International Journal of Advances in Applied Sciences (IJAAS)
Vol. 13, No. 3, September 2024, pp. 737~746
ISSN: 2252-8814, DOI: 10.11591/ijaas.v13.i3.pp737-746

Enhancement performance of the Naïve Bayes method using AdaBoost for classification of diabetes mellitus dataset type II

I Gusti Agung Putu Mahendra[1], I Made Agus Wirawan[2], I Gede Aris Gunadi[3]
[1]Department of Computer Science, Faculty of Postgraduate, Universitas Pendidikan Ganesha, Singaraja, Indonesia
[2]Department of Informatics Engineering, Faculty of Engineering and Vocational Studies, Universitas Pendidikan Ganesha, Singaraja, Indonesia
[3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Pendidikan Ganesha, Singaraja, Indonesia

Article Info

Article history:

Received Jan 12, 2024
Revised May 31, 2024
Accepted Jun 19, 2024

Keywords:

AdaBoost
Confusion matrix
Cross-validation
Diabetes
Naïve Bayes

ABSTRACT

In using technology, especially in health sciences, machine learning modeling can make it easier to predict disease treatment. Naïve Bayes optimization with AdaBoost is needed because even though Naïve Bayes has the advantage of minimal parameters, its accuracy is susceptible to too many features. AdaBoost is used to overcome sensitivity to an excessive number of features and optimize its ability to handle complex datasets. This research aims to analyze the classification results of the Naïve Bayes method with the help of the AdaBoost method. This data comes from Community Health Centers I, II, and III Mengwi District, Bali Province patient medical records. The classification process uses the Naïve Bayes method and Naïve Bayes with AdaBoost, which is then evaluated using a confusion matrix. Two scenarios were used in testing: Naïve Bayes and AdaBoost-based Naïve Bayes. The algorithm is implemented on the dataset and tested directly using cross-validation. The evaluation results show that the Naïve Bayes method experienced an increase in accuracy of 5.92% at 5-fold and 5.93% at 10-fold on a dataset with 890 data. The addition of the AdaBoost method to diabetes classification has been proven to improve the accuracy performance of the Naïve Bayes method.

Corresponding Author:

I Made Agus Wirawan
Department of Informatics Engineering, Faculty of Engineering and Vocational Studies
Universitas Pendidikan Ganesha
Udayana Street No.11 Singaraja-Bali 81116, Indonesia
Email: imade.aguswirawan@undiksha.ac.id

1. INTRODUCTION

Diabetes mellitus is a metabolic syndrome caused by abnormal blood glucose levels (hyperglycemia) [1]. One of the causes of diabetes mellitus is changes in people's lifestyles and habits, from traditional to modern, which are unhealthy and vulnerable to the risk of non-communicable diseases such as diabetes mellitus. This phenomenon reflects the challenges in preventing and treating diabetes mellitus, so early detection of diabetes is critical. If detected early, patients can avoid diabetes [2]. The World Health Organization (WHO) estimates that the number of diabetes mellitus sufferers in Indonesia will increase from 8.4 million in 2000 to around 21.3 million in 2030 [3]. Data from Community Health Centers I, II, and III in Mengwi District shows an increase in diabetes mellitus cases from 480 cases in 2021 to 634 cases in 2022.

Several machine learning-based methods can diagnose the disease: random forest, Naïve Bayes, ID3, C4.5, and others. Of the several machine learning methods used, the Naïve Bayes method produces higher accuracy and uses fewer parameters. However, the Naïve Bayes method is weak because it is

susceptible to too many features, which can reduce accuracy. This algorithm works using general probability. Probability is the chance and level of possibility of an event occurring [4]. However, in complex medical data, these models are prone to overfitting, requiring special techniques to make the models more resistant to overfitting. One technique that can be used to overcome this problem is boosting using the AdaBoost algorithm [5]. This algorithm is suitable for use on complex medical datasets because it can improve model accuracy and make it more resistant to overfitting [6]. The AdaBoost method has the advantage of reducing the error rate of weak classifiers to increase the accuracy of existing learning algorithms and can be easily combined with data mining classification methods [7].

Studies comparing the effectiveness of algorithms in diagnosing diabetes mellitus show variations in results. Research comparing classification and regression tree (CART) and Naïve Bayes shows that Naïve Bayes has lower accuracy, with a value of around 73.75% [8]. The research compares the logistic regression, Naïve Bayes, decision tree classifier, and K-nearest neighbor (K-NN) classifier methods. From this comparison, the Naïve Bayes method obtained an accuracy of 74.5% [9]. Furthermore, there was research using the Naïve Bayes method, which obtained results of 67.71% this low accuracy result was because the Naïve Bayes method was more suitable for the dataset category [10]. However, there is research that improves the performance of Naïve Bayes by calculating the absolute error between the prediction and the actual class, which increases accuracy to around 71.5% [11]. There is also research using the Naïve Bayes method to predict diabetes, with results of 76.30% [12]. Previous research concluded that the Naïve Bayes method still has low accuracy because it is greatly influenced by the number of features [13]. Based on research problems and studies, this research contributes to improving the performance of the Naïve Bayes method by using the AdaBoost method to classify type II diabetes mellitus. The data used in this research is medical record data of type II diabetes mellitus sufferers at Mengwi I, II, and III Community Health Centers, Bali Province.

