Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques

Chapter · January 2020

DOI: 10.1007/978-981-13-8798-2_12

4 authors:

M M Faniqul Islam

4 PUBLICATIONS **174** CITATIONS

SEE PROFILE

Rahatara Ferdousi

University of Ottawa

21 PUBLICATIONS **313** CITATIONS

SEE PROFILE

Sadikur Rahman

Metropolitan University

1 PUBLICATION **174** CITATIONS

SEE PROFILE

Humayra Bushra

Metropolitan University

1 PUBLICATION **174** CITATIONS

SEE PROFILE

Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques

M M Faniqul Islam[1], Rahatara Ferdousi[2], Sadikur Rahman[3], and Humayra Yasmin Bushra[4]

[1] Queen Mary University of London, United Kingdom
m.islam@smd17.qmul.ac.uk
[2] Metropolitan University Sylhet, Bangladesh
rahatara@metrouni.edu.bd
[3] Metropolitan University Sylhet, Bangladesh
rahmansadik004@gmail.com
[4] Metropolitan University Sylhet, Bangladesh
humayrabushra234@gmail.com

Abstract. Diabetes is one of the most fastest growing chronic life threatening diseases that have already affected 422 million people worldwide according to the report of World Health Organization(WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long term asymptomatic phase.The early diagnosis of diabetes is only possible by proper assessment of both common and less common sign symptoms, which could be found in different phases from disease initiation up to diagnosis. Data mining classification techniques have been well accepted by researchers for risk prediction model of disease.To predict the likelihood of having diabetes require a dataset which contains the data of newly diabetic or would be diabetic patient. In this work, we have used such a dataset of 520 instances which has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. We have analyzed the dataset with Nave Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm and after applying 10 Fold Cross Validation and Percentage Split evaluation techniques, Random forest has been found having best accuracy on this dataset. Finally, a commonly accessible, user-friendly tool for the end user to check the risk of having diabetes from assessing the symptoms and useful tips to control over the risk factors has been proposed.

Keywords: Diabetes Risk, Symptom, Early Stage, Data Mining, KDD, Dataset, Evaluation Model, Supervised Learning Algorithms, Unsupervised Learning Algorithms, Dataset, Mining Tools

[1] Corresponding Author: Rahatara Ferdousi, email: rahatara@metrouni.edu.bd, Phone: +881718656060, address: 131/8, Tootpara Taltola Hospital Road Khulna

1 Introduction

Diabetes Mellitus, a chronic metabolic disorder, is one of the fastest growing health crises of this era regardless of geographic, racial, or ethnic context. Commonly, we know about two types of diabetes called type 1 and type 2 diabetes. Type 1 diabetes occurs when the immune system mistakenly attacks the pancreatic beta cells and very little insulin is released to the body or sometimes even no insulin is released to the body. On the other-hand, Type 2 diabetes occurs when our body doesn't produce proper insulin or the body becomes insulin resistant. Some researchers divided diabetes into Type 1, Type 2, and gestational diabetes [5]. Gestational diabetes is a type of diabetes which occurs only in pregnancy due to hormonal changes. The common symptoms of diabetes are polyuria, polydipsia, polyphagia, sudden weight loss(usually Type 1), weakness, obesity(usually Type 2), delayed healing, visual blurring, itching, irritability, genital thrush, partial paresis, muscle stiffness, alopecia, etc [5][2].

