Journal of Theoretical and Applied Information Technology
31st December 2022. Vol.100. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CART METHOD APPROACH AND HIGH DIMENSION SIMULATION DATA SELECTION AND RANDOM UNDER-SAMPLING METHOD IN STUNTING CASE

WIDIARNI GINTA SASMITA[1*], WAEGO HADI NUGROHO[2], AND ANI BUDI ASTUTI[3]

[1]Research Scholar, University of Brawijaya, Malang, Indonesia

[23]Professor, University of Brawijaya, Malang, Indonesia

Email: [1]widiarnisasmita@student.ub.ac.id, [2]whn@ub.ac.id, [3]ani_budi@ub.ac.id

ABSTRACT

Stunting has become the main problem in every area for repair and investigation. Data modeling with To do classification could help for however many variables big is problematic in modeling classification and can complicate the interpretation process. Data modeling with an amount of enough significant variables could handle the selection process stepwise method. This study will create a classification model using tree decision *Classification and Regression Tree* (CART) with challenge amount variable predictor. A total of 26 variables and 650 observations were applied using the simulation data taken from real stunting data in Java east and the data used on the variable predictor which is stunting and normal categories. The result of the study is a method used that could propose selected variables in the stunting process with high accuracy. The acquired model has decisive information related to influencing factors stunting incidents by grouping family internal factors namely the health of parents who divides the knot root model decision as factor main in stunting incident.

Keywords: *CART, Stepwise, Stunting, Height Dimension*

1. INTRODUCTION

Problem stunting in children toddlers, so that have bodies short compared child her age, the thing Becomes challenge big deal in the Indonesian nation. Global Nutrition Report 2018 shows The prevalence of stunting in Indonesia among 132 countries is ranked 108th. So that has become a national concern in problem stunting. So that thing Becomes the focus President of the Republic of Indonesia in the development year 2024, the repair process can be handled through factor direct in the form of the intake of food child as well as the mother's pregnancy, and focus on a supportive environment is worthy or no, so that reduce the impact of stunting on toddlers.

Based on the results of Verawati Simamora's research in 2019, there are many factors that can cause cases of stunting in children. The causes of stunting can be caused by direct factors or even indirect factors. Where on average the direct causes of stunting are due to lack of nutritional intake and the presence of infectious diseases while indirect factors cause stunting, namely education, low mother's knowledge, family economy, nutritional status, water sanitation, and the environment. The problem of malnutrition often gets attention in various developing countries including underweight, stunting, wasting, and micronutrient deficiencies. (Maulina, 2021).

Other research states that the causal factors for stunting occur since pregnancy as a result of a lack of nutrition during that pregnancy, early initiation of breastfeeding less than 1 hour after birth or not at all, breastfeeding stopping for about 6 months and the frequency of breastfeeding is not long enough. as well as providing food that accompanies breastfeeding for approximately 6-12 months, and the food given does not vary with frequency and texture that is not suitable for age (Anggryni et al, 2021).

Variables used in the study this use factors direct and not direct, focus environment in the form of MCK, use of clean water, as well as a factor of the economy that affects incident stinking for improving sustainability in a country. Besides, Factors that influence stunting are the area where stay, pattern care, and economic status impact the nutritional quality of the child. Problem the many factor possibility come up with ideas for see factor biggest incident stunting in East Java with

classifying factor with merging method statistics that is tree classification.

Tree decision is one method of frequent classification used because have many variables free. The regular algorithm used in tree decisions includes ID3, CART, and C4.5 [3]. Tree classification uses data mining to extract data as information and pattern, that is, decision trees and Neural Networks. Decision tree technique C.45 in the learning process can change the fact in form tree more decisions easily understood next validation against testing data to find the best models from decision tree [4]. CART Algorithm with GUIDE approach to building tree model regression, however method the applies to missing value data points time in the variable predictor, and tree classification single have traits that don't stable [5]. Instability because changes small on learning will influence results high accuracy, so develop technique bagging on a tree CART classification to increase accuracy the resulting classification [6]. One particular strategy in dimensional data classification tall is to reduce dimensions. The approach in reduction dimensions shared become two : extraction variable and selection variable. Several techniques of statistics could use in the selection process variable before the classification process is carried out. One of them is stepwise regression with numerical data characteristics. Based on function, stepwise regression is a technique of possible regression chosen with a careful variable predictor that will influence the accuracy of the variable response. Research studies develop method classification with some influencing variables stunting so that the selection process is carried out variable. Selection process variable conducted To use make it easy to interpret simple models as form from development technology in the field of health as well as statistics. Destination from the study this is how effectiveness use method Stepwise and Classification and Regression Tree (CART) on simulation data.

