IJSA (2024), Vol. 8, No. 2, 166-177
DOI: https://doi.org/10.29244/ijsa.v8i2p166-177

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Handling Unbalanced Data with Smote for Unemployment Classification in Lima Puluh Kota Regency Using CART

Aldwi Riandhoko[1], Nonong Amalita[2‡], Dodi Vionanda[3], and Admi Salma[4]

[1,2,3,4]Department of Statistics, Universitas Negeri Padang, Indonesia
‡corresponding author: nongmat@fmipa.unp.ac.id

Abstract

Unemployment is a problem that occurs in the labor force, where high unemployment is caused by the low ability of the labor force. A region that is still experiencing unemployment problems in West Sumatera is Lima Puluh Kota Regency. Unemployment in Lima Puluh Kota Regency is caused by the low competence of human resources to fulfill employment market requirements. Based on the results of the Sakernas survey in August 2023, Lima Puluh Kota Regency has more employed labor force than unemployed labor force, so this results in unbalanced data. A method that can overcome unbalanced data is Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a technique with addition of synthetic data in minority class so that the proportion is balanced. Data imbalance conditions need to be handled so as to improve the performance of the classification model. Classification and Regression Trees (CART) is a classification technique with a decision tree method that can obtain the characteristics of a classification. The purpose of this research is to compare the CART model before and after applying SMOTE which can be measured by comparing the highest Area Under Curve (AUC) value. The AUC value in the CART method before SMOTE applied has a value of 62.1% while the AUC value in the CART method after SMOTE applied has a value of 70.2%. Therefore, it can be concluded that the CART classification analysis after SMOTE applied is able to provide better performance compared to the CART classification analysis before SMOTE applied.

Keywords: AUC, CART, Lima Puluh Kota Regency, SMOTE, Unemployment.

* Received: Nov 2024; Reviewed: Nov 2024; Published: Dec 2024

1. Introduction

National development aims to improve the welfare of society. Indonesia as a developing country has made many changes to support national development. It aims to create national stability, a good investment climate, and high economic growth (Pratiwi & Zain, 2014). Stable and sustainable economic growth is important for national development because it can increase employment, higher incomes for individuals and households. One of the determinants of an economy's success in the field of human resources is the availability of labor (Fajriati & Syafriandi, 2022).

Labor is the population in working age who are ready to do work, including those who have worked, those who are looking for work, those who are in school, and those who take care of the household (BPS, 2023). The workforce is divided into two parts, namely the labor force and non-labor force. The labor force is the population in working age who do work to obtain income or profit including those who already have a job but did not work a week ago. Meanwhile, the non-labor force is the population in working age who do not work and do not have a job, because they are doing activities such as attending school, taking care of the household, retired, and disabled.

Some of the problems associated with employment are unemployment, minimum wage, and lack of jobs (Pratiwi & Zain, 2014). High unemployment in a region can be caused by the low ability of the labor force. Unemployment is someone who is a person of the labor force who has entered working age and is not undergoing formal education or who does not yet have a job (BPS, 2023). Unemployment consists of people of working age who actively seeking work, those without a job who are in the process of starting a business, and individuals who are unemployed but not job hunting because they believe finding a job is impossible in another case, a person who has a job or business but has not yet started it is also considered unemployed.

Regions that is still experiencing unemployment problems in West Sumatera is Lima Puluh Kota Regency. Where the unemployment rate in Lima Puluh Kota Regency has increased in the last few years. Unemployment in Lima Puluh Kota Regency is caused by the low competence of the population's human resources to access the workforce due to the low competitiveness of human resources in Lima Puluh Kota Regency (Bappeda Kabupaten Lima Puluh Kota, 2024).

Based on the results of the Survei Angkatan Kerja Nasional (Sakernas) in August 2023, Lima Puluh Kota Regency has more employed labor force than unemployed labor force, so this results in unbalanced data. The condition where a class group has a much different amount of data compared to other classes is said to be unbalanced data (Wijayanti et al., 2021). When dealing with imbalanced data, most classification algorithms tend to yield significantly higher accuracy for the majority class compared to the minority class. Addressing data imbalance is essential to enhance the performance of the classification model.

