Implementation of Machine Learning C4.5 Algorithm to Forecast Regional Economic Development Classification

Hendra M Setiawan
Department of Economic and Business
Tanjungpura University, Pontianak, Kalimantan Barat
Tel: 08115757212, Email:hendra.msetiawan1991@gmail.com

Aditya Wedha
Department of Informatics and Engginerring
Telkom University, Bandung, Indonesia
Tel: 081296642999, Email: awedha1993@gmail.com

Abstract Regional economic development is measurable through several classification models. One of the models most frequently used by the government and researchers is classen typology. In machine learning, the classification function can be done using decision tree algorithm. The most frequently used algorithm is C4.5. This research will discuss the implementation of C4.5 algorithm for forecasting regional economic development based on classen typology. One of the findings of this research is that we managed to convert numerical continuous attribute into discrete attribute and eventually this improved the accuracy of C4.5 classification. Another important finding of this paper is that the results of classen typology was highly dependent to accumulation of the RGDP of the whole regencies which made up the RGDP of a province; therefore, in order to prevent any bias in determining the classification and to give a broader perspective in conducting the analysis, we had to compare a regency with a province that had relatively similar economic performance by using decision tree we could reduce the steps to comparing a regency to other province. And the last finding is that we could make a forecast on the classen typology classification results without complete data.

Keywords: regional economic development, C4.5 algorithm, classen typology, classification

1. INTRODUCTION

Classen typology is usually used for the classification of regional economic growth of a regency/municipality through the comparison of Regional Gross Domestic Product and Regional Gross Domestic Product Per Capita to the province serving as the reference. The results of Classen typology consist of 4 different classes, "Professor Sjafrial (2008) showed that with classen typology we could classify regency into 4 different result class by using RGDP growth and RGDP Per Capita as an indicator (determining factor)". The field of machine learning has some algorithms that function in classification as same as classen typology "Raphael Bost, Raluca Ada Popa Stephen Tu, and Sha Goldwasser. (2014) showed that Nowaday machine learning classification is used for numerous task, such as medical or genomics predictions, spam detection, face recognition and financial prediction." This research aimed to implement C4.5 algorithm, one of the machine learning algorithms for forecasting regional economic development classification based on classen typology. The areas that became the target in this research were the regencies/municipalities in West Kalimantan Province.

There are some previous researches that have implemented the thing similar to that of this research, namely comparing machine learning technique to classen typology. "Regional Development Classification Model using Decision Tree Approach by Tb. Ai Munandar Eng. Informatics Dept. Universitas Serang Raya Serang – Banten – Indonesia and Edi Winarko Computer Science and Electronic Dept. Universitas Gajah Mada Yogyakarta – Indonesia". This research discusses one of machine learning techniques when compared to classen typology. The difference from the previous researches lies on the technique used for obtaining the dataset that would be processed by C4.5 application (weka).

The main objective of this research was to implement the decision tree model obtained from C4.5 algorithm and use the decision tree model to forecast the classen typology classification results for the next year. By using the decision tree that we have obtained, we can easily change the province serving as a reference without having to repeat all steps of classen typology. Another significance of this research is that the results of the decision tree can be used as a recommendation based on data for decision makers in formulating the decisions that can drive the regional economic growth into a better direction.

This research consists of five sections. The first section contains the background of the research topic. The second section contains the resume of classen typology, C4.5 Algorithm, and related fields. The third section contains research methods such as the research framework, data sources, instruments and technique that were used. The fourth section contains the results of the data analysis. And the fifth section is conclusions.

2. Overview

2.1 Classen Typology

Classen typology is used to present the classification of regencies or municipalities in determining the regional economic development. This technique is based on 4 matrix quadrants which are divided based on the RGDP or RGDP-per-capita of the regency/municipality and province.

