Variables Selection Affecting Indonesian Human Development Index Using LASSO

Etis Sunandi[1*]**, Titin Siswantining**[2]
[1]Department of Mathematics, University of Bengkulu, Bengkulu, Indonesia
[2] Department of Mathematics, University of Indonesia, Depok, Indonesia
*Corresponding author: esunandi@unib.ac.id

Received: 24 May 2025
Revised: 18 June 2025
Accepted: 1 July 2025

ABSTRACT – According to Statistics Indonesia, the Human Development Index (HDI) is a measure that reflects the level of human development achievement in a region, based on three basic dimensions: a long and healthy life, knowledge, and a decent standard of living. There are many factors that are suspected to influence HDI in Indonesia. Another hand, estimation of parameters in regression analysis using the Least Squares Method will experience problems, if the number of independent variables is greater than the number of observations. One method that can be used to overcome this problem is to use the Least Absolute Shrinkage and Selection Operator (LASSO) method. The purpose of this study is the selection of variables that affect Indonesia's Human Development Index (HDI) in 2023 using the LASSO. The LASSO method is known as a model used to select independent variables while overcoming multicollinearity problems. The ridge regression model is used as a comparison model. The results showed that LASSO Analysis is better than Ridge Regression. This can be seen from the Mean Squared Error of Prediction (MSEP) of LASSO (0.34) is smaller than the ridge regression (3.61). In addition, the r-squared value of LASSO is higher, which is 97.6%.

Keywords – LASSO, MSEP, R-Squared, Ridge Regression

I. INTRODUCTION

Linear regression analysis is one of the statistical methods used to model, analyze the strength of relationships, and predict the effect between one or more independent variables on a dependent variable [1]. In linear regression modeling, the least squares method is commonly used to estimate regression parameters. However, this method cannot be applied when there are issues such as the number of independent variables exceeding the number of observations, or when multicollinearity exists among the independent variables. In such cases, the method becomes unsuitable as it leads to inaccurate estimation results. To overcome these problems, alternative regression methods such as LASSO regression and Ridge regression can be used.

The Least Absolute Shrinkage and Selection Operator (LASSO) regression was developed by Tibshirani in 1996 as a method for variable selection while simultaneously optimizing parameter estimates [2]. LASSO regression can automatically perform variable selection by shrinking the coefficients of less relevant variables to zero, which helps prevent overfitting in the model. In addition to LASSO regression, Ridge regression can also be used as a comparison to address these issues. Ridge regression has an estimation method similar to that of LASSO regression. While LASSO estimates parameters based on the sum of squared errors constrained by the sum of the absolute values of the coefficients, Ridge regression estimates parameters based on the sum of squared errors constrained by the sum of the squared coefficients [3].

LASSO regression is widely used in various scientific fields, especially when the analysis involves many predictor variables and aims to select important variables [4]. One of the fields where this method can be applied is social science and development. The Human Development Index (HDI) is an important indicator in the field of social science and development. According to Statistics Indonesia (BPS), the Human Development Index (HDI) is a measure that reflects the level of human development achievement in a region, based on three basic dimensions: a long and healthy life, knowledge, and a decent standard of living. Moreover, HDI achievements across regions can guide policymakers in determining development priority scales within a country. By monitoring HDI trends and its component indicators over time, the government can assess whether the implemented policies have had a positive impact on the well-being of the population.

Indonesia's Human Development Index (HDI) has been categorized as high (above 70) since 2016 and has shown consistent growth from 2020 to 2023. It started at 72.81 in 2020, increased to 73.16 in 2021, continued to rise to 73.77 in 2022, and eventually reached 74.39 in 2023. The rate of HDI growth has also accelerated each year, beginning with a 0.48 percent increase from 2020 to 2021, rising to 0.83 percent from 2021 to 2022, and recording the fastest growth of 0.84 percent from 2022 to 2023. On average, Indonesia's HDI grew by 0.72 percent per year during the 2020–2023 period [5].

DOI: 10.12962%2Fj27213862.v8i2.22891

There are many factors that are thought to affect the Human Development Index (HDI) in Indonesia. The number of factors is often referred to as the concept of high-dimensional data, where overfitting or multicollinearity often occurs during analysis. Therefore, the LASSO regression method can be applied to overcome this. Based on this background, this study aims to select variables that affect the Human Development Index (HDI) in 2023 using the LASSO regression method. In this study, a comparison of the goodness of the model between LASSO and ridge regression was carried out. The research data comes from Statistics Indonesia (BPS). There are 48 independent variables and HDI as the dependent variable.

II. LITERATURE REVIEW

Linear Regression

Regression analysis is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables, with the aim of understanding the effect of the independent variables on the dependent variable. In linear regression, there are two types: simple linear regression and multiple linear regression [6].

