International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST5231057

Machine Learning-Based Diabetic Risk Prediction Model for Early Detection

Mohammed Khalid Hossen[1*], Anika Tabassum[2], Dr. Jannatul Ferdaus[3]
[1]Department of Computer Science and Engineering, Sylhet Agricultural University, Bangladesh.
[2]Department of Computer Science and Engineering, American International University of Bangladesh.
[3]Department of Medicine, IBN SINA Medical College, Dhaka, Bangladesh.
*Corresponding Author : khalid@sau.ac.bd

ARTICLEINFO

Article History:
Accepted: 30 Sep 2023
Published: 18 Oct 2023

Publication Issue
Volume 10, Issue 5
September-October-2023
Page Number
532-542

ABSTRACT

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and effective management of diabetes can significantly reduce the risk of complications and improve the quality of life of individuals with diabetes. In recent years, machine learning techniques have been applied to predict the risk of diabetes and to develop personalized treatment plans. In this study, we propose a machine learning-based diabetic risk prediction model for early detection and management. The proposed model uses various clinical and demographic variables such as age, gender, BMI, blood pressure, and fasting blood glucose levels to predict the risk of developing diabetes. We evaluated the performance of the proposed models using a dataset of patients with diabetes and non-diabetic individuals. Machine learning techniques including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Random Forest are evaluated using the confusion matrices. The experimental results show that the Random Forest classifier achieved an accuracy of 80%, sensitivity of 82%, specificity of 80% in predicting the risk of diabetes. However, Increasing the accuracy rates of machine learning algorithms to 90% to 100% will be the challenging part of this study.

Keywords: Machine Learning Techniques, Diabetes, Linear Regression, Support Vector Machine, K-Nearest-Neighbors, Random Forest, Confusion Matrix.

I. INTRODUCTION

Diabetes is a major public health concern worldwide, affecting over 400 million individuals, and its prevalence is projected to increase in the coming years. Early detection and management of diabetes are critical in preventing complications such as heart disease, stroke, blindness, and kidney disease

(American Diabetes Association, 2020). Machine learning (ML) algorithms have shown great potentials in predicting the risk of developing diabetes in individuals. Several studies have investigated the use of machine learning algorithms for diabetic prediction.

A systematic review by Liu et al. (2021) evaluated the performance of machine learning algorithms in predicting the risk of diabetes. The review included 21 studies and found that machine learning algorithms achieved an accuracy ranging from 71% to 90% in predicting the risk of diabetes. The review concluded that machine learning algorithms have great potential in diabetic risk prediction and can assist in early detection and management of diabetes [1].

One such study by Khanam, J et al. (2021) proposed a machine learning-based diabetics risk prediction model using demographic, clinical, and lifestyle factors. The study used logistic regression, decision tree, random forest, and neural network algorithms to predict the risk diabetes in individuals. The model achieved an accuracy of 88.6% in predicting the risk of diabetes, demonstrating the potential of machine learning algorithms in diabetic risk prediction [2]. Another study by Aishwarya Mujumdar et al. (2019) proposed a machine learning-based diabetic risk prediction model using clinical and lifestyle factors. The study used decision tree, random forest, and support vector machine algorithms to predict the risk of diabetes in individuals. The model achieved an accuracy of 96% in predicting the risk of diabetes, demonstrating the potential of machine learning algorithms in diabetic risk prediction [3].

In order to predict the risk of diabetes in Chinese patients, Zhao et al. (2019) developed a deep learning technique. In identifying the risk of diabetes, the model obtained an accuracy rate of 82.1% [4]. An investigation by Deberneh et al. (2021) of a machine learning-based model for predicting the risk of type 2 diabetes using electronic health records and lifestyle data. the model was 73.2% accurate in terms of forecasting the chance of diabetes [5]. Other study by Kavakiotis et al. (2017) developed a machine learning-based model for prediction the risk of type 2 diabetes using electronic health records. The accuracy of 84.9% achieved by the model in predicting the risk of diabetes of a patient [6].

Overall, the literatures suggest that machine learning algorithms can be effective in predicting the risk of diabetes in individuals. The proposed machine learning-based diabetic risk prediction model can help healthcare providers make informed decisions for patient care and assist in the early detection and management of diabetes.

