A Novel Approach to Predict Diabetes Mellitus by Statistical Analysis and using Advanced Classification Algorithm

Saima Sultana, Mahmudul Hasan Khandaker, Abdullah Al Momen, Mohoshi Haque, Nazmus Sakib
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology, Dhaka-1208, Bangladesh

ABSTRACT

Diabetes is a severe, enduring disorder with a huge impact on the existence and health of individuals and the people around them. It happens due to insufficient production of insulin in human body. After a thorough research on this disease, it can be said that diagnosing diabetes at the early stage can help patients to control it and also knowing the probability of having the disease can be useful to the patients for taking necessary steps. So, for the prediction of this disease, a different approach has been taken which is developing a mathematical equation. To develop this equation, some basic medical information of a person have been used as parameters. Using this equation, 80% accuracy has been achieved. Three machine learning algorithms have been used: Logistic Regression, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) on the dataset to verify the credibility of this equation. The accuracy attained for Logistic Regression, SVM and KNN is 86%, 91% and 83% respectively.

Keywords

Diabetes Mellitus, Logistic Regression, SVM, KNN, Machine Learning Algorithms, Prediction System, Age, Body Mass Index (BMI), Blood Pressure, Blood Sugar, Exercise Time and Sleeping Time.

1. INTRODUCTION

Nowadays healthcare monitoring data is growing immensely and it has different properties such as structured, semi-structured and unstructured. In recent days, data-science and information technology play a vital role in providing services in the area of healthcare [1].

Diabetes Mellitus known as diabetes is an organ disorder, occurs when the pancreas does not produce adequate insulin hormone to digest carbohydrate food or the body cells does not respond to insulin anymore. Foods that contain an extensive amount of glucose are oxidized for the production of energy in human body cells. By using this type of energies, the human body operates its physical activities. Insulin controls blood sugar in the human body [2]. It is produced in a natural way in a healthy human body. But in the case of a diabetic patient, the production of this insulin hormone decreases and therefore, blood sugar level increases in comparison with a sound body. This high blood sugar often causes nerve damage, kidney damage, eye damage, heart disease, skin hearing impairment, etc. Loss of limb, blindness is also some of the effects [3].

Diabetes is one of the topmost 10 reasons of death in grownups, and it was predicted that it had affected four million deaths worldwide in 2017. Severe and deadly medical conditions can easily be developed by diabetic patients which may eventually result in unexpected death. So, diabetics is not curable, but it can be controlled. And to control it should be detected as early as possible. Therefore, a method to detect the disease at an early stage would be really beneficial [4].

In this paper, an approach is suggested to predict the probability of developing diabetes with the help of basic medical information which are directly and highly responsible for the development of diabetes. In this approach, a mathematical equation have been developed using eight basic medical information of a person as the parameters. And to verify the equation's reliability, three machine learning algorithms (Logistic Regression, SVM, and KNN) have been used.

The rest of the paper is structured as follows. In section 2 and 3 describe literature review and proposed solution respectively. Section 4 contains detailed impression of the approach. Section 5 covers the result and comparison analysis. Lastly, section 6 and 7 draw the conclusion of this paper with some comments and ideas to improve this approach as future works.

2. LITERATURE REVIEW

2.1 Types of Diabetes

Though diabetes has a number of different types, it is mainly divided into three categories: type 1, type 2, and gestational diabetes.

2.1.1 Type 1 Diabetes (T1D)

Type 1 diabetes (T1D) named as insulin-dependent diabetes, is a severe situation in which the pancreas produces little or no insulin. The actual reason of T1D is unidentified [4]. Generally, the human immune system which normally fights against harmful bacteria or viruses kills the insulin-producing cells in the pancreas by mistake. This type of diabetes cannot be cured efficiently with oral medications alone and the patients need to take insulin therapy [5].

2.1.2 Type 2 Diabetes (T2D)

Type 2 diabetes (T2D) is the most widespread type of diabetes among all. Because, 90% of the diabetes are T2D. It is generally described with insulin resistance, where the body is unable to react with insulin. As, insulin is incapable to perform properly, blood sugar levels keep escalating and releasing even more insulin. For some people with type 2 diabetes, this can sooner or later exhaust the pancreas, causing in the body producing less and less insulin, resulting in even higher blood sugar levels (hyperglycemia) [7]. Older adults are very likely to have T2D, but is gradually happening to people of all ages. A lot of people with T2D have shown signs of prediabetes (impaired fasting glucose and/or impaired glucose tolerance) before meeting the criteria for type 2 diabetes. Lifestyle factors and genetics are main reasons of developing T2D. Some lifestyle factors are responsible for T2D, including physical inactivity, obesity, and lack of sleep, stress and urbanization [6].

2.1.3 Gestational Diabetes Mellitus (GDM)

Gestational diabetes mellitus (GDM) is the most severe and neglected threat to the mother and the unborn child. It occurs when a hormone produced by the placenta forces the body from not using insulin efficiently. It is fully curable, but requires careful medical observation throughout the pregnancy [7].

After a detailed research, it can be said that among these three types of diabetes, T2D is the most common as well as severe one and requires more attention than other two types.

