Application and Prediction of Machine Learning Algorithm in Predicting Diabetes Mellitus

Bingyu Ke

Faculty of Humanities and Social Sciences, University of Portsmouth, Hampshire, PO1 2SP, United Kingdom
Bingyu.Ke@Myport.ac.uk

Abstract. Diabetes mellitus (DM) is a major worldwide health problem since it is characterised by persistently elevated blood sugar levels. Predictive analysis of DM is crucial for early detection and prevention, optimal resource allocation, development of personalized treatment plans, cost reduction, and formulation of effective public health strategies. Based on data from Kaggle, this study assesses how well the algorithms Random Forest (RF), Support Vector Machines (SVM), and Logistic Regression (LR) predict diabetes. These machine learning (ML) algorithms are assessed for their accuracy, robustness, and overall utility in identifying diabetes risk factors and early detection of the disease. This paper employed these models to analyze a large dataset from Kaggle, assessing their predictive capabilities based on accuracy, sensitivity, specificity, and generalizability. The results indicated that RF outperformed other models with an Area Under The Curve (AUC) score of 0.96361, highlighting its robust predictive power. Significant predictors across all models included hemoglobin A1c (HbA1c) level, blood glucose level, age, body mass index (BMI), hypertension, and smoking history. Additionally, chronic periodontitis and lipid levels were identified as important factors influencing diabetes risk. This research emphasizes how crucial it is to use a variety of health markers when predicting DM in order to improve early diagnosis and treatment approaches, which will eventually improve patient outcomes and save healthcare expenditures.

Keywords: Diabetes Prediction, Machine Learning Algorithms, Random Forest, Logistic Regression, Support Vector Machines.

1 Introduction

DM is a group of endocrine diseases typically characterized by persistently elevated blood glucose levels and stands as a significant public health concern across the globe. The latest statistics from the International Diabetes Federation (IDF) for the year 2023 reveal a stark increase in both the prevalence and complexity of the disease, with the number of adults living with DM reaching nearly half a billion [1]. DM has several negative effects; in addition to lowering a patient's quality of life, it can cause life-threatening complications, such as cardiovascular disease, kidney disease, and retinopathy, which can significantly increase the risk of death. Therefore, strengthen-

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_22

ing the prediction and early diagnosis of DM is of great significance for preventing complications, reducing medical costs, and improving the survival rate of patients. This trend underscores an urgent need for advanced predictive tools that can facilitate early detection and proactive treatment strategies [2].

The creation of these technologies has been enabled by the advent of ML, which holds the potential to revolutionize the prediction and treatment of diseases. ML algorithms provide a viable way to detect people who are at risk of acquiring DM before they exhibit clinical signs because of their capacity to recognize complex patterns within large datasets. Numerous ML algorithms have shown to be successful in making DM predictions. The accuracy of the RF models was high, at 80.87% [3]. Neural networks were shown to be the best in other experiments, with accuracy rates as high as 96% [4] and 78.57% [5]. SVM and neural network fused models were utilized in another investigation to get 94.87% accuracy [6].

This study applies and evaluates three cutting-edge ML algorithms for DM prediction using current data from 2023. The algorithms used in this paper are SVM for managing complex patterns and high-dimensional data, LR for handling linear relationships with robustness, and RF, an ensemble technique that can handle a large number of predictors and capture non-linear interactions [7]. This study's main objective is to evaluate these algorithms' prediction power using the most recent data, paying particular attention to their generalizability, accuracy, sensitivity, and specificity. By doing this, we want to ascertain the best method for forecasting DM and pinpointing the most significant risk variables in light of the most recent advancements in medicine and worldwide health trends.

2 Data and methods

2.1 Dataset

The dataset of this article is diabetes-related data from Kaggle and contains 100,000 pieces of sample information. As shown in Table 1, several demographic and age factors were used to predict DM. The dataset provides ample data for training and evaluating predictive models, improving the statistical significance of the findings. Given the large number of samples, the dataset is likely balanced in terms of gender, age groups, and medical conditions, which enables more generalizable insights.

