Application of SMOTE on CART Method to Handle Imbalanced Data
(Study Case: Labor Force Classification in Banten Province)

A Anindya¹, Indahwati², B Susetyo³
¹Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia
²Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia
³Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia

E-mail: andinanindya@gmail.com

Abstract. Publication of the Central Bureau of Statistics (BPS) show that Banten is province with the highest unemployment rate in Java Island during the period 2006 to 2016. One of the efforts in resolving this issue is to do a classification of labor force into unemployment and employment as well as identify its characteristics, so that later the government policy will not be mistaken. The classification method used in this research is CART that has the ease of interpretation against the results of the analysis. The accuracy of a classification tree can be viewed from the value of specificity, sensitivity, and the AUC. Unfortunately, the imbalanced data causes low value of sensitivity and AUC. One alternative way to increase the accuracy is to conduct SMOTE on pre processing data. Classification tree with SMOTE is more accurate compared to the classification tree without SMOTE, because it can increase sensitivity from 6.35% to 77.78% and has higher AUC, as much as 0.7846. While AUC of classification tree without SMOTE is 0.7507. The best classification tree produce 18 terminal knots that influenced by 6 variables, which are age, gender, marital status, status in the household, level of education completed, and the residence area.

Keywords: CART, imbalanced data, SMOTE, unemployed

Introduction
Indonesia is one of developed countries which projected to receive demographic dividend by the year 2020 to 2030. Demographic dividend is the chance of one country’s economic prosperity caused by the large proportion of productive age (15-64 years old) in population evolution with cyclical pattern once in a century [1]. If Indonesia could not optimize the opportunity, there will be demographic disaster and the amount of unemployment will be increased. It may happen if Indonesia did not have good quality human resources and adequate job opportunities.

Unemployment is one of serious problems that must be solved by Indonesia government. Province with the highest unemployment rate in Java Island during 2006 to 2016 is province of Banten, which actually had high potential in economic. Unemployment rate is the number of unemployed people as a percentage of the labor force. One of the efforts to solve the problem is by identify the factors and the characteristics of unemployed people, so that policies which made by government would not be misdirected. Previous study done by [2] indicate that age, status in...
household, marital status, education, and residence area are significantly affected the probability of someone being unemployed.

Classification and Regression Tree (CART) is one of statistical method which can be used to classify labor force into unemployed and employed people. CART method can be applied on big dimension data and has an ability to find the most important variables and interaction between variables in purpose to determine the response variable. According to [3], CART can provide low error rates estimation and the result is easy to interpret. The main purpose of CART method is to obtain a group of homogenous data as characteristics of a classifier [4]. The result of this classification are expected to be used to determine the characteristics of each labor force categories.

The formation of classification tree is susceptible to unbalanced number of observation in each class of response variables. There is a class which more dominant (majority) and a class which has much less number of observation (minority). This condition may cause misclassified in minority class, which in this research is the unemployment class. Therefore, one of the alternatives to increase the accuracy of that category is by performing Synthetic Minority Oversampling Technique (SMOTE) in preprocessing stage of analyzing data.

The objective of this research is to compare the classification tree of employed and unemployed people from CART method, before and after SMOTE was applied. The best classification tree will show the factors which affected unemployment and characteristics of each categories of labor force.

Literature Review

1. **Classification and Regression Tree (CART)**
   CART is one of the methods with nonparametric approach which developed for classification analysis topic [4]. CART method yields a regression tree when the response variable is continuous, whereas categorical response variable yields a classification tree. This method has an ability to sort out the most important variables and interactions of variables in determining response variable. The main purpose of CART is to obtain an accurate data set which able to represent the characteristic of a classifier.

   CART is constructing the tree by recursively binary partitioning the data set. It started from splitting the main node into two child nodes based on the most significant explanatory variables to represent the response variable. Each child node will split into another two child node, and this partitioning will be repeated until it reaches the final nodes (terminal nodes). Those terminal nodes contain observation groups with relatively homogenous features based on its explanatory and response variables. The classification tree forming algorithm consists of 4 stages, namely: splitting, terminal node determination, labeling class label, and pruning.