Synthetic Minority Oversampling Technique (SMOTE) technique is one method of handling unbalanced data by generating syntetic data for minority class so that the proportion of major and minor data classes becomes more balanced (Chawla et al., 2002). The advantages of the SMOTE method in general are that it does not cause missing information, avoids overfitting, builds a larger decision region, and is able to improve the accuracy of minority class predictions. SMOTE generates additional samples that are more representative of the minority class, so that the classifier has greater coverage in learning the minority class (Wijayanti et al., 2021). Therefore, data imbalance conditions need to be handled so as to improve the performance of the classification model.

Decision tree is a prediction model using a tree structure or hierarchical structure.

The concept of decision tree is to convert data into decision trees and decision rules (Ratniasih, 2014). One of the decision tree methods to classify the labor force into unemployment and non-unemployment is the Classification and Regression Tree (CART) method. By using the CART method, it can get an accurate data group as a characteristic of a classification. The results of this classification can be used to determine the characteristics of each labor force category. Then we will compare the CART model before and after applying SMOTE which can be measured by comparing the highest Area Under Curve (AUC) value.

2. Research Methods

2.1 Data and Variables

The type of data used in this research is secondary data. The data used was obtained from the results of the Survei Angkatan Kerja Nasional (Sakernas) in Lima Puluh Kota Regency in August 2023 with 1312 respondents. The variables in this research are divided into two, which are dependent variables and independent variables. The dependent variable of this research consists of non-unemployment and unemployment. The independent variables used are factors that affect unemployment in Lima Puluh Kota Regency.

Table 1: Research Variables

Variables	Variables Name	Variables Categories
Y	Employment Status	0 : Non-unemployment 1 : Unemployment
X_1	Age	0 : Young age 1 : Productive age 2 : Non-productive age
X_2	Gender	0 : Female 1 : Male
X_3	Marital status	0 : Unmarried 1 : Married 2 : divorced alive/divorced dead
X_4	Status in the family	0 : Not head of household 1 : Head of household
X_5	Education level	0 : Junior high school graduate and below 1 : High school graduate 2 : Graduated from an academy or college
X_6	Work experience	0 : No work experience 1 : Have work experience
X_7	Region of residence	0 : Rural 1 : Urban

2.2 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE technique is a way of boosting the sample size of the minority class so as to produce the same amount of data as the data in the majority class, namely by

replicating data in the minority class randomly so that the proportion between classes can be balanced (Sofyan & Prasetyo, 2021). The problem of data imbalance if ignored will cause the dominance of major classes and ignore minor classes in data processing (Wijaya et al., 2018). The method used in the SMOTE algorithm is the k-nearest neighbors method. All variables used in this research are categorical variables. The distance between minority classes is calculated using the Value Difference Metric (VDM) (Cost & Salzberg, 1993).

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

Where N is the number of independent variables, $\delta(V_{1i}, V_{2i})$ is the distance between categories for each variable. To determine the distance between categories for each variable, the equation is used.

$$(V_{1i}, V_{2i}) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_i} - \frac{C_{2i}}{C_i} \right| \tag{2}$$

Where:
n : the number of categories in the variable i
C_{1i} : the number of categories 1 that belong to the class i
C_{2i} : the number of categories 2 that belong to the class i
C_i : the number of observations in class i

2.3 Classification and Regression Tree (CART)

CART method was developed in 1984 by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. CART produces a classification tree if the dependent variable is categorical, and produces a regression tree if the dependent variable is continuous (Breiman et al., 1984). CART is used to represent decisions in the form of binary trees. The purpose of the CART method is to obtain accurate data groups as a characteristic of a classifier.

The CART method is a classification technique using a binary recursive partitioning algorithm. The term “binary” means that the sorting occurs within a dataset organized in a space known as nodes, which is categorized into two groups called internal nodes, while the term “recursive” means that the binary partitioning procedure is done repeatedly. Each internal node obtained from splitting the root node can be split back into two more nodes, and so on until certain criteria are met. The term “partitioning” means that the classification process is carried out by sorting a data set into several parts or partitions.

The following is the algorithm for creating a classification tree for the CART method (Sumartini & Purnami, 2015). Classification tree formation is done by finding a splitter from each node that can produce the highest impurity value. The impurity function of the splitter is chosen based on the Gini index rule which is evaluated using the goodness of split criterion. The greater the impurity value of a node, the better the node becomes the root node (Breiman et al., 1984).

