2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579799

Early Predictive Analytics in Healthcare for Diabetes Prediction Using Machine Learning Approach

Md. Mehedi Hassan
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
mh_ashiq@yahoo.com

Md. Al Mamun Billah
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
almamunnwu@gmail.com

Md. Mushfiqur Rahman
Department of Statistics
University of Dhaka
Dhaka, Bangladesh
mdmushfiqur-2017014005@stat.du.ac.bd

Sadika Zaman
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
sadikazaman@yahoo.com

Md. Mehadi Hasan Shakil
Department of Computer Science and Engineering
Northern University Of Business & Technology Khulna, Bangladesh
mhshakil.cse@gmail.com

Jarif Huda Angon
Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
angonzarif@gmail.com

***Abstract*—Diabetes is a metabolic disorder in the world today. The rate of production of diabetic patients is rising day by day. Diabetic disease occurs when the blood glucose level gets high, leading inevitably to other health conditions such as heart disease, kidney disease, etc. Symptoms of diabetes are increased appetite and urination, increased hunger, fatigue, blurred vision, sores that do not heal, unexplained weight loss. People with diabetes are at high risk for diseases such as eye problems, nerve damage, etc. In this paper, we proposed a diabetes prediction model for better diabetes classification that includes a model of a few external diabetes factors along with normal factors such as glucose, age, gender, Blood Pressure, Sugar, Red Blood Cells, Hemoglobin, Blood Urea, etc. We have a dataset that contains 250 variants that individually hold 16 unique attributes. We have used Logistic Regression, Support Vector Machine, and Random Trees for this prediction.10-fold cross-validation had applied for training the data and the accuracy for Logistic Regression is 94.5 %, Support Vector Machine is 96.5% and Random Tree is 97.5%.**

Keywords- ***Data Mining, Diabetes Prediction, Data Analysis, Predictive Analysis, Diagnostics***

I. Introduction

Diabetes Mellitus (DM) is the fourth highest disorder mortality rate in the world [1]. Around 422 million people worldwide have diabetes, according to the World Health Organization (WHO), the majority of whom live in low- and middle-income countries, and 1.6 million deaths are directly related to diabetes per year [2]. About 75% of patients with diabetes have blood pressure levels (BP) of around 130/80 mm Hg or use antihypertensive drugs [3]. Diabetes Mellitus is persistent, caused by the excessive level of sugar in the circulatory system, an endless disease. It is induced by the inappropriate functioning of beta-pancreatic cells. It influences various parts of the body that includes pancreatic glitch, heart disease risk, hypertension, kidney disappointments, pancreatic disorders, nerve damage, foot problems, ketoacidosis, visual disruption, and other eye problems, waterfalls, and glaucoma, etc. Diabetes Mellitus is categorized as Type-1, Type-2, and Type-3. Type-1 diabetes once referred to as juvenile diabetes or insulin-dependent diabetes, is a chronic disease in which little to no insulin is produced by the pancreas. Insulin is a hormone that is required to allow sugar (glucose) to reach energy-producing cells. It can occur in adults, while Type-1 diabetes typically occurs during childhood or adolescence. Type-1 diabetes has no cure, despite successful studies. To avoid complications, treatment focuses on controlling blood sugar levels with insulin, diet, and lifestyle. Type-2, diabetes is a type of diabetes induced by high blood sugar, insulin resistance, and relative insulin absence previously referred to as adult-onset diabetes. Increased appetite, frequent urination, and unexplained weight loss are common symptoms [4]. Type-3 diabetes is a name coined for Alzheimer's disease that results from brain resistance to insulin. It is not a medical term or a known disorder yet, but it is a term now used in studies exploring the causes of Alzheimer's disease. Also, drug abuse behaviour is so much harmful to diabetes diseases [5]. We proposed a diabetes prediction model for better diabetes classification in this paper. We assure Systolic Blood Pressure, FBS, PPBS, Urine cave collected these features such as Age, BMI, Duration of Diabetes, Diastolic of FBS, Urine color of PPBS, Type of medicine, class, gender, and so on. These types of features are used for our prediction to get the best

performance. The analysis also generalizes the collection of optimal data set features to enhance the precision of the classification. Our model will help us to predict diabetes diseases so early by medical data. In the future, we want to operate with more info. We also planned to improve the dataset analysis more feasible, accurate, and effective.

