HYBRID OF ADABOOST ALGORITHM AND NAÏVE BAYES CLASSIFIER ON SELECTION OF CONTRACEPTION METHODS

Nurul Faridah[1], Candra Dewi[2], Arief Andy Soebroto[3]

[1,2,3] Informatics Department, Universitas Brawijaya, Malang, Indonesia
Email: nurulfaridah038@student.ub.ac.id[1], dewi_candra@ub.ac.id[2], ariefas@ub.ac.id[3]

ABSTRACT

Stunting is a growth failure in children. Stunting can be avoided by adjusting birth spacing or implementing a Family Planning program by using appropriate contraception. Therefore, it is necessary to develop appropriate and rapid contraceptive selection techniques to assist family planning programs. This study develops a model for determining contraceptive methods using a Naïve Bayes Classifier. In addition, an Adaboost algorithm was used to handle the independent between attributes on Naïve Bayes. The performance evaluation of model was measured by combining k-fold cross validation and confusion matrix. Based on the results testing was obtained an optimal parameter of learning rate was 0.1 and the number of iterations was 50. The evaluation using optimal parameters produce the best accuracy of 87.5%, precision of 87.6%, recall of 87.5%, and f1-measure of 87.5%. This result was better than applying the Naïve Bayes without implementing Adaboost, which had 70% accuracy. The used of Adaboost was proven to increase the accuracy of Naive Bayes by 17.5%.

Keywords: *adaboost, naïve bayes classifier, contraception method, k-fold cross validation, confusion matrix*

1. INTRODUCTION

The stunting rate in Madiun City reached 10.18% in the year of 2020. This rate shows that there are 814 children out of a total of 7,996 suffer from stunting (RRI, 2021). Stunting can be prevented by adjusting birth spacing or implementing a Family Planning program, by using appropriate contraception for couples of childbearing age. Therefore, contraception can play a role in preventing stunting in children (BKKBN, 2021).

The contraception data currently available at village famili planning assistant (PPKBD-Pembantu Pembina Keluarga Berencana Desa) in Sumberbening village, Balerejo District, Madiun Regency is currently not used optimally in the selection of contraception. Even though, the data can be utilized to develop a model to aid in the selection of a contraceptive method for potential users.

Research on the prediction of contraceptive methods has been carried out previously using various methods. Research conducted by Naafian, et al. (2017) used the Naïve Bayes algorithm to create a decision support system at Puskesmas II Colomadu. This study resulted in an accuracy of 82.2%. Then research by Nugroho (2015) created a decision support system for the selection of contraceptive methods in couples of childbearing ages using the K-Nearest Neighbor (KNN) algorithm. This study obtained a fairly high level of accuracy, that was 95%. In another study, Wardhani, et al. (2014) used Fuzzy Logic to create a decision support system in determining contraceptives for family planning. From the results of the study, the validation level was almost 80%. Then, research conducted by Abdillah (2016) predicts the determination of the family planning method using the Naïve Bayes Classifier with a case study of the Muara Rumbai Health Center, Pekanbaru. The results of these studies obtained an average level of accuracy with a value of 81.364%.

Previous studies showed that Naïve Bayes most commonly used due to the fact that Naïve Bayes is a relatively simple and can be implemented for categorical and non categorical data (Zaki & Jr., 2013). However, the assumption of independent conditions between attributes in Naïve Bayes is a weakness that must be addressed (Jahromi & Taheri, 2018). This independence assumption can sometimes lead to loss of accuracy in Naïve Bayes (Netti & Netti, 2015). Therefore, it is

necessary to optimize Naïve Bayess to deal with this problem.

Adaboost is one of the boosting algorithms that can be used to increase the accuracy of several learning methods (Han & Kamber, 2006). As in the research conducted by Nurlaela (2020), Adaboost was used to improve the accuracy of Naïve Bayes in predicting film sales revenue with a 1.02% improvement in accuracy. In another study conducted by Rohman, et al. (2017), Adaboost-based C4.5 algorithm was used to predict heart disease and obtained 6.42% improvement in accuracy.

Based on description that has been done, this paper applied the hybrid of Naïve Bayes and Adaboost algorithm for the selection of contraceptive methods in the Family Planning program. Naïve Bayes was proven can handle categorical and continuous data. Meanwhile, the Adaboost algorithm is used to improve the accuracy of the Naïve Bayes.

2. DATA AND METHOD

2.1 Data

This study used the secondary data were taken from the documentation of PPKBD Sumberbening Village, Balerejo District, Madiun Regency. The data contains information about the use of contraception methods at year 2020. This data consists of some attributes such as age, the number of children, being or not breastfeeding, desire to have children, blood pressure, heart disease, diabetes, and contraception methods used (MKJP and Non-MKJP). The total number of data was 200 records, including 98 records for the MKJP class and 102 records for the Non MKJP class. The description of each attribute of the data is displayed at Table 1.

