Implementing GCV and mGCV to Determine Optimal Knot in Spline Regression for East Java Life Expectancy

Amanda Ayu Dewi Lestari[1], Aviolla Terza Damaliana[2], Dwi Arman Prasetya[3]
[1,2,3]Faculty of Computer Science, Pembangunan National Veteran University of East Java, Indonesia

Article Info

Article history:

Received May 10, 2025
Revised Jun 02, 2025
Accepted Jun 29, 2025

Keywords:

Generalized Cross Validation
Life Expectancy
Nonparametric Spline Regression
Modified Generalized Cross Validation

ABSTRACT

Life Expectancy is a vital indicator for evaluating population's overall welfare and health status within a specific region. According to data published by Badan Pusat Statistik (BPS) National, East Java Province ranks 10th nationally in terms of life expectancy in 2024, with male life expectancy recorded at 70.39 years and female life expectancy at 74.4 years. This research focuses on examining four key factors that are believed to influence life expectancy in East Java during 2024 including the Percentage of the Poor Population (X_1), the Percentage of Individuals Aged 5 and Above Who Regularly Smoke Tobacco (X_2), the Expected Years of Schooling (X_3), and the Open Unemployment Rate (X_4). To determine the optimal knot points in the nonparametric truncated spline regression model, the study utilizes Generalized Cross-Validation (GCV) and the modified Generalized Cross-Validation (mGCV) techniques by minimizing their respective error values. The findings indicate that all four variables significantly impact life expectancy. Among the methods applied, the mGCV approach demonstrates good performance, achieving the lowest error value of 0.100 and a coefficient of determination of 82.91%.

Corresponding Author:

Aviolla Terza Damaliana,
Faculty of Computer Science,
Pembangunan National Veteran University of East Java,
Jalan Raya Rungkut Madya, No 1, Surabaya, Jawa Timur, Indonesia
Email: aviolla.terza.sada@upnjatim.ac.id

1. INTRODUCTION

Human development is an important aspect in realizing social welfare. Development is a continuous change process actively pursued to improve the population's welfare [1]. Life expectancy is a key indicator that reflects the quality of life within a community and is used to assess the effectiveness of development policies through existing programs. Life expectancy at birth, commonly known as life expectancy, is a demographic indicator that signifies the average number of years an individual is projected to live, starting from birth, based on the mortality rates prevailing in a particular year [2].

According to data released by the Badan Pusat Statistik (BPS) in 2024, Life Expectancy in Indonesia exhibits considerable variation across provinces. East Java Province ranks 10th nationally, with a life expectancy of 70.39 years for males and 74.4 years for females. These figures indicate that females in East Java have a higher life expectancy compared to males a pattern that is relatively common in many other regions as well [3].

One significant factor influencing life expectancy in a region is the proportion of the population living below the average economic standard. In East Java Province, five regencies recorded the highest poverty rates between 2018 and 2023. In 2023, Sampang Regency ranked first, with a poverty rate of 21.76 percent [4]. High poverty levels are often associated with limited access to healthcare services, education, and other basic needs, ultimately impact life expectancy. Also Indonesia, a significant proportion of the impoverished population resides in rural areas and depends primarily on subsistence-based, low-productivity agricultural activities for their livelihood [5]. In addition to poverty, education is also a crucial indicator in evaluating the quality of life in East Java Province. Indonesia's current education system remains far from meeting the targets set by the Sustainable Development Goals (SDGs), based on existing conditions [6]. Considering that Indonesia has the highest proportion of illiterate individuals compared to other countries, the quality of education in the country remains significantly lower than that of other Southeast Asian nations [7]. In Indonesia, 15% of adolescents under the age of 15 are illiterate, whereas other countries report youth illiteracy rates of less than 10% [8]. Several educational policies face similar challenges, including inadequate infrastructure, unequal distribution of resources, and resistance to change among teaching personnel [9].

Health is another crucial determinant of life expectancy, as maintaining optimal health enables individuals to live longer and lead more productive lives [10]. Data from the East Java Provincial Statistics Agency indicate a consistently high prevalence of smoking among adolescents, with smoking rates recorded at 28.72% in 2021, 28.83% in 2022, and 28.51% in 2023 [11]. This high prevalence poses a significant public health concern, as smoking is associated with increased risks of severe health conditions, including lung cancer, cardiovascular diseases, and respiratory disorders, all of which contribute to reduced life expectancy [12]. Previous studies employing multiple linear regression have identified several factors such as healthy lifestyle practices, nutritional adequacy, and educational attainment as significantly impacting life expectancy [13]. In addition, a study employing multiple regression analysis to predict life expectancy trends from 2021 to 2030 revealed that factors such as the number of educational institutions, the unemployment rate, and GDP per capita did not exhibit statistically significant effects according to the F-test. However, GDP per capita was identified as having a notably positive impact on the average life expectancy in Central Java [14].

Accordingly, this research develops a life expectancy model for the population across the regencies and cities of East Java in 2024 by applying nonparametric truncated spline regression. This method is chosen because the relationship pattern between life expectancy and the predictor variables is assumed to be non-linear and random, reflecting differences in demographic and socio-economic characteristics across regions. In Truncated Spline Regression, the modeling process is carried out by dividing the regression curve based on specific knot points, which capture changes in data patterns more flexibly compared to parametric regression methods [15].

This study utilizes two methods to determine the optimal knot points: Generalized Cross-Validation (GCV) and modified Generalized Cross-Validation (mGCV). These methods are selected due to their ability to balance model complexity and prediction error. GCV is used because it can estimate optimal knot points by considering the trade-off between bias and variance, while mGCV provides adjustments for smaller sample sizes, thus offering greater stability in determining the optimal number of knot points [16]. Through this method, the model is expected to yield high accuracy and more accurately describe the relationship pattern between life expectancy and its influencing factors. It is hoped that the findings of this study can contribute to the planning of development policies focused on improving community welfare and quality of life in the region.

2. RESEARCH METHOD

The research follows a series of key stages, beginning with descriptive statistical analysis and proceeding to the interpretation phase. The steps of this analysis are illustrated in the flowchart presented in Figure 1.