Simple linear regression is a statistical method used to describe and model the relationship between one dependent variable and one independent variable, whereas multiple linear regression is used to model the relationship between one dependent variable and two or more independent variables [7]. The following is the model equation for linear regression:

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\boldsymbol{X} + \boldsymbol{\varepsilon} \qquad (1)$$

Where:

Y : Dependent variable
β_0 : Intercept (constant)
β_1 : Regression coefficient for variable X
X : Independent variable
ε : Error (residual)

The model equation for multiple linear regression is as follows [8]:

$$\boldsymbol{Y} = \beta_0 + \beta_1\boldsymbol{X_1} + \beta_2\boldsymbol{X_2} + \cdots + \beta_p\boldsymbol{X_p} + \boldsymbol{\varepsilon} \qquad (2)$$

Where $\beta_i X_i$ represents the regression coefficient for the i independent variable.

Least Absolute Shrinkage and Selection Operator (LASSO) Regression

LASSO regression was first introduced by Robert Tibshirani in 1996 that states that LASSO is an estimation method that minimizes the sum of squared errors, constrained by the sum of the absolute values of the coefficients [9]. LASSO regression is capable of eliminating variables from the regression model by shrinking the coefficients to zero, thus performing variable selection and addressing multicollinearity issues [10]. This regression method can also help overcome overfitting [11], [12]. Overfitting occurs when the model fits the training data too closely, thereby reducing its ability to generalize to testing or new data [13].

LASSO regression is an extension of the OLS method by shrinking some coefficients and setting others to zero, thereby selecting the best variables (selection operator) from both subset selections. Therefore, the LASSO method automatically selects relevant and significant variables [14]. In LASSO, a penalty parameter is given with the following constraint [15]:

$$||\boldsymbol{\beta}||_1^1 \leq t \qquad , t \geq 0 \qquad (3)$$

The value of t above represents the magnitude of shrinkage on the LASSO coefficient estimator, where $t \geq 0$. If the estimator $\boldsymbol{\beta}$ is the least squares estimator and $t_0 = ||\boldsymbol{\beta}||_1^1$, then values of $t < t_0$ will lead to a solution where the regression with the OLS estimator shrinks towards zero, and allows some coefficients to shrink exactly to zero.

The parameter estimation in LASSO regression is as follows [16]:

$$\begin{aligned}(\boldsymbol{\beta}, \boldsymbol{\lambda})_{LASSO} &= arg_{\beta}^{min} L_{LASSO}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \qquad (4)\\ &= arg_{\boldsymbol{\beta}}^{min}\left(||\boldsymbol{y} - \boldsymbol{\beta_0} - \boldsymbol{X\beta}||_2^2 + \boldsymbol{\lambda}\textstyle\sum_j^k |\boldsymbol{\beta_j}|\right)\\ &= arg_{\boldsymbol{\beta}}^{min}\left(||\boldsymbol{y} - \boldsymbol{\beta_0} - \boldsymbol{X\beta}||_2^2 + \boldsymbol{\lambda}\textstyle\sum_j^k |\boldsymbol{\beta}||_1^1\right)\end{aligned}$$

Where λ represents the penalty, term applied in LASSO analysis, with a value of lambda greater than or equal to zero, constrained by $D = \sum_{j=1}^{p} |\beta^L| \leq t$ (contant). The value of λ can be adjusted according to the researcher's preference. However, in general, the optimal lambda is determined using the cross-validation method [17].

DOI: 10.12962%2Fj27213862.v8i2.22891

Ridge Regression

Ridge regression was first introduced by Hoerl and R.W. Kennard in 1962. Essentially, Ridge regression is an extension of the least squares method [18]. Ridge regression is a method that can address issues of multicollinearity [19] and overfitting [16]. In Ridge regression, a bias term c is added to the diagonal of the matrix $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ to estimate the regression coefficients. The following is the Ridge regression model equation [20]:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^{\boldsymbol{R}} + \boldsymbol{\varepsilon} \tag{5}$$

Where β^R is the coefficient of the Ridge regression? The regression coefficients are influenced by the bias constant c as follows:

$$\boldsymbol{\beta}^{\boldsymbol{R}}(c) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + c\boldsymbol{I})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}\,, \quad c > 0 \tag{6}$$

Where $\beta^R(c)$ is the Ridge regression coefficient with the regularization parameter c, X^T is the transpose of the matrix X, I is the identity matrix, and c is the regularization constant, with a value greater than zero.