2. RESEARCH METHOD

This study uses the parameters of gender, age, alcohol consumption, smoking habits, body mass index (BMI) results, systole, diastole, blood sugar, fasting blood sugar, and 2-hour blood sugar to obtain data that can help classify diabetes mellitus. These parameters were obtained from interviews with several doctors. The research methodology includes data collection, data preprocessing, Naïve Bayes and AdaBoost modeling methods, performance evaluation methods (cross-validation), and performance evaluation parameters (confusion matrix). The following is presented in Figure 1 regarding the research flow.

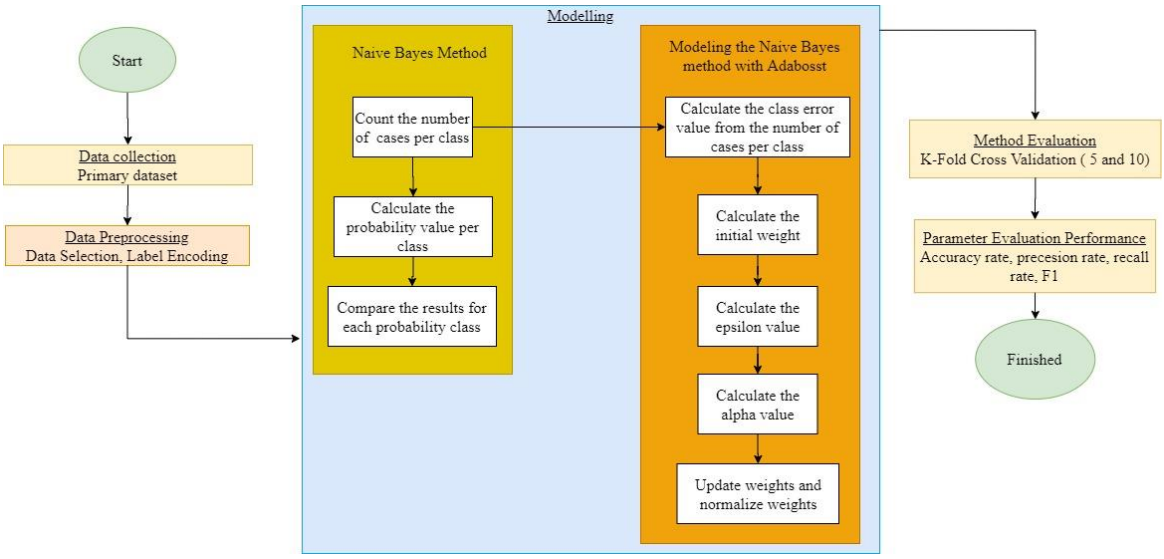

Figure 1. Research stages

2.1. Data collection

The data collected was obtained from medical record data at Community Health Centers I, II, and III Mengwi District. with a total of 890 data, the Mengwi I District Health Center has 300 data, the Mengwi II

District Health Center has 350 data and the Mengwi III District Health Center has 240 data. Data collection is carried out by collecting medical records, then the medical records that have been collected are first selected according to needs so that training data is obtained that meets the criteria, namely data originating from medical records of patients who have diabetes mellitus. These parameters were obtained from an interview with Dr. I Made Ariyoga Budiana.

2.2. Data preprocessing

The data that has been collected is carried out at the preprocessing stage. If empty values are present, steps are taken to delete rows containing empty values. Data selection is the selection (selection) of data from a set of operational data. This data selection is carried out from the patient's medical record by taking the parameters of gender, age, alcohol consumption, smoking habits, BMI results, systole, diastole, blood sugar, fasting blood sugar, 2-hour blood sugar to obtain data that can help the diabetes mellitus classification process. Label coding converts data to simpler forms to correct spelling errors and non-uniform formats to simplify classification.

2.3. Modeling

At this stage, a classification process will be carried out using the Naïve Bayes method and combined with the AdaBoost method. AdaBoost works by allocating weights to each training example so that misclassified examples are given greater weight. Then, a Naïve Bayes model is generated from a subset of the training data selected by considering the weights of the training examples. AdaBoost produces a new Naïve Bayes model at each iteration that can overcome errors made in the previous model. In this way, AdaBoost can improve Naïve Bayes performance by considering the relationships between features in the data.

