Assessment of Risk of Type 2 Diabetes Mellitus with Stress as a Risk Factor using Classification Algorithms

Rohini Patil, Kamal Shah

Abstract — Rising prevalence of type 2 diabetes mellitus is a vital health concern today, not only in India but across the world. Several factors including dietary habits, genetics, lack of physical exercise and stress are known to affect the risk of type 2 diabetes. Although awareness has increased to some extent, many people with diabetes have limited knowledge about the risk factors before the diagnosis of disease. For chronic disease prevention there is a necessity to find out such risk factors and manage them appropriately. Statistical techniques can be employed to understand the risk of type 2 diabetes in different age group of people. The objective of the research was to evaluate relationship among stress and type 2 diabetes in people with different age groups by a statistical tool. The proposed method uses three machine learning classifiers namely Support Vector Machine (SVM), Logistic Regression and Random Forest (RF) to detect type 2 diabetes at an early stage. To develop an adaptive model the preprocessing step has been applied. The accuracy of predicting diabetes using SVM, Random Forest and Logistic Regression was 80.17%, 79.37%, 78.67% respectively. The results suggest that as compared to Random Forest and Logistic Regression, SVM is better in predicting occurrence and progress of type 2 diabetes mellitus with stress as a risk factor.

Keywords—Diabetes Mellitus,Support Vector Machine,Random Forest,Logistic Regression.

I. INTRODUCTION

Diabetes is often called as silent killer.Today, diabetes represents a significant health burden among families and in turn on the country. The incidence and prevalence of both are on rise worldwide and in India. Type 2 diabetes mellitus is a chronic condition which occurs because of decreased synthesis of insulin by the pancreas or ineffective utilization of insulin. Diabetes mellitus is broadly divided into three key types; type 1 (insulin dependent) diabetes, type 2 diabetes and gestational diabetes.

Type 1 diabetes is because of the deficiency of insulin, there is increase in blood glucose level. Because of inability of pancreas to produce insulin it has to be given from outside hence patients need lifelong injections of insulin.

Type 2 diabetes occurs due to ineffective utilization of insulin [1]. Type 2 diabetes mellitus is more prevalence than insulin dependent diabetes mellitus. Type 2 diabetes can be prevented by adapting healthy life-style.

Gestational diabetes occurs during pregnancy [2]. Women with gestational diabetes are at more risk of developing type-2 diabetes. Gestational diabetes is diagnosed by regular checking of blood glucose during pregnancy.

In pre-diabetes, blood glucose levels are elevated but not as high as diabetes mellitus [3]. Pre-diabetes is a risk factor for diabetes so precaution need to be taken from early stage.

The two well-known international association's i.e. World Health Organization (WHO) and International Diabetes Federation (IDF) provide important information and data related to epidemiology. According to the report of the WHO published in 2016, diabetes is the seventh leading cause of death [4].

According to the IDF, India is the second largest country with an estimated 73 million adults between the age group of 20-79 years living with the disease. Current burden of diabetes in India is 49% of world burden and this is expected to reach 151 million by 2045, posing a major health challenge for the country [5]. Diabetes is associated with several complications including those related to eye, nerves, arteries, kidney and heart [6]. Stress is one of the important risk factors or trigger for development of type 2 diabetes mellitus. Presence of long term stress can lead to rise in blood glucose level. Stress can either be physical or mental or both [7, 8, 9].

Healthcare is one of the areas where predictive analytics can be of significant value. Machine learning algorithms can build a model from data used for prediction, decision making or solving some task. Predictive analytics uses statistical, analytical queries or machine learning methods for predicting potential outcomes [10]. Past and current data can be used for predicting future activity, behavior and trends. This research is carried out to predict the risk of type 2 diabetes mellitus based on presence of stress factors and create a simple diagnostic application which may also be useful for improving the health.

II. REVIEW OF LITERATURE

Several factors including age, genetics and glucose level are considered to be important in diabetes. These factors may differ among different types of diabetes. For predicting risk of diabetes among people with different age groups based on life style of a person data mining approach is used. Decision trees model is used for predictive analysis. The results have suggested higher prevalence of type 2 diabetes among people with age above 45 years. They may also have risk factors included hypertension, lipid level, body mass index (BMI) and the history of myocardial infarction [3].

