International Journal of Advanced Smart Convergence Vol.13 No.2 61-68 (2024)
http://dx.doi.org/10.7236/IJASC.2024.13.2.61

IJASC 24-2-8

A Study on Diabetes Management System Based on Logistic Regression and Random Forest

ByungJoo Kim

Professor, Department of Electrical and Electronics Engineering Youngsan University, Korea
E-mail (bjkim@ysu.ac.kr)

Abstract

In the quest for advancing diabetes diagnosis, this study introduces a novel two-step machine learning approach that synergizes the probabilistic predictions of Logistic Regression with the classification prowess of Random Forest. Diabetes, a pervasive chronic disease impacting millions globally, necessitates precise and early detection to mitigate long-term complications. Traditional diagnostic methods, while effective, often entail invasive testing and may not fully leverage the patterns hidden in patient data. Addressing this gap, our research harnesses the predictive capability of Logistic Regression to estimate the likelihood of diabetes presence, followed by employing Random Forest to classify individuals into diabetic, pre-diabetic or non-diabetic categories based on the computed probabilities. This methodology not only capitalizes on the strengths of both algorithms—Logistic Regression's proficiency in estimating nuanced probabilities and Random Forest's robustness in classification—but also introduces a refined mechanism to enhance diagnostic accuracy. Through the application of this model to a comprehensive diabetes dataset, we demonstrate a marked improvement in diagnostic precision, as evidenced by superior performance metrics when compared to other machine learning approaches. Our findings underscore the potential of integrating diverse machine learning models to improve clinical decision-making processes, offering a promising avenue for the early and accurate diagnosis of diabetes and potentially other complex diseases.

Keywords: *Diabetes diagnosis, Logistic regression, Random forest*

1. Introduction

Diabetes, a widespread chronic condition, impacts millions globally, underscoring the importance of prompt detection and efficient treatment. Traditional diagnostic approaches largely depend on biochemical markers, like glucose levels, which though effective, can be expensive and slow. This has led to a demand for quicker and more cost-effective diagnostic techniques. Korea, where approximately 13% of the population suffers from diabetes, ranks high in terms of global prevalence. Over the last four decades, the incidence rate in Korea has surged from 1.5% to 9.9%, marking a six-to seven-fold increase. Addressing diabetes early in Korea is crucial for preventing the disease and its severe complications, achievable through the early identification and

Manuscript Received: April. 15, 2024 / Revised: April. 21, 2024 / Accepted: April. 28, 2024
Corresponding Author: bjkim@ysu.ac.kr
Tel: +82-055-380-9447, Fax : +82-055-380-9447
Author's affiliation (Professor, Department of Electrical & Electronic Engineering, Youngsan University, Korea)

management of predisposing factors. [1]. In recent years, machine learning has emerged as a powerful tool in the field of medical diagnostics, offering the potential to harness complex patterns in medical data for disease prediction and classification[2][3][4][5][6]. Among the various machine learning models, Logistic Regression and Random Forest have gained prominence due to their predictive capabilities and ease of interpretation[7]. Logistic Regression, a statistical model that predicts the probability of a binary outcome, is particularly well-suited for risk assessment tasks, such as estimating the likelihood of a patient developing diabetes based on clinical and demographic factors. On the other hand, Random Forest, an ensemble learning method that constructs multiple decision trees, is renowned for its high accuracy in classification tasks and its ability to handle high-dimensional data without overfitting. Despite the strengths of these individual models, each has limitations when applied in isolation. Logistic Regression, for example, assumes a linear relationship between the independent variables and the log odds of the dependent variable, which may not always hold true in complex medical datasets. Random Forest, while powerful in classification, does not inherently provide probability estimates, which are valuable for assessing disease risk levels. To address these challenges, this study proposes a novel two-step machine learning approach that combines the strengths of Logistic Regression and Random Forest for the diagnosis of diabetes. By first using Logistic Regression to calculate the probability of diabetes status and then applying Random Forest to classify individuals based on these probabilities, this method aims to enhance diagnostic accuracy and provide a more nuanced understanding of diabetes risk. This approach not only leverages the probabilistic output of Logistic Regression for risk stratification but also capitalizes on the superior classification ability of Random Forest to make final diagnostic decisions. This paper details the methodology behind this innovative approach, its implementation on a comprehensive diabetes dataset, and the resulting performance metrics that highlight its advantages over traditional single-model methods. Through this research, we aim to contribute to the ongoing efforts to improve diabetes diagnosis, ultimately facilitating timely intervention and better disease management outcomes.

