Jurnal Varian
Vol. 8, No. 2, April 2025, pp 151–164
e-ISSN: 2581-2017

Lasso Regression and Elastic Net in Analysing Factors Affecting the Open Unemployment Rate

Anita Mustikasari, Andi Daniah Pahrany
Universitas Negeri Malang, Malang, Indonesia

Article Info

Article history:

Received : 07-31-2024
Revised : 03-25-2025
Accepted : 07-23-2025

Keywords:

Elastic Net;
Lars Algorithm;
Lasso Regression;
Unemployment.

ABSTRACT

This study aims to identify and analyze the variables that affect the open unemployment rate in Banten Province, Indonesia. The analyzed variables include population density, average years of schooling, labor force participation rate, minimum wage, Provincial GRDP, total labor force, and the number of poor people. The method used in this study is multiple linear regression analysis with secondary data from the Central Bureau of Statistics (BPS) for the period 2017–2022. The analysis revealed multicollinearity in the average years of schooling variable, with a Variance Inflation Factor (VIF) > 10. To address this issue, Lasso regression and Elastic Net regression were applied. The results of this study show that Lasso regression produces a model with a Mean Squared Error (MSE) of 1.3234857. In contrast, Elastic Net regression yields a model with a lower MSE of 0.180683, indicating better predictive performance. The best model for predicting the open unemployment rate in Banten Province is the Elastic Net regression. The variables that significantly affect the open unemployment rate are average years of schooling, labor force participation rate, minimum wage, Provincial GRDP, total labor force, and the number of poor people. The conclusion of this study is that Elastic Net regression is more effective in predicting the open unemployment rate than other methods. The implication of these findings is that the generated model can serve as a basis for formulating more effective labor policies to reduce the unemployment rate in Banten Province.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v8i2.4332

Corresponding Author:

Andi Daniah Pahrany,
Department of Mathematics, Universitas Negeri Malang,
Email: andi.daniah.fmipa@um.ac.id

How to Cite:
Mustikasari, A., & Pahrany, A. D. (2025). Lasso Regression and Elastic Net in Analysing Factors Affecting the Open Unemployment Rate. *Jurnal Varian*, *8*(2), 151-164.

A. INTRODUCTION

Demographic situations pose a challenge for Indonesia. As a developing country, Indonesia has a large population. One of the issues currently focused on by the government is unemployment. Unemployment is not only a phenomenon that will have an impact on individual aspects, but will also have an impact on economic aspects (Septiana & Hasyim, 2025). One of the factors is Indonesia's large population which creates a new labor force every year and has an impact on the unemployment rate (Fatkhu Rokhim et al., 2023). Unemployment is caused by an imbalance between the number of job openings and the labor force (Adriyanto et al., 2020). Open unemployment refers to the labor force that is not employed or is actively seeking employment (Marini & Putri, 2019). Open

unemployment will always exist in any country. However, the problem arises when the open unemployment rate is high, as it impacts the welfare of society (Pohlan, 2024). By mid-2023, Indonesia's population is expected to increase rapidly to 278.69 million. With the increase in population, the demand for employment also rises. According to data from Trading Economics, Indonesia has the second-highest unemployment rate in Southeast Asia after Brunei Darussalam, at 5.45%. The Central Bureau of Statistics (BPS) recorded that the number of unemployed individuals in Indonesia reached 7.99 million. The highest open unemployment rate in Indonesia is found in Banten Province at 7.52%, followed by West Java Province at 7.44%, and Riau Islands Province at 6.8%. Currently, the government aims to achieve an open unemployment rate of 5% by 2024. Moreover, Banten is one of the provinces with the highest migration rates in Indonesia, which contributes to an increased labor force.

Previous research has indicated that factors influencing the open unemployment rate include the minimum wage, average years of schooling, and Gross Regional Domestic Product (GRDP) (Setiawan et al., 2023). Additionally, increased population density, the size of the labor force, and the workforce can also contribute to a higher open unemployment rate (Rusydan & Wijaya, 2024). A higher minimum wage can reduce the open unemployment rate. Furthermore, a higher level of education can also reduce the open unemployment rate, implying that an increase in the average length of schooling can lower the unemployment rate. According to Putra & Arka (2016), a lower number of poor people leads to a lower unemployment rate. A strong GRDP in a region also plays a significant role in reducing open unemployment. This is in line with Romhadhoni et al. (2019), who state that increasing GRDP can create new jobs. The gap between this study and previous research lies in the fact that, despite identifying various factors, the specific relationship between these variables in the context of open unemployment in Banten Province has not been thoroughly explained. While these factors may influence the open unemployment rate, this research does not provide a detailed explanation of how each factor interacts, particularly in Banten. Additionally, no research explicitly investigates the direct and significant impact of these factors on open unemployment in Banten between 2017 and 2020. Therefore, it is crucial to understand the relationship between these factors to formulate more effective policies.

The difference between this study and previous research is in its methodological approach. While previous studies primarily rely on ordinary linear regression methods, these methods often fail to address multicollinearity among correlated variables, which can affect the accuracy of the results. This study employs advanced regularization techniques, including Lasso and Elastic Net regression, which are specifically designed to handle multicollinearity and produce more reliable models. High multicollinearity results in the regression parameter estimators tending to have a large diversity (Saputro et al., 2025). Additionally, while previous research has mostly focused on national data or other regions, this study specifically examines open unemployment in Banten Province between 2017 and 2020, filling a gap in the literature. The objective of this study is to identify and analyze the variables that affect the open unemployment rate in Banten Province, Indonesia. By employing Lasso and Elastic Net regression, this study aims to improve prediction accuracy and identify the most significant factors contributing to open unemployment in the region. The contribution of this study is to provide deeper insights into the relationships between the variables affecting the open unemployment rate in Banten Province, particularly between 2017 and 2020. Additionally, this study contributes to the development of a more accurate and valid prediction model for open unemployment using advanced regularization techniques. These methods can help policymakers design more effective strategies to reduce the open unemployment rate in the province.

