Diabetes Prediction using Logistic Regression and Random Forest Algorithm: A Comparative Study

Anuja Bokhare
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University)
Pune, India
anuja.bokhare@gmail.com

Vandan Raj N.
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University)
Pune, India
vnr2041060@sicsr.ac.in

2023 International Conference for Advancement in Technology (ICONAT) | 978-1-6654-7517-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICONAT57137.2023.10080370

***Abstract*—The objective of this research is to predict if the person has diabetes or not based on the certain diagnostic medical measurements. Few constraints were considered while picking up the data set, data collected for this research belongs to females whose age is at least 21 from Pima Indian heritage. The attributes are segregated as the input variables for the model and the output attribute is outcome. The goal of this project is to propose a prediction model for diabetic diagnosis using machine learning classification algorithms, and evaluate models using different metrics. In current study Logistic regression and random forest algorithm is used to implement and evaluate the model. The considered dataset has all patients are females at least 21 years old of Pima Indian Heritage. The comparison between Logistic regression and random forest algorithm is highlighted through accuracy calculation from both the algorithms.**

Keywords—Diabetes, Machine learning, Random Forest, Logistic Regression.

I. INTRODUCTION

Diabetes is one such disease in which our blood glucose, or the blood sugar levels will be very high. Glucose will be generated in the body by the food we intake, and Insulin is the hormone that will help glucose to get into our body cells to provide them energy. Although diabetes is heard very commonly in our era, it is a matter of surprise to know that the very first written record about this will date back to 1500 BC, and the record has the symptoms documented in it. Diabetes will cause several other side effects in humans like risk of dying from heart failure, nerves weakness, risk of having gingivitis which will in turn leads to tooth loss, blindness, amputations, increases risk of having Alzheimer's disease, vaginal infections can occur in women, improper or irregular mensuration, ovulational cycles this will in turn lead towards the risk of having hyperglycaemia and diabetic ketoacidosis in girls. Knowing all these facts will stress upon the importance of diagnosing the diabetes at the early stage itself and start taking proper measures to handle that. Having a healthy lifestyle will definitely help in facing this disease. Hence, the research paper is focused on coming up with ML model which will predict if a person has diabetes by considered different parameters related to health

II. PREVIOUS STUDY

Artificial Since over 30 million people in India are suffering from this disease, statistics show that it has become more than common in today's life style. Diabetes is caused when the pancreas in the body is not able to produce insulin at the quantities required by the body to function normally or when the cells and tissues in the body fail to utilize the insulin produced in the body resulting in high sugar content in the blood. Diabetes mellitus exists in three forms, they are as follows:

Diabetes Mellitus Type-1 is occurred when the pancreas generates less insulin than the requirement of the body to function normally. This is situation is also called as "insulin-subordinate diabetes mellitus" (IDDM). People need insulin dosages to maintain the levels based on their insulin secretions.

Diabetes Mellitus Type-2is a condition where there will be no insulin in the body because the body cells will be resisting it. This is more critical than the Type 1 diabetes. Otherwise called as “adult starting diabetes" or “non-insulin subordinate diabetes mellitus" (NIDDM) this is found in people with high BMI or those who lead an inactive lifestyle.

Gestational diabetes is one of the type that is identified mostly during their pregnancy in women.

The critical factors that play a vital role in diagnosis of DM are Body mass index (BMI) and age. Sometimes even the blood tests will not be able to diagnose the disease accurately. Nonlinear SVM classifier is used to predict the diagnosis of the diabetes by taking BMI, age, blood glucose concentration as the inputs and history of DM in patients or their relatives (no diabetes, predisposition to diabetes and diabetes) factors of Colombia patients with different backgrounds. They tested the model with 10-fold cross validation method by dividing the data into 80:20 ratios. This model gave an accuracy of 95.36 [1].