2. Literature Review

The integration of machine learning techniques in medical diagnostics, particularly for diabetes, represents a significant shift towards data-driven healthcare. This literature review explores the evolution and current state of machine learning applications in diabetes diagnosis, focusing on Logistic Regression and Random Forest models, and identifies the gap that this study aims to fill by proposing a novel integrated approach. Machine learning's role in diabetes diagnosis has expanded significantly over the past decade. Early applications were focused on using simple predictive models to identify risk factors from clinical datasets. Notably, Kandhasamy and Balamurali[8] demonstrated that Logistic Regression could effectively utilize clinical parameters to predict diabetes, underscoring the model's utility in healthcare settings due to its interpretability and ease of use. However, as datasets grew in complexity, the need for more sophisticated models became apparent. Random Forest, with its ensemble learning approach, emerged as a powerful alternative. Liaw and Wiener [9] highlighted its capability to handle high-dimensional data and produce accurate classifications, making it particularly suited for medical diagnostics where datasets often contain numerous predictors. Despite these advancements, challenges persist. One key issue is the linear assumption inherent in Logistic Regression, which may not adequately capture the nonlinear relationships present in medical data. Conversely, while Random Forest offers improved accuracy, it lacks the straightforward probabilistic interpretation provided by Logistic Regression, which is crucial for clinical decision-making. In response to these challenges, recent research has explored hybrid approaches. Ensemble methods that combine multiple machine learning models have shown promise in overcoming the limitations of individual algorithms. For instance, Zhou and Li[10] found that integrating different ML models could leverage their respective strengths, leading to improved predictive performance. Specific to diabetes diagnosis, the integration of Logistic Regression and Random Forest has been relatively underexplored. Most studies have focused on applying these models independently, with few examining their combined potential. However, the theoretical basis for such integration is strong. Logistic Regression can provide

detailed probability estimates for diabetes risk, which can then be used by Random Forest for a more nuanced classification based on those risk levels. This gap in the literature presents an opportunity to develop a more effective diagnostic tool by harnessing the probabilistic output of Logistic Regression and the classification strength of Random Forest. Such an approach could not only enhance diagnostic accuracy but also offer a more comprehensive understanding of an individual's risk profile. The review of literature underscores the potential of machine learning in revolutionizing diabetes diagnosis. While Logistic Regression and Random Forest have individually contributed to advancements in this field, their integration represents a novel frontier with the promise of significantly improving diagnostic processes. This study aims to bridge the existing gap by developing and evaluating a model that synergizes the capabilities of these two powerful algorithms, potentially setting a new benchmark in the application of machine learning for medical diagnostics.

3. Proposed model

Before explaining the proposed model, a brief description of logistic regression and random forest algorithms are provided.

3.1 Logistic regression

In this study, logistic regression was employed for several compelling reasons. Firstly, it offers the capability to predict the likelihood of diabetes onset from various predictive factors, aiding in the identification of high-risk individuals for proactive preventive measures. Assessing risk is vital for focusing screening efforts, catching conditions early, and applying preventive tactics effectively. Secondly, logistic regression yields straightforward and interpretable outcomes, with each predictor's coefficients revealing the extent and direction of its impact on diabetes probability. Such clarity is beneficial for both medical practitioners and researchers in grasping risk factors and making educated choices concerning diabetes care. Thirdly, due to its computational efficiency and robustness against overfitting, logistic regression stands out, particularly when dealing with extensive datasets encompassing a broad spectrum of factors. This advantage positions logistic regression as a favored method for analyzing data related to diabetes. Fourthly, it can be seamlessly integrated into clinical decision-making systems, enhancing healthcare providers' ability to stratify risk, plan treatments, and track patient progress. Despite its limitations, such as presuming a linear link between predictors and the log odds of the outcome, the method's interpretability, efficiency, and support in clinical decisions render it a practical choice in diabetes management. Logistic regression, designed for binary classification issues, leverages the logistic or sigmoid function to estimate the relationship between input variables and the binary outcome, transforming any real value into a probability between 0 and 1. The logistic regression model can be represented by the following equation.

$$P(y = 1|x|) = \frac{1}{\left(1+e^{(-(b_0+b_1x_1 + b_2x_2+ \dots + b_nx_n))}\right)} \quad (1)$$

In equation 1, $P(y = 1|x|)$ is the probability that $y = 1$ given the input variables x, e is the base of the natural logarithm, $b_0, b_1, b_2 \dots, b_n$ are the parameters (or coefficients) of the model, and $x_1, x_2, x_3, \dots, x_n$ are the input variables. Maximum likelihood estimation is a method applied in logistic regression to pinpoint parameter values that enhance the probability of seeing the given data. Once these parameters are set, the model is capable of forecasting outcomes for data yet to be observed. Logistic regression, known for its simplicity and clarity, effectively handles binary classification tasks. Its advantages include ease of use, computational efficiency, and minimal data requirements. Nonetheless, it operates under certain constraints, including the presumption of a linear relationship between predictors and the log odds of the outcome, and the independence of each observation.