2. **Synthetic Minority Oversampling Technique (SMOTE)**
   Imbalanced data occurs when the number of objects in a class is much larger than the other class. Class with more objects is referred to majority class, whereas class with fewer objects is referred to minority class. According to [5], processing a data set without fix the imbalanced data problem will tend to predict an object as majority class and ignore the
minority class. In some cases, misclassification in minority class will be more disadvantageous than the misclassification in majority class.

There are many various ways to handle imbalanced data, one of them is by using SMOTE in pre data processing stage. SMOTE is able to handle imbalanced data by generating synthetic data for minority class based on its k-nearest neighbor until the proportion of minority and majority class become more balanced. All variables used in this research are categorical, so the calculation of distance between minority class samples is done by following Value Difference Metric (VDM) formula [6]:

$$\Delta(A, B) = \sum_{i=1}^{N} \delta(V_{1i}, V_{2i})$$

where $\Delta(A, B)$ is the distance between observation A and B, N is the number of explanatory variable, and $\delta(V_{1i}, V_{2i})$ is the distance between observation A and B for each explanatory variable which calculate by following equation:

$$\delta(V_{1i}, V_{2i}) = \sum_{i=1}^{n} \left[ \frac{C_{1i}}{C_{1}} \cdot \frac{C_{2i}}{C_{2}} \right]^a$$

$\delta(V_{1i}, V_{2i})$ is the distance between observation A and B that belong to $i^{th}$ explanatory variable, n is the number of category in $i^{th}$ explanatory variable, $C_{1i}$ is the number of 1$^{st}$ category that belong to $i^{th}$ explanatory variable category, $C_{2i}$ is the number of 2$^{nd}$ category that belong to $i^{th}$ explanatory variable category, $C_{1}$ is the number of 1$^{st}$ category that happened, $C_{2}$ is the number of 2$^{nd}$ category that happened, and a is constant.

Data Source

This research used data from labor force sample that obtained from National Labor Force Survey (NLFS) which conducted by Central Bureau of Statistics (BPS) in August 2014. According [7], labor force is persons of 15 years old and over who in previous week were working, temporarily absent from work but having jobs, or those who did not have work and were looking for work. Those who are students, or housewives, or retired, or physically disabled and were not working in previous week are not in the labor force.

The data which processed by researcher generates response variable in the form of labor force classification that divided into 2 categories, which are employment and unemployment that lived in 8 districts/cities in Banten. Employment is labor force who worked for pay or assisted others in obtaining pay or profit for the duration at least one hour during the survey week, include an unpaid worker who help an economically activity/business. Unemployment is consist of person who did not have a job but looking for one; person without a job who have established a new business/firm; person without a job who were not looking for one, because they do not expect to find it; and person who have made arrangements to start work on a date subsequent to the reference period [7].

There are 7 explanatory variables used in this research, which are age, gender, marital status, status in household, level of educatio completed, working experience, and residence area. All of the variables are categorical and explained on Table 1.
Table 1 Research Variables

| Variables | Variable Names | Description |
|-----------|----------------|-------------|
| Y         | The classification of labor force | 0 = Employment 1 = Unemployment |
| X₁        | Age | 0 = 15-24 years old 1 = 25-50 years old 2 = >50 years old |
| X₂        | Gender | 0 = Female 1 = Male |
| X₃        | Marital status | 0 = Single 1 = Married 2 = Divorced/widowed |
| X₄        | Status in household | 0 = Not head of household 1 = Head of household |
| X₅        | Level of education completed | 0 = Not attending school/elementary school/Junior High School (≤SLTP) 1 = Senior High School (SLTA) 2 = University/academy |
| X₆        | Working experience | 0 = Do not have 1 = Have |
| X₇        | Residence area | 0 = Rural area 1 = Urban area |

Methods

The procedure of data analysis in this research divided into 9 parts that can be explained as follows:

