Diabetes Prediction Model for Better Clarification by using Machine Learning

Conference Paper · April 2023

DOI: 10.1109/ICICT57646.2023.10134235

6 authors, including:

Kaliappan s Seeniappan
KCG College of Technology
74 PUBLICATIONS **380** CITATIONS
SEE PROFILE

Shobha Aswal
Uttarakhand Technical University
9 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Khalid Ali Al-Salehi
Airlangga University
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Proceedings of the International Conference on Inventive Computation Technologies (ICICT 2023)
IEEE Xplore Part Number: CFP23F70-ART; ISBN: 979-8-3503-9849-6

Diabetes Prediction Model for Better Clarification by using Machine Learning

J.Lysa Eben[1]
Department of Computer Applications, School of Computing Sciences
Vels Institute of Science Technology and Advanced Studies
Chennai, Tamilnadu, India
lysaeben.scs@velsuniv.ac.in

R.Jayasudha[2]
Department of Mathematics
Dr.N.G.P.Institute of Technology
Coimbatore, India
rjayasudha98@gmail.com

S. Ramya[3]
Department of Computer Science and Engineering
Kongu Engineering College
Erode, Tamilnadu, India,
sramya.cse@kongu.edu

S.Kaliappan[4]
Division of Research and Development
Lovely Professional University,
Phagwara, Punjab (India) - 144411
kaliappan.197526@gmail.com

Shobha Aswal[5]
Assistant Professor, School of Computing,
Graphic Era Hill University
Dehradun.India
shobhaaswal@gehu.ac.in

Khalid Ali Salem Al-Salehi[6]
Biomedical engineering study program, faculty of science and technology, Airlangga University, Surabaya, Indonesia
Khalid-ali-21@fst.unair.ac.id

***Abstract*— Diabetes mellitus is one of the most pressing health concerns because so many people are afflicted by its disabling symptoms. Factors such as age, excess body fat, insufficient physical activity, a history of diabetes in one's family, a sedentary lifestyle, an unhealthy diet, hypertension, etc., all increase the likelihood of developing diabetes mellitus. Health complications are more common in people with diabetes, including cardiovascular disease, renal failure, stroke, blindness, and nerve injury. To validate a diagnosis of diabetes, hospitals typically perform a battery of procedures on the patient. Big data analytics has many vital applications in the healthcare sector. Numerous large computer systems are used in the healthcare sector. With the help of big data analytics, researchers can sift through mountains of data in search of previously unseen patterns and insights. Current techniques have a poor degree of precision in classification and forecast. While previous research has focused on factors such as glucose, body mass index, age, insulin, etc., the proposed model takes these into account and also the other factors that may be more relevant to the development of diabetes. The newer sample is superior to the older one based on categorization accuracy. A workflow algorithm for diabetes prognosis is also required to improve the accuracy.**

Keywords*— *diabetes prediction, Machine learning models, precision, health care

I. Introduction

High diabetes numbers persist, however, in nations as varied as Canada, China, India, and others. India has a populace of over 100 million people, and it is believed that 40 million of them have diabetes. There is a strong correlation between diabetes and increased death [1],[2].

Investigating diabetes prognosis using several variables related to the condition has the potential to save human lives through early diagnosis and treatment of illnesses like diabetes [18]. We use the Pima Indian Diabetes Dataset with several Machine Learning categorization and ensemble Methods to generate illness prognoses. Stated, Machine Learning is a method that can be used to educate computers and other devices officially. Many Machine Learning Methods help in knowledge collecting by using the collected data to build categorization and ensemble models [19]. The detection of diabetes could benefit from these files. Predictions can be made using various Machine Learning techniques, and deciding which to use can be difficult. Consequently, we employ commonly-used categorization and ensemble methods for projection on the dataset.

