Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

# Impact of Lifestyle Parameters for Type 2 Diabetes Prediction: A Machine Learning Paradigm

Rohini Patil
PhD student, TCET, Mumbai;
Assistant Professor, TEC, Navi Mumbai, India

Kamal Shah
Professor and R& D Head,
TCET, Mumbai

***Abstract*— Lifestyle changes are known to contribute to increased risk of type 2 diabetes mellitus. Considering morbidity and mortality associated with diabetes, its early prediction is necessary to prevent complications. In this research, performance of 6 machine learning classifiers i.e. logistic regression (LR), support vector machine (SVM), and random forest (RF), decision tree (DT), gradient boosting classifier (GBC) and K-nearest neighbor (K-NN) was evaluated for early prediction of type 2 diabetes. Through a questionnaire based study, lifestyle related parameters were collected from 374 adults. For training purpose a total of 80% data were used while testing was done on remaining 20% data. Accuracy of models was calculated for training and testing dataset. Then 10 fold cross validation was done. Accuracy of models was calculated by using different selected features. Grid search method was used for optimization. Precision, recall and F1-score and receiver operating characteristic (ROC) curve were estimated. Training accuracy of RF and DT was 100%. Highest testing accuracy of GBC and LR was 85.33%. K-fold (10 fold) accuracy of GBC was highest i.e.81.82%. Precision value of RF was highest i.e. 0.88. Similarly, recall values of LR, RF and KNN were highest i.e. 0.85 and SVM had lowest i.e. 0.81. DT and KNN gives best F-1 scores i.e. 0.84 while SVM had lowest F1-score i.e. 0.80. According to ROC curve analysis, best performance was observed with RF and GBC.**

***Keywords*— *Lifestyle, Diabetes mellitus, Machine Learning, Prediction***

## I. INTRODUCTION

Many chronic diseases affecting health of several people across the world. These illnesses are the main contributors to disability and death. In term of cost, prevalence and physical as well as psychological burden, type 2 diabetes is large healthcare burdens worldwide. Diabetes represents one of the challenging diseases due to its psychosocial and behavioral components. This is characterized by increase in blood glucose level. The reported global prevalence of diabetes in 2019 of 9.3% is expected to reach 10.9% by 2045 [1]. According to the International Diabetes Federation (IDF), India represents the second largest country for the number of diabetes cases. India contributes about 49% of world's burden and the current prevalence of 77 million is projected to reach 135 million by 2045 with first largest country with diabetes cases, a major challenge [2].

Diabetes mellitus can be cause significant morbidity and mortality due to its micro-vascular and macro-vascular complications. Several factors including genetics, growing age, race, ethnicity and lifestyle factors can rise the risk of diabetes. Health is not merely the condition of being free from illness, injury or pain; it is overall state of wellness of a person on all levels [1]. Modern way of lifestyle is contributing to huge burden on healthcare. Sleep deprivation, poor eating habits and sedentary lifestyle has contributed to growth of lifestyle diseases. Lifestyle measure refers to the personal habits, attitudes, profession, economic level, etc., that together constitute the mode of living of an individual or group. Rapid urbanization, unhealthy lifestyles and population ageing contributes to lifestyle diseases. Such diseases are preventable, and their incidence can be minimized with modifications in dietary pattern and physical exercise. Type 2 diabetes is an example of lifestyle related disease which may take years to develop, and once encountered it is difficult to manage in some patients [1, 2]. Lifestyle measures are integral part of care for preventing or delaying complications of type 2 diabetes.

ML allows to learn from prior examples and to identify patterns from large, noisy or complex data sets that can be used to formulate hypotheses [3]. Supervised learning algorithms are used for diabetes prediction. It is important to identify the relevant attribute used for prediction. In this regards, feature selection becomes important, especially for more number of feature. It helps to removes unimportant variables and increase the accuracy and performance of classification. With this background, we planned to study the usefulness of ML in prediction of type 2 diabetes based on lifestyle related parameters. Paper is organized as below:

Introduction discussed in section 1 whereas section 2 highlights related work with discussion of various machine learning and feature selection. Section 3 describes the methodology with details of 6 classifiers used in the study and also provides insights on feature selection techniques. The results of the methodology discussed in section 4. The results have been compared and analyzed. Section 5 summarizes the research followed by providing future scope.