2.3.1. The Naïve Bayes method

The Naïve Bayes classification method, proposed by the British scientist Thomas Bayes, is based on probability and statistics. This method estimates the possibility of future events based on previous experience, so it is often called Bayes' theorem [14]. One of the main characteristics of the Naïve Bayes classification method is the strong assumption of the independence of each condition or event [15]. In Naïve Bayes, if there are two separate events (for example, X and H), then Bayes' theorem is formulated as follows:

i) Overall calculation probability (P0)

Calculated by dividing the amount of data by a specific class (positive/negative) and the total amount of data. The Integer Probability Formula is shown in (1).

$$P0 = \frac{total\ data\ (positive/negative)}{total\ amount\ of\ data} \quad (1)$$

ii) Partial probability (PP) calculation

Calculated by the amount of data together with specific criteria plus one with the total positive/negative data plus the number of criteria. The partial probability formula is shown in (2).

$$PP = \frac{Number\ of\ criteria\ data+1}{Total\ data\ (positive\ or\ negative) + number\ of\ criteria} \quad (2)$$

iii) Calculation of natural logarithms (LN)

It is helpful to avoid underflow or overflow when multiplying Lot signs if the probability is too low. The natural logarithm formula is shown in (3).

$$LN\ (partial\ probability\ results) \quad (3)$$

iv) Calculation probability class

The probability of a particular class of things is calculated using the exponential natural logarithm of the total class probability. The formula for calculating probability classes is shown in (4).

$$P(class) = exp\ (\Sigma\ LN) \quad (4)$$

Where i) P0 is overall probability; ii) PP is partial possibility; ii) LN is natural logarithm; iii) P(class) is class probability; and vi) Exp is exponential function; ∑LN is total value of the natural logarithm of class probability.

2.3.2. The AdaBoost method

Bagging and boosting techniques increased classification accuracy on artificial and real datasets. The boosting algorithm is generally considered superior to bagging, although this is not always the case. They have proven effective in improving classification performance in many situations, even when the data is imbalanced. AdaBoost has the potential to reduce errors. When applied to Naïve Bayes, AdaBoost improves performance by 33.33% and produces accurate classification results by reducing errors through iterative improvement. AdaBoost is a popular boosting technique in ensemble learning. Boosting algorithms can be used with various classifiers to improve classification accuracy. AdaBoost is a machine learning algorithm comprising a group of weak classifiers combined into a robust classifier. AdaBoost comes from the words "adaptive" because it can adapt to data and other classifiers and "boosting" because it increases the accuracy of each learning given [16]. The main goal of AdaBoost is to reduce errors in the learning process. The AdaBoost algorithm is built using certain equations. Where in the classification, rules are given with labels (X_1, Y_1), …, (X_n, Y_n). Element Y has values 1 and -1. A value of +1 is given to rules with more outstanding positive data than a negative data value of -1 for rules with more significant malicious data, while a value of 0 is given to rules with the same amount of positive data and malicious data. Find the initial weight with the formula $\frac{1}{n}$, where $n$is the amount of data. Next, the alpha count value is shown in (5) [17].

$$\alpha = 0.5 \,.\ln(\frac{1-\epsilon}{\epsilon}) \tag{5}$$

Where α is t^{th} model weights and ϵ is error value at t^{th} iteration.

After getting the alpha value, the next step is to carry out a weight update process to emphasize samples that are difficult to classify by the previous weak model so that the robust model that is formed can focus on more complex cases. The normalization factor Z_t ensures that the total sample weight at each iteration remains 1. The weight update is shown in (6) [18].

$$\omega_i{}^{(t+1)} = \frac{\omega_i^{(t)}.exp\,(-\alpha_t.\gamma_i.h_t(x_i))}{z_t} \tag{6}$$

Where *wi* (*t*+1) is weight of the i^{th} sample in the t^{th} iteration+1, *wi* (*t*) is weight of the i^{th} sample in the t^{th} iteration, αt is weight of the t^{th} weak model, *Yi* is the actual label of the i^{th} data, *ht* (*xi*) is prediction of the t^{th} weak model on the i^{th} data, and *Zt* is normalization factor.

2.3.3. The proposed method (Naïve Bayes with AdaBoost)

Our proposed classifier, Naïve Bayes with AdaBoost, is a novel and powerful hybrid algorithm for solving classification problems. This algorithm consists of two classifications: Naïve Bayes and AdaBoost. Figure 2 explains how the merger method improves the accuracy of the Naïve Bayes method.

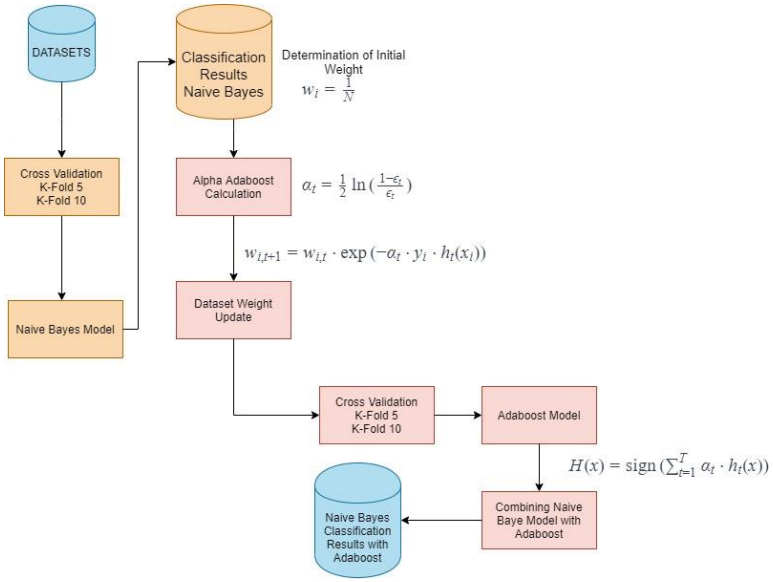

Figure 2. Architectural classifier

AdaBoost can help better with the error rate, alpha value, epsilon value, and weights of each sample data [19]. The error value measures the extent to which the AdaBoost model made errors on the classification data samples in the previous iteration. The error rate measures how "hard" or "easy" the sample was misclassified in the last iteration. The alpha value is the weight given to each classifier used in AdaBoost. The alpha value is used to provide more significant "voting" to classifiers performing well in classifying samples misclassified in the previous iteration. The epsilon value measures the accuracy of the AdaBoost model in a given iteration [20]. The smaller the epsilon value, the greater the alpha value, which means that the classifier in that iteration contributed significantly to correcting errors. Weights provide the importance of sample data in the AdaBoost learning process [21].