Manuscript published on November 30, 2019.
* Correspondence Author

Rohini Patil*, Research Scholar Thakur college of Engineering and Technology Mumbai, India Assistant Professor, Terna Engineering College, Navi Mumbai. rohinipatil@ternaengg.ac.in

Dr.Kamal Shah, Professor& Dean (R & D) Thakur college of Engineering and Technology Mumbai. kamal.shah@thakureducation.org

*Retrieval Number: D9509118419/2019©BEIESP
DOI:10.35940/ijrte.D9509.118419
Journal Website: www.ijrte.org*

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

Stress is also a significant risk parameter in the pathogenesis of type 2 diabetes mellitus by direct or indirect mechanisms. Stress can have impact on the immune system or endocrine system. It may also affect the lifestyle and eating pattern, all resulting in higher risk of developing type 2 diabetes. [7, 8, 9]

Researchers have used traditional and hybrid approach as a predictive model and showed hybrid model to be more accurate than traditional model [11].

In another model, diabetic patients were classified in three age groups i.e. 18 to35, 36 to55 and more than 55 years, using adaboost and decision tree model. They used J48 decision tree model [12].

For improvement of accuracy, researcher developed a model with two level algorithm i.e. unsupervised k-means and logistic regression applied on 14 attributes for predicting type 2 diabetes risk [13].

The Chi-squared test and used classification and regression tree (CART) model has been shown to provide accuracy of 75% for the prediction of diabetic based on daily lifestyle. Blood Pressure was identified as a major factor for the development of diabetes along with junk food, sleep, family history, rice intake and the physical activities performed [14].

For prediction of risk of diabetes investigators have developed a web application by using classification algorithm. Different models like Artificial Neural Network, Decision Tree, Logistic Regression, Naïve Bayes and Random Forest algorithm have been used for prediction analysis. Out of these, Random Forest Algorithm (RFA) was selected for the prediction of diabetes risk as it gives best performance. In this approach, input variables are randomly selected and used for creating the application [15].

Harris et al. shows for type 2 diabetes how perceived stress is a strong risk factor. They shows Direct and indirect effect of

Stress on diabetes through risk factors, like BMI, hypertension, and physical activity. This study is only female centric [16].

Martinez et al. shows how stress affect diabetes. They found an increased risk for T2D in people exposed to stressful conditions, traumatic events, mental health problems, depression. [17].

III. PROPOSED METHODOLOGY

Figure 1 below shows conceptual framework used in our study. Demographic and other data were collected both from diabetic and non-diabetic people. For building a prediction model data analysis was performed by applying preprocessing step with a suitable machine learning model which improves the accuracy.

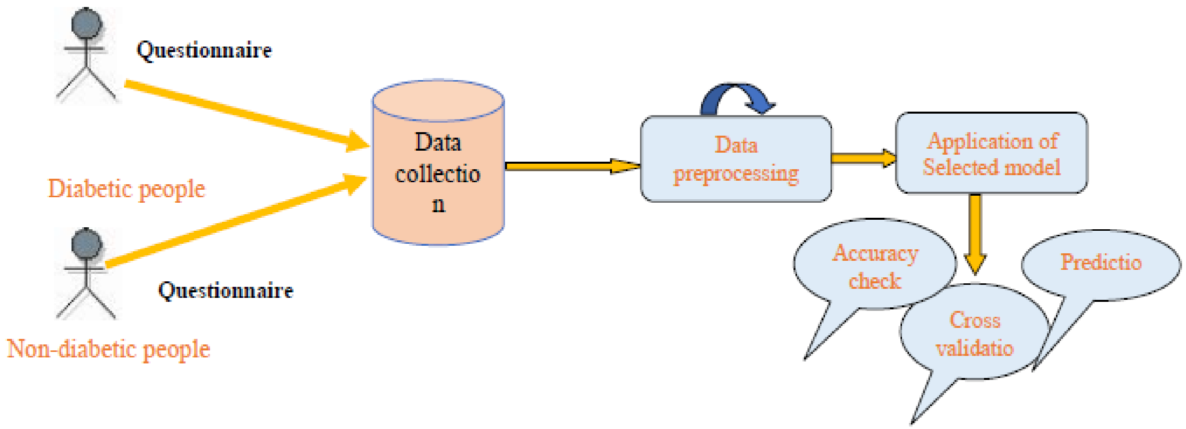

Figure3.1.Proposed Conceptual Framework

A.Data Set Description:

