Machine Learning With Factor Scoring To Predict Diabetes Risk Level In Bangladesh

Nusrat Jahan, Aminul Islam, Abdullah Al Mamun

Abstract: Diabetes is a familiar disease in our country. When blood glucose or blood sugar levels are crossed a standard level is called diabetes. Meanwhile, glucose is the main source of energy in our body. Glucose produces from different foods that we take every day in our life. On the other side, having much glucose in our blood is a serious difficulty. High level of glucose can create threat of heart disease, stroke, kidney disease, vision problems, as well as nerve complications. In this paper, our goal was to identify a person, who will be in risk with diabetic. Here, we focused only type-2 diabetes risk prediction. Our model worked with nine attributes of 555 instances. We set score against each attribute through data analysis, after that trained our machine with algorithm to get the diabetes risk in 3 levels- low, medium, and high. Here, we used Weka 3.8.0 tools for analyzing data with three different machine learning algorithms- Multilayer perception, Decision tree and IBK. We got best result for 12 fold cross validation after applying IBK that has 98.73% classification accuracy. Finally, we proposed an Android application to predict the diabetes risk level easily for making people aware about diabetes.

Index Terms: Glucose, Diabetes, Machine learning, Factor Scoring, Weka.

1. INTRODUCTION

IN this era diabetes is one of the measure diseases that have no cure but if we take necessary steps as soon as possible then it would be in a stable stage. Worldwide, over 246 million people affected with diabetes, and majority percentages of them are women. Based on the WHO report, it will be over 380 million by 2025. In 2017, a predictable 8.8 percent of the grownup population worldwide had diabetes and by the year of 2045 it increase to 9.9 percent [1]. There are three types of diabetes all over the world namely, Type 1, Type 2, and Gestational diabetes. Type 1 diabetes is insulin dependent, or child onset, Type 2 diabetes is generally called non-insulin dependent diabetes, or adult onset and it is resulted from the ineffective use of insulin and Gestational diabetes occurring during pregnancy period [2]. Type 2 diabetes is referred to as “non-insulin dependent diabetes” or “adult onset diabetes” and finds almost 90–95% of all diabetes [3]. In the last era, the frequency and prevalence of type 2 diabetes has increased histrionically, mainly in racial and ethnic minority populations [4]. Diabetes patients should have awareness and need to monitor their blood glucose levels for regulates insulin levels. Always need to have motivation to keep blood glucose levels as near to normal level as possible. Blood glucose levels which is differ from the normal level can create serious short-term and long-term complications. Today people are busy with their daily life; they have not enough time to monitor their health condition regularly for keep good health but the patients with different diseases increases day by day in remarkable levels. Diabetes is one of the major diseases. The disease has been named the fifth deadliest disease in the United States with no imminent cure [5]. Information technology increases dramatically and its pointed initiation into the medical service also into healthcare sector, that is why the cases of diabetes as well as its symptoms are well documented for everyone in the world. This unexpected rises of diabetes we can minimize if we take proper care about our health regularly. This process will be easy if people receive a smart way in their surrounded to know the risk level. In this research, we address this problem to solve. We proposed a quicker and more efficient technique of monitor this disease, leading to timely monitor of the patient’s condition based on some daily questions. For this reason, all over the world researchers worked with smart diagnosis system to minimize diseases that is rises. In this paper, we focused a smart diagnosis system for diabetes prediction. Our goal was to motivate people about their daily health monitoring process through a smart application and it focused the recent phenomena “diabetes”. We organized this paper as follows. At first, we reviewed some related and motivational work that is presented in Section 2, after that demonstrated our methodology and result in Section 3, in the next step we discuss about implementation part which is in section 4, and at last summary with future work in section 5.

2 LITERATURE REVIEW

In this section we are going to discuss some related works that was motivated us to work with diabetes risk prediction based on local data. We also noted some previous works where machine learning and data analysis applied to predict the diabetes risk level or used different machine learning approaches for classifying diabetes dataset by using different machine learning tools. Sarwar N. et al. worked on diabetes, blood glucose and vascular disease. Where, they considered age, sex, smoking, systolic blood pressure, and body-mass index for calculating different parameters to present all those three diseases. They found adults with diabetes have two times more risk of heart attacks and strokes than others [6]. Md. Aminul Islam and Nusrat Jahan in 2017 worked with Pima Indian Dataset for predicting diabetes risk level. They discussed different machine learning algorithms to classify the dataset and compared those outcomes. Their observation was to predict diabetes in early stage that shows a vital role for a diabetic patient’s [7]. Veena Vijayan V. and Aswathy Ravi kumar worked with Pima Indian Diabetic dataset. EM algorithm, KNN algorithm, K-means algorithm, amalgam KNN algorithm and ANFIS algorithm was the main focused algorithm in this paper to predict diabetes. Here, they considered 9 attributes and 768 instances and