2. STUDY LITERATURE

Machine Learning

Machine learning is a study that studies computational algorithms for several purposes, such as filtering, classifying, or detecting images or videos. Machine learning is divided into three categories, namely unsupervised learning, supervised learning, and reinforcement learning. Reinforcement learning is one of the techniques in machine learning that learns something by taking certain actions and seeing the results of those actions (learning based on previous experiences). In machine learning, SVM falls into the category of supervised learning with certain algorithms that analyze data for classification (Mitchell, 1997).

- Structure Decision Tree

Tree decisions build a classification model or regression in the form structure tree. Tree decision split the data set into set more parts small and homogeneous. Tree decisions make a classification model or regression in the form structure tree. Tree decision split the data set into set more parts small and homogeneous.

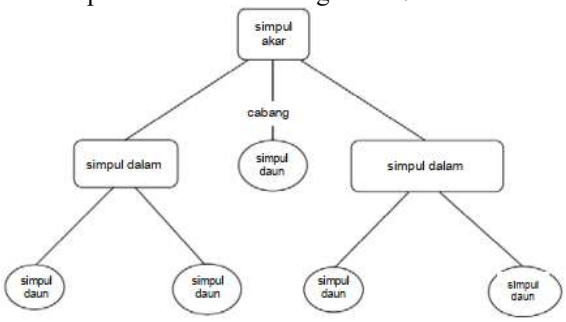

Figure 1 . Decision Tree Structure (Source: [7])

Classification And Regression Tree (Cart) Algorithm

CART is one of the algorithms used for technique tree decisions. The CART algorithm looks for all possible value that gives sorting the best with drop level the highest heterogeneity. An election was conducted to sort data into two groups: knot left and knot right. The election knot will Continue until obtained terminal node loading relative data set is homogeneous.

Election Knot

In forming tree classification, there are stages: stages election knot or election sorter. The CART algorithm is presented in summary as follows [8]

Step 1: Calculate the Gini index using the formula

$$Gini\ (D) = 1 - \sum_{i=1}^{m} p_i^2 \quad (2.1)$$

Step 2: Calculate the Gini index if binary partitioned by a separator using the formula

$$Gini_s(D) = \frac{|D_1|}{|D|} Gini\ (D_1) + \frac{|D_2|}{|D|} Gini\ (D_2) \quad (2.2\ ).$$

Step 3: Calculate the *impurity reducer* ($\Delta\ Gini\ (s)$) by formula

$$\Delta\, Gini\ (s) = Gini\ (D) - Gini_s(D)$$ (2.3).

Step 4: Repeat steps 2 and 3 on all variable predictors and find $\Delta\, Gini\ (A)$ the maximum. Variable The predictor that has a maximum is used as a separator.

Step 5 : Repeat step 2 to step 4 until the maximum tree is obtained.

Labeling Knot

Stage next is the determination terminal node if there is drop heterogeneity which means in sorting, only there is one observation (n=1) on each knot child or existence minimum limit, and there is a limit on the number of levels, or maximum levels tree regression.

Opportunity from class labelling on node presented in equation (2.1).

$$p(j_0, t) = \max_j \frac{N_j(t)}{N(t)}$$ (2 .1)

Where

$p(j_0, t)$: class proportion j on node$t$

$N_j(t)$: number of class observations $j$at node

$N(t)$: the number of observations at the node$t$

Then the label for the terminal node t is j_0.