The dataset was developed by collecting information from people through paper based and web-based questionnaires. The collected information was entered in the excel sheet and thoroughly cleaned for examining homogeneity. A dataset of 375 people comprising of 7 attributes was prepared.
The questions were related to stress related factors such as workload, salary, deadlines, travel, repetitive work, unplanned work, career, powerlessness and job security and satisfaction for prediction of type 2 diabetes. Table 1.describes different attribute used in dataset.

Table1: Description of attributes used in dataset

No.	Attribute	Description	Possible value
1	Gender	Male or female	0: Male 1: Female
2	Age	Age of a person	Numeric
3	Weight	Weight (kg)	Numeric
4	Height	Height (cm)	Numeric
5	BMI	Body Mass Index (Kg/m2)	Numeric
6	Anxiety	Presence of anxiety	0: no 1:yes

Retrieval Number: D9509118419/2019©BEIESP
DOI:10.35940/ijrte.D9509.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

7	Stress	Workload Poor Salary Deadline Travel Repetitive Work Unplanned Work Career Powerlessness Job Security No Satisfaction	1: Very low 2: Low 3: Average 4: High 5: Very high
8	Class	CLASS 0: Normal group 1: Diabetes risk group	

B. Data preprocessing

In this cross-sectional study (single visit), data were collected from 390 people including both diabetic and non-diabetic people. Out of 390 people, 375 responses were eligible for analysis. Other 15 responses were not considered because of the incomplete entries or some missing information. Outliers from dataset were also removed. Body mass index (BMI) and height required transformation. Height was collected in feet and then converted to centimeters. By dividing weight (kg) by height (meters) squared BMI was calculated.

C.Brief Description of Classification Algorithms:

A. Logistic regression model

The aim of the machine learning classification algorithm is, to develop a machine learning model which can map input data to a given class. For analyzing a dataset, a statistical approach with one or more independent class variables can determine an outcome using logit function. Logistic regression algorithm is called as logit model used for predicting and classifying medical problem, for automatic disease diagnosis. Logistic Regression is binary classification problem. The outcome is measured in the form of dichotomous variable i.e. a person is non-diabetic or diabetic. It uses linear equation with independent predictors. [18, 19, 20].

Consider x_1, x_2,… ,x_n. are an input variables.

Let z be the probability of occurrence of an event and

1-z be a probability of non-occurrence of an event.

Logit model is shown in the following equation,

$$Z= \frac{e^y}{(1+e^y)} \qquad (1)$$

$$\log\left[\frac{Z}{1-Z}\right] =Y \qquad (2$$

$$\log\left[\frac{Z}{1-Z}\right] =logit(z)= \beta_{0+}\ \beta_1 x_{1+}\beta_2 x_{2+}\ldots +\beta_n x_n \qquad (3)$$

Where β_0 represents intercept and regression coefficient are represented as $\beta_{1,}$ β_2… β_n

B. Support Vector Machine (SVM)

SVM is supervised machine learning discriminative classification techniques. For regression and classification this method have an application in. Given with two-class training sample, to find the best highest-margin separating hyper plane between the two classes is shown below in Fig. 2. It comprises of two steps: identification of the optimal hyper plane and map the objects to the specified boundaries. For better generalization hyper plane should be away from the data points belonging to the other class. The points closest to

the hyper plane margin are the support vectors [21, 22].