## II. RELATED WORK

Rajappa T. et.al gives awareness study among type 2 diabetic patients in rural population based on diet, exercise, and lifestyle. They shows 74% were aware about avoidance of food item. 54% were familiar with the food proportion, 29% was lifestyle modifications and practice followed by 15% [4].Tigga and Garg designed a risk predictive model.The proposed model used 6 different classifiers, LR, SVM, KNN, RF, DT and Naïve Bayesian (NB) and compared with PIMA diabetes dataset. The result showed RF achieved an accuracy

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

of 94.1% [5].Sneha N and Gangil T developed a model using optimal feature selection algorithm. According to the results of their study, DT showed highest specificity of 98.20% whereas NB has best accuracy of 82.30% [6].Ayush A.et.al. done study on personal indicators. The author used CART model accuracy of 75% based on lifestyle and identified as blood pressure is a significant factor for the development [7].Olivera AR, et.al. Developed a model using 4 step and identified result of all ML algo and showed RF gives less accuracy while ANN, LR gives best result[8]. Dagliati A, et al. developed model for predicting complications due to diabetes using LR with stepwise feature selection algorithm having an accuracy of 83.8%[9].Hasan, et.al. proposed framework by applying preprocessing steps and feature selection for prediction of diabetes using classifiers DT, KNN, AdaBoost ,RF, XGBoost ,NB , and Multilayer Perceptron(MLP). Author developed weighted ensemble model on Pima dataset and showed ensemble classifier is the best classifier with highest AUC value as 0.950 [10].

Haq, et.al designed a system on the clinical diabetes data set. For feature selection RF, filter based DT and Ada Boost used. The results are compared with wrapper feature selection and showed that the proposed feature selection achieved optimal accuracy [11].Chen, et.al used 3 different feature selection methods as RF-variable importance, Boruta, and RFE and The author used RF, SVM, KNN, and Linear Discriminant Analysis classifiers. Compared the result with and without feature selection methods and showed that random forest provided better result among them [12].Jahan, et.al designed a system of 555*9 data size using MLP, DT and IBK algorithm to find the different levels of risk of diabetes on weka tool. The result showed that IBK gives best result for 12 fold cross validation with an accuracy 98.73% also proposed an Android application for awareness among people[13].Lama, et.al performed research for prediction of diabetes risk in middle aged people and they also highlighted the importance of stress along with other parameters including BMI, diet and tobacco consumption[14].Le,et.al have used SVM,DT,RFC,NBC,KNN and LR algorithms with feature selection using Adaptive Particle Swam Optimization and Grey Wolf Optimization method for diabetes prediction[15].Kumari, et.al. used PIMA dataset. In this study author used soft voting ensemble classifiers, RF,LR and Naïve Byes. Proposed algorithm provided an accuracy of 79.04% [16].As per the research performed by Birjais R, et.al. GB, LR and NBC are useful for prediction and diagnosis of diabetes. In this research GB provided an accuracy of 86% whereas accuracy of LR and NBC was 79% and 77% respectively on the PIMA dataset [17].Similar work has also done by other authors [18,19,20].However, only a few studies have focused dietary habit and lifestyle related information as well as most of the study done on PIMA diabetes dataset. We are addressing on these gaps in our proposed method.

## III. PROPOSED METHODOLOGY

### *3.1 Data Description*

In this research, we collected data from general adult population with more than 18 years of age using a self-developed pre-validated questionnaire. The collected information was divided into 4 sections i.e. personal information including gender, age, weight, height and body mass index (BMI), exercise situation, eating habits and other lifestyle related parameters. The dataset consisted of a total of 374 instances out of which 87 were patients with diabetes and 287 were people without diabetes. Only personal information (without any identifying information) was collected for analysis with their consent. Table 1 shows detailed description of the dataset.