1. Classify labor force based on the definition of employment and unemployment from BPS after processed the data cleaning. This stage is done in purposed to make the data set suitable with the data set needed in analysis process that performed by R 3.3.3.
2. Conduct data exploration and descriptive analysis. This stage aims to determine the characteristics of employment and unemployment in each district/city in Banten.
3. Divide data set into training and testing data with ratio 80%:20%.
4. Build classification tree using CART method gradually, which explained as follows:
   a) Splitting
      A classification tree is made by finding a split of each node that able to minimize the heterogeneity. The heterogeneity of a node can be measured by its impurity value using Gini index. According to [4], the greater impurity value of a node, the more heterogeneous the node. Impurity value of node \( t \) can be defined as follows:

\[
i(t) = 1 - \sum_{j} p^2(j|t)
\]

(1)
where \( p(j|t) \) represent the probability of the observation in class \( j \) of node \( t \) which can be calculate by following equation:

\[
p(j|t) = \frac{\pi_j N_j(t)/N_j}{\sum_j \pi_j N_j(t)/N_j}
\]

where \( \pi_j \) is initial probability of \( j \)th class, \( N_j \) is the number of observation in \( j \)th class, and \( N_j(t) \) is the number of observation in \( j \)th class of node \( t \). If there is split \( s \) that will split \( t \) into left node \( (t_L) \) with \( P_L \) proportion and right node \( (t_R) \) with \( P_R \) proportion, then the goodness of \( s \) is defined as the decrease of impurity value which can described as follows:

\[
\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)
\]

A split that yields higher value of \( \Delta i(s,t) \) will be the best split, because it is able to decrease the impurity value more significantly. If the node is not homogeneous yet, then the same procedure will be repeated until the classification tree meets:

\[
\Delta i(s^*,t) = \max_{s \in S} \Delta i(s,t_1)
\]

After obtaining the main node, the child node will be defined with the same way.

b) Stopping criterion

Node \( t \) will be terminal node when the number of observation is less than 5 and cost-complexity reaches 0.001. If the stopping criterion is reached, exit. Otherwise, apply step a) to each node \( t \).

c) Predicted class

Predicted class of one terminal node is assign by majority vote. If \( P(j_0|t) = \max_j P(j|t) \), then the predicted class of terminal node \( t \) is \( j_0 \) [4].

d) Pruning

According to [4], a large classification tree will provide the smallest error, but it will make the complexity value higher because the data structure described is more complex. Thus, it is necessary to find the optimal tree which is simpler but has a fairly small error. It can be found by pruning the less important part of classification tree. The interest rate of the tree is measured by the size of cost-complexity with the following equation:

\[
R_\alpha(T_k) = R(T_k) + \alpha |T_k|
\]

where \( R_\alpha(T_k) \) is the relative value of a sub tree \( T_k \), \( R(T_k) \) is the misclassification rate of sub tree \( T_k \) when \( k = 1 \), \( |T_k| \) is the number of terminal node on \( T_k \), and \( \alpha \) is cost-complexity parameter. The result of pruning process is a series of classification trees \( T_k \) and the optimal tree can be defined by using cross validation as follows:
\[ R^{CV}(T_{kn}) = \min_k (R^{CV}(T_k)) \]

e) Confusion matrix

To evaluate the result of CART, this research is using confusion matrix which described in Table 2 and applied on both training and testing data set. Confusion matrix is consists of accuracy, specificity, and sensitivity. Accuracy describes the ability of tree to classify both classes correctly. Specificity describes the ability of tree to classify employment as employment, while sensitivity describes the ability of tree to classify unemployment as unemployment.

| Actual Condition | Predicted Condition | Accuracy Rate (%) |
|------------------|---------------------|-------------------|
| Negative         | True negative (a)   | Specificity       |
| Positive         | False negative (c)  | Sensitivity       |
|                  | False positive (b)  |                   |
|                  | True positive (d)   |                   |

Those can be obtained using the following equation:

\[ \text{Accuracy} = \frac{a+d}{a+b+c+d} \]

\[ \text{Specificity} = \frac{a}{a+b} \]

\[ \text{Sensitivity} = \frac{d}{c+d} \]

5. Handle the imbalance on training data set by using SMOTE through the following steps:

a) Determine the number of k-nearest neighbor. This research uses 5-nearest neighbor. It means the generated data is derived from 5 adjacent data in minority class.