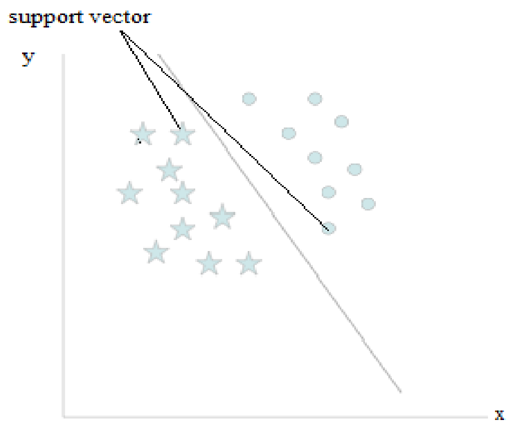

Figure 2: Support vectors representation over the hyper line

C. Random Forest :

Random Forest is one of the supervised machine learning algorithm, used for classification as well as regression. Random forest is bagging technique to generate random features. Based on attributes to classify an object, each tree does a classification it means the tree “votes” for that class. The steps are shown below

a. Load the dataset containing of “m” features.

b. Randomly select ‘n’ feature from m features for generating random samples, where n<<m.

c. Using the best split calculate node ‘d’. Split the node into sub-nodes using random samples.

d. Repeat the above steps, to find ‘n’ number of trees.

e. Calculate the total number of votes of each tree. The highest voted class is the random forest. [22]

IV. RESULT AND DISCUSSION

The performance of the classification algorithm was evaluated through the various measures such as accuracy, precision, recall and F-measure mentioned in Table 2. The research paper focuses on the three classification techniques i.e. SVM, Logistic Regression and Random forest. Table 3 shows results of the classification technique. Using Python toolkit, we analyzed and evaluated the model with following classifiers .

a. Support vector machine: The accuracy of this technique when applied on diabetic dataset was 80.17%.

b. Random forest classifier: The accuracy of random forest classifier was 79.37%.

c. Logistic Regression: The accuracy was 78.63%.

A. *K-fold cross validation*

To reduce over fitting cross -validation is used. We applied 10- fold cross validation. Training and testing of model was done 10 times. From the initial partitioned 10 sub-samples, 9 were used for training. Each subsample was retained as validation data.

B. *Performance Metrics*

To assess the performances of models, TP is considered as true positive (correctly predicted number of instances), FP as false positive (incorrectly predicted no of instances), TN as true negative (the no of instances are negative but predicted correctly) and FN as false negative (the no of instances are negative but predicted incorrectly). From the outcome of accuracy, some substantial performance measures are as shown in Table 2.

Table 2: Performance Metrics

Performance Measures	Definition	Formula
Accuracy(A_c)	Determines the accuracy of an classifier in predicting instances.	A_c=(TP+TN) /(TP+FP+TN+FN)
Precision(P_r)	Positive predicted value	P_r=TP/(TP+FP)
Recall (R)	Sensitivity	R=TP/(TP+FN)
F-measure	Weighted mean of precision and recall.	2*(P_r *R)/(P_r +R)
ROC	Receiver Operating curve	

Table 3 Comparative Performance analysis of Different Classifiers on Various Measures.

Classifier	Accuracy (%)	Precision	Recall	F-Measure
Support Vector Machine	80.17%	0.782	0.801	0.784
Logistic Regression	78.63%	0.832	0.840	0.833
Random Forest	79.37%	0.743	0.783	0.744

In our study, SVM showed the maximum accuracy (Table 3). Based on this, it can be concluded that SVM classifier predict the risk of diabetes better as compared to other classifiers (Figure 3).