2.2 Diabetes Risk Factors

As it is intended to predict the risks of T2D, so, only the risk factors of T2D has been discussed. Several factors combinedly affect the probability of developing T2D such as genes and lifestyle. Although it is impossible to change the risk factors such as age, ethnicity or family history but lifestyle risk factors like eating, physical activity, sleep time and weight can be changed. These lifestyle changes can help to control the chances of developing T2D.

- **Age:** The risk of T2D increases with age, especially after the age 45. That is probably because people incline to exercise less, lose muscle mass and gain weight as they are aging. But type 2 diabetes is also increasing intensely among children, teenagers and younger adults too [8].

- **BMI:** BMI is a numerical value that can be obtained by dividing a person’s weight in kilograms (kg) by his /her height in meters squared. Overweight is a BMI of 27.3 or more for women and 27.8 or more for men. Obesity is a BMI of 30 or more for either gender [9].Obesity has been found responsible for almost 55% of cases of T2D [10].

- **Blood Pressure (BP):** Having high blood pressure or hypertension appears to accelerate the risks of T2D. Hypertension has two types of stages and they are:

    - **Stage 1:** Systolic 130–139 mm Hg and diastolic 80–89 mm Hg

    - **Stage 2:** Systolic 140+ mm Hg and diastolic 90+ mm Hg [11]

  Whereas, 120/80 mm Hg is the normal blood pressure which should be maintained to avoid diabetes. A blood pressure (BP) of 130/80 mm Hg or less is the recommended therapeutic BP target for T2D patients. Because, diabetic patients can also develop hypertension and both of them can increase cardiovascular risk [12].

- **Blood Sugar:** High blood sugar (hyperglycemia) is another risk factor which contributes to the possibility of having diabetes. Blood sugar level below 200 mg/dl (randomly taken) and below 100 mg/dl (after fasting for 8 hours) is normal. But, when a person is diagnosed with blood sugar level equal to 200 mg/dl or more (randomly taken) and equal to 126 mg/dl or more (after fasting for 8 hours) have diabetes [13].

- **Sleep:** Sleep deprivation is a habitually ignored but it is a major risk factor for T2D. The link may seem unimaginable. Specifically, with constant sleep loss, fewer insulin is released in the body after eating. Meanwhile, human body secretes more stress hormones which makes it impossible to sleep and makes it tougher for insulin to do its job effectively. To sum up, a lot of glucose remains in the blood, which can raise the risk of developing T2D [14]. It is essential to get 7 to 9 hours of uninterrupted sleep regularly so that the body can work properly and lessen the risk of developing T2D and other health problems.

- **Physical Activity:** Less physical activity is responsible for 7% of the burden of T2D in the European Region. T2D was until recently seen as a disease of middle-aged and elderly people, but it is now increasingly seen in youngsters and children. Exercise improves blood sugar control in T2D, reduces cardiac risks and helps in to weight loss process. Regular exercise may be able to prevent or slow down the T2D development. Everyday exercise, or at least 5 times a week for 20-30 minutes, is recommended to reduce the insulin resistance [15].

2.3 Machine Learning Algorithms

A lot of Supervised Machine Learning algorithms are used in prediction system. So, ML algorithms has been used on the selected dataset. These algorithms will be discussed below.

2.3.1 Logistic Regression

Logistic regression is one of the most common ML algorithms, which comes under the Supervised Learning technique. It predicts the result of a categorical dependent variable. So, the result have to be either a discrete or a categorical value. It can be 0/1, true/False, Yes/No, etc. but instead of giving the precise value as 0 and 1, it gives the probabilistic values which is between 0 and 1. It is considered as one of the easiest Machine Learning algorithms that can be used for several classification problems such as cancer detection, Diabetes prediction, etc. [16] [17].

2.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) which is an algorithm of Supervised Machine Learning used for both regression and classification. But, for most of the cases, it is used in classification problems. The goal of this algorithm is to generate the best decision boundary that can separate n-dimensional space (where n is number of features) into classes so the new data points can be put easily in the accurate group in the future. This best decision boundary is known as a hyperplane. Then, classification is executed by finding the hyperplane that segregates the two classes properly. SVM selects the extreme points or vectors that help in generating the hyperplane. These extreme points or vectors are known as support vectors, and so the algorithm is named as Support Vector Machine [18] [19].

2.3.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is one of the easiest Supervised Machine Learning algorithms. It assumes the similarity between the new data point and available data points and place the new data point into the class that is most similar to the available classes. It saves all the available data and categorizes a new data point depending on the similarity. So, when new data arrives, it can be easily categorized into a perfectly suited class. As, this algorithm does not make any statement on primary data so it is known a non-parametric algorithm. It is also called as a lazy learner algorithm because it does not learn from the training set straightaway, rather it saves the dataset and when it gets new data, then it classifies that data into a class that is much similar to the new data [20].

2.4 Related Works

There is a lot of research on diabetes happening all around the world. Diabetes prediction is really tough but now-a-days with the help of healthcare monitoring data and machine learning algorithm prediction is less tough than before.

A model grounded on data mining methods for predicting type 2 diabetes was proposed. The model could categorize patients into either confirmed patients or suspected patients in 5 years from the first checkup time for T2D. Based on a series of preprocessing procedures, the model was consisted of double-level algorithms. In the initial level, the improved K-means algorithm was used to eliminate inaccurately clustered data and the improved dataset was used as input for next level. Then, the logistic regression algorithm was used to classify the remaining data. The 10-fold cross validation was used to improve the accuracy of the prediction model and to make a model adaptive to more than one dataset. The kappa statistic was used to judge the consistency of the model. It was showed that the model attained a 3.04% higher accuracy of estimation than those of other researchers [21].