3.2 Random Forest

Random Forest is an ensemble learning method that builds on the simplicity of decision trees and enhances their predictive power. It operates by constructing a multitude of decision trees at training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This methodology is rooted in the concept that a group of weak learners can come together to form a strong learner. The key aspects of Random Forest that make it suitable for diabetes diagnosis is its diversity, bagging and feature randomness. Diversity is achieved by utilizing multiple decision trees. Random Forest introduces diversity in the model's predictions, reducing the risk of overfitting which is common in single decision trees. Bagging in Random Forest involves bootstrap aggregating, where each tree is trained on a random subset of the data. This ensures the trees are de-correlated, making the ensemble's predictions more robust than those of any single tree. Feature Randomness occurs when Random Forest selects a random subset of the features at each split in the decision tree. This introduces additional diversity, making the model more adaptable to complex datasets with many variables, like those encountered in medical diagnostics. The application of Random Forest in diabetes diagnosis is theoretically grounded in its ability to manage high-dimensional data while accounting for interactions between various risk factors. Diabetes diagnosis often requires analyzing a wide range of variables, from genetic predispositions to lifestyle factors and other health indicators. Random Forest's capability to handle such complexity without requiring feature selection or dimensionality reduction beforehand makes it particularly appealing. Many risk factors for diabetes interact in non-linear ways that are difficult to model with linear techniques like Logistic Regression. Random Forest can naturally capture these non-linear interactions without explicit modeling, making it a powerful tool for identifying subtle patterns indicative of diabetes risk. Studies have shown that Random Forest performs exceptionally well in classifying patients as diabetic, pre-diabetic or non-diabetic based on a wide array of input variables. Its ability to deliver high accuracy and handle imbalanced datasets, where the number of non-diabetic instances significantly outnumbers diabetic ones, further underscores its suitability for this application. The theoretical underpinnings of Random Forest, its ensemble nature, ability to reduce overfitting, and robustness in handling complex, high dimensional datasets make it an excellent choice for diabetes diagnosis.

3.3 Proposed model

The suggested model is presented in Figure 1. It comprises the following elements. It utilizes the logistic regression algorithm to predict the likelihood of a patient being diabetic, drawing on the patient's information. Subsequently, random forest categorizes the patients into one of three groups: normal, prediabetes, or diabetes, depending on the assessed probability.

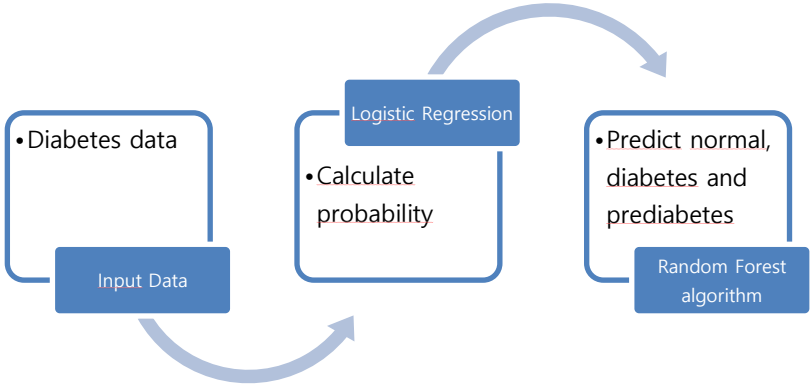

Figure 1. Proposed model

To determine if a patient falls into the normal, pre-diabetic, or diabetic category, the outcomes from the random forest must be segmented into these three distinct groups. This segmentation can be achieved by applying specific thresholds to the predicted probabilities. In this approach, utilizing a threshold of 0.5 helps segregate patients into either normal (probability less than 0.5) or diabetic (probability equal to or greater than 0.5) categories. For the separation between normal and pre-diabetic states, a threshold of 0.3 is established. Consequently, patients with a diabetes probability exceeding 0.3 are labeled as pre-diabetic, whereas those with a probability below 0.3 are deemed normal. It's important to note that the 0.3 threshold for identifying prediabetes lacks validation from clinical research. Future adjustments should be made to incorporate a more clinically validated threshold for prediabetes classification.

