The Classification of Informal Worker using Classification and Regression Tree (CART) and Non-Linear Support Vector Machine (SVM)

R Fitriani1\* and Y Andriyana2

1Master Student of Applied Statistics, Universitas Padjadjaran, Jalan Dipati Ukur No.35, Lebakgede, Kecamatan Coblong, Kota Bandung, Jawa Barat 40132, Indonesia

2Department of Statistics, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang KM. 21 Jatinangor 45361, Sumedang, Jawa Barat, Indonesia

\*rinda19001@mail.unpad.ac.id

**Abstract**. The problem of decent work availability is still be big challenges in Indonesia. A key element of decent work is employment opportunities where people works in informal sectors is the main indicator. However, the informal sectors will derive new problems on long time period because informal workers are very vulnerable to economic shocks. They do not have a certain income, lack of basic social or legal protections such as health insurance or accident insurance, and no pension plan. There are various factors that influence who works in informal sectors. In this study we want to classify the type of worker who are working on informal and formal sectors. To answer such problem there exists a classification technique called Classification and Regression Tree (CART) and Non-linear Support Vector Machine (SVM). Implementing both techniques to worker status in Blitar district in 2019. We have that the SVM method perform better than CART due to the classification accuracy score where the CART score is 79.56 percent while the SVM reaches 85.32 percent

1. Introduction

One of the big challenges that faced by Indonesia at the moment is employment opportunities problem. Not only about number of jobs that can accommodate a lot of workers but the employment created should have quality, be able to guarantee workers develop their capacity, respect human rights, and provide prosperity [1]. The International Labour Organization (ILO) stated this concept in a decent work agenda that is the main aspect in poverty eradicate, equity achievement, and inclusive sustainable development [2]. This agenda also stated in the eighth goals of Sustainable Development Goals (SDGs) that promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all [3].

The substantive employment opportunities, a key element of decent work, was comprised of indicators that provide insights regarding the quantity of labour demand and supply in an economy [2]. One of the indicators is people who works in informal sector which is defined based on their main work status. Own account worker, employer assisted by temporary worker, casual worker, and family/unpaid worker are include in informal worker [1].

Informal sectors give significance role in an economy which provide employment for the people who cannot be accommodated by formal sectors [4]. It showed that informal sector can help government to reduce unemployment. However, the informal sectors will derive new problems on long time period because informal workers are very vulnerable to economic shocks. They don’t have a certain income, lack basic social or legal protections as health insurance or accident insurance, and no pension plan [5]. Most of the informal workers are unskilled worker which low productivity job in marginal, small-scale, and often family-based activities [6].

According to National Labor Force Survey 2019, the lowest unemployment rate was Blitar district (3.11 percent) which compared to six other districts/municipalities in Karisidenan Kediri [7]. Meanwhile, the result of Economic Census 2016 showed that 87.56 percent of employees in Blitar is involved in Micro Small Enterprises (UMK) activities with small-scaled characteristics, simple technology, family based, low education/skill, and relatively low wage which were informal sectors [8]. It indicates the role of informal sector in Blitar still leaves the issue whether the available employment are decent work that can improve people standard living.

There are various factors that influence someone work in informal sector. According to [9] the main determinants of the probability to be informally employed are males, workers with low educational attainment, and workers who are employed in construction, retail trade, and the restaurant business are particularly affected. This result is supported by studies that conducted by Hartmut and Anzelika [10] that young worker, males, worker with primary education or less, persons with low skills, workers in construction and trade and related services have a substantially higher likelihood of being informally employed.

The classification method becomes important instruments to evaluate and take conclusion whether a person is a formal or informal worker. Worker status classification can be done by several methods including Classification and Regression Tree (CART) and Non-linear Support Vector Machine (SVM) method. Therefore, this study aims to classify the type of worker who are working on informal and formal sectors by several factors excluding gender, age, marital status, education level, residence, wage, and main industry.