TABLE I. : DATASET DESCRIPTION

| Features | Description |
|---|---|
| Gender | M/F |
| Age | Person age |
| Height | Height (cm) |
| Weight | Weight (kg) |
| BMI | Body mass index (Kg/m2) |
| Profession | Profession of a person |
| Smoke | Specifies Yes/No |
| Exercise | Specifies the exercise levels of a person |
| Cereal grains consumption | Specifies consumption quantity of cereals |
| Salad consumption | Specifies consumption quantity of salad |
| Cooked Vegetables | Specifies consumption quantity of cooked vegetables |
| Sweet | Specifies Yes/No |
| Frequency of sweet | Specifies frequency of sweet consumption |
| Refined Sugar | Yes/No |
| Milk Product Consumption | Yes/No |
| Milk quantity | Specifies consumption quantity of milk |
| Class Label | Diabetes / No diabetes |

### *3.2 Model Architecture*

Overall methodology was divided in 3 phases. In phase I, preprocessing of data was done. A predictive model using 6 machine learning classifiers namely support vector machine (SVM), logistic regression (LR), random forest (RF), decision tree (DT) K-nearest neighbor (K-NN) and gradient boosting classifier (GBC) was used. In phase II , we applied feature selection techniques namely filter method-Select-K best method, Feature importance technique, Information gain, Correlation technique and hybrid method as recursive feature elimination (RFE).As selecting important features is the key for success of early diagnosis of diabetes, we calculated the performance of models using different selected features. The number of features providing highest accuracy for each model using all five feature selection methods was noted along with highest accuracy. Comparative analysis of feature selection technique was done.

In phase III, the model optimization was done using grid search method for all selected classifiers. Cross validation was done with 10 fold of the data in the grid search. The optimal hyper parameters were ranked based on their accuracy and "Area Under the Receiver Operating Characteristic Curve (ROC- AUC curve)". Figure 1 summarizes our study methodology.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

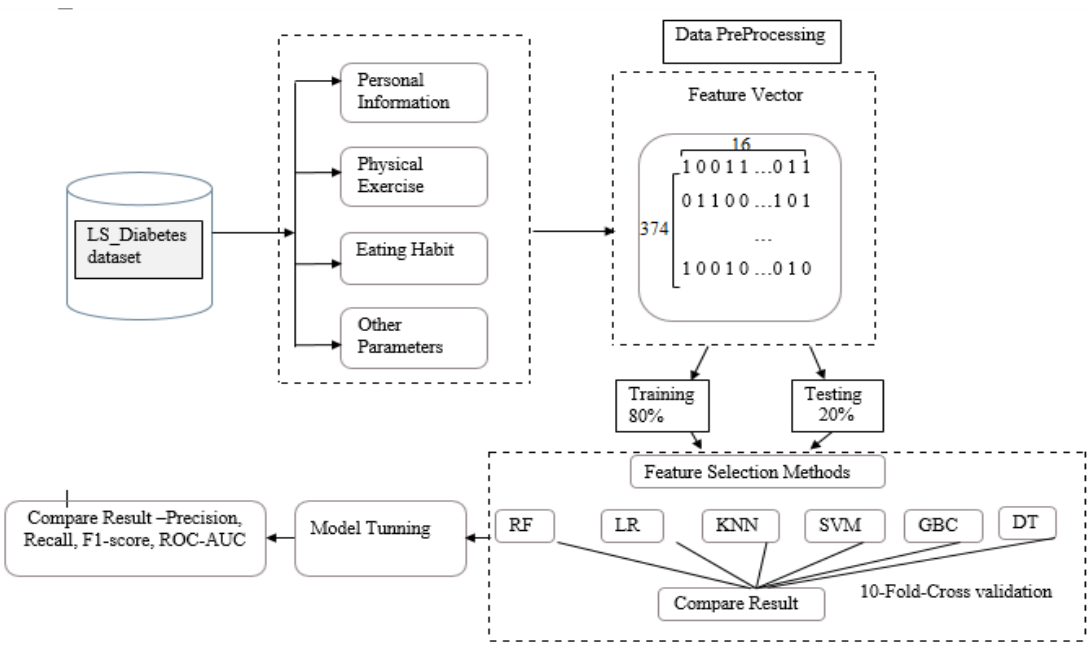


Fig. 1: Proposed methodology for lifestyle parameters

Performance measures including precision, recall and F1-measure were calculated and ROC curve was plotted for checking the robustness of the algorithms. The results of 6 models were compared based on these performance measures.