Since Type2 is way more critical than the type1, Many Machine learning algorithms were implemented to predict the occurrence of type 2 diabetes on the data collected via question are from people belonging to the different backgrounds related to their health, age and other specific things. They used different models like Logistic regression, K- Nearest neighbor classifier, Naïve bayes classifiers, SVM algorithm, decision tree method, random forest classification. After constructing the confusion matrix and doing comparative analysis, Random forest was able to predict with highest accuracy of 94.10% [2]. Although A1C is now one of the recommended ways to diagnose diabetes, using this for diagnosis and prognosis of diabetes is still uncertain. Therefore, in this paper [3], they tried to exactly predict the performance of the A1C against single and repeated glucose measurements which will help in the diagnosis of the diabetes which is prevalent and for predicting the pre-diabetes on 12,485 participants in the Atherosclerosis Risk in Communities. They found that the results were quite good and dependable. Different classification algorithms like K-Means, Artificial Neural Networks (ANN) and Random forest were used to understand the factors that are associated

Authorized licensed use limited to: Stanford University. Downloaded on May 04,2023 at 22:20:28 UTC from IEEE Xplore. Restrictions apply.

with diabetes and they found highest co-relation was between the BMI and glucose level from Apriori algorithm. Confusion matrix was again drawn to do check of accuracy comparatively, and the highest turned out to be 79.5% for ANN [4]. Although there are many ML algorithms that are being used in the prediction of diabetes and classification of the types/diagnosis involved, ANN is used more widely than any other algorithm and CNN is used for Deep learning [5].

Gestational diabetes is seen in the pregnancy women, there was no efficient way to predict the occurrence of this in the first trimester of pregnancy in China. Therefore, the authors considered taking up this challenge, gathered data, and extracted 17 variables through feature selection methods for early GDM prediction that will help the women to take necessary steps during pregnancy. They build neural network model and logistic model consisting of the seven variables [6]. This helped to predict the chances of GD in Chinese population. Different ML models have been mentioned in many papers, which can be used in prediction of the diabetes and about how the features are selected. The models the accuracy was different case by basis, in this paper also, the author have taken the data from PIMA Indians Diabetes (PID) and then told about how will have to do data pre-processing which consists of filling the missing values with means, outliers' rejection and identification, standardization of the data and feature selection of data. They have spoken about the approach that can be taken using PCA, Correlation Based and ICA Based Feature selection approach along with different ML Models like Adaboost, Random Forest, Naïve bayes and KNN [7]. The dataset is taken from the University of California, Irvine (UCI) database. Probabilistic-based naïve Bayes (NB), function-based multilayer perceptron (MLP), decision tree-based random forests (RF) models were used in this paper to predict the diagnosis of diabetes. To test the same, author has used 10-fold cross validation and did comparative study. It is observed that the NB gave better results and the processed data must be fed to get results more accurately [8]. In 2012 diabetes caused death of 1.5 million and high blood glucose caused the death of another 2.2 million. If not treated in time, diabetes can damage the heart, blood vessels, eyes, kidneys, and nerves which makes the early detection of diabetes a crucial matter [12]. In this model, the authors have used only Neural networks and then build a model which will tell if a person is having diabetes or not based on the input factors considered in building the model like pregnancies, Plasma glucose concentration, Diastolic BP, Tri Fold Thick, Serum Ins, BMI, Diabetes Pedigree function, Age [10]. They have been more focused on reducing the average error function of the Neural Network. The model built was consisting of front feed network with one input layer, which is having eight inputs to that, three hidden layers, and one output layer. With this they have achieved 0.01 average error functions with accuracy of 87.3. As an output of the model it says whether the person is healthy or sick [11]. Most of the authors have done comparative study to do the comparative study of how efficient the model is, similarly in this, they have applied Decision tree, KNN, Random forest, Naïve Bayes, Logistic regression, SVM in Weka tool and then used cross validation method to validate the data. They have also used NN in Jupiter notebook and built the neural model consisting of two hidden layers with accuracy of 88.6 % [11].

It is known fact that the outliers in the data set will be causing problems if not handled during data processing that results in model making wrong predictions/classifications. In this paper the author has spoken about the way by which outliers can be handled. Instead of removing them, can use Interquartile range algorithm after outliers are detected in the data set, this algorithm will measure the dispersion of the data and then rank the data set into four quartiles. Once quartiles are formed, Synthetic Minority Oversampling Technique (SMOTE) is used to create artificial instances by using KNN algorithm around the outlier data points. This will result in small clusters in the outlier region and makes data smoother and efficient for handling classification problems [9]. In this paper the author has focused on using principal component analysis (PCA), to improve the accuracy of the K-means clustering algorithm and logistic regression model. It is seen that by using PCA, the accuracy is increased by 1.98%. PCA will transform the set of features considered for the building the algorithm initially by eliminating the problem of correlation. This will make sure that the data is free of unwanted features and increases the accuracy, speed and reliability of the model. Post this K-means will help us in handling the outliers and they are eliminated from the data, finally the transformed data is being fed to logistic regression in this proposed model [14].

