EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Enhancing Diabetes Prediction with Data Preprocessing and various Machine Learning Algorithms

Gudluri Saranya[1], Sagar Dhanraj Pande[2, *]

[1,2, *]School of Computer Science & Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.

Abstract

Diabetes mellitus, usually called diabetes, is a serious public health issue that is spreading like an epidemic around the world. It is a condition that results in elevated glucose levels in the blood. India is often referred to as the 'Diabetes Capital of the World', due to the country's 17% share of the global diabetes population. It is estimated that 77 million Indians over the age of 18 have diabetes (i.e., everyone in eleven) and there are also an estimated 25 million pre-diabetics. One of the solutions to control diabetes growth is to detect it at an early stage which can lead to improved treatment. So, in this project, we are using a few machine learning algorithms like SVM, Decision Tree Classifier, Random Forest, KNN, Linear regression, Logistic regression, Naive Bayes to effectively predict the diabetes. Pima Indians Diabetes Database has been used in this project. According to the experimental findings, Random Forest produced an accuracy of 91.10% which is higher among the different algorithms used.

Keywords: Accuracy, Diabetes, Machine Learning, Naive Bayes, Random Forest

Received on 11 December 2023, accepted on 01 March 2024, published on 08 March 2024

doi: 10.4108/eetiot.5348

*Corresponding author. Email: sagarpande30@gmail.com

1. Introduction

Diabetes is a medical illness caused due to elevated glucose levels in the body's bloodstream. Although there is no cure for it, we can detect it and get treatment at an early stage to prevent serious complications. In this project various machine learning algorithms are used to build a model which helps us in detecting diabetes. There are mainly three types of diabetes.

Type-1:
This diabetes occurs when the immune system malfunctions by destroying insulin producing pancreatic cells. Mostly, children and young adults are the ones that are diagnosed with it; however, it can occur at any age [1]. Genes and potential triggers in the environment, such as viruses, can develop type 1 diabetes.

Type-2:
It is a medical disorder where our body doesn't utilize insulin in an effective manner [2]. Due to this the glucose levels will be elevated in the blood which leads to diabetes. Type 2 diabetes can also be caused by being overweight and obese, lack of physical activity, and having insulin resistance.

Gestational diabetes:
Hormones released during pregnancy limit the action of insulin, which results in gestational diabetes. This type of diabetes only happens during pregnancy. Those who already have pre-diabetes and have a history of diabetes in their families are frequently affected by it [2].

Diabetes is often characterized by a few key signs and symptoms, such as:

- Feeling hungrier than usual.
- Thirstier than normal
- Losing weight.
- Urinary frequency
- Hazy eyesight
- Feeling exhausted
- Non-healing wounds

- Hands or feet that feel numb or tingly.

The major contribution of this proposed approach is as follows:

1. At first complete data cleaning is done which includes fixing the data set by removing the data which is duplicated or incorrectly formatted and then SMOTE Pre-processing is performed for handling imbalanced data set i.e., production of quality training datasets by oversampling the minority class. If oversampling is not done, it may lead to biased results.

2. This paper made a novel contribution by proposing a hybrid approach for scaling the features. It is used to standardize the feature data range. This paper concentrated on using a combination of two feature scaling techniques namely StandardScaler() function and RobustScaler() function. This helps in overcoming the individual limitations of both functions and in reducing the influence of outliers on training the model.

3. The research significantly advances the accuracy of diabetes prediction models by utilizing machine learning techniques, ensemble methods, and intensive hyper-parameter tuning. For every algorithm and ensemble model, different hyper-parameter tuning techniques like grid search and random search are explored. By incorporating all these techniques together, the accuracy of the model has increased significantly.

The remaining paper is divided as: Section 2 furnishes the Literature survey. Section 3 outlines the methodology implemented for the research. It provides information regarding data preprocessing techniques like SMOTE and StandardScaler(). It also provides descriptions of various algorithms used. Section 4 covers the experiment results and comparison with other research studies. Section 5 wraps up the paper by pointing out the future scope in this area.