In this study, we will construct a machine learning-based diabetic risk prediction model, which will help us to diagnose the diabetes dataset. Patient medical records and attributes are gathered from the UCI repository. Then the dataset is considered for determining whether the patients have diabetes or not. To analyze the diabetes dataset, this article considers several attributes of the individuals. To continue this study, this article considers attributes of individuals such as age, gender, BMI, blood pressure, fasting blood glucose levels, etc. Four different machine learning-based techniques, such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Random Forest, have been used to classify and predict diabetes. Under these ML algorithms, the attributes of the provided dataset are trained. Some advanced beneficial results are shown in this research. A comparative result has been shown regarding the rate of accuracy. All the considered machine learning-based algorithms in this article are showing efficient based on their rate of accuracy, sensitivity, and specificity. But the most efficient algorithm in this study is Random Forest, which gives an accuracy rate of approximately 80%.

This article is rearranged sequentially. The methodology is discussed in Section 2. The basic concept of machine learning algorithms has been discussed briefly in Section 3. Section 4 displays the results analysis. In this section, all the algorithms are compared regarding the confusion matrix and the rate of accuracy. Finally, a conclusion and the future scope of this study have been shown in Section 5.

II. METHODS AND MATERIAL

Data collection, data preprocessing, feature selection, model selection, model training, model evaluation, model optimization, and deployment are some of the procedures involved in developing a machine learning-based diabetic risk prediction model for early detection and management. Fig. 1. Here is an in-depth explanation of each step:

Architecture of Machine Learning Technique Procedures:

Data Collection: The first step to developing a model to predict diabetic risk is to gather pertinent information. Demographic details, medical history, the outcomes of clinical exams, and laboratory test findings should all be included in this material. It is crucial to make sure the statistics are accurate representations of the population under study.

Data Pre-processing: Once the data is collected, it needs to be pre-processed. This involves cleaning the data, removing missing values, and transforming it into a format that can be used by machine learning algorithms. Pre-processing is critical as it can impact the accuracy of the model.

Feature Selection: In this step, relevant features that can predict the risk of diabetes are selected. This is important because using irrelevant features can lead to overfitting, which can result in poor performance of the model. Feature selection can be done using techniques such as correlation analysis, principal component analysis, and recursive feature elimination.

Model Selection: An appropriate machine learning algorithm is then picked for the task after the features have been chosen. Numerous methods are available, including K-Nearest Neighbors, decision trees, random forests, and logistic regression. The nature of the problem and the data at hand determine which algorithm is best.

Model Training: The preprocessed data is used as a training set for the chosen algorithm. In order to train the model to make predictions, a subset of the data is called the training set. Gradient descent, back-propagation and support vector machines are just a few of the approaches used to train the model.

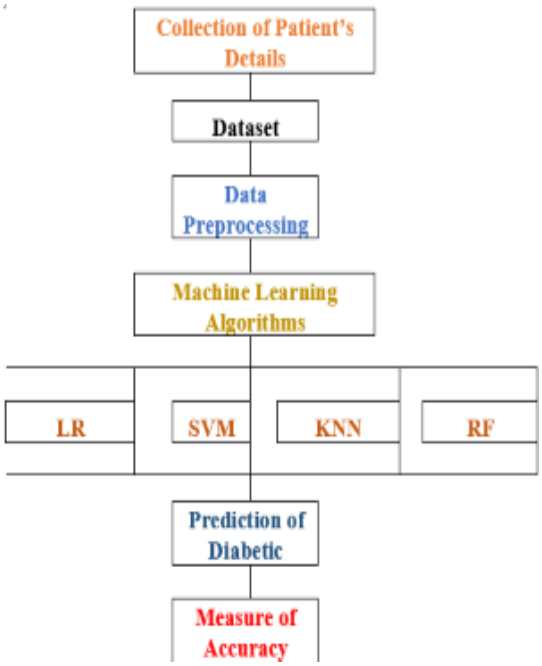
Model Evaluation: A test set must be used to evaluate the model once it has been trained. The data utilized for the test set is a portion of the data that wasn't used for training. Accuracy, precision, recall, and F1 score are some of the measures used to evaluate the model's performance. The evaluation is essential to define the effectiveness of the model in predicting diabetic risk.

