IJECA
International Journal of
Education & Curriculum Application
http://journal.ummat.ac.id/index.php/IJECA
ISSN 2614-3380 | Vol. 1, No. 1, April 2018

Support Vector Regression for Modeling Effect of Education Rate onLife Expectancy Rate in Indonesia

Muhammad Ghazali[1], Ita Fitriati[2], Ramdani Purnamasari[3]
[1,2,3]STKIP Taman Siswa Bima, Bima, Indonesia
muhammadghazali@gmail.com[1], itafitriati@gmail.com[2], ramdanipurnamasari@gmail.com[3]

Keywords:
Support Vector Regression; Modeling Effect; Education Rate

ABSTRACT

Life Expectancy Rate is the average number of years of life that is lived by someone who has reached a certain age. Life Expectancy is a tool to evaluate the government performance in improving the prosperity of the people. Studies on the factors that influence Life Expectancy Rate are needed to reach more accurate mathematics model to provide a better consideration for the government to determine the direction of future development policies. The data used in this study were derived from SUSENAS data with the objects of observations are all districts/cities in Indonesia in 2012. In this research, Support Vector Regression (SVR) method is used to estimate the effect of education factor which is represented by length of education by years (X) on Life Expectancy Rate (Y). Support Vector Regression (SVR) model in this research used several different kernels such as polynomial kernel, RBF and Exponential RBF (ERBF) to find the best model. The best model criterion is the model that produces the largest R2 value. The best model resulted in this research is a model that uses Exponential RBF kernel.

Article History:
Received: 23-01-2018
Revised : 17-02-2018
Accepted: 27-03-2018
Online : 30-04-2018

Crossref
https://doi.org/10.31764/ijeca.v1i1.1986

———————————— ◆ ————————————

A. INTRODUCTION

According to the definition of Indonesian Central Bureau of Statistics (BPS), Life Expectancy Rate is the average year of life that a person who has reached a certain agein the prevailing mortality situation in their community lives. Life Expectancy Rate is a tool for evaluating government performance in improving the welfare of the people, andimproving health status in particular. If life expectancy rates in a region is not good enough, the government should improve it with health development programs, and other social programs including environmental health, nutritional adequacy and poverty eradication programs.

Other paragraphs are indented this study uses data sourced from the National Socio-Economic Survey (SUSENAS) organized by Indonesian Central Bureau of Statistics (BPS). Previous studies using SUSENAS data include Damayanti and Ratnasari modeling variables affecting poverty in East Java using Geographically Weighted Regression (GWR) method, and then Anuraga and Otokusing Structural Equation Modeling-Partial Least Square to mathematically modeling the poverty in East Java[1]. Then Nur and Widjanarko used Meta-Analytic Structural Equation Modeling (MASEM) to modeling poverty in every districts in Java.

Other study in poverty data is Ghazali in 2016 using Generalized Methods of Moments (GMM) approach and Data Panel regression concludes that significant factors affecting the Poverty Depth Index in all districts / cities in Indonesia during the periodof 2008-2012 are among the old average variables school and life expectancy variable. But there is a strong correlation between the average length of school variables and life expectancy variables so that research needs to be done to find out how the relationship between these two variables.

The Support Vector Regression (SVR) method was used in this study because in several previous studies using SUSENAS data, modeling using the SVR method provided a better level of accuracy by Ghazali and Fitriati in 2016 then continued by Fitriati & Ghazali in 2018. SVR is the development of the Support Vector Machine (SVM) was first introduced by Vapnik in 1992 as a series of harmonious concepts in the field of pattern recognition. SVR is a regression-based modeling developed from SVM which is a classification-based method. The SVR model has advantages over the Ordinary Least Square (OLS) regression model in terms of implicit nonlinear model utilization through the application of kernel functions that map the vector of x feature data points to higher dimensional spaces allowing the use of models as in linearly separable cases.

B. METHOD

In this study, the method we use is support vector regression while to measure the best model we use the coefficient of determination as criterion. Support Vector Regression (SVR) probably has greatest use when the dimensionalityof the input space and the order of the approximation creates a dimensionality of a feature space representation much larger than that of the number of examples. The classification problem can be restricted to consideration of the two-class problem without loss of generality. SVMs were developed to solve the classification problem, but recently they have been extended to the domain of regression problems. Support Vector Regression (SVR) is an advanced application of the Support Vector Machine (SVM) in regression cases. SVM which was originally a classification method where the response variable was an ordinal variable while the SVR using the independent variable was a numerical variable in the form of real and continuous numbers. Suppose there is a set of data (1) with linear function (2), the optimal function of regression from the above equation is (3), where C is a predetermined value, and $(£^{-}, £^{+})$is a slack variable indicating the upper and lower bounds of the output in the system.