## IV. PERFORMANCE ANALYSIS

The study included 374 participants with 16 feature columns. In order to check the efficiency and robustness of the model performance matrices including Accuracy, Precision, Recall and F1 score were used. These matrices are derived using following equations.

$$Accuracy = \frac{Tp + Tn}{Tn + Tp + Fp + Fn} \quad (1)$$

$$Precision = \frac{Tp}{Tp+Fp} \quad (2)$$

$$Recall = \frac{Tp}{Tp+Fn} \quad (3)$$

$$F1score = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Where,

True positive ($Tp$) the instances are true (T) while they are (T)
True negative ($Tn$) the instances are false (F) while they are (F).
False negative ($Fn$) the instances are (F) while they are (T).
False positive ($Fp$) the instances are s (T) while they are (F).
Receiver operating characteristic (ROC) curve was plotted for different algorithms for comparison of true positive rate (TPR) to the false positive rate (FPR).

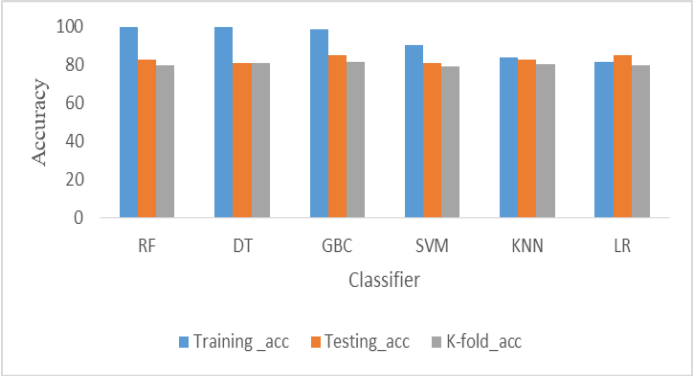


Fig. 2: Accuracy comparison of different classification algorithms

TABLE II. FEATURE SELECTION ACCURACY COMPARISON ON VARIOUS METHODS

| Classifier | RFE | Select K-Best | Feature mportance | Correlation | Information Gain |
|---|---|---|---|---|---|
| RF | 80.75 | 80.24 | 80.53 | 78.62 | 80.51 |
| GBC | 82.92 | 79.16 | 79.96 | 78.62 | 80.23 |
| SVM | 80.23 | 81.31 | 81.85 | 81.31 | 82.39 |
| KNN | - | 80.24 | 80.24 | 80.24 | 80.24 |
| LR | 80.24 | 79.95 | 80.48 | 80.48 | 80.48 |
| DT | 81.57 | 76.22 | 78.07 | 77.04 | 76.74 |

The accuracies after feature selection methods for corresponding machine learning algorithms are shown in Table 2. Accuracy comparison by applying feature selection and K-fold comparison are shown in Fig. 3.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

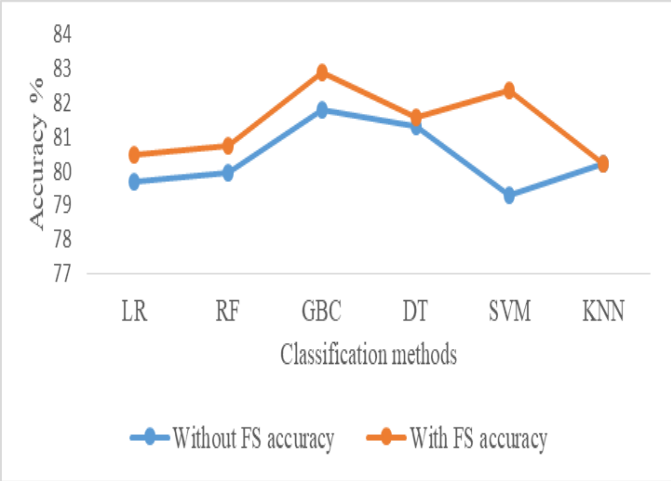


Fig. 3: Comparison of with and without feature selection accuracy

After feature selection, LR, RF, GBC, DT, SVM, K-NN provided accuracy of 80.48%, 80.75%, 82.92%, 81.57%, 82.39% and 80.24%.

