Indonesian Journal of Mathematics Education
Vol. 4, No. 2, October 2021, pp: 66~73
p- ISSN: 2654-3907, e-ISSN: 2654-346X
e-mail: ijo m e@untidar.ac.id, website: journal.untidar.ac.id/index.php/ijome

Comparison and Prediction of Data Mining Models to Determine the Classification of Family Planning Program User Status

Ariska Kurnia Rachmawati[1a)], Minhayati Saleh[2b)], Muhammad Nur Ramdani[3c)]
[1 ,2,3] Walisongo Islamic State University Semarang, Jl. Prof. Hamka No. 3-5, Semarang, Indonesia
e-mail: [a)] ariskakurnia@walisongo.ac.id, [b)] minhayati@walisongo.ac.id
[c)] muhammadnurramdani9@gmail.com ,

Abstract

Data mining is the process of analyzing a sample to determine the best performing algorithm. An easy way to extract information or insights from large amounts of data is by using the techniques involved in data mining. There are several methods of classification which can be used to determine the level of a certain acuity. In Indonesia, family planning program is the most common program in the government to control population growth. A decision tree, a logistic regression, a naive Bayes model, and a gradient boosted model are used in this study. To perform classification family planning program User Status in Mangunharjo sub-district, the variable used is the wife's age, age of the youngest child, stages of prosperous family, and number of children. The training data comparison testing is 70:30. This study was tested by using the AUC value and *t*-test. The best value for accuracy is the Decision Tree algorithm with a percentage of 94.2% and an AUC value of 0.939. From the results of this test, it can be concluded that for a comparison of all tests performed on the dataset, the Decision Tree algorithm model can be said to be better than the other three algorithm models.

Keywords: classification, data mining, decision tree, family planning program, nave Bayes, rapidminer

INTRODUCTION

Indonesia is the fourth most populous country in the world after China, India and the United States. Based on SP2020, the total population of Indonesia in September 2020 was 270.20 million people. The annual population growth rate in Indonesia during 2010 to 2020 averaged 1.25 percent, slower than the period 2000 to 2010 of 1.49 percent (Badan Pusat Statistik, 2021). Although the rate of population growth in Indonesia continues to decline, the population continues to increase. The consequence of high population growth but not accompanied by adequate facilities and infrastructure is the inability to achieve a prosperous life.

Efforts to control population growth are carried out through the Population, Family Planning and Family Development Program in order to create quality small families, and are also expected to contribute to changes in population quantity marked by changes in the number, structure, composition and distribution of a balanced population in accordance with the carrying capacity and environmental capacity (*Rencana Kerja (Renja) Dinas Pengendalian Penduduk Dan KB Kota Semarang Tahun 2020*, 2020) .

Family Planning is a government program that was first formed on June 29th, 1970, in conjunction with the establishment of the National Family Planning Coordinating Board. The family planning program aims to be one of the government's efforts to control and suppress the rate of population growth and improve maternal and child health. In this study, based on the results of the 2020 Mangunharjo Village apparatus survey taken in August 2021, 76% or 2378 women have not participated in the family planning program and 24% or 751 women have participated in the family planning program. There are many factors that influence people in using family planning programs. including the age of the wife, the number of children they have, the age of the youngest child, the employment status of husband and wife, education level of husband and wife, and stages of a prosperous family.

In addition, due to the large number of people in Mangunharjo Village, which can reach hundreds for family planning user data, it is important to reveal valuable information from the data. Then, to assist in observing this important data, a technique is needed in order to be able to explore the data. Data mining is defined as the process of extracting or extracting the required knowledge from a large amount of data. In this process, data mining will extract valuable insights by analyzing certain patterns or relationships from big data (Han et al., 2011).

The purpose of this study is to compare the four data mining classification algorithms used, namely Decision Tree (C4.5), Naïve Bayes, Logistic Regression and Gradient Boosted Trees. These four algorithms are used to classify the status of family planning users in Mangunharjo village, Semarang city.