2.4. Evaluation model

At the evaluation stage, compare the facet marks' accuracy, recall, and precision with cross-validation using k of 5 and 10. This test has 445, 668, and 890 data samples evaluated. This evaluation was carried out because large amounts of data also contribute to producing the best machine-learning algorithm [22]. The following, presented in Figure 3. explain scheme evaluation testing.

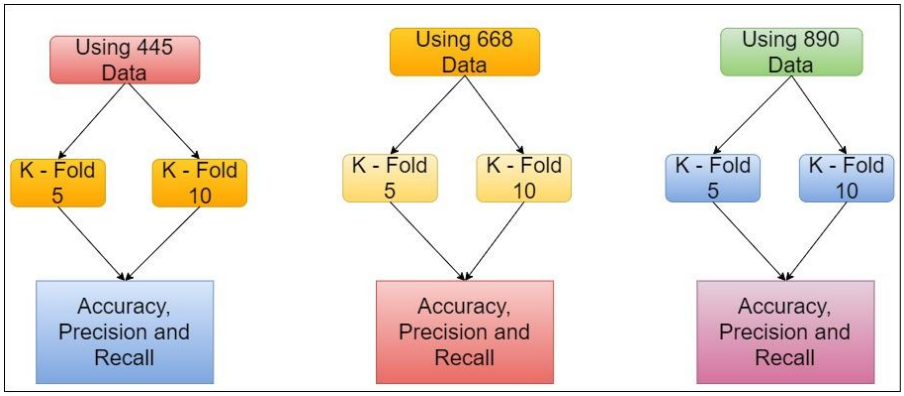

Figure 3. Test scheme

Cross-validation, or what can be called rotation estimation, is a model validation technique to assess how the results of a statistical analysis will generalize to independent data sets. One cross-validation technique is k-fold cross-validation, which breaks the data into k parts of the data set of the same size. In 10-fold, the data is divided into 10-folds of approximately equal size so that we have ten data subsets to evaluate the performance of the model or algorithm [23]. i) The 1st iteration (1-fold as Testing, 2-fold until 10-fold as training); ii) The 2nd iteration (2-fold as Testing, 1-fold, 3-fold until 10-fold as training); iii) For the10th iteration (10-fold as Testing, 1-fold until 9-fold as training); iv) Evaluation matrix calculation; and v) After all iterations are complete, all accuracies are calculated to find the average to determine overall model performance. Confusion matrices are a helpful tool for analyzing the extent of classification performance, with the ability to identify tuples from different classes. Evaluate the model by calculating accuracy (Acc), precision, recall, and F1 scores.

3. RESULTS AND DISCUSSION

This study looked into the effects of optimizing the Naïve Bayes method with AdaBoost to classify diabetes mellitus type II. The classification results compare the actual and predicted labels, which are presented as a confusion matrix. Model evaluation involves measuring model performance using the metrics of accuracy, recall, precision, and F1-score. All experiments were conducted using cross-validation methods to ensure consistent results.

3.1. Dataset

Type II diabetes mellitus is a condition of hyperglycemia caused by cells' unresponsiveness to insulin. Insulin levels may decrease slightly or remain normal. Because pancreatic beta cells still produce insulin, type II diabetes mellitus is considered non-insulin-dependent diabetes mellitus. This metabolic disorder is characterized by increased blood sugar due to decreased insulin secretion by beta cells [24]. These parameters interact with each other and may contribute to the development of type II diabetes mellitus.

Therefore, adopting a healthy lifestyle with a balanced diet, exercising regularly, not smoking, and consuming moderate amounts of alcohol can help reduce the risk of type II diabetes. In addition, it is crucial to maintain normal blood pressure and monitor blood sugar levels regularly to quickly detect the risk or development of disease. Based on literature studies and interviews with experts, the features used to diagnose type II diabetes mellitus are: gender (male, female), age (adult=26-35, late adult=36-45, late elderly=46-55, elderly, senior=56-65), alcohol consumption (yes, no), smoke (yes, no), BMI results (thin=<18.5, ideal=18.5-24.9, fat=25-29.9, very fat=>30), systole (normal=120, pre hypertension=120-139, hypertension 1=140-159, hypertension 2=160), diastole (normal=80, pre hypertension=80-89, hypertension-1=90-99, hypertension-2=100), blood sugar (no=<100, not sure=100-199, definitely=>200), fasting blood sugar (no=<100, not sure=100-125, definitely=>126), two hour blood sugar (no=110-144, not sure=145-179, definitely=>180), class (positive, negative).