Another study was about the emergence of some dependable initial detection systems and several healthcare associated systems from the medical and diagnosis data done by both the data mining and healthcare industry. The data mining models (predictive and descriptive) in the reviewed paper were identified. The tasks such as clustering, association rules, correlation analysis used for descriptive models and classification, regression and categorization used for predictive models, derived from the papers were reviewed. The methods broadly used for classification were statistical, discriminant analysis, decision tree, Markov based, swarm intelligence, KNN, genetic classifiers, artificial neural network, support vector, association rule, Bayesian classifier and logistic regression. As the conclusion, it was said that it was important to design a hybrid model which could resolve the mentioned issues and the future directions was to enhance the predictions using hybrid models [1].

This paper was concentrated upon the predictive analysis of diabetic cure using a data mining technique to discover patterns that identified the best mode of treatment for diabetes across dissimilar age clusters in Saudi Arabia. Oracle Data Miner (ODM), support vector machine algorithm was used. Dataset was analyzed into five age groups. Six types of treatment of diabetes were described. The pattern of the researchers of the work indicated that drug treatment was operative for both groups of patients but more operative for patients in the old age group [22].

The emerging optimization algorithms and machine learning algorithms were summarized in this paper. Three types of optimization algorithms and four types of machine learning algorithms were discussed. The four applications in the field of diseases diagnosis were considered and the important challenges in the deployment of disease were discussed [23].

Study on various prediction techniques and tools for machine learning in practice were discussed in this paper. How to get pattern from a large dataset and specific steps on how to apply machine learning to data were evaluated. Various algorithms, some tools and libraries were talked over. Some diseases were also considered and how to process its data while doing diagnosis and predictions using machine learning were showed [24].

3. PROPOSED METHODOLOGY

Diabetes is a common but a chronic disease with serious consequences such as long-term damage and disabling body parts. As, it disturbs many parts of the body so, if left untreated, can create severe health issues for example kidney failure, heart disease, blindness, stroke, and lower-limb amputation and so on [4]. It currently affects 200 million people and is the 5th cause of death throughout the world [25]. 1 in 2 (232 million) people with diabetes are untreated. It has affected 4.2 million deaths till now. Diabetes caused minimum 760 billion dollars USD as health expenses in 2019 which now makes it an expensive disease too [7].

All this facts and severe consequences of diabetes motivated us to create a system which can help people to predict the possibilities of having diabetes with the basic medical information. Because, if a person knows that he has a chance of being a diabetic patient at an early stage, he may be able to control the severity of the disease. Also, he will be warned in advance and let himself diagnosed thoroughly.

3.1 The Workflow of the Proposed Method

To understand the research topic in-depth, several papers were studied. After a comprehensive study, some methods and ideologies were nearly similar to the selected research topic. But there are criticisms on those methods. Therefore, a new and different method is proposed to identify the risk percentage of diabetes more accurately. Steps taken to propose the method is provided below (figure 01):

- At first, the dataset was finalized which has been used for the method. Then, a deep research has been done on the attributes of the dataset and their relevance with the risk of diabetes.
- Then, an attempt to find the co-relations between different attributes of the selected dataset has been taken by precisely analyzing the properties, relational dependencies between two or more single attributes.
- From the background analysis, it has been identified that age, BMI, high blood pressure, high sugar intake habit of patients have a proportional relationship with the risk of diabetes while low physical activity and regular sleep deprivation of the patients have an inversely proportional relationship with the risks of diabetes .
- The focus of research has been placed to develop a mathematical equation that could satisfy all the proportional and inversely proportional relationship between each and every identified attribute to predict the risk of diabetes.
- While developing the mathematical equation, it was realized that some attributes may affect the risk percentage more than the others. To be precise, a person with a higher value of BMI than the average is a potential diabetic patient. So, BMI attribute alone increases the risk here. That is why, the percentage of risk each

attribute possess has been considered individually. Because, these attributes could affect the total percentage drastically.

- Lastly, to check the accuracy and productivity of the derived equation, the attributes of derived mathematical equation has been replaced with appropriate data from the dataset.

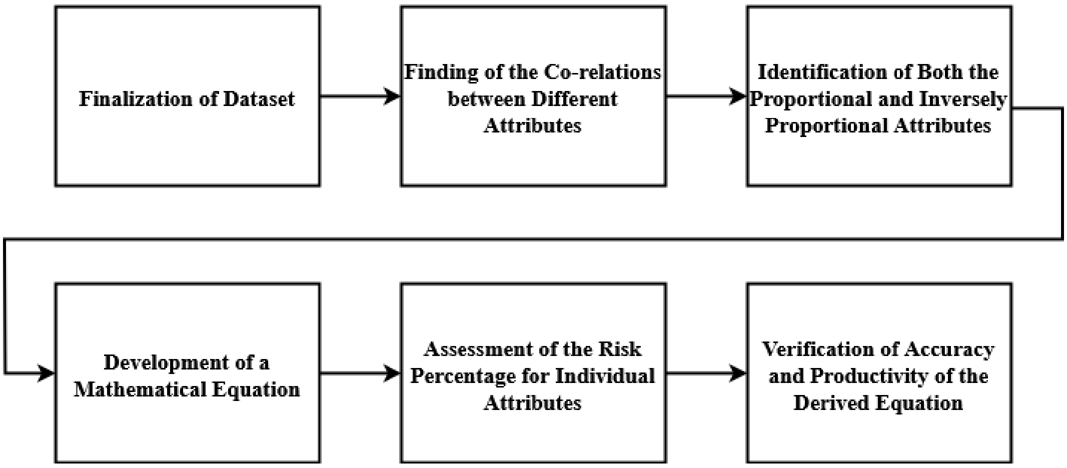
Finalization of Dataset
Finding of the Co-relations between Different Attributes
Identification of Both the Proportional and Inversely Proportional Attributes
Development of a Mathematical Equation
Assessment of the Risk Percentage for Individual Attributes
Verification of Accuracy and Productivity of the Derived Equation

