Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 1 Iss. 1 February 2024, pp: 41-46
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Comparison of Logistic Regression, Random Forest and Adaboost Algorithms for Diabetes Mellitus Classification

Alfi Syahri[1*], Umi Fariha[2], Rival Afandi[3], Intan Nurliyana[4]

[1,2,3]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]Department Information Management, Faculty Collage of Computing Informatics and Mathematics,
MARA University, Malaysia

Email: [1]12050322472@students.uin-suska.ac.id, [2]12050314787@students.uin-suska.ac.id,
[3]120500326126@students.uin-suska.ac.id, [4]2022454274@student.uitm.edu.my

Received Dec 27th 2023; Revised Jan 24th 2024; Accepted Feb 20th 2024
Corresponding Author: Alfi Syahri

Abstract

Diabetes mellitus is a chronic disease that affects the way the body regulates sugar (glucose). High blood sugar levels can lead to health complications including heart problems, eye disorders, nerve damage, kidney and blood vessel disorders. It is important for early detection of diabetes by utilizing data mining technology. Data mining has various classification models that can be used to detect diabetes, including logistic regression, random forest and adaboost. The comparison of the three algorithms aims to find out which algorithm is most appropriate in the classification of diabetes. From the results obtained, the random forest algorithm has the best performance in the classification of diabetes mellitus compared to other algorithms.

Keyword: Adaboost, Classification, Diabetes Mellitus, Logistic Regression, Random Forest

1. INTRODUCTION

Diabetes mellitus is a group of chronic diseases that affect the way the body regulates sugar (glucose) [1]. High blood sugar levels over a long period of time can lead to various health complications, including heart problems, eye disorders, nerve damage, kidney disorders and blood vessel problems [2][3].

According to World Health Organization (WHO), the number of people with diabetes in the world in 2022 reached 442 million. Thereport shows an increase in people with diabetes mellitus from the previous year [4]. IDF estimates that diabetics in 2030 will increase to 578 million people and by 2045 will reach 700 million people [5]. Because the frequency of diabetes is increasing, this disease must be identified in order to reduce the risk of health problems. To identify these problems, data mining technology is needed [6].

Data mining is the process of finding patterns or information in certain data using appropriate techniques or methods [7]. Data mining has several functions, including: association function, classification function, clustering function, prediction function, estimation function [8]. To predict the right category or label requires the application of a classification pattern or function. Classification is the grouping or categorization of data based on certain classes and attributes to produce new information [9][10].

To perform classification, a data mining method or algorithm is needed. This research uses Logistic Regression, Random Forest and Adaboost algorithms. Of the three algorithms, a data mining comparison is carried out to find out which algorithm has high accuracy in the classification of diabetes mellitus.

Based on research entitled "A Comparative Study of Different Machine Learning Tools in Detecting Diabetes", a comparative study of various machine learning algorithms in detecting diabetes through the Pima Indians dataset is discussed. The four algorithms evaluated were Gradient Boosting, Support Vector Machine, AdaBoost, and Random Forest, using all features in the dataset and features selected using the Minimum Redundancy Maximum Relevance, Feature Selection algorithm. The best results were obtained using the Random Forest algorithm with an accuracy of 99.35%. In addition, this study also discusses previous research on diabetes prediction models using various machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), LogitBoost, and others [11].

In 2021, Saloni Kumari from the Department of Electronics and Communication Engineering at Bharati Vidyapeeth Institute of Technology in New Delhi, India, conducted a comparison of algorithms for diabetes. In this study, different classification algorithms were used for comparison. classification algorithms include

Logistic Regression, KNN, SVM, Naive Bayes, Soft Voting Classifier, and several other algorithms. In this study, the soft voting classifier achieved the highest accuracy rate (79.08%) [12].

Research conducted by Yitayeh Belsti et al, with the title Comparison of Machine Learning (ML) and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash Gestasional Diabetes Mellitus (GDM) Machine learning model This study found that the Monash GDM Model, a ML model, has better predictive performance than traditional logistic regression models in predicting the development of Gestational Diabetes Mellitus. The model uses data such as age, Body Mass Index (BMI), history of GDM, family history of diabetes, previous history of adverse obstetric events, and ethnicity. Involving more than 48,000 pregnant women at Monash Health obstetric hospitals in Australia, the study used various ML and logistic regression techniques in the development and validation of the model. Theresults showed that the use of ML can improve the accuracy of GDM prediction, providing potential in the prevention and management of this condition [13].