Tabel 1 :Classen typology Matrix

RGDP Growth Rate (R)	**RGDP Per Capita (Y)**	
	Yi>Y	**Yi<Y**
Ri>R	Advance and Fast Growing Regency/City (Quadrant I)	Rapidly Growing Regency /City (Quadrant II)
Ri<R	Advancebut Pressured Regency /City (Quadrant III)	Disadvantages Regency /City (Quadrant IV)

Remark Yi= RGDP Percapita in Regency/City; Ri= RGDP Growth Rate in Regecy/City; Y= RGDP Percapita in West KalimantanProvince; R= RGDP Growth Rate In West Kalimantan Province;

1) Quadrant I, is an advanced and fast-growing Regency/City. This category is indicated by the level of RGDP growth and RGDP per capita higher than reference region.

2) Quadrant II,is an rapidly growing Regency/City. Indicated by a higher rate of RGDP growth but RGDP per capita income is lower than the reference region.

3)Quadrant III is an advanced but pressured Regency/City. Indicated by a lower rate of RGDP growth but RGDP per capita is higher than reference region.

4) Quadrant IV, is a relatively underdeveloped regions, the rate of RGDP growth and RGDP per capita, lower than reference region.

To obtain growth rate we are using this equation

$$R(i) = \frac{GDPRegency(t) - GDPRegency(t-1)}{GDPRegency(\ t-1)} x100\% \quad (1)$$

$$R = \frac{GDPProvince(t) - GDPProvince(t-1)}{GDPProvince\ (\ t-1)} x100\% \quad (2)$$

Meanwhile, the data of RGDP and RGDP-per-capita (Y(i) and Y) growth were obtained from the Central Bureau of Statistics of West Kalimantan Province.

As an illustration, let us assume that the data of the regency/municipality RGDP is 10,000 and the Province RGDP is 9,000, while the RGDP per capita for regency or municipality is 2,000 and for province is 2,000 in 2010. In 2011, there is an increase in the RGDP for regency and municipality into 11,000 and for province into 9,500 as well as an increase in the RGDP-per-capita for regency and municipality by 3,000 and province by 2,000. Based on the assumption above, we enter the data for classen typology calculation with 2010 (t-1) and 2011 (t). From the assumption that we have obtained using equation 1 we obtained that the R(i) is 10% and R is 5%, and we also figured out that Yi is higher than Y. Using the constraint from classen typology matrix in which R(i) > R and Y(i) > Y, we found a result that x regency/municipality is located in Quadrant I – Advanced and Fast Growing Regency/City. Classen typology has some weakness, the first of which is that the results of classen typology classification highly depend on the RGDP and RGDP per capita value of the province of the regencies/municipalities that would be classified. Therefore, if the regencies/municipalities classified are located in an extremely advanced province economically and supported by regencies/municipalities that mostly have advanced economy as well compared to the regencies/municipalities that are located in a relatively developing province, the result may be different (biased). For example, if x regency/municipality is classified into Quadrant II and if the province is changed into a more advanced province, the classification result may change to

Quadrant II or Quadrant III. Therefore, to obtain a broader classen typology analysis base, it is recommended to make classifications to two different provinces, namely a province with more advanced economic performance and a province that has relatively similar economic performance.

2.2 C4.5 Algorithm

"Ron Kohavi and Ross Quinlan (1999) showed that C4.5 belongs to a succession of decision tree learners that trace their origins back to the work of Hunt and others in the late 1950s and early 1960s".Its immediate predecessors of ID3.C4.5 have some weakness "Quinlan (1996) showed that Result of C4.5 tree-induction algorithm provides good classification accuracy and fastest process compared with another machine learning algorithms but for nominal continous attribute C4.5 have some weakness. Several authors have recently noted that C4.5's performance is weaker in domains with a prepon-derance of continuous attributes than for learning tasks that have mainly discrete attributes".

C4.5 result will be decision tree "Salvatore Ruggieri (2002) showed that A decision tree is a tree data structure consisting of decision nodes and leaves. A leaf is a class value. A decision node show a test running on some attribute. For each possible outcome of the test, a child node could present or a leaf will present. A decision tree function is to classify a set of data, i.e. to assign a result value to some data depending on the values of the attributes of the data. A performance measure of a decision tree over a set of cases is called classification error. It is the percentage of missclassified, for every data that classified".C4.5 algorithm main method in contructing decision tree is divide and conquer strategy. Every node associated with a set of cases. At the beginning, only the root is present, with associated the whole training set.

