Grenze International Journal of Engineering and Technology, June Issue

Diabetes Prediction using Machine Learning

Vicky Kumar Sharma[1], Shubham[2], Alok Chaurasiya[3], Sonia Rani[4] and Vikas Chaudhary[5]
[1-5]School of Computer Applications and Technology, Greater Noida, India
Email: vickykumarsharma555@gmail.com, subhampanchalchhata@gmail.com, alokchaurasiya949@gmail.com, soniarani@galgotiasuniversity.edu.in, vikas.jnvu@yahoo.co.in

***Abstract*—Diabetes mellitus is a long-term metabolic disorder, which sometimes causes serious complications in the heart, kidneys, nerves, and eyes when not detected in time. In light of the effect of the high-risk factors like obesity, age, insulin resistance, hypertension and lifestyle imbalance, intelligent diagnostic support syste5s are necessary to identify risks early. This paper will suggest a Pima Indians Diabetes Dataset-based and machine learning-based smart diabetes prediction model. Preprocessing of data was done to deal with missing and invalid values and to deal with class imbalance by making use of SMOTE and ADASYN. GR, SVM, K-NN, RF, GB, and XGBoost were selected as multiple classifiers and trained and optimized with the help of GridSearchCV and tested on the basis of the accuracy, precision, recall, F1-score, and ROC-AUC measures. It was experimentally found that the ADASYN with XGBoost had the highest performance with an accuracy of 82.47% and an AUC of 0.90. Explainable AI methods including SHAP and LIME were used to improve clinical interpretability through risk factors identification. The results show that the proposed machine learning-based diagnostic support can be very helpful in detecting diabetes at an earlier stage and aid in healthcare decision-making.**

***Index Terms*— Machine Learning, Class Imbalance, SMOTE, ADASYN, Pima Indians Dataset, Explainable AI, Decision Support System.**

I. INTRODUCTION

Diabetes mellitus, or better said diabetes, is a fast spreading health concern in the whole world, and has far reaching consequences to people, health facilities, and even to the economies of nations [1], [2]. Chronic hyperglycemia is the main symptom of the disease, and it is widely categorized as several types where Type 1, Type 2, and gestational diabetes are the commonest types of the disease [3]. Its multifaceted character, as well as the silent character of development in most cases, makes early detection and proper diagnosis essential in preventing further complications in the long term [4], [5], [6]. As per the world statistics, an estimated 451 million individuals are having diabetes in 2017 and this figure is expected to grow to 693 million cases by 2045 [7]. According to more recent research, the world has already surpassed half a billion diabetic patients, which admits of the frightening rate of the disease expansion. It is projected that this may increase by another 25% in the future adding more reasons to why proactive and cost- effective diagnostic solutions are necessary in the first place [8]. However, such traditional diagnostic approaches as the fasting blood glucose test and the HbA1c test are important as they work only based on biochemical parameters and cannot reveal the insidious and insidious development of the disease.

Grenze ID: 01.GIJET.12.2.198_20

Consequently, more medical professionals are beginning to consider more complex, less-invasive and biomedical-based methods of risk detection of diabetes at an earlier age. Machine learning and predictive analytics, specifically, have potential tools that would analyze complex patterns in patient data, which allows to conduct a risk assessment faster and more accurately. These new technologies can revolutionize the medical treatment and prevention of diabetes and help it through early interventions and custom healthcare plans.

II. Related Work

There were numerous related works done in the field of automatic healthcare processes previously. In this part the research topic and the extent of the study are clearly outlined and a literature review is carried out by utilizing the databases of scholars to appraise the relevance of each source. The remaining part of this section contains the latest and related literature. We would categorize the sources according to the themes and the concepts they discuss and the influence of the sources, basing on the level of credibility and the reputation of the authors [10].To do so, several ML methods are performed to build correct predictive systems that can support healthcare experts in making clinic-related decisions. The experimental findings prove that the suggested models have a high level of performance concerning accuracy and other measures. The performance improvement also substantiates the efficiency of these ML-based systems as decision support systems used to identify the factors exposing diabetes risks and help medical professionals diagnose diabetes mellitus at an initial stage [9]. Multiple researches have tried to construct risk prediction models of type 2 diabetes by applying the machine learning method. In one of them, Pima Indians Diabetes Dataset (Public) were used, and various ML models were trained to make predictions. To overcome the problem of imbalance in the class distribution that exists in the dataset, Synthetic Minority Oversampling Technique (SMOTE) was used as a technique to avoid bias in the models. The experimental results showed that the Random Forest (RF) classifier is superior to other ones, and its accuracy is 86.97.