II. Related Literature

In this paper [6], is aimed at making use of important features, developing a prediction algorithm using machine learning, and finding the best classifier compared to clinical outcomes to provide the closest result and proposed approach aims to use Predictive Analysis to pick the characteristics that are responsible for the initial detection of Diabetes Mellitus. They utilized the Decision Tree (DT), Random Forest, Naïve Bayes Classifier (NBC) and% accuracy by using Decision Tree (DT), 98.00% accuracy by using the Random Forest Algorithm, and 82.30% accuracy by using Naïve Bayes Classifier (NBC). In this paper [7], some features that are liable for diabetes have been used, such as BMI, glucose, age, insulin, and so on. They have used Decision Tree, Gaussian NB, LDA, SVC, Random Forest, Extra Trees, AdaBoost, Logistic Regression, for classification. And got AdaBoost classifier as the best model of 98.8 percent accuracy. This paper [8], has built a deep neural network for the prediction of diabetes, and this analysis used a deep neural network technique for diabetes recognition. Based on several medical variables, diabetes. In addition, for ten-fold cross-validation, accuracy is 97.11 percent, sensitivity is 96.25 percent, and specificity is 98.80 percent. In this paper [9], is aimed at developing a model that can predict the risk of diabetes in patients with optimum accuracy. In this research, three classification algorithms for machine learning, namely Decision Tree, SVM, and Naive Bayes, are therefore used to identify diabetes at a beginning period. Various metrics such as Precision, Accuracy, F-Measure, and Recall test the efficiency of all three algorithms. Accuracy is calculated against instances classified correctly and incorrectly. The results indicated that Naive Bayes exceeds other algorithms with the highest accuracy of 76.30 percent. In this paper [10], they used artificial neural networks to determine whether an individual is diabetic or not. Using a neural network model, the criterion has been to minimize the error function in neural network training. The average error function of the neural network is equivalent to 0.01 after training the ANN model and the accuracy of the prediction of whether an individual is diabetic being 87.3%. This prediction model will help for taking proper medicine and it will give a

III. Data Mining Framework

Figure - 1 has presented the proposed data mining system for diabetes analytics. To distinguish outliers and missing data, the raw data is first extracted and coded and then preprocessed. We used the following criteria to deal with the missing data: firstly, we have deleted it if the variable has more than 15 % missing data. Secondly, we hold it and use missing data imputation techniques to impute the missing data if the

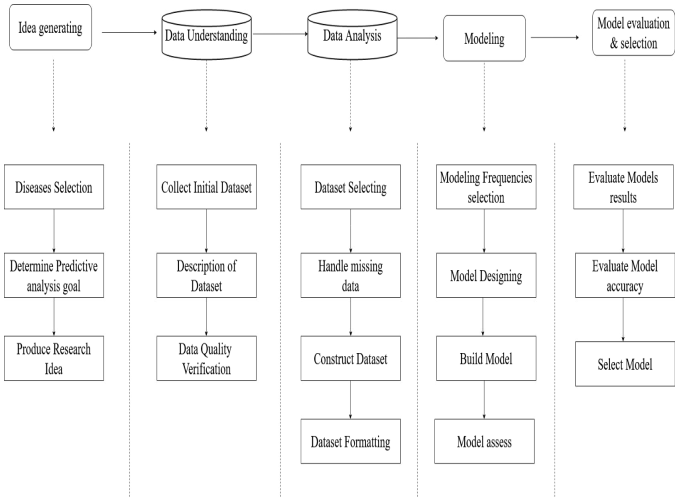

Fig. 1. Propose a framework for Diabetes analysis and prediction.

variable has 15 percent or less missing data [11]. Five missing methods of data imputation are assumed: fixed using mean, fixed using mid-range, random uniform, random normal, and Classification, and Regression Trees (C&RT) algorithm.
In comparison to the original data, the imputation models are done based on the variability in the imputed data [12]. To classify the features to be used in the prediction models, feature selection is then used. Here are the names of the prediction models used are: Logistic Regression, Random Trees, and Support Vector Machines. Based on precision, sensitivity, and specificity, the models were analyzed. This data was collected from the University of California-Irvine repository for this analysis.

IV. Methodology

The following steps have been followed to do this analysis.

A. Data collection
B. Data preprocessing
C. Predictive Analytics Models for Diabetes.

A. Data collection

The data has been collected from Shaheed Sheikh Abu Naser Specialised Hospital, Khulna. The dataset contains 289 participants and 14 variables regarding diabetes have been collected from them, that contain Type-2 diabetics. Table- 1 shows the respondent's current type of diabetes. The following table shows the description of the dataset.

We have visualized our dataset to 3D plots which are shown in fig-2.