Figure 1: Flowchart for the Proposed Method of Diabetes Prediction

4. IMPLEMENTATION

4.1 Derivation of the Mathematical Equation

The target is to generate the highest probability of having diabetes based on the factors (attributes) of the dataset. The factors are age, Body Mass Index (BMI), Blood Pressure (Systolic and Diastolic), Blood Sugar (Fasting and random), Exercise Time and Sleeping Time. The **R**isk **o**f **D**iabetes has been denoted as **RoD.**

Among all the factors, some of them are proportional with the risk. It means if the value of these factors increases then the probability of risk also increases. Again, some of the factors are inversely proportional with the risk. . It means if the value of these factors decreases then the probability of risk also increases. So, the relationship between the individual proportional and inversely proportional factors with the risk (RoD) is shown below. Here, SBP, DBP, BSL, ET and ST stand for Systolic Blood Pressure. Diastolic Blood Pressure, Exercise Time and Sleeping Time respectively.

Now, for all the eight factors [26] [27]:

$$\boldsymbol{RoD} \propto \boldsymbol{Age} \tag{1}$$

$$\boldsymbol{RoD} \propto \boldsymbol{BMI} \tag{2}$$

$$\boldsymbol{RoD} \propto \boldsymbol{SBP} \tag{3}$$

$$\boldsymbol{RoD} \propto \boldsymbol{DBP} \tag{4}$$

$$\boldsymbol{RoD} \propto \boldsymbol{BSL(random)} \tag{5}$$

$$\boldsymbol{RoD} \propto \boldsymbol{BSL(fasting)} \tag{6}$$

$$\boldsymbol{RoD} \propto \frac{\boldsymbol{1}}{\boldsymbol{ET}} \tag{7}$$

$$\boldsymbol{RoD} \propto \frac{\boldsymbol{1}}{\boldsymbol{ST}} \tag{8}$$

After combining the proportional and inversely proportional relationships with RoD, the mathematical equation is derived. And it is given below.

$$\boldsymbol{RoD} \propto \frac{\boldsymbol{Age} * \boldsymbol{BMI} * \boldsymbol{SBP} * \boldsymbol{DBP}}{\boldsymbol{ET}} * \frac{\boldsymbol{SBL(random)} * \boldsymbol{SBL(fasting)}}{\boldsymbol{ST}} \tag{9}$$

In equation 9, to replace the proportional sign with equal sign, a constant K is used and finally the desired mathematical equation is derived.

$$\boldsymbol{RoD} = \boldsymbol{K} * \frac{\boldsymbol{Age} * \boldsymbol{BMI} * \boldsymbol{SBP} * \boldsymbol{DBP}}{\boldsymbol{ET}} * \frac{\boldsymbol{SBL(random)} * \boldsymbol{SBL(fasting)}}{\boldsymbol{ST}} \tag{10}$$

4.2 Computation for the Constant K

In the previous section 4.1, an equation with both proportional and inversely proportional factors with respect to the risk of diabetes has been created. The value of individual factors was collected from authentic sources that are responsible for producing highest risk of diabetes [15] [28] [29] [30] [31] [32]. The risk values for each individual factors are shown in the table 1.

Table 1: Highest and Lowest Risk Value of Each Factors

Attributes	Lowest Factor	Highest Factor	SI Units
Age	0.5 years	47 years	s
BMI	18 kg/m^2	35 kg/m^2	kg/m^2
Systolic Blood Pressure	120 mmHg	140 mm Hg	pa
Diastolic Blood Pressure	80 mmHg	90 mm Hg	pa
Blood Sugar (Fasting)	5.5 mmole/L	7 mmole/L	m/L
Blood Sugar (Random)	7.8 mmole/L	11 mmole/L	m/L
Sleep Time	4 hours		s
Exercise Time	20 minutes		s

For these highest risk value of each individual factor, the probability of having diabetes is considered as 100% and then the value of constant K has been generated. All the Factors need to be converted into SI units.