3.2. Overall evaluation results using Naïve Bayes based on AdaBoost

In this subchapter, we will evaluate the performance of the model using two different approaches, namely Naïve Bayes without AdaBoost and Naïve Bayes using AdaBoost, with varying amounts of data and using different k-fold values to find out the extent to which AdaBoost can help in improving classification accuracy [25]. With 5-fold totaling 445 data in Table 1, the Naïve Bayes method using AdaBoost produces an accuracy of 81.78%. On the other hand, the Naïve Bayes method produces an accuracy of 76.40%. With a 10-fold, the Naïve Bayes method using AdaBoost produces an accuracy of 84.40%, while the Naïve Bayes method produces an accuracy of 81.14%. With a 5-fold total of 668 data in Table 1, the Naïve Bayes method uses AdaBoost to produce an accuracy of 80.16%. In contrast, the Naïve Bayes method produces an accuracy of 75.79 %, and with 10-fold, the Naïve Bayes method using AdaBoost produces an accuracy of 82.66%. As a comparison, the Naïve Bayes method produces an accuracy of 80.61%. From experiments with 890 data, the Naïve Bayes method with AdaBoost has the highest accuracy, with a 5-fold of 85.36%, as shown in Table 1. With a 10-fold, it is 87.96% for the Naïve Bayes method, which has an accuracy of 79.44% with a 5-fold and 82.02% with a 10-fold. There is an increase in accuracy of 5.92% with 5-fold and 5.93% with 10-fold.

Precision and recall values on data 445 and 668 obtained low results because Naïve Bayes assumes conditional independence between features. Even though dataset features are correlated, this assumption can lead to reduced precision or recall [26]. With a total of 448 positive data and a total of 442 dangerous data, a total of 890 precision and recall data values have increased. This result is because adding positive and negative data can balance the class distribution and increase the representation of the minority (positive) and majority (negative) classes from the previous data. With more data, models can have more information to learn and improve.

The F1-Score in Table 1 produces the highest value from the previous experiment. These results show that the model has a good balance between precision (the extent to which positive results are correct) and recall (the extent to which the model can find all positive cases). This condition means the model can provide accurate predictions with a low risk of false positive and false negative errors [27]. AdaBoost is highly correlated with the improved performance of Naïve Bayes methods. The method proposed in this research can increase accuracy, precision, and recall compared to the Naïve Bayes method without AdaBoost.

Table 1. Overall results of the experiment

Fold	Data	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Computing time (s)
5-fold	445	Naïve Bayes	76.40	57.45	48.84	52.79	0.0590
5-fold	445	Naïve Bayes+AdaBoost	81.78	56.87	57.31	57.08	3.3480
5-fold	668	Naïve Bayes	75.79	57.58	48.04	52.24	0.1281
5-fold	668	Naïve Bayes+AdaBoost	80.16	52.67	56.99	55.81	4.3620
5-fold	890	Naïve Bayes	79.44	70.75	66.07	68.37	0.1350
5-fold	890	Naïve Bayes+AdaBoost	85.36	69.63	71.40	70.50	4.9360
10-fold	445	Naïve Bayes	81.14	54.44	51.36	52.85	0.1298
10-fold	445	Naïve Bayes+AdaBoost	84.40	55.84	54.00	54.90	4.4730
10-fold	668	Naïve Bayes	80.61	52.67	52. 67	52.63	0.1630
10-fold	668	Naïve Bayes+AdaBoost	82.66	51.21	52.86	52.02	6.9928
10-fold	890	Naïve Bayes	82.02	66.30	67.55	66.92	0.2550
10-fold	890	Naïve Bayes+AdaBoost	87.96	73.42	67.10	70.18	7.8261

3.3. Comparison of our proposed method with the Pima Indians dataset

In the context of the Pima Indians diabetes data analysis, previous studies have used Naïve Bayes methods to predict diabetes risk. From Table 2, it is observed that the proposed model achieved better results

on the dataset. In previous research, Naïve Bayes showed a relatively low level of accuracy when compared with the proposed method. This condition is caused by the weakness of Naïve Bayes, which is very sensitive to an excessive number of features and ultimately causes a decrease in classification accuracy. The proposed method uses cross-validation with 10-fold. It uses an error limit value of 0.3 to select the number of features from 8 features to 5 features to obtain an accuracy of 81.5%. AdaBoost works by allocating weights to each training example so that misclassified examples are given greater weight. Then, a Naïve Bayes model is generated from a subset of the training data selected by considering the weights of the training examples. AdaBoost produces a new Naïve Bayes model at each iteration that can overcome errors made in the previous model. Based on the experiments presented in Table 2, we analyze that Naïve Bayes optimization with AdaBoost also produces higher performance than Naïve Bayes without AdaBoost.

