**M – 12**

# CART BAGGING FOR CLASSIFICATION OF CHILD LABOR'S CASE IN CENTRAL SULAWESI PROVINCE

**Mohammad Fajri, Muhammad Mashuri**
*Institut Teknologi Sepuluh Nopember*

**Abstract**

Classification is statistical method that used to classify the systematical data. In statistics, there are several classification methods, one of the methods is CART (Classification and Regression Trees) that resulted classification trees model. CART has several advantages related with the model and the classification result, although it has a weakness on stability model that resulted. To solve this weakness, bagging (bootstrap aggregating) technique was applied on CART method to increase the stability and classification accuracy, that called CART bagging method. In this research, CART bagging is using to classify child labor's case in Central Sulawesi province. Beside that, this research also aims to determine the influence factors of child labor's case and the most dominant factor. The result shows that child get involved in work activity influenced by child's school participation, child's age, child's sex, number of household member, income per capita and head of household's education level, with child's school participation become the most dominant factor of child labor's case in Central Sulawesi Province. Classification accuracy used to see how good the classification model. Bagging technique on CART resulted higher classification accuracy.
**Keywords**: classification, CART bagging, child labor, classification accuracy.

## 1. INTRODUCTION

Classification is one of statistical method that used to classify a systematic data. In statistic, there are several familiar classification method like discriminant analysis and logistic regression which in classification literature called the classic model. Both of this method has some assumptions related with predictors scale, relationship between predictors and another classic assumptions that are a weakness because many data can't fill up this assumptions.

Science development also influence the classification method. The classification methods developing to solve classic method's weakness. Classification and Regression Trees (CART) is one of developing method. CART is an unique method, because resulted a visual classification model. It also has an efficient computation and easier model interpretation. (Breiman, L., Friedman, J., Olshen, R. and Stone, C., 1984).

CART introduced by Breiman et al (1984) is an inovative method to large data analysis through binary partitioning procedur that used to describe relationship between response variable and predictor variable. CART resulted tree model. Tree model will be a regression tree if the response variable is a continue data and the tree model will be a classification tree if the response variable is categoric data.

CART has several adventages and also weakness. The adventages are not attached with any predictors assumptions, resulted visual model, can used to many variable, simple classification result, efficient and easy interpretation. While the weakness is susceptible resulted unstable model, the resulted model change according the number of learning and testing data, and the election process is not suitable if applied in complex data structure (Lewis, R.J., 2000).

To solve this weakness, Breiman (1996) developing a method called CART Bagging (Bootstrap Aggregating) to fix the stability and the accuracy of prediction power by variance

reduction from the predictor. The simple idea from bagging is using bootstrap resampling to generate a multi version predictor, when if it combine, has a better result than a single predictor that generate to finish a same problem.

CART Bagging will applied to child labor's case in Sulawesi Tengah. According to National Commission on Child Protection in 2013, explain that the highest child labor provinces located in eastern Indonesia, include Papua and Sulawesi. In Central Sulawesi, number of child labor get 10% from the total labor. This thing surely related with many factors. Central Sulawesi as a developing region has a pretty large economic growth in 2013, 10,33% with the central resource come from farm, mining, service, construction, trade, hotel and restaurant. Ironically from Labour Ministry's data, several of this central resource like farm, mining, service and manufactur are the place where child labor appears. Even the child also found work as rock cracker, which very danger to this child, mentally and physically.

## 2. RESEARCH METHOD

### 2.1. Classification and Regression Trees (CART)

CART introduced by Breiman et al (1984) is an innovative method to large data analysis through binary recursive partitioning procedure that used to describe relationship between responsese variable and predictor variable. CART resulted tree model. Tree model will be a regression tree if the responsese variable is a continue data and the tree model will be a classification tree if the responsese variable is categorical data (Breiman *et al*, 1984). CART algorithm has several steps, there are.

1. Classification Trees Formed

a. Splitting Rule

Splitting rule on learning data aims to get every homogen derivative set than the main set. This thing doing by detemine the impurity function in *t* node. Index gini criteria used to split with impurity function

$$i(t) = \sum_{i \neq j} p(j|t)p(i|t)$$

Goodness of Split is using to evaluate the splitting result that defined by

$$\phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Tree's development doing by looking for every possible split on $t_1$ node, and then get $s^*$ split that give a maximal impurity derivative

$$\Delta i(s^*, t_1) = \max_{s \subset S} \Delta i(s, t_1)$$

b. Terminal Node

A *t* node will be terminal node if:

- There are no a impurity derivative significantly.
- The minimum limit fulfill.
- There are depth limit for the tree.

c. Class Label

Class label determine based of the most number rule.