- Nusrat Jahan, Senior Lecturer, Department of CSE, Daffodil International University, Dhaka, Bangladesh. E-mail: nusratjahan.cse@diu.edu.bd.
- Md. Aminul Islam, Department of Health Informatics, Bangladesh University of Health Sciences(BUHS), Dhaka, Bangladesh. E-mail: aminul.buhs@gmail.com.
- Abdullah Al Mamun, Student, Institute of IT, University of Dhaka, Bangladesh. E-mail: auvyauvy@gmail.com.

Amalgam KNN algorithm produced greater than 80% accuracy [8]. Vinaytosh Mishra, Dr. Cherian Samuel, and Prof. S.K.Sharma they used logistic regression classification techniques to predict the disease in early stages [9]. Meraj Nabi, Abdul Wahid, and Pradeep Kumar also worked with Pima Indian Diabetes data from UCI Repository. In this paper they explored evolutionary performance of Naive Bayes, Logistic Regression and Decision tree, Random forest using Pima Indian Diabetes data [10]. In 2015, Aiswarya Iyer, S. Jeyalatha, and Ronak Sumbaly diagnosed the diabetes in early stage by examining the patterns of dataset from Pima Indian Diabetes dataset. Through, they proposed Decision Tree (DT) and Naive Bayes algorithms for their diagnosis model but naive bayes gave the least error rate [11]. In 2017, K-means used for classifying diabetes data by Wenqian Chen and others. They considered Pima Indian Dataset for Type 2 diabetes diagnosis. Furthermore, they compared with some previous studies and found their proposed model provided better accuracy for Type 2 diabetes prediction. They got 90% accuracy [12]. In 2017 Md. Maniruzzaman et al. used Gaussian process (GP) for classifying Indian Pima diabetes dataset. In this paper they followed 3 types of kernel techniques for GP based classification. Here, authors' also compared with existing techniques namely- LDA, QDA and NB. Finally, they got 81.97% accuracy for their model [13]. In 2018, Mir A. and Dhage S. N. proposed a classifier model using weka 3.8.2 data analysis tool. In this paper 9 attributes used for predicting diabetes disease. Here, they used 4 types of machine learning algorithms: Naive Bayes, SVM, Random forest, and Simple CART. SVM worked better in their study. They found 79% accuracy level for SVM among 4 algorithms [14]. A. G. Januszewska et al. in their paper they focused on BMI for prevention of diabetes Type 2. They studied 175 participants who have higher risk level of Type 2 diabetes and they observed them around 12 months. In the next level, they found successful weight loss improve the risk level of a patient with high level of glucose and also increase physical activity as well [15]. Multilayer Perceptron (MLP) is the most widely used neural network classifier as well as J48 algorithm is a form of Decision tree algorithm. On the other hand, IBK is one of the most popular classifiers. In our study we found better result after applying IBK with 12 fold cross validation approach and the accuracy was 98.73%. In this paper, we collected data from our country and proposed model is also for our country. Here, we analyzed the data to predict the diabetes in early stage. In this paper, we used machine learning techniques to find out our outcomes and we set threshold to measure the diabetes risk level. In our study we found IBK approach generated better result (98.73%) for our dataset. This paper highlights the effectiveness to predict the diabetes in early stage that is more advantageous to reduce the long term complications.

3 METHODOLOGY

In this section we discussed about proposed method for predicting diabetes risk levels. Here, we explored the data collection way, data processing, worked flow procedure, factor scoring procedure and accuracy comparison of different algorithms.

3.1 Data Preparation

For making a suitable dataset we used a survey form to collect data from patient directly and total 555 instances/responses. For this reason we study and found 9 attributes which was more effective to find diabetes risk level. There are many more attributes which are also effective but in our study we considered 9 as showed in Table 1. Analyzing and classifying the prepared dataset, we applied the machine learning algorithms in three categories: Lazy, Function, and Trees to classify our dataset.