2. Literature Survey

Table 1. Analysis of various papers discussed about the prediction of diabetes

Year of Pubication	Purpose	Methodology Used (Approach)	Results	Gaps/ Limitations
2022 [3]	To create a DNN-based decision support system for diabetes diagnosis.	Artificial Intelligence based Decision Support Systems (DSS) are being implemented with a Bi-LSTM system to make decisions more accurate and efficient.	The testing results for this suggested model were encouraging with an accuracy of 93.07%.	The proposed method used the embedding model rather than a pre-trained model.
2022 [4]	To predict diabetes using Ensemble Techniques	Used ensemble techniques by Combining algorithms like Random Forest, Gradient Boosting, Adaptive Boosting, and XGboost by applying Ensemble Methods.	When compared to other models, XG Boost had the best performance in the test results.	Some data preprocessing techniques like feature scaling are missing.
2022 [5]	A Case Study of Bandipora District.	Six distinct algorithms - RF, SVM, DT, Gradient Boosted Classifier, Logistic Regression (LR) and Multilayer Perceptron (MLP) were applied.	Of all the algorithms used, Random Forest has achieved the best accuracy rate.	Integration of various algorithms is not examined.

2021 [6]	For an effective Diabetes Prediction using ANN	The ABP-SCGNN approach was used in the paper to build an improved ANN model.	According to the experimental findings, 93% accuracy was achieved on the testing set.	It could have been possible to produce a more thorough analysis by incorporating a wider variety of algorithms.
2020 [7]	Research was made on Diabetes Prediction based on ML.	This paper applied algorithms such as LightGBM, Naive Bayes classifier and SVM.	SVM has the highest rate of accuracy of nearly 96.54%.	Various other machine learning algorithms like Random Forest, Naive Bayes theorem could have been employed. The processes required to preprocess the data before analysis are not sufficiently discussed in detail.
2020 [8]	To classify Diabetes Using PPG Waveform	In this study, through the analysis of photoplethysmogram (PPG) wave forms, predictive model based on logistic regression is constructed for accurately classifying diabetes.	Due to its total prediction accuracy of 92.3% it is a trustworthy substitute for categorizing diabetes and its prediction in clinical context.	More additional features relevant to diabetes prediction and classification can be included for the performance of the classifier to be enhanced.
2020 [9]	To classify diabetes Patients using Kernel-based SVM.	Here we apply kernel functions in SVM.	SVM using linear kernel displayed the best accuracy of all.	Limited exploration of other machine learning algorithms for analysis.
2019 [10]	Analyzing of ML techniques to Predict Diabetes Mellitus.	Support Vector Machine, KNN, Naive Bayes classifier, C4.5 DT are applied.	C4.5 decision tree gave best accuracy among all other machine learning methods.	The employed algorithm's accuracy in predicting the diabetes is relatively low.
2018 [11]	Predicting Diabetes Mellitus with Various Machine Learning methods.	Five-fold cross-validation procedure, Principal component analysis (PCA), a dimensionality reduction method and mRMR algorithm are used.	The findings demonstrated that when all the variables were employed, random forest prediction got highest accuracy of 80.84%.	Multiple predictors can be considered further to improve the accuracy of predictions.

2013 [12]	To classify diabetes through SVM algorithm.	SVM is employed in this method for finding diabetes.	The accuracy of the Diabetes data set's training set is 65.8 whereas its testing set accuracy is 78.2. From cross-validation accuracy, as the quantity of training samples rises, accuracy significantly improves.	Absence of feature subset selection process which can enhance the performance of the SVM classifier.

3. Methodology

The methodology used in this paper will be seen in this section. We will see the data preprocessing procedure along with the seven different algorithms used in this paper. The output gives the accuracy of these machine learning models. The algorithm with highest accuracy will then be applied to make predictions.

3.1. Data set Description

Data set: Pima Indian Diabetes data set was used for the analysis.

Pregnancies: No. of times pregnant.
Woman's risk of acquiring insulin resistance may increase due to Several pregnancies which can raise the risk of her becoming diabetes.

Glucose: Glucose concentration through oral glucose tolerance test (two hours)
Diabetes occurs when the glucose levels in the blood are elevated.

Blood Pressure: Diastolic BP measured in mm Hg.
The chance of occurrence of type 2 diabetes is more to those with high BP.

Skin Thickness: TSF thickness in mm
Triceps skin-fold thickness (TSF) is a measure of subcutaneous fat located on the back of the upper arm. Obesity is a risk factor for having diabetes (Type-2) and high TSF levels are often linked to increased body fat.