TABLE I
METRICS OF THE PERFORMANCE TO TEST MODELS FOR ANALYTICS

Attribute	Type	Description, Units, and Values	Min	Max	Mean	Std. Dev	Unique	Valid
Age	Continuous	Age in years	18	85	49.92	11.897		289
Height	Continuous	Height in cm	138	180	155.89	7.228		289
Weight	Continuous	Weight in kg	32	100	62.36	10.611		289
BMI	Continuous	Body Mass Index (BMI)	14.4	42.2	25.615	4.073		289
Duration of Diabetes	Continuous	Diabetes Duration in Years	0	20	1.746	3.208		289
Diastolic Blood Pressure	Continuous	Diastolic Blood Pressure in mmHg	90	200	127.11	13.815		289
Systolic Blood Pressure	Continuous	Systolic Blood Pressure in mmHg	60	130	81.71	8.284		289
FBS	Continuous	Fasting Blood Sugar in mg/dL	4.3	24	10.076	4.32		289
PPBS	Continuous	Post Prandial Blood Sugar in mg/dL	5	28.5	14.71	5.889		289
Urine color of FBS	Nominal	Urine Color of FBS-(Blue, Green, Green Yellow, Lemon Green, Orange, Red, Yellow)					7	252
Urine color of PPBS	Nominal	Urine color of PPBS-(Blue,Brick Red, Green, Green Yellow, Lemon Green, Orange,Red, Yellow)					8	252
Type of medicine	Nominal	Type of Medicine - (Insulin, Tablet)					2	288
Class	Nominal	Class-(1,2)					2	289
Gender	Nominal	Gender - (Male, Female)					2	289

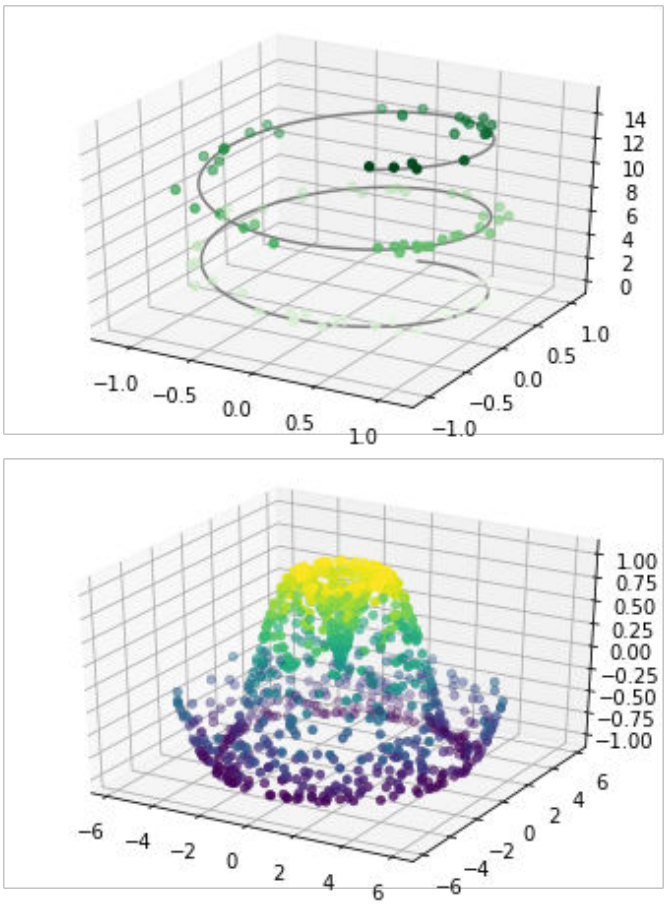

Fig. 2. 3D plots of our dataset.

B. Data preprocessing

We have used 3 types of methods have been used to impute the missing data (1) Fixed Median (2) Random Normal (3) Random Uniform. We have prepared all datasets for training and building models.

A. Tools and techniques:

Logistic Regression, Random Trees, and Support Vector Machines these 3 methods have been used to predict diabetes types. The confusion matrix has been built for each method for training as well as testing sets and the performance matrices: Accuracy, specificity, precision, recall, F1-score have been used to evaluate/compare the models.

True Positive(TP): The number of predictions where the classifier correctly predicts the classes as positive.

True Negative(TN): The number of predictions where the classifier correctly predicts the classes as negative.

False Positive(FP): The number of predictions where the classifier incorrectly predicts the negatives as positive.

False Negative(FN): The number of predictions where the classifier incorrectly predicts the positive as negative.