b) Set the oversampling number as 800%, which means the minority class data will be raised by 8 times.

c) Calculate the distance between each observation using VDM formula.

d) Pick one of minority class samples randomly and determine the 5-nearest neighbor by sorting the distance between each selected sample and every minority class samples.

e) Choose the majority category in main vector and 5-nearest neighbor for every variable. If there is more than one majority category, then choose randomly. It would be the synthetic data made by SMOTE.

f) Apply step d) and e) until the number of oversampling is reached.
g) The new data set consists of original minority class data, synthetic data made by SMOTE, and the original majority class data.

6. Apply step 4 on the new data set which made in step 5.
7. Form the Receiver Operating Characteristic (ROC) curve by using classification tree which made in step 4 and 6. ROC is a curve that able to describe the classification performance in two dimensions [8]. The curve is created by plotting the false positive rate (1-specificity) against true positive rate (sensitivity). ROC can be converted into Area Under Curve (AUC) when used to compare some performances.
8. Determine the best tree by comparing the value of AUC from ROC curve in step 7 and the confusion matrix on testing data set while considering the accuracy of training data set.
9. Interpret the best classification tree in step 8.

Results and Discussion

Descriptive

The objects of this research are sample of labor force in Banten from NFLS as many as 7290 individuals which spread over 4 districts and 4 cities, namely Pandeglang District, Lebak District, Tangerang District, Serang District, Tangerang City, Cilegon City, Serang City, and South Tangerang City. Figure 1 illustrates that out of 7290 individuals in Banten, 91.07% of them are employment, while the percentage of those who belonging to unemployment category was only as much as 8.93%. This condition is called the imbalanced data.

![Percentage of each labor force category in Banten](image)

Table 3 shows that the district/city with the largest number of observation is Tangerang District with 1234 samples, while the area with the least number of observation is Serang City with 644 samples. Other than that, it can be seen that Serang District has the highest of unemployment rate in Banten Province on August 2014 as much as 14.27%. The lowest unemployment rate is in South Tangerang City as much as 6.71%.
Table 3 Number of samples and unemployment rate in each district/city in Banten

| District/City         | Number of Samples | Unemployment Rate (%) |
|-----------------------|-------------------|-----------------------|
| Pandeglang District   | 833               | 6.72                  |
| Lebak District        | 968               | 8.68                  |
| Tangerang District    | 1234              | 8.02                  |
| Serang District       | 953               | 14.27                 |
| Tangerang City        | 1135              | 7.40                  |
| Cilegon City          | 688               | 11.19                 |
| Serang City           | 644               | 9.16                  |
| South Tangerang City  | 835               | 6.71                  |

The characteristic of labor forces in Banten is described in Table 4 by looking at the number of observation column for each explanatory variable. According to age variable, labor force samples is dominated by individuals aged 25-50 years old, while the gender variable is dominated by male. Based on the residence area, the labor force samples in Banten mostly lives in urban area rather than rural area.

Table 4 Unemployment based on each category in explanatory variables

| Explanatory Variables | Category           | Number of Unemployment | Number of Observation | Unemployment Rate (%) |
|-----------------------|--------------------|-------------------------|-----------------------|-----------------------|
| Age                   | 15-24 years old    | 415                     | 1359                  | 30.54                 |
|                       | 25-50 years old    | 210                     | 4624                  | 4.54                  |
|                       | >50 years old      | 26                      | 1307                  | 1.99                  |
| Gender                | Female             | 213                     | 2516                  | 8.47                  |
|                       | Male               | 438                     | 4774                  | 9.17                  |
| Marital Status        | Single             | 466                     | 1659                  | 28.09                 |
|                       | Married            | 156                     | 5144                  | 3.03                  |
|                       | Divorced/widowed   | 29                      | 487                   | 5.95                  |
| Status in Household   | Not head of household | 568                | 3851                  | 14.75                 |
|                       | Head of household  | 83                      | 3439                  | 2.41                  |
| Level of Education    | ≤SLTP              | 320                     | 4217                  | 7.59                  |
|                       | SLTA               | 295                     | 2234                  | 13.21                 |
|                       | University/academy | 36                      | 839                   | 4.29                  |
| Working Experience    | Do not have        | 340                     | 3515                  | 9.67                  |
|                       | Have               | 311                     | 3775                  | 8.24                  |
| Residence Area        | Rural              | 234                     | 2413                  | 9.70                  |
|                       | Urban              | 417                     | 4877                  | 8.55                  |
In addition, Table 4 also shows that there is a significant difference of unemployment rate between each category of age variable, with 30.54% in 15-24 years old category, 4.54% in 25-50 years old category, and 1.99% in more than 50 years old category. As someone getting older, their responsibility will be bigger. They will increase their productivity and decide to be employment in order to achieve a better life. Therefore, the unemployment rate of married samples is only 3.03%, because someone who is married must be having a greater responsibility. Furthermore, according to gender variable, the unemployment rate between female and male are not much different, with 8.47% among female samples and 9.17% among male samples.