METHOD

Data Types and Sources

The sample is an element of a population and has special characteristics for that population (Abdullah, 2015). The data source used is secondary data. The secondary data used in this study are documents sourced from the results of the 2020 Mangunharjo Village officer survey obtained in August 2021, amounting to 751 people. The sample size will be calculated using the Slovin Formula. The Slovin includes an element of inaccuracy due to sampling errors that can still be tolerated (Abdullah, 2015). The tolerance value used in this study is 5%. Sample calculation using the slovin formula is obtained from the following calculation,

$$n = \frac{N}{1\,(\,N\,\alpha^2)} = \frac{751}{1\,(\,751\,.0{,}05^2)} = \frac{751}{1\,(\,1{,}8775)} \quad (1)$$

$$n = 400\,.$$

The sampling technique used in this study is a simple random sampling technique with a total sample of 400 people obtained from the results of calculations using the Slovin formula. The sample data is divided into two classifications, namely 62% or a total of 248 people with the status of family planning program users and as many as 38% or a total of 152 people.

Data Variable

The variables used in this study are divided into two, namely the dependent variable and the independent variable. The dependent Variable is Y = KB User Status, where 0 = Not a KB user; 1 = User KB; and the independent variables show in Table 1.

Table 1. Independent Variables

No	Variable	Description
1.	X1 : Wife's Age	1 = x ≤ 25 Years 2 = 25 < x ≤ 35 Years 3 = x > 35 Years
2.	X2 : Number of Children	0 = x ≤ 2 Children 1 = x > 2 Children
3.	X3 : Age of the Smallest Child	0 = x ≤ 10 Years 1 = 10 < x ≤ 20 Years
4.	X4 : Prosperous Family Stage	1 = Stages of a prosperous family 1 2 = Stages of a prosperous family 2 3 = Stages of a prosperous family 3

Data Processing Techniques

Data Mining is a slice of several scientific fields that unites techniques from machine learning, pattern recognition, statistics, databases, and visualization for handling problems of retrieving information from large databases (Larose & Larose, 2014). Data Mining is an automatic analysis of large or complex data with the aim of obtaining patterns or trends that are often not aware of their existence. In simple terms, data mining is the process of extracting or exploring the existing knowledge in a set of data (Romero et al., 2010). Data mining uses a pattern matching approach and other algorithms used to determine key relationships in the explored data.

The stages of data mining according to Sumathi & Sivanandam (2006) are as follows: 1) Data cleaning: The steps taken in data cleaning are the process of removing noise and inconsistent or irrelevant data; 2) Data Integration: This process is the process of merging data from various databases into a new database; 3) Data Selection: The data contained in the database is often not entirely used. Thus, only the appropriate data for analysis will be retrieved from the database; 4) Data Transformation: This stage is the stage where the data is converted or combined into an appropriate format for processing in data mining; 5) Mining Process: It is the main process when methods are applied to find important and hidden knowledge from data. The mining process carried out is the selection of the Decision tree, (C4.5), Naïve Bayes algorithms, Logistic Regression and Gradient Boosted Trees to find patterns or knowledge obtained from beyond the family planning program user data. In this process the calculation is assisted using RapidMiner Studio 9.10 software.

Decision Tree

Decision Tree is one of the most widely used classification methods and has a clear and easy-to-understand concept. The decision tree method converts very large data into a decision tree that represents the rules (Suyanto, 2017). The decision tree generated from the training process can explain how the data classification process works and is easy to implement using a recursive algorithm.

Decision Tree algorithm consists of a collection of nodes (*nodes*) connected by branches, where the branch moves from under the *root node. and ends at the leaf* node. The *leaf* node contains a final decision or target class for a *decision tree* . While the *root node* is the starting point of a *decision tree*. And there is one important node, namely an intermediate node that connects to a question or test (Rokach & Maimon, 2014). The stages of tree formation for both regression and classification are determining the root node, placing the dataset on the root node, dividing (splitting) the dataset into subsets, determining the decision node, the process is repeated until the stopping criteria is reached (Han et al., 2011).

Naive Bayes Algorithm

The Naïve Bayes algorithm is one of the classification algorithms used to predict the probability of membership of a class (Han et al., 2011). The Naïve Bayes algorithm or the so-called Naïve Bayes Classifier comes from the Bayes theorem discovered by Thomas Bayes in 1770. The Bayes theorem is a theorem with two different interpretations. Bayes' theorem states how far the degree of subjective belief must be rationally changed when given new instructions (Melinda et al., 2020).

Bayes rule is used to calculate the probability of a class. The Naïve Bayes algorithm provides a way of combining the previous probabilities with possible conditions into a formula that can be used to calculate the probability of each possibility that occurs (Jadhav & Channe, 2016).

Logistic Regression

Binary logistic regression is a data analysis method used to find the relationship pattern between the dependent variable (y) in the form of a binary variable and the independent variable (x) (Hosmer Jr et al., 2013). Because logistic regression predicts probabilities, rather than just classes, the method can be used is likelihood method.