This study uses the updated diabetes dataset from Kaggle. Kaggle is a platform that organizes various kinds of data competitions to solve complex problems using data analysis and machine learning techniques. The dataset can be accessed at https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset. In this dataset there are 9 attributes and 768 data entries. The calculation is done using Google Collab. Each process is repeated until all parts get a portion of the testing data. The last process is testing using confusion matrix. This matrix produces the classification accuracy level of all algorithms. The accuracy level is obtained from the percentage of data calculation results that match the actual conditions [14]. The more the accuracy rate value means that the algorithm or method is getting better [15].

This research is very important to do in order to choose the optimal model in improving public health services. This research can also increase the understanding of machine learning in the health sector and provide information on influential factors. With the right algorithm model for diabetes data, it is hoped that the next research can create a decision support system for early detection of diabetes. the system can recognize early human behavior and lifestyle in order to reduce the risk of severity due to diabetes. This research is different from previous studies that only compare 2 methods and mostly use KNN and naive bayes.

2. MATERIAL AND METHOD

The stages in this research are shown in the following figure 1.

.

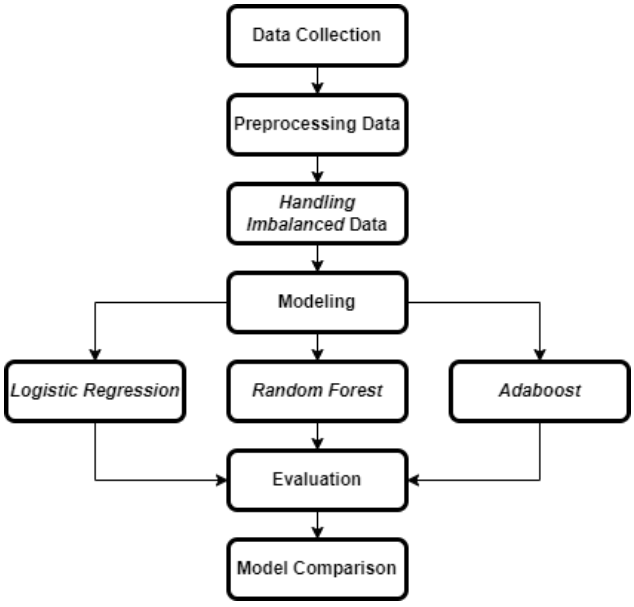

Figure 1. Research Methodology

2.1. Data Collection

The data used in this study is a diabetes dataset taken from the Kaggle.com website sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset has 768 data entries with 9 attributes, the last attribute has 2 values. Value 1 for patients affected by diabetes and value 0 for patients who are not affected by diabetes. The following is a description of the attributes in tabulated form. The formula is written clearly using an equation with an index like the following example of table 1.

Table 1. Data Collection

Atribute	Data
Pregnancies	768 non-null int64
Glucose	768 non-null int64
BloodPressure	768 non-null int64
SkinThickness	768 non-null int64
Insulin	768 non-null int64
BMI	768 non-null int64
DiabetesPedigreeFunction	768 non-null int64
Age	768 non-null int64
Outcome	768 non-null int64

2.2. Preprocessing Data

Preprocessing is the first step before processing data or analyzing data. One of the functions of preprocessing is to identify missing values in the dataset [16]. In the diabetes dataset used in this study, there are no missing values so that the data is effective for processing.

2.3. Handling Imbalanced Data

Handling unbalanced data is a set of techniques or strategies designed to address the problem of unbalanced class distribution in a data set. Imbalance occurs when one class has more or less samples than another class [17]. Some common methods for handling imbalanced data involve resampling techniques, weight adjustment, and the use of specialized algorithms

2.4. Logistic Regression

Logistic regression is a statistical method commonly used in various fields including social science, biostatistics and machine learning [18]. Although the name is similar to linear regression, logistic regression is more suitable for classification processes and modeling probabilities. In the case of binary classification, logistic regression uses a sigmoid function to represent the relationship between the input (feature) and output (class) variables. The sigmoid function is defined as equation 1.

$$P(y = 1) = \frac{1}{1 + e^{-x}} \quad (1)$$

The logistic regression equation is used to predict the probability of the target class (e.g., 0 or 1) based on some input features. P(Y=1) is the probability that the output (Y) is a positive class and e is the natural algorithm base.

2.5. Random Forest

Random forest is an ensemble algorithm used for classification tasks in machine learning .Ensemble methods combine results from multiple models to improve the performance and stability of predictions [19]. Random forest is known to be reliable in overfitting, and performs well on various types of data. Random forest is an ensemble learning that can build multiple decision trees. Random forest algorithm isalso a useful decision algorithm for regression and classification. The output generated from random forest forclassification is the result of all the trees in the ensemble. The equation of random forest is defined as equation 2.

