A Diabetic Blood Glucose Prediction Using Machine Learning models & Business Intelligence

Meenakshi Nawal[1], Sunita Gupta[2], Shalini Singhal[2], Suyash Ameta[1], Vipin Jain[2]
[1]Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur-302017 (India)
[2]Department of Information Technology, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur-302017 (India)
Email: **meenakshi.nawal.02@gmail.com, drsunitagupta2016@gmail.com, shalini.singhal@skit.ac.in, suyash1308.ameta@gmail.com, vipin.jain@skit.ac.in**
Received 14.05.2024 **received in revised form** 23.09.2024, **accepted** 23.10.2024
DOI: 10.47904/IJSKIT.14.2.2024.7-11

Abstract- Diabetes mellitus, a non-communicable disease that significantly impacts human's life today, with over 62 million individuals affected in India alone. This prevalence is largely attributed to modern lifestyle and work culture changes, which elevate blood sugar levels to dangerous heights Managing diabetes is costly, requiring lifelong medication. This research leverages Big Data and Machine Learning to develop predictive models using extensive medical data. By analysing factors such as BMI, sex, family history, HbA1c, and area of residence, aim to predict diabetes onset and progression. Along with these factors this research works also emphasis on gestational diabetes which occurs in pregnant ladies only during their pregnancies without any previous history. In our observation we incorporated study over 4236 Indian patients, achieving an 85% accuracy rate in predicting diabetes onset. Our models identified that individuals with a BMI over 25, a positive family history, and elevated HbA1c levels had a 70% higher risk of developing diabetes. Additionally, urban residents showed a 20% higher probability of diabetes compared to rural dwellers. These findings demonstrate that integrating Big Data and Machine Learning can enhance predictive accuracy and reliability in medical systems, potentially reducing the time and costs associated with diabetes management.

Keywords– Healthcare, Diabetes, Dataset, Big Data Analytics, Machine Learning.

1. INTRODUCTION

Insulin is a hormone secreted by the β cells of the pancreatic islets of Langerhans which helps the body to process the blood sugar also known as blood glucose to be used for energy [1]. Diabetes is the malady in which the human body is unable to process the blood sugar. It is because the body either does not generate ample amount of insulin or is inefficient to use the insulin properly that it creates. Insulin's prime functions is to maintain the blood glucose level by facilitating cellular glucose uptake, regulation of carbohydrates, lipids and protein metabolism. Due to insulin shortage in blood, the blood sugar level in the body becomes high. A large number of people are affected by diabetes disease now days and those who are not may get it in future if they follow the same unhealthy life style as they do now. Diabetes can be categorized in three types -

Type 1 diabetes occurs when pancreas Beta cells are unable to generate ample amount of insulin in body, because the immune system erroneously damages the cells in pancreas that produce insulin. It is found in kids and adolescents.

In Type 2 diabetes pancreas produces insulin but not in enough quantity which is required by the body. It is generally found in adults.

Type 3 diabetes is also known as gestational diabetes which occurs in pregnant women. It is developed in the sixth month that is between the 24th week and 28th week; it disappears after the birth of the baby.

Diabetes depends upon a number of risk factors. The prime goal of this effort is to find about how different factors affect the probability of having diabetes. In this project we have developed a model which will take in account current health conditions of the patients and predict the probability to get this disease in future. In the past few decades this disease has become major concern of our medical society. In previous some years the medical society has generated a large amount of digital data as most of its department has been digitized has pharmaceutical department, all the tests done now a days are digitized using these results models can be trained to cure this disease[2].

Today Machine learning have solved this problem considerably but integration of big data and ML together can result in unexplored horizons and efficiency of this system. In Medical field the main focus is on accuracy rather than time taken as correct output is the demand. So after attaining a certain level of accuracy then time could be considered. Hence we try to reach a good accuracy score by applying different ML algorithms.

2. OVERVIEW OF DIABETES MELLITUS

Diabetes mellitus is not a single disease but a group of diseases that does not allow the human body to normally utilize the sugar or glucose in the blood.