Suppose we have a sample of n independent observations of the pair $(x_i, y_i), i = 1,2,\dots,n$ where y_i denotes the value of a dichotomus outcome variable and x_i is the value of the independent variable for the i^{th} subject. The probability of that classes was either π if $y_i = 1$, or $1 - \pi$ if $y_i = 0$. The probability function for each observation is given as follows:

$$f(\beta, x_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} ; y = 0,1 \qquad (2)$$

where $\pi(x_i) = \frac{e^{(\sum_{j=0}^{p} \beta_j x_{ij})}}{1+e^{(\sum_{j=0}^{p} \beta_j x_{ij})}}$; if $j = 0$ then $x_{ij} = x_{i0} = 1$.

Thus, the log likelihood probability function is

$$L(\beta) = \ln\big(l(\beta)\big) = \ln\left(\prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}\right)$$

$$= \ln\left\{\prod_{i=1}^{n}\left(1 + e^{\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}\right)^{-1}\right\} e^{(\sum_{j=0}^{p} \beta_j x_{ij})} \qquad (3).$$

Gradient Boosted Trees

The Gradient Boosted Trees method was introduced by J.H. Friedman for building decision trees using a greater degree of freedom in structure and having a higher level of learning for optimizing outcomes and minimizing overfitting (Friedman, 2001). Gradient Boosted Trees are supervised learning methods based on decision trees. An algorithm gradient boosted tree works in a sequential way to improve existing predictors that are not consistent with current predictions, thereby verifying that adjustments made earlier are effective.

The negative log likelihood is the loss function that the gradient boosted classification tree algorithm attempts to reduce. The following formula expresses the negative log likelihood for the ith training example in a classification problem with two classes, where p denotes the probability that the example belongs to class 1, as follows:

$$-[y_i * log(p) + (1 - y_i) * \log(1 - p)] \quad (4)$$

To determine the overall loss, the loss is added across all training cases. The odds ratio, also known as the ratio p/(1-p), is sometimes stated in terms of log(odds) by taking the odds ratio's natural log. The following equation relates the log(odds) to the probability p of an event

$$p = \frac{e^{\log(odds)}}{1+e^{\log(odds)}} \quad (5)$$

In this method, a new test is produced by analyzing a test's asymmetry, and then the new test is made by minimizing the function of the test (Natekin & Knoll, 2013):

$$-\log L1 = -\sum_{i=1}^{N} y_i log(odds) + log\,(1 + e^{\log(odds)}) \quad (6)$$

Validation Model

This study uses a dataset distribution ratio of 70%: 30%, where the training data is 280 and the testing data is 120.

Pattern Evaluation

The goal is to identify interesting patterns into the knowledge found. In this research, the evaluation test of the classification model uses the accuracy and AUC. The accuracy of the classification model is measured using formula as follows (Agarwal, 2014),

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (7)$$

The accuracy value in this research was obtained from the confusion matrix table of processing results using RapidMiner.

Secondly, AUC (Area Under the ROC Curve) is used as a performance measure in this study to evaluate the quality of the classification process. Area under the ROC curve (AUC) is a useful metric for classifier performance since it is independent of the decision criterion selected and prior probabilities (Rokach & Maimon, 2014). The AUC indicates that areas adjacent to the ROC will employ the AUC to increase cross-sectional comparables. The scale of AUC performance qualification is 0 to 1, where the number 0 indicates negative level and the number 1 indicates the positive level.

This study used a *t*-test as a method of evaluation. The *t*-test is used to determine the results of the differences in the performance of the algorithm being compared (Hui & Zongfang, 2013). Based on the precondition that the variables collectively follow a normal distribution, the t-test is used to determine if the means of the variables in default and non-default are equal. It aims to determine whether the variables are predictive. The formula of t-test as follows,

$$t = \frac{D}{S/\sqrt{n}} \quad (8)$$

Where *D* is the mean difference, *S* is the sample variance.

Knowledge Presentation

In the knowledge presentation, visualization and presentation of knowledge about the methods used to obtain the knowledge gained by the user is carried out.

RESULTS AND DISCUSSION

In this study, the pattern of family planning program user status will be classified using decision tree, logistic regression, Naïve Bayes and gradient boosted trees. Classification is a very important part in data mining. An objective of the model comparison is to compare the most reliable algorithms and to determine the highest level of accuracy by using *t*-tests and RapidMiner 9.10.