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} fi\,(x) \quad (2)$$

Description:

N : the number of trees in the forest
$f_i(x)$: the tree prediction for the input
$\frac{1}{N}$: the equal weight given to each tree

2.6. AdaBoost

Adaboost or adaptive boosting is also a machine learning algorithm for improving classification models [20]. It works by giving more weight to instances that are difficult to predict in advance, allowing the model to focus more on difficult samples and improve overall classification accuracy. AdaBoost also uses a decision tree, but can give each tree a different weight. These weights are used to place more emphasis on samples that were misclassified in the previous tree. The output of the AdaBoost model is as equation 3.

$$F(x) = \sum_{t=1}^{T} \alpha_t f_t(x) \quad (3)$$

Description:

F(x) : final model output
T : number of booting rounds (iterations)
a_t, : alpha weight for tth weak learner.
ft (x) : prediction of the t-th weak learner for input x

The AdaBoost model as a whole consists of a linear combination of decision tree outputs, with poorly performing trees being given greater weight in subsequent iterations. Therefore, although logistic regression, random forest, and AdaBoost differ significantly in approach, each of these algorithms can be represented with mathematical formulas that show how each models the relationship between features and outputs.

3. RESULTS AND DISCUSSION

Modeling of Logistic Regression, Random Forest and Adaboost Algorithms is carried out for the classification of diabetes. Based on the results of the model experiment, the confusion matrix results are obtained as figure 2.

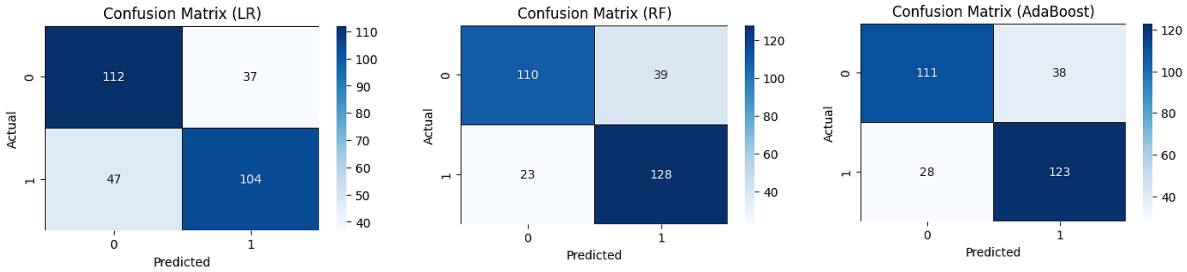

Figure *2*. Result Evaluation Logistic Regression

Evaluation of the processing results of the Logistic Regression algorithm with the confusion matrix above shows 216 true prediction data and 84 false prediction data, with the following detailed results: True Positive = 112; True Negative = 104; False Positive = 47 and False Negative = 37. For the evaluation of the processing results of the Random Forest algorithm with the confusion matrix above, it shows 238 correct prediction data and 62 wrong prediction data, with the following detailed results: True Positive = 110; True Negative = 128; False Positive = 23 and False Negative = 39. While the evaluation of the processing results of the adaboost algorithm with the confusion matrix above shows 234 correct prediction data and 66 wrong prediction data, with the following details: True Positive = 111; True Negative = 123; False Positive = 28 and False Negative = 38.

Based on the evaluation results with the confusion matrix, then the accuracy comparison is carried out on the three algorithms. The following Figure 4. is the result of a comparison between logistic regression, random forest and adaboost in the classification of diabetes, which in this comparison or comparison the random forest algorithm has the highest accuracy value among adaboost and logistic regression algorithms.

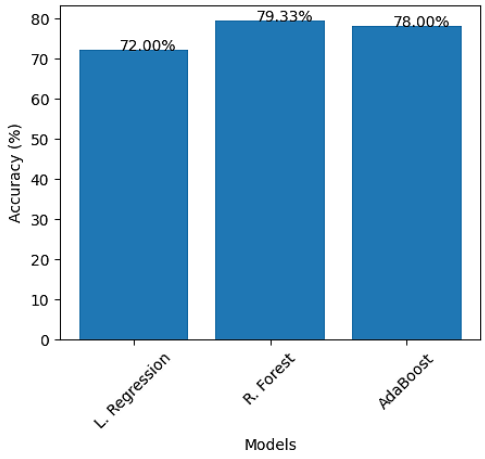

Figure 3. Accuracy comparison result diagram

Evaluation with confusion matrix resulted in ideal comparison values, where some data processing outcomes achieved high accuracy. It is evident that all three algorithms have accuracy above 70%, indicating that the algorithm performance is quite good, especially Random Forest, which has accuracy approaching 80%. The availability of the dataset may influence the evaluation conducted; further research could compare it with other evaluations such as precision and recall.