III. Methodology

Diabetes is attributed to the fact that the pancreas does not produce the required insulin or the body does not consume the produced insulin. The level of the number of individuals with diabetes in the US is high according to reports which have been ascertained by WHO. Continuous glucose monitors (CGMs) are emerging as a highly desired method of blood glucose monitoring. The procedure involves the insertion of a right needle wearable that measures the level of blood sugar (tissue levels) into the skin. The CGM monitors the direction of the blood glucose and raises an alarm to prevent hypoglycemia and give the optimal glycemic monitoring [10]. The information is related to the smartphone through Bluetooth, and the information is linked to the cloud computing (internet or network) server. In the case of diabetics, when the sugar level increases in the system, numerous alerts are made immediately. Once the percentage of diabetes is greater than the allowed limit, particularly with diabetic patients, then the emergency will be made. This piece of work demonstrates the use of machine learning in order to predict diabetes. We have a dataset of 769 records and 8 features which give the likelihood of whether a patient has diabetes or not. The current status of the patient is automatically transferred to the intelligent healthcare system and saved in an EHR to process it further. As it may be observed, this figure shows that the patients data is processed in the cloud and trained based on the pretrained models and the necessary action can be activated automatically based on the results obtained in the cloud. Figure 1 demonstrates the suggested system of diabetes prediction/monitoring in smart health cities to be implemented in the cloud [11].

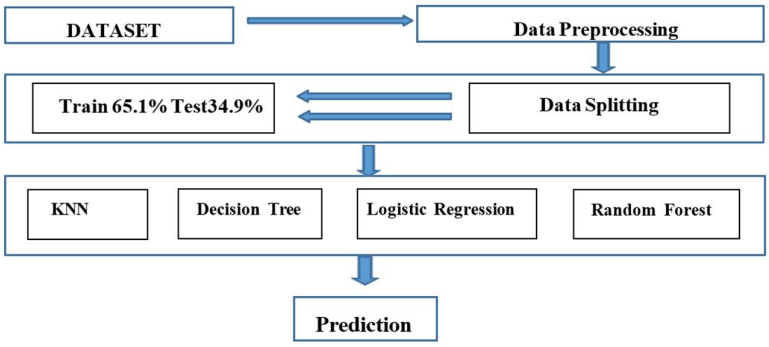

Figure 1. The Proposed Diabetes Prediction/Monitoring System in Intelligent Health Cities [12].

A. Dataset

To ensure that the Smart Diabetes Prediction System is reliable and relevant in the real world, the Pima Indians Diabetes Dataset, a popular open-source medical dataset in diabetes prediction research, will be used. The dataset is a collection of clinical records of 769 Pima Indian descent female respondents aged 18 to 77 years who willingly gave their data to be used in the research. Out of these, 269 of them are diagnosed to have diabetes and the non-diabetic patients make up the rest leading to an unequal distribution of classes as shown in Figure 2[12]. The records contain the important health- related objects, including the number of pregnancies, the level of glucose in the blood, blood pressure, the thickness of the skin, body mass index (BMI), age, and the outcome of diabetes. The data were collected using standard medical tools, such as a GlucoLeader Enhance glucose meter, OMRON HEM-7156T blood pressure monitor, and a digital LCD body fat caliper, and is thus appropriate to be used in the training and evaluation of machine learning-based diabetes prediction models.The dataset is also giving some statistical information like minimum, maximum and mean values for each of the attributes which are summarised in Table 1[13].