Table 2. Comparison of the algorithm with the proposed one improved algorithm Naïve Bayes classifier on Pima Indian dataset

Reference	Model	Accuracy (%)
RB-Bayes algorithm for the prediction of diabetes in "Pima Indian dataset" [10]	Naïve Bayes	67.71
Diabetes diagnosis system using modified Naïve Bayes classifier [11]	Naïve Bayes	71.5
Comparison of Cart and Naïve Bayesian algorithm performance to diagnose diabetes mellitus [8]	Naïve Bayes	73.75
The prediction of diabetes: a machine learning approach [9]	Naïve Bayes	74.5
Prediction of diabetes using classification algorithms [12]	Naïve Bayes	76.30
Our proposed method	Naïve Bayes+AdaBoost	81.5

3.4. Error limit determination experiments and computing time comparison

The error value measures the extent to which the AdaBoost model makes errors in classifying samples. Table 3 shows the experimental results for each error value to find the best accuracy in improving the Naïve Bayes method with AdaBoost. Table 3 shows that with an error value of 0.3, the accuracy results are the best compared to values of 0.4 and 0.5. This error limit value is obtained by comparing the amount of data in the opposite result class and the amount of positive and negative data. AdaBoost is the most popular boosting method and has been theoretically and empirically proven to improve baseline Naïve Bayes performance [28].

Table 3. Experimental error limit values

No	Sign error limits	Number of features	NB accuracy+AdaBoost 5-fold (%)	NB accuracy+AdaBoost 10-folds (%)
1	0.3 (Age, BMI results, diastole, fasting blood sugar, two-hour blood sugar)	5	85.36	87.95
2	0.4 (Age, BMI results, systole, diastole, fasting blood sugar, two-hour blood sugar)	6	80.36	83.60
3	0.5 (Gender, age, alcohol consumption, smoke, BMI results, systole, diastole, blood sugar, fasting blood sugar, two-hour blood sugar)	11	78.76	75.73

The main idea behind AdaBoost is the construction of a robust classifier using a group of weak classifiers. Although AdaBoost is very powerful if the error rate is more significant than 0.5, it may fall short of the model's ability to apply the knowledge gained from the training data to the training data with a high degree of accuracy, and the overall accuracy of the model may be reduced [29]. Setting a minimum error rate threshold value can provide stringent criteria for including features in the model. As a result, only features that meet these criteria are included, and other features are ignored [30].

The test results show that error rates of 0.4 and 0.5 can be seen in Table 3 contribute negatively to the performance of the ensemble model. With an error value of 0.4, six features were obtained: age, BMI results, diastole, systole, fasting blood sugar, and 2-hour blood sugar, producing an accuracy of 80.36% with 5-fold and 83.60% with 10-fold. This accuracy is lower if an error rate of 0.3 is set. This condition is caused by the Siatole feature (blood pressure) not being able to help diagnose type II diabetes mellitus. Type II diabetes mellitus is caused by two things, namely, a decrease in peripheral tissue response to insulin (insulin resistance) and a decrease in the ability of pancreatic cells to secrete insulin in response to a glucose load [31]. Meanwhile, eleven features were obtained with an error rate of 0.5, namely gender, age, alcohol consumption, smoking, BMI results, systole, diastole, blood sugar, fasting blood sugar, and 2-hour blood sugar of 78.76% with 5-fold, and 75.73% with 10-fold. This accuracy is lower than using error levels of 0.3 and 0.4, where there is no feature reduction at an error level of 0.5, so the model tends to become more

complex. Features with higher errors (high error rates) may be excluded from consideration. This result means that AdaBoost tends to select and focus on more accurate features when classifying data. This approach will provide a more precise final prediction [32]. Computing results using a laptop with Intel Core i7 7700HQ 3.80 GHz Processor specifications with 16 GB memory. Computation times are shown in Table 1. The regular Naïve Bayes method has faster computation times than AdaBoost-based Naïve Bayes. Ordinary Naïve Bayes may be a good choice if the research focuses more on computational time and relatively simple data. However, if looking for higher accuracy in complex data processing, AdaBoost-based Naïve Bayes may be a better choice. This research has succeeded in comprehensively investigating the improvement of Naïve Bayes performance with AdaBoost. However, additional and in-depth research may be needed to confirm this, especially regarding increasing the training data not limited to just three districts. Our research shows that optimizing the Naïve Bayes method with AdaBoost is more robust than the original method. Even though the AdaBoost method can improve Naïve Baye's performance, this method only focuses on feature selection. Therefore, future research must examine the classification process using ensemble learning methods (bagging, stacking, boosting) [33], [34]. The latest observations show that the AdaBoost method can determine optimal features for classifying type II diabetes mellitus using the Naïve Bayes method. Our findings provide definite evidence that this phenomenon is related to changes in features used rather than caused by an increase in the number of features.

4. CONCLUSION

From the evaluation results using cross-validation from experiments with a total of 890 data, the Naïve Bayes method with AdaBoost has the highest accuracy with 5-fold and 10-fold, namely 85.36% and 87.95%, respectively. In contrast, the Naïve Bayes method has an accuracy of 79.44% with 5-fold and 82.02% with 10-fold. There is an increase in accuracy of 5.92% for 5-fold and 5.93% for 10-fold. The addition of the AdaBoost method to diabetes classification provides a higher accuracy value compared to the Naïve Bayes algorithm without AdaBoost. So it is clear that applying the AdaBoost method to the Naïve Bayes algorithm can increase accuracy. AdaBoost combines several weak classification models into a more robust model that better combines information from multiple features. Naïve Bayes has naive assumptions regarding feature independence, meaning that the Naïve Bayes method assumes that features do not influence each other. Regarding computing time, the Naïve Bayes method produces a faster time than the Naïve Bayes method with AdaBoost and can carry out further comparisons with different methods.