Model Optimization: The performance of the model can be developed by optimizing its parameters or by using more advanced techniques such as ensemble learning or hyperparameter tuning. To obtain the best performance, the model is modified during the optimization process.

Deployment: The model can be deployed in the real world when it has been trained and optimized. This involves incorporating the model into an application that can be used by healthcare professionals to estimate the risk of diabetes in their patients. Thorough documentation, monitoring, and evaluation of the model's efficiency must occur simultaneously with the deployment. In conclusion, building a machine learning-based diabetic risk prediction

model for early detection and management is a difficult process that calls for careful thought about data gathering, preprocessing, feature selection, model selection, model training, model evaluation, and model deployment. To make sure the model is accurate, dependable, and effective for early detection and control of diabetes, a partnership between healthcare specialists, data scientists, and software engineers is necessary.

Table 1. Architecture of prediction models



III.Machine Learning-Based Algorithms

Machine learning algorithms are a collection of mathematical and statistical methods used to train models to recognize patterns and forecast future events. In supervised learning, each sample has a known target value, therefore the model is trained on labeled data. By minimizing a predetermined loss function, the model learns to translate input data to output values [7, 8, 9]. Support vector machines, logistic regression, K-Nearest Neighbors, and random forests are a few examples of supervised learning algorithms. Four alternative machine learning (ML) methods are covered in this article to get the highest accuracy.

Logistic Regression: Logistic regression is a supervised learning algorithm used for binary classification tasks, where the goal is to predict the probability of an instance belonging to a particular class (e.g., yes or no, true or false, 0 or 1).

The logistic regression algorithm works by modeling the relationship between the input features and the output variable using a logistic function. The logistic function, also known as the sigmoid function, is an S-shaped curve that maps any input value to a value between 0 and 1. The logistic function is defined as:

$$g(z) = 1/(1 + e^{-z})$$

where z is a linear combination of the input features and their associated weights:

$$z = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

The logistic regression model learns the weights w_0, w_1, w_2, ... ,w_n that minimize a loss function, such as the binary cross-entropy loss, see mathematical details in [10, 11, 12, 13]. The weights are updated iteratively using an optimization algorithm, such as gradient descent, until the model converges to a minimum value of the loss function.

Once the model is trained, it can be used to predict the probability of an instance belonging to a particular class. The output of the logistic regression model has a probability value between 0 and 1, which can be thresholder to make a binary classification decision. A commonly used threshold is 0.5, where values greater than or equal to 0.5 are classified as positive (belonging to the target class) and values less than 0.5 are classified as negative (not belonging to the target class).

Support Vector Machine: A supervised learning algorithm called the support vector machine (SVM) is utilized for classification and regression problems.

The goal of SVM is to identify the appropriate decision boundary for classifying the input data.

Building a hyperplane in a high-dimensional space is how the SVM algorithm works. It is chosen that the hyperplane maximally separates the two groups of data. Support vectors are utilized to define the hyperplane and are the points closest to the hyperplane. The SVM algorithm attempts to maximize the margin, which is the separation between the hyperplane and the support vectors [14, 15, 16, 17].

The input data is transformed into a higher dimensional space using a method known as the kernel trick by SVM in situations where a hyperplane cannot linearly separate the data. The SVM can implicitly transfer the input data into a higher dimensional space by using the kernel trick, which entails replacing the dot product of the input data with a kernel function.

SVM can be used for both binary and multi-class classification tasks. For binary classification, the SVM algorithm learns a hyperplane that separates the two classes. For multi-class classification, the SVM algorithm learns multiple hyperplanes to separate the various classes. One common popular method for multi-class classification with SVM is the one-vs-one, where a different binary classifier is trained for every pair of classes.

SVM is widely utilized in many fields, including bioinformatics, text classification, and picture classification. They have a reputation for their proficiency in working with non-linear data as well as their capacity to manage high-dimensional data. SVM can be computationally expensive to train on large datasets, and they can be sensitive to the regularization parameter and kernel function selection.