II. Proposed Methodology:

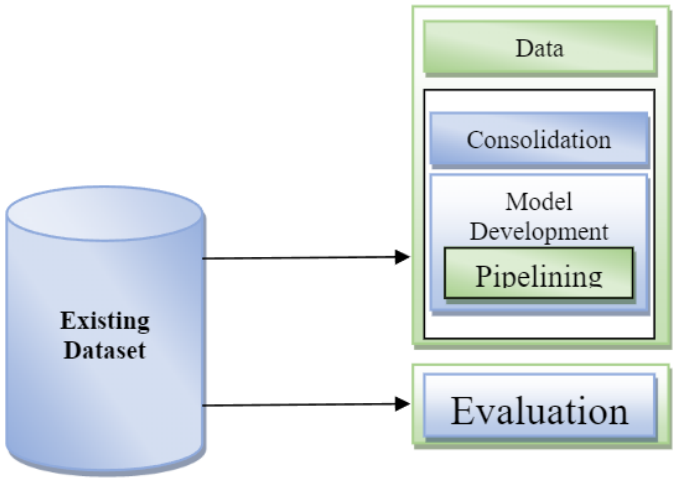

Fig. 1. Diabetes Prediction Model Diagram

It has different kinds of modules, and important five models include:

Existing Dataset

Data Customization

Consolidation

2023 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-9849-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICT57646.2023.10134235

Model Development Evaluation [3].

We shall take a look at each model briefly:

A. Existing Dataset

This section focuses on collecting and analyzing data to identify patterns and trends that may be used for forecasting and assessing outcomes. Here's a quick rundown of the data we're working with: [4].

There are 800 entries in this diabetes dataset, with ten different characteristics.

TABLE I. : DATA SET CHARACTERISTICS

Name of the dataset characteristic	Types of characteristics
Total number of carried women	N
Body glucose level	N
BP	N
Age of the patient	N
Insulin	N
Thickness of the skin	N
Type of the job wether it is office work/ field work/ any other)	NO
Body mass index value	N
Sex(Male/ Female/other)	N
Outcome	C

B. Data Customization

In this stage of the model, we deal with erroneous information to get more reliable outcomes. As you can see, there are blanks in this data collection. Because features like glucose level, blood pressure, skin thickness, body mass index, and age cannot be zero, we imputed missing values for these characteristics. We apply a scaling factor to the dataset to make all values comparable [5].

C. Consolidation

To classify each patient as either diabetic or non-diabetic, we used K-means clustering on the dataset. Found When Glucose and Age were strongly correlated, K-means clustering took place. [6-8] The K-means clustering method used these two factors as the clustering determinants. Because we used this clustering tool, each file has a category name (0 or 1) that tells what it is.

Algorithm:

Step 1: Select K clusters and collect data points.

Step 2: Randomly place centroids .

Step 3: Repeat Steps 4 and 5 for a set number of iterations.

Step 4: At every datapoint

-Select the closest point to the centroid

-assign correct point of the particular cluster

Step 5: $(\text{centroid})_n$ = Average of all centroid points assigned to the cluster.

Step 6: Stop

D. Model Development

Model development is the most crucial stage since it involves developing a model to foretell diabetes. We include several machine-learning techniques for predicting diabetes. [9-11].

```
Algorithm 1: Machine learning algorithms predict diabetes:

Randomize training and test sets
Specify Model Mn employs the following algorithms:
 KNN(); DTC(); GaussianNB();
LDA(); SVC(); LinearSVC(); AdaBoost();
RandomForestClassifier(); Perceptron();
ExtraTreeClassifier(); Bagging();
LogisticRegression(); GradientBoostClassifier();
End.
```

Fig. 2. machine learning algorithm to predict the diabetes

E. Evaluation:

We reached the last stage of the prediction model. Here, we use

Assessment measures

Classification accuracy

Confusion matrix,

F1-score to assess the efficacy of the predictions [12],[13]

1) Classification Accuracy:

It indicates how many input samples led to correct predictions.

The format is as follows-

$$Accuracy\ of\ the\ predection = (Correct\ predections)/(total\ predections)$$

2) Confusion Matrix:

In these models, the overall performance is being summarized in a matrix that is the final output.