This could be a clear evidence that, according to WHO, the number of the diabetic patient had been sharply increased from 108 million in 1980 to 422 million in 2014 [1]. The most alarming fact is that more than 80% of diabetic people were from low and middle-income countries in 2013 and the prevalence is surging up in these countries. Recently Diabetes Australia has published that, Diabetes even may exist up to 7 years before clinical diagnosis[3], which was even up to twelve years previously noted by Harris et al.[21]. Within this time frame, people may gradually suffer from fatal complications like heart attacks, strokes, eye damage resulting in blindness, foot ulcer, amputation of the affected limb, kidney damage and other forms of multi organ damage [21]. Most of the cases, these complications would be easily controlled or even prevented in some cases with early detection and treatment initiation that could possibly save around 1415 AUD [3]. The degree of diabetic complication is more when the period between onset of disease and treatment initiation is longer [21]. According to Diabetes Australia, failure in early detection of TYPE 2 could cost Australian health care system more than 700 million dollars each year [3]. In 2017, the total expenditure of diagnosed diabetes in the United States alone was 327 billion USD [2].In [12], in the year 2011, China had experienced 90 million (9% of the population), India had 61.3 million (8% of the population) and Bangladesh had 8.4 million(10% of the population). Comparing to developed countries like Australia and The USA, low and middle-income countries cannot afford the burden of managing such a costly disease like diabetes, the prevalence of which is increasing in an alarming rate. Therefore, early diagnosis and initiation of appropriate therapeutic management may play a pivotal role in the patient outcome and reduce the gross national expenditure and production lost. Another considering issue is that globally, OGTT(Oral Glucose Tolerance Test), HbA1c are widely accepted method of diagnosis of diabetes which are usually referred by the physician after developing patient's sign symptoms. However, these tests are not so cheap, lab reagent and technician dependent as well as time-consuming, these tests are not available in remote settings. As the protocol of the treatment is not only long term but also expensive, the earlier detection of diabetes is beneficial in terms

of patient's health, individual and national expenditure, as well as productivity [11].

In this modern era of technology, computer technology can help us to detect diseases accurately and can save our time and money. Data mining is an important field of computer science which is used for prediction. It is the process of discovering new data from previously known data through data analysis [21]. To predict a disease using data mining approaches, we need its symptoms along with clinical data. Symptoms are a very important factor for new patients and early stage prediction since, they have no data except symptoms. We also need clinical data for analyzing and discovering new data.

Early assessment of symptoms can be possible by creating mass awareness, manual assessment by health workers/assistants (where doctor/facilities are not available due to remoteness) in the rural setting, or by some user-friendly and cost-effective system. This system should be designed for specific target users so that it is easily accessible for mass people. As a result, early diagnosis of diabetes, pre-diabetes, risk of diabetes through symptom assessment by any means not only can prevent fatal outcomes of diabetes but also can save up such a huge financial expenditure as stated above, as well as increase the national productivity level, which could bring fruitful outcomes in low and middle income countries. Thus in this paper, we are providing analysis on a newly created dataset of 520 instances using different data classification algorithm to find one that provides better accuracy. Finally, we have proposed a tool for the end users to predict the likelihood of diabetes risk at its early stage, using patients symptoms with the help of data mining techniques.

2 Literature Review

In this section different research works that were envisioned to predict diabetes using data mining have been provided with their remarkable contribution.

In [13], authors collected 865 data with 9 attributes called Sex, Diastolic B.P, Plasma glucose, Skin fold thick, BMI, Diabetes Pedigree type, No. of times Pregnant, 2 hour Serum Insulin and Diabetes probability and used WEKA 3.6.6 for the experiment. They found 100%accuracy with J48 (C4.5), 98.48% with the Decision Tree, 97.85% with the Neural Network, 96.54% with JRip and 95.85% with Nave Bayes algorithm. They also calculated the performance over time.

In [9], the author used 738 patient's data for experimental analysis.To predict diabetes they introduced algorithms like CNN, KNN, SVM, SVM+LDA, NB, SVM, ID3, C4.5, CART for comparing the analysis on the dataset. The best accuracy at 88.10% was achieved using SVM and LDA algorithm together.

In [4], the author compared three machine learning algorithms to predict diabetes. They introduced SVM, Logistic regression, ANN to seven attributes of their data including the Glucose, Blood Pressure, Skin, Thickness, Insulin, BMI, Diabetes Pedigree Function, and the age. After comparing their features, the researcher opinionated that the Support Vector Machine (SVM) found SVM as the best classification method.

In [17], the author used Artificial Neural Network for predicting diabetes They collected 250 diabetes patients data from Pusat Perubatan University Kebangsaan Malaysia, Kuala Lumpur and between 25 to 78 years old. They used MATLAB to train data. They had done Regression analysis using different algorithms, BFGS Quasi-Newton, Bayesian Regulation, Levenberg-Marquardt. They found 88.8% accuracy with Bayesian Regulation algorithm.