Table 1 Data source information

Attribute	Description
Gender	Indicates the gender of the individual (e.g., male or female).
Age	The age of the individual in years.
Hypertension	A binary indicator (0 or 1) showing whether the individual has hypertension.
Smoking History	A binary indicator (0 or 1) indicating if the individual has a history of heart disease.
BMI	The Body Mass Index of the individual, calculated from height and weight.
HbA1c Level	The level of glycated hemoglobin in the blood, which is indicative of blood sugar levels.
Blood Glucose Level	The current blood glucose level of the individual.
Diabetes	The target variable (0 or 1), indicating whether or not the individual has diabetes.

2.2 Model

The following is an overview of the three models used in this paper.

The statistical technique for binary classification is called logistic regression (LR), which calculates the likelihood of a binary outcome based on one or more predictors. It maps anticipated values to probabilities between 0 and 1 using the logistic function. LR works well for binary or multi-class classification problems [8]. It assumes linear relationships between features and the target. It is robust with large datasets and interpretable due to the coefficient.

The logistic function is as follows:

$$P(Y = 1|X) = \frac{1}{1+e(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+B_n X_n)} \quad (1)$$

Where βi are the coefficients learned from the data.

With several decision trees constructed and their forecasts combined, Random Forest (RF) is an ensemble learning technique that enhances overall performance.

It uses bagging (bootstrap aggregating) to create subsets of training data. Each subset is used to train a decision tree, and final predictions are made by averaging (regression) or majority voting (classification) across all trees. RFs handle large datasets and high-dimensional feature spaces well. They provide feature importance rankings and reduce the risk of overfitting compared to individual decision trees. They can be used for both classification and regression tasks [9].

As part of the integrated learning process, the RF algorithm constructs several decision trees during training and outputs the average prediction (regression) or class pattern (classification) for each tree. Combining the forecasts from each decision tree is the primary formula utilized by the RF algorithm.

For classification, the prediction ŷ for an input x is given by:

$$\hat{y} = \text{mode}\{h_{t(x)}\} \quad (2)$$

Where $h_t(x)$ is the prediction of the t-th decision tree.

For regression, the prediction ŷ for an input x is given by:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (3)$$

Where T is the total number of trees in the forest, and $h_t(x)$ is the prediction of the t-th decision tree.

The goal of the supervised learning method SVM is to localize hyperplanes in the feature space most suitable for classification. The margin between data points of various classes is maximized. The hyperplane that maximizes the distance to the closest training points—also known as support vectors—is selected to divide the classes. A kernel trick is applied to convert data into a higher-dimensional space in non-linear scenarios. When using kernel functions, it is efficient for high-dimensional spaces and non-linear boundaries. SVMs need to be carefully tuned because they are sensitive to the kernel and parameter choices. Although they might be computationally demanding for large datasets, they perform well for binary and multi-class problems [10].

2.3 Evaluation Index

To assess the performance of the classification models in this study, several evaluation metrics are employed to provide comprehensive insights into various aspects of model effectiveness. The selected metrics include Accuracy, Precision, Recall, F1 Score, AUC-ROC, Confusion Matrix, and Log Loss.

Accuracy represents the proportion of correctly predicted instances out of all instances. Suitable if the classes are balanced. It can be misleading for imbalanced datasets.

The formula is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

The percentage of true positive forecasts among all positive predictions is known as precision. Beneficial when the expense of false positives is significant (e.g., medical diagnosis).

The formula is as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

Recall represents the proportion of actual positives correctly predicted. Essential when the cost of false negatives is high (e.g., missing a diagnosis).

The formula is as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

The F1 score represents the reconciled mean of accuracy and recall, balancing the two metrics. It is ideal when both accuracy and recall are important and there is a category imbalance.

The formula is as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

The AUC-ROC plots true positive and false positive rates at different thresholds to determine the ability of the model to distinguish between categories. The improved model is represented by a higher AUC, where 1.0 indicates perfection and 0.5 indicates random guessing. It is useful to evaluate the performance of binary classifiers.

A Confusion Matrix is a table that displays the counts of actual versus predicted classifications. Helps visualize classification results and calculate other metrics.

Log Loss represents the performance of a classifier by comparing predicted probabilities against actual class labels. Suitable when the output is probabilistic (e.g., in logistic regression).