Table 1 Contraception Method Data Attributes

No	Attribute	Description
1	Age (F1)	Age of contraception method user (more than 20 years)
2	Number of children (F2)	Number of children owned by contraception method users (1 or more)
3	Breastfeeding (F3)	Users of contraception methods are or are not breastfeeding 1. Yes (Y) 2. No (T)
4	Desire to have children (F4)	Time interval of desire to have children 1. Less than 2 years (K) 2. More than 2 years (M) 3. Don't want to have any more children (T)
5	Blood pressure (F5)	Blood pressure of users of contraception methods 1. Normal (N) 2. Hypertension (H)
6	Heart Disease (F6)	Have or are currently suffering from heart disease 1. Yes (Y) 2. No (T)
7	Diabetes (F7)	Currently suffering from diabetes 1. Yes (Y) 2. No (T)
8	Contraceptive Methodsi (C)	Types of contraception methods used 1. MKJP 2. Non-MKJP

2.2 Method

The process for the selection of contraception methods begins with training on the classification model. For this purposed, the data was divided into two for training data and test data by comparison of 80% training data and 20% test data.

Classification model training using Adaboost with Naïve Bayes Classifier. In this process, the Naïve Bayes Classifier acts as a weak learner by calculate the probability of each training data in order to obtain the classification results. After that, Adaboost will perform boosting by generating a combination of weak learners. From several weak learners obtained, majority voting will be conducted to obtain the recommendation for choosing a contraception method. The process of selecting contraception is described in Figure 1.

2.2.1 Initialization of Number of Iterations (T) and Weight

The initialization of the T parameter aims to determine the maximum number of iterations that will be carried out in the training process. The weight value will determine the probability of data in the training data resampling process. Initialization of weight for each training data based on Equation 1 (Wu, et al., 2007).

$$W_i = \frac{1}{N}, i = 1, \cdots, N \quad (1)$$

Where W_i is the weight of training data and N is total number of training data

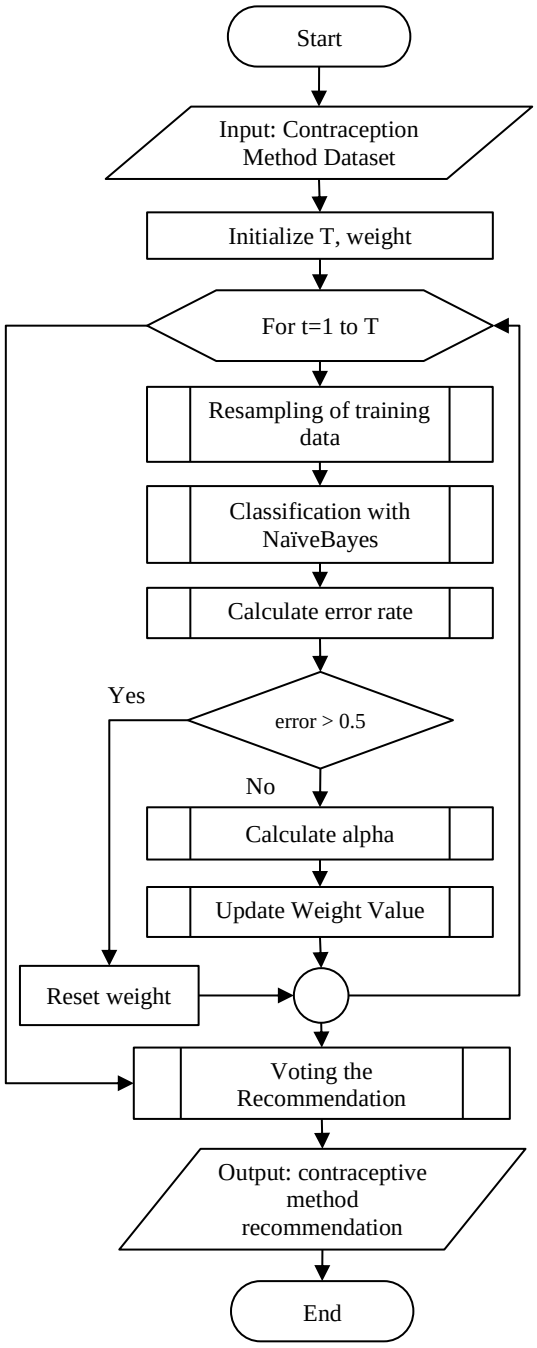

Figure 1 Flowchart of contraception method selection

2.2.2 Resampling of Training Data

The first step for the training process was resampling the training data based on the weight value of each data. The training data will remain the same if the weight values of each data on the training data are all the same. Meanwhile, if the weight values were different, resampling was done at random but still paying attention to the training data's weight value. The resampling data in each iteration will become training data during the training stage, while the test data will remain the initial training data.