Insulin: 2-Hour serum insulin(muU/ml)
Diabetes develops when your body isn't producing or using insulin effectively.

Diabetes pedigree function:
The diabetes pedigree function can be useful in identifying patterns of inheritance for diabetes, such as whether the condition is being passed down through multiple generations of a family or is limited to a specific branch of the family tree.

Age:
The risk of having diabetes increases as people age, making age a key risk factor for the disease.

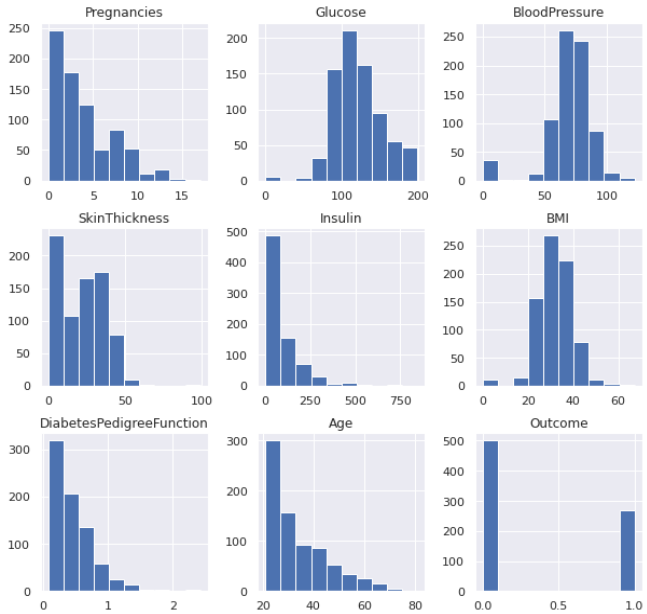

Figure 1. Histogram of important features of the dataset

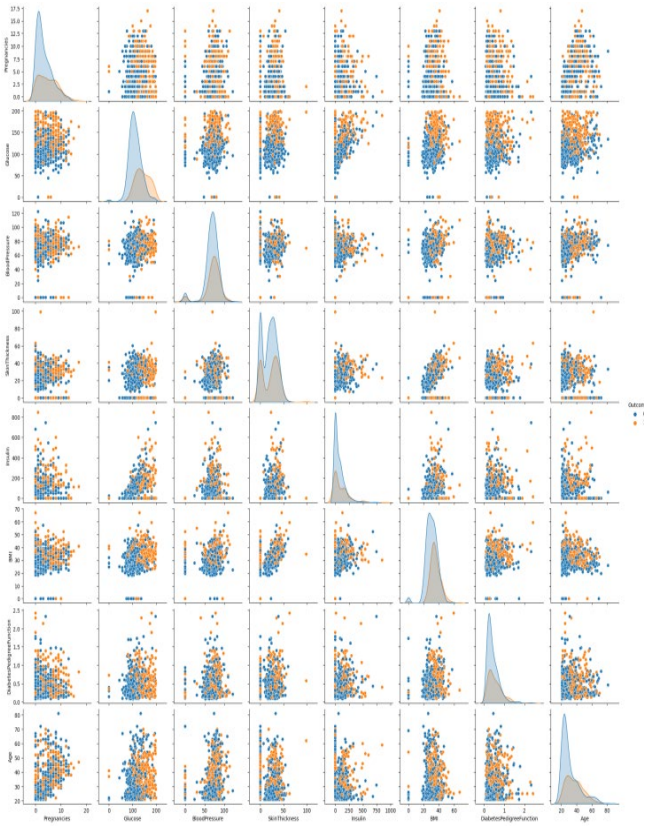

Figure 2. Pair plot of the important attributes

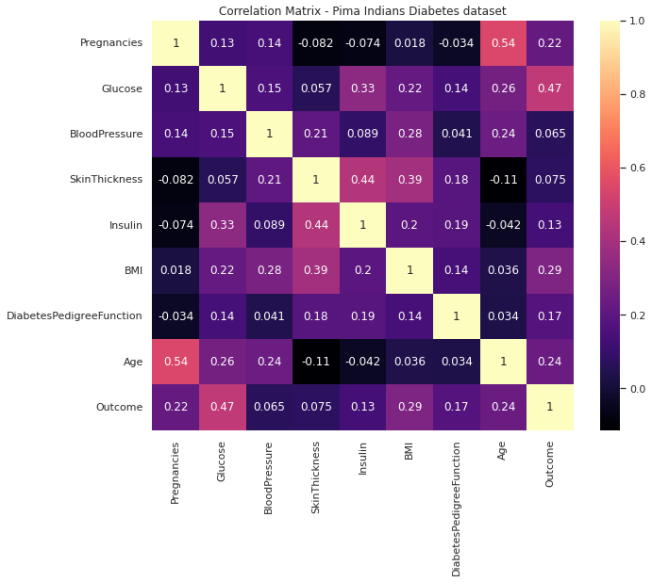

Figure 3. Correlation Matrix

3.2. Dataset Preprocessing

Various methods are implemented to clean the data which aids in getting higher output accuracy. In the first step the features that are not required for solving the problem are removed. In the following steps all the null values in the data set are handled and finally duplicate values are removed. Finally, the data set is validated.

3.3. SMOTE (Synthetic Minority Oversampling Technique)

As the data set is imbalanced (number of positive and negative outcomes are not equal) the algorithms to be used may be partial towards the majority class, resulting in poor performance for the minority class. So SMOTE was used which builds some artificial instances of the minority class by interpolating between existing examples. Specifically, SMOTE selects an example from the minority class and identifies its k nearest neighbors. It then generates synthetic examples by randomly selecting one of the neighbors and creating a new example along the line connecting the two exam-ples. This process is repeated for a specified number of synthetic examples, resulting in an augmented data set with a balanced class distribution. SMOTE has been shown to improve the performance of classification algorithms on imbalanced datasets. Fig. 4 and Fig. 5 signifies the impact of SMOTE before its application and after its application respectively.

