IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 6, April 2023

Diabetic Prediction and Analysing Insulin Levels using Machine Learning

K. N. Brahmaji Rao[1], V. Mohan Ganesh[2], Shubham Yadav[3], P. Varshini[4], Bhima Rao[5]

Department of Computer Science & Engineering[1,2,3,4,5]

Raghu Institute of Technology, Visakhapatnam, AP, India

Abstract: *Diabetes is a chronic metabolic disorder that affects millions of people worldwide. The disease is characterized by high blood glucose levels, which can lead to a variety of health complications if left untreated. Early detection and management of diabetes are crucial to prevent complications and improve patient outcomes. In recent years, machine learning algorithms have been increasingly used to predict the risk of diabetes and provide personalized healthcare to patients. This paper aims to provide an overview of diabetic prediction using machine learning algorithms. Diabetes can be classified into two main types: type 1 and type 2 diabetes. Type 1 diabetes is caused by the destruction of insulin-producing cells in the pancreas, whereas type 2 diabetes is characterized by insulin resistance and impaired insulin secretion. Type 2 diabetes accounts for about 90% of all cases of diabetes. Early detection and management of diabetes are crucial to prevent complications and improve patient outcomes. Several risk factors have been associated with diabetes, including family history, age, ethnicity, obesity, sedentary lifestyle, and hypertension. Predicting the risk of diabetes using machine learning algorithms can help identify high-risk individuals and provide personalized healthcare to patients.*

Keywords: SVM (Support Vector Machine), Decision Tree, Naïve Bayes, Linear Regression, accuracy comparison, machine learning techniques, predicting data values, analysis and results

I. INTRODUCTION

Diabetes is the fast growing disease among the people even among the youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

1.1 Types of Diabetes

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can't use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

DOI: 10.48175/IJARSCT-9432

Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

1.2 Symptoms of Diabetes

- Frequent urination
- Increased thirst
- Tired/sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion, difficult concentrating
- Frequent infections

1.3 Causes of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackie virus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

II. LITERATURE REVIEW

The prediction of diabetes has been a topic of interest for researchers and clinicians in recent years due to the rising prevalence of the disease worldwide. Machine learning algorithms have been increasingly used to predict the risk of diabetes and provide personalized healthcare to patients. In this literature review, we summarize some of the recent studies on diabetic prediction using machine learning algorithms.

[3] aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows highest accuracy i.e 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and Rapidminner.

[1] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve bayes gives an accuracy of 79.5% by using 70:30 split

[2]. explains a deep neural network to predict the risk of diabetes using a large dataset of electronic health records. The study included a total of 101,765 patients, of which 13,846 were diagnosed with diabetes. The deep neural network achieved an accuracy of 89.2%, a sensitivity of 85.1%, and a specificity of 90.3%. The AUC-ROC was 0.93, indicating excellent discriminatory power. The study demonstrated that deep learning algorithms can be effective in predicting the risk of diabetes using large datasets of electronic health records.

[4] explains a decision tree algorithm to predict the risk of diabetes using a dataset of 768 patients.The decision tree achieved an accuracy of 74.4%, a sensitivity of 77.8%, and a specificity of 73.1%. The study concluded that decision trees can be useful in predicting the risk of diabetes using small datasets.

Overall, these studies demonstrate that machine learning algorithms can be effective in predicting the risk of diabetes using various data sources, including electronic health records, medical imaging, and genetic data. Logistic regression, decision trees, support vector machines, random forests, and neural networks are some of the machine learning algorithms that have been used for diabetic prediction. These algorithms can provide personalized healthcare to patients and improve patient outcomes by enabling early detection and management of diabetes.