1. Method

|  |
| --- |
| **Table 1.** Predictor Variables |
| 6 pt |  |
| Variable | Categorization |
| Gender $\left(X\_{1}\right)$ | 1: Male2:Female |
| Residence $\left(X\_{2}\right)$ | 1: Urban2: Rural |
| Age $\left(X\_{3}\right)$ | 1: Productive (15-64 years old) 2: Unproductive (more than 64 years old) |
| Marital Status $\left(X\_{4}\right)$ | 1: Single, 2: Married, 3: Divorce, 4: Death Divorce |
| Education level $\left(X\_{5}\right)$ | 1: Not going to school, 2: Elementary school3: Junior High school, 4: Senior High school,5: University |
| Main Industry $\left(X\_{6}\right)$ | 1: Agriculture, 2: Manufacture, 3: Services |
| Wage $\left(X\_{7}\right)$ | 1: Unpaid; 2: 1-999,999; 3: 1,000,000-1,999,999; 4: 2,000,000-4,999,999; 5: more than 5,000,000 |

This study use data from National Labor Force Survey August 2019 in Blitar district which consist of 1,465 respondents who work. The concept of work is economic activity carried out by someone with the intention of obtaining or helping to obtain income or profit, at least 1 hour (uninterrupted) during previous week. These activities include unpaid worker activities [11]. The response variable is worker status that is formal or informal worker while there are seven predictor variables which is listed in the Table 1.

* 1. Classification and Regression Tree (CART)

CART analysis is a tree-building technique which is unlike traditional data analysis method. It is ideally suited to the generation of clinical decision rule. CART is often able to uncover complex interactions between predictors which may be difficult or impossible to uncover using traditional multivariate techniques [12]. CART aims to get an accurate data group as a characteristic of a classification, besides CART is also used to describe the relationship between response variables with one or more predictor variables. The tree building model based on scale of response variables, if response variables has continuous scale then tree model is regression tree, if response variables has categorical scale then tree model is classification trees [13].

CART analysis has a number of advantages over other classification methods, including multivariate logistic regression. First, it is inherently non-parametric. In other words, no assumptions are made regarding the underlying distribution of value of the predictor variables. CART can be applied to large datasets with various variable scale, relatively simple to interpret, more accurate and faster calculation [12]. The classification method involves learning data set and testing data set. Learning data set is used to create optimal classification trees while testing data set is required to validate a model how able the model can predict new data.

The steps in implementing CART algorithm are as follows [12]:

1. Tree Building begins at the root node, which includes all observation in the learning data set.
2. Classifier selection is used to get classifier which able to generate node with the highest homogenous value of response variables. It is measured by impurity measure $i(t)$. The impurity measure function is used in this study is the “Gini”.
3. Determination of terminal node

If a node $t$ does not significantly decrease impurity is called terminal node.

1. Class labeling on terminal node is based on a maximizing rule.
2. Tree Pruning in order to generate a sequence of simpler trees, each of which is a candidate for the appropriately-fit final tree, the method of “cost-complexity” pruning is used. This method relies on a complexity parameter, denoted $α$, which is gradually increased during the pruning process.

$R\_{α}\left(t\right)=R\left(t\right)+α\left|\tilde{T}\right|$ (1)

where $R\left(t\right)$ is error proportion of sub-tree, $\left|\tilde{T}\right|$ is size of number of terminal node tree T

1. Optimal Tree Selection

The goal in selecting the optimal tree, defined with respect to expected performance on an independent set of data, is to find the correct complexity parameter $α$ so that the information in the learning data set is fit but not overfit. Then, cross validation is required to validate the model from learning data set.

* 1. Support Vector Machine (SVM)

SVM is a technique which have been developed by Vapnik in 1995 to make predictions in both classification and regression [14]. SVM provides a powerful method for classification which is classified in the class of supervised learning [15]. This machine learning techniques was considered due to its performance in predicting the class of new data.