Figure 4. Class Imbalance Problem in the Original Data set

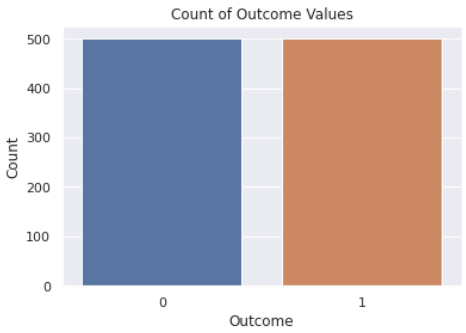

Figure 5. After Applying SMOTE

3.4. Standard Scaler and Robust Scaler

To handle the outliers and to improve the model performance two different scal-ing techniques are applied. Standard Scaler assists in making a distribution standard-ized i.e., making mean as zero and variation as one. Standard scaling is done using this formula:

$$z = x_i - mean(x) / sd(x) \quad (1)$$

Here,

z = scaled data
x = data sample

Robust Scaler is used to bring the features of the data to a similar scale. It removes the median and makes the Inter Quartile Range (IQR) value (which is the range be-tween the 1st and the 3rd quartile) as one. Python sklearn library offers with these functions.

3.5. Algorithms

SVM

Support Vector Machine algorithm aims in finding the hyperplane which will divide a given space into various classes. To do this, SVM selects extreme points called support vectors which are then used to create a hyperplane. Hence algorithm is named as Support Vector Machine.

This approach makes it easier for future data points to be classified quickly and ac-curately.

The hyperplane is of the form.

$$W^T X + b = 0 \quad (2)$$

w = weight vector
x = Vector that is to be inputted
b =bias

If equation-2 evaluates to a value greater than 0, then it's considered a positive point; if not, it's considered negative.

Decision Tree classifier

This classifier has an internal hierarchy with nodes that indicate the characteristics of the data, branches symbolizing the decision rules, and finally external nodes showing what result is achieved. In this way, it allows for a comprehensive yet sim-ple representation of information. Here the algorithm begins at the root node and uses that data to forecast the class of the new data set. This algorithm traverses the tree by comparing the attribute values of the record and each node respectively. When a match is found, it then jumps to the following node and repeats this process with all other sub-nodes until it reaches the leaf node. Fig. 6 depicts the architecture of decision tree classifier.