2.3. Gross Domestic Product

GDP is main variable in this research "Gregory Mankiw (2015) showed that Gross domestic product (GDP) is the market value of all final goods and services produced within a country in a given period of time." In this research we used the RGDP or Regional Gross Domestic Bruto of West Kalimantan Province. "Gregory Mankiw (2015) showed that GDP is the most closely watched economic statistic because it is thought to be the best single measure of a society's economic well-being." GDP consists of Real GDP and Nominal GDP. Nominal GDP is the production of goods and service based on the price applied currently, while real GDP is the value of goods and serviced based on the constant price. "AAUI (2003) showed that The Federal Reserve uses data such as the real GDP and other related economic indicators to adjust its monetary policy". This research used Real GDP that assessed based on the basic year 2000. Besides, GDP-per-capita served as one of important variables in this research because the growth of the GDP-per-capita can always capture the development of economic growth. "Lill Andersen and Ronald Babula (2008) showed that The economic theory distinguishes between two sources of GDP-per-capita growth: capital accumulation (physical and human) and productivity growth".

2.4. Kalimantan Barat Province

Under the 1956's Law No 25, Kalimantan Barat has obtain a new status as Otonomous Province with pontianak as capital city since 1 january 1957. Later, this date considered as birthday of Kalimantan Barat Province. Kalimantan Barat Province consist of 14 regencies it is, Sambas Regency, Bengkayang Regency, Landak Regency, Pontianak Regency, Sanggau Regency, Ketapang Regency, Sintang Regency, Kapuas Hulu Regency, Sekadau Regency, Melawai Regency, Kayong Utara Regency, Kubu Raya Regency, Pontianak Regency and Singkawang Regency. All regency in kalimatan barat used as data in this research. Decision tree models will used to forecast classen typology result in all 14 regency in Kalimantan Barat Province.

2.5. Macroeconomic Forecasting

Macroeconomic Forecasting have a very high rates of error. As stated by Gregory Mankiw "If forecasters could accurately predict the condition of the economy a year in advance, then monetary and fiscal policymakers could lookahead when making policy decisions. In this case, policymakers could stabilize the economy despite the lags they face". That is the reason why we need some precise data driven forecasting. "Francis X. Diebold (1997) showed that Understanding the future of macroeconomic forecasting requires understanding the interplay between measurement and theory, and the corresponding evolution of the nonstructural and structural approaches to forecasting. Nonstructural macroeconomic forecasting methods attempt to exploit the reduced-form correlations in observed macroeconomic time series, with little reliance on economic theory. Structural models, in contrast, view and interpret economic data through the lens of a particular economic theory."

3. RESEARCH METHODOLOGY

The research method used in this research can be seen in Figure 1 below. The first stage of this research was started from the collection of RGDP data for 14 regencies/municipalities in West Kalimantan Province. The data series used in this research was from 2011 to 2014 and it would be used as data training for making the decision tree model. After the first stage was done, the research would proceed to the second stage that contained two steps. The first step was finding the results of classen typology. Then, the the results from this step were used as the benchmark of the decision tree and dataset accuracy in making the decision tree using weka application.

In this research, the dataset was not directly input into weka. Firstly, a conversion from numerical continuous attribute into discrete attribute was done. The purpose was to obtain a more accurate decision tree