2.2.3 Classification with Naïve Bayes Classifier

The Naïve Bayes algorithm interprete that the value of one particular attribute independent to other attributes. Naïve Bayes Classifier is a simplification of the Bayes Theorem and can be calculated using Equation 2 (Bramer, 2007).

$$P(C_i|X) = P(C_i)\prod_{j=1}^{n} P(X_j|C_i) \quad (2)$$

Naïve Bayes combines Prior probability and Likelihood probability in single formula to calculate the posterior probability of each class. Posterior probability will be used to determine the final class of a prediction. The class prediction formula in Naïve Bayes is shown in Equation 3 (Zaki & Jr., 2013).

$$\hat{y} = \arg\max_{i}\{P(C_i|X)\} \quad (3)$$

To calculate the likelihood probabilities $P(x_1|C_i)$, $P(x_2|C_i)$, …, $P(x_n|C_i)$, the type of each attribute should be considered as follows (Han & Kamber, 2006).

1. If the feature is a data category, then P(xj|Ci) is the number of Ci classes in the training data that has a value of xj for the Aj attribute divided by the number of data belonging to the Ci class.
2. If the feature is continuous, it will have a Gaussian distribution with the mean μ and standard deviation σ parameters as in Equation 4.

$$P(X_j|C_i) = g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

Calculation of Likelihood probability needs the value of Prior probability, mean, and standard deviation. The mean and standard deviation are needed if the data attribute is continuous. The calculation of prior probabilities can be done using Equation 5.

$$P(C_i) = \frac{N_c}{N} \quad (5)$$

2.2.4 Calculating the Error Rate

The error rate was obtained by adding the weight values of the data that have misclassification (error weight), then divided by

the amount of data as shown in Equation 6 (Wu, et al., 2007).

$$\varepsilon_t = \frac{1}{n}\sum_{i=1}^{N} w_i^t, y_i \neq h_t(x_i) \quad (6)$$

Denote:
ε_t = error rate value
n = the number of data that has misclassification
$w_i^t, y_i \neq h_t(x_i)$ = weight value of data that has misclassification

The error value has an impact on the weight value determination and iteration process during training. If the error value is greater than 0.5, then the weight value will be reset as Equation 1 and the iteration will proceed to the next iteration without performing the process after the condition selection.

2.2.5 Calculating the Weight of the Classification Model

The weight value of the classification will be used in the process of updating the weight and the majority voting process to determine the results of the recommendations. The formula to calculate the weight value of the classification model is shown in Equation 7 (Wu, et al., 2007).

$$a_t = \frac{1}{2} ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (7)$$

The constant value in Equation 7 is the learning rate. The most commonly value was used as learning rate is 0.5 and 1. This value can be changed as needed.

2.2.6 Updating The Weight Values

The new weight value was computed based on the result (true or false) of classification results. The weight value of data that has errors in the categorization process will increase, and vice versa. The updating of the weight value was performed using Equation 9 (Wu, et al., 2007).

$$w_i^{t+1} = \frac{w_i^t}{Z_t} \cdot \begin{Bmatrix} exp^{-a_t} \; if \; y_i = h_t(x_i) \\ exp^{a_t} \; if \; y_i \neq h_t(x_i) \end{Bmatrix}, i = 1, \cdots, N \quad (9)$$

Denote:
w_i^{t+1} = new weight value
w_i^t = old weight value
Z_t = normalization constant

2.2.7 Voting The Recommendation

If process has reached the maximum number of iterations, the results of the recommended contraceptive method option will be determined by a majority vote. A sample of training data and the weight value of the classification model for each iteration will be obtained from the training process. The training data sample will be used to classify the test data, while the classification model's weight value will be used for majority voting. The majority voting can be done using Equation 10 (Wu, et al., 2007).

$$H(x) = sign(\sum_{t=1}^{T} a_t h_t(x)) \quad (10)$$

Denote:
$H(x)$ = recommendation results
a_t = the weight of the classification model in the t-th iteration
$h_t(x)$ = the data classification in the t-th iteration

3. RESULTS AND DISCUSSION

This study performs testing of parameters that affect the performance of Adaboost. The performance evaluation was measured using K-Fold Cross Validation and the accuracy was measured using a Confusion Matrix.