Mean Squared Error of Prediction (MSEP)

The Mean Squared Error of Prediction (MSEP) is one of the evaluation metrics for regression models used to assess the accuracy of predictions made by the model on new or test data. MSEP measures the average squared error between the actual values and the predicted values [21]. A smaller MSEP value indicates that the model provides more accurate results. The equation for calculating the MSEP is as follows:

$$MSEP = \sum_{\mathrm{j}=1}^{n} \frac{(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2}{n} \tag{7}$$

R-Squared

R-Squared is an evaluation metric used to measure how well the model explains the variability in the data [22]. An R-Squared value of 0 indicates that the model performs no better than a simple mean, while a value of 1 indicates that the model provides an accurate fit. The equation for calculating the R-Squared value is as follows:

$$R - Squared \; = 1 - \frac{\sum_{i=1}^{n}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2}{\sum_{i=1}^{n}(\mathrm{y_i} - \bar{y})^2} \tag{8}$$

Where $\mathrm{y_i}$ is the actual value of the i-th data point, $\hat{\mathrm{y}}_\mathrm{i}$ is the predicted value from the model for the i-th data point, and $\bar{y}$ is the mean value of all the y values or the dependent variable.

III. METHODOLOGY

The research method is an empirical study. The application of statistical methods is conducted on socio-economic data in Indonesia. This study uses secondary data sourced from the Statistics Indonesia (BPS). The variables in this study consist of dependent variables and independent variables. The dependent variable is the Human Development Index (HDI) data. While the independent variables are covariates that tend to influence HDI. The independent variables used are 48 variables. The research variables are tabulated in Table 1.

Table 1 Research Variables

No	Variable name	Scale
Y	Human Development Index (HDI)	Ratio
X_1	Percentage of Women Aged 15–49 Years Who Have Been Married and Gave Birth in the Last Two Years with Midwifery Assistance	Ratio
X_2	Percentage of Population with Health Insurance – BPJS Contribution Assistance Recipients (PBI)	Ratio
X_3	Percentage of Population with Health Insurance – BPJS Non-Contribution Assistance Recipients (Non-PBI)	Ratio
X_4	Percentage of Population with Health Insurance – JAMKESDA	Ratio
X_5	Percentage of Population with Health Insurance – Private Insurance	Ratio
X_6	Percentage of Population with Health Insurance – Office/Company	Ratio
X_7	Percentage of Women Aged 15–49 Years Who Are Married and Using Family Planning	Ratio
X_8	Percentage of Smokers Aged ≥ 15 Years by Province	Ratio
X_9	Unmet Health Service Needs by Province (Percent)	Ratio
X_{10}	Percentage of Population Who Went to Government Hospitals in the Last Month	Ratio
X_{11}	Percentage of 6-Month-< Babies Who Get Exclusive Breastfeeding by Province	Ratio
X_{12}	Percentage of Population Aged 15 and Over Who Smoked in the Last Month (Age Group 15–24 Years)	Ratio
X_{13}	Life Expectancy (AHH)	Ratio
X_{14}	Percentage of Illiterate ≥ 10-Year-Old Population	Ratio
X_{15}	School Participation Rate (APS) Ages 19–23	Ratio

X_{16}	Percentage of ≥-10-Year-Old Population Who Never Went to School	Ratio
X_{17}	Percentage of Literate ≥ 15-Year-Old Population	Ratio
X_{18}	High School Pure Participation Rate	Ratio
X_{19}	Education Graduation Rate by Province and Level of Education	Ratio
X_{20}	Community Literacy Development Index	Ratio
X_{21}	Open Unemployment Rate (TPT) by Province (Percent)	Ratio
X_{22}	Open Unemployment Rate (TPT) – August	Ratio
X_{23}	Labour Force Participation Rate (TPAK) – August	Ratio
X_{24}	Percentage of Children Aged 10–17 Years Working by Province (Percent)	Ratio
X_{25}	Number of Job Seekers Registered – Total	Ratio
X_{26}	Number of Registered Job Vacancies – Total Placements/Fulfillment of Workforce	Ratio
X_{27}	Workforce Placement – Total	Ratio
X_{28}	Access to Drinking Water Services	Ratio
X_{29}	Access to Basic Sanitation Services	Ratio
X_{30}	Access to Basic Health Facilities	Ratio
X_{31}	Proportion of Households with Home Ownership Status	Ratio
X_{32}	Percentage of Households with Own Defecation Facilities	Ratio
X_{33}	Percentage of Households without Defecation Facilities	Ratio
X_{34}	Percentage of Households That Use the Main Fuel for Cooking in the Form of Gas/LPG	Ratio
X_{35}	Percentage of Households with a Floor Area < 19 m²	Ratio
X_{36}	Percentage of Households with Protected Drinking Water Sources (Protected Wells)	Ratio
X_{37}	Percentage of Households with Occupancy Area ≤ 7.2 m² per Capita by Province and Type of Region	Ratio
X_{38}	Percentage of Households by Province and Source of Decent Drinking Water (Percent)	Ratio
X_{39}	Percentage of Households with Access to Decent and Affordable Housing by Province	Ratio
X_{40}	Percentage of Households with Proper Sanitation	Ratio
X_{41}	Percentage of Households with the Largest Non-Land Floor	Ratio
X_{42}	P1 (Indicators of Inequality in Poor Population's Expenditure)	Ratio
X_{43}	Gini Ratio by Province and Region	Ratio
X_{44}	Percentage of Poor Population	Ratio
X_{45}	Average Expenditure per Capita per Month	Ratio
X_{46}	Number of Divorces	Ratio
X_{47}	Child Mortality Rate (CMR)	Ratio
X_{48}	Maternal Mortality Rate (MMR)	Ratio