A series of stages of data preparation and initial data processing have been passed to prepare data that is truly valid before being processed at the next stage. The series of stages from the beginning to the end of the classification process are listed in Figure 1, and the stages of comparison of 4 data mining algorithms is shown in Figure 2.

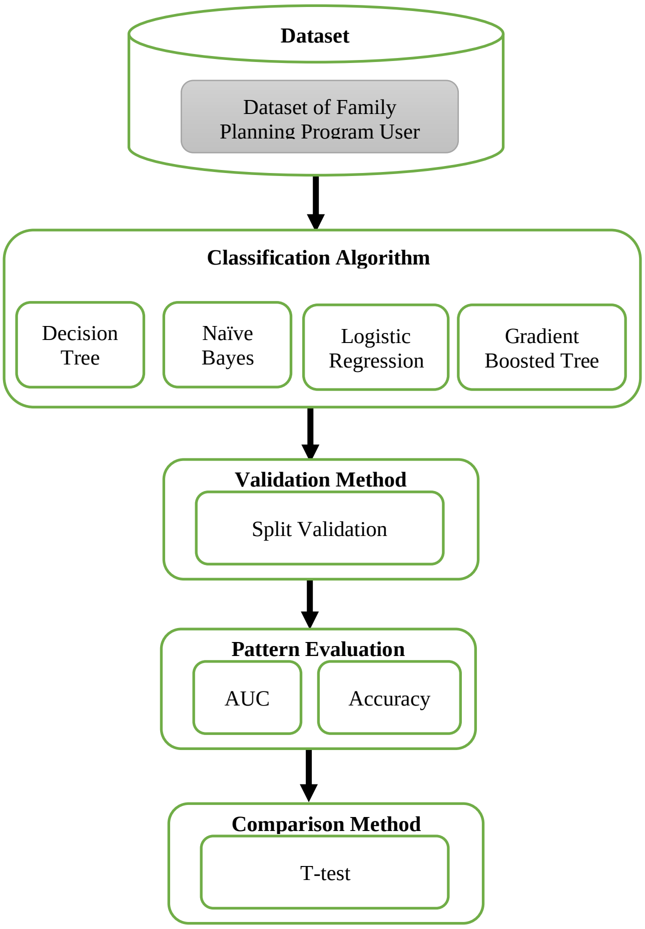


Figure 1. Classification Process Flowchart

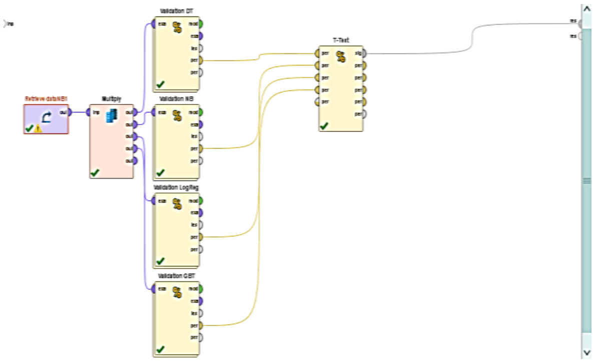


Figure 2. Stages of comparison of 4 data mining algorithms with RapidMiner 9.10

The next step is to classify using four data mining models, where the results of the classification process are evaluated using a confusion matrix and ROC Curve to measure performance or accuracy. To compare the best data mining model, Figure 3 shows that the Decision Tree algorithm produces the best classification model compared to the other three algorithms.

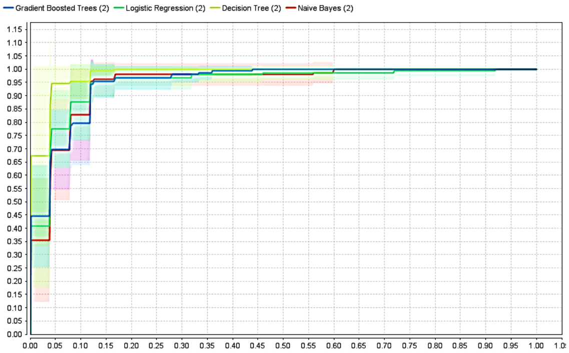


Figure 3. ROC Comparison of Data Mining Models

The test results with RapidMiner software produce ROC curves on the Decision Tree algorithm showing an accuracy value of 92,4% with an AUC value of 0.939.