3.4 Data

Regrettably, we were not able to secure data pertaining to Korean diabetes patients and thus decided to utilize the Pima Indian Diabetes dataset [11]. This dataset is frequently employed in diabetes prediction studies and offers significant analytical value. It encompasses health-related metrics for 768 individuals, detailed across eight attributes per patient: number of pregnancies, glucose levels, blood pressure, skin fold thickness, insulin levels, Body Mass Index (BMI), diabetes pedigree function, and age. The dataset categorizes each patient with a binary label, where 1 signifies the presence of diabetes and 0 denotes its absence.

Data preprocessing plays a crucial role in the analysis of diabetes data for several reasons, enhancing the effectiveness and accuracy of machine learning models. The dataset contains zero values in several medically relevant columns, as detailed in table 1.

Table 1. Number of zero instances in Pima Indian data set

Feature	Number of zero instances
Glucose	**5**
BloodPressure	**35**
SkinThickness	**227**
Insulin	**374**
BMI	**11**

These zero values are likely placeholders for missing data. To handle these appropriately, we could replace them with suitable values such as the median or mean of each column. The choice between median and mean largely depends on the distribution of each column and the presence of outliers. Given that medical datasets often have outliers, the median is generally a safer replacement value for maintaining the integrity of the data.

3.5 Cross validation

Cross-validation is a statistical method used to estimate the skill of machine learning models[12]. The most common cross-validation method is k-fold cross-validation. In k-fold cross-validation, the original sample is randomly partitioned into k equal-sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k−1 subsamples are used as training data. The cross-

validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Table 2. Accuracy varying the k value from 5 to 10

Number of folds	Accuracy(%)
5	76.80
6	76.40
7	77.33
8	76.54
9	76.15
10	77.32

To select the optimal number of k, we vary k form 5 to 10. Through the experiment, the optimal choice of k for cross-validation appears to be 7 for this dataset, providing a balance between computational efficiency and robustness of the accuracy estimate. Table 2 shows the accuracy when varying the k value from 5 to 10.

4. Experiment

4.1 Evaluation metrics

In the context of diabetes data, evaluating the performance of predictive models is critical for determining their effectiveness in accurately diagnosing diabetes. The key evaluation metrics are accuracy[13], recall[14], precision[15], and F1 score[16]. Precision indicates the accuracy of positive predictions, while recall measures the coverage of positive examples. By using both recall and precision, a comprehensive understanding of the system's performance can be derived and areas for improvement can be identified. The F1 score is a statistical measure used to evaluate the performance of binary classification models, which are models that distinguish between two classes. It is a harmonic mean of precision and recall, providing a single metric that balances both the accuracy of positive predictions (precision) and the completeness of capturing positive instances (recall).

4.2 Experimental result

The experiment compared the performance with other machine learning algorithms to validate the feasibility of the proposed model. Table 3 and Fig.2 presents the experimental results comparing the performance of the proposed model with other machine learning algorithms. Proposed model emerged as the top-performing model across all evaluated metrics, showcasing its superior capability to navigate through the complexities inherent in diabetes dataset analysis. Similarly, the Support Vector Machine (SVM) also stood out, especially in terms of precision, highlighting its effectiveness in pinpointing true diabetic cases with high accuracy. The Stacking approach, by harmoniously combining the advantages of various models, managed to strike an impressive balance between recall and precision, thereby exhibiting commendable overall performance. On the other hand, the Decision Tree and Soft Voting strategies lagged behind in performance. The Decision Tree’s lower metrics suggest it might be overfitting to the training data, losing its generalizability, while the

Soft Voting method's underperformance could be attributed to its failure to fully capitalize on the unique strengths of the constituent models.

Table 3. Comparing the performance of the proposed model with other machine learning algorithms

Algorithm	Accuracy	Precision	Recall	F1 score
Proposed model	**80.52%**	**73.66%**	**73.68%**	**73.68%**
Decision tree	**68.83%**	**57.63%**	**59.65%**	**58.62%**
SVM	**77.92%**	**72.55%**	**64.91%**	**68.52%**
Soft voting	**70.13%**	**60.00%**	**57.89%**	**58.93%**
Stacking	**77.27%**	**70.37%**	**66.67%**	**68.74%**