K-Nearest Neighbors: A supervised machine learning technique called k-Nearest Neighbors (k-NN) is utilized for classification and regression problems [18, 19, 20, 21].

A data point in k-NN is identified or predicted according to the category or value of its k closest neighbors. Based on their distance from the given data point, the algorithm calculates the k nearest neighbors. This distance can be calculated using a variety of metrics, including the cosine distance, Manhattan distance, and Euclidean distance.

In classification tasks, the data point is classified according to the majority of the k nearest neighbors, whereas in regression tasks, the value for the data point is predicted using the average of the k nearest neighbors.

One of the strengths of k-NN is that it is a simple and easy-to-understand algorithm that requires no training process. However, it can be computationally expensive when working with large datasets or when calculating distances using complex metrics. It is also sensitive to the choice of k and the distance metric used and may not perform well in high-dimensional spaces.

Random Forests: For both classification and regression issues, the Random Forest (RF) algorithm is a common supervised machine learning method [22, 23, 24, 25]. It is an ensemble learning technique that constructs a number of decisions trees and integrates the results of each to produce predictions.

In a Random Forest, several decision trees are built using randomly selected portions of the training data, and the combined forecasts of all the trees are then used to make the final prediction. In the Random Forest, each decision tree is trained using a different subset of the data, which lowers correlation across the trees and enhances performance.

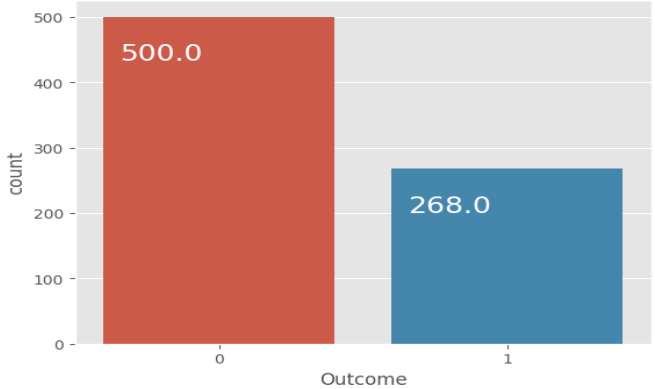
Large datasets with high dimensionality can be handled by the Random Forest technique, which is renowned for its accuracy and robustness. More intricate non-linear correlations between features and objectives can be captured, and it is less prone to overfitting than a single decision tree. It can also offer estimations of feature importance, which can be helpful in the feature selection process.

IV. RESULT ANALYSIS

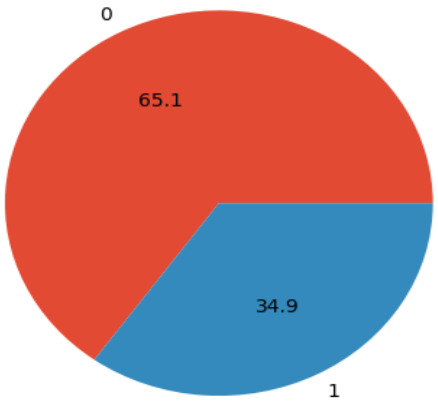
The analysis of a Machine Learning-Based Diabetic Risk Prediction Model for Early Detection would

typically involve evaluating the model's performance on a test dataset and interpreting the results.

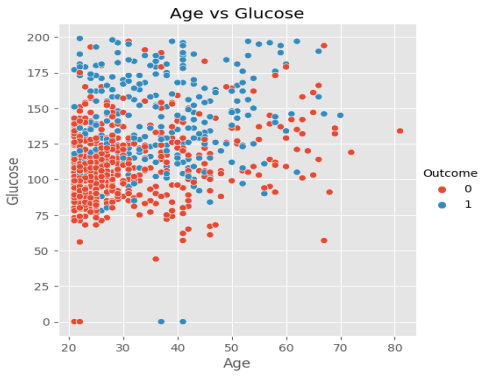
The evaluation metrics used to analyze the model's performance would depend on the specific problem being addressed and the type of machine learning algorithm used. For a classification problem, common evaluation metrics include accuracy, precision, recall, and F1 score. The analysis of a Machine Learning-Based Diabetic Risk Prediction Model for Early Detection would aim to assess the model's performance, identify any areas for improvement, and provide insights into the underlying factors contributing to diabetic risk. It is crucial to evaluate the model's performance on a diverse set of patients and populations to ensure its generalizability and usefulness in clinical practice.