Accuracy: The fraction of the total samples that have been correctly classified.

$$Accuracy = \frac{TP \ + \ TN}{TP \ + \ TN \ + \ FP \ + FP} \tag{1}$$

Precision: The fraction of predictions as positives that have been actually positive.

$$Precision = \frac{TP}{TP + \ FN} \tag{2}$$

Recall: The fraction of all positive samples that have been correctly predicted.

$$Recall = \frac{TN}{TN \ + \ FP} \tag{3}$$

Specificity: The fraction of all negative samples that have been correctly predicted as negatives.

$$Specificity = \frac{TP}{TP + \ FP} \tag{4}$$

F1-score: Combines the precision and recall into a single measure (Harmonic mean of precision and recall).

$$F1 - score = \frac{2TP}{2TP + \ FP \ + \ FN} \tag{5}$$

The dataset has been randomly partitioned by 80% (training set) and 20%(testing set). From the total 36 types I diabetes patients 31 have been allocated in the training set and 5 in the testing set [13].

From the total 70 and 183 of type II and type III diabetes patients consecutively, 51 of type II and 149 of type III have been selected for training and the rest for testing.

C. Predictive Analytics Models for Diabetes

To predict diabetes mellitus three different methods were used: Logistic Regression, Support Vector Machine (SVM), Random Trees. Those algorithms are used to find predictive results [14] To evaluate models three different performance metrics are used, they are accuracy, sensitivity, and specificity. We are following the process fig-3 to complete this study.

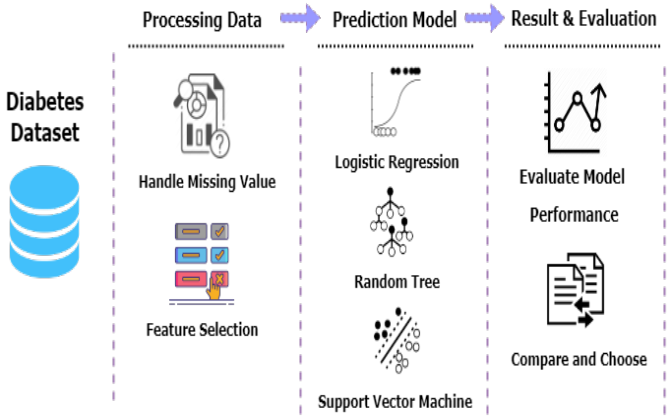

Fig. 3. System Architecture of this study.

Those three matrix details and descriptions are shown in table-3. When someone has diabetes the classification result will be positive for negative classification result he has not diabetes.

TABLE II
METRICS OF THE PERFORMANCE TO TEST MODELS FOR ANALYTICS

Metric	Description	Equation
	Tests the model's ability to predict the mark of latest or unknown data properly.	(TP+TN)/(TP+TN+FN+FP)
Sensitivity	Tests the proportion of positive ones (or Yes's) accurately defined as such.	TP/(TP+FP)
Specificity	The proportion of negatives (or No's) that are properly identified as such is calculated.	TP/(TP+FN)
Precision	The fraction of predictions as positives that have been positive.	TN/(TN+FP)
F1-score	Combines the precision and recall into a single measure (Harmonic mean of precision and sensitivity).	2TP/(2TP+FP+FN)

TABLE III
ABBREVIATION OF THE PERFORMANCE TO TEST MODELS FOR ANALYTICS

Abbreviation	Name	Description
TP	True Positives	The number of right categories predicted to be positive (or Yes)
TN	True Negatives	The number of right categories predicted to be positive (or No)
FP	False Positive	The number of cases that are incorrectly predicted to be positive if it is negative.
FN	False Negative	The number of cases that are incorrectly predicted to be negative if it is positive.

We have applied 3 machine learning algorithms, which are Logistic Regression, Support Vector Machine and Random Trees for training our dataset. Overall we have used 80% of the total dataset for building the model. We have gotten the best accuracy rate for the Random Forest algorithm which is 97.5%. Also for Logistic Regression and Support Vector Machine algorithm, the accuracy rate is 94.5% and 96.5%.

TABLE IV
PERFORMANCE MEASURES FOR THE TRAIN DATA

Method	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Logistic Regression	44	145	7	4	94.5	91.67	95.39
Support Vector Machine	44	149	0	7	96.5	86.27	100
Random Trees	46	149	0	5	97.5	90.20	100

For testing our dataset we have used 20% of our dataset. In random forest we have gotten 100% accuracy rate for Random Forest. In table-6, we have described the performance of 3 different algorithms.

TABLE V
PERFORMANCE MEASURES FOR THE TEST DATA.