III. METHODOLOGY

Figure 1 describes the proposed model. It shows that the model consists of both Logistic regression and Random forest model to predict if the person will be diabetic or not based on the input attributes. The data consists of the many medical attributes that will help in prediction. Therefore, the cleaned and processed data is fed into the supervised ML models and have done comparative study. Also in this paper, the co-relation between the attributes is initially identified, which will help in capturing the information about how they are related to each other in a healthy person and in a diabetic person. By knowing this, the risk factors can be highlighted

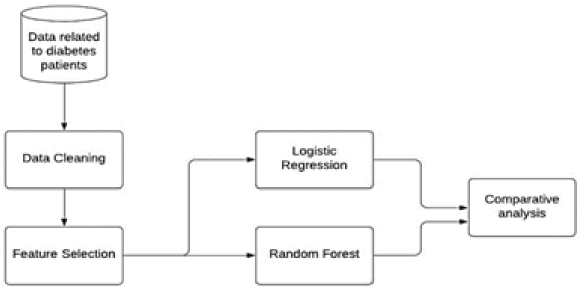

Fig. 1. Diabetic Prediction Model

IV. DATASET AND EXPERIMENT DISCUSSION

This section describes about the dataset consider during the study.

Dataset : The PIMA dataset taken consists of records belonging to 768 people, and it is composed of diabetic and non-diabetic patients.

Data contains the below mentioned attributes:

1. Pregnancies: The frequency of woman's pregnancy, this will help us in understanding if the woman can get gestational diabetes

Authorized licensed use limited to: Stanford University. Downloaded on May 04,2023 at 22:20:28 UTC from IEEE Xplore. Restrictions apply.

2. Glucose: Plasma glucose concentration is taken in a two hours gap to understand an oral glucose tolerance

3. BloodPressure: Diastolic blood pressure (mm/Hg)

4. SkinThickness: Triceps skin fold thickness (mm)

5. Insulin: 2-Hour serum insulin (mu U/ml)

6. BMI: Body mass index which is calculated basically using the formula: (weight in kg/(height in m)^2)

7. DiabetesPedigreeFunction: The probability of a person getting diabetes is more if the relatives are having the history, hence this needs to be considered as one of the features.

8. Age: Age (years)

9. Outcome: 1 if diabetes, 0 if no diabetes

A. *Experiment Steps*

1) Data cleaning and Feature Selection:

Understanding the data is the first process to build any model and then clean the data if necessary. The sample of the data considered for the study shown in Table 1. In the Outcome column, 1 represents the person is diabetic and 0 represents the person is non-diabetic.

2) Checking for Null values and Handling Missing Data:

The data does not contain null values, but some are imputed as 0 which is indicated here as 0.head() function is used to understand the data and found that some features contain 0, and it does not make sense in the data, hence this indicates missing value. Missing values are taken care of by replacing 0 values by NaN. Table 2 shows the number of rows that an attribute having entry as zero:

TABLE I. NULL VALUE COUNT IN DATA

Attribute	Number of entries
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
Age	0
Pregnancies	111
DiabetesPedigreeFunction	0

Therefore, there are 768 records in total, and It is taken care that the data is not containing null values or any duplicate values.

B. *Correlation matrix*

Diabetic patients form 34.9% of the whole data set taken for study while non-diabetic patients represent 65.1%. Class imbalance is one of the supervised learning problems, which can happen if one, fail to understand the spread of the data and its attributes. It happens when there is significant difference between the minority and the majority classes which in our case is diabetes person's data and healthy person's data respectively or mostly in classes with binary values, as observed in our dataset. Since we are having huge number of diabetes person's data, we will have to make sure that this will not lead to poor performance of classification model.

This is demonstrated using co-relation matrix and heat map as shown in Figure 2. It shows correlation coefficients between sets of the attributes we have considered for the model. Each attributes (X) correlated with each of the other attributes in the Table 2 i.e. (Y).