In this research, the number of covariates exceeds the number of observations. This study modeled HDI using the LASSO method. The ridge regression method is used as a comparison model. Ridge regression and LASSO modeling on data is done using the glm.net package in R software. The steps to be followed in the analysis for this study are as follows:

1. Conduct data exploration, including descriptive statistical analysis.
2. Standardize the data.
3. Perform LASSO analysis uses eq.4
4. Conduct Ridge regression analysis as a comparison to the LASSO method uses eq.6
5. Calculate the MSEP and R-Squared values use eq. 7 and 8.
6. Evaluate the models by observing the smallest values of MSEP and R-Squared for both the LASSO and Ridge regression methods performed.
7. Select the best model.

IV. RESULTS AND DISCUSSIONS

This research began with data exploration. The Human Development Index (HDI) in Indonesia is very diverse. Some provinces have a higher HDI than the national. These provinces are DKI Jakarta, DI Yogyakarta, Riau Islands, East Kalimantan, Bali, and Banten. DKI Jakarta Province has the highest HDI at 83.55. while the lowest is Papua Province at 63.01. HDI distribution per province in Indonesia can be seen in Figure 1.

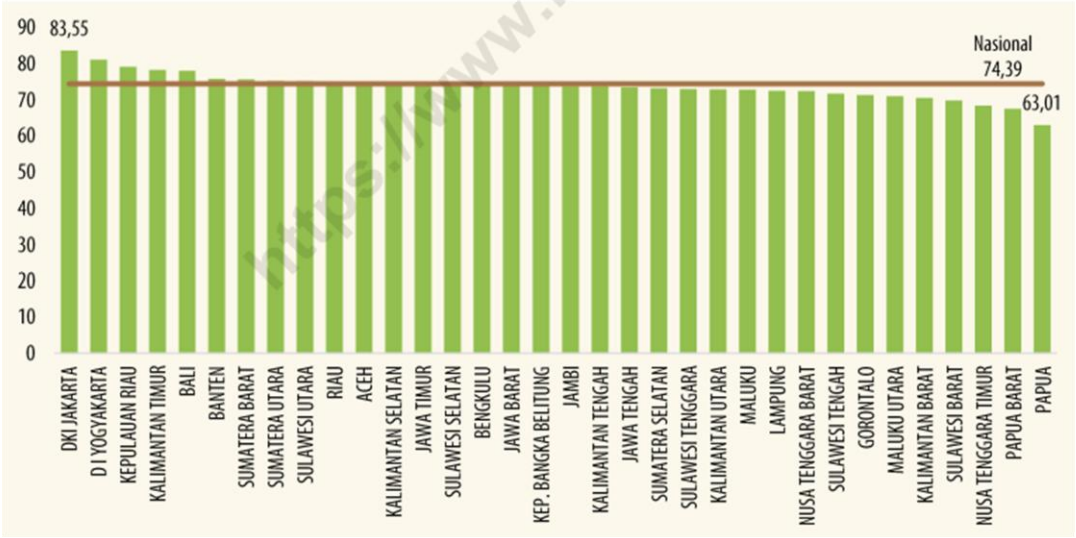

Figure 1 HDI by Province, 2023
*Source: Statistics Indonesia (BPS)

The ridge and LASSO regression models overcome multicollinearity by shrinking the regression coefficients close to zero as well as variable selection. The most important thing about LASSO regression is choosing the optimal lambda (λ). The optimal $\boldsymbol{\lambda}$ can be determined using the k-fold validation process. From the results of the cross-validation process, the optimal λ value in the ridge regression model is 34,219 by looking at the smallest MSE value. In LASSO, the optimal λ value obtained through the cross-validation process that minimizes MSE is 0.186. The LASSO shrinkage parameter (λ) was obtained by the cross-validation method that resulted in the minimum MSE. The scatter plot of LASSO lambda values, Figure 2, shows that the larger the log ($\boldsymbol{\lambda}$) value, the smaller the number of independent variables used. In addition, it can be seen that by using the optimum $\boldsymbol{\lambda}$, the number of independent variables used is 18 variables.