The probability of observations entering the right and left nodes is calculated with the following equation (Breiman et al., 1984).

$$P_L = \frac{left\ node\ candidate}{data\ training} \tag{3}$$

$$P_R = \frac{right\ node\ candidate}{data\ training} \quad (4)$$

The impurity value at node t is defined as follows.

$$i(t) = 1 - \sum_{j=1} p^2(j|t) \quad (5)$$

Where $p(j|t)$ is the proportion of class j at node t. The partioning (s) at node t is as a decrease in impurity (Breiman et al., 1984).

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (6)$$

Where:
$i(t)$: Impurity function at the node
$i(t_R)$: Impurity function at the right internal node
$i(t_L)$: Impurity function at the left internal node
P_R : Probability of observation at the right node
P_L : Probability of observation at the left node

A node t will become a terminal node or will not be split again, if the number of observations is less than the minimum split. A node t will become a terminal node or not, if the number of observations is less than the minimum split limit and if the depth of the tree is maximum. If these are reached, the tree growing process will be stopped.

The class labeling of a terminal node is based on the majority voting rule, i.e. if $P(j_0|t) = max_j\ P(j|t)$, then the class label for terminal t is j_0 (Breiman et al., 1984).

$$P(j_0|t) = max_j\ P(j|t) = max_j\ \frac{N_j(t)}{N(t)} \quad (7)$$

Where:
$P(j|t)$: Probability of class j at node t
$N_j(t)$: Number of observations of class j at node t
$N(t)$: Number of observations at node t

Tree pruning is used to prevent the formation of large and complex classification trees, so as to obtain the optimal tree size. To obtain the optimal tree size, tree pruning is carried out based on the minimum cost complexity measure where the pruned branch is the branch that has the smallest value using the following formula (Breiman et al., 1984).

$$g_m(t) = \frac{R(t) - R(T_k)}{|T_k| - 1} \quad (8)$$

Where:
$g_m(t)$: Compexity Parameter
$R(t)$: Misclassification at node t
$|T_k|$: Number of terminal nodes in the kth subtree
$R(T_k)$: Misclassification in the tree

The misclassification at the node is obtained by the following formula (Breiman et al., 1984).

$$R(t) = r(t).P(t) \tag{9}$$

Where:

$r(t)$: Probability of misclassification at classifier node t with $r(t) = 1 - max_j\, P(j|t)$

$P(t)$: Probability of the number of objects at node t

The best pruned subtree is made based on the cross validation estimator as follows.

$$R(T_t{}^v) = \frac{1}{N_v} \sum X\left(d^{(v)}\right) \tag{10}$$

Where $X\left(d^{(v)}\right)$ is the result of the classification and N_v is the number of observations in tree.

2.4 Confusion Matrix

Confusion matrix describes the performance of the model through a table (Saputro & Sari, 2019). Each row of the matrix represents the actual classification of the data and each column of the matrix represents the predicted classification of the data or otherwise.

Table 2: Confusion Matrix

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Based on the results of the confusion matrix, calculations can also be made to measure the performance of the model, namely the accuracy, sensitivity, and specificity values. Accuracy means how accurately the model can classify the data correctly. Sensitivity is a criterion in the confusion matrix used to measure the accuracy of positive classes that are classified as also positive.

Specificity is a criterion used to measure the accuracy of negative classes that are classified as negative classes as well. The classification value is obtained using the following equation.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \tag{11}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{12}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{13}$$

2.5 Receiver Operating Characteristics (ROC)

Receiver Operating Characteristics (ROC) curve is a method used to visualize and select the best classification model based on performance. On the ROC curve, the True Positive (TP) rate is plotted on the Y-axis while the False Positive (FP) rate is plotted on the X-axis. ROC has an area called Area Under Curve (AUC) which can be used to compare the performance of multiple classification models to find the best model. The AUC value ranges from 0 to 1, where if it is close to 1, it can be said that the model is able to classify well (Sari et al., 2020). The calculation of AUC is defined as follows:

$$\text{AUC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (14)$$

3. Result and Discussion

Based on the Sakernas results in Kabupaten Lima Puluh Kota in August 2023, Lima Puluh Kota regency has an unbalanced number of unemployed and non-unemployed which is presented as follows.