Each of training data expressed by ($x\_{i},y\_{i}$) with $i=1,2,…,n$ and $x\_{i}=\left\{x\_{i1},x\_{i2},…,x\_{iq}\right\}^{T}$ is an attribute set for data$ i$ in related to it $y\_{i}\in \left\{-1,+1\right\}$ represent class label. Hyperplane linear classification SVM denoted as

$w.x\_{i}+b=0$ (2)

with $w$ is weight vector and $b$ is bias, $w$ and $b$ are parametric model and $w.x\_{i}$ is an inner-product between $w$ and $x\_{i}$. Data $x\_{i}$ that classified as (-1) class is a data that qualified for inequality

$w.x\_{i}+b\leq -1$ (3)

Data $x\_{i}$ that classified as (+1) class is a data that qualified for inequality

$w.x\_{i}+b\geq +1$ (4)



**Figure 1.** Linear Model of Support Vector Machine.

The equation (3) and (4) can be illustrated in figure 1. Substracting equation (3) to (4), it can be obtain that $w\left(x\_{b}-x\_{a}\right)=2 $where $x\_{b}-x\_{a}$ is a vector parallel in hyperplane position. Margin hyperplane was given by a distance between two hyperplane from those two classes. The presented notation is summarized into

$\left‖w\right‖.d=2$ or $d=\frac{2}{\left‖w\right‖}$ (5)

* 1. Hyperplane SVM

Hyperplane SVM is separable function between two data set from two different class. Data class classification on equation (3) and (4) can be combined with notation:

$y\_{i}\left(w.x\_{i}+b\right)\geq 1, i=1,2,…,N$ (6)

Optimal margin is counted by maximizing the distance between hyperplane and the closest data which formulated in (5). This problem can be solved with optimization method namely as Quadratic Programming (QP) by minimizing inverse equation (5). Thi**s** optimization can be solved by Lagrange multiplier:

$L\_{p}=\frac{1}{2}\left‖w\right‖^{2}-\sum\_{i=1}^{n}α\_{i}.y\_{i}\left(w.x\_{i}+b\right)$ (7)

$α\_{i}$ is lagrange multiplier that correspond with $x\_{i}$. The value of $α\_{i}$ is either zero or positive. The optimization problem given above is still difficult to solve due to numerous parameters. Hence, Langrange optimization equation must be transformed to functional lagrange multiplier itself (known as problem duality). Lagrange multiplier equation can be elaborated into:

 $L\_{p}=\frac{1}{2}\left‖w\right‖^{2}-\left(\sum\_{i=1}^{n}α\_{i}.y\_{i}\left(w.x\_{i}\right)-b\sum\_{i=1}^{N}α\_{i}.y\_{i}+\sum\_{i=1}^{N}α\_{i}\right)$ (8)

Optimal requirement located at the third term on the right segment in the equation and force this term to be as equal as 0. By replacing $w$ and $\left‖w\right‖^{2}=w\_{i}.w\_{j}$term, equation above will transform into Lagrange multiplier duality in form of $L\_{d}$ where $x\_{i}x\_{j}$ is a dot-product, two data in training data.

$L\_{d}=\sum\_{i=1}^{N}α\_{i}-\frac{1}{2}\sum\_{i,j}^{}α\_{i}α\_{j}y\_{i}y\_{j}x\_{i}x\_{j}$ (9)

* 1. Non-Linear SVM

Most of data in the real life are non-linear, data rarely to be linear separable [16]. To solve non-linear problem, SVM is modified by adding Kernel Function [17]. In non-linear SVM, first data $x$ is mapped by the function $Ф(x)$ to a higher dimensional vector space. In this new vector space, the hyperplane that separates both classes can be constructed. Generally, the transformation of $Ф$ unknown, and very difficult to understand, then dot-product calculation can be changed with kernel function which is known as kernel trick.

Kernel trick provides a variety of conveniences, because in the SVM learning process, to determine the support vector, it can be identified by knowing the kernel function being used, and does not need to know the form on non-linear functions itself. This study will use four type of kernel functions such as linear, Radial Basis Function (RBF), sigmoid and polynomial.