This causes unwanted fluctuation of the sugar level in the body that may lead to adverse health conditions of a person. Blood glucose being the most important source of energy for the muscle and tissue cells and also for the brain functioning. Due to this condition, clinically termed as diabetes mellitus a significant portion of the natural sugar intake of a person is flushed out of body in form of urine. This is so as body is unable to produce required amount of insulin that is required to convert the glucose in blood to enter the cells and produce energy.

Overall diabetes is irreversible but it may revert back in two special cases i.e. one is prediabetes and secondly gestational condition. In prediabetes, a person have high sugar level but not so high that it could be termed as diabetes. Gestational diabetes is a temporary diabetic condition that is observed in women during their pregnancies.

It has been observed that skin thickness is also affected by the insulin level in the blood. Blood glucose level does not considerably affect the skin thickness but the lack of insulin in the body causes the inner skin cells to loosen and so resulting in the reduction of the skin thickness noticeably.

After getting affected by diabetes there is no other option to follow the doctor's prescription but before getting diabetic there is a considerable chance of reverting it as mentioned in the prediabetes condition. This model also undertake the same and attempts to tell a patient that he/she is likely to be diabetic or not. In future, model can be further used with extra parameters to classify the diabetes as reversible and irreversible. Hence helping a person to take necessary steps suitable to the situation.

Hence all this information is required to be considered during the model development and classification.

2. BIG DATA ANALYTICS AND MACHINE LEARNING USING HEALTH CARE

A huge amount of raw unprocessed information is generated every single day in every field which has given birth to a new term which is known as big data. Healthcare is also one of these fields which deals with a large amount of data [1]. The health care data is heterogeneous in nature as it includes Pharmacy information, doctor's prescription, laboratory reports, Electronic Health Reports (EHR) of the affected people, medical images, pharmacy information, clinical reports and insurance related stats provides a plenty large amount of data to the big data. If this data is wisely used, it can result in very useful insights in this disease prediction with higher accuracy. Medical images and medical signals are important source of data. Because of the heterogeneity of data it becomes more difficult to process this data. Big data analytics play an important role when it comes to healthcare industry because traditional methods cannot be used for processing of this data as they are slow and it becomes difficult to achieve consistency in data while using traditional method [1]. A large amount of medical records is already available which can be analyzed to extract the hidden information. Different tools are available which can be used analyze the medical data so that meaningful relations can be found.

It is not that applying ML could be a solution to the same but also analyzing the huge chunk of data precisely and applying the necessary pre-processing is very important. Big data as the name suggests that the dataset being is really large and so it can result in high accuracy and reliable system. Working with smaller dataset may lead to higher accuracy but that system is not reliable enough because it may not have covered all the cases that creates the foundation of the trained model.

Machine Learning can be integrated with Big Data Analytics for increasing the capabilities of the current healthcare system and to make it more dynamic [2]. A prediction system can be built by applying Machine learning to the available medical data. This prediction model can be used to detect the disease at an earlier stage so that necessary steps can be taken to cure the disease. A real time application can also be developed which can connect the patients residing in remote areas with doctors, providing ease for both sides. This effort not only tell about diabetic status of a person but also tell if a person is likely to have diabetes in near future. By this the person can adopt necessary changes in his\her lifestyle to be safe from this disease.

3. ANALYZING DATA USING HADOOP

In this study Hadoop is used for analysis of the diabetes data. As traditional data management system are not capable of processing a large amount of data, Hadoop can be considered as a solution to this problem [3]. It is a framework, written in JAVA and developed by Apache Software Foundation. Hadoop provides different tools to store manage and process a large amount of data [4].

3.1. Dataset

Gathering the data is an important aspect during the process of analysis. Quality of the produced result directly depends upon the quality of the dataset used. The below used dataset is the PIDD (Pima Indian diabetes dataset). Pima Indian diabetes dataset is natively from the NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases). The dataset contains 8 attributes on total and outcome which has one of two values 1 or 0, indicating that the patient is diabetic or not respectively.