TABLE 1

CONSIDERED ATTRIBUTES

No	*Name of attributes*	*Type*
1	Age	Numeric
2	Gender	Numeric
3	Blood pressure(mm Hg)	Numeric
4	Exercise	Numeric
5	Body mass index(weight in Kg/(height in m)^2)	Numeric
6	Stress Level	Numeric
7	Genetic	Numeric
8	Sleeping Hrs.	Numeric
9	Smoking	Numeric
10	Class	Numeric

For this paper work we collected local data from different people by providing a survey form with nine attributes, Table 1 represents the attributes which was we considered for diabetes risk prediction process. Our total instance was 555, each with 9 attributes and one attribute for result/class level. We observe previous worked in the time to select attributes. We selected 9 attributes for predicting diabetes risk level which was also considered in our application. We try to avoid all those data which would be difficult to put in our application to calculate the risk level of diabetes. Fig. 1, prepared for presenting a flow chart that was illustrated our proposed model and we also followed this flow chat to implement our provided application. If we look at the flow chart then at first we found that have to train our machine with dataset and our data set classified into three categories based on three risk levels. Our application was implemented based on threshold value to predict diabetes risk in three levels. To implement the model we used IBK model to train our machine as IBK provided best result.

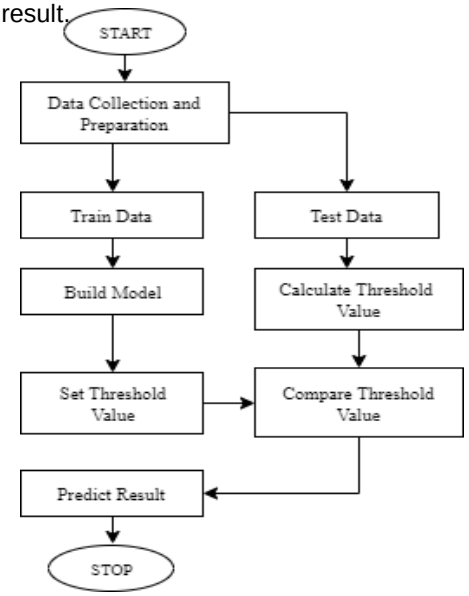

Fig. 1*. Flow chart to predict diabetes risk level*

For describing the threshold value we presented Table 2. This table indicates the attributes with a risk score, where we considered each attributes as a factor and set a risk score to measure the risk level of diabetes in three levels. Total risk score ranging from 0-14. Scoring is a challenging issue and in that case we considered total instances to set a score against each attribute. We used logistic regression approach to find a risk score. We considered every attribute individually with class and found out the risk score for a specific attribute. To the next, we decided the level as follows –

1. If the total risks score are less than or equal 6 then the suggestion is: It is risk free level or low level.
2. If greater than 6 but less than ten, then the suggestion is: To take necessary steps for controlling blood glucose and start proper exercise for medium level risk.
3. Finally, if greater than 10 then predict level is: High risk with diabetes.

At the time to implement an Android application we considered these three threshold values to predict the diabetes risk level of a user. Our three classification model also followed Table 2 to classify the dataset with better result as we compared with some previous study in next section.

TABLE 2

RISK SCORE CONSIDERED FOR EACH ATTRIBUTES

Name of attributes	Risk Scoring	Risk Prediction
Age(Years) <=39 40-49 >=50	 0 1 2	 Low Medium High
Gender Female Male	 2 1	 Low Medium
Blood pressure(mm Hg) No Yes	 0 1	 Low Medium
Body mass index(Kg/(m)^2) Underweight Normal Overweight	 0 1 2	 Low Medium High
Exercise <15 15-30 >30	 2 1 0	 High Medium Low
Stress Level Normal Medium High	 0 1 2	 Low Medium High
Genetic No Yes	 0 1	 Low High
Sleeping Hrs. <=5 5-7 >7	 2 1 0	 High Medium Low
Smoking No Yes	 0 1	 Low High
Risk Level (0 -14)	<=6 6-10 >10	Low Medium High

3.2 RESULT ANALYSIS

Now, it is time to discuss about classification approach. We applied three machine learning algorithms: Multilayer perception, Decision tree and IBK. Here, multilayer perception is a neural network base approach; Decision tree is a level based algorithm which is produce root and child to find the required data and finally IBK is an algorithm that is followed nearest neighbor selection approach. In the next step we trained the machine only using cross validation approach. 5, 10, 12, and 15 fold we considered for training the machine. In our study, cross validation worked better for our dataset. Now, Table 3 to understand the classification result of the model according to different parameters. Here, we presented True Positive (TP) rate that is showed the correctly predicted result, False Positive (FP) for presenting actual false result predict as true, Precision actually follow the following equation –

$$\text{Precision} = TP / (TP+FP) \quad (1)$$

Finally the Accuracy presented the overall result of a model. Accuracy for presenting the model performance at a glance. Equation (2) for represent the accuracy.