$$D = \{(x^1, y^1), (x^2, y^2), \dots, (x^l, y^l)\}, x \in R^n, y \in R \tag{1}$$

$$f(x) = \langle w, x \rangle + b \tag{2}$$

$$\Phi(w, £) = \underline{1}\|\mathrm{w}\|^2 + C \textstyle\sum_{\mathrm{i}}(£^{-} + £^{+}) \tag{3}$$

The factor $\|\mathrm{w}\|^2$ is called the regularization factor. Minimizing $\|\mathrm{w}\|^2$ will make thefunction as thin as possible, so that it can control the function capacity. Using the idea of the insensitive loss function introduced by Vapnik (1995) we minimize the norm fromw to get a good generalization for the regression function ƒ. The ε-insensitive loss function equation, to solve the optimization equation (3) is

$$L(y) = \{ {}_{|f} \quad 0 \quad for \ |f(x) - y| < s \ (4)$$
$$(x) - y| \quad - s \quad other$$

so the solution is as follows

$$\max_{\alpha,}{}^{*} \ \mathrm{W}(\alpha, \alpha^{*}) = \max_{\alpha,\alpha}{}^{*} \ - {}^{1}\textstyle\sum \quad \bar{} \quad \sum^{l} \quad (\alpha_{\mathrm{i}} - \alpha^{*})(\alpha_{\mathrm{j}} -$$

2 i=1 j=1 i

$\alpha^{*})(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{\mathrm{j}}) - \sum^{l}\ \alpha_{\mathrm{i}}(y_{\mathrm{i}} - \mathrm{s})\alpha^{*} - \alpha^{*}(y_{\mathrm{i}} + \mathrm{s})(5)$

j i=1 i i

with restrictions

$0 \leq \alpha_{\mathrm{i}}, \alpha^{*} \leq C,^{\mathrm{i}}\ \mathrm{i} = 1, \ldots .l,$

$\sum^{l}\quad (\alpha_{\mathrm{i}} - \alpha^{*}) = 0$

(6)

j=1 i

Complete equation (5) with boundary (6) using Lagrange multiplier then theoptimal condition of the regression function is written as follows:

$\bar{w} = \sum^{l}\quad (\alpha_{\mathrm{i}} - \alpha^{*})K(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{\mathrm{j}})$ (7)

i=1 i

$\bar{b} = -\,{}^{1}\langle \bar{w}\langle(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{r}) + K(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{s})\rangle\rangle$ (8)

If ε = 0 then we get the optimization of loss function in the form of a simpler equation as follows

$\min_{\beta}\ {}^{1}\sum \qquad \sum^{l}\quad (\beta_{\mathrm{i}})(\beta_{\mathrm{j}})K(\mathrm{x}_{\mathrm{i}},\ \mathrm{x}_{\mathrm{j}})\ - \sum^{l}\quad \beta_{\mathrm{i}}y_{\mathrm{i}}$ (9)

2 i=1 j=1 i=1

with restrictions

$$-C \leq \beta_{\mathrm{i}} \leq C, \mathrm{i} = 1, \ldots, l$$

$$\sum_{\mathrm{i}=1}^{l} \beta_{\mathrm{i}} = 0$$

where$(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{\mathrm{j}})$is the kernel function of $(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{\mathrm{j}})$

The optimal regression function of the equation (2) written as follows

$\langle \bar{w}x\rangle = \sum^{l}\qquad {}_{\mathrm{i}=1}\beta_{\mathrm{i}}(\mathrm{x}_{\mathrm{i}},\ \mathrm{x}_{\mathrm{j}})$ (10)

$\bar{b} = -\,{}^{1}\langle \bar{w}\langle(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{r}) + K(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{s})\rangle\rangle$ (11)

$\bar{2}$

In this study, optimization is assisted with kernel functions including polynomial kernels and Gaussian Radial Basis Function (RBF) and Exponential Radial Basis Function (eRBF).

The polynomial kernel function equation is

$, \qquad , \qquad d$

$(x, x\ ) = ((x, x\ ) + 1)$ (12)

The RBF kernel function equation is

$(x, x') = exp(-\frac{\|x-x'\|^2}{2d^2})$ (13)

While the equation of the Exponential RBF kernel function is

$(x, x') = exp(-\frac{\|x-x'\|}{2d^2})$ (14)

where d is the kernel degree[8][12].

To choose the best model, we used the model validation procedure using coefficient of determination (R^2) which is the percentage of influence of the predictor variable to the response variable. The equations are written as follows:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (15)$$

y_i denotes the i-th observational object and $\hat{y}$ is the prediction of the i-thdata. Then the best model is the model that has the maximum R^2 value.