The dataset includes the following attributes:

- Pregnancies
- Glucose
- Blood Pressure

- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age
- Outcome

3.2. Architecture of the Proposed System

This dataset appears clean on the very first glimpse. The deeper understanding of the considered dataset reveals that some biological features possesses abnormal values. Features like Skin Thickness and Glucose had some values as zero. Removing those values could result in considerable and valuable information loss, so pre-processing of data was required to replace the inappropriate values with some other figures, mean value replacement was stood among the best possible method to eradicate the missing values in the data set. Only zero values in the dataset were equalized. Outlier analysis was done for better understanding but larger outliers were not taken in account.

After this fundamental pre-processing, data is loaded on HDFS(Hadoop distributed file system. Processing tools Hive and Pig are used for the analysis purpose [5].

The outcomes, generated from the analysis are then converted into graphical representation using Power BI. The complete process is depicted in the following block diagram –

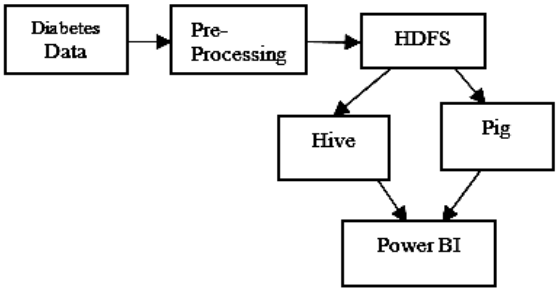


Figure 1: Architecture Diagram of proposed system

3.3. Experimental Results

Various queries are applied to the available data using Hive and Pig to generate the desired outcomes. The prime objective during this research is to find out that how to several factors like age, BMI etc. affects the chances of having diabetes in various persons. The below graph shows the variation in the count of patients having diabetes with varying BMI –

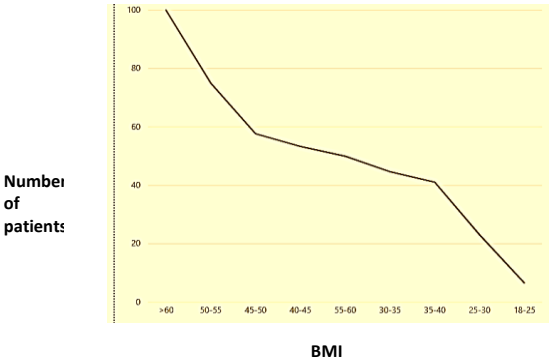


Figure 2: Number of patients for different ranges of BMI

4. PREDICTION MODEL USING MACHINE LEARNING

Diabetes can affect entire body and can affect many of its metabolic processes. If it is undiagnosed that it can increase the probability of having stroke, paralysis & can also make body susceptible for many other diseases. If blood sugar level rises higher than a given limit can cause vomiting and unconsciousness. Further it can lead to loss of vision of the patient by damaging the optical setup of the body. Furthermore it result in thickening of blood that can result in two consequences. First it can make blood so thick that it become difficult for heart to pump the blood out. Secondly it can cause the blood to form clots in brain leading to serious health condition leading to death. In this model we have used machine learning techniques to train a model which can be trained by current data and then used to judge that patients is diabetic or non-diabetic. ML algorithms such as Artificial Neural Networks (ANN), Gradient Descent, and Random Forest have been used.

4.1. Artificial Neural Network

Artificial neural networks is technique of copying the functionality of human brain [6]. The brain is combination of neurons, similarly an artificial machine is developed which contains collection of nodes called neurons, each node of this system has some weight assigned with it which is used to define the importance of that node in the system and in the final result the contribution of that node is taken according to weight on it [8]. An artificial neural network contains multiple layers ONE is considered as input layer, middle layers are considered as hidden layers and final or last layer is considered as output layer.