Tree complex decisions and havemost terminal node called with tree maximum (T_{max}). The parameter to measure the complexity of a decision tree is called the *complexity parameter* (CP). Moreover, searching for score *resubstitution estimate* and *relative error to get CP value. Resubstitution estimate* is symbolized by calculated using the formula in equation Tree complex decisions and have most terminal node called with tree maximum. The parameter to measure the complexity of a decision tree is called the *complexity parameter* (CP). Moreover, searching for score *resubstitution estimate* and *relative error to get CP value. Resubstitution estimate* of T_t, which is symbolized by $R(T_t)$ calculated using the formula in equation

$$R(T_t) = \frac{1}{N}\sum_{i=1}^{N} I(T_t(x_i) \neq y_i)$$

The *relative error value* is defined in equation (2.2) .

$$Re(T_t) = \frac{R(T_t)}{R(T_1)}$$ (2.2)

Where

$Re(T_t)$: *relative error* on *subtree*T_t

$R(T_t)$: *resubtition estimate* on *subtree* T_t

$R(T_1)$: *resubtition estmate* in the first decision tree (a decision tree consisting of only the root node)

The measure *of complexity parameter* is defined by equation (2.3) .

$$cp_t = \frac{Re(T_t) - Re(T_{t+1})}{nsplit\ (T_{t-1}) - nsplit\ (T_T)}$$ (2.3)

Where

cp_t : complexity *parameter t*

$Re(T_t)$: *relative error* in *subtree*T_t

$nsplit(T_t)$: number of selections on *subtree*T_t

The value shows no pruning, which means the *subtree* is the maximum tree. According to Han, et al . (2011), getting an Optimum tree decision is to use the 1 SE rule.

The method used for validation cross tree the decision on R is *K-fold cross-validation* with K=10. *Relative error* result cross calculated use formula equation (2.4). Temporary deviation standard and *standard error* in *the subtree* are computed using equations (2.5) and (2.6).

$$CV(T_t) = \frac{1}{K}\sum_{k=1}^{K} Re(T_t^{(k)})$$ (2.4)

$$SD(T_t) = \sqrt{var\left(Re\left(T_t^{(k)}\right)\right) \ldots Re(T_t^{(k)})}$$ (2.5)

$$SE(T_t) = \frac{SD_k(T_t)}{\sqrt{K}}$$ (2. 6)

The formula 1 SE rule is presented in equation (2. 7)

$$CV(T_t) \leq CV\left(\hat{T}_t\right) + SE\left(\hat{T}_t\right)$$ (2.7)

where $\left(\hat{T}_t\right)$= argmin$CV_k(T_t)$

Regression logistics is one method used to carry out the analysis process connection Among variables not free who have nominal or ordinal scale with variable free. Variable response in analysis logistics in the form of a categorical or qualitative and variable predictor in the form of qualitative and quantitative, defined as follows.

$$Z = \begin{cases} 1, sucess \\ 0, fail \end{cases}$$

By opportunities that occur $Pr\ Pr\ (Z = 1) = \pi$and $Pr\ Pr\ (Z = 0) = 1 - \pi$. If there are many $n$mutually independent $Pr\ Pr\ (Z_i = 1) = \pi_i$random variables with $Z_1, \ldots, Z_n$, then the joint chance is expressed in equation (2.8).

$$\prod_{i=1}^{n} \pi_i^{Z_i}(1-\pi)^{1-\pi} = \exp\left(\sum_{i=0}^{n} z_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \log(1-\pi_i)\right)$$ (2.8)

Basically general , regression model logistics specified as function x in equation (2.9) [9].

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \quad (2.9)$$

For simplify the process of estimating the regression parameters $\pi(x)$so that it becomes the logit form in equation (2.10).

$$L(\beta) = \ln[l(\beta)]$$
$$= \sum_{i=1}^{n}\{y_i\, ln[\ \pi(x_i)\ ] + (1-y_i)\ln[1-\pi(x_i)]\}$$
$$= \sum_{i=1}^{n}\{y_i\, ln[\ \pi(x_i)\ ] + \ln\big(1-\pi(x_i)\big) - y_i \ln[1-\pi(x_i)]\}$$
$$= \sum_{i=1}^{n}\{y_i\, ln[\ \pi(x_i)\ ] - y_i \ln[1-\pi(x_i)] + \ln\big(1-\pi(x_i)\big)\}$$
$$= \sum_{i=1}^{n} y_i[\, ln\, \pi(x_i)\ -\ \ln(1-\pi(x_i))] + \ln\big(1-\pi(x_i)\big)$$
$$= \sum_{i=1}^{n} y_i[\, ln\frac{\pi(x_i)}{1-\pi(x_i)}] +\ \ln(1-\pi(x_i))$$
$$= \sum_{i=1}^{n} y_i[\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j] + \ln(\frac{1}{1+e^{\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j}})$$
$$= \sum_{i=1}^{n} y_i[\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j] -\ \ln(1+ e^{\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j}) \quad (2.10)$$