Table 2. Performance Vector with Decision Tree Method

Accuracy : 92.4% ± 3.87%			
	True KB user	*True* Not KB User	*Class precision*
Pred. KB User	221	5	97.79%
Pred. Not KB User	27	147	84.48%
Class recall	89.11%	96.71%	

The ROC curve generated by the Naïve Bayes algorithm shows an accuracy value of 90.8% with an AUC value of 0.930.

Table 3. Performance Vector with Naïve Bayes Method

Accuracy : 90.8% ± 5.41%			
	True KB user	*True* Not KB User	*Class precision*

Pred. KB User	221	10	95.67%
Pred. Not KB User	27	142	84.02%
Class recall	89.11%	93.42%	

The Gradient Boosted Tree algorithm shows an accuracy value of 88.3% with an AUC value of 0.938.

Table 4. Performance Vector with Gradient Boosted Tree Algorithm

Accuracy : 88.3% ± 4.09%			
	True KB user	*True* Not KB User	*Class precision*
Pred. KB User	221	20	91.70 %
Pred. Not KB User	27	132	83.02%
Class recall	89.11%	86.84%	

While the logistic regression model shows an accuracy value of 91.3% with an AUC value of 0.939.

Table 5. Performance Vector with Logistic Regression

Accuracy : 91.3% ± 4.22%			
	True KB user	*True* Not KB User	*Class precision*
Pred. KB User	221	9	96.40 %
Pred. Not KB User	27	143	85.50%
Class recall	89.11%	94.54%	

Of all the algorithms, the AUC value is in the range 0.90 – 1.00, which means that the data mining algorithm produces a very good predictive model.

Table 6. Comparison of Algorithm Performance

	Decision Tree	Naive Bayes	Logistic Regression	Gradient Boosted Tree
Accuracy	0.924	0.908	0.913	0.883
AUC	93.9%	93%	93.9%	93.8%

The results of the comparison of the AUC and accuracy values to the four data mining algorithms in classifying the status of family planning program users are listed in Figure 4 and Figure 5.

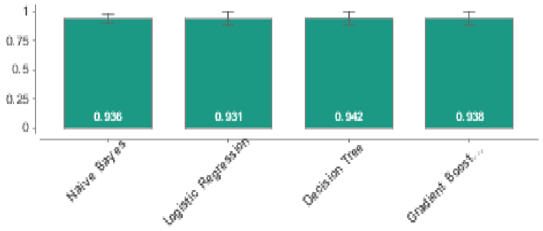


Figure 4. Comparison graph of AUC values

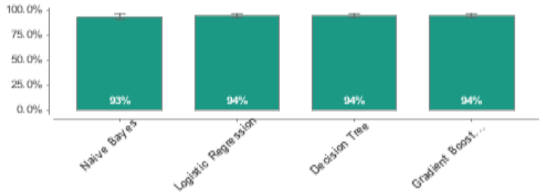


Figure 5. Comparison graph of accuracy values

The next *t*-Test test is carried out with the aim of testing which classification algorithm is the best, where in the test until the smallest value 0.05 is declared as the best test result (Agarwal, 2014).

Table 7. *t*-test statistic test

Algorithm	Decision Tree	Naive Bayes	Logistic Regression	Gradient Boosted Tree
Decision Tree	-	0.501	0.041	1.000
Naive Bayes	0.501	-	0.127	0.518
Logistic Regression	0.041	0.127	-	0.055
Gradient Boosted Tree	1.00	0.518	0.055	-

In the significance test the value uses a statistical significance level of 0.05 which means that if it is statistically less than 0.05 it shows a significant difference between the average values, thus it must reject the null hypothesis (H_0). Based on Table 7, it is known that there is a significant difference (H_1) between Decision Tree and Logistics Regression because the significance value is 0.041 < 0.05. H_0 explains that there is no significant difference between the Naïve Bayes algorithm and the Gradient Boosted Tree.

From the *t*-test and AUC tests above, the Decision Tree algorithm has the best performance compared to the other three algorithms in classifying user data for family planning programs in Mangunharjo Village, Semarang.

In terms of testing data on training data, it is said to be quite successful because the resulting percentage reaches more than 90%. This occurs because the data used for testing is the training data, which is the data that is processed to create the classification model. The next step is to predict the testing data for the four algorithms. Subsequent testing was carried out on testing data which amounted to 30% or 120 data.