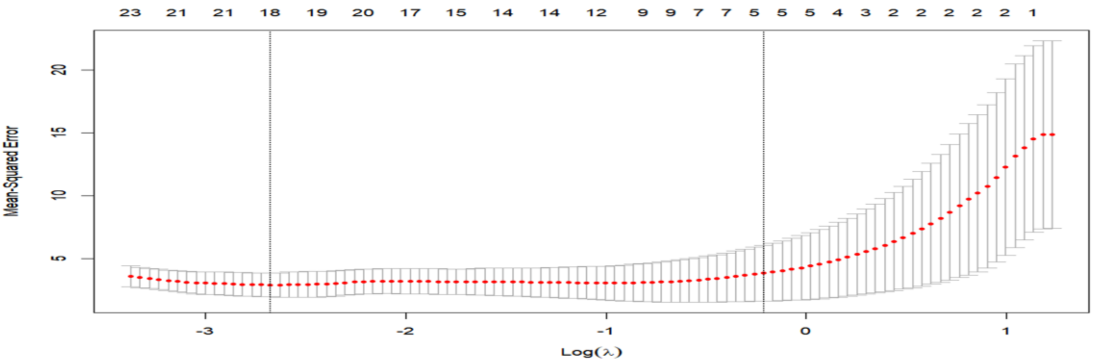

Figure 2 Plot of LASSO lambda values

Based on Table 2, the LASSO regression coefficient shrinks to exactly zero, resulting in a simpler regression model as well as overcoming multicollinearity problems because the independent variables that have a high correlation value will be selected from the model. There are 18 selected variables. Thus, the independent variables that affect the HDI are Percentage of Women Aged 15–49 Years Who Have Been Married and Gave Birth in the Last Two Years with Midwifery Assistance (x1), Percentage of Population with Health Insurance – BPJS Non-Contribution Assistance Recipients (Non-PBI)(x3), Percentage of Population Who Went to Government Hospitals in the Last Month (x10), School Participation

Rate (APS) Ages 19–23 (x15), Percentage of Literate ≥ 15-Year-Old Population (x17), Education Graduation Rate by Province and Level of Education (x19), Access to Drinking Water Services (x28), Access to Basic Sanitation Services (x29), Access to Basic Health Facilities (x30), Proportion of Households with Home Ownership Status (x31), Percentage of Households by Province and Source of Decent Drinking Water (x38), Percentage of Households with Proper Sanitation (x40), and Gini Ratio by Province and Region (x43).

Table 2 Coefficients of LASSO model

variable	estimate	variable	estimate	variable	estimate	variable	estimate
$(\boldsymbol{Intercept})$	72.62	X_{13}	.	X_{25}	.	X_{37}	.
$\boldsymbol{X_1}$	−0.06	X_{14}	.	X_{26}	.	X_{38}	0.37
$\boldsymbol{X_2}$	.	X_{15}	0.49	X_{27}	.	X_{39}	.
$\boldsymbol{X_3}$	0.69	X_{16}	.	X_{28}	−0.07	X_{40}	0.01
$\boldsymbol{X_5}$	.	X_{17}	0.23	X_{29}	0.26	X_{41}	.
$\boldsymbol{X_6}$	.	X_{18}	.	X_{30}	0.14	X_{42}	.
$\boldsymbol{X_7}$	.	X_{19}	0.79	X_{31}	−0.04	X_{43}	0.17
$\boldsymbol{X_8}$	.	X_{20}	.	X_{32}	.	X_{44}	.
$\boldsymbol{X_9}$	.	X_{21}	.	X_{33}	.	X_{45}	.
$\boldsymbol{X_{10}}$	0.04	X_{22}	.	X_{34}	0.44	X_{46}	.
$\boldsymbol{X_{11}}$	.	X_{23}	.	X_{35}	0.70	X_{47}	−0.63
$\boldsymbol{X_{12}}$	.	X_{24}	−0.34	X_{36}	.	X_{48}	−0.73

LASSO regression modeling uses the optimum λ. The LASSO regression model formed is:

$$\boldsymbol{\hat{y} = 72.62 - 0.06X_1 + 0.69X_3 + 0.04X_{10} + 0.49X_{15} + 0.23X_{17} + 0.79X_{19} - 0.34X_{24} - 0.07X_{28} + 0.26X_{29}}$$
$$\boldsymbol{+ 0.14X_{30} - 0.04X_{31} + 0.44X_{34} + 0.70X_{35} + 0.37X_{38} + 0.01X_{40} + 0.17X_{43} - 0.63X_{47} - 0.73X_{48}}$$

The best model selection from LASSO and ridge regression is chosen based on the smallest MSEP value and the largest R-squared value. The MSEP value shows the size of the error, so the smaller the value, the better the model. Meanwhile, the R-squared value shows how much diversity is, so the larger this size, the better the model. The goodness measures of these two models can be presented in Table 3.