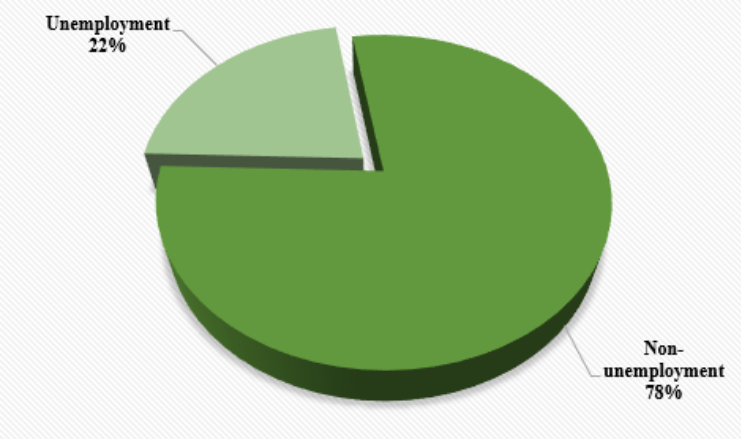

Figure 1: Description of Unemployment Data for Lima Puluh Kota Regency 2023

Based in Figure 1, it can be seen that the unemployed labor force is much larger than the unemployed labor force. So it can be seen that the data is not balanced. To balance the data, it is necessary to handle it using the SMOTE method.

The SMOTE method creates synthetic data based on the nearest neighbor using the value difference metric (VDM) distance for the minor class, namely the unemployed workforce. Where there are 283 respondents who are unemployed so that replication is needed twice to get the number of respondents balanced with the major class. The distribution of data after replication with the SMOTE method is as follows.

Table 3: Data Distribution Using SMOTE

Before SMOTE		After SMOTE		Number of Replication
Mayor	Minor	Mayor	Minor	
Non-unemployment 1029 (78%)	Unemployment 283 (22%)	Non-unemployment 1029 (55%)	Unemployment 829 (45%)	2 times

In Table 3, it can be seen that the amount of data that originally amounted to 41 data will increase to 492 so that it can be said that the number of variables in each class is balanced. In addition to the amount of data on the dependent variable that will increase, the amount of data on each independent variable will also increase following the amount of data on the dependent variable. By balancing the number of members in each type of data in the dependent variable, it is hoped that there will be no cases of underfitting or overfitting and produce a good level of accuracy.

3.1 CART Analysis before Applying SMOTE

The formation of classification trees with the CART method begins by dividing the data into training data and testing data. Training data is used to form a classification tree, while testing data is used for model validation. In this study, the proportion between training data and testing data is 80% and 20%, or respectively 1073 samples for training data and 239 samples for testing data. The variable that has the highest impurity value will be the main separator or root node.

The first stage in applying the CART method is to determine the best parser as the main parser. The main divider obtained is age (X_1). The next stage is the process of pruning the maximum classification tree so that a simpler tree is obtained, or called the optimal tree. The best classification tree produces 4 terminal nodes with 3 independent variables, namely age (X_1), work experience (X_6), and education level (X_5). The diagram of the CART method can be seen as follows.

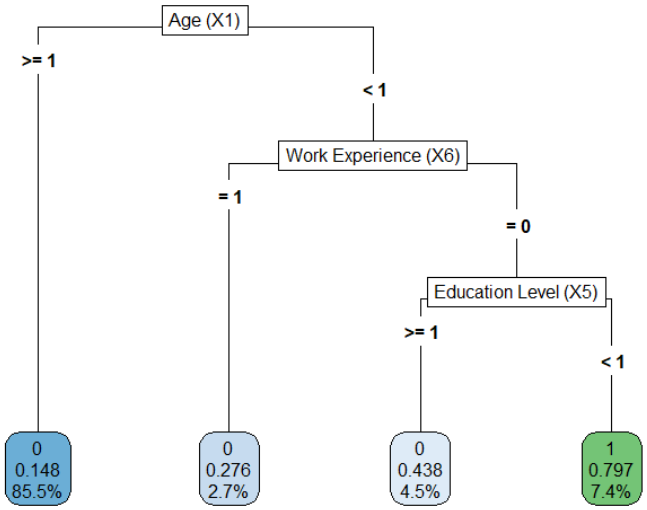

Figure 2: Optimal Classification Tree before Applying SMOTE

Based on Figure 2, it can be seen that the resulting classification tree has 4 terminal nodes, of which 3 are categorized as 0 or unemployment. Forming a model on unbalanced data will result in a large number of terminal nodes on the major data in the classification tree. Thus, the resulting sensitivity will be low on minor data, which means that the classification model ignores minor classes in the classification.