From the proposed pipeline i.e. After applying the grid search as an optimization technique, the respective accuracies of classifiers were 85.33%, 84%, 85.33%, 85.33%, 84% and 81.33% respectively. The best performance for the prediction of diabetes is achieved by GBC,LR and KNN model shown in below fig.4.

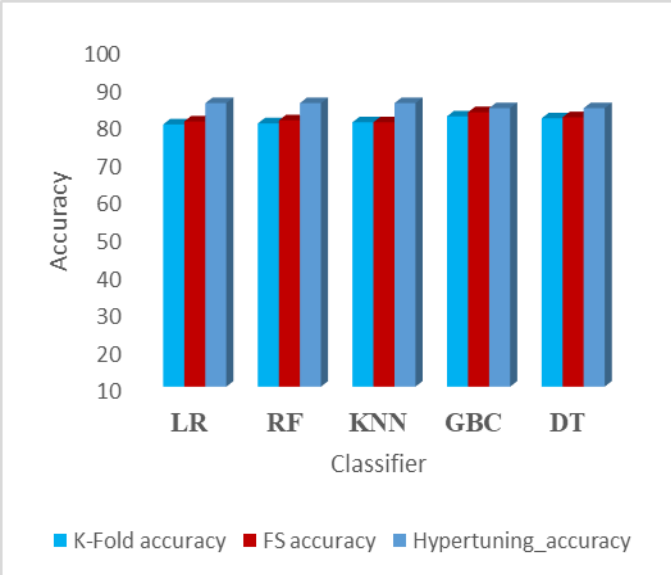


Fig. 4: Comparison of k-fold, feature selection and hypertuning accuracy

TABLE III. : STATISTICAL ANALYSIS OF VARIOUS CLASSIFIERS

| Methodology | Precision | Recall | F1-score | Roc-Auc score |
|---|---|---|---|---|
| RF | 0.88 | 0.85 | 0.82 | 92.8% |
| GBC | 0.84 | 0.84 | 0.81 | 92.49% |
| LR | 0.86 | 0.85 | 0.83 | 87.52% |
| K-NN | 0.85 | 0.85 | 0.84 | 85.55% |
| DT | 0.84 | 0.84 | 0.84 | 82.5% |
| SVM | 0.80 | 0.81 | 0.80 | 81.74% |

Table 3 shows statistical analysis of various classifiers. Here weighted average values are to be considered. Precision value of RF was highest i.e. 0.88 while that of SVM was 0.80. Similarly, recall values of RF,LR and KNN were highest i.e. 0.85. Recall values of SVM were lowest i.e. 0.81. SVM had lowest F1-score i.e. 0.80 whereas highest F1- score was observed with KNN and DT i.e.0.84 Overall, RF and GBC gives good result. Below fig.5 shows ROC curve analysis with highest value achieved by RF and GBC model.

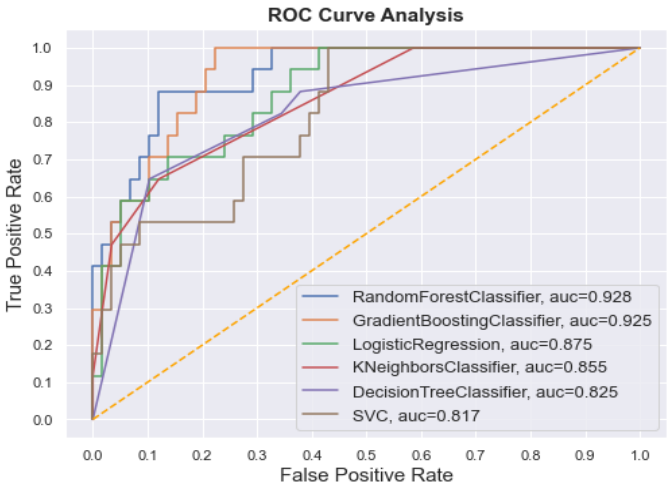


Fig. 5: Comparison of with and without feature selection accuracy

## V. CONCLUSION AND FUTURE WORK

We performed a study to develop predictive model for estimation the risk of type 2 diabetes mellitus using machine learning algorithms based. The dataset consisted of demographics and lifestyle related parameters of 374 people with or without diabetes. In this paper, we compared 6 classifiers method Random Forest (RF), Gradient Boosting Classifier (GBC), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT). We combine those ML method with various features selection method Select K-best, Feature importance, Correlation, Information gain and Recursive feature elimination to select the best classifiers method based on various measures. GBC with RFE as feature selection method provided an accuracy of 82.92%. RF showed best performance model by achieving an AUC value of 92.8%. Overall, RF, LR and KNN provided best accuracy of 85.33%.