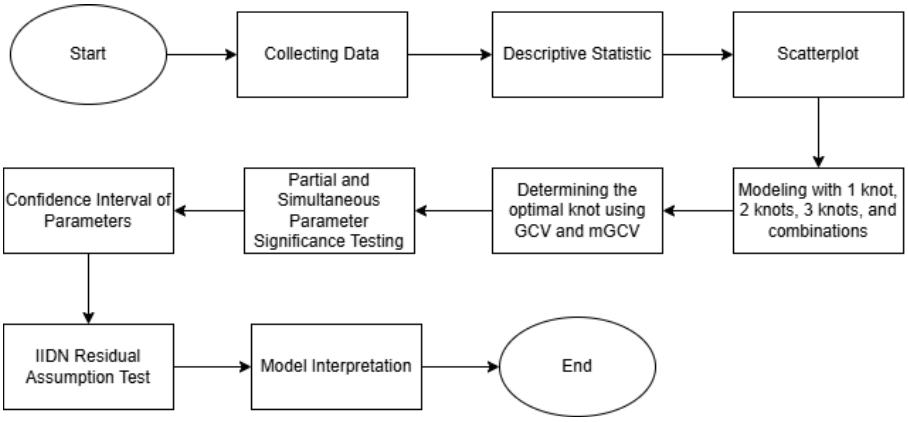
Start
Collecting Data
Descriptive Statistic
Scatterplot
Modeling with 1 knot, 2 knots, 3 knots, and combinations
Determining the optimal knot using GCV and mGCV
Partial and Simultaneous Parameter Significance Testing
Confidence Interval of Parameters
IIDN Residual Assumption Test
Model Interpretation
End

Figure 1. Research Flow Diagram

2.1. Dataset

This study relies on secondary data from the Badan Pusat Statistik of East Java Province (BPS) for the year 2024. The units of analysis are the districts and cities within East Java Province. The study involves five main variables: Life Expectancy (Y), the Percentage of Poor Population (X_1), the Percentage of Population Aged 5 Years and Over Who Regularly Smoke Tobacco (X_2), the Expected Years of Schooling (X_3), and the Open Unemployment Rate (X_4).

2.2. Regresi Nonparametric Spline Truncated

Among the various nonparametric regression techniques, truncated spline regression is widely utilized due to its flexibility. Its key characteristic lies in the piecewise nature of the spline function, allowing it to effectively model shifts in data trends across different intervals [17]. In truncated spline nonparametric regression, knot points are used to aid in the regression analysis [18]. The nonparametric regression model is generally expressed through the following equation:

$$y_i = f(x_i) + \varepsilon_i \quad ; i = 1,2,\dots,n$$

A spline function of degree p with specified knot points K_1, K_2, K_3,…,K_i is used to approximate the regression curve $f(x_i)$ which can be expressed through the following equation.

$$f(x_i) = \sum_{j=0}^{p} \beta_o x_i^j + \sum_{j=1}^{r} \beta_{p+j}(x_i - K_j)_+^p$$

Inserting equation (1) into equation (2) yields the following expression for the nonparametric spline regression model.

$$y_i = \sum_{j=0}^{p} \beta_o x_i^j + \sum_{j=1}^{r} \beta_{p+j}(x_i - K_j)_+^p + \varepsilon_i \quad ; i = 1,2,\dots,n$$

The function $(f_i - K_j)_+^p$ is a truncated function defined as follows:

$$(x_i - K_j)_+^p = \begin{cases} (x_i - K_j)^p, & x_i \geq K_j \\ 0 \quad , & x_i \leq K_j \end{cases}$$

In this model, p denotes the spline order, and K represents the number of knots used in the model. The parameters of the truncated spline nonparametric regression are estimated using the Ordinary Least Squares (OLS) method, as expressed below [19].

$$\widehat{\beta_{ols}} = (X^T X)^{-1} X^T y$$

2.3. Determining the Optimal Knot Locations

In the process of determining the optimal knot points, specific criteria are used to balance the model's bias and variance. The Generalized Cross-Validation (GCV) method is employed to measure prediction error while accounting for model complexity. Additionally, the Modified Generalized Cross-Validation (mGCV) can be used as an alternative to address the limitations of GCV under certain conditions.

2.3.1. Generalized Cross Validation

The effectiveness of the truncated spline regression model is maximized when optimal knot points are used, as these locations correspond to changes in the underlying functional structure. A commonly employed technique for determining these optimal knots is Generalized Cross-Validation (GCV), valued for its asymptotic optimality [20]. To obtain the best-fitting model, knot selection is guided by minimizing the GCV criterion.

$$GCV = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{(n^{-1}Tr(I-G))^2}$$

In this case, I denotes the identity matrix, n indicates the total number of observations, and $K = (K_1, K_2, K_3, \dots, K_r)$ refers to the knot points.

2.3.2. Modified Generalized Cross Validation

The Modified Generalized Cross-Validation (mGCV) method enchances of the standard GCV approach, designed to improve model selection stability in nonparametric regression. It introduces a correction factor to account for the model's degrees of freedom, helping to minimize the bias that may occur in conventional GCV calculations [21]. The formulation of the mGCV method is given as follows:

$$mGCV = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{(n^{-1}\rho.Tr(I-G))^2}$$

Where $0 < \rho < 1$

The parameter ρ serves as a correction factor aimed at adjusting the penalty for model complexity.

2.4. Parameter Significance Testing

Parameter testing serves as a crucial step in assessing the significance and suitability parameters within a statistical or regression model.

2.4.1. Simultaneous Test

To evaluate the overall significance of the model, the following hypotheses are formulated [22]:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: at least one coefficient $\beta_p \neq 0$ for $p = 1,2, \dots, n$

The rejection region is defined when $F_{hitung} > F_{a,(n-p-1)}$ or reject H_0 if $p - value < a$, indicating the rejection of the null hypothesis. This indicates that at least one of the predictor variables significantly influences the response variable, or that at least one parameter is statistically distinct from zero.

2.4.2. Parsial Test

The following hypothesis is used to test the model partially:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$, j = 1, 2, …, p + r

Where $se(\widehat{\beta_j})$ represents the standard error of $\widehat{\beta_j}$, and the rejection criterion for the t-test is to reject H_0 jika $|t_{hitung}| > t_{\frac{a}{2},(n-p-1)}$ or reject H_0 jika $p - value < a$. The analysis results confirm that the predictor variable has a significant impact on the response variable [22].

2.5. Confidence Interval

A confidence interval (CI) defines a range of plausible values for an unknown population parameter, derived from sample data [23].

$$CI = Sample\ mean\ \pm z\ value\ \times Standart\ error\ of\ mean\ (SEM)$$

The confidence interval can be calculated at a 90% or 99% confidence level, with the critical value or z-value depending on the level of confidence.

2.6. Residual Assumption Test

Residual assumptions or model fit testing is an important stage in the regression analysis process aimed at validating that the mathematical model developed appropriately represents the data pattern being studied.

2.6.1. Identical Assumption Test

The Glejser Test is a statistical method used to detect the presence of heteroscedasticity in a regression model. It involves regressing the absolute values of the residuals from the original model on one or more explanatory variables [24].

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$

H_1: at least one $\sigma_i^2 \neq \sigma^2$; i = 1, 2, …, n

The null hypothesis is rejected if $F_{hitung} > F_{a;(p-1,n-p-1)}$ or $p - value < a$ indicating evidence of heteroscedasticity. If $F_{hitung} < F_{a;(p-1,n-p-1)}$ or $p - \ value < a$ then H_0 fails to be rejected, meaning the null hypothesis is retained, indicating that the data meets the assumption of homoscedasticity.

2.6.2. Independent Assumption Test

The independence assumption states that the errors in a regression or ANOVA model are free from any specific pattern or dependency among them. In other words, the residuals should not exhibit autocorrelation [25]. The Durbin-Watson test is commonly used to assess whether this assumption is met, with the evaluation based on the following criteria.