TABLE II. DESCRIPTION OF DATA

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

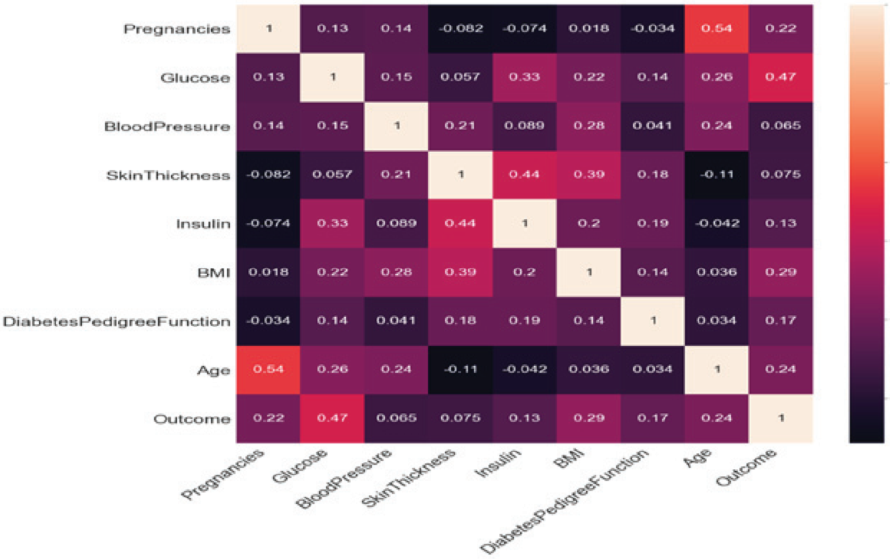

Fig. 2. Correlation Matrix

Authorized licensed use limited to: Stanford University. Downloaded on May 04,2023 at 22:20:28 UTC from IEEE Xplore. Restrictions apply.

C. *Histogram*

Since all the attributes considered are numerical in nature and not categorical, histogram for all the attributes will help in knowing the spread of the data.

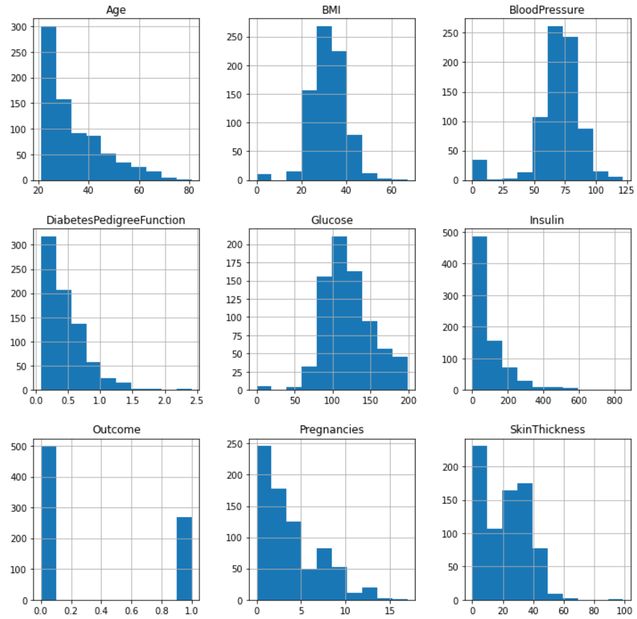

Fig. 3. Histogram of all the attributes considered for the model

From Figure 3. can be said that the data is normally distributed and skewed. From this, we can see that there are no outliers as well in the data. Still, it will not tell anything about the distribution of the attributes in case of a healthy person and diabetic person.

V. Result Analysis

This section gives highlights of result analysis of the study.

TABLE III. Statistical Inferences Related to Features

Attribute	Median value for the healthy person	Median value for the diabetic person
Insulin target value	102.5	169.5
Glucose	107.0	140.0
Skinthickness	27.0	32.0
Blood Pressure	70.0	74.5
BMI	30.1	34.3

A. *Logistic Regression*

The model is implemented using Logistic regression i.e. used as the classifier, It will help in classifying the data based on the input parameters, and then also based on that it will predict for test data, shown in Figure 4, if the patient is diabetic or not. The cost function value will be between 0 and 1 for this.