In the CART method, the accuracy of classification will be measured through the accuracy, sensitivity, and specificity values where it can be calculated using confusion matrix.

Table 4: CART Classification Assignment before Applying SMOTE

Actual	Prediction		Provision
	Non-unemployment	Unemployment	
Non-unemployment	177	7	96,1%
Unemployment	41	14	25,4%
Accuration			79,9%

Based on Table 4, the accuracy of CART classification before SMOTE is applied produces an accuracy value of 79,9%, sensitivity of 96,1%, and specificity of 25,4%. The accuracy of classification in the non-unemployment category was predicted as non-unemployment by 177 respondents, the non-unemployment category was predicted as unemployment by 7 respondents, the unemployment category was predicted as non-unemployment by 41 respondents, and the unemployment category was predicted as unemployment by 14 respondents.

3.2 CART Analysis after Applying SMOTE

The initial percentage of observations in the minority category is 22%, while for the majority category it is 78%. SMOTE was then performed on the minor data with some oversampling possibilities. Synthesis data is generated twice from minor data. So that new data is obtained on minor data as many as 849 respondents. The percentage result in the minor category is 45% and in the major data is 55%. Data that has gone through the SMOTE stage becomes more balanced than before SMOTE is applied.

The first stage in applying the CART method is to determine the best parser as the root node. The main disaggregator obtained is age. The next stage is the process of pruning the maximum classification tree so that a simpler tree is obtained, or called the optimal tree. The age variable has the highest impurity value compared to other independent variables. So that the classification tree obtained can be seen as follows.

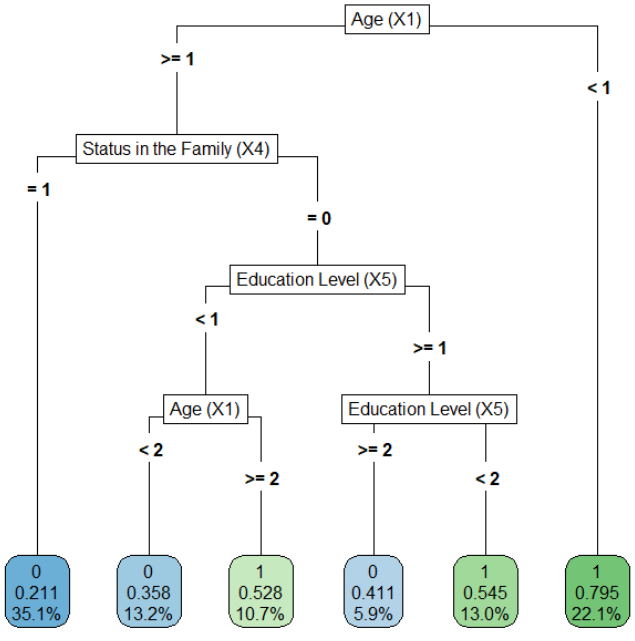

Figure 3: Optimal Classification Tree after Applying SMOTE

The best classification tree produces 6 terminal nodes with 3 independent variables, namely age (X_1), status in the family (X_4), and education level (X_5). The optimal classification tree produces 3 terminal nodes that classify unemployment. Measuring the accuracy of the decision tree classification in classifying unemployment using the CART method after applying SMOTE

Table 5: CART Classification Assignment after Applying SMOTE

Actual	Prediction		Provision
	Non-Unemployment	Unemployment	
Non-unemployment	135	46	71,8%
Unemployment	53	119	72,1%
Accuration			72,1%

Based on Table 5, the accuracy of CART classification after SMOTE is applied produces an accuracy value of 72,1%, sensitivity of 71,8%, and specificity of 72,1%. The accuracy of classification in the non-unemployment category was predicted as non-unemployment by 135 respondents, the non-unemployment category was predicted as unemployment by 46 respondents, the unemployment category was predicted as non-unemployment by 53 respondents, and the unemployment category was predicted as unemployment by 119 respondents.

