IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

Diabetes Prediction using Artificial Intelligence and Machine Learning

R. A. Jamadar[1], Atharv Damle[2], Om Patil[3], Prajwal Zarekar[4]
Department of Information Technology[1,2,3,4]
AISSMS Institute of Information Technology, Pune, Maharashtra, India

Abstract: *Diabetes is a fatal disease and its developments must be monitored continuously. If one is affected with this disease, it may stay throughout one's life, depending upon the stage and severity. Furthermore, having too much glucose in the blood can cause health issues including kidney disease, heart disease, stroke, eye problems, dental disease, foot problems, nerve damage. So, one must take steps to avoid these complications and oversee one's diabetes. The most common type of diabetes is type 1 and type 2. In this type of diabetes, the patient faces problems like the body is not able to produce or use insulin. In other kinds of diabetes, like gestational diabetes, which crop up during pregnancy. Gestational diabetes causes high blood sugar that can affect pregnant women's and baby's health. For diagnoses and administration of diabetes various Machine Learning and Data Mining methods are used. This study focuses on new developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes. In this study, the machine learning algorithms are used to classify diabetes patients.*

Keywords: Diabetes, Classification, Prediction, Machine learning, Accuracy

I. INTRODUCTION

Diabetes mellitus (DM) is defined as the collective metabolic malfunctioning in which humans have increased blood sugar, either because the pancreas is not capable of generating sufficient insulin, or owing to incompetence of cells to respond to insulin produced. This results in multiple medicinal occurrences like polyphagia, polyuria, and polydipsia. Till date, diabetes continues to be a public health dilemma all over the world. In developed countries, this is becoming the prominent reason of death and has become the fourth or fifth most common non-communicable diseases worldwide. It is estimated that by 2025, 3000 lakh individuals will be diabetic or pre-diabetic in the entire world. The maximum increase in DM is seen in developing countries such as India in the last few years. DM is a state where the blood glucose level, also known as blood sugar of our body becomes too high. Diabetes can occur in people at any age and have three foremost categories of diabetes specifically type 1, type 2, and gestational diabetes. Gestational diabetes is believed to arise due to many hormonal and other changes, which occur in the body during pregnancy, while some women develop insulin resistance in the body. Type 1 diabetes mellitus (T1DM), also known as juvenile diabetes, occurs most frequently in children; however, T1DM can also progress in adults. In T1DM, the body becomes incapable of producing insulin or sufficient insulin due to an autoimmune response of the body that destroys the cells which produce insulin. Adult-onset diabetes, also known as type 2 diabetes mellitus (T2DM), can strike anyone at any age, including children. But it commonly occurs in older and middle-aged and overweight people. In T2DM, insulin resistance in fat, muscle, and liver cells occurs due to which cells are not able to use insulin to transmit glucose into the body's cells for consumption of energy. T1DM and T2DM are chronic diseases that have no treatment. Cardiovascular disease, chronic renal failure, and diabetic retinopathy are examples of long-term impairments. As a result, in the current situation, recurrent disease medication is essential, as is quitting smoking and maintaining healthy body weight. Furthermore, having diabetes raises the level of glucose in our blood. High blood glucose levels damage the small blood vessels in our kidneys, heart, neurological system, and eyes. Owing to these reasons, DM is responsible for heart problems, kidney disease, stroke, nerve damage, and increases the chances of sexual dysfunction. To avoid this, early detection of the DM is critical, as are effective techniques. Among all the reported cases of people with diabetes, only 5% have T1D whereas T2 diabetes accounts for 95% of cases and is commonly associated with physical inactiveness, obesity, older age, and family history of T2DM, or particular antiquity of gestational diabetes.