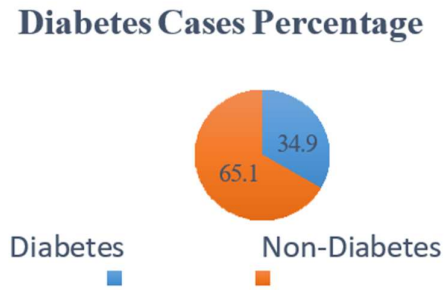

Figure 2. Pima Indian dataset

TABLE I: FEATURES OF THE PIMA INDIAN DATASET

Features	Minim um	Maximu m	Averag e
Pregnancies	0	8	1.61
Glucose(mg/dl)	52.2	274	109.39
Blood Pressure (mm Hg)	5.9	115	71.09
Skin thickness(mm)	2.9	23.3	10.78
BMI(Kg/m2)	2.61	41.62	22.69
Age(years)	17	77	27.02

B. Dataset Preprocessing

The consolidated dataset was checked during the pre processing to determine inconsistencies and physiologically invalid values. The occurrence of zero values in some of the attributes like BMI and skin thickness which do not serve a medical purpose was substituted with the mean of the corresponding features in order to maintain the consistency of data without any significant change in the overall distribution. The dataset was partitioned after the data cleaning with the holdout method, 80 percent of the data was used to train and 20 percent to test to ensure that the data has been learned and the performance is also assessed without bias. Mutual Information (MI) was used to evaluate the feature relevance as it quantifies the dependency between features in the inputs and the target class. The higher the MI scores, the higher the contribution of prediction.

C. Semi-supervised learning

The dataset of Pima Indians Diabetes has fully filled class labels with missing or invalid values in some of the clinical features especially insulin which is essential in diabetes analysis. Eradicating these records would erode the quality of data and accuracy of models hence to fill the gaps in insulin values a semi-supervised regression method was utilized. Valid insulin measurements were considered as labelled data and the ones with missing or zero measurements were assumed to be unlabeled. The regression models were trained to estimate the relationship between insulin and other clinical characteristics including the glucose level, BMI, age, and blood pressure.The data was divided into 80 percent training and 20 percent validation data. Root Mean Square error (RMSE), which is defined as:

$$\text{RMSE} = \sqrt{\left(\Sigma(\mathbf{Predicted}_i - \mathbf{Actual}_i)^2 \boldsymbol{Ni}=\mathbf{1}/\ \text{N}\right)} \quad (1)$$

D. Pre-Treatment of Data Preparation and Class Imbalance

The Pima Indians Diabetes Dataset was ready to be trained and examined using models after all preprocessing procedures, including missing-value treatment and feature-importance analysis. Using mutual information analysis, the Diabetes Pedigree Function was established as the least informative variable, and it has been eliminated to simplify the model and eliminate noise. The normalized data had only clinically significant attributes to predict diabetes. Nevertheless, some imbalance in the data in terms of class was observed, that is, almost 500 non-diabetic and 269 diabetic cases, which may bias the models towards the majority group and lower sensitivity to diabetic cases in case it was not addressed.

E. Class Imbalance SMOTE and ADASYN

Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) were used in the training data only to minimize the imbalance between the classes but the test sample remained intact to enable an unbiased assessment. SMOTE creates synthetic minority samples, which is done by interpolating between the existing minority samples and their nearest neighbors. ADASYN expands this method to more challenging to learn minority samples, which produces additional synthetic data in areas with greater class imbalance. In comparison to random oversampling, the two methods do not repeat existing samples therefore causing overfitting to be minimized. These methods yielded a balanced dataset on training, which allowed better generalization and predictive accuracy of the trained machine learning types of classifier.

F. Minmax normalization

The minmax normalization which has been used in this study is the one used. The data is brought to the equal range using the equation below:

$$Xscaled = (X - X_{min}) / (X_{max} - X_{min}) \quad (2)$$

IV. Comparative Analysis of Machine Learning and Ensemble Techniques for Diabestes Prediction

The article employs a number of machine learning and ensemble techniques to develop an automatic predicting system of diabetes. The combination of hyperparameters and overfitting is taken to the best possible level with the assistance of the GridSearchCV framework. The Decision Tree model is an instance of rule based learning device to determine discrete values of a target based on the choice of a node depending on the information gain that may be the Gini impurity or the entropy.

$$Gini_i = 1 - \sum_{k=1}^{n} (p_{i,k})^2 \quad (3)$$

$$Entropy = \sum_{i=0}^{n} -p_i \log_2(p_i) \quad (4)$$

In the current study, the researcher uses various machine learning and ensemble models to forecast diabetes including KNN, Random Forest, Logistic Regression, Gradient Boosting, and XGBoost. KNN uses nearest neighbors (K = 5) to make predictions whereas a random forest improves prediction based on combination of additional decision trees and the removal of variance. Logistic regression offers the estimate of the probability of the class in binary classification and gradient boosting (building of powerful models) by minimizing the errors. Another way of enhancing the performance of XGBoost is by optimizing the boosting plus the ability to control the depth of the trees which are ensuring the improved generalization and reduced complexity.