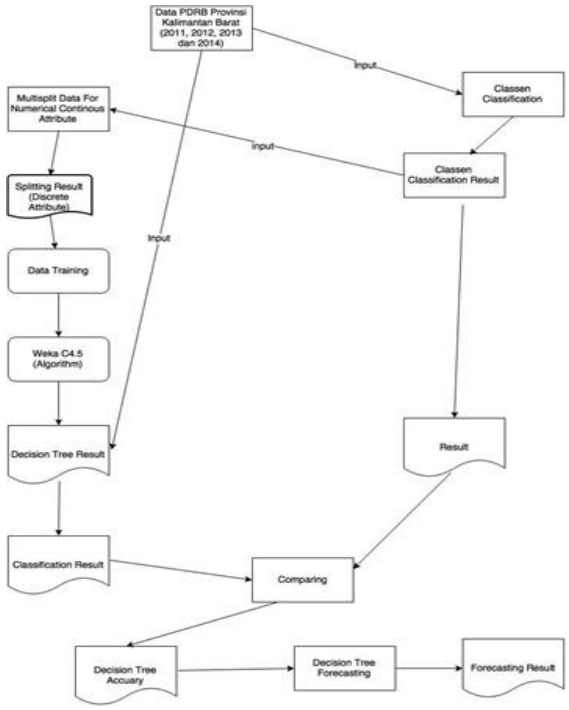

Figure 1: Research Methodology

After the discrete attribute that served as the data training set was obtained, the data input process into Weka was carried out. Weka is an application that can be used for making a decision tree based on C4.5 algorithm. After weka finished the decision tree modeling process, the next process was comparing the results from the decision tree classification to the classen typology classification to find out the accuracy of the decision tree classification in the same year. The last stage was forecasting the classen typology for the next year using the decision tree model that had been obtained.

4. RESULT

4.1 Classen Typology Result

From all diagram of 2012, 2013 and 2014 we could conclude that, There are exist 3 Regency that always classified in first quadran or classified as advance and fast growing regency it was Kubu Raya Regency, Pontianak Regency and Singkawang Regency. This three regency know as most advance and famous regency in Kalimantan Barat Province. In the other hand also exist some regency that always classified as disadvantages regency it was bengkayang regency and landak regency. But in 2014 kalimantan barat province facing an economic slowdown that made kalimantan barat province RGDP growth rate fall into 5 % in 2014. Because of this economic slowdown two regency that usually classified as disadvantage regency in 2012 changes into Rapidly Growing Regency it was sambas regency and mempawah regency. That prove if economic slowdown or economic boom happens it could increasing errors in forecasting. "Barry Eichengreen, Donghyun Park and Kwanho Shin (2011) showed that all fast growing economies eventually slow down. The question is when."

4.2 Decision Tree Model Result and Variable Analysis

This research succesfully transform numerical continous attribute of RGDP into discrete attribute. Result show in table 3,4 and 5.

Tabel 2: Result For Every Regency And City 2013 Year

Data Set	RGDP Growth	RGDP Percapita	Discrete – RGDP Growth	Discrete – Percapita	Typology Result	Typology Result (Quadran)	Decision Tree Result	Forecasting Same Year
Sambas Regency	6,18	19,722	Rendah	RR	KabupatenTertinggal	Q IV	Q IV	Right
Benglayang Regen	5,90	18,793	Rendah	RR	KabupatenTertinggal	Q IV	Q IV	Right
Landak Regency	5,23	15,042	Rendah	RR	KabupatenTertinggal	Q IV	Q IV	Right
Mempawah Regen	5,44	15,213	Rendah	RR	KabupatenTertinggal	Q IV	Q IV	Right
Sanggau Regency	5,98	23,931	Rendah	TT	KabupatenMajuTapiTertekan	Q III	Q III	Right
Ketapang Regency	4,55	27,331	Rendah	TT	KabupatenMajuTapiTertekan	Q III	Q III	Right
Sintang Regency	6,47	18,472	Tinggi	RR	KabupatenBerkembangCepat	Q III	Q III	Right
Kapuas Hulu Rege	5,23	20,629	Rendah	S	KabupatenTertinggal	Q IV	Q IV	Right
Seladau Regency	6,56	16,106	Tinggi	RR	KabupatenBerkembangCepat	Q II	Q II	Right
Melawi Regency	4,85	13,551	Rendah	RR	KabupatenTertinggal	Q IV	Q IV	Right
Kayong Utara Reg	5,36	18,303	Rendah	RR	KabupatenTertinggal	Q IV	Q IV	Right
Kubu Raya Regen	6,66	24,263	Tinggi	TT	KabupatenMajudanCepatTum	Q I	Q I	Right
Pontianak Regency	7,86	31,899	Tinggi	TT	KabupatenMajudanCepatTum	Q I	Q I	Right
Singkawang Regen	6,62	25,102	TInggi	TT	KabupatenMajudanCepatTum	Q I	Q I	Right