The unemployment rate is able to show the distribution of unemployment in each category of explanatory variables. The unemployment rate in gender, work experience, and residence area variables are not significantly different between each category, or can be called homogeneous. The distribution of unemployment that looks more diverse can be found in age, marital status, status in household, and level of education completed variables.

**CART without SMOTE**

In CART method, training data set is used to create classification tree, while testing data set is used for model validation. The maximum tree formed from training data set has 6 terminal nodes and involves 5 explanatory variables, which are age (X1), gender (X2), marital status (X3), working experience (X6), and residence area (X7). The next stage is pruning maximum tree into optimal tree. The result of the pruning process consists of several trees with cost-complexity parameter, error proportion, and \( R^{CV} \). \( R^{CV} \) is the relative error which obtained from 10-fold cross validation. However, due to some predetermined stopping criterion in classification tree making, the number of classification trees generated in training data processing is only one. Therefore, in this case, the pruning stage will not be applied.

![Figure 2 Plot between R^{CV} and the number of terminal node on CART without SMOTE](image)

In CART method, the goodness of a classification will be measure by the value of accuracy, specificity, and sensitivity. Figure 2 shows that the more the terminal node is, the higher the \( R^{CV} \) value will be. This happens because of the imbalanced data problem, where the number of observations in one class is much larger than the other class. The increased number of terminal nodes tends to increase the sensitivity value, but decrease the overall accuracy value. The classification tree formed can be seen in Figure 3, where the red number under terminal node describes serial number of terminal nodes.
Labor forces which classified as employment are described in 5 terminal nodes, while labor forces which classified as unemployment are described in 1 terminal node only. Confusion matrix of classification tree formed using the original training and testing data set are given in Table 5 and Table 6. The ability of CART method to classified labor force correctly is indicated by total accuracy, which are 91.14% correct for training data set and 91.63% correct for testing data set. However, misclassification rate of unemployment class in both training and testing data set is way too high. Therefore, it can be concluded that classification tree formed by using original data set isn’t able to classify unemployment well.

Table 5 Confusion matrix of CART without SMOTE on training data set

| Observation     | Predicted Training Data Set | Accuracy Rate (%) |
|-----------------|-----------------------------|-------------------|
|                 | Employment                  | Unemployment      |
| Employment      | 5279                        | 28                | 99.47 |
| Unemployment    | 489                         | 36                | 6.86  |
| Total Accuracy  |                             |                   | 91.14 |

Table 6 Confusion matrix of CART without SMOTE on testing data set

| Observation     | Predicted Testing Data Set | Accuracy Rate (%) |
|-----------------|----------------------------|-------------------|
|                 | Employment                 | Unemployment      |
| Employment      | 1328                       | 4                 | 99.70 |
| Unemployment    | 118                        | 8                 | 6.35  |
| Total Accuracy  |                             |                   | 91.63 |
Imbalanced Data Handling