3.1 Learning Rate Parameter Testing

The purpose of this test is to find the optimal learning rate for contraception method recommendations using the Adaboost algorithm with the Naïve Bayes Classifier. This test had been done on 5 iterations. The test result of the learning rate parameter is shown at Table 2. According to this result was obtained the average accuracy value of 78.5% from a test with a learning rate of 0.1.

Table 2 The Result of Learning Rate Parameter Testing

Learning Rate	Trial Accuracy (%)					Average Accuracy (%)
	1	2	3	4	5	
0,1	75	82,5	77,5	75	82,5	**78,5**
0,2	80	80	70	70	70	74
0,3	72,5	70	67,5	80	67,5	71,5
0,4	65	67,5	70	65	70	67,5
0,5	62,5	65	70	70	67,5	67
0,6	67,5	70	70	67,5	70	69
0,7	60	77,5	70	72,5	70	70
0,8	72,5	70	72,5	72,5	72,5	72
0,9	70	72,5	75	72,5	75	73
1	72,5	72,5	72,5	70	72,5	72

3.2 Iteration Testing

The purpose of this test is to find the optimal number of iterations for contraceptive method recommendations using the Adaboost algorithm with the Naïve Bayes Classifier. The learning rate value in this test was 0.1, which corresponds to the results of the last learning rate test. Table 3 shows the results of testing the number of iteration parameters.

Table 3 The Result of Iteration Testing

Number of Iterations	Trial Accuracy (%)					Average Accuracy (%)
	1	2	3	4	5	
5	70	72,5	72,5	75	82,5	74,5
10	75	72,5	75	80	72,5	75
20	77,5	82,5	75	80	77,5	78,5
50	77,5	87,5	80	85	77,5	**81,5**
100	70	65	67,5	85	77,5	73
200	65	67,5	72,5	82,5	75	72,5

According to Table 3, the second trial with a number of iterations equal to 50 had the highest accuracy value of all the experiments, which was 87.5 percent. Meanwhile, the highest average accuracy value is also obtained from testing using the number of iterations with a value of 50, which is 81.5%.

3.3 Optimal Parameter Testing

This test aims to determine the effect of using optimal parameters on the accuracy of recommendations for choosing a contraceptive method using the Adaboost algorithm implementation with the Naïve Bayes Classifier. The test was repeated five times. The test employs the best parameter determined in the previous test, learning rate of 0.1 and the number of iterations is 50 iterations. The training and testing data of each experiment will be of fixed value. This aims to determine whether each stage of the algorithm has an impact on the final recommendation.

Table 4 The Result of Optimal Parameter Testing

Test	Result (%)			
	Accuracy	Precision	Recall	F_1-Measure
1	87,5	87,5939849	87,5	87,4921826
2	75	75,2525252	75	74,9373433
3	80	80,3030303	80	79,9498746
4	87,5	87,5939849	87,5	87,4921826
5	82,5	82,5814536	82,5	82,4890556

According to Table 4, the first and fourth experiments achieved the highest results, with accuracy of 87.5%, precision of 87.6%, recall of 87.5%, and f1-measure of 87.5%. Furthermore, the boosting by resampling in Adaboost algorithm's causes a change in accuracy in each experiment. This resampling process was carried out randomly so that the results from each experiment will be different. This procedure occasionally receives a sub-optimal sample of training data, resulting in low accuracy, and vice versa. Because the algorithm used in combination with Adaboost is a Naïve Bayes Classifier, the final result is determined by the training data.

3.4 K-Fold Cross Validation

By Using the hybrid of Adaboost and Naïve Bayes Classifier intends to the examinination of the effect of changes in training data and test data on the accuracy of recommendations for choosing a contraceptive method. The K value utilized in this test is 5. According to the results of the previous parameter test, the T parameter is 50 iterations, and the learning rate is 0.1.

Table 5 The Average Result of 5-Fold Cross Validation

Test	Average (%)			
	Accuracy	Precision	Recall	F_1-Measure
1	81	81,6	81,2	80,6
2	82	82,4	81,8	81,4
3	80	80	79,8	79,8
4	81	81,2	80,8	80,6
5	81,5	81,8	81,8	81,2
Average	**81,1**	**81,4**	**81,08**	**80,72**

Table 5 shows that there is no significant difference in the evaluation value of each experiment. The final average is 81.1% accuracy, precision of 81.4%, recall of 81.08%, and f1-measure of 80.72%.