In equation (2.10) $g(x)$is the form of logit

Regression logistics do not assume that variable predictor multivariate normal distribution with matrix covariance; however, on the contrary, regression logistics assume that variable response binomial distribution describes distribution from same *error* with the score. In addition, logistic regression requires that the predictor variables are independent, not multicollinear, which means that the predictor variables are not linearly related.

Size association could show dependencies or dependency Among variable usual response called with *odds ratio*the value used to interpret coefficient on regression model binary logistics. The equality *odds ratio* for for $x = 1$and $x = 0$ and can be written in equation (2.11).

$$\Psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \quad (2.11)$$

The value of the *odds ratio* can be written in equation (2.12).

$$\Psi = \frac{(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}})/(\frac{1}{1+e^{\beta_0+\beta_1}})}{(\frac{e^{\beta_0}}{1+e^{\beta_0}})/(\frac{1}{1+e^{\beta_0}})} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_0} \quad (2.12)$$

Analysis *stepwise* done with model connection Among variable predictor with variable response. Method *Stepwise Regression* or called with regression purposeful level for election variable gradually. Stage election variables include adding and removing variables based on score significance statistics. Stages in the selection process are as follows [10]

a. Shaping matrix coefficient correlation Among Y by $X_i$using equation (2.12).

$$r_{yx_i} = \frac{\sum(X_{ij}-\bar{X}_i)(Y_j-\bar{Y})}{\sqrt{\sum(X_{ij}-\bar{X}_i)\sum(Y_j-\bar{Y})^2}} \quad (2.12)$$

b. Chosen variable predictor with highest correlation with Y

c. Count equality regression logistics with $\bar{Y} = f(X_1)$ whether the variable is real; if it is not real, then the process stops and takes the model from $Y = \bar{Y}$ as the best model. If the variable is real, look for the next variable with the highest correlation value after the el variable first for testing.

d. Count equality regression $\bar{Y} = f(X_1, X_2)$ looking at the increase in the decrease in the residual square, $R^2\ adj$and the F partial value for the regression significance test by testing the regression coefficient β_i, using the F test in equation (2.13) with the hypothesis :

H_0:$\beta_i = 0$
H_0:$\beta_i \neq 0$

$$F_{hitun} = \left(\frac{\beta_i}{s(\beta_i)}\right)^2 \quad (2.13)$$

With decision when $F_{hitung} < F_{tabel}$accept H_0then the process is terminated, if $F_{hitun} \geq F_{tabel}$rejected H_0then the variable X_2remains in the model.

e. Check return level significance stats moment is at in models. If have significance more from threshold with use comparison $F_{hitung} \geq F_{tabel}$then it will be maintained in models

The stages above are repeated until there is an appropriate variable put in and out of the formed model.

Regression logistics stepwise available good in regression logistics binary, multinomial and linear regression. The stepwise method has superiority for adding and reducing variable predictors by the level significance of the formed model, so to do this required amount of iteration that will be processing and analyzing subtraction as well as an additional variable.

3. DATA AND METHODS

Data

Research data is the generation data as much as 200 and 10,000. Data generated from actual data

that comes from a questionnaire study related to stunting in East Java with scenario

Table 1. Scenario Data

No	Proportion	Observation	
		$n_1 = 200$	$n_2 = 10.000$
Stunting	20%	40	2,000
Normal	80%	160	8.000

Method

Stepwise

To do selection variable predictor use method stepwise :

a. Correlate all variable predictors to the variable response.
b. Choose variable owned predictor correlation high and enter to in models.
c. Calculate AIC.
d. Choose variable explanatory having correlation highest second and enter to in the model and return to the stage (b).
e. If the AIC value in step next is more minor than the previous AIC, so variable explanation the enter in models.
f. Repeat steps (b) – (e) until all variable explanations are tested.