Method	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
Logistic Regression	15	31	4	0	92	100	88.57
Support Vector Machine	18	29	2	1	94	94.74	93.55
Random Trees	19	31	0	0	100	100	100

In Fig-4, we have visualized our predicted result.

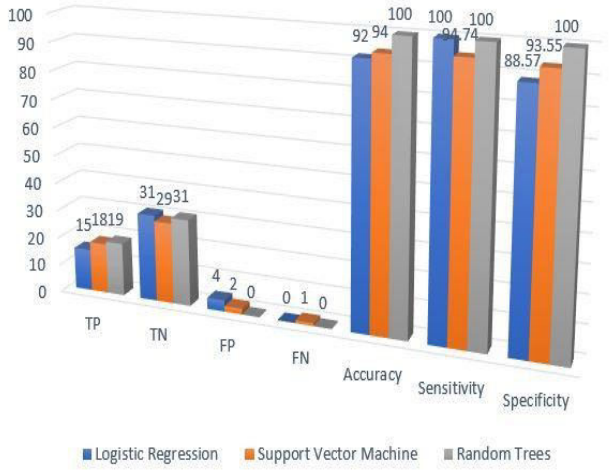

Fig. 4. Performance of different algorithms

V. CONCLUSION

The application of data mining and analytics techniques to predict diabetes has been presented in this paper. It is a matter of sorrow that the rate of diabetes is growing day by day. Here, several missing data imputation techniques are being used because the data had missing values. There are already three ML algorithms that we used. By applying the algorithms, we have got accuracy for Logistic Regression is 94.5 %, Support Vector Machine is 96.5% and Random Tree is 97.5% and the result is medical approved. Various limitations have found during the research. The dataset used in this analysis would have a relatively small number of instances.The main limitation of this study, we have analysed on limited number of data also this prediction model is only for a certain type of diabetics. We have faced many challenges. We also planned to improve the dataset analysis more feasible, accurate, and

effective. We have also planned in the future we will use deep learning for this research purpose. For instance, Random Forest (RF), and Naive Bayes Classifier (NBC) we would like to use the most useful methods to preprocess the dataset.

REFERENCES

[1] F. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques", 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 2018. Available: 10.1109/icoei.2018.8553959.

[2] "Diabetes", Who.int, 2021. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1.

[3] G. Bakris and J. Sowers, "ASH Position Paper: Treatment of Hypertension in Patients With Diabetes-An Update", The Journal of Clinical Hypertension, vol. 10, no. 9, pp. 707-713, 2008. Available: 10.1111/j.1751-7176.2008.00012.x.

[4] S. Watts, "What are the Symptoms of Type 1 Diabetes?", EndocrineWeb, 2021. [Online]. Available: https://www.endocrineweb.com/conditions/type-1-diabetes/type-1-diabetes-symptoms.

[5] M. Hassan, Z. Peya, S. Zaman, J. Angon, A. Keya and A. Dulla, "A Machine Learning Approach to Identify the Correlation and Association among the Students' Drug Addict Behavior", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020. Available: 10.1109/icccnt49239.2020.9225355.

[6] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", Journal of Big Data, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0175-6.

[7] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science, vol. 165, pp. 292-299, 2019. Available: 10.1016/j.procs.2020.01.047.

[8] S. Islam Ayon and M. Milon Islam, "Diabetes Prediction: A Deep Learning Approach", International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21-27, 2019. Available: 10.5815/ijieeb.2019.02.03.

[9] D. Sisodia and D. Sisodia, "Prediction of Diabetes using Classification Algorithms", Procedia Computer Science, vol. 132, pp. 1578-1585, 2018. Available: 10.1016/j.procs.2018.05.122.

[10] K. Tripathi, "Diabetes Classification and Prediction Using Artificial Neural Network", International Journal of Computer Engineering & Technology, vol. 10, no. 3, 2019. Available: 10.34218/ijcet.10.3.2019.018.

[11] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets", 2016 International Conference on Data Science and Engineering (ICDSE), 2016. Available: 10.1109/icdse.2016.7823957.

[12] F. Javed Mehedi Shamrat, P. Ghosh, M. Sadek, M. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease", 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020. Available: 10.1109/inocon50539.2020.9298026.

[13] B. Pranto, S. Mehnaz, E. Mahid, I. Sadman, A. Rahman and S. Momen, "Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh", Information, vol. 11, no. 8, p. 374, 2020. Available: 10.3390/info11080374.

[14] [1]P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques", IEEE Access, vol. 9, pp. 19304-19326, 2021. Available: 10.1109/access.2021.3053759