* 1. SVM Parameter Tuning

In the supervised learning method, the datasets is divided into two parts, namely training and testing data set. Training data set is used to determine model parameter which is necessary to optimize. Grid search method was used to find the best parameters. Grid search is generally a combination of parameters tested to the SVM model to get the classification error value in the training set. The value of classification error is tested on the training set by using cross validation technique.

Cross validation is the simplest and most widely used method for estimating prediction error [18]. This study will use 10-fold cross validation method. Training data sets are randomly partitioned into 10 equal-sized subsets. Subset 1 is used as testing data set and the other subsets (2-9) are used as training set. That process is done 10 times. The results of each validation process are the averaged to obtain a single classification error value.

|  |
| --- |
| **Table 2.** Grid search points on the rough grid. |
| 6 pt |  |
| Parameter | Value |
| Gamma | $$2^{-3},2^{-2},2^{-1},2^{0}, 2^{1}, 2^{2},2^{3}$$ |
| Cost | $$10^{-1},10^{0}, 10^{1}, 10^{2}$$ |

The selection of grid search points on cross validation becomes a crucial point in determining the best model. Selecting wrong grid search point will decrease the generalization ability of SVM. This study will do two steps grid search which are rough grid search in the first step (Table 2) and finer grid search in the second step (Table 3).

|  |
| --- |
| **Table 3.** Grid search points on the finer grid. |
| 6 pt |  |
| Gamma on the rough grid | Precision |
| $$2^{-3}$$ | $$0.05,0.1,0.125,0.15,0.175$$ |
| $$2^{-2}$$ | $$0.2, 0.25,0.3,0.35,0.4$$ |
| $$2^{-1}$$ | $$0.3,0.4,0.5,0.6,0.7$$ |
| $$2^{0}$$ | $$0.75,1,1.25,1.5,1.75$$ |
| $$2^{1}$$ | $$1.5,2,2.5,3,3.5$$ |
| $$2^{2}$$ | $$3.5,4,4.5,5,5.5$$ |
| $$2^{3}$$ | $$7,7.5,8,8.5,9$$ |

The evaluation of classification accuracy can be showed by value of Apparent Error Rate (APER). APER value is proportion of error classification sample by classification function [19]. The APER value can be calculated by classification error table.

1. Result and Discussion
	1. CART Analysis

Learning data set is used for generating tree classification while testing data set is used to validate model. The combination of learning data set and testing data set which provide the highest accuracy classification value is 75 percent learning data set and 25 percent testing data set (Table 5), and hence it is used in the next analysis.

The classification trees can be divided into maximum classification trees and optimal classification trees. Maximum classification trees is a classification tree with the highest number of terminal node. All of predictor variables are included in the maximum classification trees. Whereas, the optimal classification tree is the pruning of maximum classification tree that has been generated based on the rule of cost complexity minimum.

|  |
| --- |
| **Table 5.** The accuracy of classification given data combination. |
| No. | Data Combination (%) | Accuracy |
| Learning | Testing | Learning | Testing |
| 1 | 70 | 30 | 0.7923 | 0.7799 |
| 2 | **75** | **25** | **0.7952** | **0.7960** |
| 3 | 80 | 20 | 0.7970 | 0.7855 |
| 4 | 85 | 15 | 0.7872 | 0.7690 |
| 5 | 90 | 10 | 0.7869 | 0.7632 |
| 6 | 95 | 5 | 0.7945 | 0.7857 |

The optimum classification tree from CART method for worker in Blitar district is showed in Figure 2. From seven predictor variables, only four variables exist in the classification tree. That variables are wage, education, age, and marital status. Whereas the three other variables not significantly influence in classification tree building.