Cart

1. Comparing real data into training data and test data randomly.
2. Normalize − training data and obtain normalized training data.
3. Undersampling with RUS at for balanced training data. In the data selection process using the stepwise method.
4. Selecting predictors using the stepwise method.
   a. Correlate all explanatory variables to the response variables.
   b. Select the explanatory variable that has the correlation and enter it into the model.
   c. Calculating AIC.
   d. Choose the explanatory variable that has the second highest correlation and enter it into the model and return to step (b).
   e. If the AIC value in the next step is smaller than the previous AIC, the explanatory variable is included in the model.
   f. Repeat steps (b) – (e) until all explanatory variables are tested.
5. Determine the maximum iteration value using K-Fold Cross-validation. The set of selected predictors that have a maximum value is called 1.
6. Modeling the decision tree with the CART algorithm on with predictor 1 until the maximum tree is obtained.
7. Pruning trees according to the 1-SE rule. Trees obtained after pruning are called optimum trees.

4. EMPIRICAL RESULTS

The study process simulation carries out learning performance from the Stepwise method on CART with 650 observations with 26 variables. Shared data sets Become two, with an 80:20 portion of 520 as test data and 130 as training data. To reduce the problem is in the process of classification on dimensional data tall, so the selection process required using stepwise regression to do subtraction variable.

After the simulation, the next step is to look for correlation to see mutual variables relate and have closeness. The correlation model obtained as follows:

Table 2. Correlation Model

Variable	Estimate	Z-Value
X_2(BB during pregnancy)	3.2	0.659
X_3(Lila)	207.079	0.014
X_4(Blood pressure)	1,742	0.405
X_5(HB levels)	-2.462	0.000
X_6(Urine Protein)	019,622	0.000
X_7(gestational age)	52,562	0.000
X_9(Vegetable Consumption)	-321,504	-0.015
X_{10}(animal consumption)	22,153	0.005
X_{11}(Milk Consumption)	-130,354	-0.018
X_{15}(Father's tension)	-6,135	-0.008
X_{20}(TB)	-6.189	-0.009

AIC of the resulting model through an iterative process on correlation is of 31. 88

Journal of Theoretical and Applied Information Technology
31st December 2022. Vol.100. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

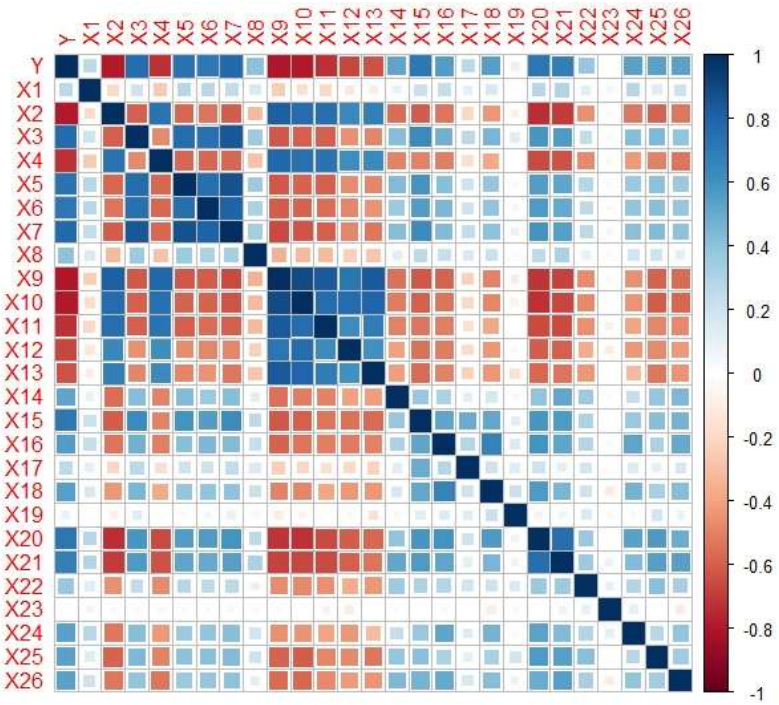

Figure 2. Correlation Visual All Variable

Figure 1 indicates the level of closeness in each variable by visualizing the correlation between variables with color density in each variable, the value of each variable can be seen in table 2.