B. *Random Forest*

Random Forest supervised learning algorithm chosen to do the prediction if the person is diabetic or not. The number of tree structures present in the data is directly proportional to the accuracy of the model. Each internal node within the tree corresponds to an attribute and every leaf node represents class label. The confusion matrix related to the random forest model is shown in Figure 5

In this paper, two machine learning classification models have been implemented and Table 4 shown the comparison between them i.e., Logistic Regression and Random Forest algorithm in terms of performance metrics. Logistic regression has done much better job in predicting if a person is suffering from diabetes or not and also in classifying the data.

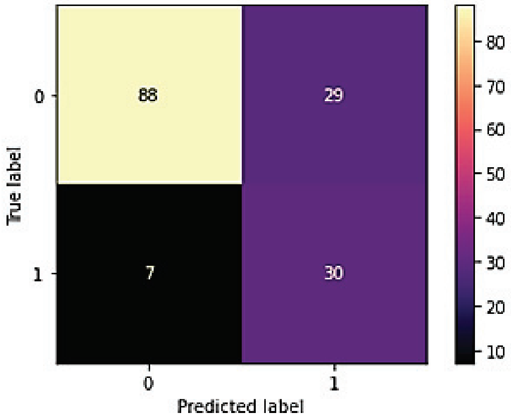

Fig. 4. Confusion Matrix from Logistic Regression

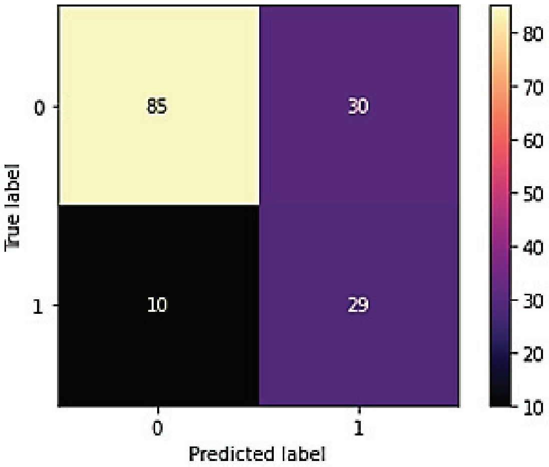

Fig. 5. Confusion Matrix from Random Forest

TABLE IV. Performance Comparison of Algorithms

	Logistic Regression	Random Forest
Accuracy	76.66 %	74 %
Precision	0.50	0.49
Recall	0.81	0.74
F-Measure	0.618	0.589

Authorized licensed use limited to: Stanford University. Downloaded on May 04,2023 at 22:20:28 UTC from IEEE Xplore. Restrictions apply.

VI. Discussion

Since there is having, many attributes considered in the model, in this section the correlation of few selected attributes using scatter plots is shown.

From Figure 6. can draw a conclusion that the Healthy persons are concentrate with an age less than or equal to 30 and glucose levels lesser than or equal to 120.

Since gender attribute is not considered and research data is having female gender only, it is important to understand the co-relation between the age and number of pregnancies in women.