So, the attributes of the equation 10 are replaced with highest risk value for each factors. And for the highest risk value of each individual factor, let the probability of having diabetes be 100%. So, the equation transforms as:

$$\mathbf{100 \;= K * \frac{47yr * 35kg/m^2 * 140mm\,Hg}{20min} * \frac{90mm\,Hg * 11mmole/L * 7mmole/L}{4hr}} \qquad \textbf{(11)}$$

Transferring all the values into SI units:

$$\mathbf{100 = K * 1.48 * 10^9 s * 35kg/m^2 * \frac{18.665 * 10^3 pa * 11.998 * 10^3 pa}{1200\,s} * \frac{11 * 10^{-3}\,mole/L * 7 * 10^{-3}\,mole/L}{14400\,s}} \qquad \textbf{(12)}$$

After solving the above equation, a value of constant K is attained. And the value is

$$\mathbf{K = 1.931 * 10^{-6}\;\; sm^2L^2kg^{-1}mole^{-2}pa^{-2}} \qquad \textbf{(13)}$$

4.3 The Deployment of Machine Learning Algorithms

To verify the integrity of the mathematical equation, three machine learning algorithms have been applied on the dataset. The machine learning algorithms are Logistic Regression, SVM, and KNN. Before applying the algorithms on a dataset, the dataset has to go through some series of steps and these steps are combinedly known as ML methods.

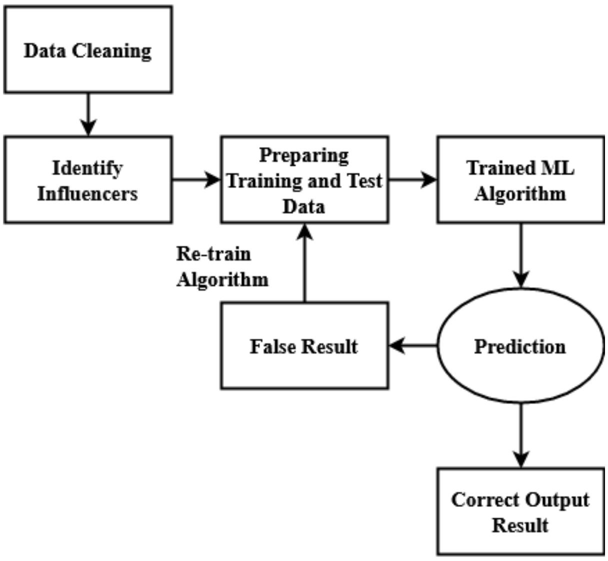

Figure 2: Flowchart for Machine Learning Algorithms

At first, the dataset has been cleaned by removing the null and infinite values. From the cleaned dataset, the highly weighted influencers (attributes) has been identified. In the next step, the dataset has been prepared for training and testing process. Then, the dataset has been trained by the ML algorithms for prediction. If the false result is generated from the prediction then the algorithms needs to be retrained by preparing the dataset for training and testing process. And if the true result is generated then the desired outcome will be achieved (figure 2). For the SVM model, kernel = linear and for KNN model, K = 2 have been considered.

5. RESULT ANALYSIS

In this chapter, the experiment's result has been analyzed. For the prediction, a mathematical equation has been formed. To get the predictive risk percentage by solving the equation, a dataset which was suitable for the research was selected.

5.1 Dataset

All the information was collected from different web sources where more than 1000 people submitted their all health indexes in real-time (current status of Diabetes). Only a few participant's data has been considered in this paper.

5.2 Prediction of the Risk of Diabetes

5.2.1 The Prediction Process

The selected dataset contains some necessary factors that have strong relations with the risk of diabetes than the others. And the derived equation has been created based on the factors that generates an initial risk as an output for a given input containing all the considered factors. And the output is the basic risk probability without considering the individual factors that are very strongly associated with and can change the risk probability significantly. So, checking one by one all the factors that are strongly associated with the risk and changing the risk probability is necessary. Following these concepts, the final output risk probability has been generated.

5.2.2 Calculation for Risk Percentage

For a given input into the equation, it gives an output of general risk probability. After that, some cases where the factors are most responsible for resulting the risk percentage of diabetes need to be considered. To do this step, some conditions needs to be followed and they are given below:

- When the risk is less than 25%, it means one has less probability of having diabetes. Here, sleeping time that appears as a less risk factor of having diabetes comparatively than the other attributes has been considered. In this case, if the diabetes risk of a person is less than 25% and sleeping time is less than 4.5 hours, then the initial risk value has been increased to 5%.

- When the risk is less than 30%, the factor exercise time has been considered. If the exercise time is less than 20 minutes, then again, the initial risk value has been increased to 5%.

- Now, one of the most important cases where the risk value is less than 50% and systolic blood pressure is less than 120mm Hg. It was confirmed that a systolic blood pressure (SBP) is associated with type 2 diabetes and a 1mm Hg increase is associated with a 1%–4% increase in type 2 diabetes risk. So, during this described case, the initial risk has been multiplied with 3 for each 1mm Hg increase.