In [8], authors used Pima Indian Diabetes dataset and WEKA as their software tool for dataset testing. They tested their dataset with Nave Bayes (NB), Random Forest (RF), and function-based Multilayer Perceptron (MLP) algorithms and used different test methods called FCV, PS, UTD. They also predicted with pre-processed and without pre-processed data and made a convenient table on their result. They found 100% accuracy with Random Forest algorithm with UTD method. However, the author stated that pre-processed data can give more accuracy in the Nave Bayes algorithm.

In [22], the author has created a new model for type 2 diabetes patients treatment. He collected 318 medical records with 9 nominal attributes including the patient's Gender, Age, Smoking, History of hypertension, Renal problem, Cardiac problem, Eye problem. Duration of Diabetes Basic control was used as a class level attribute. He used the J48 algorithm and found an accuracy rate of 70.8% and ROC (Receiver operating characteristic) rate was 0.624.

In [23], authors have predicted diabetes with supervised and unsupervised learning. They used the software tool WEKA to find a better prediction algorithm in machine learning. Finally, they concluded that ANN or Decision tree is the best way for diabetes prediction.

In [10], the author used Logistic Regression to predict diabetes. In their data, they used Age, Smoking, Parental Diabetes Mellitus, Hypertension & Waist Circumference, Sex, BMI and HBA1C information as the attribute. The data analysis was conducted using the software tool IBM SPSS 20.0. In result, they found the likelihood 78.5565%, Cox & Snell R Square Nagelkerke Square .628, and Nagelkerke R Square 0.839.

In [21], the author aimed to forecast whether the patient has been affected by diabetes or not using the data mining tools and the MV dataset. This dataset contains 1024 complete instances of 26 Parameters. MV dataset was collected from various districts people using Questionnaires. They experimented Decision Trees to predict diabetes for local and systemic treatment.

3 Proposed System Architecture

The proposed system architecture is shown in Fig. 1. The dataset containing the information about the symptoms of the patients will be feed to the prediction algorithms like Nave Bayes, Decision Trees, Logistic Regression and Random forest algorithm. Then the performance of the algorithms will be tested with appropriate evaluation model, in particular, 10 fold Cross-validation and percentage split techniques. Then the best algorithm chooses to build the system

for the end users using the dataset as Database. Taking the symptom from the user as input the system will support the user for risk prediction.

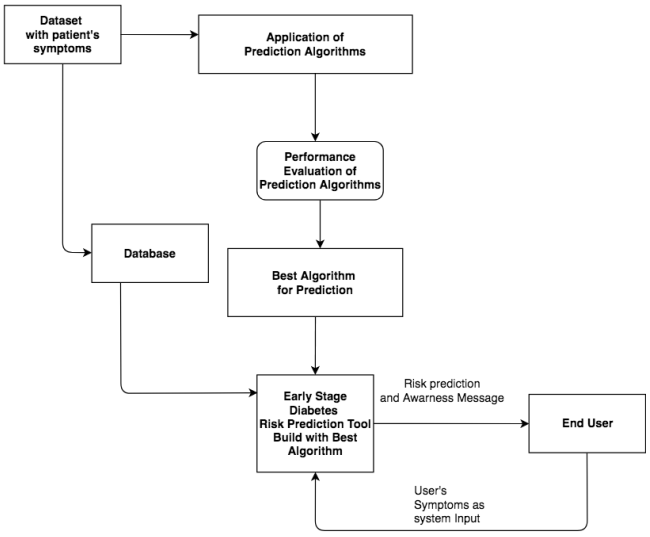

Fig. 1: Proposed System Architecture

.

4 Methodology

The dataset was analyzed using following classification algorithms. The data analysis procedure can be formulated according to the algorithm 1.

Data: Diabetes Symptom Dataset
Result: Classification Technique with Best Accuracy
i=1;
bestAccuracy=0;
maxAccuracy=accuracy(1st algo);
while *i is less than or equal to numberOfAlgorithm* **do**
  if *accuracy(ith algo) greater than maxAccuracy)* **then**
    maxAccuracy=accuracy(ith algo);
    i++;
  end
  bestAccuracy=maxAccuracy;
end

Algorithm 1: Algorithm for Dataset Analysis

4.1 Naive Bayes(NB)

Naive Bayes uses a probabilistic algorithm. The algorithm assumes the features and variables provided are independent to one another. It is carried out by

using a probabilistic approach which determines class probabilities and predicts most probable classes. The following equation from (1) to (3) represent the classification formula, where Pos and Neg represent person with diabetes risk and without diabetes risk, which are the values of the class attribute for this dataset. X is the instances of the dataset as well as person.