The formula is as follows:

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (8)$$

3 Results

Model Performance. Table 2 shows the performance of SVM, LR, and RF models for predicting diabetes. The table shows that RF has the highest overall performance with an AUC score of 0.96361, an accuracy of 0.97, and the best balance between accuracy and recall for positive and negative classes. SVM achieved a high AUC score of 0.92626 and an accuracy of 0.97 but showed lower recall for the positive class, indicating a tendency to miss some positive cases. LR provided the fastest evaluation time with an AUC score of 0.96129 and an accuracy of 0.96 but had lower precision and recall for the positive class compared to RF.

Table 2 Classification Model Performance Comparison

Metric	SVM	Logistic Regression	Random Forest
AUC Score	0.92626	0.96129	0.96361
Accuracy	0.97	0.96	0.97
Precision (0)	0.97	0.97	0.97
Precision (1)	0.95	0.86	0.96
Recall (0)	1.00	0.99	1.00
Recall (1)	0.64	0.61	0.68
F1-score (0)	0.98	0.98	0.98
F1-score (1)	0.76	0.72	0.80
Elapsed Time	148.96s	0.10s	2.88s
Log Loss	0.2488	0.3226	0.2132

Data Visualization Results of Logistic Regression. The confusion matrix shown in Fig. 1 provides a detailed breakdown of the performance of the LR model by showing the counts of actual versus predicted classifications. The confusion matrix in Fig. 1 shows that the LR model performs well in identifying negative instances with a high number of true negatives (27,206) and a relatively low number of false positives (247). However, the model had more difficulty in correctly identifying positive instances, with a higher number of false negatives (985) than true positives (1562). This suggests that while the model is effective in excluding non-positive instances, there is room for improvement in correctly predicting positive instances, which may depend on the key of the application.

Overall, the confusion matrix shows areas where the model's prediction performance could be improved and offers insightful information about the model's advantages and disadvantages.

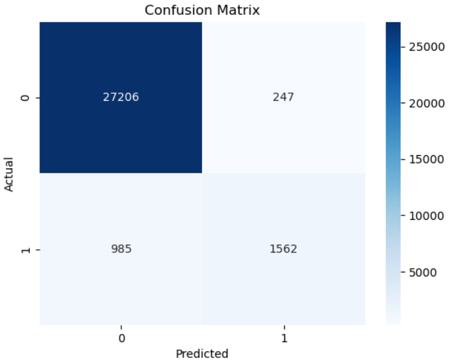

Fig. 1. Confusion Matrix (Logistic Regression)

The ability of the LR model to distinguish between positive and negative categories is evidenced by the receiver operating characteristic (ROC) curve, which is based on the false positive rate at different threshold settings, as shown in Fig. 2. One way to assess the overall performance of the model is to use the AUC. Plotting the true positive rate (sensitivity) versus the false positive rate (1-specificity) for various thresholds is called an ROC curve.

The AUC is 0.96, indicating a high level of model performance.

The ROC curve in Fig. 2 shows that the LR model has strong discriminatory power, with high true positive rates and low false positive rates within a certain threshold range. The model seems to be very successful in discriminating between positive and negative categories with an AUC of 0.96. The test is more accurate if the curve closely matches the top and left boundaries of the ROC space. On the other hand, if the curves are close to the diagonal or 45-degree line, the test is less accurate.

Overall, the high AUC values indicate that the LR model performs very well in distinguishing between categories, making it a reliable tool for prediction tasks.

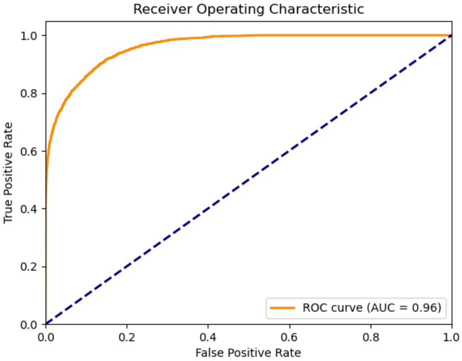

Fig. 2. ROC Curve (Logistic Regression)

The feature importance graph shown in Fig. 3 illustrates the significance of various features in the LR model by displaying the magnitude of their coefficients. This helps to understand which features have the most impact on the model's predictions. The feature importance graph in Fig. 3 indicates that HbA1c level and blood glucose level are the most influential features in predicting the target variable, which is consistent with medical understanding of diabetes indicators. Age, heart disease, and hypertension also play significant roles, though to a lesser extent. Features like BMI, gender, and smoking history have relatively lower importance in the model's predictions.