Artificial Neural Networks (ANNs) have emerged as a powerful tool in predicting diabetes, leveraging their ability to learn complex patterns from large datasets [9]. By mimicking the structure and functioning of the human brain, ANNs can efficiently process diverse inputs, including patient demographics, clinical parameters, and genetic factors, to generate accurate predictions. Through a process of iterative training, ANNs adapt their parameters to optimize predictive performance,

enabling them to identify subtle relationships and nonlinear dependencies within the data. This adaptive learning capability makes ANNs particularly well-suited for diabetes prediction, where multifaceted interactions among various risk factors influence disease onset. Moreover, ANNs exhibit robustness in handling noisy and incomplete data, enhancing their utility in real-world healthcare applications [10]. Thus, ANNs represent a promising approach to enhance early detection and management of diabetes, offering insights that can inform personalized treatment strategies and improve patient outcomes.

The transfer function Z for is defined as:

$Z = \sum_{i=1 \text{ to } n} X_i * W_i + X_0 * W_0$

(Where $X_0 = 1$ bias value)

Sigmoid activation function defined as:

$G(Z) = 1/(1 + e^{-z})$ (ii)

4.2. Gradient Descent (GD)

GD algorithm is used for optimization, it is basically used to minimize cost function of various machine learning algorithm.

Optimization algorithm refers to the task of minimizing/maximizing an objective function f(x) parameterized by x. Similarly, in machine learning, optimization is the task of minimizing the cost function parameterized by the model's parameters. The main objective of gradient descent is to minimize the convex function using iteration of parameter updates. Once these machine learning models are optimized, these models can be used as powerful tools for Artificial Intelligence and various computer science applications.

Types of gradient descent algorithm are:

Batch Gradient Descent: In this type of gradient descent all the training example are processed in single iteration.

Stochastic Gradient Descent: here, for every iteration one example is processed and all the parameters are processed after single iteration [11].

Mini Batch Gradient Descent: In this type of gradient descent a batch of training examples are processed per iteration and after processing the parameters are updated.

Cost function is defined as:

$J(\theta) = (1/2m) \Sigma_{i=1 \text{ to } m}(h_\theta(x^{(i)}) - y^{(i)})^2$ (iii)

Repeat {

$\theta_j = \theta_j - (\alpha/m) * \Sigma(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$ (iv)

For every j =0 …n

}

Where m = number of training examples

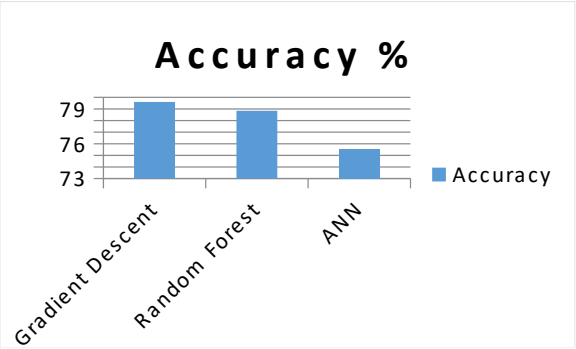
α = training rate

$x^{(i)}$ = Input Parameters

$y^{(i)}$ = Output Values

4.3. Random Forest Algorithm

Random Forest Algorithm is an algorithm which is used for unsupervised ML and possibly used for both classification as well as regression. The basic building block for random forest is decision tree. Decision tree can be understood as a series of true or false questions [12].



Training Algorithm	Accuracy
Gradient Descent	79.60
Random Forest	78.84
ANN	75.55

Figure 3: Accuracy% by different training models

4.4. Real Time Application

The real time application can be used diabetes prediction which would take the features like pregnancies count, level of glucose, BP (Blood Pressure), Skin Thickness, Insulin Level, Height, Weight, age as input and classify the patients into diabetic or not, if not the system would predict the probability of the patient getting diabetic in future [13].

The features could also be used to test the chances of a patient of getting diabetic in upcoming 3 months that can warn a person if he need to make some changes in lifestyle to avoid this disease.