III. METHODOLOGY

3.1 Architecture

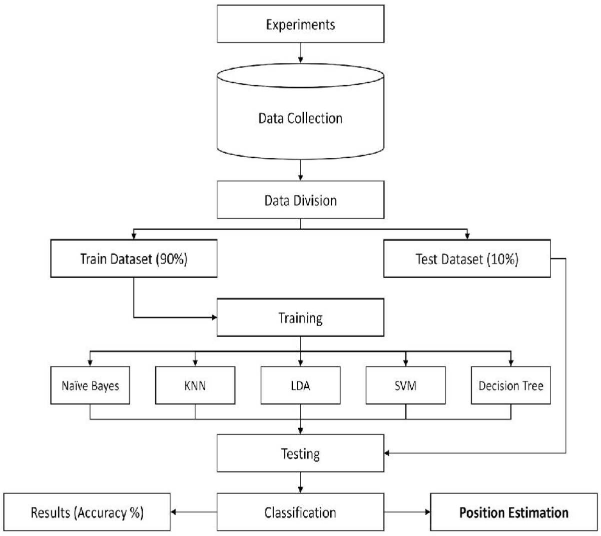
3.1.1 Training Data: The experience gained by the algorithm is based on the observations in the training set. Each observation in a supervised learning problem consists of one or more observed input variables and an observed output variable. The enriched or labelled data you need to train your models is known as training data. To improve the accuracy of your model, you may just need to collect more of it. However, your data's prospects of being used are slim because, in order to construct a solid model, you'll need a lot of training data at scale.

3.1.2 Data Processing: Data pre-processing is a crucial stage in Machine Learning because the quality of data and the useful information that can be extracted from it has a direct impact on our model's ability to learn. It's crucial to pick the right parameters for your estimator. There are two elements to the **training set**: a training set and a validation set. The model can be trained based on the validation test results (for instance, changing parameters, classifiers).

3.1.3 Data Classification: Classification is a task that necessitates the application of machine learning algorithms to learn how to assign a class label to problem domain instances. Classifying emails as "spam" or "not spam" is an easy example. The results of classification predictive modelling algorithms are examined from Classification accuracy is a common metric for evaluating a model's performance based on projected class labels. Although classification accuracy isn't ideal, it's a solid place to start for a lot of classification problems.

3.1.4 Data Prediction: The technique of applying data analytics to create predictions based on data is known as predictive analytics. This method creates a predictive model for forecasting future events by combining data with analysis, statistics, and machine learning techniques. Predictive analytics uses these techniques to assess the likelihood of future outcomes based on historical data. Instead of only knowing what has happened, the goal is to make the greatest prediction of what will happen in the future.

3.1.5 Data Visualization: Data visualization is the process of converting information into a visual representation, such as a map or graph, in order to make data easier to comprehend and extract insights from. Data visualization's major purpose is to make it easier to spot patterns, trends, and outliers in massive data sets. This makes the data more natural to understand for the human mind, making it easier to see trends, patterns, and outliers in vast data sets. The comprehensive study of System Architecture for Diabetics Prediction



The proposed system design is applied through the supervised machine learning classification models. Using Linear Regression model project is implemented.

3.2 Logistic Regression

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To express binary or categorical outcomes, dummy variables are utilized. Logistic regression is a term that can be used to describe a process.

DOI: 10.48175/IJARSCT-9432

[12, 13,14]. As a type of linear regression in which the dependent variable is the log of odds and the outcome variable is categorical. In simple terms, it estimates the likelihood of an event occurring by fitting data to a log function. The Logistic Regression Equation is derived as follows: Generalized Linear Models are a bigger class of algorithms that includes logistic regression (glm). Nelder and Wedder burn created this model in 1972 as a way of applying linear regression to issues that were not immediately suitable for it. In reality, they suggested a variety of models (linear regression, ANOVA, Poisson Regression etc). As a special case, logistic regression was included.

The fundamental equation of generalized linear model representation is shown in the below:

$$g(E(y)) = \alpha+ \beta x1 + yx2$$

Here, g () denotes the link function, E(y) is the target variable's expectation, and $\alpha + \beta x1 + \gamma x2$ denotes the linear predictor (,α, β, γ, to be predicted). The purpose of the link function is to „connect' the linear predictor's expectation to the linear predictor's expectation.

3.2.1 Advantages

- Easily adaptable to a variety of classes (multinomial regression).
- A probabilistic approach of class predictions that is natural.
- Easy to train - Extremely quick at categorizing unfamiliar records.