Table 1. Durbin-Watson Test Criteria

H$_0$	Decision	If
No positive autocorrelation is found	H$_0$ is rejected	$0 < d < d_L$
No positive autocorrelation is found	No decision	$d_L \leq d \ \leq d_U$
No negative autocorrelation is found	H$_0$ is rejected	$4 - d_L < d < 4$
Testing is inconclusive	No decision	$4 - d_U \leq d \leq 4 - d_L$
No positive or negative autocorrelation is found	H$_0$ is accepted	$d_U < d < 4 - d_L$

2.6.3. Normality Test

One commonly used method to assess data normality is the Kolmogorov-Smirnov Test. This method operates by examining how the distribution of the research data aligns with the properties of a standard normal distribution [26]. When implementing this test, the hypotheses can be structured as follows:

$H_0: F_n(\varepsilon) = F_0(\varepsilon)$, the residuals is normal distribution

$H_0: F_n(\varepsilon) \neq F_0(\varepsilon)$, the residuals is not normal distribution

The null hypothesis H$_0$ is rejected if the calculated significance value surpasses the critical value q$_{(1-a)}$ obtained from the Kolmogorov-Smirnov distribution table. Alternatively, rejection occurs when the p-value associated with the test statistic falls below the predetermined significance level α\alphaα. This indicates that the observed data significantly deviate from the expected distribution, providing statistical evidence against the assumption of normality.

2.7. Coefficient of Determination

The coefficient of determination, commonly denoted as R^2, quantifies the extent to which the variation in the dependent variable can be accounted for by the independent variables in a regression model.

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

A value of 0 indicates that the model fails to explain any of the variability in the response variable, while a value of 1 signifies a perfect fit, meaning all observed variations are fully accounted for by the predictors [27].

3. RESULTS AND DISCUSSION

3.1. Characteristics of Life Expectancy

This study employs data from 38 observational units, encompassing all regencies and cities within East Java Province. Each observational unit corresponds to an administrative region at the regency or city level and is analyzed concerning the response variable, namely life expectancy. The predictor variables consist of various social, economic, and environmental indicators hypothesized to influence the life expectancy levels across the respective regions. Below are the observations or descriptions of all 38 districts/regencies.

Table 2. Number of Observation

No	City/Regency	Life Expectency (Y)	Poor Population (X_1)	Individuals Aged 5 and Above Who Smoke Tobbaco (X_2)	Expected Years of Schooling (X_3)	Open Unemployment Rate (X_4)
1	Kab Pacitan	74,74	13,08	22,45	12,69	1,56
2	Kab Ponorogo	75,28	9,11	25,47	13,78	4,19
3	Kab Trenggalek	75,35	10,5	22,99	12,63	3,9
⋮	⋮	⋮	⋮	⋮	⋮	⋮
37	Kota Surabaya	76,02	3,96	17,24	14,87	4,91
38	Kota Batu	75,36	3,06	23,23	14,58	3,63

All variables used in this study were examined for the presence of missing values. The examination was conducted using descriptive and exploratory methods. No missing values were found in the main variables, including life expectancy and the predictor variables (X_1 to X_4); therefore, no imputation or case exclusion was necessary. Data cleaning was performed prior to analysis, including range checks, detection of duplicate entries, and standardization of district/city names to ensure consistency across variables. The following presents the data characteristics of four factors that are presumed to influence life expectancy in East Java Province.

Table 3. Descriptive Statistics

Variable	Mean	Varians	Min	Max
Y	74,8	5,52	73,31	76,02
X1	9,78	17,75	3,06	20,83
X2	22,84	8,08	17,24	27,86
X3	13,58	0,81	11,98	15,79
X4	4,04	1,38	1,56	6,49

Referring to Table 3, Surabaya City recorded the highest Life Expectancy in 2024, with a value of 76.02 years. Meanwhile, Bondowoso Regency was noted as the region with the lowest life expectancy, at 73.31 years. The average LE for all regencies/cities in East Java in 2024 reached 74.8 years, with a variance of 5.52. Regarding the poverty factor (X1), Sampang Regency ranked the highest, while Batu City recorded the lowest. For the percentage of the population aged 5 and over who habitually smoke tobacco (X2), Sumenep Regency had the highest rate at 27.86%, while Surabaya City had the lowest at 17.24%.

Regarding education, Malang City recorded the highest expected years of schooling (X3), while Bangkalan Regency was at the lowest position. As for the open unemployment rate (X4), Pacitan Regency had the highest rate, while Sidoarjo Regency reported the lowest. The Madura region appears dominant in several unfavorable social indicators, such as high poverty levels, a high prevalence of smoking habits, and low expected years of schooling. This can be attributed to several factors. The high poverty rate in Madura is driven by limited economic resources and low access to employment in the formal sector. The majority of the population relies on agriculture and small-scale trade, which often result in inconsistent income.

3.2. The Relationship Between Life Expectancy Data Patterns and the Suspected Influencing Variables

Regression analysis typically begins with the construction of a scatterplot to explore the nature of the relationship between predictor variables and the response variable, Life Expectancy. This visual representation helps in identifying potential patterns or trends, as illustrated in Figure 2.

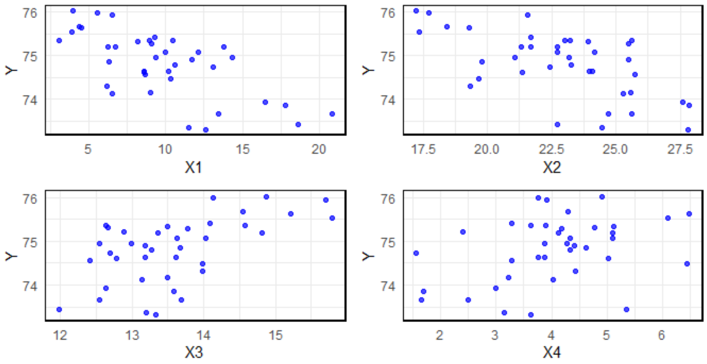

Figure 2. Scatterplot of predictor variables against response variables

Figure 2 illustrates that the relationship between Life Expectancy and each of predictor variables deviates from a linear pattern. However, based on previous studies, these variables have been found to be associated with life expectancy. Therefore, all variables will be treated as components in a nonparametric regression, and the analysis will be conducted using the truncated spline nonparametric regression method.

3.3. Life Expectancy Modeling

This research aims to develop a model for Life Expectancy in East Java utilizing nonparametric regression with linear truncated splines. The modeling process incorporates various spline configurations, including models with one, two, and three knots, as well as combinations of these knot placements to explore their effectiveness.

3.3.1. Estimation Using the GCV Method

This study using the Generalized Cross Validation (GCV) method with varying numbers of knots specifically one, two, three, and combinations thereof. The purpose of this variation is to assess the impact of different knot quantities on the accuracy and performance of the model's estimation outcomes. The estimated nonparametric truncated spline regression model with a single knot point is given below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(x_{i1} - K_{11}) + \beta_3 x_{i2} + \beta_4(x_{i2} - K_{21}) + \beta_5 x_{i3} + \beta_6(x_{i3} - K_{31}) + \beta_7 x_{i4} + \beta_8(x_{i4} - K_{41}) + \varepsilon_i$$

In the process of determining the optimal knot point for the model with a single knot, the GCV (Generalized Cross Validation) values were obtained as described below.