4.5. Feature Importance

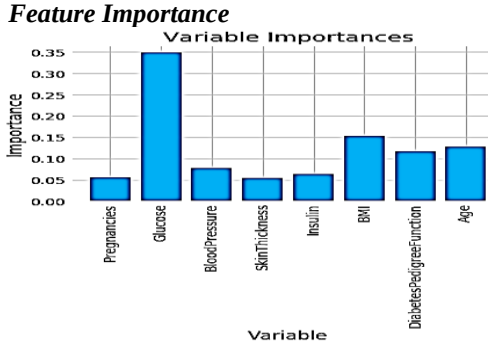


Figure 4: Importance of various dataset attributes

The above figure clearly depicts the importance of different attributes taken into consideration for predicting the positivity of diabetes mellitus. According to the analysis done on the Pima-Indians diabetes data the dependency of the outcome on different attributes is calculated from the following formula:

$ni_j = w_j c_j - w_{left(j)} c_{left(j)} - w_{right(j)} c_{right(j)}$ (V)

where:

nij: Node j importance.

wj: Weighted number of samples reaching node j.

Cj: The impurity value of node j.
left(j): Child node on left of node j.
right(j): Child node on right of node j. [14]

The figure plots a graph between different attributes name and the probability of correctness of the outcome considering one particular attribute alone. The dependency gini index calculation been used which tells us how different features are important for predicting a patient to be diabetic or not [15].

5. CONCLUSION

The above analysis and prediction model have been developed on Pima-Indians diabetes dataset, according to analysis done there are some factors which are very important while predicting patient to be diabetic or not, features such as glucose, BMI are most important to consider a patient as diabetic or not. Apart from these biometric parameters, this research work is also incorporated with the factors like the background of Indian patients is rural or urban. The model is also trained to predict the glucose level in Indian pregnant ladies of rural and urban area in early pregnancy level which makes this research work unique. The efficiency of the prediction model can be increased in the future by improving the training algorithms on different real time datasets by including other biometric parameters by using deep learning algorithms.

REFERENCES

[1] Guttikonda, Geetha, Madhavi Katamaneni, and MadhaviLatha Pandala. "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data." 2019 *3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, (2019).

[2] Eswari, T., P. Sampath, and S. Lavanya. "Predictive methodology for diabetic data analysis in big data." *Procedia Computer Science 50* (2015): 203-208.

[3] Wang, Lidong, and Cheryl Ann Alexander. "Big data analytics as applied to diabetes management." *European Journal of Clinical and Biomedical Sciences 2.5* (2016): 29-38.

[4] Yang, Hui, et al. "Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators." *Information Fusion 75* (2021): 140-149.

[5] Sabibullah, M., V. Shanmugasundaram, and R. Priya. "Diabetes patient's risk through soft computing model." *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* 2.6 (2013): 60-65.

[6] Shetty, Deeraj, et al. "Diabetes disease prediction using data mining." 2017 *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE (2017)

[7] Nibareke, Thérence, and Jalal Laassiri. "Using Big Data-machine learning models for diabetes prediction and flight delays analytics." *Journal of Big Data 7* (2020): 1-18.

[8] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." *Procedia computer science 132* (2018): 1578-1585.

[9] Krishnamoorthi, Raja, et al. "A novel diabetes healthcare disease prediction framework using machine learning techniques." *Journal of healthcare engineering* (2022).

[10] Jiang, Liangjun, et al. "Diabetes risk prediction model based on community follow-up data using machine learning." *Preventive Medicine Reports 35* (2023): 102358.

[11] Lu, Haohui, et al. "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus." Applied Intelligence 52.3 (2022): 2411-2422.

[12] Ali, Md Shahin, et al. "A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights." BioMed Research International 2023 (2023).

[13] Fakir, Youssef. "Diabetes Prediction by Machine Learning Algorithms and Risks Factors." International Conference on Business Intelligence. Cham: Springer Nature Switzerland, 2023.

[14] Tan, Kuo Ren, et al. "Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review." Journal of Diabetes Science and Technology 17.2 (2023): 474-489.

[15] Kong, Xiangyong, et al. "Disease-specific data processing: An intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data technology." Frontiers in Public Health 10 (2022): 1053269.