Hence, it is significant to implement various approaches that can be applied in predicting the future T2DM outbreak. In the research community, computational intelligence techniques have sparked a lot of interest.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

Machine learning approaches, according to numerous recent researches, have the ability to provide high classification accuracy when compared to other algorithms for data classification. Attaining prominent accuracy in prediction is crucial because it can lead to a suitable precaution programmer. Prediction accuracy may vary depending on different learning techniques and approaches. As a result, it's critical to find systems that can anticipate diabetes outbreaks with high accuracy. The accuracy of prediction achieved in the current project is compared to that of earlier research The goal of this research is to provide a framework of classification algorithms for DM diagnosis that is based on computational intelligence techniques. Computational intelligence techniques are the most effective decision-making approaches for the real world and scientific problems. The goal is also to see how well different computational intelligence algorithms work when it comes to classifying diabetic and non-diabetic samples. Moreover, performance of these techniques has been evaluated on different classification performance measurements. For this, four computational intelligence techniques were used namelysupport vector machine (SVM), Logistic Regression, Random Forest and XGBoost.

Moreover, the performance was compared using receiver operating characteristic (ROC) and calibration graph. Machine learning has been applied previously in the biomedical field, for association of heart disease and diabetes, in analysis of diabetes proteins, etc.

II. LITERATURE SURVEY

We have briefly discussed some of the existing research articles that are connected to our work in this part. For details on related research publications, see table 1.

Sr. No.	Title	Conclusion	Limitations
1.	Diabetes Prediction using Machine Learning Techniques [1]	The main objective of the paper was to design and implement various ways to Predict Diabetes Using Machine Learning and Analyze Performance of those methods. In the proposed system SVM, KNN, Random Forest, Decision Tree and Gradient Boosting are used.	The proposed system has achieved only 77% classification accuracy.
2.	A Review of Diabetic Prediction Using Machine Learning Techniques [2]	The study focuses on how different performance values of classification algorithms are calculated on various measures. Training and testing the datasets like Pima Indians Diabetes Dataset. This study collected various classification techniques for the purpose of improving accuracy, specificity and sensitivity.	In this work, the key objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For several classification problems, the higher number of samples chosen but it doesn't lead to higher classification accuracy.
3.	Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women [3]	Diabetes Detection is done using SVM Classifier. This can be further improved by using Hybrid approaches of multiple Classifiers as well as by incorporating Fuzzy Logic. Similarly, Diabetes in Women is predicted using a Decision Tree. Hence, both prediction and detection using the proposed approach will be effective.	The proposed system deals with small data sets and has not been tested with big data sets.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

4.	Machine Learning Based Unified Framework for Diabetes Prediction [4]	The study compares performance of the six-machine learning classification technique and uses 10-fold validation technique to evaluate their performance. It proposes a framework for diabetes prediction, monitoring and application (DPMA). Multiple machine learning classifiers results better as compared to a single machine learning classifier.	Most of the studies do not consider the F-score, precision, and recall. However, our study provides average prediction of classification model by considering the Fscore, recall and precision.
5.	Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran [5]	The paper focuses on performance based on conventional data mining classification techniques that calculates and compares a dataset of patients of type 2 diabetes in the city of Tabriz, Iran. SVM, ANN, Decision tree, Nearest Neighbors, and Bayesian network techniques were used. It was concluded that methods effectiveness depends on the complexity and nature of the dataset used.	Many of the cases in this study were identified as type 2 diabetics before the screening process, thus the dataset is an unbalanced dataset with many diabetic patients.
6.	Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes [6]	Mass diabetic screening was organized to detect undiagnosed and pre-diabetic, in which only those people who scored high on IWDRS, were pathologically tested for high blood sugar. It is a cost-effective approach where late diagnosis is a major problem and can be used in under developed countries.	Randomly selected Test Dataset not used to derive IWDRS was 61.5%
7.	Machine learning for medical diagnosis history, state of art and perspective [7]	The study compares the performances of 8 different medical datasets. The Decision tree algorithm has the most appropriate subset of attributes from the compared algorithm.	The paper does not provide a comprehensive overview but rather describes some subareas. It also tries to verify some unexplained phenomena from complementary medicine.
8.	A Decision Support System for Diabetes Chronic Care Models Based on General Practitioner Engagement and EHR Data Sharing [8]	The study presents and tests use-case scenarios in real time and provides a comprehensive framework for supporting the GPs during the diabetes early detection & enrollment stage.	The proposed DSS for the management does not manage different chronic diseases.
9.	An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques [9]	The objective of the study is to predict the presence of diabetes more accurately. The study used three classification algorithms namely Naive Bayes, Random Forest, and NB-Tree	It uses a smaller number of input attributes. To mine a huge amount of unstructured data available in the healthcare industry database Text and web mining can be used.
10	Analysis of computational intelligence techniques for	The use of computational intelligence techniques for predicting diabetes is described in the study. The study uses various clinical parameters based on different classification approaches in the identification of diabetes.	