The result of CART method applied on original data indicates that the imbalanced data problem causes the observation in minority class tends to be ignored in labor force classification. Therefore, this problem will be solved by applying SMOTE on training data set. Synthetic data is generated as much as 8 times of minority class data set, which are $525 \times 8 = 4200$ new observations. The new data is added to the original data set, so total observation in the new data set becomes $4200 + 525 = 4725$. Table 7 represents proportion of observation in each class in training data set before and after SMOTE was applied. It shows that the number of observation in each class in the new data set became more balance, with 52.90% unemployment and 47.10% employment.

| Class       | Without SMOTE (%) | SMOTE (%)  |
|-------------|-------------------|------------|
| Employment  | 5307 (91.00)      | 5307 (52.90)|
| Unemployment| 525 (9.00)        | 4725 (47.10)|
| Total       | 5832 (100.00)     | 10032 (100.00)|

CART with SMOTE

Formation of the second classification tree was done by using the new training data set where SMOTE has been applied. The first stage in the establishment of a classification tree is splitting. The best split that is uses as main split is marital status variable ($X_3$) which divides population into two child nodes, the left node consists of married category and the right node consists of single and divorced/widowed category. $X_3$ has the highest decrease of impurity value compared to the other variables as much as 0.075. This shows that $X_3$ is the most dominant variable in the making of classification tree with addition of SMOTE data. Calculation of impurity value can be seen in Appendix 1. The maximum classification tree formed has 21 terminal nodes.

After maximum classification tree was obtained, the next step is to pruning to get optimal tree. Pruning process was done by using minimum $R^2_C$ criterion. According to Figure 4, it can be seen that $R^2_C$ value tends to decrease as the number of terminal node increases. The minimum $R^2_C$ value obtained is when the number of terminal node is 18 as much as 0.442, then it will be pruned at that point.

Figure 4 Plot between $R^2_C$ and the number of terminal node on CART with SMOTE
Labor force which classified as employment was formed on 6 terminal nodes, while labor force which classified as unemployment was formed on 12 terminal nodes. According to Table 8 and Table 9, the accuracy obtained between training and testing data set is not much different, either on total accuracy, sensitivity, or specificity. In addition, sensitivity and specificity tend to be more balanced, so it can be said that the optimal tree is able to classify labor force quiet well.

### Table 8 Confusion matrix of CART with SMOTE on training data set

| Observation | Predicted Training Data Set | Accuracy Rate (%) |
|-------------|----------------------------|------------------|
| Employment  | 4029                       | 75.92            |
| Unemployment| 1278                       |                  |
| Total Accuracy | 780                  | 83.49            |
| Unemployment | 3945                       |                  |
| Total Accuracy | 79.49                  |                  |

### Table 9 Confusion matrix of CART with SMOTE on testing data set

| Observation | Predicted Testing Data Set | Accuracy Rate (%) |
|-------------|----------------------------|------------------|
| Employment  | 1011                       | 75.90            |
| Unemployment| 321                        |                  |
| Total Accuracy | 28                   | 77.78            |
| Unemployment | 98                         |                  |
| Total Accuracy | 76.06                  |                  |

### Comparison of Classification Trees

Comparison of two classification trees which obtained before was done by comparing values of total accuracy, sensitivity, specificity, and AUC. This comparison will be performed on testing data set. According to Table 10, total accuracy of classification tree without SMOTE is bigger than classification tree with SMOTE, but classification tree without SMOTE was not able to classify unemployment properly because its sensitivity is very small, which only 6.35%. Mistakenly classify unemployment as employment will be certainly fatal, because policies that were made by government in order to decrease unemployment rate can be misdirected. After SMOTE was conducted on pre-processing data, the sensitivity obtained increased drastically to 77.78%.

### Table 10 Comparison between classification tree obtained with and without SMOTE

| Criterion      | Without SMOTE | SMOTE   |
|----------------|---------------|---------|
| Total Accuracy | 91.63%        | 76.06%  |
| Sensitivity    | 6.35%         | 77.78%  |
| Specificity    | 99.70%        | 75.90%  |
| AUC            | 0.7507        | 0.7846  |

Classification tree without SMOTE yields higher specificity, but the reduction of specificity which happened after SMOTE was conducted is not too significant compared to how its sensitivity increased. In addition, misclassification which happened because of classifying employment as unemployment tends to be less risky than incorrectly classify unemployment as
employment. Although SMOTE has decreased specificity as much as 23.80%, AUC of classification tree with SMOTE is 0.0339 higher than AUC of classification tree without SMOTE. The value of AUC was obtained from area under ROC curve which shown in Figure 5. This indicates that classification tree with SMOTE is more accurate than the classification tree without SMOTE, so the best classification tree chosen is classification tree with the addition of synthetic data.