B. RESEARCH METHOD

This research employs a quantitative approach, providing an overview of the problem using numerical data. This approach aims to measure the relationship between various variables and identify patterns or trends within the data. Quantitative research allows for the use of statistical techniques to analyze and interpret numerical data, providing an objective basis for drawing conclusions and making predictions. The data used is secondary data sourced from the publication of the Central Bureau of Statistics for 2017-2022, obtained through the page https://banten.bps.go.id/. The data in this study are presented as follows:

a. Dependent Variable
   Y : Open Unemployment Rate of Banten Province (percent).
b. Independent Variable
   X_1 : Population Density of Banten Province (people), X_2 : Average Years of Schooling in Banten Province (years), X_3 : Participation Rate labour force Banten Province (percent), X_4 : Banten Province Minimum Wage (rupiah), X_5 : GRDP of Banten Province 2017-2022 (billion rupiah), X_6 : Total Labor Force of Banten Province 2017-2022 (people), X_7 : Number of Poor People 2017-2022 (people).

The data processing process in this study uses applications such as R Studio and Minitab 19. The steps of the data processing process in this study are as follows:

1. Descriptive Analysis

Descriptive analysis is the first step in the data processing process that presents and interprets the data. This aims to provide a detailed and clear description of the data.

2. Estimation of multiple linear regression coefficients using the Least Squares Method (MKT)

Multiple linear regression has a model, which is displayed in Equation (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{1}$$

One way to estimate the parameters $\beta 0,\ \beta 1,\ \beta 2,\ \ldots,\ \beta k$ in regression analysis is the Least Squares Method (MKT). The way MKT works is by minimizing the Sum of Squared Error (Auqino et al., 2019). Parameter estimates can be calculated using Equation (2).

$$\beta^{MKT} = \left(X^T X\right)^{-1} \left(X^T Y\right) \tag{2}$$

3. Classical Assumption Test

Classical Assumption Test consists of Residual Normality Test, Multicollinearity Test, Heteroscedasticity Test, and Autocorrelation Test. The data was standardized using Z-score to ensure the dataset has a uniform scale. Z-score standardizes data by setting the mean of data set to zero and the variance to one (Frans L et al., 2022). The Z-score formula can be seen in Equation (3).

$$Y_{new} = \frac{Y_{old} - \bar{Y}}{\sigma},\ X_{new} = \frac{X_{old} - \bar{X}}{\sigma} \tag{3}$$

The Y_{new} symbol is the standardized data value, the Y_{old} symbol is the original data value, the Y symbol is the average of the Y variable data, and the σ symbol is the standard deviation value of the Y or X variable data. The X_{new} symbol is the standardized data value, the X_{old} symbol is the original data value, and the X symbol is the average of the X variable data.

4. Lasso Regression

Lasso regression (least absolute shrinkage and selection operation) is one of the regression methods used to overcome multicollinearity. In 1966, Tibshirani introduced the Lasso for the first time. Lasso regression selects independent variables that affect the dependent variable. The LASSO approach trades off potential bias in estimating individual parameters for a better expected overall prediction (Ranstam & Cook, 2018). The way Lasso works is that the regression coefficient of the predictor variable with a high correlation to the error is reduced to zero or nearly zero. Lasso can be used on normally distributed continuous data (Fanny et al., 2018). The Bayesian adaptive lasso regression improves classical adaptive lasso by using a Gibbs sampler for tractable posteriors, addressing high-dimensional challenges without requiring OLS estimates (Mubasher et al., 2024). According to Andana et al. (2017), lasso has the following general Equation (4):

$$Y^{LASSO} = X\beta + \varepsilon \tag{4}$$

Y^{LASSO} is the dependent variable vector, X is the independent variable matrix, β is the lasso coefficient vector, and ε is the error vector. According to Hastie Trevor and Qian Junyang (2014), the lasso estimation process is as follows (Equation (5)):

$$\hat{\beta}^{lasso} = argmin \sum_{i=1}^{N} \left(y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{5}$$

with the criterion $\sum_{j=1}^{p} |\beta_j| \leq s$. The tuning parameter s value governs the shrinkage of the Lasso coefficients with $s \geq 0$ and $s = \frac{t}{\sum_{j=1}^{k} \left|\hat{\beta}_j^0\right|}$, $\hat{\beta}_j^0$ is the MKT estimator, $t = \sum_{j=1}^{k} \left|\widehat{\beta}_j\right|$, β_j is the regression coefficient on Lasso (Andana et al., 2017). A small s value causes the coefficient to be zero, so that influential variables will be included in the model (**NurfitriImroah2020**). The s value is obtained from the cross-validation process. The value $s < t$ causes the coefficient to be almost zero or exactly zero. Equation (6) can be written in the form of a Lagrange equation as follows:

$$\hat{\beta}^{lasso} = argmin \sum_{i=1}^{N} (y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda |\beta| I \tag{6}$$

The estimation of $\hat{\beta}^{lasso}$ can be used the following Equation (7):

$$\hat{\beta}^{lasso} = Sign(\hat{\beta}_j^0)\left(\hat{\beta}_j^0 - \frac{\lambda}{2}\right) \tag{7}$$

The steps of the LARS algorithm are as follows (Restu Ningsih et al., 2023):

1. Standardize the data so that it has a mean of 0 and a variance of 1.
2. Initial estimation of regression coefficient $\beta = 0$.
3. Adding the highest correlated predictor variable to the model.
4. Removing the predictor variable if the coefficient is zero and then recalculating the least squares coefficient.
5. Perform iterations on all variables until they are included in the lasso regression model.