Tabel 3 : Discrete Variable 2012

Income Percapita	Nominal	Growth	Nominal
High	=>19,969-	High	=>6,415
Medium	<19,959	Medium to high	6,04 - 6,414
		Medium	5,97 - 6,03
		Medium to Low	=<5,96

Tabel 4 :Discrete Variable 2013

Income Percapita	Nominal	Growth	Nominal
High	>=22,8	High	>6,32
Medium To High	20,05 - 22,7	Medium	<=6,32
Medium	<=20,604		

Tabel 5 :Discrete Variable 2014

Income Percapita	Nominal	Growth	
High	>=22,994	High	>5,15
Medium	<22,994	Medium	<=5,13

The accuracy of decision tree model made by weka (C4.5) compare with classen typology result on the same year is in tabel 6 and detail result for 2013 in tabel 2;

Tabel 6 :Accuracy Of Decision Tree Models

Year	Accuracy	Wrong Case
2012	92%	1 Case
2013	100%	0 Case
2014	100%	0 Case

For 2012 year we are using decision tree models below ;

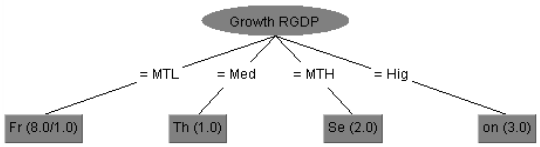

Figure 2 :2012 Decision Tree Model

For 2013 year we are using decision tree models below ;

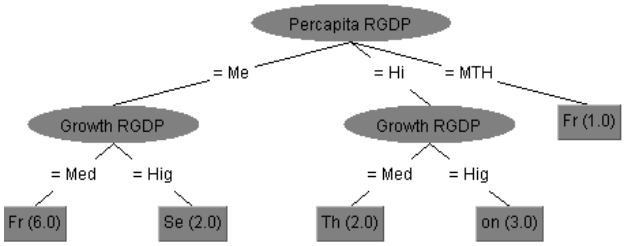

Figure 3 : 2013 Decision Tree Model

For 2014 year we are using decision tree models below ;

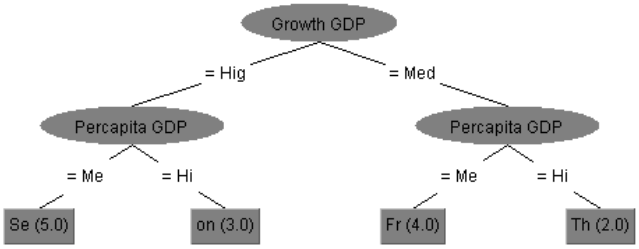

Figure 4 : 2014 Decision Tree Model

For 2012, 2013, and 2014 same year experiment using decision tree, error case only occurs in 2012 (Ketapang Regency). Reason behind that error based on our hypotesis are an economic shock/slowdown in ketapang regency.

Second reason there is a huge gap between RGDP percapita and RGDP growth of ketapang regency. This huge gap could be reason behind error in prediction made by decision tree model of 2012. RGDP Percapita Of Ketapang regency was second largest in kalimantan Barat Province but RGDP growth only place in ninth (9) position. That makes ketapang regency RGDP growth below average of all dataset and income percapita above average of all data set.