![ROC curve of (a) CART without SMOTE and (b) CART with SMOTE](image)

**Figure 5 ROC curve of (a) CART without SMOTE and (b) CART with SMOTE**

**Interpretation of Best Classification Tree**

The best classification tree, which is the classification tree with the addition of SMOTE data, can be seen in Figure 6. The classification tree is formed by 6 explanatory variables, which are age ($X_1$), gender ($X_2$), marital status ($X_3$), status in household ($X_4$), level of education completed ($X_5$), and residence area ($X_7$), and generates 18 terminal nodes. Characteristic of labor force which classified as unemployment is formed on 12 terminal nodes and described in Table 11, while labor force which classified as employment is formed on 6 terminal nodes and presented in Appendix 2. In Figure 6, there is also an explanation about the meaning of each number and nodes formed in the classification tree.
The explanation of numbers in terminal nodes:

- The number in first line describes class label. If the class label is 0 (employment), then the terminal node will be colored in blue, while 1 (unemployment) will be colored in green. The darker the color of terminal node, the greater the proportion of class label observation.

- Second line numbers from left to right describes number of employment observations and number of unemployment observations in the node.

- The number in third line describes number of observations in terminal node divided by total observations.

Figure 6 Classification tree with SMOTE
Table 11 Characteristics of unemployment in Banten Province

| Terminal Node | Unemployment Proportion | Characteristics                                                                 |
|---------------|-------------------------|--------------------------------------------------------------------------------|
| 2             | 0.885                   | Married, aged over 25 years old, head of household, and female                  |
| 5             | 0.682                   | Married, aged over 50 years old, not head of household, male, and live in urban area |
| 6             | 0.615                   | Married, aged over 25 years old, not head of household, male, and live in rural area |
| 7             | 0.853                   | Married and aged 15-24 years old                                               |
| 9             | 0.765                   | Single of divorced/widowed, live in urban area, high level of education (university/academy), and aged over 50 years old |
| 10            | 0.563                   | Single of divorced/widowed, live in urban area, high level of education (university/academy), and aged 15-24 years old |
| 11            | 0.611                   | Single of divorced/widowed, live in urban area, level of education completed is senior high school or below, and not head of household |
| 13            | 1                       | Divorced/widowed, live in urban area, level of education completed is senior high school or below, head of household, and aged 15-24 years old |
| 14            | 0.908                   | Single, live in urban area, level of education completed is senior high school or below, and head of household |
| 16            | 0.941                   | Divorced/widowed, live in rural area, aged over 50 years old, and level of education completed is senior high school or above. |
| 17            | 0.984                   | Single, live in rural area, and aged over 50 years old                          |
| 18            | 0.824                   | Single or divorced/widowed, live in rural area, and aged 15-50 years old        |

Proportion of unemployment in Table 11 from the number of unemployment in node divided by number of observations included in the terminal node. Terminal node 13 which has characteristics as follows: divorced/widowed, live in urban area, level of education completed is senior high school or below, head of household, and aged 15-24 years old, can be regarded as a node with main characteristic of unemployment in Banten Province, because its unemployment proportion is 1. According to Figure 6, terminal node 13, 1 is obtained from $41 \div (0+41)$, in other
words, all observations in terminal node 13 (41 observations) are correctly classified as unemployment. In addition, unemployment proportion in terminal node 17 is also high, as much as 0.984. The characteristic explained in terminal node 17 are single, live in rural area, and aged over 50 years old.