Table 4. Minimum GCV for One Knot

GCV	Knots			
	X_1	X_2	X_3	X_4
0.3640	12.58	22.93	14.02	4.20
0.3647	12.85	23.09	14.07	4.27
0.3665	13.21	23.30	14.15	4.37
0.3670	13.57	23.52	14.23	4.47

The minimum GCV value for a single knot point from the iteration process is 0.3639. Following the identification of the optimal knot in the single-knot scenario, as presented in Table 4, the analysis is extended by incorporating two knots into the model. The estimated model with four predictor variables for the two-knot case is as follows.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(x_{i1} - K_{11}) + \beta_3(x_{i1} - K_{12}) + \beta_4 x_{i2} + \beta_5(x_{i2} - K_{21}) + \beta_6(x_{i2} - K_{22}) + \beta_7 x_{i3} + \beta_8(x_{i3} - K_{31}) + \beta_9(x_{i3} - K_{32}) + \beta_{10} x_{i4} + \beta_{11}(x_{i4} - K_{41}) + \beta_{12}(x_{i4} - K_{42}) + \varepsilon_i$$

In the stage of determining the optimal knot points for the model with two knots, the GCV values were obtained as described below.

Table 5. Minimum GCV for Two Knot

GCV	Knots			
	X_1	X_2	X_3	X_4
0.3324	3.06	17.24	11.98	1.56
	20.83	27.86	15.79	6.49
0.3985	3.06	17.24	11.98	1.56
	20.467	27.643	15.712	6.389
0.4434	3.422	17.456	12.057	1.66
	3.785	17.673	12.13	1.76
0.4667	3.422	17.456	12.057	1.66
	4.147	17.890	12.21	1.86

Table 5 shows that the lowest GCV value obtained for the two-knot configuration through the iterative process is 0.3324. Once these optimal knot locations are identified, the analysis advances to the three-knot model for further evaluation. The estimated with four predictor variables for the three-knot case is as follows.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2(x_{i1} - K_{11}) + \beta_3(x_{i1} - K_{12}) + \beta_4(x_{i1} - K_{13}) + \beta_5 x_{i2} + \beta_6(x_{i2} - K_{21}) + \beta_7 x_{i2}(x_{i2} - K_{22}) + \beta_8(x_{i2} - K_{23}) + \beta_9 x_{i3} + \beta_{10}(x_{i3} - K_{31}) + \beta_{11}(x_{i3} - K_{32}) + \beta_{12}(x_{i3} - K_{33}) + \beta_{13} x_{i4} + \beta_{14}(x_{i4} - K_{41}) + \beta_{15}(x_{i4} - K_{42}) + \beta_{16}(x_{i4} - K_{43})$$

In the stage of determining the optimal knot points for the model with three knots, the GCV values were obtained as described below.

Table 6. Minimum GCV for Three Knot

GCV	Knots			
	X_1	X_2	X_3	X_4
0.3007	5,961	18.973	12.602	2.364
	6,686	19.407	12,757	2,566
	7,774	20.057	12.990	2.867
0.3485	6.323	19.190	12.679	2.465
	6.686	19.407	12.757	2.566
	9.950	21.357	13.457	3.471
0.3461	5,598	18.757	12.524	2.264
	6.686	19.407	12.757	2.566
	7.774	20.057	12.990	2.867
0.3907	4.510	18.106	12.291	1.962
	9.225	20.924	13.301	3.270
	10,675	21.791	13.612	3.672

As presented in Table 6, the iterative process yields a minimum GCV value of 0.3007. The analysis then proceeds by evaluating various knot combinations to identify the configuration that further minimizes the GCV, as each knot arrangement can produce distinct modeling results. Below are the GCV values corresponding to the truncated spline nonparametric regression models with multiple knot configurations.

Table 7. Minimum GCV for Combination Knot

GCV	Knots			
	X_1	X_2	X_3	X_4
0.3007	5.961	18.973	12.602	2.364
	6.686	19.407	12,757	2,566
	7.774	20.057	12.990	2.867
0.3681	12.489	18.973	14.001	1.560
		19.407		6.490
		20.057		
0.3792	5.961	18.973	14.001	1.560
	6.686	19.407		6.490
	7.774	20.057		
0.3847	5.961	18.973	11.980	4.175
	6.686	19.407	15.790	
	7.774	20.057		

Table 7 indicates that the lowest GCV value among the knot combinations is achieved with the configuration (3,3,3,3), yielding a value of 0.3007. After obtaining the optimal knot results, a comparison is made as the basis for selecting the best model.

Table 8. Minimum GCV of Optimal Knot

Jumlah Knot	GCV minimum	R-Square	MSE
1	0.3640	60.553	0.236
2	0.3324	53.354	0.232
3	0.3007	82.914	0.205
Kombinasi (3,3,3,3)	0.3007	82.914	0.205

Referring to Table 8, among the models employing one, two, three, and combined knot configurations, the most optimal model is identified by the lowest GCV value. This occurs when each variable is assigned three knots, resulting in a GCV of 0.3007 and an MSE of 0.205. These values suggest that the model effectively captures the nonlinear relationship between the predictor variables and the response, demonstrating a high level of estimation accuracy.

3.3.2. Estimation Using the mGCV Method

The optimal knot point is determined based on the minimum estimated value of the modified Generalized Cross Validation (mGCV) criterion. Below is the nonparametric truncated spline regression model estimated using a single knot and four predictor variables. The corresponding mGCV values for this one-knot configuration are summarized in Table 9:

Table 9. Minimum mGCV for One Knot

ρ	Minimum mGCV	ρ	Minimum mGCV
0.1	0.222	0.6	0.287
0.2	0.233	0.7	0.304
0.3	0.245	0.8	0.322
0.4	0.258	0.9	0.342
0.5	0.272		

As presented in Table 9, the lowest mGCV value is obtained when ρ=0.1, yielding a value of 0.222. The specific positions of the corresponding optimal knot points are detailed below:

Table 10. Knots of Minimum mGCV for One Knot

X_1	X_2	X_3	X_4
12.488	22.875	14.001	4.175

Following the identification of the optimal single-knot configuration, the analysis proceeds to select the most suitable two-knot points. This stage involves evaluating various two-knot combinations, with their corresponding (mGCV) values summarized in Table 11:

Table 11. Minimum mGCV for Two Knot

ρ	Minimum mGCV	ρ	Minimum mGCV
0.1	0.180	0.6	0.266
0.2	0.193	0.7	0.290
0.3	0.208	0.8	0.313
0.4	0.225	0.9	0.322
0.5	0.244		

According to Table 11, the optimal modified Generalized Cross Validation (mGCV) value is achieved when ρ=0.1 resulting in the lowest observed value of 0.222. The precise locations of the corresponding knot points associated with this optimal result are detailed below.:

Table 12. Knots of Minimum mGCV for Two Knot

X_1	X_2	X_3	X_4
9.225	20.924	13.301	3.270
11.038	22.008	13.690	3.773

Upon identifying the optimal configuration with two knots, the analysis advances to selecting the most effective three-knot arrangement. This phase involves assessing multiple three-knot configurations, with their corresponding modified Generalized Cross Validation (mGCV) values presented in Table 13:

Table 13. Minimum mGCV for Three Knot

ρ	Minimum mGCV	ρ	Minimum mGCV
0.1	0.100	0.6	0.171
0.2	0.110	0.7	0.194
0.3	0.122	0.8	0.222
0.4	0.136	0.9	0.257
0.5	0.152		

Table 13 indicates that the minimum mGCV value occurs at ρ = 0.1, with a value of 0.100, along with the detailed knot positions as follows:

Table 14. Knots of Minimum mGCV for Three Knot

X_1	X_2	X_3	X_4
5.961	18.973	12.602	2.364
6.686	19.407	12.757	2.566
7.774	20.057	12.990	2.867

After establishing the optimal knot configuration, a comparative analysis is conducted to assess and contrast the performance of the models derived from various knot arrangements. This evaluation aims to identify the model that offers the most accurate and reliable representation of the underlying data patterns.

Table 15. Minimum mGCV of Optimal Knot

Jumlah Knot	mGCV minimum	R-Square	MSE
1	0.222	60.486	0.232
2	0.180	68.705	0.232
3	0.100	82.914	0.205
Kombinasi (3,3,3,3)	0.100	82.914	0.205

Table 15 presents a comparative analysis of models incorporating one, two, and three knots. Among these, the model employing three knots for each predictor variable emerged as the most optimal, as indicated by the minimum mGCV value of 0.100. This model also demonstrated strong explanatory power with an R-squared of 82.914 and exhibited the best fit through the lowest Mean Squared Error (MSE) of 0.205.

3.3.3. Determining the Optimal Knot Locations

Table 16 presents a comparative analysis between the GCV method and the mgcv package-based approach. This comparison aims to evaluate the consistency and performance of both methods in selecting optimal knot configurations and fitting the regression model.

Table 16. Knot Selection Performance

Method	Minimum Value	R-Square	MSE
GCV	0.3007	82.914	0.205
mGCV	0.100	82.914	0.205

Based on the comparison between GCV and mGCV methods in modeling life expectancy using a nonparametric regression spline truncated, it can be concluded that the mGCV method yields a more optimal result. This is evidenced by the lower minimum value produced by mGCV (0.100) compared to GCV (0.3007). Although both methods produce the same R-square and Mean Squared Error (MSE) values, namely 82.914 and 0.205 respectively, the advantage of mGCV lies in its ability to select more appropriate knot points, resulting in a model that better fits the data and minimizes the risk of overfitting. While the iteration time for mGCV is slightly longer (27.655 seconds) than that of GCV (25.7575 seconds), the difference is still acceptable and proportional to the improvement in model quality. Therefore, the mGCV method is selected as the best model for this study.

3.4. Model Parameter Estimation using the OLS Method

Following the selection of optimal knot points based on the criterion of the lowest mGCV value, the subsequent stage involves parameter estimation using the Ordinary Least Squares (OLS) method. The following sections provide explanation of the developed nonparametric spline regression model, which utilizes three knots for each of the four predictor variables.

$$\hat{y} = 26{,}595 + 0{,}558x_{i1} - 4{,}02(x_{i1} - 5{,}961) + 4{,}408(x_{i1} - 6{,}686) - 1{,}037(x_{i1} - 7{,}774) + 0.267x_{i2} - 4{,}197(x_{i2} - 18{,}973) + 7{,}692(x_{i2} - 19{,}407) - 3{,}887(x_{i2} - 20{,}057) + 3{,}027x_{i3} - 12{,}454(x_{i3} - 12{,}602) + 11{,}106(x_{i3} - 12{,}757) - 1{,}215(x_{i3} - 12{,}990) + 2{,}336x_{i4} - 19{,}207(x_{i4} - 2{,}364) + 21{,}517(x_{i4} - 2{,}566) - 4{,}363(x_{i4} - 2{,}867)$$

3.5. Significance Testing of Parameters in the Nonparametric Spline Regression Model

3.5.1. Simultaneous Test

This simultaneous testing is conducted to evaluate the model's parameter estimates collectively. The significance level (alpha) used is 0.05. The results of the simultaneous test are presented in Table 17.

Table 17. ANOVA Parameter Test

Source of Variation	df	*Sum of Square (SS)*	*Mean Square*
Regresi	16	16.936	1.058
Eror	21	3.49	0.166
Total	37	20.426	

Based on Table 16, the $F_{\text{calculated}}$ value is 6.369, and the $F_{\text{table}(0.05,16,21)}$. Therefore, the decision is to reject H_0, meaning that at least one parameter has a significant effect on Life Expectancy. From this conclusion, individual or partial tests can be conducted further.

3.5.2. Partial Test

The t-test is employed to perform partial or individual significance testing. Table 18 presents the outcomes of the individual parameter evaluations derived from the spline regression model.

Table 18. Partial Test Result

Variable	Parameter	Koefisien	*p-value*	Decision
	β_0	26.595	2.196	Not Sig
X_1	β_1	0.558	0.0407	Significant
	β_2	-4.023	0.012	Significant
	β_3	4.408	0.010	Significant
	β_4	-1.037	0.014	Significant
X_2	β_5	0.267	0.552	Not Sig
	β_6	-4.197	0.075	Not Sig
	β_7	7.692	0.004	Significant
	β_8	-3.887	0.0003	Significant
X_3	β_9	3.027	0.0058	Significant
	β_{10}	-12.454	0.0107	Significant
	β_{11}	11.106	0.065	Not Sig
	β_{12}	-1.215	0.552	Not Sig
X_4	β_{13}	2.336	0.014	Significant
	β_{14}	-19.207	0.018	Significant
	β_{15}	21.517	0.039	Significant
	β_{16}	-4.363	0.189	Not Sig

Based on Table 18, there are several parameters that are not significant with respect to the variables, including β_0, β_5, β_6, β_{11}, β_{12}, and β_{16}. In contrast, the remaining 10 parameters are significant, as indicated by their p-values being less than 0.05. Although some parameters do not show a significant effect on life expectancy, the significance of the other 10 parameters suggests that the four predictor variables still have an overall influence on the life expectancy variable.