Table 3 The goodness measures

Model	MSEP	R-Squared
LASSO	0.34	0.976
Ridge Regression	3.61	0.750

Based on Table 3, The MSEP value shows the size of the error which so the smaller the value, the better the model. Where LASSO has an MSEP value of 0.34 which is smaller or can be said to be very good than the MSEP value of Ridge Regression which is 3.61. R-squared value shows how “good” the model is at explaining the data diversity. The range of R-squared values is from 0 to 1, where the closer the R-squared value is to 1, it means that the model is very good at explaining the data diversity. LASSO has an R-square value close to 1, which is 0.976 and largest than the R-square of Ridge Regression which is 0.750. thus, the LASSO model is better than ridge regression. This is in line with what [3] stated, that the coefficients in ridge regression will be depreciated towards zero, while in lasso regression some regression coefficients can be depreciated exactly to zero. The shrinkage of the coefficient exactly to zero is what makes lasso regression superior in terms of variable selection. The reduction of some variables results in a more efficient model interpretation.

Interpretation and discussion of its implementation based on the best model, the Lasso Regression, as follows:

1. Each one percent increase in the number of women aged 15-49 who have ever married and given birth in the last two years with the help of a midwife leads (X_1) will reduce HDI by 0.06. This means that regions with many births tend to have a low HDI, so to overcome this, it can be done by increasing the socialization of family planning programs to suppress the decline in HDI.
2. Each one percent increase in the number of Population with Health Insurance – BPJS Non-Contribution Assistance Recipients (Non-PBI) (X_3) will increase HDI by 0.69. This means that improving access to basic health services, especially for the non-poor, has a significant impact on improving HDI. Therefore, the government needs to ensure that all non-poor people are registered as non-PBI and the government can conduct annual evaluations.
3. Each one percent increase in the number of Population Who Went to Government Hospitals in the Last Month

(X_{10}), will increase HDI by 0.04. This means that the high number of visits indicates that people are starting to realize the importance of formal treatment and have more confidence in the services of government hospitals. Therefore, the government needs to increase service capacity, such as medical personnel, beds, medicines, and medical equipment.

4. Each one increase in the number of School Participation Rate (APS) Ages 19–23 (X_{15}) will increase HDI by 0.49. This age group is generally associated with higher education, such as colleges or universities. The government needs to build new PTNs in 3T areas, strengthen the KIP-College scholarship program, and expand distance education (PJJ) and socialization to high school graduates to be able to continue higher levels of study.
5. Each one percent increase in the number of Literate ≥ 15-Year-Old Population (X_{17}) will increase HDI by 0.23. This means that regions with high literacy rates tend to have a better quality of life, which can be improved by expanding access to basic education and illiteracy eradication programs for adults.
6. Each one percent increase in the number of Education Graduation Rate by Province and Level of Education (X_{19}) will increase HDI by 0.79. This means that the success of the population in completing formal education significantly impacts on improving the quality of life, so to improve it can be done by encouraging the continuation of education and reducing the dropout rate at every level.
7. Each one percent increase in the number of Children Aged 10–17 Years Working by Province (Percent) (X_{24}) will reduce HDI by 0.34. This means that the high rate of child labor has a negative impact on the quality of life of the community, so to overcome it can be done by strengthening supervision of child labor and increasing access to education and social protection for poor families.
8. Each one increase in the number of Access to Drinking Water Services (X_{28}) will reduce HDI by 0.07. This means that areas with high water service coverage do not necessarily have a good quality of life, so to overcome this can be done by improving the quality and sustainability of clean water services, not just the quantity.
9. Each one increase in the number of Access to Basic Sanitation Services (X_{29}) will increase HDI by 0.26. This means that areas with better basic sanitation tend to have a higher quality of life, so to improve it can be done by expanding the development of decent and safe basic sanitation infrastructure.
10. Each one increase in the number of Access to Basic Health Facilities (X_{30}) will increase HDI by 0.14. This means that regions with broad access to basic health services have a better HDI, so to improve it can be done by expanding the coverage of health facilities, especially in disadvantaged areas.
11. Each one increase in the number of Proportion of Households with Home Ownership Status (X_{31}) will reduce HDI by 0.04. This means that home ownership does not guarantee a good quality of life, so to overcome this can be done by improving the quality of houses owned by the community so that they are livable.
12. Each one percent increase in the number of Households That Use the Main Fuel for Cooking in the Form of Gas/LPG (X_{34}) will increase HDI by 0.44. This means that households using clean energy tend to have a better quality of life, which can be improved by expanding the energy conversion program to LPG evenly.
13. Each one percent increase in the number of Households with a Floor Area < 19 m² (X_{35}) will increase HDI by 0.70. This means that even if the house is small, it does not necessarily reflect a low quality of life, so to overcome this, it can be done by comprehensively reviewing housing eligibility indicators that consider basic facilities and residential density.
14. Each one percent increase in the number of Households by Province and Source of Decent Drinking Water (X_{38}) will increase HDI by 0.37. This means that access to safe drinking water is very important for improving the quality of life, so improving it can be done by expanding the distribution of safe and standardized drinking water.
15. Each one percent increase in the number of Households with Proper Sanitation (X_{40}) will increase HDI by 0.01. This means that proper sanitation still makes a positive contribution to HDI, albeit a small one, so increasing it can be done by ensuring all households have access to sanitation that meets health standards.
16. Each one increase in the number of Gini Ratio by Province and Region (X_{43}) will increase HDI by 0.17. This means that income inequality can occur in fast-growing regions, which can be addressed by keeping economic growth inclusive and equitable to all levels of society.
17. Each one increase in the number of Child Mortality Rate (CMR) (X_{47}) will reduce HDI by 0.63. This means that areas with high child mortality rates have a low quality of life, which can be addressed by improving maternal and child health services and strengthening nutrition and immunization interventions.
18. Each one increase in the number of Maternal Mortality Rate (MMR) (X_{48}) will reduce HDI by 0.73. This means that maternal mortality is an important indicator of the quality of life in a community, and can be addressed by improving access to safe delivery and strengthening reproductive health services.