**Conclusion**

The imbalanced data causes low sensitivity in the classification, so the tree could not be used to classify unemployment correctly. Application of SMOTE method in pre-processing stage of data analysis can increase sensitivity as much as 71.43%, but decrease specificity value as much as 23.80%. However, classification tree with the addition of SMOTE data has bigger AUC value than the classification without SMOTE. This indicates that SMOTE is able to improve the performance of CART method which used in imbalanced data case. Classification tree with SMOTE is formed by 6 explanatory variables and generates 18 terminal nodes, which divided into 6 employment nodes and 12 unemployment nodes.

**References**

[1] Sugiharto and Setiawan D 2015 Utilization of demographic dividend trough improving human development index in North Sumatera *Journal of Social Sciences* 7 1-12

[2] Yuliati, Huseno T, and Febriani 2011 The effect of population characteristics on unemployment in West Sumatera *Journal of Management and Entreprenership* 2 15-43 (Padang: Universitas Tamansiswa Padang)

[3] Sartono B and Syafitri UD 2010 Ensemble tree method: selected solution to handle the weakness of single regression and classification tree *Statistics and Computation Forum* 15 1-7

[4] Breiman L, Friedman JH, Olshen RA, and Stone CJ 1993 *Classification and Regression Trees* (New York: Chapman and Hall)

[5] Chawla VN, Bowyer KW, Hall LO, and Kegelmeyer WP 2002 SMOTE: Synthetic Minority Oversampling Technique *Journal of Artificial Intelligence Research* 16 321-357

[6] Cost S and Salzberg S 1993 A weighted nearest neighbor algorithm for learning with symbolic features *Machine Learning* 10 57-58 (Boston: Kluwer Academic Publisher)

[7] [BPS] Central Bureau of Statistics 2016 Labor force situation in Indonesia August 2016 [internet] [downloaded 2017 Feb 16] available on: https://bps.go.id/website/pdf_publikasi/Keadaan-Angkatan-Kerja-di-Indonesia-Agustus-2016--.pdf

[8] Fawcett T 2006 An introduction to ROC analysis *Pattern Recognition Letter* 27 861-874
Appendix
Appendix 1
An example of calculating impurity impairment in order to determine the main split will use a classification tree which shown in Figure 6. The explanatory variable to be calculated is marital status (X3). The first stage is to calculate Gini index of main node by using Equation (1):

\[
i(t) = 1 - \left( \frac{5307}{10032} \right)^2 - \left( \frac{4725}{10032} \right)^2 = 0.498317
\]

Calculate Gini index of left node \((i(t_L))\) and right node \((i(t_R))\) from one of \(X_3\) split probability:

| Y Category     | Number of Observations |
|----------------|------------------------|
| Employment (0) | 5307                   |
| Unemployment (1)| 4725                   |
| **Total**      | **10032**              |

\[
i(t_L) = 1 - \left( \frac{3987}{5718} \right)^2 - \left( \frac{1731}{5718} \right)^2 = 0.422168
\]

\[
i(t_R) = 1 - \left( \frac{1320}{4314} \right)^2 - \left( \frac{2994}{4314} \right)^2 = 0.424713
\]

Calculate impurity impairment by using Equation (2):

\[
\Delta i(s,t) = 0.498317 - \left( \frac{5718}{10032} \right) 0.422168 - \left( \frac{4314}{10032} \right) 0.424713 = 0.075055
\]

Perform calculation on each split probability from all of explanatory variables. Explanatory variable with the highest \(\Delta i(s,t)\), will become the main split.
Appendix 2

| Terminal Node | Employment Proportion | Characteristics                                                                 |
|---------------|-----------------------|---------------------------------------------------------------------------------|
| 1             | 0.902                 | Married, aged over 25 years old, head of household, and male                     |
| 3             | 0.827                 | Married, aged over 25 years old, not head of household, and female               |
| 4             | 0.546                 | Married, aged over 25 years old, not head of household, male, and live in urban area |
| 8             | 0.674                 | Single or divorced/widowed, high level of education (university/academy), and aged 25-50 years old |
| 12            | 0.709                 | Divorced/widowed, live in urban area, level of education completed is senior high school or below, head of household, aged over 25 years old |
| 15            | 0.792                 | Divorced/widowed, live in rural area, aged over 50 years old, and low education level (junior high school or below) |