V. Used Performance Parameters

The present section describes the results of the experiment and analysis of the developed automated diabetes prediction model. The effectiveness of various machine learning classifiers is initially checked and compared. Subsequent to this, how the system is implemented on the web interface and the Android-based mobile application is also briefly described.In order to evaluate and compare predictive performance of the models, the common evaluation criteria were used, which were precision, recall, F1-measure, area under the ROC curve (AUC), and accuracy of classification. All these measures are aimed at measuring the accuracy, dependability and sensitivity to prediction of the classifiers.

A. Formula

$$Precision = TP / (TP + FP) \quad (5)$$

$$Recall = TP / (TP + FN) \quad (6)$$

$$F1\ score = (2 \times Recall \times Precision) / (Recall + Precision) \quad (7)$$

Table 2. Compares the various performance measures of various classifiers of the merged data using SMOTE synthetic oversampling method.
Based on this table, the Random forest classifier performed the best with an overall accuracy of 0.7922 and F1 score and AUC of 0.8697 respectively.
Table 3 provides the overall performance of all the classifiers based on ADASYN technique on combination dataset. As depicted in Table 4, XGBoost had the highest results with an accuracy of 0.8247 and a greater AUC of 0.9087, and the Decision Tree classifier was the worst performer, especially on accuracy and F1-score.

TABLE II: CLASSIFICATION PERFORMANCE WITH SMOTE TECHNIQUE

Model	Precision	Recall	F1- Score	Accu racy	AUC
XG Boost	0.724	0.781	0.751	0.805	0.884
Random Forest	0.702	0.759	0.729	0.792	0.869
Gradient Boosting	0.678	0.737	0.706	0.772	0.852
Logistic Regression	0.645	0.722	0.682	0.753	0.834
SVM	0.657	0.694	0.675	0.746	0.821
KNN	0.639	0.740	0.686	0.753	0.815

TABLE III: CLASSIFICATION PERFORMANCE WITH ADASYN TECHNIQUE

Model	Precis ion	Recall	F1- Score	Accuracy	AUC
XG Boost	0.752	0.805	0.778	0.824	0.908
Random Forest	0.729	0.787	0.757	0.811	0.892
Gradient Boosting	0.705	0.759	0.731	0.792	0.875
Logistic Regression	0.672	0.745	0.707	0.772	0.856
SVM	0.685	0.717	0.701	0.766	0.843
KNN	0.667	0.763	0.712	0.772	0.838

A. Precision

Secondly, a domain adaptation method was used whereby the machine learning model is trained on a source dataset and tested on a target dataset. The mater of automatic diabetes prediction model in this study was first trained on a larger Pima Indian open-source data.
The trained model was subsequently tested on the smaller, privated RTML data, the findings of which is provided in Table 6.
It is important to note that in the training phase, XGBoost model was used in conjunction with the ADASYN technique in resolving the issue of class imbalance.
Performance Metrics for Random Forest: True Positives (TP) = 31: Correctly identified diabetic cases True Negatives (TN) = 86: Correctly identified non-diabetic cases False Positives (FP) = 14: Non-diabetic cases incorrectly classified as diabetic (Type I error) False Negatives (FN) = 23: Diabetic cases incorrectly classified as non-diabetic (Type II error) that shown in table 4.