$$P(Pos|X) = P(x_1|pos) * P(x_2|pos) * \cdots * (x_n|pos) * P(Pos) \tag{1}$$

$$P(Neg|X) = P(x_1|neg) * P(x_2|neg) * \cdots * (x_n|neg) * P(Neg) \tag{2}$$

$$P(x_i|Pos) = \frac{(TotalPos|x_i)}{TotalPos} \tag{3}$$

where i is an increment until it reaches n(total attributes for our data)

4.2 J48 Decision Tree(J48 DT)

J48 algorithm is a kind of decision tree which belongs to the supervised learning algorithm. It is one of the most important classifiers as it is easy and simple to implement. Using the decision tree, a dataset is broken down into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.The algorithm uses equation (4) to (6) to find information gain for our dataset to predict the outcome.

$$E(P) = -\sum_{j=1}^{n} \frac{|P_j|}{|P|} log \frac{|P_j|}{|P|} \tag{4}$$

$$E(j|P) = \frac{|P_j|}{|P|} log \frac{|P_j|}{|P|} \tag{5}$$

$$Gain(P, j) = E(P - E(j|P)) \tag{6}$$

P represents total instance, n represents total number of classes, and j represents total number of attributes in the dataset.

4.3 Logistic Regression(LR)

The LR classifier works with the class and uses multinomial logistic regression model with a ridge estimator. For k number of classes and for instances n with attributes m, the parameter matrix B can be calculated with the matrix given in equation (7).

$$B = m * (k - 1) \tag{7}$$

The probability for class j with the exception of the last class is stated in (8). and the last class probability given in (9).

$$P_j(X_i) = \frac{exp^{\sum_{j=1}^{k-1} X_i B_j}}{(1 + exp)^{\sum_{j=1}^{k-1} X_i B_j}} \tag{8}$$

$$P'_j(X_i) = \frac{1}{(1 + exp)^{\sum_{j=1}^{k-1} X_i B_j}} \tag{9}$$

Thus the negative multinomial log-likelihood is -

$$\begin{aligned} L = -\sum_{i=1}^{n} [\Sigma_j^{k-1} (Y_{ij} * ln(P_j(X_i))) + (1 - \Sigma_{j=1}^{k-1}) * \\ ln(1 - \Sigma_{j=1}^{k-1} P_j(X_i))] + ridge * B^2 \end{aligned} \tag{10}$$

In order to determine accuracy B, L is kept minimized as much as possible.

4.4 Random Forest(RF)

Random forest uses bagging method to train the dataset. For a training set of $X = x_1, , x_n$ and $Y = y_1, ...y_n$ it selects random sample B times with replacement of the training set and fits trees to these samples. After training, it predicts unseen samples x' by averaging the predictions from all the individual regression trees on x' as shown in equation (11) and also by taking the majority vote in the case of classification trees.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{11}$$

5 Experimental Analysis

Dataset details and the result analysis is represented in this section.

5.1 Dataset Details

This dataset contains reports of diabetes-related symptoms of 520 persons. It includes data about peoples including symptoms that may cause diabetes. This dataset has been created from a direct questionnaire to people who have recently become diabetic, or who are still non-diabetic but having few or more symptoms. The data has been collected from the patients using direct questionnaire from Sylhet Diabetes Hospital of Sylhet, Bangladesh.

Table 1: DESCRIPTION OF DATASET

	Number of Attributes	Number of Instances
Diabetes Symptom Dataset	16	520

Table 2: DESCRIPTION OF ATTRIBUTE

Attribues	Values
Age	1.20-35, 2.36-45, 3.46-55,4.56-65, 6.above 65
Sex	1.Male, 2.Female
Polyuria	1.Yes, 2.No.
Polydipsia	1.Yes, 2.No.
sudden weight loss	1.Yes, 2.No.
weakness	1.Yes, 2.No.
Polyphagia	1.Yes, 2.No.
Genital thrush	1.Yes, 2.No.
visual blurring	1.Yes, 2.No.
Itching	1.Yes, 2.No.
Irritability	1.Yes, 2.No.
delayed healing	1.Yes, 2.No.
partial paresis	1.Yes, 2.No.
muscle stiffness	1.Yes, 2.No.
Alopecia	1.Yes, 2.No.
Obesity	1.Yes, 2.No.
Class	1.Positive, 2.Negative.