2. Pruning

Pruning aims to get the optimal tree with using cost complexity minimum (Breiman, *et al*. 1984).

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

Where

$R(T)$ = Re-substitution estimate

$\alpha$ = Complexity parameter

$|\tilde{T}|$ = Number of terminal node *T*

$R_\alpha(T)$ is a linear combination of re-substitution estimate and its complexity. Breiman *et al*, (1984) state that cost complexity pruning determine a $T(\alpha)$ sub tree that minimize $R_\alpha(T)$ in all sub trees, or for each $\alpha$, search sub tree $T(\alpha) < T_{max}$ that minimize $R_\alpha(T)$.

$$R_\alpha(T(\alpha)) = \min_{T<T_{max}} R_\alpha(T)$$

When *R(T)* is used as derivative optimal tree criteria, $T_1$ will be chosen, a larger tree means a smaller *R(T)*.

3. Optimal Tree

The optimal tree that chosen is the right size tree and has a small re substitution estimate cost. Large tree size considering has high complexity because the data structure describe complexly. Re-substitution estimate that used is.

$$R^{ts}(T_k) = \frac{1}{N_2} \sum_{x_n, j_n \in L_2} X(d(x_n) \neq j_n)$$

$T^*$ is the chosen optimal tree with criteria $R^{ts}(T^*) = \min_k R^{ts}(T_k)$.

**2.2. Bootstrap Aggregating (Bagging)**

Bagging is a technique that suggested by Breiman (1996) to fix the stability and accuracy prediction power of regression and classification tree by variance reduction from the predictor. Basic idea from bagging is using bootstrap resampling to generate a multi version predictors. Bagging is intensive procedure to fix the unstable classification.

CART Bagging is CART method which bagging technique added in. In classification, B sample of bootstrap take from learning, then in every bootstrap sample CART applied to result class prediction. Classification accuracy is depending on number of bootstrap, so number of bootstrap replication is very important. Sutton (2004) suggested 25 or 50 times replication, while Hestie et al (2001) suggested 50, 100 and 200 times replication.

**2.3. Data**

Used data in this research is secondary data from SUSENAS in Central Sulawesi Province. The data will be divide into two part, learning data and testing data. Learning data used to model verification and testing data used to model validation. The data divide subjectively. (Briman *et al*, 1984).

Variable in this research are response variable and predictors variable. Response variable is child work status with nominal scale:

$$Y = \begin{cases} 0, \text{if child not work} \\ 1, \text{if child work} \end{cases}$$

While the predictors are

$X_1$ = Child's age, interval scale.

$X_2$ = Child's sex, nominal scale.
1 = Boy, 2 = Girl.

$X_3$ = Child's School Participation, nominal scale.
0 = No School, 1 = School

$X_4$ = Head of Household's age, interval scale.

$X_5$ = Head of Household's education, ordinal scale.
0 = No School
1 = Elementary School
2 = Junior High School
3 = Senior High School
4 = College

$X_6$ = Head of Household's Sex, nominal scale.
1 = Male, 2 = Female

$X_7$ = Head of Household's Occopation Sector, nominal scale.
1 = Farm
2 = Industry
3 = Trade
4 = Service
5 = Other