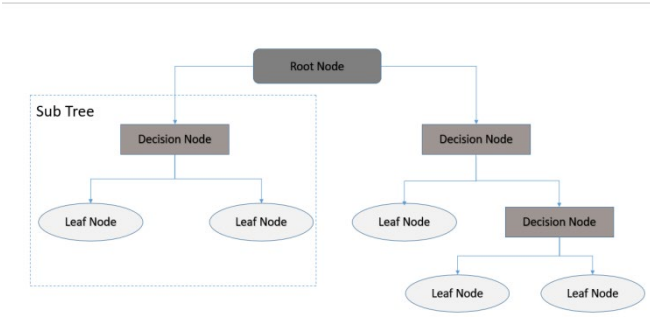

Figure 6. A typical decision tree architecture

Random Forest

To boost accuracy of the input data set, Random Forest classifier employs multiple decision trees on distinct subsets of the original data set. The Random Forest is then created by combining N Decision Trees and then predictions are made for each tree formed in the initial step and takes the average of those results.

KNN

KNN relies on previously known data to make predictions about unknown inputs. It employs the idea of similarity between unknown and existing cases and the new instance is placed in the group that closely resembles the existing categories [13].

The KNN working algorithm is:

1. Load the training and testing data set.
2. Choose the value of nearest data points i.e., k.
3. Sort the data in increasing order after finding the distance between train and test sets of the data.
4. Select top k rows.
5. The new data point will be classified based on the most frequent class of the top k rows.

Distance functions:

1. Euclidean distance:

$$\sqrt{\sum_{i=1}^{n} (xi - yi)^2} \quad (3)$$

2. Manhattan distance:

$$\sum_{i=1}^{n} | xi - yi | \quad (4)$$

Linear Regression

In linear regression, a variable's value ("dependent variable" (a)) is predicted based on the other variable ("independent variable" (b)). It tries to find the best relationship between these two variables.

Simple linear regression:

$$a=(\text{slope})b + \text{intercept} \tag{5}$$

Logistic Regression

Logistic regression is mainly used for binary classifications, i.e., it uses the given input data and predicts output which is binary. Hence the output will be discrete values (0 or 1), (yes or no) etc. The output is determined using Sigmoid function.

Sigmoid function:

$$p = 1/(1+e^{-(b_0+b_1x)}) \tag{6}$$

Naive Bayes

It is an algorithm that follows a probabilistic approach that measures the probability of an event, given an event has already happened. It is used for classification purposes.

XG Boost

eXtreme Gradient Boosting, commonly known as XGboost is a supervised machine learning technique. It makes a series of models, and a final model is prepared by combining them. It provides high execution speed and model performance. Fig 7 provides the details of the proposed architecture.

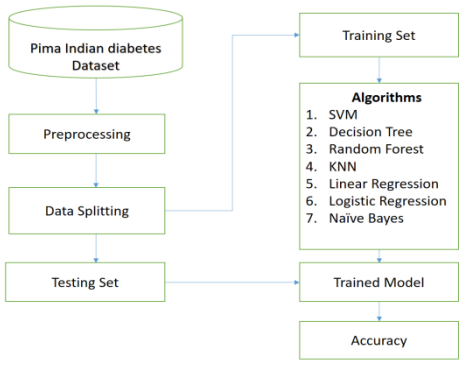

Figure 7. Overall workflow of the predictive models

4. Results

The highest accuracy achieved in this experiment is 91.10%. It was achieved through the Random Forest Algorithm. Grid search, a hyper-parameter tuning technique, is applied to the random forest algorithm. It aids in identifying the best hyper-parameter for a given model. Table 2 represents the results obtained using various machine learning algorithms. Table 3 represents the accuracy score of existing and proposed machine learning algorithms.

Table 2. Accuracy of the model when respective algorithm is implemented.