This analysis helps to prioritize features in further model improvements and provides insights into which factors are most critical in the context of the prediction task.

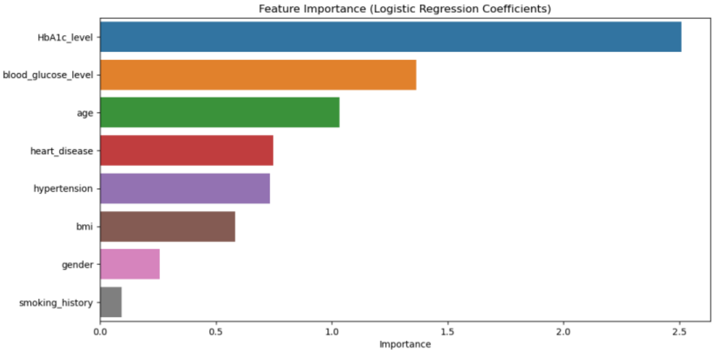

Fig. 3. Feature Importance (Logistic Regression Coefficients)

Data Visualization Results of Decision Tree. Fig. 4's confusion matrix, which compares the counts of actual and projected classifications, offers a thorough analysis of the Decision Tree (DT) model's performance. The confusion matrix in Fig. 4 shows that, with a very high number of genuine negatives (27,373) and a very low number of false positives (80), the DT model is effective at identifying negative cases. With more true positives (1,742) than false negatives (805), the model also performs rather well in identifying positive situations.

Overall, the confusion matrix suggests that the DT model has a strong ability to correctly classify both negative and positive instances, although there is still some room for improvement in reducing the number of false negatives. This performance makes the DT model a reliable tool for classification tasks in this context.

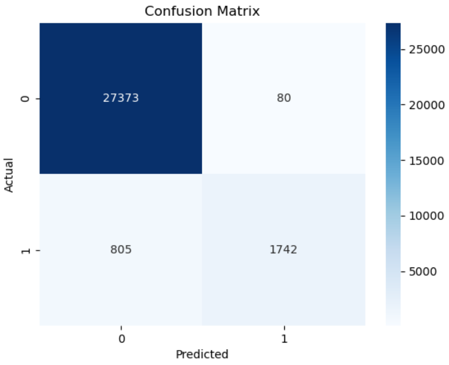

Fig. 4. Confusion Matrix (Decision Tree)

The feature importance plot shown in Fig. 5 illustrates the importance of various features in the RF model by showing the magnitude of their importance scores. This helps to understand which features have the greatest impact on the model's predictions. The feature importance plot in Fig. 5 shows that HbA1c level and blood glucose level are the features that have the greatest impact on predicting the target variable, which is consistent with the medical understanding of diabetes indicators. BMI and age also play an important role, albeit to a lesser extent. Characteristics such as smoking history, hypertension, heart disease, and gender were of relatively low importance in model predictions.

This analysis helps to prioritize features in further model improvements and provides insights into which factors are most critical in the context of the prediction task.

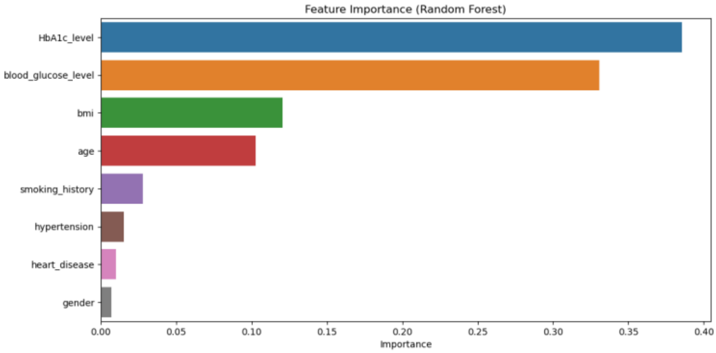

Fig. 5. Feature Importance (Random Forest)

The ability of the RF model to distinguish between positive and negative classes is evidenced by the ROC curve, which corresponds to the rate of false positives at the different thresholds in Fig. 6. One way to assess the overall performance of the model is to use the AUC. The ROC curve in Fig. 6 shows that the RF model has a strong ability to discriminate across a range of thresholds, with a high rate of true positives and a relatively low rate of false positives. The model appears to be very successful in distinguishing between positive and negative categories with an AUC of 0.96. The test is more accurate if the curves closely match the top and left boundaries of the ROC space. On the other hand, if the curve is close to the diagonal or 45-degree line, the test is less accurate.