The data pre-processing has been conducted by handling the missing values following the technique of ignoring the tuples with incomplete values. After pre-processing, 500 instances have been remained in total. Among them, 314 are positive values and 186 are negative values. The detail description of the attributes is shown in Table 2. Two class variables are used to find whether the patient is having a risk of diabetes (positive) or not (negative).

5.2 Result Analysis

Performance of different Data Mining techniques on our dataset with detailed accuracy information is represented in the following tables. Although Nave Bayes classifier is one of the most popular algorithms for data prediction, in case of our dataset, the accuracy of it was the lowest for both the Cross-validation method and also for the Percentage split. However, the best result was achieved using Random Forest Algorithm where using 10 fold cross validation 97.4% instances were classified correctly and using percentage split technique it could classify 99% of the instances correctly as shown in Table 3. For the more semantic view of the performance of used algorithms using both evaluation techniques are depicted in graphs. In Fig. 2, the performance of the algorithms using Cross-validation evaluation is depicted and in Fig. 3, the results from percentage split have been shown to represent the comparative accuracy of the used algorithms.

Table 3: COMPARISON OF EVALUATION METRICS USING 10 FOLD CROSS-VALIDATION AND PERCENTAGE SPLIT(80:20)

Evaluation Metrics	Cross-Validation				Percentage Split			
	NB	LR	J48	RF	NB	LR	J48	RF
Total Number of Instances	500	500	500	500	100	100	100	100
Correctly Classified Instances	437	462	478	487	88	91	95	99
	87.4%	92.4%	95.6%	97.4%	88%	91%	95%	99%
Incorrectly Classified Instances	63	38	22	13	12	9	5	1
	12.6%	7.6%	4.4%	2.6%	12%	9%	5%	1%

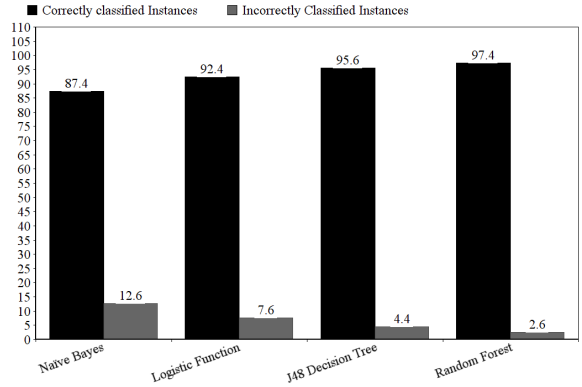

Fig. 2: Performance of Classification Algorithms Using Cross-Validation Technique

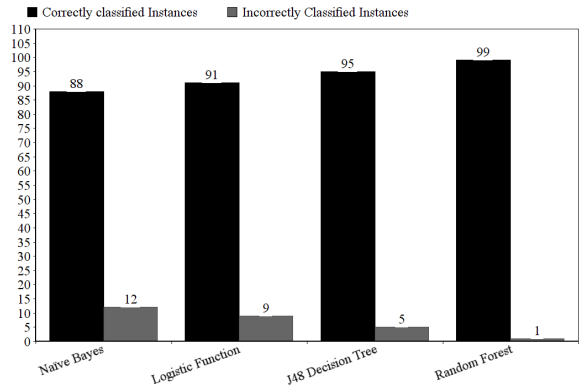

Fig. 3: Performance of Classification Algorithms Using Percentage Split Technique

Table 4: COMPARISON OF PERFORMANCE PARAMETERS USING 10 FOLD CROSS-VALIDATION

Performance Parameters	Class	Weighted Average			
		NB	LR	J48	RF
TP Rate	Positive	0.869	0.936	0.949	0.978
	Negative	0.886	0.903	0.968	0.968
	Weighted Average	0.874	0.924	0.956	0.974
FP Rate	Positive	0.118	0.097	0.032	0.032
	Negative	0.131	0.064	0.051	0.022
	Weighted Average	0.123	0.084	0.039	0.029
Precision	Positive	0.925	0.942	0.980	0.981
	Negative	0.800	0.894	0.918	0.963
	Weighted Average	0.879	0.924	0.957	0.974
Recall	Positive	0.869	0.936	0.949	0.978
	Negative	0.882	0.903	0.968	0.968
	Weighted Average	0.874	0.924	0.956	0.974
F-measure	Positive	0.897	0.939	0.964	0.979
	Negative	0.839	0.898	0.942	0.965
	Weighted Average	0.875	0.924	0.956	0.974