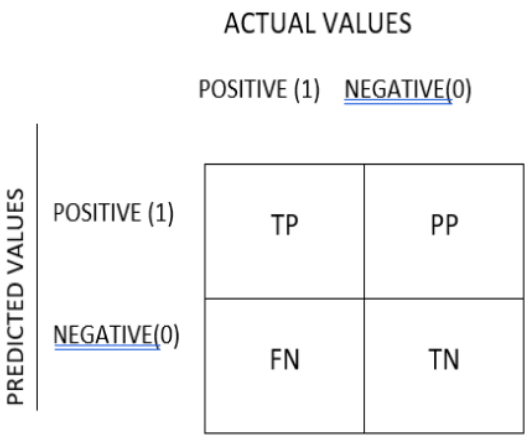

Fig. 3. Confusion matrix

Where,

TP:True/ Positive value of the diagnosis

FP: False/Positivevalue of the diagnosis

FN: False/Negativevalue of the diagnosis

TN:True/Negativevalue of the diagnosis

The accuracy of the matrix may be determined by averaging its values along the principal diagonal.

Predict the value of F1 score:

It's how you find out how reliable a test is. The F1 Score balances accuracy with recall using a harmonic mean [14,15]. F1 Score is between 0 and 1, which shows how accurate and stable your system is at classifying. Specifically, its mathematical expression is-

$$F1\ score\ value = (1/(1 /(precesion\ value\ of\ the\ predection) + 1/(recall\ value)))$$

F1 Score searches for a middle ground between accuracy and recall.

Precision Value:

It is the proportion of true positives to total true positives predicted by the classifier.

$$Precision = TP/((TP + FP))$$

Recall:

Correct positive findings as a percentage of total relevant samples The formula is as follows:

$$Precision = \mathbf{TP}/((\mathbf{TP} + \mathbf{FN}))$$

III. RESULTS

We obtained the following results after applying many machine-learning algorithms to the dataset. The best reliability, 97%, is achieved using logistic regression.