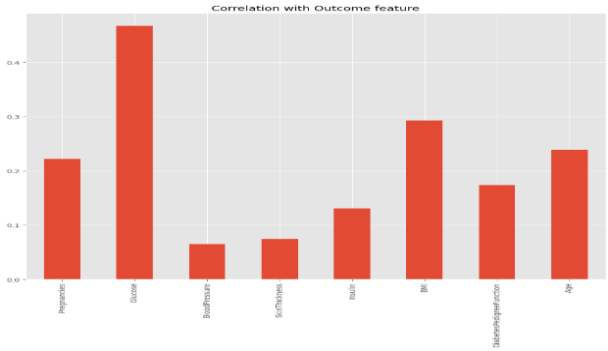
(a)



(b)



(c)



(d)

Figure 1. (a) Target class for diabetic dataset, (b) Percentage of non-diabetic (‘0’) and diabetic (‘1’), (c) Age v/s Glucose with the Outcome feature, and (d) Correlation with the Outcome feature

Analysis of Diabetes Dataset

Now, we will focus on the analysis of the features of the diabetes dataset. As we can see in figure 1(a), the total number of observations in the outcome feature is 768, where not having diabetes 500, denoted by ‘0’,

and having diabetes 268, represented by '1'. So, the percentage of not having diabetes is 65.1% and the percentage of diabetes number is 34.9%, see figure 1(b). It is clearly depicted that the rate of non-diabetic numbers is higher than the rate of diabetic numbers. The total number of features in the diabetic dataset is 08. In this dataset, Age and Glucose are two important features with respect to Outcome target. As in figure 1(c), age between 20 to 30 and the glucose level between 75 to 120 is considered for the non-diabetic patients, where diabetic is showing when the age between 40 to 55 and glucose between 125 to 200. In figure 1(d), it shows the correlation efficiency between the features with respect to Outcome column. As we can see in figure 1(d), Glucose, BMI and Age have a strong correlation with the Outcome feature; however, BloodPresure has a less correlation among them.

Performance Analysis

In this paper, four different machine learning techniques like Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest-Neighbors (KNN), Random Forest (RF) classifiers are studied broadly to predict the diabetes. The accuracy rate of technique has been measured and detected the high accuracy of the selected algorithm. The accuracy rate is predicted through the ratio of the correct prediction to the total number of a given data set. Mathematically it is written as,

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Where, TP, TN, FP, and FN consider as True Positive, True Negative, False Positive, and False Negative, respectively. After diagnosing the considered ML techniques for training and testing dataset, we can predict the best technique by defining the high accuracy rate, see figure 2. But the rate of accuracy is calculated with the help of confusion matrices. As we can see in the figure 2 and 3, the Random Forest and Logistic Regression algorithms gives approximately the same accuracy; however, RF is the best one to compare with the others machine learning algorithms, where RF has an accuracy of nearest 80%.

We will focus more by studying the ML algorithms through precision, recall and f1-score, see figure 4. Precision, recall, and F1-score are three commonly used evaluation metrics in machine learning and information retrieval to assess the performance of classification or binary decision-making models. These metrics are used to evaluate how well a model can predict the correct class labels.

Precision: Precision is a measure of the accuracy of the positive predictions made by the model. It is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP). Precision is often interpreted as the ability of the model to minimize false positives, or in other words, the ability to correctly identify positive samples without misclassifying negative samples.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: Recall, also known as sensitivity or true positive rate (TPR), is a measure of the ability of the model to correctly identify all the positive samples, including both true positives and false negatives (FN). Recall is often interpreted as the ability of the model to minimize false negatives, or in other words, the ability to correctly identify positive samples without missing any.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-score: F1-score is the harmonic mean of precision and recall and provides a balanced measure of both precision and recall. It is often used when both precision and recall are important and need to be considered together. F1-score is calculated as,