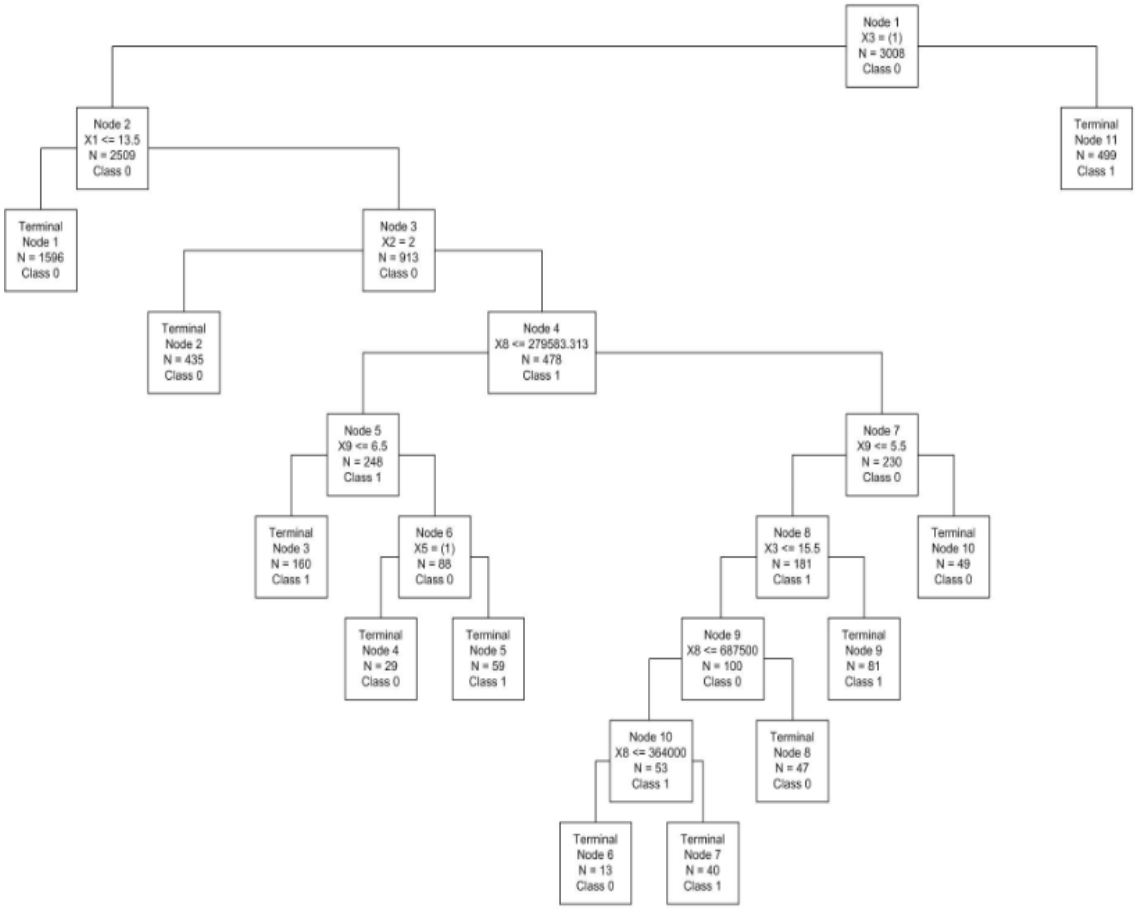
$X_8$ = Percapita Income (Rupiah), interval scale.
$X_9$ = Number of Household's Member, interval scale.

## 3. RESULT AND DISCUSSION

### 3.1. Child Labor Classification With CART

Maximum tree is a classification tree with the most terminal nodes. Maximum tree has 239 terminal nodes with the first splitting in child's age variable that has main role to classification tree formation and the most dominant variable in child labor classification. All of the predictors include to maximum tree model. Although the maximum tree has the high classification accuracy, but this tree has a very complex structure with 239 terminal nodes, for that classification tree's pruning necessary to resulted the optimal tree.

The optimal tree resulted from pruning process not built by all predictors. From 9 variable, just 6 variable contain in optimal tree, there are child's school participation $(X_3)$, child's age $(X_1)$, child's sex $(X_2)$, number of household's member $(X_9)$, per capita income $(X_8)$, and head of household's education $(X_5)$.




**Picture 3.1.** Optimal Tree

Terminal node is the last point from a structural split in classification tree. This nodes can't split become another nodes. This means terminal nodes contain of homogeny observations and finally categorized as a certain class. Based on optimal tree, there are 11 terminal nodes.

- Terminal node 1 labeled as class 0, this means observations in this node predicted as not working child. The sequential structure shows that the member of this node is child under 13,5 years old and schooling. This node has 1596 observations.
- Terminal node 2 has 435 observations which predicted as not working child. The sequential structure showing schooling girl 13,5-15,5 years old.