To obtain the optimal model, cross-validation is used as an assessment method. Cross validation approach used to estimate prediction error and improve accuracy in model selection (Andana et al., 2017). One of the cross-validation techniques is k-folds with $k = 5$ or $k = 10$ to produce cross-validation with high bias and low (Frans L et al., 2022).

5. Elastic Net Regression

Elastic Net regression is one of the regressions to overcome the existence of assumption deviations, namely multicollinearity. Elastic Net regression is a combination of ridge regression and lasso regression. This method can overcome multicollinearity by selecting coefficients. Elastic Net regression has the advantage of ridge and lasso regression, namely greater flexibility in controlling the alpha parameter. It can also overcome the problem of underfitting and overfitting (Nur et al., 2023). Elastic Net regression is able to overcome the shortcomings of ridge regression and lasso regression. Lasso regression only selects the most correlated variables from the group and ignores other variables. The elastic net with α set to 0 is equivalent to ridge regression (Waldmann et al., 2013). Elastic Net selects variables simultaneously and the correlated variables are selected (Vilanasari, 2023). According to Hastie Trevor and Qian Junyang (2014), Elastic Net has the following penalty:

$$L = \alpha|\beta|_1 + (1-\alpha)|\beta|^2 \tag{8}$$

with α being the penalty lasso and $(1-\alpha)$ being the penalty ridge. Elastic Net has the following equation:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2 + \lambda_2 \sum_{j=1}^{p}\beta_j^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| \tag{9}$$

With $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$, $0 \leq \alpha \leq 1$, the parameter estimator can be used as follows:

$$\beta^{net} = \arg\min \sum_{i=n}^{n}\left[\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2 + \lambda\left((1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\left(\sum_{j=1}^{p}|\beta_j|\right)\right)\right] \tag{10}$$

After obtaining the parameter estimator β, the Elastic Net regression model is obtained as follows: obtaining the parameter estimator β, the Elastic Net regression model is obtained as follows:

$$\bar{Y}^{net} = \beta_0 + \beta_1\bar{X}_1 + \ldots + \beta_{p-1}\bar{X}_{p-1} + \varepsilon \tag{11}$$

$\bar{Y}$ is the dependent variable, $\beta 0$ is the constant, $\beta 0$ is the regression coefficient, and ε is the error. The step to estimate the Elastic Net regression parameters is as follow (Handayani & Wachidah, 2023) :

1. Distribution of testing data and training data with a ratio of 20%: 80%.
2. Cross-validation was performed to determine the optimal λ and α
3. Specify variables independent variables that contribute to variable dependent variable by using variable importance.
4. Perform parameter estimation based on the λ and α values obtained.
5. Test for multicollinearity in the Elastic Net model.

6. The Best Selection

The criteria for selecting the best model can be seen in the mean square error (MSE) value. MSE is one of the methods to measure the accuracy of the model by looking at the error rate. The smaller the MSE value, the more accurate the prediction and model error. The calculation of the MSE value uses the following formula (Katemba & Djoh, 2017) :

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y - \hat{Y}\right)^2 \tag{12}$$

Where Y is the observation data, while $\hat{Y}$ is the prediction of the observation data, and n is the amount of data. A good model has a small MSE value.

C. RESULT AND DISCUSSION

1. Descriptive Analysis

From the data that has been collected, a descriptive analysis is obtained as in Table 1.

Table 1. Descriptive Analysis

Variable	Total Count	Mean Deviation	Standard	Minimum	Maximum
Open Unemployment Rate (Y)	48	8,959	1,866	4,670	13,060
Population Density (X_1)	48	4,201	4,561	376	14,486
Average Years of Schooling (X_2)	48	8,781	1,800	6,200	11,840
Participation Rate labor force (X_3)	48	63,331	2,458	57,020	69,970
Minimum Wage (X_4)	48	3,511,004	675,949	2,127,112	4,340,254
GRDP (X_5)	48	68,625	67,442	19,009	268,165
Total Labor Force (X_6)	48	67.722	48.436	18.562	239.788
Poor Population (X_7)	48	92.67	65.77	13.20	272.35

Based on Table 1, the highest open unemployment rate is 13.06% in Serang Regency, and the lowest is 4.67% in South Tangerang City. The highest population density is 14,486 people per square kilometer in Tangerang City, and the lowest is 376 people per square kilometer in Lebak Regency. The highest average years of schooling in Banten Province is 11.840, or equivalent to senior high school, and the lowest is 6.2, or equivalent to elementary school. The maximum Participation rate of the labor force is 69.97% and the minimum is 57.02%. The average minimum wage in Banten Province is Rp. 3,511,004 with an average GRDP of Rp. 68,625 billion. The maximum number of labor force is 239,788 million people in 2020, and the minimum number of labor force is 18.562 million people in 2018. The largest number of poor people is 272.35million in Tangerang Regency, and the smallest number of poor people is 13.20 million in Cilegon City.

2. Multiple Linear Regression Coefficient Estimation

Based on the data obtained, multiple linear regression analysis was conducted to determine the relationship between variables. To estimate the coefficient of multiple linear regression, the least squares method (MKT) can be used. The coefficient estimation results using Minitab 19 software are as follows:

Table 2. Multiple Linear Regression Coefficients

Variables	Coefficient
Constant	27.33
Population Density (X_1)	0.000035
Average Years of Schooling (X_2)	-1.242
Participation Rate labor force (X_3)	-0.1929
Minimum Wage (X_4)	0.000001
GRDP (X_5)	0.000013
Total Labor Force (X_6)	0.000041
Poor Population (X_7)	-0.0247

Table 2 shows the multiple linear regression results with a constant of 27.33, which represents the base value when all independent variables are zero. The coefficients of population density, minimum wage, GRDP, and total labor force are positive,

indicating that an increase in these variables tends to increase the open unemployment rate, although the effect is very small. In contrast, the average years of schooling, labor force participation rate, and the number of poor people have negative coefficients, which means that increases in these variables actually contribute to lowering the unemployment rate. This finding suggests that education and active participation in the labor market play an important role in reducing unemployment.