Algorithm	Accuracy (%)
SVM	83.11
Decision Tree Classifier	87.01
Random Forest	91.10
KNN	79.22
Linear Regression	80.52
Logistic Regression	84.41
Naive Bayes	81.81
XGboost	75.32

Table 3. Comparative analysis of various existing and proposed algorithms

Algorithm	Other models experimental accuracy (%)	Dataset Used by them	Our Model's Accuracy for Proposed approach (%)
SVM [12]	78.2	Pima Indian diabetes	83.11
Decision Tree [14]	71.81	Pima Indian diabetes	87.01
Random Forest [15]	75.00	Pima Indian diabetes	91.10
KNN [15]	78.00	Pima Indian diabetes	79.22
Linear Regression [16]	78.00	Pima Indian diabetes	80.52
Logistic Regression [10]	77.30	Pima Indian diabetes	84.41
Naive Bayes [10]	76.00	Pima Indian diabetes	81.81

5. Conclusion and Future Scope

In this paper, various machine learning algorithms were implemented to predict diabetes. The proposed method approach starts with data cleaning, then applying feature scaling techniques and finally using various algorithms. Finally, it is found that the Random Forest produced the maximum accuracy of all the algorithms used. In the future, an app or a website can be built that uses the proposed algorithms for early detection of diabetes and various other diseases.

References

[1] Kharroubi, A. T., & Darwish, H. M. (2015). Diabetes mellitus: The epidemic of the century. World journal of diabetes, 6(6), 850–867. https://doi.org/10.4239/wjd.v6.i6.850

[2] American Diabetes Association (2010). Diagnosis and classification of diabetes mellitus. Diabetes care, 33 Suppl 1(Suppl 1), S62–S69. https://doi.org/10.2337/dc10-S062

[3] Rabie, O., Alghazzawi, D., Asghar, J., Saddozai, F. K., & Asghar, M. Z. (2022). A Decision Support System for Diagnosing Diabetes Using Deep Neural Network. Frontiers in public health, 10, 861062. https://doi.org/10.3389/fpubh.2022.861062

[4] Alluri, R. P., & Hemavathy, R. (2021). Diabetes Prediction Using Ensemble Techniques. International Journal of Applied Engineering Research, 16(5), 410-415. Retrieved from http://www.ripublication.com https://www.ripublication.com/ijaer21/ijaerv16n5_12.pdf

[5] Salliah Shafi Bhat, Venkatesan Selvam, Gufran Ahmad Ansari, Mohd Dilshad Ansari, Md Habibur Rahman, and Mamoon Rashid. 2022. Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandi-pora. Intell. Neuroscience 2022 (2022). https://doi.org/10.1155/2022/2789760

[6] Siri, Adel & Ullah, Syed Sajid. (2021). An Improved Artificial Neural Network Model for Effective Diabetes Prediction. Complexity. 2021. 1-10. 10.1155/2021/5525271.

[7] Xue, Jingyu & Min, Fanchao & Ma, Fengying. (2020). Research on Diabetes Prediction Method Based on Machine Learning. Journal of Physics: Conference Series. 1684. 012062. 10.1088/1742-6596/1684/1/012062.

[8] Yousef K. Qawqzeh, Abdullah S. Bajahzar, Mahdi Jemmali, Mohammad Mahmood Otoom, Adel Thaljaoui, "Classification of Diabetes Using Photoplethysmogram (PPG) Waveform Analysis: Logistic Regression Modeling", BioMed Research International, vol. 2020, Article ID 3764653, 6 pages, 2020. https://doi.org/10.1155/2020/3764653

[9] G. A. Pethunachiyar, "Classification of Diabetes Patients Using Kernel Based Support Vector Machines," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-4, doi: 10.1109/ICCCI48352.2020.9104185.

[10] M. F. Faruque, Asaduzzaman and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangla-desh, 2019, pp. 1-4, doi: 10.1109/ECACE.2019.8679365.

[11] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in genetics, 9, 515. https://doi.org/10.3389/fgene.2018.00515

[12] Jegan, Chitra. (2013). Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. 3. 1797 - 1801.

[13] Zhang Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11), 218. https://doi.org/10.21037/atm.2016.03.37

[14] Shafi, Salliah and Ansari, Gufran Ahmad, Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach (May 25, 2021). Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), Available at SSRN: https://ssrn.com/abstract=3852590 or http://dx.doi.org/10.2139/ssrn.3852590

[15] AlZu'bi S, Elbes M, Mughaid A, Bdair N, Abualigah L, Forestiero A, Zitar RA. Diabetes Monitoring System in Smart Health Cities Based on Big Data Intelligence. Future Internet. 2023; 15(2):85. https://doi.org/10.3390/fi15020085

[16] Khanam, Jobeda Jamal & Foo, Simon. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express. 7. 10.1016/j.icte.2021.02.004.