Table 5: COMPARISON OF PERFORMANCE PARAMETERS USING PERCENTAGE SPLIT

Performance Parameters	Class	Weighted Average			
		NB	LR	J48	RF
TP Rate	Positive	0.930	0.947	0.965	1.000
	Negative	0.814	0.860	0.930	0.977
	Weighted Average	0.880	0.910	0.950	0.990
FP Rate	Positive	0.186	0.140	0.070	0.023
	Negative	0.070	0.053	0.035	0.000
	Weighted Average	0.136	0.102	0.055	0.013
Precision	Positive	0.869	0.900	0.948	0.983
	Negative	0.897	0.925	0.952	1.000
	Weighted Average	0.881	0.911	0.950	0.990
Recall	Positive	0.930	0.947	0.965	1.000
	Negative	0.814	0.860	0.930	0.977
	Weighted Average	0.880	0.910	0.950	0.990
F-measure	Positive	0.898	0.923	0.957	0.991
	Negative	0.854	0.892	0.941	0.988
	Weighted Average	0.879	0.910	0.898	0.980

6 Proposed Tool For the End Users

To provide instant help to the mass people for diabetes risk prediction, regardless location age or educational background an easy and globally accessible system is required. As the web technology has quickly become the worlds most common way of searching data and services, a simple website could be undertaken to check the risk of the diabates using users symptom as input.This website should provide both prediction of likelihood of having diabetes and some useful health tips for both the diabetic and non-diabetic. Useful health tips for a non-diabetic can reduce or delay the risk of him/her to have diabetes. A demo homepage of our proposed tool is shown in Fig. 4

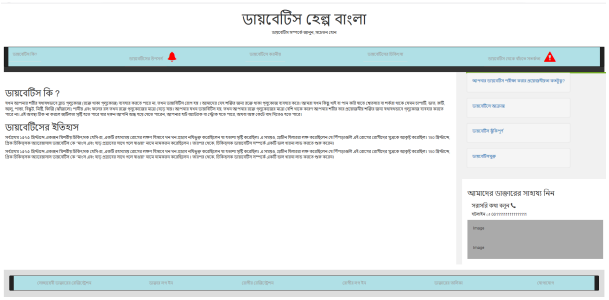

Fig. 4: Homepage of Proposed tool

7 Conclusion

The potentiality of diabetes is increasing among people of all age. The present study says that detection of diabetes at its early stage can play a pivotal role for treatment. Simple awareness measures such as low sugar diet, regular physical activity and healthy lifestyle can avoid obesity. As the Data mining methods, techniques and tools are becoming more promising to predict diabetes and eventually number of patients reduce the treatment cost, its role in this medical health care is undeniable. The main contribution is to find out the best algorithm for the prediction on newly created datasets made for diabetic risk prediction. We found that the Random Forest algorithm had performed with the best accuracy in percentage split evaluation test. Finally, a tool for the marginal user has been proposed, which can be used for diabetic risk prediction, awareness creation and instant help. However, this research can be updated regularly with a dataset with more instances and can apply other widely accepted other data mining technologies for prediction purpose. As the system has been only prototyped, a deploying version of this system can be considered as a sustainable outcome of this research.

Ethical Approval:
All procedures performed in studies involving human were in accordance with the ethical standards of the institution at which the studies were conducted and ethical approval was obtained from Sylhet Diabetic Hospital, Sylhet Bangladesh.

Ref: S.D.A/88

Informed Consent:
Informed consent was obtained from all individual participants included in the study.