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

	diabetes mellitus prediction [10]		
11	Prediction of Diabetes using Classification algorithm [11]	The study explains systematic efforts made in designing a system which helps in the prediction of diabetes. The study is evaluated on various measures using 3 machine learning algorithms namely Decision Tree, SVM and Naive Bayes.	The designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases

III. METHODOLOGY

The proposed model will use SVM, Logistic Regression, Random Forest and XGBoost to predict diabetes. Figure 1 shows how the assessment technique is executed in a sequence of phases.

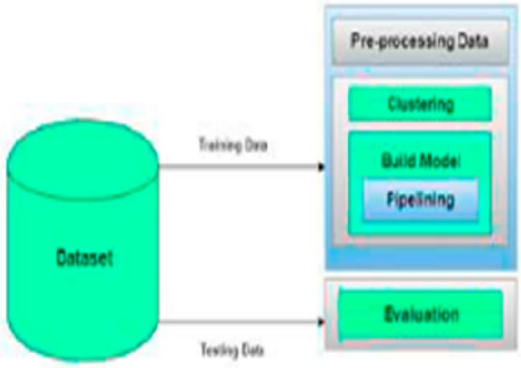

Figure 1: System Architecture

Refer Figure 2 for use case diagram.
The model has five different modules. These modules include:

3.1 Dataset Collection

This includes data collection and analysis in order to investigate patterns and trends, which aids in forecasting and evaluating outcomes.

- `Pregnancies` – Number of times pregnant
- `Glucose` – Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- `BloodPressure` – Diastolic blood pressure (mm Hg)
- `SkinThickness` – Triceps skinfold thickness (mm)
- `Insulin` – 2-Hour serum insulin (mu U/ml)
- `BMI` – Body mass index (weight in kg/(height in m)^2)
- `DiabetesPedigreeFunction` – Diabetes pedigree function
- `Age` – Age (years)
- `Outcome` – Class variable (0 or 1) 268 of 768 are 1, the others are 0

3.2 Data Pre-processing:

This phase of the model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. Because these attributes cannot have zero values, we imputed missing values for a few selected attributes such as glucose level, blood pressure, skin thickness, BMI, and age. The dataset is then scaled to normalize all values.

3.3 Clustering

On the dataset, clustering is utilized to categorize each patient as diabetes or non-diabetic. Before performing clustering, highly correlated attributes were found which were, glucose and age. Clustering will be performed on these two attributes. After implementation of this clustering will get class labels (0 or 1) for each of our records.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

3.4 Model Building

This is the most crucial phase, which includes the development of a diabetes prediction model. We used a variety of machine learning methods to predict diabetes in this study. These algorithms include Support Vector Classifier, Random Forest Classifier, Logistic Regression, and K-Nearest Neighbor.

3.5 Evaluation

This is the final step of the prediction model. We use numerous evaluation measures, such as classification accuracy, to evaluate the prediction findings.