F1-score = 2 * (Precision * Recall)/ (Precision + Recall)

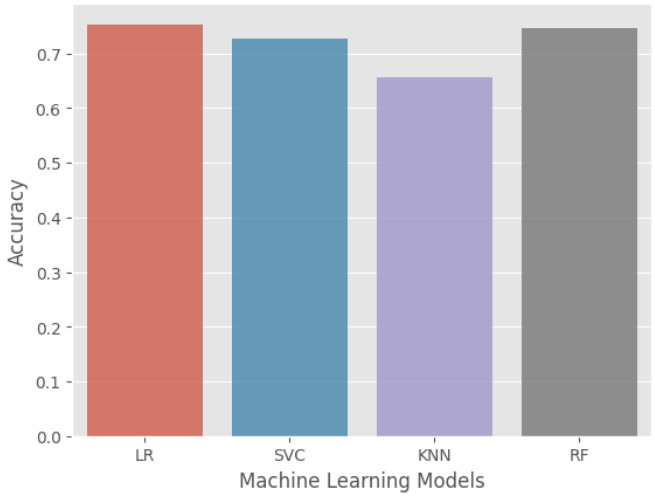
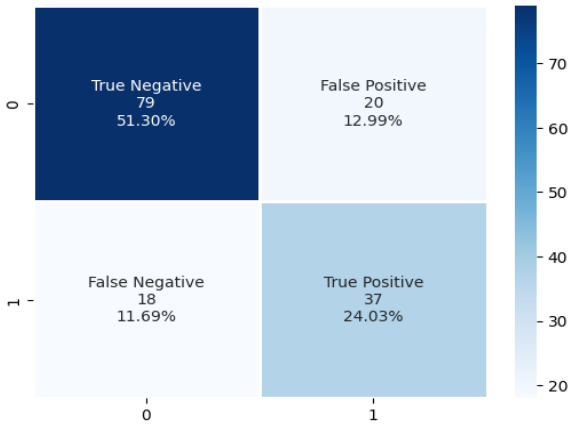
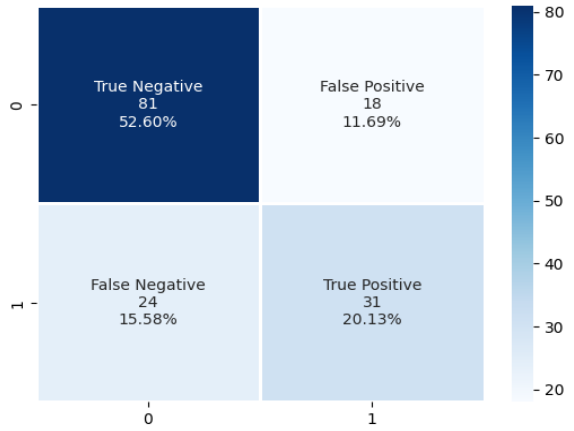


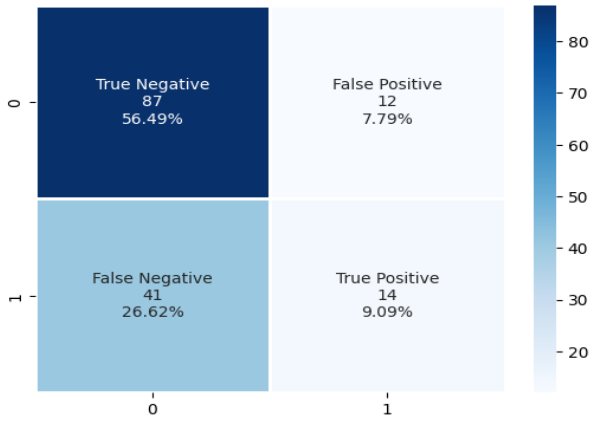
Figure 2. Accuracy comparison of machine learning techniques