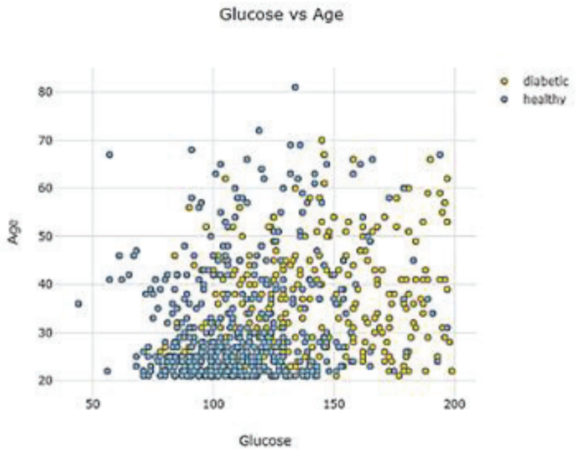

Fig. 6. Scatter plot for Glucose v/s Age

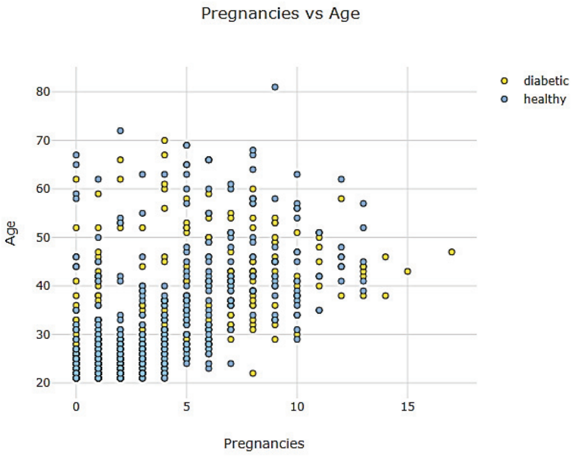

Fig. 7. Scatter plot for Pregnancies v/s Age

Figure 7. Shows that there is higher probability of getting diabetes if the age is greater than 25 OR if a woman is having pregnancies more than 4. Hence, if any one of the criteria is true, then the person should be extra careful during the pregnancy period.

Logistic regression has done much better job in predicting if a person is suffering from diabetes or not and also in classifying the data.

VII. Conclusion

Current study aims to make sure that diabetes can be predicted as early and as accurately as possible by using different Machine learning algorithms. Logistic Regression and Random Forest models are used to do the same. Various data pre-processing methods are implemented during the study for outlier detection, feature relation, visualization of the correlation during the study. The outcome of the study highlights that the logistic regression algorithms helps to classify 76.66% the diabetic or health cases. The random forest algorithm performs 74% accuracy in the detection of the same in the current study. Also, study describes how attributes are affecting each other and the range of the attributes with respect to a healthy person and a diabetic person.

VIII. Future Scope

Since the data did not have outliers in current data, model did not have any step of handling outliers, but as a future scope, outliers will be treated using Interquartile algorithm and then create clusters of outlier data by using Synthetic Minority Oversampling Technique.

References

[1] Viloria, Amelec, et al. "Diabetes diagnostic prediction using vector support machines." *Procedia Computer Science* 170 (2020): 376-381.

[2] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." *Procedia Computer Science* 167 (2020): 706-716.

[3] Selvin, Elizabeth, et al. "Performance of A1C for the classification and prediction of diabetes." *Diabetes care* 34.1 (2011): 84-89.

[4] Alam, Talha Mahboob, et al. "A model for early prediction of diabetes." *Informatics in Medicine Unlocked* 16 (2019): 100204.

[5] Larabi-Marie-Sainte, Souad, et al. "Current techniques for diabetes prediction: review and case study." *Applied Sciences* 9.21 (2019): 4604.

[6] Wu, Yan-Ting, et al. "Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning." *The Journal of Clinical Endocrinology & Metabolism* 106.3 (2021): e1191-e1205.

[7] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." *IEEE Access* 8 (2020): 76516-76531.

[8] Singh, D. A. A. G., E. Jebamalar Leavline, and B. Shanawaz Baig. "Diabetes prediction using medical data." *Journal of Computational Intelligence in Bioinformatics* 10.1 (2017): 1-8.

[9] Nnamoko, Nonso, and Ioannis Korkontzelos. "Efficient treatment of outliers and class imbalance for diabetes prediction." *Artificial Intelligence in Medicine* 104 (2020): 101815.

[10] El Jerjawi, Nesreen Samer, and Samy S. Abu-Naser. "Diabetes prediction using artificial neural network." *International Journal of Advanced Science and Technology* 121 (2018).

[11] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." *ICT Express* 7.4 (2021): 432-439.

[12] Harz, H. H., Rafi, A. O., Hijazi, M. O., & Abu-Naser, S. S. (2020). Artificial Neural Network for Predicting Diabetes Using JNN. International Journal of Academic Engineering Research (IJAER), 4(10).

[13] Srivastava, Suyash, et al. "Prediction of diabetes using artificial neural network approach." *Engineering Vibration, Communication and Information Processing*. Springer, Singapore, 2019. 679-687.

[14] Zhu, Changsheng, Christian Uwa Idemudia, and Wenfang Feng. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques." *Informatics in Medicine Unlocked* 17 (2019): 100179.

Authorized licensed use limited to: Stanford University. Downloaded on May 04,2023 at 22:20:28 UTC from IEEE Xplore. Restrictions apply.