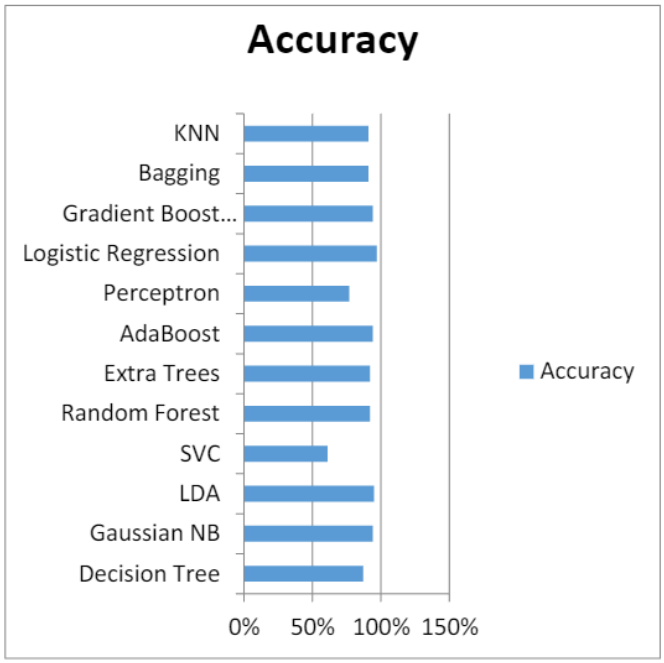

Fig. 4. Accuracy Chart

For Logistic Regression, Here Is the Confusion Matrix

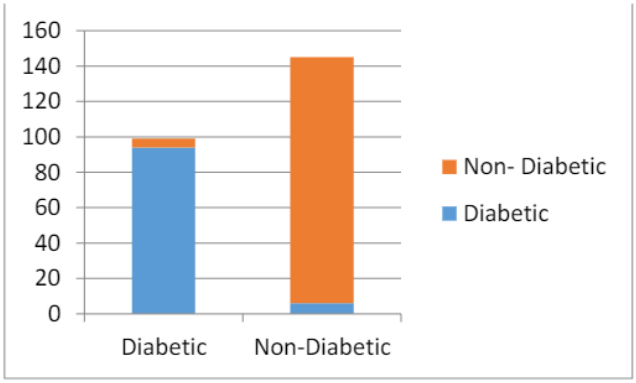

Fig. 5. An Error Matrix for Logistic Regression

Compare Accuracy, F1-Score, Precision, and Recall are performance metrics [16],[17]. Fig 3 lists the confusion matrix values for the most precise method. Seeing the range of values enables us to comprehend the differences between them better.

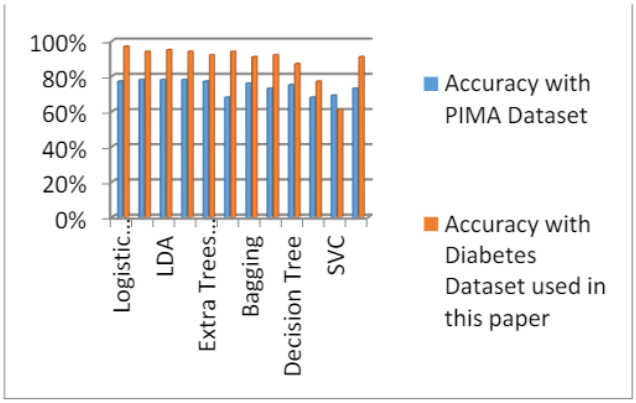

Fig. 6. Accuracy comparisons across a range of machine learning techniques

With Pipelining, we improved Logistic Regression accuracy to 97.2%.

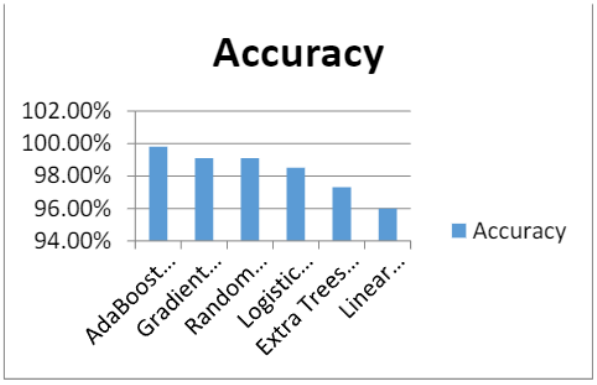

Fig. 7. Predicting Outcomes

IV. CONCLUSION:

In this research, we use many machine learning algorithms to analyze and classify the data, with the most incredible accuracy coming from logistic regression at 98.50%. After going through the pipeline, the AdaBoost classifier was the best model, with 99.80% accuracy. We have seen machine learning algorithms' accuracy in comparing two datasets. This dataset makes the model more accurate and precise at predicting diabetes than other datasets. In the coming years, this study could expand to determine how common diabetes is among the general population.

REFERENCES

[1] Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019, Pages 292-299, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.

[2] Lai, H., Huang, H., Keshavjee, K. *et al.* Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord* **19**, 101 (2019). https://doi.org/10.1186/s12902-019-0436-6

[3] Fregoso-Aparicio, L., Noguez, J., Montesinos, L. *et al.* Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr* **13**, 148 (2021). https://doi.org/10.1186/s13098-021-00767-9

[4] Shivlal Mewada, Anil Saroliya, N. Chandramouli, T. Rajasanthosh Kumar, M. Lakshmi, S. Suma Christal Mary, Mani Jayakumar, "Smart Diagnostic Expert System for Defect in Forging Process by Using Machine Learning Process", *Journal of Nanomaterials*, vol. 2022, Article ID 2567194, 8 pages, 2022. https://doi.org/10.1155/2022/2567194

[5] Aggarwal, A., Kumar, S., Bhatt, A. and Shah, M.A., 2022. Solving User Priority in Cloud Computing Using Enhanced Optimization Algorithm in Workflow Scheduling. *Computational Intelligence and Neuroscience*, *2022*.