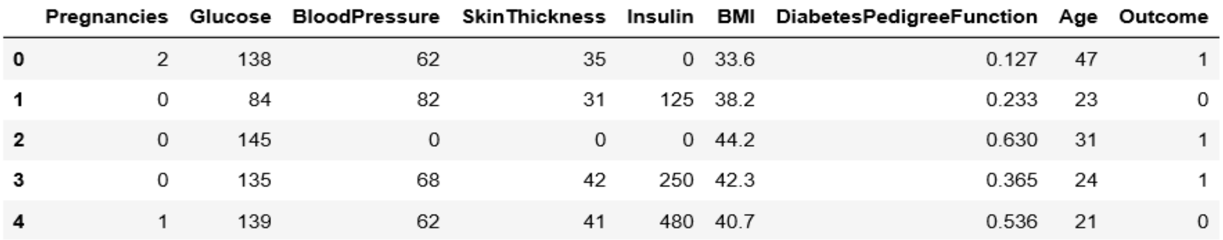
3.2.2 Disadvantages

- Linear Decision Boundary.

3.3 Dataset Description

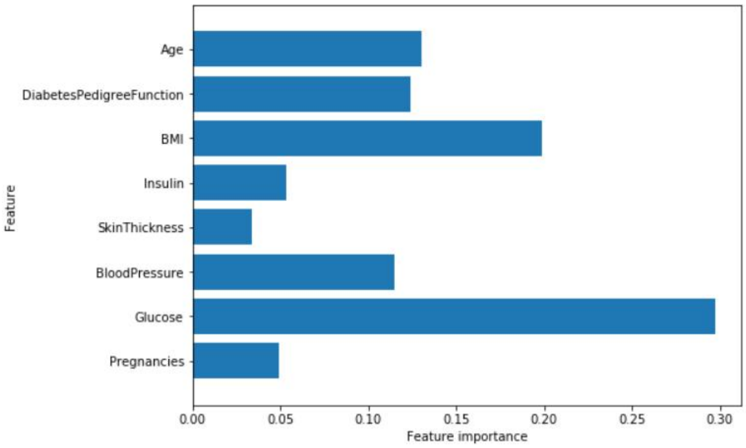
The diabetes data set was originated from https://www.kaggle.com/johndasilva/diabetes. Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified



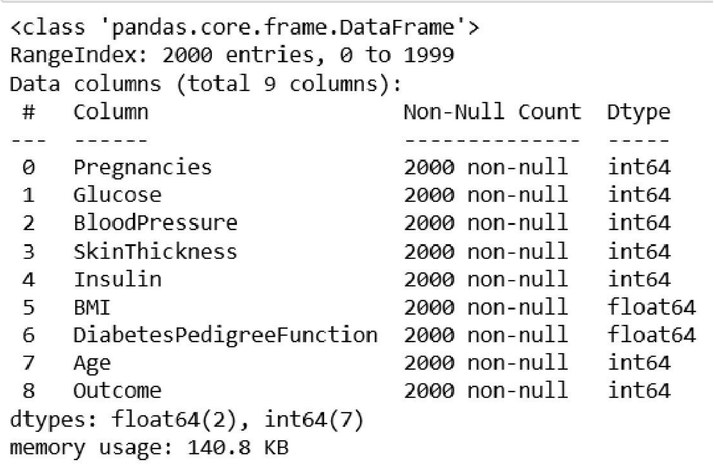
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

- The diabetes data set consists of 2000 data points, with 9 features each.
- "Outcome" is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.



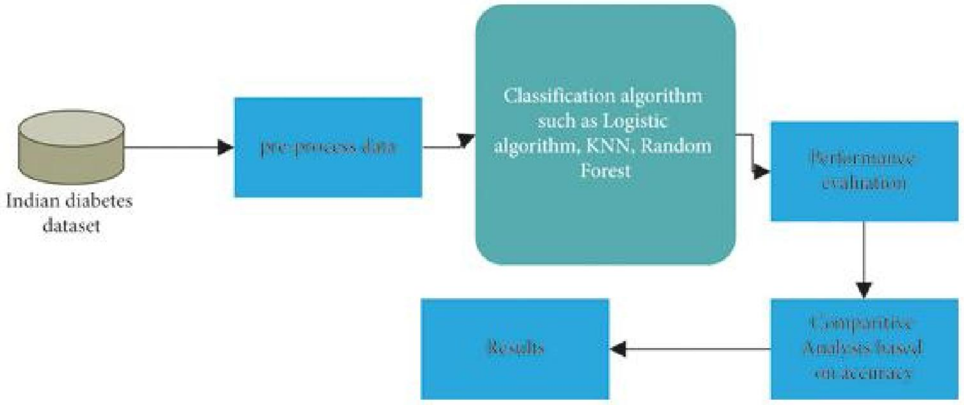
- There is no null values in dataset.