3. Classical Assumption Test

After obtaining the regression coefficient, a classical assumption test is performed to verify the assumptions that must be met. The assumption test includes:

a. Residual Normality Test

Based on the plot image of the normality test, the $P - Value > 0.05$ is obtained. which means the data is normally distributed. Thus, fulfilling the residual normality test.

b. Multicollinearity Test

1) By using Equation (4), the correlation value between variables is obtained as follows:

Table 3. Correlation Value

	Y	X_1	X_2	X_3	X_4	X_5	X_6
X_1	**-0.507**						
X_2	-0.375	0.831					
X_3	-0.004	-0.105	-0.291				
X_4	0.164	0.424	0.662	0.019			
X_5	0.208	0.074	0.434	-0.076	0.46		
X_6	0.282	0.125	-0.049	0.299	0.288	-0.316	
X_7	**0.067**	-0.078	-0.349	0.436	-0.077	-0.468	0.831

Based on Table 3, there is a significant difference between the correlation value and the regression coefficient for variables X_1 and X_7, indicating potential multicollinearity in the model. The correlation value of X_1 is -0.507, but the regression coefficient is positive at 0.000035, while X_7 has a positive correlation of 0.067, but the regression coefficient is negative at -0.0247. This difference in direction indicates that the independent variables are correlated and can affect the contribution of each variable to the dependent variable in the regression model. This multicollinearity needs to be considered because it can compromise the accuracy of the model's estimation and interpretation.

2) By using Equation (5), the value of each independent variable is obtained as follows.

Table 4. Multiple Linear Regression VIF Value

Variable	VIF
Population Density	6.64
Average Years of Schooling	11.88
Participation Rate labor force	1.65
Minimum Wage	3.74
GRDP	2.22
Total Labor Force	5.98
Poor Population	5.99

Based on Table 4, This VIF (Variance Inflation Factor) table shows the level of multicollinearity between independent variables in the regression model. A high VIF value indicates a strong correlation between variables that can affect the stability of the model. From the table, it can be seen that the average years of schooling variable (X_2) has the highest VIF value of 11.88, exceeding the general limit of 10, which indicates the potential for high multicollinearity and needs to be analyzed further. While the other variables have VIF values below 10, which is still within the tolerance limit, although some such as population density (X_1), total labor force (X_6), and number of poor people (X_7) have values close to 6, indicating a moderate correlation that still needs to be watched out for. Overall, the

DOI: https://doi.org/10.30812/v8i2.4332

model is quite stable, but special attention needs to be paid to variable X_2 to avoid distorting the regression results.

3) Heteroscedasticity Test Based on the versus of fits plot which can be seen in normality test, the dots spread out not forming a certain pattern. So, it can be said that it does not contain heteroscedasticity.

4) Autocorrelation Test

Based on the Durbin Watson test, the DW statistical value is $-2 < 1.39 < 2$, which means there is no autocorrelation.

4. Data Standardization

Data is standardized using Equation (6) for the dependent variable and the independent variable. Standardization aims to equalize data units. The data is standardized until the mean value is zero and the variance is one.

5. Lasso Regression

After standardized data is obtained, estimation will then be carried out for the estimator of the lasso coefficient with the LARS algorithm. Estimation of the LARS algorithm lasso regression coefficient using R Studio software. The following are the stages of independent variables that enter the model.

Table 5. LARS Sequence

Var	X_1	X_6	X_4	X_5	X_3	X_2	X_7	$-X_1$	X_1
Step	1	2	3	4	5	6	7	8	9

Table 5 shows the order of variable selection at each step in the LARS (Least Angle Regression) regression process. In the first step, the variable X_1 (Population Density) comes first, followed by X_6 (Total Labor Force), X_4 (Minimum Wage), and so on until X_7 (Number of Poor People) in the 7$^{\text{th}}$ step. Interestingly, in the 8th and 9th steps, there is a re-selection of X_1 with a negative sign ($-X_1$) and then re-entry as X_1, which reflects a shift in the direction of the variable's contribution in the model. This shows that LARS dynamically adjusts the selection and direction of variable influence based on partial correlation when new variables enter the model.

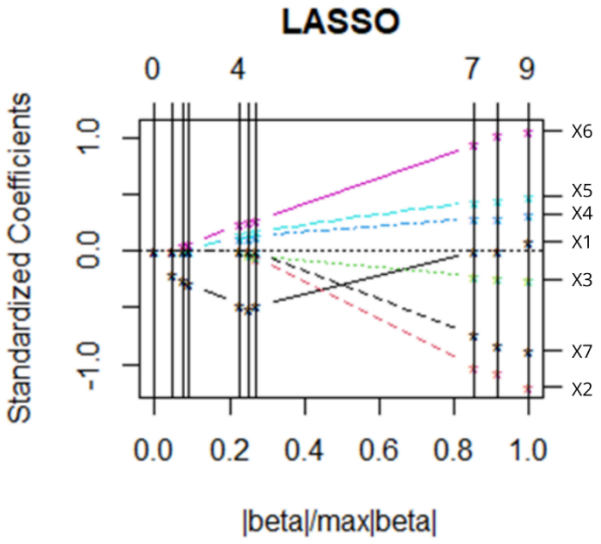

Figure 1. Lasso LARS Plot

Figure 1 is a plot of the stages of variables entering the model. The first variable entered into the model is population density (X_1) because it has the highest correlation value of 0.507. Furthermore, the number of labor force variables (X_6), minimum wage variables minimum (X_4) and so on. The variable X_1 is also the last variable to enter the model. The next step is estimating the coefficients at each iteration step. The following presents the estimated coefficients at each stage:

Table 6. LARS Sequence

Step	X_1	X_2	X_3	X_4	X_5	X_6	X_7	$t = \sum \lvert\widehat{\beta_j}\rvert$	$s = \frac{\sum \lvert\widehat{\beta_j}\rvert}{\max \sum \lvert\widehat{\beta_j}\rvert}$
1	0	0	0	0	0	0	0	0	0
2	-0.1999	0	0	0	0	0		0.1999	0.046597
3	-0.2574	0	0	0		0.5744	0	0.8318	0.193893
4	-0.2870	0	0	0.0257	0	0.0707	0	0.3834	0.089371
5	-0.4783	0	0	0.1111	0.1512	0.2406	0	0.9811	0.228695
6	-0.5035	0	-0.0240	0.1195	0.1706	0.2688	0	1.0864	0.25324
7	-0.4772	-0.0520	-0.0415	0.1395	0.1868	0.2725	0	1.1689	0.272471
8	0	-1.0235	-0.2216	0.3002	0.4377	0.9483	-0.7378	3.6691	0.855268
9	0	-1.0736	-0.2332	0.3035	0.4539	1.0390	-0.8283	3.9315	0.916434
10	0.0859	-1.1985	-0.2541	0.3292	0.4829	1.0700	-0.8707	4.29	1

Based on Table 6, the first step in the lasso regression model starts by setting all initial coefficients to zero. The population density variable (X_1) is the first variable to enter the model. Next, the variable number of labor force (X_6), minimum wage (X_4), and the process continues until all variables enter the model. The value of s is a tuning parameter that shrinks the coefficients towards almost zero or exactly obtained at a value of $s < t$ so that the coefficient will shrink to almost zero or exactly zero. In determining the optimal model, cross validation is carried out, k-fold cross validation is applied using fraction mode and step mode. The cross validation in fraction mode used is 5-fold with using the LARS algorithm package in R Studio software.

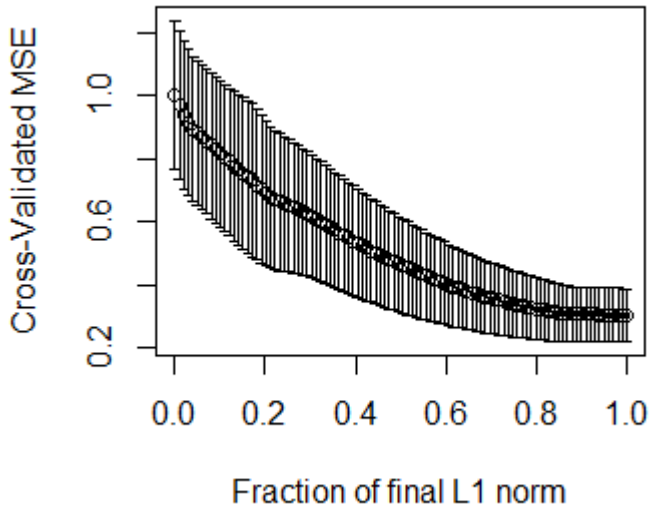

Figure 2. Mode Fraction LARS

Based on Figure 2, the mode fraction performs calculations on cross validation with the s value obtained from $\sum \lvert\check{\beta}_j\rvert \ / \ max \ \sum \lvert\check{\beta}_j\rvert$. In Figure 2, it can be seen that the greater the value of s, the smaller the CV MSE value is 0,0827 at s=1 in ten steps. However, at that step all variables are included in the model so that multicollinearity has not been resolved because variables have not been selected. So, it is necessary to cross-validate the step mode. Step mode aims to determine the best model from several step iterations that have been carried out. Step mode is done by calculating the cross-validation value of each stage. The selection of the lasso model is determined by the minimum CV MSE value generated in cross validation.

Based on Figure 3, it can be seen that steps 8, 9, and 10 provide almost the same CV MSE value. However, at step 9 the smallest CV MSE value is 0.04552. So, the best model using step mode is at the ninth step. In the ninth step, six variables enter in the model, namely the average length of schooling (X_2), Participation rate labor force (X_3), minimum wage (X_4), GRDP Provincial (X_5), number of labor force (X_6), number of poor people (X_7).

The lasso regression equation using the LARS algorithm is as follows:

$$\check{Y}^{LASSO} = -1,0736\ X_2 - 0,2332\ \ X_3 + 0,3035\ X_4 + 0,4539\ \ X_5 + \ 1,0391\ \ X_6 - 0,8283\ X_7$$

DOI: https://doi.org/10.30812/v8i2.4332

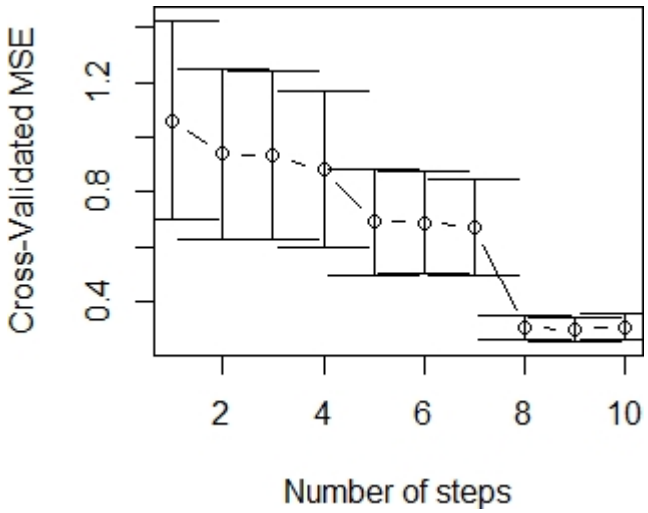

Figure 3. Step mode LARS

The lasso equation above has an R-square value of 80.34%. This value means that the dependent variable is influenced by the independent variable by 80.34%. While the rest is influenced by other variables. Furthermore, multicollinearity is checked by looking at the VIF value of each independent variable.