V. CONCLUSIONS AND SUGGESTIONS

Based on the results and discussion that have been presented, the LASSO model is able to shrink the independent variables used in HDI modeling from 48 variables to 18 independent variables. The model formed is simpler. so that the interpretation of the model is more efficient by using only 18 predictor variables, namely Percentage of Women Aged 15–49 Years Who Have Been Married and Gave Birth in the Last Two Years with Midwifery Assistance ($\boldsymbol{X_1}$), Percentage of Population with Health Insurance – BPJS Non-Contribution Assistance Recipients (Non-PBI) ($\boldsymbol{X_3}$), Percentage of Population Who Went to Government Hospitals in the Last Month ($\boldsymbol{X_{10}}$), School Participation Rate (APS) Ages 19–23 ($\boldsymbol{X_{15}}$), Percentage of Literate ≥ 15-Year-Old Population ($\boldsymbol{X_{17}}$), Education Graduation Rate by Province and Level of Education ($\boldsymbol{X_{19}}$), Access to Drinking Water Services ($\boldsymbol{X_{28}}$), Access to Basic Sanitation Services ($\boldsymbol{X_{29}}$), Access to Basic Health Facilities ($\boldsymbol{X_{30}}$), Proportion of Households with Home Ownership Status ($\boldsymbol{X_{31}}$), Percentage of Households by Province and Source of Decent Drinking Water ($\boldsymbol{X_{38}}$), Percentage of Households with Proper Sanitation ($\boldsymbol{X_{40}}$), and Gini Ratio by Province and Region ($\boldsymbol{X_{43}}$). Based on the measure of model goodness, it can be concluded that LASSO is superior to ridge regression in terms of MSEP and R-squared measures. LASSO has a smaller MSEP of 0.34 than in Ridge Regression of 3.61. in addition, LASSO also has the largest R-squared value of 97.6%.

REFERENCES

[1] P. S. Lestari, S. Martha, and N. N. Debataraja, “Penerapan Metode Regresi Ridge Pada Kasus Angka Kematian Bayi Di Provinsi Jawa Timur,” 2022.

[2] H. A. Khoirunissa, A. R. Wijaya, B. Isnaini, and K. Ferawati, “Analisis Faktor-Faktor Penyebab Inflasi di Indonesia Menggunakan Regresi Ridge, LASSO dan Elastic-Net,” *Indones. J. Appl. Stat.*, vol. 7, no. 2, pp. 121–130, 2024, doi: 10.13057/ijas.v7i2.96921.

[3] F. Rahmawati and R. Y. Suratman, “Performa Regresi Ridge dan Regresi Lasso pada Data dengan Multikolinearitas,” *Leibniz J. Mat.*, vol. 2, no. 2, pp. 1–10, 2022, doi: 10.59632/leibniz.v2i2.176.