2013 and 2014 decision tree classification accuracy result is 100%. It means decision tree models could classified correctly all off data set in same year. In this research 2012 discrete attribute of RGDP growth split into fourth different class and RGDP percapita split only into two different class. Same things happen in 2013 decision tree models, RGDP growth become two class more than RGDP percapita. Only in 2014 RGDP l growth doesn't be determining factor. This is indicating how important RGDP growth in determining classification of economic regional development "Ivan O. Kitov(2006) showed that In developed countries, real GDP per capita has to grow with time along a straight line, if no large change in the specific age population is observed". If we take into account RGDP growth importance in economic analysis we could conclude that policymakers should ensure a policy to promote RGDP growth first before aiming for RGDP Percapita.

4.3. Entropy and Gain Analysis

In converting into discrete variable author find that the best way to convert it was by manually calculate it on excel until "gain" have same exact value with "entropy", Equation that used by author is normal C4.5 equation below;

$$Entropy(S) = \sum_{j=1}^{k} (-Pj \log 2\ Pj) \quad (3)$$

This research realized that when "gain" value have same exact value with "entropy" then perfect class separation achieve. So to increase accuracy of our decision tree model author did not input RGDP growth and RGDP percapita data directly to weka but by convert it first into discrete variable. To get gain value we are using this equation;

$$\text{Gain (a)} = \sum_{i=1}^{k} Entropy(S) - \left(\frac{|Si|}{|S|}\right) * \text{Entropy(Si)} \quad (4)$$

We are keep doing this equation until we get exact same gain and entropy value in year of analysis shown in tabel below. And based on that calculation we split our data into discrete variable show before;

Tabel 7 : Gain and Entropy Value Comparison

Year	Gain Value (A)	Entropy Value (S)
2012	0,543564443	0,543564443
2013	0,877277651	0,877277651
2014	1,006736529	1,006736529

Total sequence of calculation until we reach that value show in figures below ;

Tabel 8 : Calculation Sequence To Obtain Exact Gain Value

Year	Total Sequence
2012	4 Calculation Sequence
2013	2 Calculation Sequence
2014	3 Calculation Sequence

Tabel 9 : Forecasting of 2013 Regional Economic Development Classification

Regency Name	RGDP Growth	RGDP Pecapita Growth	Classen Result	Classen Result-Quadran	Forecasting Result - Quadran	Status
Sambas Regency	6,18	19,722	Disadvantage Regency	Quadran IV	Quadran II	Wrong
Bengkayang Regenc	5,90	18,793	Disadvantage Regency	Quadran IV	Quadran IV	Right
Landak Regency	5,23	15,042	Disadvantage Regency	Quadran IV	Quadran IV	Right
Mempawah Regency	5,44	15,213	Disadvantage Regency	Quadran IV	Quadran IV	Right
Sanggau Regency	5,98	23,931	Developed but Depressed Regency	Quadran III	Quadran III	Right
Ketapang Regency	4,55	27,331	Developed but Depressed Regency	Quadran III	Quadran IV	Wrong
Sintang Regency	6,47	18,472	Rapidly Growing Regency	Quadran II	Quadran I	Wrong
Kapuas Hulu Regen	5,23	20,629	Disadvantage Regency	Quadran IV	Quadran IV	Right
Sekadau Regency	6,56	16,106	Rapidly Growing Regency	Quadran II	Quadran I	Wrong
Melawi Regency	4,85	13,551	Disadvantage Regency	Quadran IV	Quadran IV	Right
Kayong Utara Rege	5,36	18,303	Disadvantage Regency	Quadran IV	Quadran IV	Right
Kubu Raya Regency	6,66	24,263	Advance and Fast Growing Regency	Quadran I	Quadran I	Right
Pontianak Regency	7,86	31,899	Advance and Fast Growing Regency	Quadran I	Quadran I	Right
Singkawang Regenc	6,62	25,102	Advance and Fast Growing Regency	Quadran I	Quadran I	Right

4.4 Forecasting Result and Variable Analysis

Finally we could arrive at main objective of this research. It is forecasting regional economic development classification based on classen typology. Result show in table below;

Tabel 10 : Forecasting Accuracy

Year	Forecasting Accuracy
2012 Decision Tree Model Forecast 2013 Regional Economic Development Based On Classen Typology	71%
2013 Decision Tree Model Forecast 2014 Regional Economic Development Based On Classen Typology	50%

Forecasting Accuracy for regional economic development classification was not precise in 2014. As stated before if economic slowdown happens it will impact decision tree forecast accuracy. But in forecasting disadvantage regency the accuracy is astonishing.