4. CONCLUSION

Based on the research results from above, the random forest algorithm has the highest accuracy value of 79.33%, followed by adaboost with a value of 78%, and logistic regression with a value of 72%. Thus, random forest is consistently better in diabetes mellitus disease classification. Random forest and adaboost are also more resistant to outliers in the diabetes mellitus dataset because they have high accuracy values.

REFERENCES

[1] S. Safiri et al., "Prevalence, Deaths and Disability-Adjusted-Life-Years (DALYs) Due to Type 2 Diabetes and Its Attributable Risk Factors in 204 Countries and Territories, 1990-2019: Results From the Global Burden of Disease Study 2019," Front. Endocrinol. (Lausanne)., vol. 13, no. February, pp. 1–14, 2022, doi: 10.3389/fendo.2022.838027.

[2] C. Carpinteiro, J. Lopes, A. Abelha, and M. F. Santos, "A Comparative Study of Classification Algorithms for Early Detection of Diabetes," Procedia Comput. Sci., vol. 220, pp. 868–873, 2023, doi: 10.1016/j.procs.2023.03.117.

[3] S. Rammang and N. N. Reza, "Pengendalian Diabetes Melitus Melalui Edukasi dan Pemeriksaan Kadar Gula Darah Sewaktu," vol. 7, pp. 133–137, 2023.

[4] T. Mora, D. Roche, and B. Rodríguez-Sánchez, "Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms," Diabetes Res. Clin. Pract., vol. 204, no. April, 2023, doi: 10.1016/j.diabres.2023.110910.

[5] R. J. Tiurma and Syahrizal, "Obesitas Sentral dengan Kejadian Hiperglikemia pada Pegawai Satuan Kerja Perangkat Daerah," Higeia J. Public Heal. Res. Dev., vol. 5, no. 3, pp. 227–238, 2021.

[6] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Comput. Sci., vol. 216, no. 2022, pp. 21–30, 2022, doi: 10.1016/j.procs.2022.12.107.

[7] S. Kumar and K. K. Mohbey, "A review on big data based parallel and distributed approaches of pattern mining," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 5, pp. 1639–1662, 2022, doi: 10.1016/j.jksuci.2019.09.006.

[8] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," Soc. Sci. Res., vol. 110, no. April 2022, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.

[9] M. M. Rahman, Y. Watanobe, T. Matsumoto, R. U. Kiran, and K. Nakamura, "Educational Data Mining to Support Programming Learning Using Problem-Solving Data," IEEE Access, vol. 10, pp. 26186–26202, 2022, doi: 10.1109/ACCESS.2022.3157288.

[10] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Glob. Transitions Proc., vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.

[11] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, "A comparative study of different machine learning tools in detecting diabetes," Procedia Comput. Sci., vol. 192, pp. 467–477, 2021, doi: 10.1016/j.procs.2021.08.048.

[12] S. Dutta, B. C. S. Manideep, S. M. Basha, R. D. Caytiles, and N. C. S. N. Iyengar, "Classification of diabetic retinopathy images by using deep learning models," Int. J. Grid Distrib. Comput., vol. 11, no. 1, pp. 89–106, 2018, doi: 10.14257/ijgdc.2018.11.1.09.

[13] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," Int. J. Cogn. Comput. Eng., vol. 2, no. November 2020, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[14] P. A. Zandbergen and S. J. Barbeau, "Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones," J. Navig., vol. 64, no. 3, pp. 381–399, 2011, doi: 10.1017/S0373463311000051.

[15] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[16] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, Data preprocessing in predictive data mining, vol. 34. 2019. doi: 10.1017/S026988891800036X.

[17] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," Inf. Sci. (Ny)., vol. 513, pp. 429–441, 2020, doi: 10.1016/j.ins.2019.11.004.

[18] G. Di Franco and M. Santurro, "Machine learning, artificial neural networks and social research," Qual. Quant., vol. 55, no. 3, pp. 1007–1025, 2021, doi: 10.1007/s11135-020-01037-y.

[19] D. Nguyen et al., “Ensemble learning using traditional machine learning and deep neural network for diagnosis of Alzheimer’s disease,” IBRO Neurosci. Reports, vol. 13, no. September, pp. 255–263, 2022, doi: 10.1016/j.ibneur.2022.08.010.
[20] K. W. Walker and Z. Jiang, “Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach,” J. Acad. Librariansh., vol. 45, no. 3, pp. 203–212, 2019, doi: 10.1016/j.acalib.2019.02.013.