A. Classification Accuracy

It is the ratio of the number of correct predictions to the total number of input samples. It is given as

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

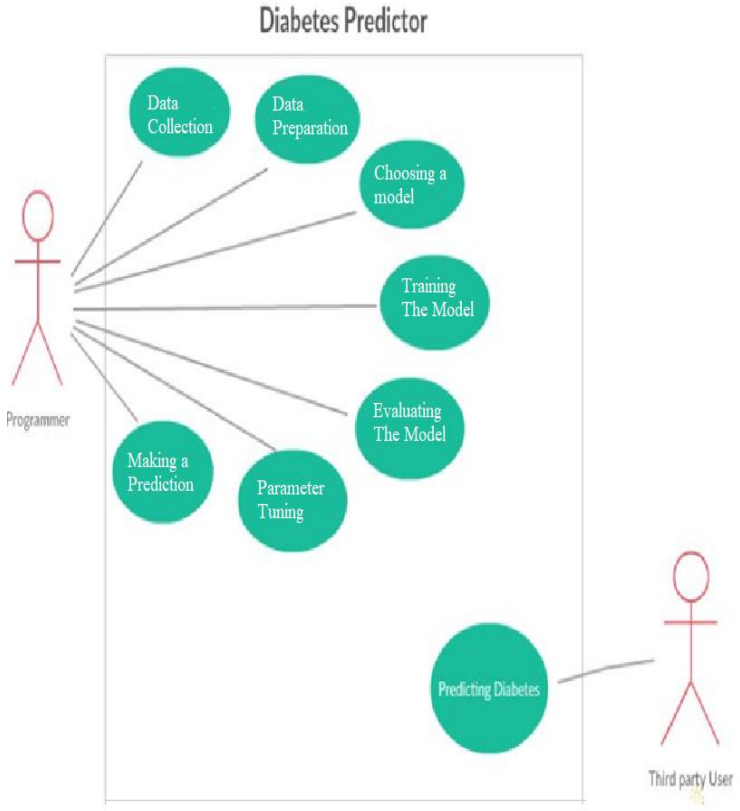

Figure 2: Use Case Diagram

The following is a general framework of steps in supervised machine learning:

1. Data Collection
2. Data Preparation
3. Choosing a model
4. Training the model
5. Evaluating the model
6. Parameter tuning
7. Making prediction

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

IV. ALGORITHMS

4.1 SVM

SVM stands for Support Vector Machine and is a supervised machine learning algorithm. SVM is the most widely used categorization method. A hyperplane is created by SVM to separate two classes. In high-dimensional space, it can generate a hyperplane or a series of hyperplanes. This hyperplane can also be utilized for regression or classification. SVM differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is accomplished via a hyperplane, which accomplishes separation to the nearest training point of any class.

A. Algorithm

1. Choose the hyper plane that best divides the class.
2. To identify the best hyper plane, you must calculate the Margin, which is the distance between the planes and the data.
3. When the distance between classes is small, the chances of miscarriage are great, and vice versa. So, we need to
4. Choose the class with the highest margin.

Margin = distance to positive point + Distance to negative point.

4.2 XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE and MPI) and can solve problems beyond billions of examples.

4.3 Logistic Regression

Logistic regression is also a supervised learning classification algorithm. It's used to figure out how likely a binary response is based on one or more predictors. They might be either continuous or discrete in nature. When we wish to categories or separate some data objects into categories, we apply logistic regression. It classifies data in binary form, that is, just in 0's and 1's, which is used to classify patients as diabetic positive or negative. The main goal of logistic regression is to find the best fit, which describes the connection between the target and predictor variables. Logistic regression is a model that is based on linear regression. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

4.4 Random Forest

It's an ensemble learning method that can be used for classification and regression.It has a higher level of accuracy when compared to other models. Large datasets are no problem with this strategy. Leo Bremen is the creator of Random Forest. It's a well-known ensemble learning method. By lowering variation, Random Forest improves Decision Tree performance. It works by training a large number of decision trees and then determining the mode of the classes, classification, or mean prediction (regression) of the individual trees.