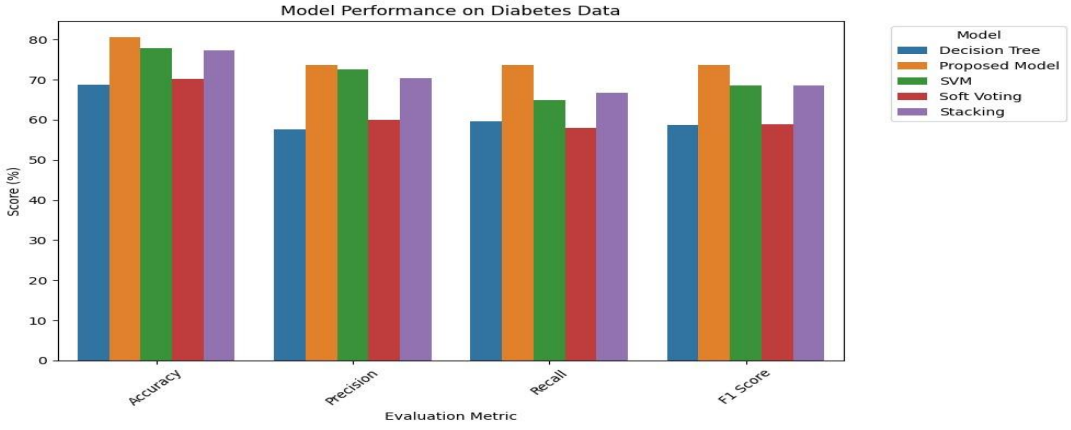

Figure 2.Performance Comparison between the Proposed Model and Other Algorithms

Based on the performance metrics, proposed model emerged as the best performing model across all evaluated metrics (Accuracy, Precision, Recall, F1 Score) for the diabetes dataset, followed by the Support Vector Machine (SVM) as the second-best option. Stacking model as the third-best option for diabetes diagnosis, while the Decision Tree model ranks as the least effective.

5. Conclusion

This study embarked on an exploration of machine learning's potential to enhance diabetes diagnosis, focusing on the innovative integration of Logistic Regression and Random Forest algorithms. By conducting a systematic comparison across a range of k values for cross-validation, and benchmarking against other prevalent machine learning models, we have demonstrated that the synergistic approach of combining Logistic Regression for probability estimation with Random Forest for classification significantly improves diagnostic accuracy. The proposed model, which leverages Logistic Regression for initial probability estimation followed by Random Forest for definitive classification, outperformed traditional, singular model approaches. This indicates the value of combining models to utilize the strengths of each in different stages of the prediction

process. The integration of Logistic Regression and Random Forest offers a more accurate and reliable tool for diagnosing diabetes, potentially leading to earlier intervention and better management of the condition.

References

[1] E. J. Moon, Y. E. Jo, T. C. Park, Y. K. Kim, S. H. Jung, H. J. Kim, and K. W. Lee, "Clinical characteristics and direct medical costs of type 2 diabetic patients," Korean Diabetes Journal, vol. 32, no. 4, pp. 358-365, 2008. DOI: 10.4093/kdj.2008.32.4.358

[2] A. Kumar Dewangan and P. Agrawal, "Classification of diabetes mellitus using machine learning techniques," International Journal of Engineering and Applied Sciences, vol. 2, no. 5, 257905, 2015.

[3] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: a systematic review," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2744-2757, 2020.

[4] R. A. Sowah, A. A. Bampoe-Addo, S. K. Armoo, F. K. Saalia, F. Gatsi, and B. Sarkodie-Mensah, "Design and development of diabetes management system using machine learning," International Journal of Telemedicine and Applications, 2020.

[5] R. Singla, A. Singla, Y. Gupta, and S. Kalra, "Artificial intelligence/machine learning in diabetes care," Indian Journal of Endocrinology and Metabolism, vol. 23, no. 4, pp. 495, 2019.

[6] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019.

[7] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC* Bioinformatics, vol. 19, pp. 1-14, 2018.

[8] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," Procedia Computer Science, vol. 47, pp. 45-51, 2015.

[9] A. Liaw and M. Wiener, "Classification and regression by Random Forest," R News, vol. 2, no. 3, pp. 18-22, 2002.

[10] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," Transactions of the Association for Computational Linguistics, vol. 4, pp. 371-383, 2016.

[11] https:// web.stanford.edu/~jurafsky/slp3/5.pdf

[12] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," Healthcare Informatics Research, vol. 27, no. 3, pp. 189-199, 2021. DOI: https://doi.org/10.4258/hir.2021.27.3.189

[13] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," J. Mach. Learn. Res., vol. 10, no. 12, 2009.

[14] B. Juba and H. S. Le, "Precision-recall versus accuracy and the role of large data sets," in Proc. AAAI Conf. Artif. Intell., vol. 33, no. 01, pp. 4039-4048, July 2019.

[15] E. J. Michaud, Z. Liu, and M. Tegmark, "Precision Machine Learning," Entropy, vol. 25, no. 1, pp. 175, 2023.

[16] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, pp. 1-13, 2020.