**Figure 2.** Worker Classification Tree in Blitar District 2019.

According Figure 2 can be obtained three node for formal worker class and four node for informal worker class. The characteristics of formal worker class as follow:

1. Worker with the wage of 5 million rupiah and above with the higher level of education is university
2. Worker with wage of less than 5 million rupiah with the higher level of education is Senior high school and above, work in other than agriculture industry
3. Worker with wage of less than 5 million rupiah with the higher level of education is less than Senior high school, on productive age and single status.

The characteristics of informal worker class as follow:

1. Worker with the wage of 5 million rupiah and above with the higher level of education is less than Senior high school
2. Worker with wage of less than 5 million rupiah with the higher level of education is Senior high school and above, work in agriculture industry
3. Worker with wage of less than 5 million rupiah with the higher level of education is less than Senior high school, on productive age, the marital status other than single
4. Worker with wage of less than 5 million rupiah with the higher level of education is less than Senior high school and unproductive age.

An evaluation of 25 percent of the testing dataset was 367 respondents. There are misclassification of 75 respondents where 19 informal workers are classified as formal workers and 56 formal workers are classified as informal workers. The classification accuracy was 79.56 percent with classification error of 20.44 percent. The classification accuracy values above 75 percent indicate that the CART method can describe the classification of worker status (formal or informal).

* 1. SVM Analysis

This study used 75 percent and 80 percent training data set and each is processed through four kernel models with the grid search method. The determinant of the best model is done by selecting rough and finer grid search points on cross validation to get optimal cost and gamma parameter value which corresponds to the smallest APER score. Table 6 shows the optimum parameters of training data set in each kernel and the APER value with finer grid search.

|  |
| --- |
| **Table 6.** Optimum score parameter and APER training data set with finer grid search. |
| Kernel | 75% Training data | 80% Training data |
| Cost | Gamma | APER | Cost | Gamma | APER |
| Linear | 4 | 100 | 21.61 | 0.5 | 100 | 21.68 |
| RBF | 0.25 | 1 | 15.55 | 0.3 | 1 | 15.90 |
| Sigmoid | 0.1 | 0.1 | 23.38 | 0.1 | 0.1 | 22.95 |
| Polynom (2) | 0.25 | 100 | 21.56 | 4 | 0.1 | 22.01 |
| Polynom (3) | 0.05 | 10 | 17.95 | 0.05 | 10 | 17.99 |

After obtaining the optimum parameter of training data set, the next step is to test the accuracy of the model parameters of each kernel in the test data set. In Table 7, showing the optimum parameter accuracy of each kernel to testing data set, where RBF kernel model with parameter cost 0.3 and gamma 1, 20 percent on testing dataset is the best model obtained in this study with 85.32 percent prediction accuracy score. The score accuracy over 70 percent can be considered to be a good illustration that SVM model can depict whether people works as formal or informal worker. There are misclassification of 43 respondents where 39 informal workers are classified as formal workers and 4 formal workers are classified as informal workers.

|  |
| --- |
| **Table 7.** Accuracy score in test data set. |
| Kernel | 25% Testing data | 20% Testing data |
| Cost | Gamma | Accuracy | Cost | Gamma | Accuracy |
| Linear | 4 | 100 | 75.48 | 0.5 | 100 | 74.40 |
| RBF | 0.25 | 1 | 83.11 | **0.3** | **1** | **85.32** |
| Sigmoid | 0.1 | 0.1 | 72.48 | 0.1 | 0.1 | 74.40 |
| Polynom (2) | 0.25 | 100 | 79.29 | 4 | 0.1 | 79.52 |
| Polynom (3) | 0.05 | 10 | 78.47 | 0.05 | 10 | 77.82 |

Based on the best SVM model selected can be applied to classify new respondents. Table 8 are presented examples of new respondents to be classified using the selected SVM model. The first respondent states that males, live in rural area, productive age, single, junior high school graduate, unpaid worker in agriculture industry will tend to be informal worker. In the second respondent states that males, live in urban area, productive age, single, university graduate, employee in services industry with wage of 5 million rupiah and above will tend to be formal worker. The third respondent states that females, live in rural area, unproductive age, death divorce status, not going to school, work in manufacture industry with wage of less than 1 million rupiah tend to be informal worker.