Overall, the RF model performs remarkably well in differentiating between the classes, as evidenced by the high AUC value, which makes it a trustworthy tool for prediction tasks.

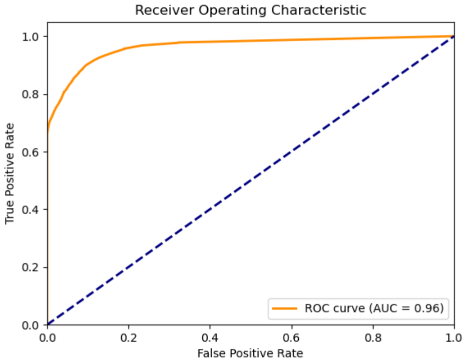

Fig. 6. ROC Curve (Random Forest)

Data Visualization Results of Support Vector Machine. Fig. 7's confusion matrix, which compares the counts of actual and projected classifications, offers a thorough analysis of the SVM model's performance. This matrix facilitates comprehension of the model's ability to discriminate between the positive and negative classifications. Fig. 7's confusion matrix, which has a very high number of true negatives (27,369) and a very low number of false positives (84), shows that the SVM model is effective at identifying negative instances. With more true positives (1,631) than false negatives (916), the model also does reasonably well in recognizing positive situations.

Overall, the confusion matrix suggests that the SVM model has a strong ability to correctly classify both negative and positive instances, although there is still some room for improvement in reducing the number of false negatives. This performance makes the SVM model a reliable tool for classification tasks in this context.

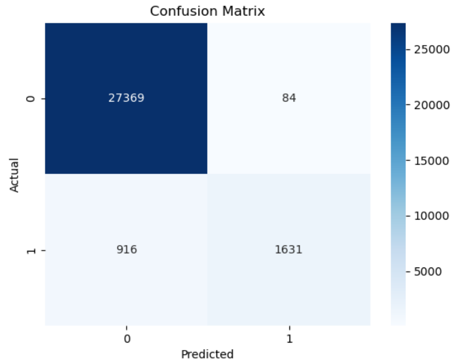

Fig. 7. Confusion Matrix (SVM)

The ability of the SVM model to distinguish between positive and negative categories is shown in the ROC curve in Fig. 8, which plots the true-positive versus false-positive rates for different threshold settings. One way to assess the overall performance of the model is to use the AUC. The ROC curve in Fig. 8 shows the good discriminative ability of the SVM model, with a high true positive rate and a low false positive rate over a range of thresholds. The model seems to be successful in distinguishing between positive and negative categories, as shown by its AUC of 0.92. The test is more accurate if the curve closely matches the top and left boundaries of the ROC space. On the other hand, if the curve is close to the diagonal or 45-degree line, the test is less accurate.

Overall, the high AUC values indicate that the SVM model performs well in distinguishing between categories, making it a reliable tool for prediction tasks.

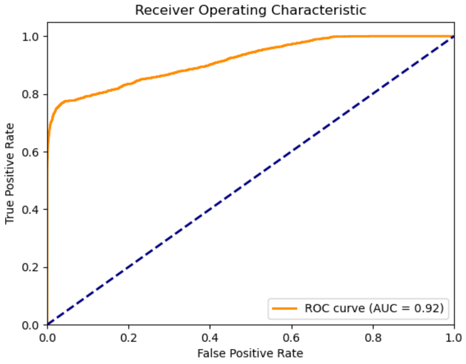

Fig. 8. ROC Curve (SVM)

The feature importance graph shown in Fig. 9 illustrates the significance of various features in the SVM model by displaying the magnitude of their importance scores. This helps to understand which features have the most impact on the model's predictions. The feature importance graph in Fig. 9 indicates that blood glucose level and

HbA1c level are the most influential features in predicting the target variable, which aligns with medical understanding of diabetes indicators. Age and BMI also play significant roles, though to a lesser extent. Features like hypertension, heart disease, smoking history, and gender have relatively lower importance in the model's predictions.