Performance measures in our study suggest lifestyle related parameters are the risk factor for the development of type 2 diabetes mellitus. Strategies for avoiding junk food, doing regular exercise should be employed in the high risk population for reducing risk of type 2 diabetes development.

## DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## REFERENCES

[1] WHO Homepage,http://www.who.int/en/news-room/fact-sheets/detail/diabetes accessed 2019/02/21.
[2] IDF Homepage, https://www.idf.org/our-network/regions-members/south-east-asia/.../94-india.html accessed 2019/02/22.
[3] Saikat Dutt, Amit Kumar Das: Machine Learning. Pearson Education, India (2018).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

[4] Rajappa T., Ponniraivan K., Kalyan H., Selvaraju K., Karunanandham S.: Assessment of degree of awareness about diet, physical exercise, and lifestyle modifications among diabetic patients. Int J Med Sci Public Health2018; 481-86.
[5] Tigga N.,Garg S.: Prediction of Type 2 Diabetes using Machine Learning Classification Methods Classification, International Conference on Computational Intelligence and Data Science, Procedia Computer Science 2020;.167:706-16.
[6] SnehaN. &GangilT.: Analysis of diabetes mellitus for early prediction using optimal features selection,J Big Data 2019;6(13):1-19
[7] Aysh A., Divya S. :Prediction of diabetes based on personal lifestyle indicators,In :1st international conference on next generation computing technology:IEEE(2015)
[8] Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo A, Barreto SM, Duncan BB.: Comparison of machinelearning algorithms to build a predictive model for detectingundiagnosed diabetes - ELSA-Brasil: accuracy study Sao Paulo Med J 2017; 135:234-46.
[9] Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R.: Machine learning methods to predict diabetes complications. Journal of Diabetes Science and Technology 2018;12: 295-302.
[10] Hasan MD. K.,Alam MD.A.,Das D.,Hossain E.,Hasan M.: Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers.IEEE Access 8(2020). https://doi.org/10.1109/ACCESS.2020.2989857.
[11] Haq A.,Li J.P.,Khan J.,Memon M.H.,Nazir S.,Ahmad S.,Khan G.A.,Ali A.,Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data, Sensors 2020;20(2649):1-21.
[12] Chen R., Dewi C.,Huang S.,Caraka R.E.:Selecting critical features for data classification based on machine learning methods.J Big Data 2020;7:1-26.
[13] Jahan N.,Islam A.,Mamun A.A.,: Machine Learning With Factor Scoring To Predict Diabetes Risk Level In Bangladesh., Int J of Scientific & Technology research 2020; 9(2):2863-67.
[14] Lama L.,Wilhelmsson O.,Norlander E.,Gustafsson L.,Lager A.et.al.: Machine learning for prediction of diabetes risk in middle-aged Swedish people.Heliyon 2021;7:1-6.
[15] T. M. Le Vo T.M.,Pham T.N.,Dao S.V.T.: Novel Wrapper—Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic. IEEE engineering in medicine and biology society section 2021;9:7869-84.
[16] Kumari S.,Kumar D.,Mittal M.:An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier International Journal of Cognitive Computing in Engineering 2021;2:40-6.
[17] Birjais R.,Mourya A.K.,Chauhan R.,Kaur H. Prediction and diagnosis of future diabetes risk: a machine learning approach,SN Applied Sciences 2019;1:1112.
[18] Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. Int. J. Environ. Res. Public Health 2021; 18,3317:1-14.
[19] Kaur H.Kumari V. :Predictive Modelling and analytics for diabetes using a machine learning approach, Applied Computing and Informatics (2018).
[20] Nai-aruna N, Moungmaia R. Comparison of classifiers for the risk of diabetes prediction. 7th International Conference on Advances in Information Technology Procedia Computer Science.2015; 69:132-42. (2015).