Tabel 11 : Forecasting Accouary For Disadvantages Regency

Year	Forecasting Accuracy
2012 Decision Tree Model Forecast 2013 Regional Economic Development For disadvantages region Based On Classen Typology	85%
2013 Decision Tree Model Forecast 2014 Regional Economic Development For disadvantages region Based On Classen Typology	100%

5. CONCLUSION

This paper managed to make a forecast on the classen typology with a fairly satisfying accuracy especially for the regencies/municipalities classified into advanced and fast growing regency and disadvantaged regency so it could give a recommendation for the decision makers in formulating monetary and fiscal decisions in West Kalimantan Province. Policy makers could determine the target of the RGDP growth that had to be achieved by the region in order to get a better classification result.

Result of decision tree classification accuracy in 2013 and 2014 is 100% for same year. It means decision tree model could classified correctly all of dataset in same year. And also based on our decision tree we could conclude that policymakers should ensure policy to promote economic growth first before aiming for RGDP Percapita. Important Finding in this paper that could help academic entity is to convert numerical continous attribute to discrete attribute in improving accuary of C4.5 algorithm for economical analysis.

Another important finding of this paper is that the classen typology was highly dependent on RGDP growth and RGDP-per-capita values of the province so in order to prevent any misleading in the classen typology analysis, it is recommended to make a comparison with another province, and by using a decision tree, the process can be accelerated. We could compare so many regency in more efficient step and time.

The next important finding is that one of the weaknesses of this typology is its limitation in making a classification if it did not have complete data and to make a classification, the RGDP Growth and RGDP-per-capita values had to be identified so it is relatively hard to make a forecast. However, by using a decision tree, the weaknesses could be addressed with a fairly reliable performance. This is because by using a decision tree, the lack of complete data can be dealt with.

Last finding is by using decision tree policymakers could set a target for his region to achieve a better classification result especially for regency that classify as disadvantages regency.

REFERENCES

AAII (2003). *The Top Ten Economic Indicators : What to Watch and Why,*Association for Investment Management and Research(AIMR), Charlottesville, USA.

Ai Munandar,and Edi Winarko. (2015).*Regional Development Classification Model using Decision Tree Approach,*International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 8.

Central Beurau of Statistic Kalimantan Barat Province. (2015).*Kalimantan Barat in Figures,Pontianak Kalimantan Barat , Indonesia.*

Francis X. Diebold.(1998).*The Past, Present, and Future of Macroeconomic Forecasting.*Journal of Economic Perspectives, 12, 175-192.

Gregory Mankiw (2015). *Principle of Macroeconomic,*Cengage Learning, 200 First Stamford Place , 4th floor Stamford, CT 06902 USA.

J. R Quinlan. (1996).*Improved Use Of Continuous Attribute in C4.5,*Journal of Artificial Intelligence.

Lill Andersen and Ronald Babula.(2008). *The Link BetweenOpenness and Long-Run Economic Growth,United States International Trade Commission Journal Of International Commerce and Economics.*

Raphael Bost, Raluca Ada Popa, Stephen Tu, and ShaGoldwasser. (2014).*Machine learning classfication over encrypted data, Cryptology ePrint Archive,* Report 2014/331.

Ross Quinlan, and Ron Kohavi. (1999).*Decision Tree Discovery,*University of New South Wales Sydney 2052 Australia.

Salvatore Ruggieri. (2002).*Efficient C4.5.Dipartimento di Informatica, Universita di Pisa Corso Italia 40, 56125 Pisa Italy.*

Sjafrizal,Prof.(2008)*EkonomiRegional :TeoridanAplikasi/Regional Economic : Theory and Application,* Baduose Media, Padang, Indonesia.