REFERENCES

[1] Y. Lin and Z. Sun, "Current views on type 2 diabetes," *Journal of Endocrinology*, vol. 204, no. 1, pp. 1–11, Jan. 2010, doi: 10.1677/JOE-09-0260.

[2] E. C. Westman, "Type 2 diabetes mellitus: a pathophysiologic perspective," *Frontiers in Nutrition*, vol. 8, Aug. 2021, doi: 10.3389/fnut.2021.707371.

[3] M. T. Sari, "Characteristics and lifestyle in diabetes mellitus patients," in *The International Conference on Public Health Proceeding*, 2021, pp. 321–327.

[4] R. Achmad and A. S. Girsang, "Implementation of Naive Bayes classifier algorithm in classification of civil servants," *Journal of Physics: Conference Series*, vol. 1485, p. 012018, Mar. 2020, doi: 10.1088/1742-6596/1485/1/012018.

[5] X. Gu and P. P. Angelov, "Multiclass fuzzily weighted adaptive-boosting-based self-organizing fuzzy inference ensemble systems for classification," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 9, pp. 3722–3735, Sep. 2022, doi: 10.1109/TFUZZ.2021.3126116.

[6] A. Sanusi, C. A. Putra, and F. A. Akbar, "Implementation of AdaBoost algorithm on C50 for improving the performance of liver disease classification," *JEECS (Journal of Electrical Engineering and Computer Sciences)*, vol. 8, no. 2, pp. 93–102, Dec. 2023, doi: 10.54732/jeecs.v8i2.1.

[7] M. Haghighi, "Data mining and machine learning: an overview of classifiers," *Ciência e Natura*, vol. 37, p. 76, Dec. 2015, doi: 10.5902/2179460X20756.

[8] I. Santiko and P. Subarkah, "Comparison of cart and Naive Bayesian algorithm performance to diagnose Diabetes mellitus," *IJIIS: International Journal of Informatics and Information Systems*, vol. 2, no. 1, pp. 9–16, Sep. 2019, doi: 10.47738/ijiis.v2i1.9.

[9] L. Kumar and P. Johri, "The prediction of diabetes," *International Journal of Reliable and Quality E-Healthcare*, vol. 11, no. 1, pp. 1–9, Mar. 2022, doi: 10.4018/ijrqeh.298630.

[10] R. Rajni and A. Amandeep, "RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, p. 4866, Dec. 2019, doi: 10.11591/ijece.v9i6.pp4866-4872.

[11] J. K. Alwan, D. S. Jaafar, and I. R. Ali, "Diabetes diagnosis system using modified Naive Bayes classifier," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, p. 1766, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1766-1774.

[12] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[13] R. Wibowo, M. A. Soeleman, and A. Affandy, "Hybrid top-k feature Selection to improve high-dimensional data classification using Naïve Bayes algorithm," *Scientific Journal of Informatics*, vol. 10, no. 2, pp. 113–120, Apr. 2023, doi: 10.15294/sji.v10i2.42818.

[14] I. M. A. Wirawan and I. W. B. Diarsa, "Mobile-based recommendation system for the tour package using the hybrid method," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 12, no. 8, p. 64, Dec. 2018, doi: 10.3991/ijim.v12i8.9483.