- When the risk is less than 90%, then both age and BMI factors have been considered. BMI according to three age groups has been considered and they are: less than 18 years, greater than 18 years and less than 45 years, greater than 45 years. At each age group, the risk has been converted into 100% and then calculated for five circumstances.
    - For the **Age less than 18 years age** group, if

- BMI is less than 18.5, then initial risk increases 7.0%
- BMI is greater than 18.5 and less than 25, then initial risk increases 10.5%
- BMI is greater than 25 and less than 30, then initial risk increases 24.5%
- BMI is greater than 30 and less than 35, then initial risk increases 38.5%
- BMI is greater than 35, then initial risk increases 52.5%
- For the **Age greater than 18 years and less than or equal to 45 years** group, if
  - BMI is less than 18.5, then initial risk increases 7.0%
  - BMI is greater than 18.5 and less than 25, then initial risk increases 8.4%
  - BMI is greater than 25 and less than 30, then initial risk increases 21%
  - BMI is greater than 30 and less than 35, then initial risk increases 24.5%
  - BMI is greater than 35, then initial risk increases 42%
- For **Age greater than 45 years** group, if
  - BMI is less than 18.5, then initial risk increases 3.5%
  - BMI is greater than 18.5 and less than 25, then initial risk increases 5.6%
  - BMI is greater than 25 and less than 30, then initial risk increases 10.5%
  - BMI is greater than 30 and less than 35, then initial risk increases 17.5%
  - BMI is greater than 35, then risk initial increases 24.5%
- When the risk is less than 95%, it means someone has highest risk of having diabetes. And in diabetes prediction, blood sugar level is one of the main factors. Because of this factor alone, someone can have highest risk. So, the blood sugar level as the highest risk factor has been considered here.
  - If someone's fasting blood sugar level is between 5.5 to 6.9 mmol/l, 5.5mmol/l from one's fasting blood sugar level has been subtracted and multiplied it with 24 and then added the result with the initial risk. And when one's fasting blood sugar level less than 7, the risk has been considered as 100%.
  - If someone's random blood sugar level is between 7.8 to 11 mmol/l, 5.5 mmol/l from one's fasting blood sugar has been subtracted and multiplied it with 12 and added the result with the initial risk. And when one's random blood sugar is less than 11, the risk has been considered as 100%.

5.3 Performance Comparison among Mathematical Equation and Machine Learning Algorithms

As, three machine learning algorithms have been applied on the dataset to verify the reliability of the mathematical equation, some comparison results have been achieved. For comparison purposes, accuracy, precision, recall, F1-score and AUC-ROC curve have been used. These comparison results are shown in form of some bar charts.