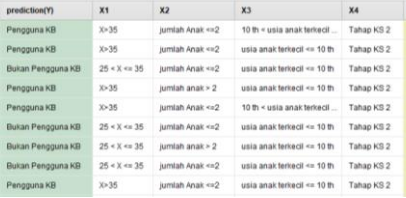
Table 8. Comparison of Algorithm Performance for Testing Data

	Decision Tree	Naive Bayes	Logistic Regression	Gradient Boosted Tree
Accuracy	0.900	0.867	0.900	0.842
AUC	89.9%	88.8%	88.8%	89.2%

Here are a few instances of predictions made using the four methods,

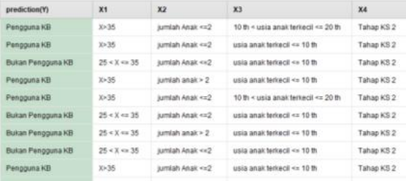
Table 9. Prediction of Fours Algorithm

Decision Tree



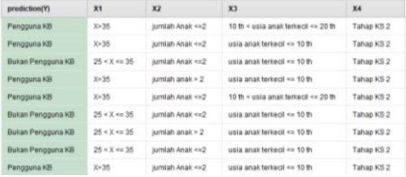
prediction(Y)	X1	X2	X3	X4
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil ...	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil ...	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2

Naïve Bayes



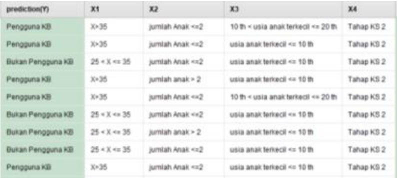
prediction(Y)	X1	X2	X3	X4
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil <= 20 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil <= 20 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2

Logistic Regression



prediction(Y)	X1	X2	X3	X4
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil <= 20 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil <= 20 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2

Gradient Boosted Tree



prediction(Y)	X1	X2	X3	X4
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil <= 20 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	10 th < usia anak terkecil <= 20 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah anak > 2	usia anak terkecil <= 10 th	Tahap KS 2
Bukan Pengguna KB	25 < X <= 35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2
Pengguna KB	X>35	jumlah Anak <=2	usia anak terkecil <= 10 th	Tahap KS 2

CONCLUSION

Comparison of classification against datasets can be said to be not easy in terms of choosing the algorithm, because not all types of data can support the algorithm model even though the model is included in the classification.

Four algorithm models used in the comparison include Decision tree, Naïve Bayes, Logistic Regression and Gradient Boosted Tree. The best accuracy value is the Decision Tree algorithm with a percentage of 92.4% and an AUC value of 0.939. From the results of this test, it can be concluded that for a comparison of all the tests carried out on the dataset, the Decision Tree algorithm model can be said to be better than the other three algorithm models for predicting the classification of family planning program user status.

REFERENCES

Abdullah, M. (2015). *Metode penelitian kuantitatif* (1st Editio). Aswaja pressindo.

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Badan Pusat Statistik. (2021). *Hasil sensus penduduk 2020*. https://www.s.go.id/pressrelease/2021/01/21/1854/hasil-sensus-penduduk-2020.html

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Waltham, MA: Elsevier.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). Hoboken, NJ: John Wiley & Sons.

Hui, L., Li, S., & Zongfang, Z. (2013). The model and empirical research of application scoring based on data mining methods. *Procedia Computer Science*, *17*, 911–918. https://doi.org/10.1016/j.procs.2013.05.116

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research*, *5*(1), 1842–1845. https://doi.org/10.21275/v5i1.nov153131

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). Hoboken, NJ: John Wiley & Sons.

Melinda, V., Primartha, R., Wijaya, A., & Jambak, M, I. (2020). Optimization naive bayes algorithm using particle swarm optimization in the classification of breast cancer. *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, pp. 362–369. https://doi.org/10.2991/aisr.k.200424.055

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(21), 1-21. https://doi.org/10.3389/fnbot.2013.00021

Rokach, L., & Maimon, O. (2014). *Data Mining With Decision Trees: Theory and applications* (2nd Ed). Toh Tuck Link, Singapore: World Scientific.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. Sj. (2010). *Handbook of educational data mining*. Boca Raton, Florida: CRC press.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining and its applications*. Heidelberg: Springer Berlin. https://doi.org/10.1007/978-3-540-34351-6

Suyanto, D. (2017). *Data Mining untuk klasifikasi dan klasterisasi data*. Bandung: Informatika Bandung.