Table 7. VIF Value of Lasso Regression

Variables	VIF
X_2	2.3018
X_3	1.4240
X_4	3.2126
X_5	1.7851
X_6	5.9148
X_7	5.9826

Based on Table 7 about the multicollinearity test in the new lasso equation, the value of $VIF < 10$ for all independent variables. Therefore, multicollinearity has been resolved. In addition, in the ANOVA analysis, each variable shows a value of $p - value < 0.05$. This shows that the average length of school variable (X_2), Participation rate labor force (X_3), minimum wage (X_4), Provincial GRDP (X_5), total labor force (X_6), number of poor people (X_7) have a significant effect on the open unemployment rate (Y).

The increase in the open unemployment rate in Banten Province is caused by the minimum wage (X_4), Provincial GRDP (X_5), the number of labor force (X_6) assuming other variables are considered constant. Increase in wages 1 point, causes an increase in the unemployment rate by 0.3035. Similarly, an increase in the Provincial GRDP by 1 point will result in an increase in the unemployment rate by 0.4539. In addition, if the number of labor force increases by 1 point, the unemployment rate will increase by 1.0391. This is in line with Astuti & Setyonaluri (2022) which states that the high labor force can increase the open unemployment rate, along with the increasing number of years the number of labor force will continue to grow, therefore the number of labor force can increase the open unemployment rate.

On the other hand, the decrease in the open unemployment rate is due to the average length of schooling (X_2), Participation rate labor force (X_3), and the number of poor people (X_7 decreasing. The increase in the average length of schooling by 1 point (X_2) can reduce the open unemployment rate by 1.0736. Increase in Participation rate labor force (X_3) by 1 point, can reduce the open unemployment rate by 0.2332. An increase of 1 point in the number of poor people (X_7) can reduce the open unemployment rate by 0.8283. This happens because many poor people work but underpaid. Poor people do not mean unemployment. The poor themselves are people whose expenditure is below the poverty line, which is 433 thousand per month. In fact, many poor people work even though the salary earned is below the Banten minimum wage. So that the number of poor people can reduce the open unemployment rate in Banten.

6. Elastic-Net Regression

Data that has been standardized is then divided into data. The data is divided into 20% testing data and 80% training data. This study uses a total of 48 data so that the training data is 38 data, and testing data is 10 data. To create a regression model, the optimal lambda and alpha values are needed. Therefore, a 10-fold cross validation was conducted to find lambda and alpha by considering the smallest RMSE and MAE values and the largest R-square. The following table presents the results of cross-validation.

Table 8. Cross Validation

α	λ	RMSE	R-Square	MAE
0.1753148	0.005074640	0.5524624	0.7793739	0.4675607
0.2945777	0.002461247	0.5541742	0.7759618	0.4683379
0.3583630	0.002357187	0.5541656	0.7759101	0.4682940
0.5054517	0.002017541	0.5543277	0.7755304	0.4683974
0.5349936	0.480098132	0.8651898	0.6399541	0.7062313
0.5488174	0.007613340	0.5495719	0.7839918	0.4653399
0.6714276	0.024560942	0.5718544	0.7840191	0.4934210
0.6842647	**0.005600899**	**0.5495462**	**0.7822873**	**0.4650407**
0.9666141	0.312619571	0.8800621	0.5763280	0.7257001
0.9925088	0.303496851	0.8781396	0.5798385	0.7239450

Based on Table 8, the most optimal α value is 0.6842647. While the most optimal λ is 0.005600899. After obtaining the value of α and λ, then find the value of variable importance with using the "`varImp()`" syntax in R Studio. The variable importance value is used to determine the variables that influence the open unemployment rate. The value is seen in the overall value where the greater the overall value means the more influential. The following table presents the variable importance.

Table 9. Importance Variable

Variable	Overall
Total Labor Force (X_6)	100
Average Years of Schooling (X_2)	98.9
Number of Poor People (X_7)	75.12
GRDP Provincial (X_5)	48.44
Minimum Wage (X_4)	28.45
Labor Force Participation Rate (X_3)	17.34
Population Density (X_1)	0

Based on Table 9, the variable Total Labor Force (X_6) has the largest overall value of 100 so that the variable has the most influence on the open unemployment rate. While the population density variable (X_1) has the smallest overall value of 0 so that the variable has no effect to the open unemployment rate. Next, we estimate the coefficient parameters using Elastic Net regression. The coefficient estimation is done using R Studio. The following table presents the Elastic Net regression coefficient parameter estimation.