3.6. Confidence Interval Parameter

In this study, the confidence interval is used to provide a range of estimates for the parameters. Below are the confidence intervals for each parameter:

Table 19. Confidence Interval Parameter

Variable	Parameter	Koefisien	Lower Limit	Upper Limit
	β_0	26.595	-4.883	58.074
X1	β_1	0.558	0.033	1.082
	β_2	-4.023	-7.061	-0.984
	β_3	4.408	1.180	7.635
	β_4	-1.037	-1.834	-0.240
X2	β_5	0.267	-0.639	1.173
	β_6	-4.197	-8.792	0.396
	β_7	7.692	2.808	12.576
	β_8	-3.887	-5.750	-2.024
X3	β_9	3.027	1.002	5.052
	β_{10}	-12.454	-21.571	-3.337
	β_{11}	11.106	-0.628	22.842
	β_{12}	-1.215	-5.344	2.913
X4	β_{13}	2.336	0.532	4.139
	β_{14}	-19.207	-34.672	-3.742
	β_{15}	21.517	1.435	41.598
	β_{16}	-4.363	-10.964	2.236

Table 19 shows the regression coefficients along with the confidence interval (CI) for each parameter. Parameters whose confidence intervals do not cross zero (either entirely positive or entirely negative) are considered statistically significant. There are 6 parameters whose confidence intervals cross zero, including β_0, β_5, β_6, β_{11}, β_{12}, dan β_{16}.

3.7. Residual Asumption Test

3.7.1. Identical Assumption Test

An identical variance test is carried out to confirm that the residuals exhibit constant variance, a condition known as homoscedasticity. This assumption implies that the distribution of error terms remains stable across all levels of the predicted values. The outcome of this test, performed using the Glejser method, is summarized in Table 20.

Table 20. Identical Assumption Test Result

Source of Variation	df	Sum of Square (SS)	Mean Square	$F_{Calculated}$
Regresi	16	0.929	0.058	1.683
Eror	21	0.725	0034	
Total	37	1.655		

Based on the results in Table 20, the F-statistic value is 1.683 with a p-value of 0.130. Since the *p-value* $> \alpha$ with $\alpha = 0.05$, the decision fails to reject H_0, indicating that there is no heteroscedasticity. This result indicates that the variance of the residuals is constant across all levels of the independent variables. Therefore, it can be concluded that the assumption of homoscedasticity is met, which supports the reliability of the regression estimates.

3.7.2. Independent Assumption Test

In regression analysis, the independence assumption requires that residuals defined as the differences between observed and predicted values are uncorrelated with one another. Table 21 presents the findings of the Durbin-Watson test, which is used to assess this assumption.

Table 21. Durbin Watson Test Result

d_{hitung}	$d_{L,0.05}$	$d_{U,0.05}$	$4\text{-}d_{U,0.05}$	$4\text{-}d_{L,0.05}$
1,9510	1,2614	1,7223	2,2767	2,7386

Table 21 shows that the Durbin-Watson test produced a $d_{calculated}$ value of 1.9510. When compared with $d_{U,0.05}$ and 4 - $d_{U,0.05}$, it can be seen that $d_{U,0.05} < d_{calculated} < 4 - d_{U,0.05}$ ($1.7916 < 1.9510 < 2.2767$). Since the calculated statistics are within the acceptance region, there is not enough evidence to reject the null hypothesis of no autocorrelation. Thus, the residuals can be considered independent, meeting the main assumption of this regression analysis and can be continued with the normal distribution test.

3.7.3. Normality Test

The Kolmogorov-Smirnov (K-S) test was utilized to assess whether the model's residuals are normally distributed. The findings of the Kolmogorov-Smirnov test are illustrated in Figure 3 below.

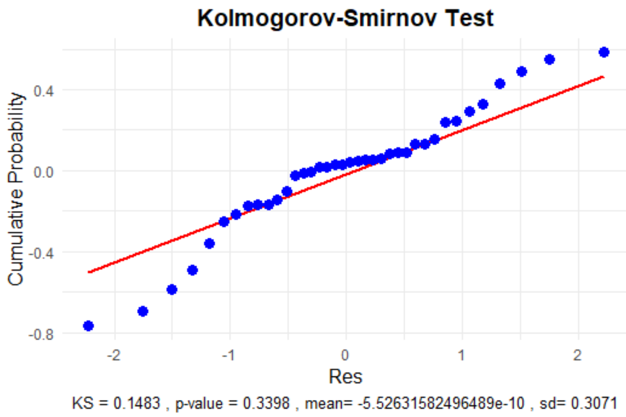

Figure 3. Kolmogorov-Smirnov Test Result

Referring to the Kolmogorov-Smirnov test results presented in Figure 2, the p-value obtained is 0.3398. Since the *p-value* > *α* with α = 0.05, the null hypothesis cannot be rejected. This outcome suggests that the residuals are approximately normally distributed, thereby meeting the assumption of normality and confirming the appropriateness of statistical inference from the model.

3.8. Coefficient of Determination

The nonparametric linear spline regression model, employing a knot combination selected through the mGCV method, produces an R² value of 82.91%. This indicates that the four predictor variables collectively account for 82.91% of the variation in Life Expectancy across East Java Province. In other words, the model demonstrates a strong goodness-of-fit, indicating that the selected predictors contribute significantly to explaining the patterns and factors influencing life expectancy in the region. The remaining 17.09% may be attributed to other variables not included in the model, random variation, or measurement errors.

3.9. Interpretation of Life Expectancy Using a Nonparametric Spline Regression Model

1. Assuming all variables are constant, the following is the relationship between Life Expectancy (Y) and the Percentage of Poor Population (X1).

$$\hat{y} = 0{,}558x_1 - 4{,}02(x_1 - 5{,}961) + 4{,}408(x_1 - 6{,}686) - 1{,}037(x_1 - 7{,}774)$$

$$= \begin{cases} 0{,}558x_1 & ; \quad x_1 < 5{,}961 \\ -3{,}425x_1 + 32{,}96 & ; \quad 5{,}961 \le x_1 < 6{,}686 \\ 0{,}946x_1 - 5{,}511 & ; \quad 6{,}686 \le x_1 < 7{,}774 \\ -0{,}091x_1 + 2{,}55 & ; \quad x_1 \ge 7{,}774 \end{cases}$$

If a regency/city with a Percentage of Poor Population value of less than 5.961 percent experiences an increase of 0.1 percent, assuming the other predictor variables remain constant, Life Expectancy will increase by 0.558 percent. The regencies/cities that fall within this interval include Batu City, Malang City, Surabaya City, Madiun City, Sidoarjo Regency, and Mojokerto City. This phenomenon reflects that in developed cities, access to health services, nutrition, and sanitation is maintained even though the number of poor people has increased slightly. Health programs such as JKN-KIS (National Health Insurance) and 24-hour health center services in big cities allow poor people to continue to obtain decent basic health services, so that the negative impacts of poverty can be reduced.

2. Assuming all variables are constant, the following is the relationship between the percentage of the population aged 5 years and over who regularly smoke tobacco (X2) and Life Expectancy (Y).

$$\hat{y} = 0.267x_2 - 4{,}197(x_2 - 18{,}973) + 7{,}692(x_2 - 19{,}407) - 3{,}887(x_2 - 20{,}057)$$
$$= \begin{cases} 0{,}267x_2 & ; \quad x_2 < 18{,}973 \\ -3{,}93x_2 + 79{,}629 & ; \quad 18{,}973 \le x_2 < 19{,}407 \\ 3{,}762x_2 - 69{,}649 & ; \quad 19{,}407 \le x_2 < 20{,}057 \\ -0{,}125x_2 + 8{,}312 & ; \quad x_2 \ge 20{,}057 \end{cases}$$

If a regency/city has a percentage of the population aged 5 years and over who smoke tobacco above 20.057 percent, and this value increases by 0.1 percent, the life expectancy within this interval will tend to decrease by 0.125 percent. The regencies/cities in this segment are Lamongan, Situbondo, Sampang, Banyuwangi, etc. In areas with very high smoking rates, such as Situbondo, Sampang, and Lamongan, it is often exacerbated by low awareness of the dangers of smoking, minimal access to health services, and social norms that consider smoking to be commonplace, especially among men of productive age. This condition causes an increase in disease burden and impacts decreasing life expectancy, although the age structure and local demographic conditions still influence the magnitude of the impact.