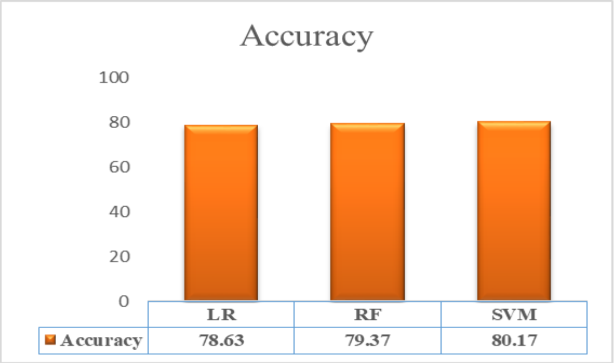

Figure 3: Comparison of classification algorithm - Accuracy

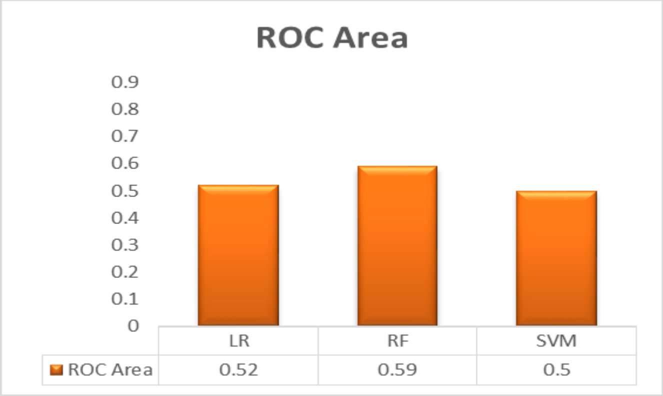

Figure 4: ROC Area of Classification Algorithm

Figure-4 represents ROC area of all classification algorithms. Statistical comparison of all classifiers are plotted in Figure 5.

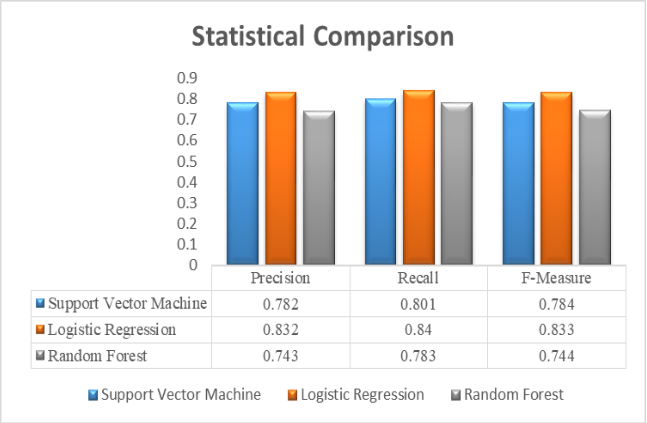

Figure 5: Statistical comparison of Classification algorithms

Retrieval Number: D9509118419/2019©BEIESP
DOI:10.35940/ijrte.D9509.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. LIMITATIONS

Small number of responses and limited number of risk factors for prediction modeling are the limitations of our study. These findings need to be evaluated by using larger data sets of diabetes and non-diabetes people with multiple prediction modeling techniques. Considering the above listed limitations readers should carefully extrapolate findings of our study

VI. CONCLUSION

Diabetes is a chronic and progressive metabolic disease. In order to become better quality of life of people with diabetes, it is important to detect and avoid the complications of diabetes at the early stage through predictive analysis by applying the classification techniques. Gender, age, body mass index, anxiety along with the stress factors such as workload, poor salary, deadlines, travel, repetitive work, unplanned work, career, powerlessness and job security and satisfaction may influence the development of type 2 diabetes. Stress is identified as an important risk factor for development of type 2 diabetes. In this work, use of SVM provided accuracy of 80.17% for the prediction of diabetes. The results suggest that as compared to Random Forest and Logistic Regression, SVM is better in predicting development of type 2 diabetes mellitus with stress as a risk factor. Further research can be done to evaluate effect of gender, age and other risk factors on the risk of diabetes.

ACKNOWLEDGMENT

Authors of this study wish to thanks Dr.Anant Patil for his contribution in explaining the medical terminologies and relevance of different risk factors of diabetes.