C. RESULT AND DISCUSSION

The data of this study are secondary data taken from the results of the data collection of the National Socio-Economic Survey (SUSENAS) for 2012 by the Central Statistics Agency (BPS). Data collected include concerns all indicators are included in the health indicators, human resources and economics. With observation data consisting of 497 districts and cities in Indonesia, the response variable is Life Expectancy Rate (Y), while the predictor variable is length of education by years (X) in each districts and cities in Indonesia in 2012. The software used is the Matlab toolbox created by Steve R. Gunn. The data description is displayed by Table 1 as follows:

Table1. Variable Description.

Variable	**Response (Y)**		**Predictor (X)**
	Units	Year	Year
Min		60.93	2.30
	Mean68.9016		8.007
Max		75.39	12.25
Stdv		2.7265	1.56370
Corr			0.4369

The correlation between the independent variable and the response variable is equal to 0.4369 which means that there is a relationship that is directly proportional (positive sign) between the average variable of school length and life expectancy. This means that if the public education factor is getting better that characterized by the high average length of schooling, then the life expectancy in Indonesia will increase. The SVR kernels used in this study are Linear, Polynomial, Gaussian Radial Basis Function (RBF) and Exponential Gaussian Radial Basis Function (eRBF) kernels. Each kernel, Polynomial and RBF are used 3 different degrees. After obtaining the prediction from the response variable *y ̂_i* the accuracy will be compared to the response variable *(y_i)*. If the value of *y ̂_i* is closer to the value of *y_i*, the greater the accuracy level indicated by the small MSE value and the large R2 value. Summary of output from this experiment is shown by Table 2 as follows:

Table 2. Summary of the experimental output of the SVR model

Methods	Kernel	Degrees	R^2
OLS			0.1909
SVR	Polynomial	1	0
SVR	Polynomial	2	0 %
SVR	Polynomial	3	0 %
SVR	RBF	1	0.2377
SVR	RBF	2	0.2377
SVR	RBF	3	0.2377
SVR	ERBF	1	0.6890
SVR	ERBF	2	0.6890
SVR	ERBF	3	0.6890

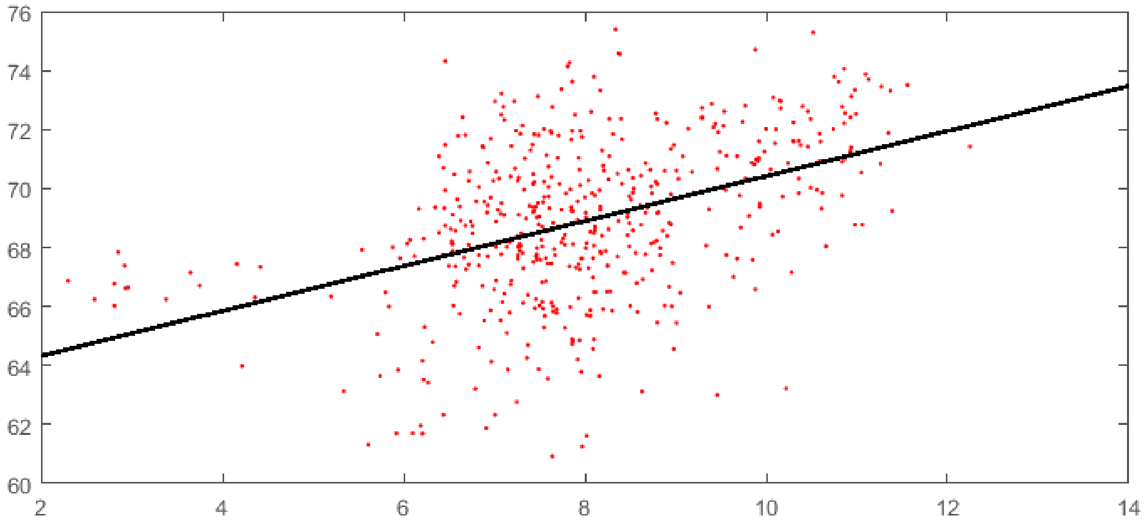

Figure 1. Plot regression of Life Expectancy (Y) vs AverageSchool Duration (X)

Regression Method (OLS) only produces R^2 coefficient of determination of 19.09% which means that the resulting model is not too good because the predictor variable is unable to influence the response variable of 19.09% while the rest is influenced by other factors. The regression plot in Figure 1 forms a random distribution pattern but the regression line shows an upward trend which means that there is a straight comparison between the average school length and life expectancy. Using the SVR method of polynomial kernel is not able to produce a more accurate model, but if using RBF kernel obtained coefficient of determination R^2 better that is equal to 23.77%. Higher coefficient of determination is obtained by using eRBF kernel that is equal to 68.90%.