TABLE IV: CONFUSION MATRIX OF RANDOM FOREST

Actual\Predicted	Non-Diabetic	Diabetic
Non-Diabetic	86	14
Diabetic	23	31

TABLE V: CONFUSION MATRIX LOGISTIC REGRESSION

Actual\Predicted	Non-Diabetic	Diabetic
Non-Diabetic	82	18
Diabetic	27	27

Performance Metrics for Logistic Regression: True Positives (TP) = 27: Correctly identified diabetic cases True Negatives (TN) = 82: Correctly identified non-diabetic cases False Positives (FP) = 18: Non-diabetic case incorrectly classified as diabetic (Type I error) False Negatives (FN) = 27: Diabetic cases incorrectly classified as non-diabetic (Type II error) that shown in Table 5.
Performance Metrics for SVM: True Positives (TP) = 26: Correctly identified diabetic cases True Negatives (TN) = 83: Correctly identified non-diabetic cases False Positives (FP) = 17: Non-diabetic cases incorrectly classified as diabetic (Type I error) False Negatives (FN) = 28: Diabetic cases incorrectly classified as non-diabetic (Type II error) that shown in Table 6.

TABLE VI: CONFUSION MATRIX OF SVM

Actual\Predicted	Non-Diabetic	Diabetic
Non-Diabetic	83	17
Diabetic	28	26

TABLE VII: CONFUSION MATRIX OF KNN

Actual\Predicted	Non-Diabetic	Diabetic
Non-Diabetic	84	16
Diabetic	25	29

Performance Metrics for KNN: True Positives (TP) = 29: Correctly identified diabetic cases True Negatives (TN) = 84: Correctly identified non-diabetic cases False Positives (FP) =16: Non-diabetic cases incorrectly classified as diabetic (Type I error) False Negatives (FN) = 25: Diabetic cases incorrectly classified as non-diabetic (Type II error) that shown in Table 7.

Figure 3 demonstrated that the glucose level and BMI became the most significant characteristics, and age and insulin came next in terms of their impact on the results of diabetes, whereas skin thickness and blood pressure demonstrated a relatively weak connection with the problem and demonstrated the ROC curve of the XGBoost using ADASYN method. The AUC of XGBoost is 0.84 as indicated in this figure – 4.

Then, the explainable AI methods are applied using SHAP and LIME models to know the way in which the model can predict the decision. The interesting feature importance of the XGBoost with ADASYN using explainable AI, SHAP library, is presented in Figure 4.

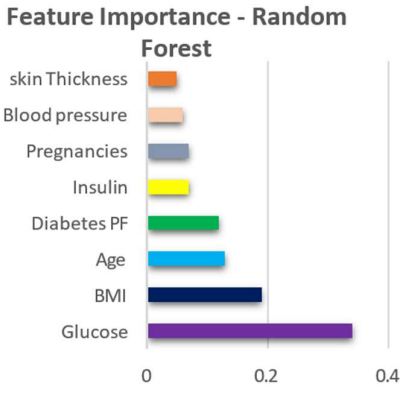

Figure 3: Features Importance - Random Forest

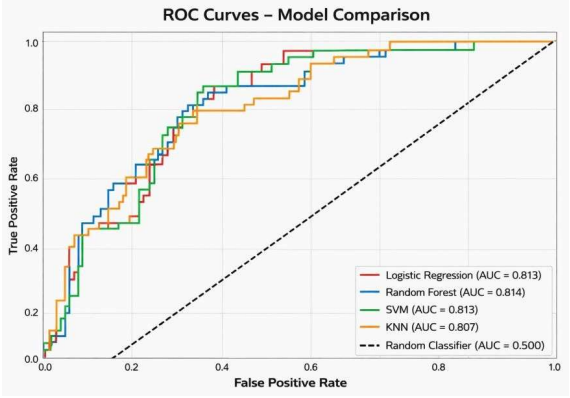

Figure - 4 : ROC Curves - Model Comparision

V. CONCLUSION

The current study was established with the intention of coming up with a system that can automatically determine the risk factors that can lead to diabetes and with an automated decision-support mechanism to help the medical practitioners in diagnosing diabetes using the risk factors. To fulfill this goal, different data preprocessing methods were used to improve the quality of data, predictive accuracy and development of reliable models. Furthermore, hyperparameter tuning was carried out using each model to identify the best parameter settings that would produce the best prediction accuracy.In-depth experiments were conducted on the processed Pima Indians Diabetes Dataset and the resulting models reviewed in the previous sections showed great performance in the various measures of evaluation such as accuracy, precision, sensitivity, specificity, F-measure and ROC/AUC score. The K-Nearest Neighbors (KNN) was found to be the most effective of the considered methods, performing better than other methods tested and also compared to several of the state-of-the-art methods reported in literature

REFERENCES

[1] A. T. Kharroubi, ‘‘Diabetes mellitus: The epidemic of the century,’’ World J. Diabetes, vol. 6, no. 6, p. 850, Jun. 2015.
[2] Q. Fu, R. Chen, S. Xu, Y. Ding, C. Huang, B. He, T. Jiang, B. Zeng, M. Bao, and S. Li, ‘‘Assessment of potential risk factors associated with gestational diabetes mellitus: Evidence from a Mendelian randomization study,’’ Frontiers Endocrinol., vol. 14, Jan. 2024, Art. no. 1276836.