3. Assuming all variables are constant, the following is the relationship between years of schooling expectation (X3) and life expectancy (Y).

$$\hat{y} = 3{,}027x_3 - 12{,}454(x_3 - 12{,}602) + 11{,}106(x_3 - 12{,}757) - 1{,}215(x_3 - 12{,}990)$$
$$= \begin{cases} 3{,}027x_3 & ; \quad x_3 < 12{,}602 \\ -9{,}427x_3 + 156{,}94 & ; \quad 12{,}602 \le x_3 < 12{,}757 \\ 1{,}679x_3 + 15{,}26 & ; \quad 12{,}757 \le x_3 < 12{,}990 \\ 0{,}464x_3 + 31{,}042 & ; \quad x_3 \ge 12{,}990 \end{cases}$$

If a regency/city has a years of schooling expectation value of more than 12.990 percent and increases by 0.1 percent, this will result in a 0.464 percent increase in life expectancy. The regencies/cities that fall within this segment are Mojokerto city, Madiun city, Batu city, etc. It shows that although increasing education at an already high level has a positive impact in the long term, the magnitude of the effect tends to be more moderate than in areas with lower levels of education. This can be explained because the benefits of primary and secondary education on health and a healthy lifestyle have already been met, so that further increases only provide relatively small additional effects.

4. Assuming all variables are constant, the following is the relationship between the open unemployment rate (X4) and life expectancy (Y).

$$\hat{y} = 2{,}336x_4 - 19{,}207(x_4 - 2{,}364) + 21{,}517(x_4 - 2{,}566) - 4{,}363(x_4 - 2{,}867)$$
$$= \begin{cases} 2{,}336x_4 & ; \quad x_4 < 2{,}364 \\ -16{,}871x_4 + 45{,}405 & ; \quad 2{,}364 \le x_4 < 2{,}566 \\ 4{,}646x_4 - 9{,}807 & ; \quad 2{,}566 \le x_4 < 2{,}867 \\ 0{,}283x_4 + 2{,}071 & ; \quad x_4 \ge 2{,}867 \end{cases}$$

If a regency/city has an open unemployment rate value of less than 2.364 percent and increases by 0.1 percent, assuming the other predictor variables remain constant, life expectancy will increase by 2.336 percent. The regencies/cities that fall within this interval are Pacitan Regency, Pamekasan Regency, and Sumenep Regency. This phenomenon can be explained by the fact that unemployment in these regions does not necessarily indicate an economic crisis, but may instead result from increased participation in education or labor migration to other areas. Furthermore, in smaller cities, the growing interest of the population in small-scale entrepreneurship and family-based economic activities means that being unemployed does not always equate to being unproductive.

4. CONCLUSION

Data for 2024 indicate that the average life expectancy in East Java Province is 74.8 years, with a variance of 5.52. The highest life expectancy is observed in Surabaya City (76.02 years), while Bondowoso Regency reports the lowest (73.31 years). These statistics highlight the health and quality of life disparities that remain to be addressed across various regions of East Java. However, overall, the province demonstrates a relatively favorable performance regarding population welfare indicators. These findings are expected to provide input to local governments. Given the statistically significant impacts of poverty, smoking, education, and unemployment on life expectancy, an integrated development strategy should be pursued. Policymakers are encouraged to prioritize access to education, strengthen tobacco control policies, and promote job creation programs, especially in districts with lower life expectancies.

The most appropriate truncated nonparametric spline regression model was obtained using three knot points, selected based on the modified Generalized Cross Validation (mGCV) approach, which yielded the lowest value of 0.100. The results of both simultaneous and partial significance tests indicate that all four predictor variables have statistically significant effects on life expectancy jointly and individually. The coefficient of determination (R^2) value of 82.91% suggests that the model can explain most of the variation in life expectancy across districts and municipalities in East Java. In comparison, the remaining 17.09% is attributed to other unaccounted-for factors.

Although the mGCV approach was employed to minimize the risk of overfitting, it should be noted that using three knot points per variable in the context of a relatively small number of observations (38 regions) still carries a potential risk of overfitting. Therefore, future research may consider more adaptive knot selection methods or regularization-based approaches to improve model generalizability and ensure robust performance in similar empirical contexts.

REFERENCES

[1] S. O. Shavira, M. Balafif, and N. Imamah, “Pengaruh Pertumbuhan Ekonomi, Upah Minimum, dan Tingkat Pengangguran terhadap Kesejahteraan Masyarakat di Jawa Timur Tahun 2014-2018,” *Bharanomics*, vol. 1, no. 2, pp. 93–103, 2021, doi: 10.46821/bharanomics.v1i2.158.

[2] T. Muhaimin, “Children ’ s quality of life,” *J. Kesehat. Masy. Nas.*, vol. 5, no. 2, pp. 1–23, 2010, doi: 10.21109/kesmas.v5i2.148.

[3] BPS, “Angka Harapan Hidup (AHH) Menurut Provinsi dan Jenis Kelamin (Tahun), 2022-2024.” [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NTAxIzI=/angka-harapan-hidup--ahh--menurut-provinsi-dan-jenis-kelamin--tahun-.html

[4] Putra, “Determinan Tingkat Kemiskinan di Provinsi Jawa Timur Periode 2009-2013,” *Skrpsi*, vol. 6, no. 3, pp. 645–654, 2015, [Online]. Available: https://books.google.co.id/books?id=D9_YDwAAQBAJ&pg=PA369&lpg=PA369&dq=Prawirohardjo,+Sarwono.+2010.+Buku+Acuan+Nasional+Pelayanan+Kesehatan++Maternal+dan+Neonatal.+Jakarta+:+PT+Bina+Pustaka+Sarwono+Prawirohardjo.&source=bl&ots=riWNmMFyEq&sig=ACfU3U0HyN3I

[5] D. A. Prasetya, A. P. Sari, M. Idhom, and A. Lisanthoni, “Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports,” *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 113–122, Feb. 2025, doi: 10.35882/skzqbd57.

[6] A. Muhaimin, T. M. Fahrudin, and S. S. Alamiyah, “Sentiment analysis in social media: Case study in Indonesia,” vol. 8, pp. 27–30, 2024, [Online]. Available: http://dx.doi.org/10.11594/nstp.2024.4106

[7] T. Trimono, A. H. Abdillah, and M. Risqi, “Analisis Perubahan Tingkat Pengangguran di Kabupaten/Kota Kalimantan Barat Tahun 2010-2018,” *Pros. Semin. Nas. Sains Data*, vol. 3, no. 1, pp. 213–221, 2023, doi: 10.33005/senada.v3i1.115.