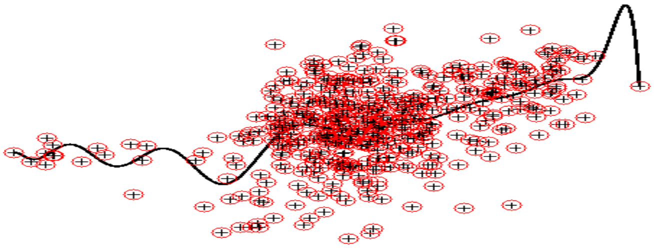

Figure 2. Plot of X vs Y using the 3rd degree RBF kernel SVR

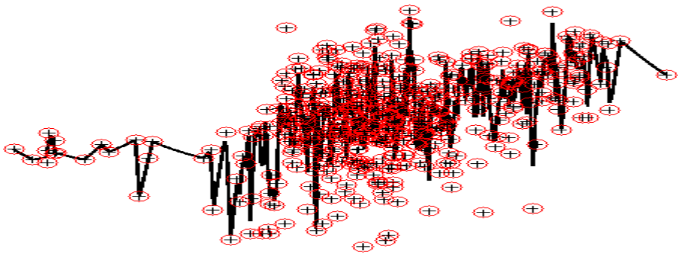

Figure 3: Plot of X vs Y using the 3rd degree eRBF kernel SVR

The best model to estimate the effect of educational factors as measured by the average length of school to the life expectancy rate in Indonesia with the Support VectorRegression (SVR) method is obtained by using the Exponential Radial Basis Function (eRBF) kernel, which is indicated by a better R^2 value than using the OLS regression method and using other SVR kernels.

D. CONCLUTION AND SUGGESTIONS

Suggestions for further research are comparing data with several different SVM kernels including the Splines kernels, B-Splines, Fouries Series, etc. There also needs to befurther research on the relationship patterns of the two variables based on quartile datato see what the range of years of education that can be said to be significant to life expectancy.

REFERENCE

G. Anuraga and B. W. Otok, *"Pemodelan Kemiskinan Di Jawa Timur Dengan Structural Equation Modeling-Partial Least Square,"* J. Stat. Univ. Muhammadiyah Semarang, vol. 1, no. 2, 2013.

Y. Damayanti, V. Ratnasari, J. Arief, and R. Hakim, *"Pemodelan Penduduk Miskin di Jawa Timur Menggunakan Metode Geographically Weighted Regression (GWR),"* J. Sains dan Seni ITS, vol. 2, no. 2, 2013.

A. Nur and B. Widjanarko, *"Meta-Analitycstructural Equation Modeling (Masem) Pada Faktor-Faktor Yang Mempengaruhi (Meta-Analityc Structural Equation Modeling (Masem) On Factors Influencing Poverty Of Java),"* Pros. Semin. Nas. Mat., no. November, pp. 51–62, 2014.

M. Ghazali, *Regresi Data Longitudinal Dengan Estimasi Generalized Method of Moments Pada Pemodelan Penduduk Miskin di Indonesia tahun 2008-2012*. Surabaya: Insititut Teknologi Sepuluh Nopember, 2016.

M. Ghazali and B. W. Otok, *"Pemodelan fixed effect pada regresi data longitudinal dengan estimasi generalized method of moments (studi kasus data pendududuk miskindi indonesia),"* J. Stat. Univ. Muhammadiyah Semarang, vol. 4, no. 1, 2016.

M. Ghazali and I. Fitriati, *Aplikasi Support Vector Regression pada Pemodelan Kemiskinan di Indonesia*. Proceeding Seminar Nasional Riset Ilmu Komputer 2016. Makassar, 2016.

I. Fitriati and M. Ghazali, *"Pemodelan Pengaruh Rata-Rata Lama Sekolah Terhadap Indeks Kedalaman Kemiskinan Di Indonesia Menggunakan Support Vector Regression,"* Pros. Semin. Nas. Ris. Kuantitatif Terap. 2017, vol. Vol. 1, No, no. April, pp. 100–105, 2017.

V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer-Verlag New York, 2000.

B. Santosa, *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*, 1st ed. Yogyakarta: Graha Ilmu, 2007.

H. Drucker et al., *"Support Vector Regression Machines,"* Bell Labs Monmouth Univ., vol. 1.

A. J. Smola and B. Scholkopf, *"On a kernel–based method for pattern recognition, regression, approximation and operator inversion.,"* Algorithmica, vol. Technical, no. Technical Report 1064, p. GMD FIRST, 1997.

S. R. Gunn, *"Support Vector Machines for classification and regression,"* ISIS Tech. Rep., p. 14.1: 5-16., 1998.

R. E. Walpole, *Introduction to Statistics*, 2nd ed. Macmillan, 1974.