|  |
| --- |
| **Table 8.** Decision consideration on SVM to new respondent profile. |
| Respondent | Gender | Residence | Age | Marital Status | Education | Main Industry | Wage | Decision |
| 1 | 1 | 2 | 1 | 1 | 3 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 2 | 5 | 3 | 5 | 0 |
| 3 | 2 | 2 | 2 | 4 | 1 | 2 | 2 | 1 |

1. Conclusion

The CART method can describe the classification of worker in Blitar district where the optimum tree classification building is 7 nodes consist of 4 variables with classification accuracy of 79.56 percent on 25 percent test set. The SVM method can depict the classification of worker in Blitar district where RBF kernel model with parameter cost 0.3 and gamma 1, 20 percent on test set is the best model obtained with 85.32 percent prediction accuracy score. Furthermore, the SVM method is a better method because it has a higher classification accuracy score than CART method.

1. References

[1] Badan Pusat Statistik 2018 *Indikator Pekerjaan Layak di Indonesia 2017* (Jakarta: BPS RI)

[2] International Labour Organization 2013 *Decent Work Indicators ILO Manual Second version* (Geneva: ILO)

[3] United Nation Development Programme 2015 *Sustainable Development Goals Booklet*

[4] Asian Development Bank and Badan Pusat Statistik 2011 *The Informal Sector and Informal Employment in Indonesia* (Phiippines: ADB)

[5] Ronald G and Dhyah H 2013 Entrepreneurial Motivation Pengusaha Sektor Formal dan Informal di Jawa Timur *Jurnal AGORA* **1**

[6] Leonardo G and Leopoldo T 2007 Labor Informality in Latin America and the Caribbean: Patterns and Trends from Household Survey Microdata *Documento de Trabajo* Nro.46

[7] Badan Pusat Statistik 2020 *Provinsi Jawa Timur Dalam Angka 2020* (Surabaya: BPS Jatim)

[8] Badan Pusat Statistik 2017 *Potensi Ekonomi Kabupaten Blitar* (Blitar: BPS Kabupaten Blitar)

[9] Vladimir G and Rostislav 2014 Between Light and Shadow: Informality in the Russian Labour Market *IZA Discussion Paper* No 8279

[10] Hartmut L and Anzelika Z 2013 Re-defining Informal Employment and Measuring its Determinants: Evidence from Russia *IZA Discussion Paper* No.7844

[11] Badan Pusat Statistik 2019 *Keadaan Angkatan Kerja di Indonesia Agustus 2019* (Jakarta: BPS RI)

[12] Roger J L 2000 An Introduction to Classification and Regression Tree (CART) Analysis *Presented at the 2000 Anual Meeting of Society For Academy Emergency Medicine in San Fransisco California*.

[13] Leo B, Jerome H F, Richard A O, Charles J S 1998 *Classification And Regression Trees* (London: Chapman & Hall/CRC)

[14] Steve R G 1998 Support Vector Machine for Classification and Regression *Technical Report* University of Southampton

[15] Stephen W H and Sam M 2007 SVM Clustering *Fourth Annual MCBIOS Conference Computational Frontiers in Biomedicine* New Orleans

[16] Muhamad B J 2018 *Perbandingan Kernel Trick Pada Non Linier Support Vector Machine* Tesis (Bandung: Universitas Padjadjaran)

[17] Jakub N and Michal K 2018 Selecting training sets for support vector machines: a review *Artificial Intelligence Review* **52**

[18] Trevor H, Robert T, Jerome F 2008 *The Elements of Statistical Learning Data Mining, Inference, and Prediction* Second Edition Springer Series in Statistics

[19] Richard A J and Dean W W 1982 *Applied Multivariate Statistical Analysis* (New Jersey: Prentice Hall)