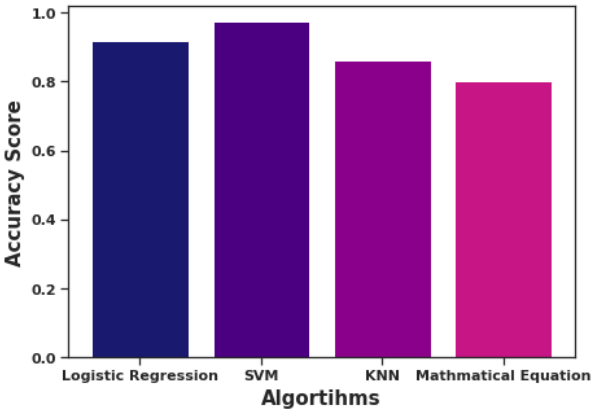

Figure 3: Bar Chart of Accuracy

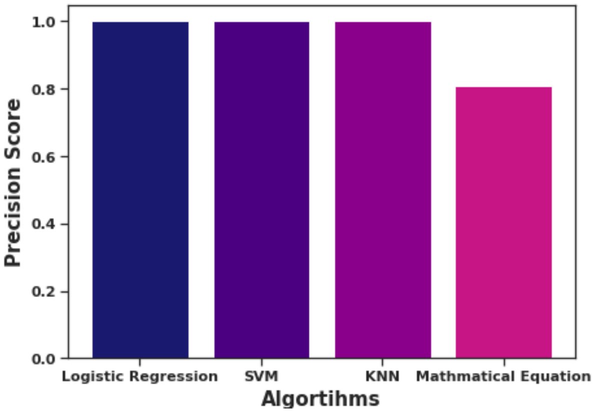

Figure 4: Bar Chart of Precision

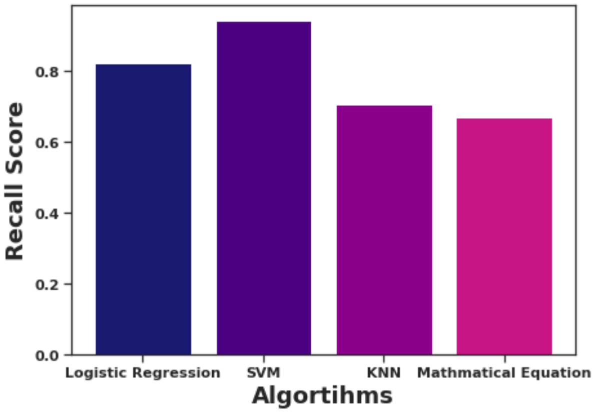

Figure 5: Bar Chart of Recall

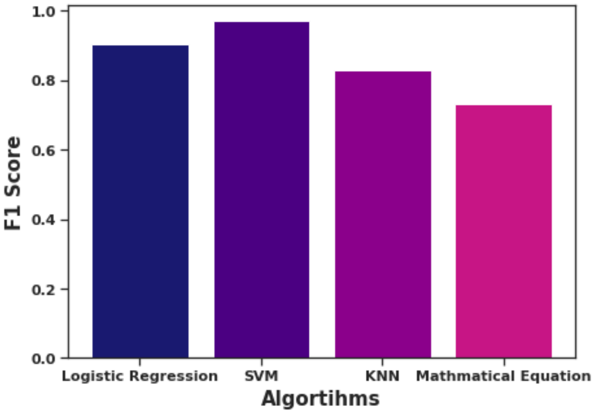

Figure 6: Bar Chart of F1 Score

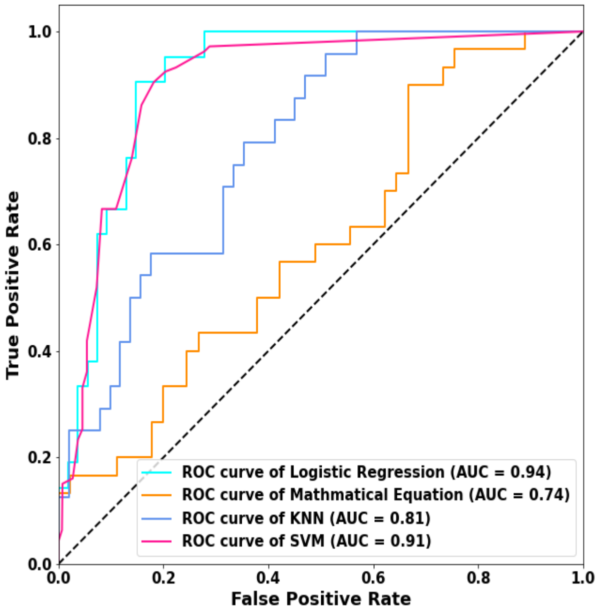

Figure 7: AUC-ROC Curve

From these figures, it can be clearly seen that, the accuracy measures of these ML algorithms and the derived mathematical equation is quite close. For a better understanding, the percentage of accuracy measures are illustrated in a tabular format in table 2.

Table 2: Performance Comparisons among Equation and ML Algorithms

Accuracy Measures	Logistic Regression	SVM	KNN	Mathematical Equation
Accuracy	86%	91%	83%	80%
Precision	97%	97%	97%	78%
Recall	81%	91%	67%	63%
F1 Score	85%	97%	81%	76%
AUC	0.94	0.91	0.81	0.74

6. CONCLUSION

As a long-lasting disease, diabetes can make its patients suffer for the rest of his life. After developing diabetes once, it cannot be cured. But, it can be controlled. The earlier it is diagnosed, the easier it will be to control this disease. And, this system can play an important role here. This system will predict the probability of a person's having diabetes and if the probability in near 60%, the person will know that and he can change the lifestyle habits which affects the possibility and then he will probably able to avoid diabetes. But, if unfortunately, the probability is over 60%, the person can go a doctor and be able to control it before it gets too late. It will be a great help to the people who may not have the financial ability to take a diabetes diagnosis whenever he/she wants. Lastly, the hope is to reduce the number of patients who are suffering from diabetes and also the increasing death rate solely because of diabetes.

7. FUTURE SCOPE

Based on a patient's percentage of risk of having diabetes, the system can be developed so that it can provide a set of suggestions using reinforcement learning algorithm for the diabetic patients to control it. Then, it will further generate suggestions based on the patient's negative or positive outcome by following the system's previous generated suggestions. The patient's clinical histories will be stored time to time so that the system can generate appropriate suggestions to gain the best possible outcome.

8. REFERENCES

[1] N. Jothi, W. Husain, et al., "Data mining in healthcare–a review," *Procedia Computer Science*, vol. 72, pp. 306–313, 2015.

[2] Rosenstock, J., Park, G., Zimmerman, J., & Glargine, U. I. (2000). Basal insulin glargine (HOE 901) versus NPH insulin in patients with type 1 diabetes on multiple daily insulin regimens. US Insulin Glargine (HOE 901) Type 1 Diabetes Investigator Group. *Diabetes care*, 23(8), 1137-1142.

[3] Lal, B. S. (2016). Diabetes: Causes, Symptoms And Treatments. book: *Public Health Environment and Social Issues in India, Edition*, 1, 55-67.

[4] Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes care*, 28(S37), S5-S10.

[5] Ndisang, J. F., Vannacci, A., & Rastogi, S. (2017). Insulin resistance, type 1 and type 2 diabetes, and related complications 2017.

[6] Dariush Mozaffarian, Aruna Kamineni, Mercedes Carnethon, Luc Djoussé, Kenneth J. Mukamal, David Siscovick." Lifestyle Risk Factors and New-Onset Diabetes Mellitus in Older Adults: The Cardiovascular Health Study." *Archives of Internal Medicine*, vol.169, issue.8, Pages.798.

[7] Atlas, D. (2015). International diabetes federation. IDF Diabetes Atlas, 7th edn. *Brussels, Belgium: International Diabetes Federation.*

[8] Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. *The Lancet*, 389(10085), 2239-2251

[9] Ray, D. E., Matchett, S. C., Baker, K., Wasser, T., & Young, M. J. (2005). The effect of body mass index on patient outcomes in a medical ICU. Chest, 127(6), 2125-2131.

[10] Centers for Disease Control and Prevention (CDC. (2004). Prevalence of overweight and obesity among adults with diagnosed diabetes--United States, 1988-1994 and 1999-2002. *MMWR. Morbidity and mortality weekly report*, 53(45), 1066.

[11] Cheung, B. M., & Li, C. (2012). Diabetes and hypertension: is there a common metabolic pathway?. Current atherosclerosis reports, 14(2), 160-166.

[12] Lipman, M. L., & Schiffrin, E. L. (2012). What is the ideal blood pressure goal for patients with diabetes mellitus and nephropathy?. Current cardiology reports, 14(6), 651-659.