[8] D. H. Ratnasari and N. Nugraheni, “Peningkatan Kualitas Pendidikan Di Indonesia Dalam Mewujudkan Program Sustainable Development Goals (Sdgs),” *J. Citra Pendidik.*, vol. 4,

no. 2, pp. 1652–1665, 2024, doi: 10.38048/jcp.v4i2.3622.

[9] D. A. Prasetya, P. T. Nguyen, R. Faizullin, I. Iswanto, and E. F. Armay, "Resolving the shortest path problem using the haversine algorithm," *J. Crit. Rev.*, vol. 7, no. 1, pp. 62–64, 2020, doi: 10.22159/jcr.07.01.11.

[10] I. Zain, E. O. Permatasari, M. Mardiyono, A. Muhaimin, and D. A. Rasyid, "ANALYSIS OF THE RELATIONSHIP BETWEEN CHARACTERISTICS OF TEENAGERS AND FAMILY FUNCTIONS ON TEENAGERS' BEHAVIOR FOR CONSUMING DRUGS IN EAST JAVA," *J. Biometrika dan Kependud.*, vol. 11, no. 02, pp. 122–133, Nov. 2022, doi: 10.20473/jbk.v11i02.2022.122-133.

[11] J. W. Kusuma, H. Hamidah, U. Umalihayati, and P. P. Rini, "Mengurai Benang Kusut Kebijakan Pendidikan Indonesia: Sebuah Literature Review Analitik," *J. Ilm. Glob. Educ.*, vol. 5, no. 2, pp. 1810–1826, 2024, doi: 10.55681/jige.v5i2.2772.

[12] K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji, and E. M. Safitri, "Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020*, no. March 2020, pp. 344–347, 2020, doi: 10.1109/ISRITI51436.2020.9315484.

[13] Wahidah Alwi, Adnan Sauddin, and Nahda Islamiah. M, "Faktor-Faktor Yang Mempengaruhi Angka Harapan Hidup Di Sulawesi Selatan Menggunakan Analisis Regresi," *J. MSA (Mat. dan Stat. serta Apl.*, vol. 11, no. 1, pp. 72–80, 2023, doi: 10.24252/msa.v11i1.32266.

[14] F. F. Muhammad, F. Abdurrahim, J. P. Gunawan, M. T. Rahma, T. A. Oktaviana, and E. Antriyandarti, "Analysis study: pengaruh faktor AHH (angka harapan hidup) pada masyarakat Provinsi Jawa Tengah tahun 2013-2021," *Kemakmuran Hijau J. Ekon. Pembang.*, vol. 1, no. 1, pp. 11–22, 2024, doi: 10.61511/jekop.v1i1.2024.744.

[15] M. Maharani and D. R. S. Saputro, "Generalized Cross Validation (GCV) in Smoothing Spline Nonparametric Regression Models," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1808, no. 1, 2021, doi: 10.1088/1742-6596/1808/1/012053.

[16] M. A. Lukas, F. R. De Hoog, and R. S. Anderssen, "Performance of Robust GCV and Modified GCV for Spline Smoothing," *Scand. J. Stat.*, vol. 39, no. 1, pp. 97–115, 2012, doi: 10.1111/j.1467-9469.2011.00736.x.

[17] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, 2nd editio. Boca Raton: CRC Press, 1999. doi: 10.1201/9781482273144.

[18] M. Palli, A, S., Jaafar, J., Gilal, A. R., Alsughayyir, A., Gomes, H. M., Alshanqiti, A., & Omar, "Journal of information and communication technology : JICT," *J. Inf. Commun. Technol.*, vol. 5, no. 1, pp. 45–62, 2002, [Online]. Available: http://e-journal.uum.edu.my/index.php/jict/article/view/8062

[19] R. A. Gunawan, D. Prasetya Zulkarmain, S. T. Arianto, F. Ekonomi, and D. Bisnis, "Socius: Jurnal Penelitian Ilmu-Ilmu Sosial Perbandingan Metode Ordinary Least Square (OLS) dan Metode Partial Least Square (PLS) Untuk Mengatasi Multikolinearitas," *J. Ilm. Bid. Sos. Ekon. Budaya, Ekol. dan Pendidik.*, vol. 1, no. 6, pp. 795–808, 2023, [Online]. Available: https://doi.org/10.5281/zenodo.10476911

[20] F. Lamusu, T. Machmud, and R. Resmawan, "Estimator Nadaraya-Watson dengan Pendekatan Cross Validation dan Generalized Cross Validation untuk Mengestimasi Produksi Jagung," *Indones. J. Appl. Stat.*, vol. 3, no. 2, p. 85, 2021, doi: 10.13057/ijas.v3i2.42125.

[21] A. T. Damaliana, I. Nyoman Budiantara, and V. Ratnasari, "Comparing between mgcv and agcv methods to choose the optimal knot points in semiparametric regression with spline truncated using longitudinal data," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 3, 2019, doi: 10.1088/1757-899X/546/3/032003.

[22] A. K. Khotimah, A. A. Rahman, M. Z. Alam, Y. H. Nur, and T. R. Aufi, "Analisis Regresi Linier Berganda Dalam Estimasi Indeks Pembangunan Manusia di Indonesia Multiple Linear Regression Analysis In Estimating The Human Development Index In Indonesia," vol. 15, no. November, pp. 90–99, 2024, doi: 10.30872/eksponensial.v15i2.1318.

[23] A. Hazra, "Using the confidence interval confidently," *J. Thorac. Dis.*, vol. 9, no. 10, pp.

4125–4130, 2017, doi: 10.21037/jtd.2017.09.14.

[24] F. A. Firdausya and R. Indawati, "Perbandingan Uji Glejser Dan Uji Park Dalam Mendeteksi Heteroskedastisitas Pada Angka Kematian Ibu Di Provinsi Jawa Timur Tahun 2020," *J. Ners*, vol. 7, no. 1, pp. 793–796, 2023, doi: 10.31004/jn.v7i1.14069.

[25] R. E. Nugroho, "Analisis Faktor – Faktor Yang Mempengaruhi Pengangguran di Indonesia Periode 1998 – 2014," *Pasti*, vol. X, no. 2, pp. 177–191, 2014.

[26] I. Sintia, M. D. Pasarella, and D. A. Nohe, "Perbandingan Tingkat Konsistensi Uji Distribusi Normalitas Pada Kasus Tingkat Pengangguran di Jawa," *Pros. Semin. Nas. Mat. Stat. dan Apl.*, vol. 2, no. 2, pp. 322–333, 2022.

[27] T. Maidarti, M. Azizah, E. Wibowbo, and I. Nuswandari, "Pengaruh pelatihan dan Motivasi Kerja Terhadap Kinerja Karyawan Pada PT. SARAKA MANDIRI SEMESTA BOGOR," *Deriv. J. Manaj.*, vol. 16, no. 1, pp. 127–145, 2022.