[6] Choudhury, A., Aggarwal, A., Rangra, K. and Bhatt, A., 2019. The Components of Big Data and Knowledge Management Will Change Radically How People Collaborate and Develop Complex Research. In *Big Data Governance and Perspectives in Knowledge Management* (pp. 241-257). IGI Global.

[7] Salliah Shafi Bhat, Venkatesan Selvam, Gufran Ahmad Ansari, Mohd Dilshad Ansari, Md Habibur Rahman, "Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2789760, 12 pages, 2022. https://doi.org/10.1155/2022/2789760

[8] Kaur, H. and Kumari, V. (2022), "Predictive modelling and analytics for diabetes using a machine learning approach", *Applied Computing and Informatics*, Vol. 18 No. 1/2, pp. 90-100. https://doi.org/10.1016/j.aci.2018.12.004

[9] Srivastava, Deepak, et al. "Analysis of Protein Structure for Drug Repurposing Using Computational Intelligence and ML Algorithm." International Journal of Software Science and Computational Intelligence (IJSSCI) 14.1 (2022): 1-11.

[10] Birjais, R., Mourya, A.K., Chauhan, R. *et al.* Prediction and diagnosis of future diabetes risk: a machine learning approach. *SN Appl. Sci.* **1**, 1112 (2019). https://doi.org/10.1007/s42452-019-1117-9

[11] Srivastava, D., Soni, D., Sharma, V., Kumar, P., & Singh, A. K. (2022). An Artificial Intelligence Based Recommender System to analyze Drug Target Indication for Drug Repurposing using Linear Machine Learning Algorithm. Journal of Algebraic Statistics, 13(3), 790-797.

[12] Wahlang, I.; Maji, A.K.; Saha, G.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Jasinska, E. Brain Magnetic Resonance Imaging Classification Using Deep Learning Architectures with Gender and Age. *Sensors* **2022**, *22*, 1766. https://doi.org/10.3390/s22051766

[13] Soni, D., Srivastava, D., Bhatt, A., Aggarwal, A., Kumar, S., & Shah, M. A. (2022). An Empirical Client Cloud Environment to Secure Data Communication with Alert Protocol. Mathematical Problems in Engineering.

[14] R. K. Singh, P. Singh and G. Bathla, "User-review oriented social recommender system for event planning," Ingénierie des Systèmes d'Information, vol. 25, no. 5, pp. 669-675, 2020.

[15] G. Bathla, P. Singh, S. Sharma, M. Verma, D. Garg and K. Kotecha, "Recop: Fine-grained Opinions and Collaborative Filtering based Recommender System for Industry 5.0," Soft Computing, p. Preprint, October 2021.

[16] A. Aggarwal, S. Gaba and P. Singh, "Character Recognition using Approaches of Artificial Neural Network: A Review," CEUR Workshop Proceedings, vol. 3309, no. 1, pp. 186-193, 2022.

[17] G. Bathla, P. Singh, R. K. Singh, E. Cambria and R. Tiwari, "Intelligent Fake reviews detection based on Aspect Extraction and Analysis using Deep Learning," Neural Computing and Applications, July 2022.

[18] L. Mohan, J. Pant, P. Suyal, and A. Kumar, "Support Vector Machine Accuracy Improvement with Classification," in Proceedings - 2020 12th International Conference on Computational Intelligence and Communication Networks, CICN 2020, 2020, pp. 477–481, doi: 10.1109/CICN49253.2020.9242572.

[19] I. Kumar, C. Bhatt, V. Vimal, and S. Qamar, "Automated white corpuscles nucleus segmentation using deep neural network from microscopic blood smear," J. Intell. Fuzzy Syst., vol. 42, no. 2, pp. 1075–1088, 2022, doi: 10.3233/JIFS-189773.