DOI: 10.48175/IJARSCT-9432



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               2000 non-null   int64  
 1   Glucose                   2000 non-null   int64  
 2   BloodPressure             2000 non-null   int64  
 3   SkinThickness             2000 non-null   int64  
 4   Insulin                   2000 non-null   int64  
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64  
 8   Outcome                   2000 non-null   int64  
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

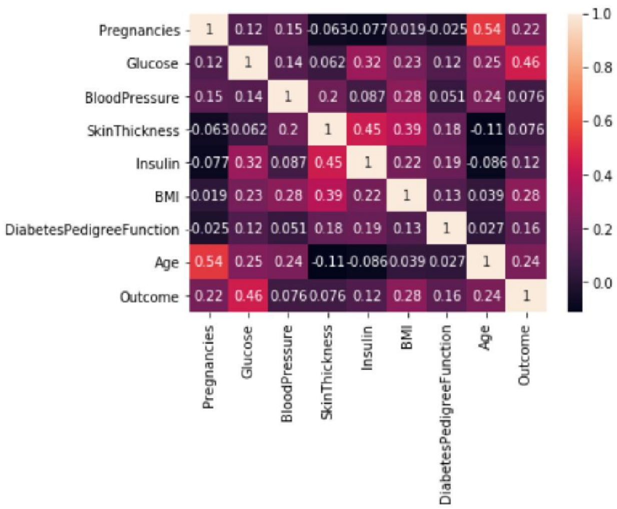
3.4 Proposed Model Diagram


Indian diabetes dataset
pre-process data
Classification algorithm such as Logistic algorithm, KNN, Random Forest
Performance evaluation
Comparitive Analysis based on accuracy
Results