3.3 Comparison of CART Analysis before and after Applying SMOTE

The comparison of the two classification trees obtained previously is done by comparing the AUC values. At this section we will describes the comparison of CART before and after applying SMOTE to the testing data.

Table 6: Comparison of CART Classification Accuracy before and after Applying SMOTE

Criteria	Before SMOTE (%)	SMOTE (%)
Accuracy	79,9%	72,1%
Sensitivity	96,1%	71,8%
Specificity	25,4%	72,1%
AUC	60,8%	71,9%

Table 6 shows that the total accuracy of the classification tree before applying SMOTE is higher than the classification tree after applying SMOTE, but the classification tree before applying SMOTE cannot classify unemployment well. This can be seen from the sensitivity value which is very large and the specificity which is small. After SMOTE is applied at the stage before data analysis, the sensitivity value obtained decreases and the specificity value increases. It can also be seen that the AUC value after applying SMOTE is higher than the AUC value before applying SMOTE, which is 71.9%. This is the same as previous research, where in the research of Fajriati and Syafriandi (2022) the specificity value of the minor class experienced a high increase, so that the AUC value before applying SMOTE increased after applying SMOTE.

4. Conclusion

Unbalanced data in the minor class category causes a high misclassification in the minor class category. The use of SMOTE applied can balance the amount of data in each unemployed labor force. The AUC value in the CART method after applying SMOTE has increased by 11%. So in this study it can be concluded that the CART classification analysis after applying SMOTE is able to provide better performance compared to the CART classification analysis before applying SMOTE. So that the CART method after applying SMOTE is the best method in classifying unemployment in Lima Puluh Kota Regency.

Reference

Badan Perencanaan Pembangunan Daerah Kabupaten Lima Puluh Kota (2024). *Rancangan Awal Rencana Pembangunan Jangka Panjang (RPJPD) Kabupaten Lima Puluh Kota Tahun 2025-2045*. Lima Puluh Kota, Indonesia: Badan Perencanaan Pembangunan Daerah.

Badan Pusat Statistik (BPS) (2023). *Keadaan Angkatan Kerja di Indonesia*, Indonesia: Badan Pusat Statistik

Breiman, L., Friedman, J. H., Olshen, R. A., dan Stone, C. J. (1984). *Classification And Regression Trees*, New York: Chapman and Hall.

Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique, *Journal Of Artificial Intelligence research*, *16*(2): 321-357.

Cost, S. dan Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, *Machine Learning*, *10*: 57-78.

Fajriati, Y. R. (2022). Pengklasifikasian Status Kerja pada Angkatan Kerja di Kabupaten Tanah Datar Menggunakan Metode CART dan Metode CHAID. *Journal Of Mathematics UNP, 7*(3): 25-33.

Pratiwi, F. E., dan Zain, I. (2014). Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara. *Jurnal Sains dan Seni Pomits*, *3*(1): 54-59.

Ratniasih, N. L. (2014). Konversi Data Training Tentang Pemilihan Kelas Menjadi Bentuk Pohon Keputusan Dengan Teknik Klasifikasi. *Eksplora Informatika 3*(2): 145-154.

Saputro, I. W. dan Sari, B. W. (2019). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Citec Journal*, *6*(1).

Sari, R. V., Firdausi, F., dan Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabisa dengan Menggunakan Algoritma SGD, Random Forest, dan Naïve Bayes. *Jurnal Pendidikan Informatika*, *4*(2): 1-9.

Sofyan, S. & Prasetyo, A. (2021). Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019. In *Seminar Nasional Official Statistics*.

Sumartini, S. H. dan Purmani, S. H. (2015). Penggunaan Metode Classification and Regression Tree (CART) Untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya. *Jurnal Sains dan Seni ITS*, Vol. *4*(2): 211-216.

Wijaya, J., Soleh, A. M., & Rizki, A. (2018). Penanganan Data Tidak Seimbang pada Pemodelan Rotation Forest Keberhasilan Studi Mahasiswa Program Magister IPB, *Xplore, 2*(2): 32-40.

Wijayanti, N. P. Y. T., Kencana, E. N., dan Sumarjaya, I. W. (2021). SMOTE: Potensi dan Kekurangannya Pada Survei. *E-Jurnal Matematika 10*(4): 235-240.