Table 10. Estimated Elastic Net Coefficient

Variable	Coefficient
Constant	0.01128497
Population Density (X_1)	0
Average Years of Schooling (X_2)	-1.08406636
Participation Rate labour force (X_3)	-0.19561554
Minimum Wage (X_4)	0.31363881
GRDP (X_5)	0.52685360
Total Labor Force (X_6)	1.03309249
Poor Population (X_7)	-0.83272081

Based on Table 10, it can be seen that the coefficient shrinks to zero. Variable X_2, X_3, X_4, X_5, X_6, dan X_7 enter the model and have a significant effect on the dependent variable. While X_1 is not included in the model because the parameter estimate is

0, and the overall value is 0. Based on the coefficient parameter estimate, the Elastic Net equation is obtained as follows:

$$\check{Y}^{NET} = 0,0112 - 1,0841X_2 - 0,1956\ X_3 + 0,3136X_4 + 0.5268\ X_5 + 1,0331\ X_6 - 0,8327X_7$$

A positive coefficient parameter estimate value will increase the dependent variable. If the independent variable increases by one point, the open unemployment rate will increase. Meanwhile, a negative parameter estimate will decrease the dependent variable. If the independent variable increases by one point, the open unemployment rate will decrease. Next, multicollinearity is tested in the new model. The following is the VIF value of each independent variable.

Table 11. VIF Value of Elastic Net Regression

Variable	Nilai VIF
Average Years of Schooling (X_2)	2.301
Participation Rate labour force (X_3)	1.424
Minimum Wage (X_4)	3.212
GRDP (X_5)	1.785
Total Labor Force (X_6)	5.914
Poor Population (X_7)	5.982

Based on Table 11, the VIF value of each variable is less than 10. This indicates that the multicollinearity case has been resolved. However, in the poor population variable (X7), there are still differences in sign. According to Lindner et al. (2020), the difference in sign does not always indicate multicollinearity. Therefore, it can be concluded that the multicollinearity case has been resolved. The R-square value in the testing data shows that about 52.4% of the dependent variable can be explained by the independent variable. Meanwhile, the rest is influenced by other variables that are not included in the model.

7. Best Model Selection

Table 12. Best Model Selection

	MSE
LASSO	1.3234857
ELASTIC NET	0.180683

Based on Table 12, the MSE value of Elastic Net is smaller than that of Lasso regression. It can be said that the best model for overcoming multicollinearity in the open unemployment rate is Elastic Net regression. The finding of this study is that Elastic Net regression is more accurate in predicting the factors affecting open unemployment. The results of this study are in line with or supported by Altelbany (2021) and Kayanan & Wijekoon (2019), which states that Elastic Net regression is better than Lasso regression in overcoming multicollinearity due to the flexibility of Elastic Net.

A comparison between the results of this study and those of previous studies that only used Lasso regression reveals significant differences. The previous study successfully overcame multicollinearity using Lasso; however, this study employs a combination of Lasso and Elastic Net, which is more effective in handling correlations between more complex variables. The use of these two methods resulted in a more accurate and stable model, with a lower MSE compared to previous studies. Elastic Net regression, as an extension of Lasso, is more flexible in dealing with multicollinearity and produces a more robust and valid model. The main findings of this study indicate that the addition of Elastic Net enhances prediction accuracy and provides more precise insights into identifying factors that affect open unemployment.

D. CONCLUSION AND SUGGESTION

The best model for predicting the open unemployment rate in Banten Province uses Elastic Net regression. Multicollinearity has been resolved, as evidenced by VIF values less than 10 and coefficients converging to zero. Factors that influence the open unemployment rate in Banten Province are the number of labour force (X_6), average years of schooling (X_2), the number of poor people (X_7), GRDP Provincial (X_5), minimum wage (X_4), and participation rate in the labour force (X_3). The factor that most influences the open unemployment rate in Banten Province is the size of the labour force. Meanwhile, population density is a factor that does not affect the open unemployment rate. The labour force size is the most significant factor affecting the open unemployment

rate. Population density does not have an impact. The government should focus on enhancing the labour force through training or capital support, especially for job seekers. Future research could explore alternative methods to address multicollinearity and incorporate additional variables that may influence the open unemployment rate. This research is expected to make a significant contribution by applying Lasso and Elastic Net regression techniques to identify factors that influence the open unemployment rate in Banten Province. Additionally, Elastic Net regression is found to be more accurate than Lasso, as it produces the smallest MSE, offering a better prediction model. This research is also expected to provide deeper insight into the interaction between these factors, which can support policies aimed at reducing unemployment in Banten.

ACKNOWLEDGEMENT

We would like to thank Universitas Negeri Malang and all parties contributing to this research.

DECLARATIONS

AUTHOR CONTIBUTION

All authors contributed to this manuscript, from exploring ideas to writing this article.

FUNDING STATEMENT

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

REFERENCES

Adriyanto, A., Prasetyo, D., & Khodijah, R. (2020). Angkatan Kerja dan Faktor yang Mempengaruhi Pengangguran. *Jurnal Ilmu Ekonomi & Sosial*, *11*(2), 66–82. https://doi.org/10.35724/jies.v11i2.2965

Altelbany, S. (2021). Journal of Applied Economics and Business. *Journal of Applied Economics and Business*, *5*(1), 131–142. https://doi.org/https://doi.org/10.34260/jaebs.517Journal

Andana, A. P., Safitri, D., & Rusgiyono, A. (2017). Model regresi menggunakan least absolute shrinkage and selection operator (lasso) pada data banyaknya gizi buruk kabupaten/kota di Jawa Tengah. *Jurnal Gaussian*, *6*(1), 21–30.