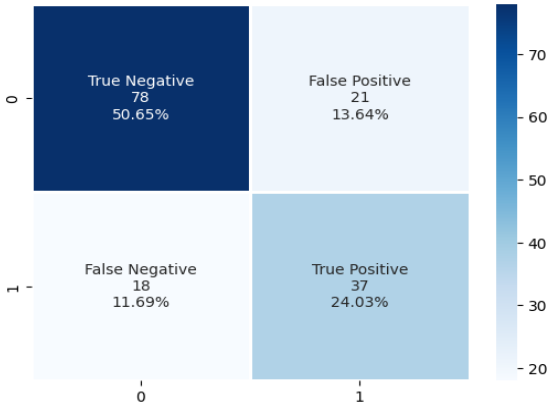
(a) LR



(b) SVM



(c) KNN



(d) RF

Figure 3. Confusion matrix of the machine learning algorithms

F1-score balances the trade-off between precision and recall, and a higher F1-score indicates a better overall performance of the model. It is a commonly used metric in machine learning to evaluate the effectiveness of a model's performance in binary classification tasks.

After evaluating the rate of precision, recall, and f1-score, we compare these measurements among the four machine learning models. As we can see in figure 4, RF and LR algorithms show reasonable score which is almost 81% in all categories. In the final section we have shown a comparison sensitivity and specificity of through ROC curve for the LR and RF algorithms. Both models show a similar performance in figure 5.

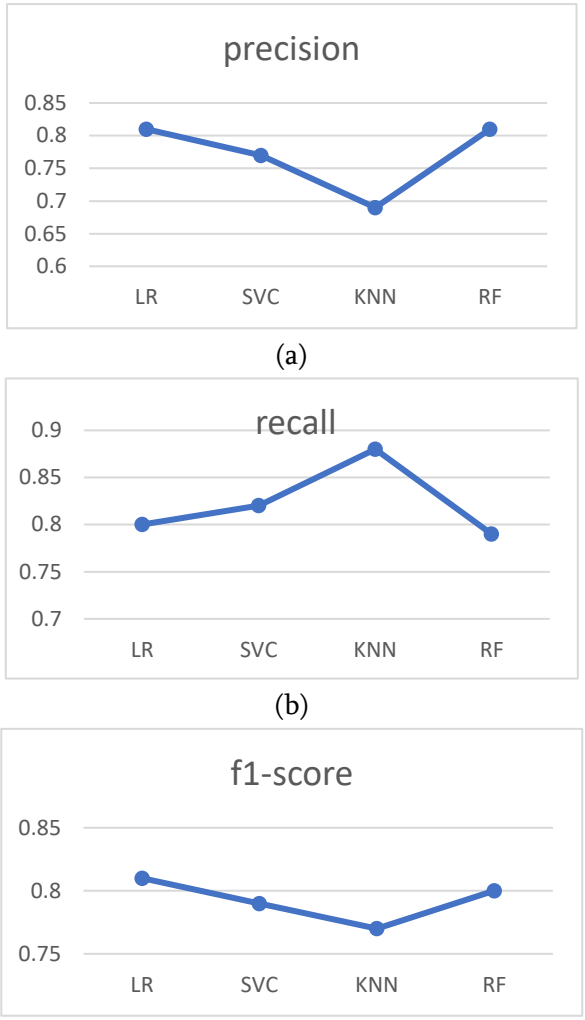


Figure 4. Comparisons of (a) precision, (b) recall, and (c) f1-score among the machine learning techniques

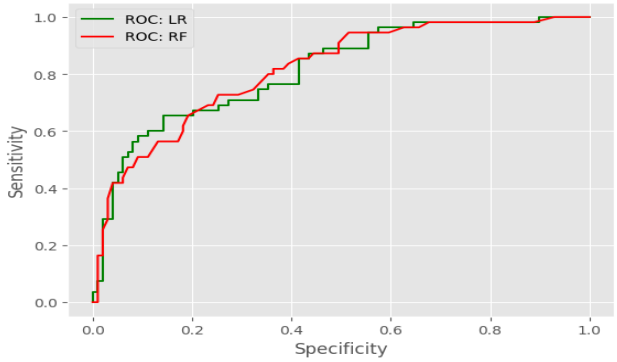


Figure 5. Comparison of sensitivity and specificity of through ROC curve.