Selection Variable

At stage selected variable stepwise regression _ as many as five variables through stages iteration 25 times with score optimum variable on table 3:

Tabel 3. Variabel Estimate

Variable	Estimate	Z-Value
X3	113.03	-0.000
X7	16.11	0.000
X9	-180.32	-0.007
X10	21.22	0.004
X11	-67.95	-0.006

Model produced :

$$y = -366.14 + 113.03X_3 + 16.11X_7 - 180.32X_9 + 21.22X_{10} - 67.95X_{11}$$

The selected variable consists of X_3 = Upper arm circumference, X_7=maternal gestational age, X_9=consumption of vegetable protein, X_{10}= consumption of animal protein, and X_{11}=consumption of milk, which are variable selected with seeing the lowest AIC value after compared with other models available through the selected AIC iteration process you amounted to 21,545 . AIC value in correlation model simulation is more considerable than model value after carrying out a stepwise model that signifies that selected variable already represents information.

Tree Classification and Regression

CART is a method of classification using training data and test data. Training data used for shape tree classification and test data as validation data ability tree classification for new data prediction

The separation process is carried out until not allowed again for breaking down tree classification maximum from training data. The separation will Keep going conducted so that it produces a complex tree called tree maximum.

Table 4. Selecting Optimal Subtree

CP	nsplit	Rel	Error	Xerror	xstd
first	0.897	1	0.000	1.122	0.071
2nd	0.010	2	0.102	0.142	0.037
3st	0.000	3	0.091	0.132	0.356

In the process of selecting the optimal subtree based on the denfan 1-SE rule, based on the table above, the 3rd subtree has a score $CV(\hat{T}_t)$ of 0.132 so that the optimum subtree can be determined using the formula:

$$CV(T_t) \leq CV(\hat{T}_t) + SE(\hat{T}_t)$$

$$CV(T_t) \leq 0.132 + 0.356$$
$$CV(T_t) \leq 0.488$$

The results above could serve optimal subtree of stunting data that has been conducted selection using stepwise as a variable predictor. Following is the optimal subtree based on cross-validation results

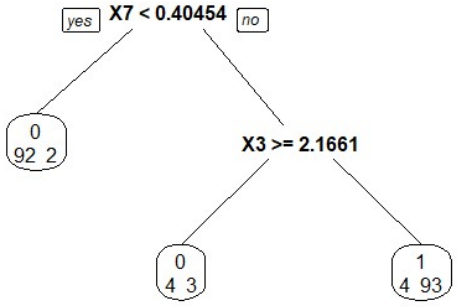

Figure 3. Optimal Subtree

Based on Figure 2 X_7 split, the root node shows that the split point of X_7 the heterogeneity maximizes the reduction. Based on the model obtained to determine whether a child is stunted or normal, first look at the value of X_7. If the value of X_7 is less than 0.04054 then the child could be categorized as not stunting if no will appear Other causes of checking variable X_3 if the expression X_3 is more than equal to 2.1661, then it can be said that it is not stunting. If it is less then it is categorized as stunting.

Model Evaluation

After getting the model shown in figure 1, a model evaluation process is carried out using training data to know whether overfitting occurs. The evaluation was conducted for measurement performance classification performed with the large area under ROC curve (AUC). AUC value for stunting data in the table under

Table 5. Value Of AUC

AUC Training Data	AUC Test Data
0.91	0.88

The model interpreted by the AUC l model on the training and test data has a value that is not too far, which signifies that the data is consistent with test data and training data. Could see from visualization AUC curve

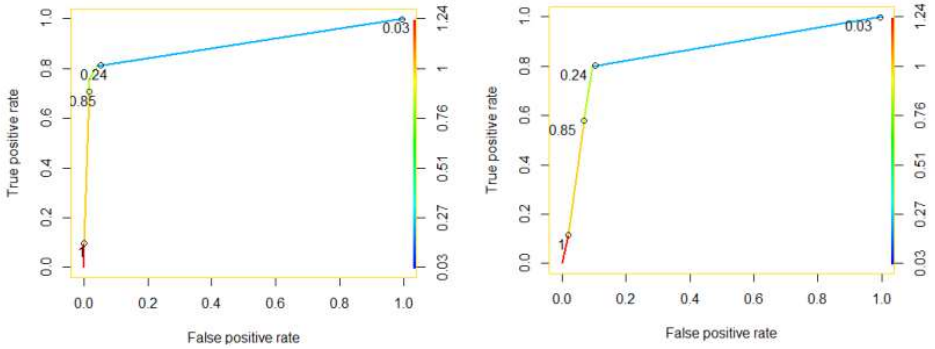

Figure 4.Grafik Of AUC Curve

Model Accuracy Is Done To See Accuracy With The Use Confusion Matrix. From The Confusion, The Matrix Could See Accuracy As Significant As 0.9179487.