[4] A. A. H. Suruddin, E. Erfiani, and I. M. Sumertajaya, “The Continuum Regression Analysis with Preprocessed Variable Selection LASSO and SIR-LASSO,” *Inferensi*, vol. 8, no. 1, pp. 45–51, 2025, [Online]. Available: https://iptek.its.ac.id/index.php/inferensi/article/view/21658

[5] Badan Pusat Statistik, *Indeks Pembangunan Manusia 2023 Volume 18*. Badan Pusat Statistik/BPS-Statistics Indonesia, 2024. [Online]. Available: https://web-api.bps.go.id/download.php?f=GOvR/J1dHI5pfCpRUWdm2nNJRWdPQWd5amZZZVdzK3JCcmM3YVBjdk16aGRocnVaN1lCSTYvaitUYXdsYnFUNzViaXROeU5OMDBZcWJZbW1BNEI1ajMvcXhzT0pTblpOWEFwdXJMUWF2Wk9QSWU5R3I3MlozMXlZUVJuZzZJSGxhc1UzclNWUTR0YzNCak1rZWVxM2pLV0dZdkk0T21iSH

[6] M. Arif and M. Faisal, “Penerapan Model Regresi Linear Untuk Estimasi Mobil Bekas Menggunakan Bahasa Python,” *EULER*, vol. 11, no. 2, pp. 182–191, 2023, doi: 10.37905/euler.v11i2.20698.

[7] A. A. Azahra, “Analisis Prediksi Jumlah Penerimaan Mahasiswa Baru Menggunakan Metode Regresi Linier Sederhana,” *Bull. Appl. Ind. Eng. Theory*, vol. 3, no. 1, 2022.

[8] M. Sinanta P.W.J, “Prediksi Harga Mobil Menggunakan Linear Regression, Ridge Regression Dan Lasso Regression,” *J. Rev. Pendidik. dan Pengajaran*, vol. 8, no. 1, pp. 3066–3071, 2025, [Online]. Available: https://journal.universitaspahlawan.ac.id/index.php/nutrihealth

[9] R. Tibshirani, “Regression shrinkage and selection via the lasso: A retrospective,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 73, no. 3, 2011, doi: 10.1111/j.1467-9868.2011.00771.x.

[10] I. Sartika, N. N. Debataraja, and N. Imro’ah, “Analisis Regresi Dengan Metode Least Absolute Shrinkage And Selection Operator (Lasso) Dalam Mengatasi Multikolinearitas,” *BIMASTER*, vol. 9, no. 1, pp. 31–38, 2020, [Online]. Available: https://jurnal.untan.ac.id/index.php/jbmstr/article/view/38029/75676584327

[11] J. H. Lee, Z. Shi, and Z. Gao, “On LASSO for predictive regression,” *J. Econom.*, vol. 229, no. 2, 2022, doi: 10.1016/j.jeconom.2021.02.002.

[12] F. Li, L. Lai, and S. Cui, “On the Adversarial Robustness of LASSO Based Feature Selection,” in *Wireless Networks (United Kingdom)*, 2022. doi: 10.1007/978-3-031-16375-3_3.

[13] O. A. Montesinos López, A. Montesinos López, and J. Crossa, “Overfitting, Model Tuning, and Evaluation of Prediction Performance,” in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, 2022. doi: 10.1007/978-3-030-89010-0_4.

[14] C. Hong *et al.*, “LASSO-Based Identification of Risk Factors and Development of a Prediction Model for Sepsis Patients,” *Ther. Clin. Risk Manag.*, vol. 20, 2024, doi: 10.2147/TCRM.S434397.

[15] P. Hu, L. Chen, and Z. Zhou, “Machine Learning in the Differentiation of Soft Tissue Neoplasms: Comparison of Fat-Suppressed T2WI and Apparent Diffusion Coefficient (ADC) Features-Based Models,” *J. Digit. Imaging*, vol. 34, no. 5, 2021, doi: 10.1007/s10278-021-00513-7.

[16] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[17] L. Zhou, F. Koehler, D. J. Sutherland, and N. Srebro, “Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression,” *ACM / IMS J. Data Sci.*, vol. 1, no. 2, 2024, doi: 10.1145/3594234.

[18] S. K. Safi, M. Alsheryani, M. Alrashdi, R. Suleiman, D. Awwad, and Z. N. Abdalla, “Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19,” *Migr. Lett.*, vol. 20, no. 6, 2023, doi: 10.59670/ml.v20i6.3468.

[19] I. S. Dar, S. Chand, M. Shabbir, and B. M. G. Kibria, “Condition-index based new ridge regression estimator for linear regression model with multicollinearity,” *Kuwait J. Sci.*, vol. 50, no. 2, 2023, doi: 10.1016/j.kjs.2023.02.013.

[20] A. Tsigler and P. L. Bartlett, “Benign overfitting in ridge regression,” *J. Mach. Learn. Res.*, vol. 24, 2023, [Online]. Available:

https://www.jmlr.org/papers/volume24/22-1398/22-1398.pdf

[21] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean Squared Error, Deconstructed," *J. Adv. Model. Earth Syst.*, vol. 13, no. 12, 2021, doi: 10.1029/2021MS002681.

[22] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, 2021, doi: 10.7717/PEERJ-CS.623.