[13] O'Sullivan, J. B., & Mahan, C. M. (1965). Blood Sugar Levels, Glycosuria, and Body Weight Related to Development of Diabetes Mellitus: The Oxford Epidemiologic Study 17 Years Later. JAMA, 194(6), 587-592.

[14] Touma, C., & Pannain, S. (2011). Does lack of sleep cause diabetes. *Cleve Clin J Med*, 78(8), 549-58.

[15] S. R. Colberg, R. J. Sigal, J. E. Yardley, M. C. Riddell, D. W. Dunstan, P. C. Dempsey, E. S. Horton, K.

Castorino, and D. F. Tate, “Physical activity/exercise and diabetes: a position statement of the american diabetes association,” *Diabetes care*, vol. 39, no. 11, pp. 2065–2079, 2016.

[16] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12), 1373-1379.

[17] Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica: Biochemia medica,* 24(1), 12-18.

[18] Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, 24(12), 1565-1567.

[19] Hua, S., & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular biology*, 308(2), 397-407.

[20] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), 1883.

[21] H.Wu, S. Yang, Z. Huang, J. He, and X.Wang, “Type 2 diabetes mellitus prediction model based on data mining,” *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.

[22] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, “Application of data mining: Diabetes health care in young and old patients,” *Journal of King Saud University-Computer and Information Sciences*, vol. 25, no. 2, pp. 127–136, 2013.

[23] K. Chui, W. Alhalabi, S. Pang, P. Pablos, R. Liu, and M. Zhao, “Disease diagnosis in smart healthcare: Innovation, technologies and applications,” *Sustainability*, vol. 9, no. 12, p. 2309, 2017.

[24] B. Nithya and V. Ilango, “Predictive analytics in health care using machine learning tools and techniques,” in 2017 *International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 492–499, IEEE, 2017.

[25] Wang, L., Kong, L., Wu, F., Bai, Y., & Burton, R. (2005). Preventing chronic diseases in China. *The lancet*, 366(9499), 1821-1824.

[26] “The link between a lack of sleep and type 2 diabetes.”https://www.sleepfoundation.org/articles/link-between-lack-sleep-and-type-2-diabetes. Accessed: 2019-12-26.

[27] “Diabetes.” https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444. Accessed: 2020-01-23.

[28] Crilly, P. (2020). Managing hypertension: the role of diet and exercise. *The Pharmaceutical Journal*, 304(7934).

[29] Spiegel, K., Knutson, K., Leproult, R., Tasali, E., & Cauter, E. V. (2005). Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes. *Journal of applied physiology*, 99(5), 2008-2019.

[30] Alva, M. L., Hoerger, T. J., Zhang, P., & Gregg, E. W. (2017). Identifying risk for type 2 diabetes in different age cohorts: does one size fit all?. *BMJ Open Diabetes Research and Care*, 5(1).

[31] Narayan, K. V., Boyle, J. P., Thompson, T. J., Gregg, E. W., & Williamson, D. F. (2007). Effect of BMI on lifetime risk for diabetes in the $_{\text{US}}$. *Diabetes care*, 30(6), 1562-1566.

[32] ” Diabetes - Diagnosis and treatment - Mayo Clinic.“ https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451. Accessed: 2020-11-20.

[33] Alehegn, M., Joshi, R. R., & Mulay, P. Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach.

[34] Duke, D. L., Thorpe, C., Mahmoud, M., & Zirie, M. (2008, March). Intelligent Diabetes Assistant: Using machine learning to help manage diabetes. *In 2008 IEEE/ACS International Conference on Computer Systems and Applications* (pp. 913-914). IEEE.

[35] Kumar, P. S., & Pranavi, S. (2017, December). Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. *In 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS)* (pp. 508-513). IEEE.

[36] Mirshahvalad, R., & Zanjani, N. A. (2017, September). Diabetes prediction using ensemble perceptron algorithm. *In 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 190-194). IEEE.

[37] Priyadarshini, R., Dash, N., & Mishra, R. (2014, February). “A Novel approach to predict diabetes mellitus using modified Extreme learning machine.” *In 2014 International Conference on Electronics and Communication Systems (ICECS)* (pp. 1-5). IEEE.

[38] M. Adam, E. Y. Ng, S. L. Oh, M. L. Heng, Y. Hagiwara, J. H. Tan, J. W. Tong, and U. R.Acharya, “Automated characterization of diabetic foot *using nonlinear features extracted from thermograms,” Infrared Physics & Technology*,vol. 89, pp. 325–337, 2018.

[39] M. R. Devi and J. M. Shyla, “Analysis of various data mining techniques to predict diabetes mellitus,” *International Journal of Applied Engineering Research*, vol. 11, no. 1, pp. 727–730, 2016.

[40] S. Bashir, U. Qamar, and F. H. Khan, “Intellihealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework,” *Journal of biomedical informatics*, vol. 59, pp. 185–200, 2016.

[41] S. R. Colberg, R. J. Sigal, J. E. Yardley, M. C. Riddell, D. W. Dunstan, P. C. Dempsey, E. S. Horton, K. Castorino, and D. F. Tate, “Physical activity/exercise and diabetes: a position statement of the american diabetes association,” *Diabetes care*, vol. 39, no. 11, pp. 2065–2079, 2016.