References

1. Diabetes,World Health Organization(WHO), 30 October 2018.[Online] Available:https://www.who.int/news-room/fact-sheets/detail/diabetes.
2. Statistics About Diabetes, American Diabetes Association,22 March 2018.[Online] Available:https://http://www.diabetes.org.
3. Failure to detect type 2 diabetes early costing $700 million per year, Diabetes Australia ,8 July 2018.[Online] Available:https://www.diabetesaustralia.com.au.
4. Tejas N. Joshi, Prof. Pramila M. Chawan - "Diabetes Prediction Using Machine Learning Techniques". S. Dewangan.et.al. *Int. Journal of Engineering Research and Application* ,ISSN: 2248-9622, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13.
5. The 6 Different Types of Diabetes. (March 5, 2018). The Diabetic Journey, Retrieved from https://thediabeticjourney.com/the-6-different-types-of-diabetes.
6. R Manimaram, De.M.Vanitha, Novel Approach to Prediction of Diabetes using Classification Mining Algorithm. *International Journal of Innovative Research in Science, Engineering and Technology*, Vol 6, Issue 7, july 2017.
7. Vrushali Balpande, Rakhi Wajgi, Review on Prediction of Diabetes using Data Mining Technique, *International Journal of Research and Scientific Innovation (IJRSI)*Volume IV, Issue IA, January 2017 — ISSN 23212705.
8. Asir A.G Singh, E.J. Leavline and B. S. Baig, Diabetes Prediction Using Medical Data, *Journal of Computational Intelligence in Bioinformatics*, vol. 10, no.1 (2017) pp.1-8.
9. Pragati Agrawal, Amit kumar Dewangan - "A BRIEF SURVEY ON THE TECHNIQUES USED FOR THE DIAGNOSIS OF DIABETES-MELLITUS". *International Research Journal of Engineering and Technology (IRJET)*. e-ISSN: 2395 -0056; p-ISSN: 2395-0072.Volume: 02 Issue: 03 — June-2015 .
10. Vinaytosh Mishra, Dr. Cherian Samuel, Prof. S.K.Sharma3 - "USE OF MACHINE LEARNING TO PREDICT THE ONSET OF DIABETES" . *International Journal of Recent advances in Mechanical Engineering (IJMECH)* Vol.4, No.2, May 2015
11. Ramachandran, A. "Know the signs and symptoms of diabetes." *The Indian journal of medical research* 140.5 (2014): 579.
12. Akter, Shamima, et al. "Prevalence of diabetes and prediabetes and their risk factors among Bangladeshi adults: a nationwide survey." *Bulletin of the World Health Organization* 92 (2014): 204-213A.
13. VelidePhani Kumar, Lakshmi Valide. A data mining approach for prediction and treatment of diabetes disease. *International Journal of Science Inventions Today*, 2014; ISSN 2319-5436.
14. Prakash Mahindrakar, Dr. M. Hanumanthappa, Data Mining In Healthcare: A Survey of Techniques and Algorithms with its Limitation and Challenges, *International Journal ofEngineering Research and Applications*, ISSn: 2248-9622, Vol.3 Issue 6, Nov-Dec 2013.

15. Dr. Sudhir B. Jagtap, Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA" *International Conference in Emerging Trends in Science, Technology and Management*-2013
16. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
17. Muhammad Akmal Sapon, Khadijah Ismail and Suehazlyn Zainudin - "Prediction of Diabetes by using Artificial Neural Network". 2011 *International Conference on Circuits, System and Simulation* IPCSIT vol.7 (2011) (2011) IACSIT Press, Singapore.
18. DIABETES: A NATIONAL PLAN FOR ACTION. THE IMPORTANCE OF EARLY DIABETES DETECTION, ASPE, 12/01/2004.
19. Tapp, Robyn J., et al. "Albuminuria is evident in the early stages of diabetes onset: results from the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab)." *American journal of kidney diseases* 44.5 (2004): 792-798.
20. S. Wild, G. Roglic, A. Green, R. Sicree and H. King. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 27(5): 1047- 1053.
21. Harris, Maureen I., et al. "Onset of NIDDM occurs at least 47 yr before clinical diagnosis." *Diabetes care* 15.7 (1992): 815-819.
22. Ahmed (2016b). "Developing a Predicted Model for Diabetes Type 2 Treatment Plans by Using Data Mining".
23. Rabina1, Er. Anshu Chopra2 - "DIABETES PREDICTION BY SUPERVISED AND UNSUPERVISED LEARNING WITH FEATURE SELECTION". ISSN: 2454-132 ,(Volume2, Issue 5).