A. Algorithm

- The first step is to select the “R” features from the total features “m” where R<<M.
- The node with the best split point among the "R" features.
- Using the best split, divide the node into sub nodes.
- Repeat steps a through c until you reach the “l” number of nodes.
- Created a forest by repeating steps a to d “a” number of times to produce “n” trees. Using the Gin-Index Cost Function, which is given by: the random forest determines the best split. The first step is to look at the options and use the foundations of each indiscriminately created decision tree to predict the outcome and store it at intervals around the target location. Second, count the votes for each anticipated target and, as a result of the random forest

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, May 2022
Impact Factor: 6.252

formula's ultimate prediction, admit the predicted target with the most votes. Random Forest has a number of options that produce accurate predictions for a variety of applications.

V. RESULT

After testing different training splits, it was observed that the best results were observed when 33% of the data was selected for testing. Random forest with 200 estimators leads to an accuracy of 99%. Since it is an unrealistically high number we conclude that this model must overfit the training data set and it is unwise to continue with it. Logistic regression yields an accuracy of 78% which is not enough for prediction. Similar is the case for SVM which yields a 77% accuracy. The best result was obtained by XG Boost algorithm with a score of 88%. This is the algorithm that was chosen for building an interactive client design.

VI. CONCLUSION

The project is driven by the necessity of predicting diabetes before it becomes malignant. Also, the proposed system uses the latest machine learning algorithms to build a fully functional system. The study takes into account important bodily factors and makes an accurate prediction as to whether the given patient's dataset is worrying or not. Because of the system it will be extremely beneficial to medical practitioners and the patients themselves to figure out whether they have a cause of worry or not. This can lead to a lot of saved lives and discomfort. As a result of the robust evaluation system, it can be determined that applying ML approaches yields satisfactory results. The accuracy of the evaluation can be improved by providing it with larger and more accurate training datasets. Finally, we can classify results using the XGBoost algorithm.

REFERENCES

[1]. AishwaryaMujumdara, Dr. Vaidehi V, "Diabetes Prediction using Machine Learning Techniques".
[2]. International Conference on Recent Trends in Advanced Computing 2019, Icrtac.
[3]. M.Rajeswari, Dr.P. Prabhu, "A Review of Diabetic Prediction Using Machine Learning Techniques". International Journal of Engineering and Techniques - Volume 5 Issue 4, July 2019
[4]. AakanshaRathore, Simran Chauhan, SakshiGujral, "Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women", Volume 8, No. 5, May-June 2017, ISSN No. 0976-5697, Available Online at www.ijarcs.info.
[5]. S M Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed HaiderNoori, Md Nazirul Islam Sarkar, "Machine Learning Based Unified Framework for DiabetesPrediction", BDET 2018, August 25–27, 2018, Chengdu, China. © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6582- 6/18/08
[6]. Mahmoud Heydari & Mehdi Teimouri & Zainabolhoda Heshmati & Seyed Mohammad Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran". © Research Society for Study of Diabetes in India 2015
[7]. Omprakash Chandrakar, Dr. Jatinderkumar R. Saini, "Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes"
[8]. Igor Kononenka, "Machine learning for medical diagnosis history, state of art and perspective"
[9]. Emanuele Frontoni, Luca Romeo, Michele Bernardini, Sara Moccia, Lucia Migliorelli, Marina Paolanti, Alessandro Ferri, Paolo Misericordia, Adriano Mancini, Primo Zingaretti, "A Decision Support System for Diabetes Chronic Care Models Based on General Practitioner Engagement and EHR Data Sharing"
[10]. B. Tamilvanan, Dr. V. Murali Bhaskaran, "An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 39-44, www.iosrjournals.org.
[11]. Ashok Kumar Dwivedi "Analysis of computational intelligence techniques for diabetes mellitus prediction." © The Natural Computing Applications Forum 2017
[12]. DeeptiSisodia, Dilip Singh Sisodia," Prediction of Diabetes Using Classification Algorithm",
[13]. www.elsevier.com/locate/procedia, Procedia computer science 132(2018) 1578-1585.