This analysis helps to prioritize features in further model improvements and provides insights into which factors are most critical in the context of the prediction task.

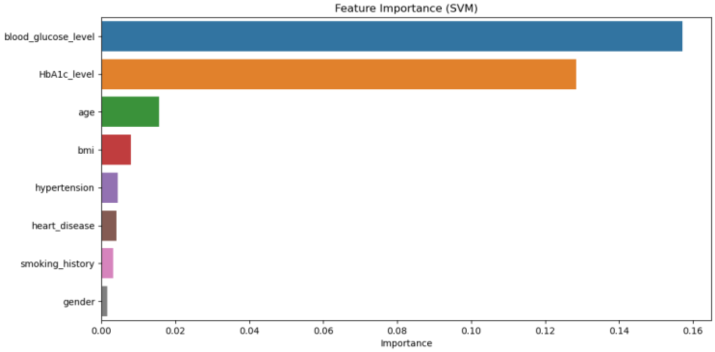

Fig. 9. Feature Importance (SVM)

4 Conclusion

The evaluation of different models, including LR, DT, RF, and SVM, reveals varying levels of performance across key metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. The feature importance analysis highlights that blood glucose level and HbA1c level consistently emerge as the most significant predictors across models, underscoring their critical role in diabetes prediction. While models like RF and SVM demonstrate high AUC scores, indicating strong discriminatory power, there remains room for improvement in reducing false negatives, particularly for models like SVM. This suggests a need for further optimization and possibly integrating additional relevant features to enhance predictive accuracy and reliability. Overall, these insights provide a valuable foundation for refining model performance and improving diabetes prediction outcomes.

In terms of feature importance, HbA1c level and blood glucose level consistently emerged as the most significant predictors of diabetes across all models, aligning with established medical understanding. Age was found to be a significant predictor, with older age increasing the likelihood of diabetes. BMI was also shown to be a crucial factor, with higher BMI values correlating with increased diabetes risk. Hypertension and smoking history, while less influential than HbA1c and blood glucose, still contributed significantly to the predictive models.

Additionally, there are some studies that identified a significant correlation between abnormal glucose metabolism indicators (like fasting glucose and HOMA-IR) and the severity of chronic periodontitis, suggesting that oral health could be an important factor in diabetes prediction [11]. However, due to the limitations of the data, the impact of other diseases on diabetes could not be analyzed.

References

1. Hossain, Md. J.; Al-Mamun, Md.; Islam, Md. R.: Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Science Reports , 7 (3), (2024).
2. Sugandh, F.; Chandio, M.; Raveena, F.; Kumar, L.; Karishma, F.; Khuwaja, S.; Memon, U. A.; Bai, K.; Kashif, M.; Varrassi, G.; Khatri, M.; Kumar, S.: Advances in the Management of Diabetes Mellitus: A focus on personalized medicine. Curēus (2023).
3. Haner E. N. , Kırğıl, B. and Ayyıldız T. E. : Predicting Diabetes Using Machine Learning Techniques, 2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE), Ankara, Turkey, 137-141, (2022).
4. Ma, J.: Machine Learning in Predicting Diabetes in the Early Stage, 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 167-172,(2020).
5. Alzboon, M. S. : A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes, 2023 IEEE Tenth International Conference on Communications and Networking (ComNet), Hammamet, Tunisia, 1-12,(2023).
6. Ahmed , U. et al.:Prediction of Diabetes Empowered With Fused Machine Learning, in IEEE Access, 10, 8529-8538, (2022).
7. Samet, S.; Laouar, M. R.; Bendib, I.; Eom, S. Analysis and prediction of diabetes disease using machine learning methods. International Journal of Decision Support System Technology, 14 (1), 1–19 (2022).
8. Sperandei, S. Understanding logistic regression analysis. Biochemia Medica, 12–18, (2014).
9. Khan, A. A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems With Applications, 244, 122778, (2014).
10. Geeksfor G.: Support Vector Machine (SVM) Algorithm. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/.
11. Choi, Y.-H.; McKeown, R. E.; Mayer-Davis, E. J.; Liese, A. D.; Song, K.-B.; Merchant, A. T.: Association between periodontitis and impaired fasting glucose and diabetes. Diabetes Care, 34 (2), 381–386, (2011).