3.5 Comparation the Adaboost-Naïve Bayes and the Naïve Bayes

The accuracy achieved from the implementation of bybrid of Naïve Bayes and Adaboost (Adaboost-Naïve Bayes) then being compared with the accuracy of Naïve Bayess (Table 6). The several tests showed that the

optimal number of iterations of Adaboost-Naïve Bayes was 50 and the learning rate was 0.1.

Table 6 Comparation of Adaboost Naïve Bayes and Naïve Bayes

Classifier	Accuracy	Precision	Recall	F_1-Measure
Naïve Bayes	0,7	0,767	0,7	0,68
Adaboost Naïve Bayes	0,875	0,876	0,875	0,875

Table 6 shows that the performance of Adaboost-Naïve Bayes Classifier outperforms Naïve Bayes Classifier. This was indicated by an increasing of accuracy, precision, recall and F1-measure value. Adaboost-Naïve Bayes Classifier method increase accuracy by 17.5%. Furthermore, the results of accuracy, precision, recall, and F-measure were greater than 80%. This means that the Adaboost-Naïve Bayes is capable of correctly classifying test data. The final findings received from this test were 87% accuracy, 87.6% precision, 87.5% recall, and 87.5% f1-measure.

5. CONCLUSION

Hybrid of Adaboost algorithm and Naïve Bayes Classifier can be used in recommending contraceptive method selection. The testing on optimal parameter (learning rate 0.1 and the number of iterations 50) results the average accuracy of 81.1%, precision of 81.4%, recall of 81.08%, and f1-measure of 80.72%. Meanwhile, the highest results obtained were accuracy of 87.5%, precision of 87.6%, recall of 87.5%, and f1-measure of 87.5%. This shows that Hybrid of Adaboost and Naïve Bayes perform better than Naïve Bayes (which has a 70% accuracy). The recommendation for contraceptive method selection with the combination of Adaboost and Naïve Bayes has boosted accuracy by 17.5%. On the other hand, this method has shortcomings in the training data resampling step in the Adaboost algorithm. The resampling process was carried out randomly and sometimes will obtain an ineffective sample of training data. For future work need to handle this problem by appliying a methods of resampling training data.

6. REFERENCE

Abdillah, I. (2016). Prediksi Penentuan Metode KB dakam Program Keluarga Berencana dengan Menggunakan Naïve Bayes Classifier (Studi Kasus: Puskesmas Muara Fajar Rumbai, Pekanbaru). *Skripsi thesis, Universitas Islam Negeri Sultan Syarif Kasim Riau*.

BKKBN. (2021). *BKKBN*. Dipetik July 22, 2021, dari https://www.bkkbn.go.id/detailpost/kontrasepsi-bisa-cegah-stunting

Bramer, M. (2007). *Principles of Data Mining*. London: Springer-Verlag.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Jahromi, A. H., & Taheri, M. (2018). A non-parametric mixture of Gaussian Naïve Bayes classifiers based on local independent features. *IEEE*.

Naafian, N. R., Siswanti, S., & Saptomo, W. L. (2017). SISTEM PENDUKUNG KEPUTUSAN PEMILIHAN METODE KONTRASEPSI DI PUSKESMAS II COLOMADU DENGAN ALGORITMA NAÏVE BAYES. *Jurnal TIKonSiN*.

Netti, K., & Netti, K. (2015). A Novel Method for Minimizing Loss of Accuracy in Naïve Bayes Classifier. *IEEE*.

Nugroho, C. G., Nugroho, D., & Fitriasih, S. H. (2015). Sistem Pendukung Keputusan Untuk Pemilihan Metode Kontrasepsi Pada Pasangan Usia Subur Dengan Algoritma K-Nearest Neighbour (KKN). *Jurnal Ilmiah SINUS*.

Nurlaela, D. (2020). Penerapan Adaboost Untuk Meningkatkan Akurasi Naïve Bayes Pada Prediksi Pendapatan Penjualan Film. *INTI NUSA MANDIRI*.

Rohman, A., Suhartono, V., & Supriyanto, C. (2017). Penerapan Algoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung. *Jurnal Teknologi Informasi*.

RRI. (2021). *rri.co.id*. Dipetik July 22, 2021, dari https://rri.co.id/daerah/1013175/ratusan-anak-kota-madiun-alami-stunting

Wardhani, D., Nurdini, Y., & Bayhaqi. (2014). Sistem Pendukung Keputusan Dalam Penentuan Alat Kontrasepsi Untuk Keluarga Berencana Dengan Pemodelan Logika Fuzzy. *Seminar Nasional Teknologi Informasi dan Multimedia*.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*.

Zaki, M. J., & Jr., W. M. (2013). *Data Mining and Analysis: Fundamental Concepts and Algorithms.* Belo Horizonte: Cambridge University Press.