5. CONCLUSION AND SUGGESTION

Previous studies used parametric analysis techniques where the conclusions from the results were more difficult to simplify. In contrast to this study, the results were in the form of a tree diagram and were easier to understand.

In the research, the selection method was carried out through two stages, namely correlation and stepwise regression, to handle the classification of high dimensions. The process that was followed was in accordance with the direction of previous research which carried out research without an initial selection process so as to produce a lower accuracy value, when going through the selection process using the stepwise method and the balance data process using the random under sampling method, a high accuracy value of 0.917 was produced and the model evaluation value the training data and testing data are not much different by 0.91 and 0.88. The results of the study using the variable selection method and the clustering process related to the factors that affect stunting are in accordance with the theory and produce a high accuracy value. Gestational age and size of the baby at birth greatly affect stunting.

The use of the Random Undersampling method can simplify the process of balancing unbalanced data so as to make research results more accurate and easier to use in future big data (data mining) and the stepwise process as a regulatory method for factor selection, but the use of the Random Undersampling process is used. on a binary predictor. So that this research is limited by the use of binary data and a stepwise selection method to reduce research variables. In future research, other methods can be used in dealing with high-dimensional problems and class differences, to maximize results in the classification process using the CART method.

6. LIMITATION

This research focuses on the problem of stunting in Indonesia. The environment and existing problems may differ in different countries.

7. ACKNOWLEDGMENTS

This work is supported by Penelitian tesis magister – Direktoral Pendidikan Tinggi (DIKTI), Republik Indonesia.

REFERENCES:

[1] Unicef Indonesia. (2018). Nutrisi Mengatasi beban ganda malnutrisi di Indonesia. Diakses pada 3 Oktober 2021 dari https://www.unicef.org/indonesia.

[2] Semba, R. D. dan Bloem, M. W. (2001). *Nutrition and Health in Developing Countries.* Tontowa, New Jersey: Humana Press

[3] Amri, K. (2013). *Data Mining untuk Menentukan Kriteria Calon Nasabah Potensial pada AJB Bumiputera 1912 Palembang.* Tesis. Teknik Informatika fakultas Ilmu Komputer Universitas Bina Darma. Palembang.

[4] Zega, A. S. (2014). Penggunaan Pohon Keputusan untuk Klasifikasi Tingkat Kualitas Mahasiswa Berdasarkan Jalur Masuk Kuliah. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI). ISSN:1907-5022.*

[5] Yin Loh, W dan Zheng, W. (2013). Regression Trees For Longitudinal and Multiresponse Data. *The Annals of Applied Statistics*. Doi : 10.1214/12-AOAS596.

[6] Srikandi, D. (2015). *Klasifikasi Anak Putus Sekolah Dengan Melibatkan Peubah Jaringan Sosial Menggunakan Cart Di Sulawesi*.Tesis. Pascasarja Jurusan Statistika. IPB. Bogor.

[7] Han, J. dan Kamber, M., 2006. *Data Mining Concepts and Techniques.* 2nd ed. US: Elsevier.

[8] Rochayani, Y. M., Sa'adah, U., dan Astuti, B. A. (2020). *Finding Biomarkers from a HighDimensional Imbalanced Dataset Using the Hybrid Method of Random Undersampling and Lasso.* Doi:10.21512/Comtech.v11i2.6452.

[9] Hosmer, S. W. dan Lemeshow, S. (2000). *Applied Logistic Regression Second edition. John Wiley & Son.*

[10] Rahmi, S. N. (2018). *Ensemble Support Vector Machine Dengan Random Undersampling Pada Klasifikasi Data DNA Microarray Untuk Menangani Multi Class Imbalance.* Tesis. Statistika. Fakultas Matematika, Komputasi dan Sains Data. Institut Teknologi Sepuluh Nopember. Surabaya.