[3] J.-M. Li, X. Li, L. W. C. Chan, R. Hu, T. Zheng, H. Li, and S.Yang, ‘‘Lipotoxicity-polarised macrophage-derived exosomes regulate mitochondrial fitness through Miro1-mediated mitophagy inhibition and contribute to type 2 diabetes development in mice,’’ Diabetologia, vol. 66, no. 12, pp. 2368–2386, Dec. 2023. Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, ‘‘Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention,’’ Int. J. Med. Sci., vol. 11, no. 11, pp. 1185–1200, 2014.
[4] D. Liang, X. Cai, Q. Guan, Y. Ou, X. Zheng, and X. Lin, ‘‘Burden of type 1 and type 2 diabetes and high fasting plasma glucose in Europe, 1990– 2019: A comprehensive analysis from the global burden of disease study 2019,’’ Frontiers Endocrinol., vol. 14, Dec. 2023, Art. no. 1307432.
[5] Y. Zhou, X. Chai, G. Yang, X. Sun, and Z. Xing, ‘‘Changes in body mass index and waist circumference and heart failure in type 2 diabetes mellitus,’’ Frontiers Endocrinol., vol. 14, Dec. 2023, Art. no. 1305839.
[6] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, ‘‘IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,’’ Diabetes Res. Clin. Pract., vol. 138, pp. 271–281, Apr. 2018.
[7] P. Peng, Y. Luan, P. Sun, L. Wang, X. Zeng, Y. Wang, X. Cai, P. Ren, Y. Yu, Q. Liu, H. Ma, H. Chang, B. Song, X. Fan, and Y. Chen, ‘‘Prognostic factors in stage IV colorectal cancer patients with resection of liver and/or pulmonary metastases: A population-based cohort study,’’ Frontiers Oncol., vol. 12, Mar. 2022, Art. no. 850937.
[8] Z. Ullah, F. Saleem, M. Jamjoom, B. Fakieh, F. Kateb, A. M. Ali, and B. Shah, “Detecting high-risk factors and early diagnosis of diabetes using machine learning methods,” Journal of Healthcare Engineering, vol. 2022, Article ID 6958202, 2022. DOI: 10.1155/2022/6958202.
[9] AlZu’bi, S.; Aqel, D.; Lafi, M. An intelligent system for blood donation process optimization-smart techniques for minimizing blood wastages. Clust. Comput. 2022, 25, 3617–3627. [Google Scholar] [CrossRef]
[10] Elbes, M.; Alrawashdeh, T.; Almaita, E.; AlZu’bi, S.; Jararweh, Y. A platform for power management based on indoor localization in smart buildings using long short-term neural networks. Abdoul Malik, & Cengiz Tepe. (2025). Improving Early Diabetes Diagnosis : A Machine Learning Approach for Class Imbalance Mitigation.
[11] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” Healthcare Technology Letters, vol. 10, no. 1–2, pp. 1–10, Dec. 2022, doi: 10.1049/htl2.12039.
[12] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.
[13] Dua, D., & Graff, C. (2019). Pima Indians Diabetes Database. machine learning at University of California, Irvine Repository. National Institute of Diabetes and Digestive and Kidney Diseases.
[14] Saru, G. S., & Subashini, J. (2019). The prediction of type 2 diabetes involves the application of competitive random forest and SMOTE hybrid. Journal of Medical Systems, 43(8), 1-12. RF-SMOTE mix error that was determined in this paper was 86.97.
[15] Zou, Q., Qu, K., Huang, Y., Liu, G., & Fu, R. (2018). Machine learning prognosticator of diabetes mellitus. Genetics 9, 515. doi.org/10.3389/fgene.2018.00515.