[15] S. Pitoglou, A. Filntisi, A. Anastasiou, G. K. Matsopoulos, and D. Koutsouris, "Exploring the utility of anonymized EHR datasets in machine learning experiments in the context of the Model health project," *Applied Sciences*, vol. 12, no. 12, p. 5942, Jun. 2022, doi: 10.3390/app12125942.
[16] J. Zhang, K. Xia, Z. He, Z. Yin, and S. Wang, "Semi-supervised ensemble classifier with improved sparrow search algorithm and its application in pulmonary nodule detection," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–18, Feb. 2021, doi: 10.1155/2021/6622935.
[17] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class AdaBoost," *Statistics and Its Interface*, vol. 2, no. 3, pp. 349–360, 2009, doi: 10.4310/SII.2009.v2.n3.a8.
[18] S. A. Alanazi *et al.*, "Public's mental health monitoring via sentimental analysis of financial text using machine learning techniques," *International Journal of Environmental Research and Public Health*, vol. 19, no. 15, p. 9695, Aug. 2022, doi: 10.3390/ijerph19159695.
[19] V. K. P. German, B. D. Gerardo, and R. P. Medina, "Implementing enhanced AdaBoost algorithm for sales classification and prediction," *International Journal of Trade, Economics and Finance*, vol. 8, no. 6, pp. 270–273, Dec. 2017, doi: 10.18178/ijtef.2017.8.6.577.
[20] R. P. Fhonna, M. K. M. Nasution, and Tulus, "Real-time detection with AdaBoost-SVM combination in various face orientation," *Journal of Physics: Conference Series*, vol. 978, p. 012015, Mar. 2018, doi: 10.1088/1742-6596/978/1/012015.
[21] O. Hornyák and L. B. Iantovics, "AdaBoost algorithm could lead to weak results for data with certain characteristics," *Mathematics*, vol. 11, no. 8, p. 1801, Apr. 2023, doi: 10.3390/math11081801.
[22] R. Goorbergh, M. Smeden, D. Timmerman, and B. Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, Aug. 2022, doi: 10.1093/jamia/ocac093.
[23] O. Chamorro-Atalaya *et al.*, "K-fold cross-validation through identification of the opinion classification algorithm for the satisfaction of university students," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 11, Aug. 2023, doi: 10.3991/ijoe.v19i11.39887.
[24] J. Son and D. Accili, "Reversing pancreatic β-cell dedifferentiation in the treatment of type 2 diabetes," *Experimental & Molecular Medicine*, vol. 55, no. 8, pp. 1652–1658, Aug. 2023, doi: 10.1038/s12276-023-01043-8.
[25] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of machine learning algorithms with different k values in k-fold cross validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: 10.5815/ijitcs.2021.06.05.
[26] D. Berrar, "Bayes' theorem and Naive Bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 403–412.
[27] K. Staszak, B. Tylkowski, and M. Staszak, "From data to diagnosis: how machine learning is changing heart health monitoring," *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 4605, Mar. 2023, doi: 10.3390/ijerph20054605.
[28] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The Elements of Statistical Learning*, New York, NY: Springer, 2009, pp. 337–387. doi: 10.1007/978-0-387-84858-7_10.
[29] O. Agboola, "Spam detection using machine learning and deep learning," Louisiana State University and Agricultural and Mechanical College, 2022.
[30] B. N. Saha, G. Kunapuli, N. Ray, J. A. Maldjian, and S. Natarajan, "AR-boost: reducing overfitting by a robust data-driven regularization strategy," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2013, pp. 1–16.
[31] Y. T. Wondmkun, "Obesity, insulin resistance, and type 2 diabetes: associations and therapeutic implications," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. Volume 13, pp. 3611–3616, Oct. 2020, doi: 10.2147/DMSO.S275898.
[32] T. Takenouchi and S. Eguchi, "Robustifying AdaBoost by adding the Naive error rate," *Neural Computation*, vol. 16, no. 4, pp. 767–787, Apr. 2004, doi: 10.1162/089976604322860695.
[33] X. Wu and J. Wang, "Application of bagging, boosting and stacking ensemble and easy ensemble methods for landslide susceptibility mapping in the three gorges reservoir area of China," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, p. 4977, Mar. 2023, doi: 10.3390/ijerph20064977.
[34] N. S. F. Putri, A. P. Wibawa, H. Ar Rasyid, A. Nafalski, and U. R. Hasyim, "Boosting and bagging classification for computer science journal," *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 1, p. 27, Mar. 2023, doi: 10.26555/ijain.v9i1.985.

BIOGRAPHIES OF AUTHORS

I Gusti Agung Putu Mahendra is a computer science graduate student at Department of Computer Science, Faculty of Postgraduate, Universitas Pendidikan Ganesha. He obtained a Bachelor's degree (S. Kom) in Computers from the Faculty of Informatics and Computers, Information Systems Study Program, Stikom Bali Institute of Technology and Business. His desire to continue developing his understanding and skills in the field of Computer Science is driven by the belief that his research can contribute significantly to technological and scientific progress: He can be contacted at email: agung.mahendra@undiksha.ac.id.

I Made Agus Wirawan is a lecturer at the Department of Informatics Engineering, Faculty of Engineering and Vocational Studies, Universitas Pendidikan Ganesha, Bali, Indonesia. He took a Bachelor's degree (S. Kom) in the Department of Computer Science and Electronics, Faculty of Mathematics & Natural Sciences, Universitas Gajah Mada, Yogyakarta, Indonesia. He obtained a Master's degree (M. Cs) in the Department of Computer Science and Electronics, Faculty of Mathematics & Natural Sciences, Universitas Gajah Mada, Yogyakarta, Indonesia. Lastly, he also obtained a Doctorate (Dr) degree in the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gajah Mada, Yogyakarta, Indonesia. His areas of research interest include machine learning, artificial intelligence, and deep learning. He can be contacted at email: imade.aguswirawan@undiksha.ac.id.

I Gede Aris Gunadi He received a B.Sc. degree in Physics (special field: introductory physics) from the Institut Teknologi Sepuluh November, Surabaya in 2000. He completed a master's program in the Informatics Engineering study program at the Faculty of Information Technology, Institut Teknologi Sepuluh Nopember Surabaya in 2008 and a Doctorate in computer science from Universitas Gajah Mada in 2016. He has 12 years of experience in research, teaching, and community service in the Department of Physics Education, four years in the Computer Science Master's Program, and one year in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Pendidikan Ganesha. The main topics of the field are data science, simulation, computing, and electronics. He is responsible for teaching courses in the data science, IoT, soft computing, computer simulation, and electronics departments. He can be contacted at email: igedearisgunadi@undiksha.ac.id.