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (2)$$

Here, TN= True negative and FN= False negative. TN is for actual false result correctly predict as false and FN to predict actual true as false. That is why FN is so dangerous among those all measurement matrix.

TABLE 3

PERFORMANCE MEASUREMENT

Cross Validation	Measurement Matrix	IBK	Decision Tree	Multilayer Perception
5 Fold	TP	0.981	0.909	0.957
	FP	0.023	0.035	0.029
	Precision	0.962	0.940	0.952
	ROC Area	0.995	0.982	0.986
	Accuracy %	97.83	94.41%	96.57
10 Fold	TP	0.981	0.923	0.938
	FP	0.023	0.037	0.037
	Precision	0.962	0.937	0.938
	ROC Area	0.996	0.983	0.973
	Accuracy	97.83	94.77	95.31
12 Fold	TP	0.995	0.981	0.966
	FP	0.017	0.020	0.029
	Precision	0.972	0.967	0.953
	ROC Area	0.996	0.997	0.980
	Accuracy	98.73	98.01	96.93
15 Fold	TP	0.995	0.981	0.957
	FP	0.017	0.035	0.037
	Precision	0.972	0.944	0.939
	ROC Area	0.996	0.994	0.976
	Accuracy	98.73	97.11	96.03

We classified the dataset using Weka 3.8.0 by applying three different machine learning algorithms as we mentioned before. Weak is a well-known tool which is actually a collection of machine learning algorithms for explaining different types of data analysis based problems. Java programming is the basic

of this tool and it can runs on almost any platform. We can also visualize our dataset and pattern through this tool [16]. After found a better threshold value for each level, we compared it with the new sample's threshold value. We used the final outcome to develop the application in an Android platform.

TABLE 4

COMPARATIVE STUDY

Feature	Applied Algorithm	Accuracy
768 instances with 8 attributes [17]	Found Linear-SVM worked better for their proposed model in 2018	89%
Pima Indian Diabetes Dataset [12]	K-means clustering performed well in this study, 2017	90.04%
Our proposed model	IBK was performed better	98.73

Table 4 for showing some recent study at the same time it is also possible to visualize the comparative study with this table. Recently, there are many work done based on diabetes prediction as it is one of the most crucial diseases which has no cure. So, it is necessary to predict it in early stage. Most of the research are trying to do this as accurately as possible.

4 PROPOSED PLATFORM

Implement the model is necessary for prevent this diseases in early stage. This section for describing the implementation approach. After analyzing the collected data we implemented an Android application that is easy to use, because it is a worldwide available mobile platform and user can predict the type-2 diabetes risk level in anyplace at any time. To implement the model we followed Fig. 1 as a work flow chart. Based on our approach here, Fig. 2 shows the first user interface.

***Fig. 2**. First user Interface*

From first interface user will have to face 9 selected questions and needed to answer all required questions that actually indicated our selected nine attributes one by one. Finally, user will be able to know about the prediction result on diabetes risk level. Here, we provided some user interfaces as a sample in Figure 3 and 4.

