbatecola and G. Paolisso, "Repaglinide Administration Improves Brachial Reactivity in Type 2 Diabetic Patients", Diabetes Care, Vol.28, (2005), pp.366– 71.
[5] Box, G. E. P.; Jenkins, G. M.; and Reinsel, G. C.2008. Time Series Analysis: Forecasting and Control, 4th Edition. Wiley).
[6] A. Morteza et al., "Inconsistency in albuminuria predictors in type 2 diabetes: a comparison between neural network and conditional logistic regression", Translational Research, vol.161, No.5, (2013), pp.397-405.
[7] M. J. Sernyak et al., "Association of diabetes mellitus with use of atypical neuroleptics in the treatment of schizophrenia", American Journal of Psychiatry, (2014).
[8] M. Thirugnanam et al., "Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach. "Procedia Engineering, Vol.38, (2012). pp.1709-118,.
[9] H. R. Marateb et al., "A hybrid intelligent system for diagnosing microalbuminuria in type 2", (2014). pp.34-42,
[10] Andreassen, S.; Benn, J. J.; Hovorka, R.; Olesen, K. G.; and Carson, E. R.1994. A probabilistic approach to glucose prediction and insulin dose adjustment: Description of metabolic model and pilot evaluation study. Computer Methods and Programs in Biomedicine 41: 153–165.
[11] Bunescu, R.; Struble, N.; Marling, C.; Shubrook, J.; and Schwartz, F.2013. Blood glucose level prediction using physiological models and support vector regression. In Proceedings of the IEEE 12th International Conference on Machine Learning and Applications (ICMLA). Miami, FL: IEEE.
[12] Diabetes Control and Complications Trial Research Group.1993. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. New England Journal of Medicine 329 (14): 977–986.
[13] Salah, M., Altalla, K., Salah, A., & Abu-Naser, S. S. (2018). Predicting Medical Expenses Using Artificial Neural Network. International Journal of Engineering and Information Systems (IJEAIS),
[14] Pima Indians Diabetes DataBase, Data Obtained From: http://www.liacc.up.pt/ML/statlog/datasets/diabetes/diabetes.doc.html
[15] T. M. Ahmed, Using Data Mining To Develop Model For Classifying Diabetic Patient Control Level Based On Historical Medical Records. Journal of Theoretical and Applied Information Technology, Vol.87 no. (2), (2016), pp.316-350.
[16] R. Ali, M. H. Siddiqi, M. Idris, B. H. Kang and S. Lee, "Prediction of diabetes mellitus based on boosting ensemble modeling". In International Conference on Ubiquitous Computing and Ambient Intelligence (pp.25-28). Springer International Publishing, (2014).
[17] Time Magazine.2013. The 25 best inventions of the year 2013. http://techland.time.com/2013/11/14/the-25-bestinventions-of-the-year-2013/slide/the-artificial-pancreas/, accessed April, 2014.
[18] A. Elzamly, S. S. Abu Naser, B. Hussin and M. Doheir, "Predicting Software Analysis Process Risks Using Linear Stepwise Discriminate Analysis: Statistical Methods", Int. J. Adv. Inf. Sci. Technol, vol.38, no.38, (2015), pp.108-115.
[19] Nesreen Samer El_Jerjawi, and Samy S. Abu-Naser, "Diabetes Prediction Using Artificial Neural Network", International Journal of Advanced Science and Technology, vol.124, (2018), pp.1-10.
[20] Gaganjot Kaur and Amit chhabra, "Improved J48 Classifiaction Algorithm for prediction of Diabetes", International journal of computer Application (0975-8887), volume 98-No.22, July 2014.
[21] Tianqi Chen, Tong He, "xgboost: eXtreme Gradient Boosting", Package version: 0.90.0.2, August 1, 2009.
[22] S. S. Abu Naser and M. H. Al-Bayed, "Detecting Health Problems Related to Addiction of Video Game Playing Using an Expert System", World Wide Journal of Multidisciplinary Research and Development, vol.2, no.9, (2016), pp.7-12.
[23] Nahla Barakat, Andrew P. Bradley and M. Nabil Barakat, "Intelligible Support Vector Machine for Diagnosis of Diabetes Mellitus", research gate publication.
[24] Abu Naser, S. S., & Alhabbash, M. I. (2016). Male Infertility Expert system Diagnoses and Treatment. American Journal of Innovative Research and Applied Sciences, 2 (4).
[25] Abu Naser, S. S., & Alawar, M. W. (2016). An expert system for feeding problems in infants and children. International Journal of Medicine Research, 1 (2), 79-82.
[26] AlZamily, J. Y., & Abu-Naser, S. S. (2018). A Cognitive System for Diagnosing Musa Acuminata Disorders. International Journal of Academic Information Systems Research, (IJAISR) 2 (8), 1-8