Astuti, N. F. V., & Setyonaluri, D. (2022). Labor Market Outcomes of Vocational High Schools (SMK) and General High Schools (SMA) during the COVID-19 Pandemic. *The Journal of Indonesia Sustainable Development Planning*, *3*(3), 278–293. https://doi.org/10.46456/jisdep.v3i3.328

Fanny, R., Djuraidah, A., & Alamudi, A. (2018). Pendugaan Produktivitas Bagan Perahu dengan Regresi Gulud, LASSO dan Elastic-net. *Xplore: Journal of Statistics*, *2*(2), 7–14. https://doi.org/10.29244/xplore.v2i2.89

Fatkhu Rokhim, Novianti, T., & Anggraeni, L. (2023). Factors Influencing Unemployment in Indonesia. *Journal of Scientific Research, Education, and Technology (JSRET)*, *2*(1), 122–131. https://doi.org/10.58526/jsret.v2i1.51

Frans L, O., Rizki, S. W., & Kusnandar, D. (2022). Pemodelan Pertumbuhan Ekonomi Kalimantan Barat Menggunakan Pendekatan Least Absolute Shrinkage And Selection Operator (LASSO). *Bimaster: Buletin Ilmuah Math, Stat dan Terapannya*, *11*(1), 111–120.

Handayani, A., & Wachidah, L. (2023). Metode Regresi Elastic-net untuk Mengatasi Masalah Multikolinearitas pada Kasus Tingkat Pengangguran Terbuka di Provinsi Jawa Barat. *Bandung Conference Series: Statistics*, *3*(1), 66–72. https://doi.org/10.29313/bcss.v3i1.5757

Hastie Trevor and Qian Junyang. (2014). Glmnet Vignette, 1–42.

Katemba, P., & Djoh, R. K. (2017). Prediksi Tingkat Produksi Kopi Menggunakan Regresi Linear. *Jurnal Ilmiah Flash*, *3*(1), 42. https://doi.org/10.32511/flash.v3i1.136

Kayanan, M., & Wijekoon, P. (2019). Performance of LASSO and Elastic net estimators in Misspecified Linear Regression Model. *Ceylon Journal of Science*, *48*(3), 293. https://doi.org/10.4038/cjs.v48i3.7654

Lindner, T., Puck, J., & Verbeke, A. (2020). Misconceptions about multicollinearity in international business research: Identification, consequences, and remedies. *Journal of International Business Studies*, *51*(3), 283–298. https://doi.org/10.1057/s41267-019-00257-1

Marini, L., & Putri, N. T. (2019). Peluang Terjadinya Pengangguran di Povinsi Bengkulu : Sebererapa Besar? *1*(1), 70–83.

Mubasher, S., Zakria, M., Shahzad, A., Ali, N., & Faisal, H. (2024). Dynamic Ridge Regression vs. Lasso Regression: A Comparative Study for Modeling Pakistan's Unemployment Rate. *Global Journal of Mathematics and Statistics*, (November 2024), 21–45. https://doi.org/10.61424/gjms

Nur, A. R., Jaya, A. K., & Siswanto, S. (2023). Comparative Analysis of Ridge, LASSO, and Elastic Net Regularization Approaches in Handling Multicollinearity for Infant Mortality Data in South Sulawesi. *Jurnal Matematika, Statistika dan Komputasi*, *20*(2), 311–319. https://doi.org/10.20956/j.v20i2.31632

Pohlan, L. (2024). Unemployment's long shadow: the persistent impact on social exclusion. *Journal for Labour Market Research*, *58*(1). https://doi.org/10.1186/s12651-024-00369-8

Putra, I. K. A. A., & Arka, S. (2016). Analisis Pengaruh Tingkat Pengangguran Terbuka, Kesempatan Kerja, Dan Tingkat Pendidikan Terhadap Tingkat Kemiskinan Pada Kabupaten / Kota Di Provinsi Bali. *EP Unud*, *7*(3), 416–444.

Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Romhadhoni, P., Faizah, D. Z., & Afifah, N. (2019). Pengaruh Produk Domestik Regional Bruto (PDRB) Daerah terhadap Pertumbuhan Ekonomi dan Tingkat Pengangguran Terbuka di Provinsi DKI Jakarta. *Jurnal Matematika Integratif*, *14*(2), 113. https://doi.org/10.24198/jmi.v14.n2.19262.113-120

Rusydan, R. M., & Wijaya, R. S. (2024). Impact of Economic Growth, Minimum Wage, Labor Force Participation Rate, and Population Size on the Open Unemployment. *Journal of Business Management and Economic Development*, *2*(03), 1186–1198. https://doi.org/10.59653/jbmed.v2i03.911

Saputro, D. R. S., Wahyu, N. L., & Widyaningsih, Y. (2025). Performance of Ridge Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Net in Overcoming Multicollinearity. *Journal of Multidisciplinary Applied Natural Science*, *5*(2), 370–382. https://doi.org/10.47352/jmans.2774-3047.251

Septiana, D., & Hasyim, S. (2025). The Effect of Labor Force , Minimum Wage , and Per Capita Income on Unemployment Rate in Five Southeast Asian Countries TALENTA Conference Series The Effect of Labor Force , Minimum Wage , and Per Capita Income on Unemployment Rate in Five Southeast Asian. *8*(2). https://doi.org/10.32734/lwsa.v8i1.2409

Setiawan, K., Haikal, M., Wicaksana, A. G., & Dermawan, D. (2023). Analisis Faktor-Faktor Yang Mempengaruhi Tingkat Pengangguran Terbuka Di Provinsi Banten 2017-2021. *Jurnal Riset Rumpun Ilmu Ekonomi*, *2*(1), 107–120. https://doi.org/10.55606/jurrie.v2i1.1112

Vilanasari, V. (2023). *Penerapan Rgresi Elastic Net Pada Faktor Yang Mempengaruhi Indeks Ketimpangan Gender Di Indonesia* [Doctoral dissertation]. Universitas Negeri Malang.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, *4*(DEC), 1–11. https://doi.org/10.3389/fgene.2013.00270

[This page intentionally left blank.]

Vol. 8, No. 2, April 2025, pp 151–164
DOI: https://doi.org/10.30812/v8i2.4332