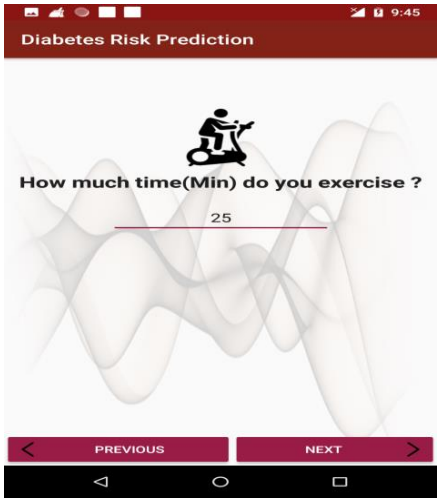

***Fig. 3**. Questionnaires format*

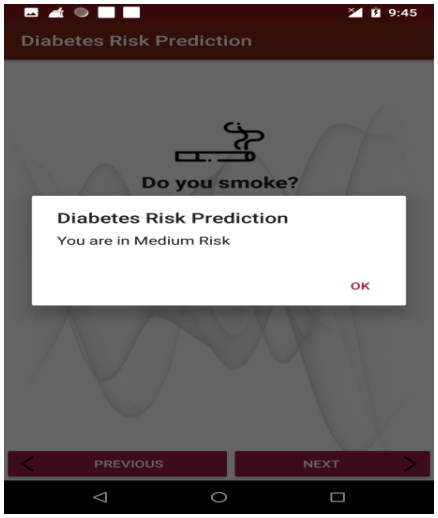

***Fig. 4**. Result Prediction*

5 CONCLUSION

Early prediction is a better approach rather than cure any diseases. Diabetes is one of the most familiar diseases which have no cure. In our country people are so much busy with their daily life activities but they have not enough time to care about their health but healthy life is one of our daily needs. This paper is for increasing awareness about diabetes because in our country percentage of diabetic patient increases terrifically. So, we should take necessary steps as soon as possible to stop this phenomenon. Here, we provided an Android application that helps us to predict the type-2 diabetes risk level that would be helpful to take necessary steps. In future this paper study can helps us to design more helpful application and to increase awareness about diabetes. We have also a future plan to develop more application for predicting other diseases which is more dangerous in our generation.

ACKNOWLEDGMENT

The authors sincerely like to thank the Daffodil International University, Dhaka, Bangladesh to motivate us to do this research work.

REFERENCES

[1] John Elflein, "Percentage of diabetics in the global adult population in 2019 and 2045", https://www.statista.com/statistics/271464/percentage-of-diabetics-worldwide/ (Dec 10, 2019)

[2] "Diabetes", http://www.who.int/mediacentre/factsheets/fs312/en/ (30 October 2018)

[3] "Classification and Diagnosis of Diabetes", American Diabetes Association, Diabetes Care 40(Supplement 1): S11-S24. http://care.diabetesjournals.org/content/40/Supplement_1/S11 (2017)

[4] Dabelea D, Mayer-Davis EJ, Saydah S, et al.; SEARCH for Diabetes in Youth Study. Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. JAMA 2014;311:1778–1786

[5] Aiswarya Iyer, S. Jeyalatha, Ronak Sumbaly, "Diagnosis of diabetes using classification mining techniques". International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.5, No.1, January 2015, pp. 1-14

[6] Sarwar N and et al., "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies". Lancet, volume 375 on page 2215, 2010

[7] Md. Aminul Islam and Nusrat Jahan,"Prediction of Onset Diabetes using Machine Learning Techniques". International Journal of Computer Applications 180(5):7-11, December 2017

[8] Veena Vijayan V., and Aswathy Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus".International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014

[9] Vinaytosh Mishra, Dr. Cherian Samuel, and Prof. S.K.Sharma, "USE OF MACHINE LEARNING TO PREDICTTHE ONSET OF DIABETES". International Journal of Recent advances in Mechanical Engineering (IJMECH) Vol.4, No.2, May 2015

[10] Meraj Nabi, A bdul Wahid, and Pradeep Kumar, "Performance Analysis of Classification Algorithms in Predicting Diabetes". International Journal of Advanced Research in Computer Science, Volume 8, No. 3, March – April 2017, ISSN No. 0976-5697

[11] Aiswarya Iyer, S. Jeyalatha and RonakSumbaly, "DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES". International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[12] Wenqian Chen, Shuyu Chen, Hancui Zhang, and Tianshu Wu, "A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree". International Conference on Engineering, Technology and Innovation, 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017

[13] Md. Maniruzzama et al., "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm" Computer Methods and 152 (2017) 23–34

[14] Ayman Mir, and Sudhir N. Dhage, "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare". Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018

[15] Aleksandra Gilis-Januszewska and et al., "Determinants of weight outcomes in type 2 diabetes prevention intervention in primary health care setting (the DE-PLAN project)". Gilis-Januszewska et al. BMC Public Health (2018)

[16] Sapna Jain 2.M Afshar Aalam3. M. N Doja,"K-MEANS CLUSTERING USING WEKA INTERFACE", Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, February 25 – 26, 2010

[17] H. Kaur and V. Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, Applied Computing and Informatics, https://doi.org/10.1016/j.aci.2018.12.004