- Terminal node 3 has 160 observations that predicted as working child. Based on the sequential structure, this node shows schooling boy 13,5-15,5 years old with per capita income not more than Rp. 279.583,313 and has 5-6 household members.
- Terminal node 4 contain schooling boy 13,5-15,5 years old with per capita income not more than Rp. 279.583,313. It has 5-6 household members and the head of household's education is equal with elementary school. This node has 29 observations and predicted as not working child.
- Terminal node 5 contain schooling boy 13,5-15,5 years old with per capita income not more than Rp. 279.583,313. It has 5-6 household members and the head of household's education is beside elementary school. This node has 59 observations and predicted as working child.
- Terminal node 6 predicted as not working child and has 13 observations. This node shows schooling boy 13,5-15,5 years old with per capita income Rp. 279.583,313-Rp 364.000 and has not more than 5 household members.
- Terminal node 7 predicted as working child and has 40 observations. According to the sequential structure, this node shows schooling boy 13,5-15,5 years old with per capita income Rp. 364.000-Rp. 687.500 and has not more than 5 household members.
- Terminal node 8 contain schooling boy 13,5-15,5 years old with per capita income more than Rp. 687.500 and has not more than 5 household members. This node predicted as not working child and has 47 observations.
- Terminal node 9 predicted as working child and has 81 observations. This node showing schooling boy above 15,5 years old with per capita income Rp.279.583,313-Rp. 687.500 and has more than 5 household members.
- Terminal node 10 has 49 observations and predicted as not working child. This node shows schooling boy 13,5-15,5 years old with per capita income Rp. 279.583,313-Rp. 687.500 and has more than 5 household members.
- Terminal node 11 predicted as working child and has 499 observations. This node showing not school child.

Testing data used to determine classification accuracy.

**Table 3.1.** Classification Accuracy

| Observation | Prediction | | Classification Error | Classification Accuracy |
|---|---|---|---|---|
| | 0 | 1 | | |
| 0 | 501 | 146 | 12,57% | 77,43% |
| 1 | 30 | 95 | 24% | 76% |
| Total | | | 22,8% | 77,2% |

### 3.2. Bagging Applied

Bagging misclassification rate used to determine classification accuracy of CART bagging. This research use 200 times replications because it has the highest classification accuracy than the other replications. The difference between before bagging applied and after bagging applied occur in following table.

**Tabel 3.2.** The difference between before bagging applied and after bagging applied

| Analysis | *Sensitivity* (%) | *Specitivity* (%) | *Accuracy* (%) |
|---|---|---|---|
| Without Bagging | 82,5 | 58,6 | 78,4 |
| With Bagging (200 times replications) | 95 | 50 | 87,4 |
| Difference (%) | 12,5 | -8,6 | 9 |

This table shows that bagging applied increase classification accuracy in CART from 78,4% to 87,4%. This means in child labor's case in Central Sulawesi Province, bagging applied on CART can increase classification accuracy 9%.

## 4. CONCLUSION AND SUGGESTION

1. Optimal tree established by child's school participation, child's age, child's sex, number of household's member, per capita income, and head of household's education. Beside that, the optimal tree also produce 6 groups which predicted as not working child and 5 groups which predicted as working child. The 5 groups are:

- Group 1 has 160 observations that predicted as working child. Based on the sequential structure, this node shows schooling boy 13,5-15,5 years old with per capita income not more than Rp. 279.583,313 and has 5-6 household members.
- Group 2 contain schooling boy 13,5-15,5 years old with per capita income not more than Rp. 279.583,313. It has 5-6 household members and the head of household's education is beside elementary school. This node has 59 observations and predicted as working child.
- Group 3 predicted as working child and has 40 observations. According to the sequential structure, this node shows schooling boy 13,5-15,5 years old with per capita income Rp. 364.000-Rp. 687.500 and has not more than 5 household members.
- Group 4 predicted as working child and has 81 observations. This node showing schooling boy above 15,5 years old with per capita income Rp.279.583,313-Rp. 687.500 and has more than 5 household members.
- Group 5 predicted as working child and has 499 observations. This node showing not school child.

2. In child labor's case in Central Sulawesi Province, bagging applied on CART can increase classification accuracy from 78,4% to 87,4% or increase 9%.
3. This research can become a reference to Central Sulawesi government to solve child labor's case in that province.

## REFERENCES

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. *Classification and Regression Trees*. New York – London: Chapman & Hall.

Breiman, L. 1996. *Bagging Predictors*. Berkeley: Statistics Department University of California.

Irwanto. 1995. *Child Labor in Three Metropolitan City, Jakarta, Surabaya and Medan*. UNICEF and Atma Jaya Research Centre Series.

Melkas., A. 1996. *Economics Incentives for Children and Families to Eliminate or Reduce Child Labour*. International Labour Office.

Sumarmi. 2009. *CART Bagging To Classify Characteristic of Drop Out Student in Jambi*. Tesis: Sepuluh Nopember Institute of Technology.