V. CONCLUSION

In conclusion, the development of a machine learning-based diabetic risk prediction model for early detection holds great promise in improving the identification and management of individuals at risk of diabetes. The use of precision, recall, and F1-score as evaluation metrics can help assess the performance of such models and guide further improvements. The model's ability to accurately predict the risk of diabetes can potentially enable early intervention and preventive measures, leading to better patient outcomes and reduced healthcare costs. For this reason, this article focuses on the diabetes prediction based on the accuracy rate of confusion matrices. The statistics of the four algorithms are used to estimate the accuracy rate through confusion matrices. When four algorithms are compared, it is found that Random Forest algorithm has been showed the best performance with high accuracy rate. So, the accuracy rate of RF model is 80% approx., which indicates that machine learning algorithms will be considered as a predefined tool to diagnose the diabetes in the near future. Other statistical measurement such as precision, recall, and F1-score have calculated for 82%, 80%, and 81%, respectively.

VI. ACKNOWLEDGEMENTS

I would like to thank my research assistant Anika and Jannatul, for their support and patience throughout the research process. We focused on this research to learn how to shape the data in machine learning algorithms, especially for diabetic. Their contributions to the data collection and analysis were instrumentals in the success of this research.

VII. REFERENCES

[1]. Song, X., Liu, X., Liu, F., & Wang, C. (2021). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. International journal of medical informatics, 151, 104484.

[2]. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432-439.

[3]. Aishwarya Mujumdar, V Vaidehi (2019), Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019, Pages 292-299, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.

[4]. Yuancheng Zhao, Lijun Zhao, Yiting Wang, Junlin Zhang, Honghong Ren, Rui Zhang, Yucheng Wu, Yutong Zou, Nanwei Tong, Fang Liu. (2023) The association of plasma NT-proBNP level and progression of diabetic kidney disease. Renal Failure 45:1.pc

[5]. Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. International Journal of Environmental Research and Public Health. 2021; 18(6):3317. https://doi.org/10.3390/ijerph18063317.

[6]. KavakiotisI. Tsave O. Salifoglou A. Maglaveras N. Vlahavas I. Chouvarda I. 2017.Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 15:104–16.

[7]. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[8]. Hossen, M. K. (2022). Heart Disease Prediction Using Machine Learning Techniques. American Journal of Computer Science and Technology, 5(3), 146-154.

[9]. Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, Hafiz Husnain Raza Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", Journal of Healthcare Engineering, vol. 2021, Article ID 9930985, 17 pages, 2021. https://doi.org/10.1155/2021/9930985.

[10]. El Morr, C., Jammal, M., Ali-Hassan, H., El-Hallak, W. (2022). Logistic Regression. In: Machine Learning for Practical Decision Making. International Series in Operations Research & Management Science, vol 334. Springer, Cham. https://doi.org/10.1007/978-3-031-16990-8_7.

[11]. Bisong, E., & Bisong, E. (2019). Logistic regression. Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners, 243-250.

[12]. Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5-6), 352-359.

[13]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[14]. Wang, L. (Ed.). (2005). Support vector machines: theory and applications (Vol. 177). Springer Science & Business Media.

[15]. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

[16]. Zhang, Y. (2012). Support vector machine classification algorithm and its application. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3 (pp. 179-186). Springer Berlin Heidelberg.

[17]. Noble, W. S. (2004). Support vector machine applications in computational biology. Kernel methods in computational biology, 71, 92.

[18]. Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 1-19.

[19]. Kadhm, M. S., Ghindawi, I. W., & Mhawi, D. E. (2018). An accurate diabetes prediction system

based on K-means clustering and proposed classification approach. International Journal of Applied Engineering Research, 13(6), 4038-4041.

[20]. Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, 5(4).

[21]. Reddy, S. K., Krishnaveni, T., Nikitha, G., & Vijaykanth, E. (2021, September). Diabetes Prediction Using Different Machine Learning Algorithms.

Cite this article as :

Mohammed Khalid Hossen, Anika Tabassum, Dr. Jannatul Ferdaus, "Machine Learning-Based Diabetic Risk Prediction Model for Early Detection ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 5, pp. 532-542, September-October 2023. Available at doi : https://doi.org/10.32628/IJSRST5231057
Journal URL : https://ijsrst.com/IJSRST5231057