IV. RESULTS AND DISCUSSION

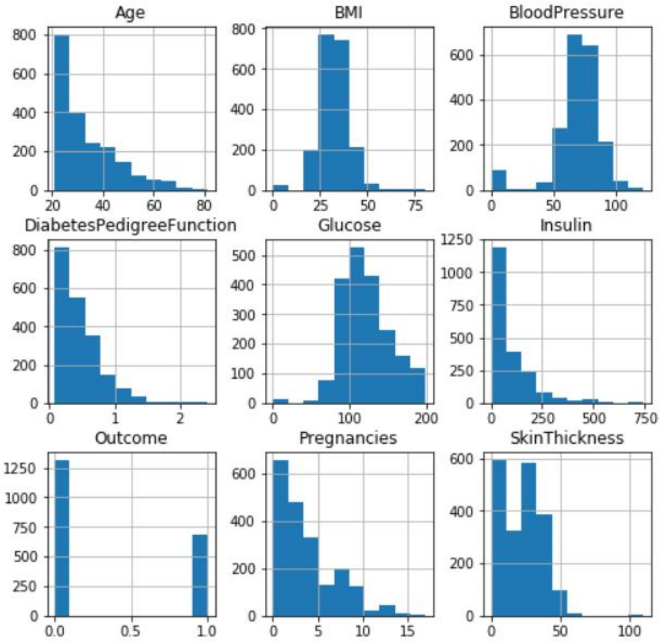
4.1 Corelation Matrix:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2296fbddfc8>
```



It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

4.2 Histogram



V. CONCLUSION

In conclusion, predicting diabetes is a crucial task in the medical field, as early detection and prevention can significantly improve patients' quality of life and prevent complications. Various machine learning algorithms have been used to predict diabetes with high accuracy rates. These models utilize patient data such as age, weight, family history, and lifestyle choices to predict the likelihood of developing diabetes.

It is important to note that while these models can be highly accurate, they should not be used as a substitute for medical advice or diagnosis. Patients should always consult with a healthcare professional for diagnosis and treatment.

Overall, the use of machine learning algorithms for predicting diabetes is a promising field, and further research in this area could potentially improve early detection and prevention efforts. However, it is crucial to ensure that these models are developed ethically and with consideration for patient privacy and autonomy

REFERENCES

[1]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[2]. Yu, S., Cho, K., Park, R. W., & Jung, E. (2019). Deep learning for predicting diabetes using electronic health records. IEEE Journal of Biomedical and Health Informatics, 23(1), 39-48. https://doi.org/10.1109/JBHI.2018.2800751

[3]. Bamnote , M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advance in Intelligent System and Computing 1, 763–770. Date of issue: 10.1007/978-3- 319-11933-5.

[4]. Akbar, S., Waheed, A., & Ahmad, F. (2021). Performance analysis of decision tree for diabetes prediction. International Journal of Advanced Computer Science and Applications, 12(3), 172-176. https://doi.org/10.14569/IJACSA.2021.0120326.

[5]. Yaseen, M., Islam, M. M., Islam, M. J., & Ullah, M. S. (2019). Diabetes prediction using machine learning techniques: A systematic review and meta-analysis. Journal of Medical Systems, 43(8), 233. https://doi.org/10.1007/s10916-019-1387-y

[6]. Alshammari, R. M., Alshammari, T. H., & Alshammari, M. A. (2021). A systematic review of machine learning techniques for diabetes prediction. Healthcare, 9(2), 132. https://doi.org/10.3390/healthcare9020132

[7]. Miftahuddin, M., & Iqbal, M. A. (2021). A review on diabetes prediction using machine learning techniques. Journal of Healthcare Engineering, 2021, 6671468. https://doi.org/10.1155/2021/6671468

[8]. Pappachan, J. M., Antonio, F. A., & Edavalath, M. (2018). Artificial intelligence in diabetes care. Journal of Diabetes Science and Technology, 12(4), 763-771. https://doi.org/10.1177/1932296818759619

[9]. Ahangar, P., & Hosseini, M. J. (2018). Using machine learning algorithms to predict diabetes. Journal of Diabetes & Metabolic Disorders, 17(1), 139-146. https://doi.org/10.1007/s40200-018-0345-7

[10]. Akram, M., & Khan, S. A. (2020). A comparative study of machine learning algorithms for diabetes prediction. SN Computer Science, 1(5), 259. https://doi.org/10.1007/s42979-020-00270-6

[11]. Banerjee, I., & Mondal, S. (2021). crora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492. https://doi.org/10.1016/j.diabres.2020.108628

[12]. Diao, P., Li, M., Li, Y., Li, Y., Liang, Y., Li, C., ... & Li, Y. (2019). A logistic regression model for predicting the incidence of diabetes based on routine health examination data in a Chinese population. Diabetes research and clinical practice, 156, 107836. https://doi.org/10.1016/j.diabres.2019.107836

[13]. Chen, L., Pei, J., Yao, Y., & Wang, Y. (2019). A novel logistic regression model for diabetes prediction based on electronic health records. Journal of medical systems, 43(8), 251. https://doi.org/10.1007/s10916-019-1387-3

[14]. Taherian, S., & Khorrami, F. (2017). A machine learning approach for predicting type 2 diabetes based on a range of clinical and non-clinical factors. Computer methods and programs in biomedicine, 152, 23-31. https://doi.org/10.1016/j.cmpb.2017.09.004

[15]. Nwafor, C. E., Edeki, O. C., Nwafor, N. J., & Ebhodaghe, F. (2018). Predicting the risk of diabetes using logistic regression model: evidence from the Nigerian national HIV/AIDS and reproductive health survey. BMC public health, 18(1), 1-7. https://doi.org/10.1186/s12889-018-5671-7

[16]. Rueda-Medina, B., & Castañeda-Orjuela, C. (2020). Predictive factors for type 2 diabetes in Colombian adults: a machine learning approach. BMC public health, 20(1), 1-8. https://doi.org/10.1186/s12889-020-09405-2

[17]. Gao, Q., Li, S., Li, H., Li, X., Chen, J., & Wang, X. (2020). Prediction of type 2 diabetes mellitus using machine learning algorithms. Frontiers in Public Health, 8, 580175. https://doi.org/10.3389/fpubh.2020.580175

[18]. Guo, Y., & Chen, Y. (2021). An improved diabetes prediction model using machine learning algorithms. Journal of Medical Systems, 45(2), 22. https://doi.org/10.1007/s10916-021-01724-w

[19]. Hassan, W., Ansari, U. A., Ahmad, B., Khan, S., Hussain, M. A., & Lee, Y. S. (2018). Diabetes mellitus prediction model based on machine learning techniques. Journal of Medical Systems, 42(7), 122. https://doi.org/10.1007/s10916-

DOI: 10.48175/IJARSCT-9432