REFERENCES

1. Sanakal R, Jayakumari T. Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. Int J Comput Trends Technol. 2014;11(2):94–8.
2. Lakshmi KR, Kumar SP. Utilization of data mining techniques for prediction of diabetes disease survivability. Int J Sci Eng Res. 2013;4(6):933–40.
3. Repalli P. Prediction on diabetes using data mining approach. Stillwater: Oklahoma State University; 2011.
4. http://www.who.int/en/news-room/fact-sheets/detail/diabetes accessed on 18th nov.2018
5. https://www.idf.org/our-network/regions-members/south-east-asia/.../94-india.html assessed on 18th Nov.2018
6. Motka R, et al. Diabetes mellitus forecast using different data mining techniques. In: Computer and communication technology (ICCCT), IEEE, 4th international conference. New York: IEEE; 2013.
7. Dinesh Luthra. A Link Between Stress And Diabetes. WebmedCentral PSYCHIATRY 2010;1(10):WMC001095:1-6
8. Shona J. Kelly and Mubarak Ismail. Stress and Type 2 Diabetes: A Review of How Stress Contributes to the Development of Type 2 Diabetes. Annu. Rev. Public Health 2015. 36:441–62
9. Gemma Falco, Piero Stanley Pirro, Elena Castellano1, Maura Anfossi, Giorgio Borretta and Laura Gianotti. Journal of Neurology and Psychology. April 2015 Vol.:3, Issue:1.1-7.
10. Brown DE, et al. Predictive analytics. Washington: IEEE Computer Society; 2015
11. Jayanthi N. , B. Babul V. Rao S.N.Survey on clinical prediction models for diabetes prediction. J Big Data 2017; 4:26 DOI 10.1186/s40537-017-0082-7
12. Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib, Karim KeshavjeecPerformance Analysis of Data Mining Classification Techniques to Predict Diabetes Procedia Computer Science 82 (2016) 115 – 121
13. Wu H,YangS,HuangZ,HeJ,WangX.Type 2 diabetes mellitus prediction model based on data mining.Informatics in medicine unlocked 2018;10:100-107
14. Aysh anand, Divya shakti Prediction of diabetes based on personal lifestyle indicators,In :1st international conference on next generation computing technology:IEEE;2015
15. NongyaoNai-aruna,,RungruttikarnMoungmaia,Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science 69 (2015) 132 – 142
16. Harris ML,Oldmeadow C, Hure A, Luu J, Loxton D, Attia J. Stress increases the risk of type 2 diabetes onset in women: A 12-year longitudinal study using causal modelling. PLoS ONE 12(2):e0172126.doi:10.1371/journal. pone.0172126
17. Martinez A, Sanchez W, Benitez R, Gonzalez Y, Mejia M, Otiz J. A job stress predictive model evaluation through classifier's algorithms. IEEE Latin America Transactions, VOL. 16, NO. 1, JAN. 2018
18. Tom Mitchell, Machine Learning. First Edition, McGraw- Hill, 1997
19. Osisanwo F. Akinsola J. ,Awodele O., Hinmikaiye J.,Olakanmi O. ,Akinjobi J. :Supervised Machine Learning Algorithms: Classification and Comparison International Journal of Computer Trends and Technology (IJCTT) 2017;vol 48 :128-138
20. Saikat Dutt,Amit Kumar Das , Machine Learning ,Pearson Education India.
21. Sisodia Sisodia D. Sisodia D.: Prediction of Diabetes using Classification Algorithms: Procedia Computer Science 132 (2018) 1578–1585
22. Sneha, N. Gangil, Tarun :Analysis of diabetes mellitus for early prediction using optimal features selection, Journal of Big Data 2019;vol 6. ://doi.org/10.1186/s40537-019-0175-6

AUTHORS PROFILE

Rohini Patil BE (Computer Science and Engineering), ME (Computer Engineering) and pursuing PhD in Information technology. She has teaching experience of over 15 years. She has authored 14 publications. Which are published in national and international journals. Her research work including machine learning and distributed systems. She is a member of ISTE.

Dr Kamal Shah (BE Electrical-REC, ME EXTC) has a vast teaching experience and more than 10 years of research experience. She has completed her Ph.D. in engineering in 2010 in the field of image processing with the major focus on vector quantization. She has 35 research papers to her credit and currently pursuing post-doctoral research in the field of Quantum computing from BARC, Mumbai. She has developed her interest in the field of Block chain technology from 2017 onwards and developed corporate training material which covers training modules, virtual laboratories and certification. Currently she is working actively in this field with her team to develop various industry ready solutions. Under her mentorship 18 companies are incubated incubation center funded by DST in the domain of Manufacturing, IOT-Block chain and clean technology