JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 6, No. 3, July 2022, pp. 530-543

Comparing Five Machine Learning-Based Regression Models for Predicting the Study Period of Mathematics Students at IPB University

Sri Nurdiati[1], Mohamad Khoirun Najib[2]
[1,2]Department of Mathematics, IPB University, Indonesia
nurdiati@apps.ipb.ac.id[1], mkhoirun_najib@apps.ipb.ac.id[2]

Article History:
Received : 14-04-2022
Revised : 16-05-2022
Accepted: 20-05-2022
Online : 16-07-2022

Keywords:
Cross-Validation;
Hyperparameter;
Machine Learning;
Multiple Linear Regression;
Tuning Constant

ABSTRACT

Grade point average (GPA) is initial information for supervisors to characterize their supervised students. One model that can be used to predict a student's study period based on GPA is a machine learning-based regression model so that supervisors can apply the right strategy for their students. Therefore, this study aims to implement and select a machine learning-based regression model to predict a student's study period based on GPA in semesters 1-6. Several regression models used are least-square regression, ridge regression, Huber regression, quantile regression, and quantile regression with l_2-regularization provided by Machine Learning in Julia (MLJ). The model is evaluated and selected based on several criteria such as maximum error, RMSE, and MAPE. The results showed that the least-square regression model gave the worst evaluation results, although the calculation method was easy and fast. Meanwhile, the quantile regression model provided the best evaluation results. The quantile regression model without regularization gives the smallest RMSE (2.31 months) and MAPE (3.56%), while the quantile regression model with l_2-regularization has a better maximum error (4.9 months). The resulting model can be used by supervisors to predict the study period of their supervised students so that supervisors can characterize their students and can design appropriate strategies. Thus, the student's study period is expected to be accelerated with a high-quality final project.

Crossref
https://doi.org/10.31764/jtam.v6i3.8408

———————— ◆ ————————

A. INTRODUCTION

Machine learning is a branch of artificial intelligence that develops a computer algorithm to adapt and evolve based on empirical data (Jalal & Ezzedine, 2019). There are three types of machine learning based on human supervision in the learning process, i.e., supervised learning, unsupervised learning, and reinforcement learning (Dey, 2016). Supervised learning is a type of machine learning in which computer algorithms are trained on input data labelled for a specific output. Examples of supervised learning are regression and classification problems (Kotsiantis, 2007). Meanwhile, unsupervised learning does not require labels in the learning process. The unsupervised learning algorithm will find natural patterns from the data without human supervision. An example of this unsupervised learning is the clustering problem (Alzubi et al., 2018). On the other hand, reinforcement learning is a type of machine learning based on rewards and/or punishments for desired and/or unwanted behavior (Sutton, 1992).

Machine learning is one of the most popular and frequently used techniques today. According to the data on scopus.com, articles about machine learning first appeared in the

1950s (Campaigne, 1959; Martens, 1959) and have overgrown from year to year. Research related to machine learning in 2021 reached 80,722 articles with the top four fields: computer science, engineering, medical science, and mathematics. Some of the applications of machine learning include forecasting (Aggarwal & Toshniwal, 2021), prediction (Liu et al., 2021), anomaly detection (Zhou, 2021), and pattern recognition (Li, 2021).

One of the uses of machine learning is to predict the length of study for undergraduate students. The accreditation of a study program is strongly influenced by the study period of its graduates, referring to the Regulation of the National Accreditation Board for Higher Education (called BAN-PT) Number 3 of 2019 concerning higher education accreditation instruments. Several studies that apply machine learning to predict a student's study period, such as the C4.5 and k-nearest neighbor (kNN) (Purwanto et al., 2019), the decision trees and artificial neural networks (Rohmawan, 2018), fuzzy k-NN (Anugerah et al., 2017), perceptron (Masykuri et al., 2021), and many others. However, from some of these studies, it was found that there are still limited who apply machine learning-based regression methods, such as ridge regression (Marquardt, 1970), Huber regression (Huber, 1964), and quantile regression (Lejeune & Sarda, 1988).

Several factors affect the study period, such as the grade point average (GPA), the suitability of the final project topic with the area of interest, other main activities, and others. However, GPA is initial information for supervisors to characterize their supervised students. A model that can predict a student's study period based on GPA is needed for characterization so that supervisors can apply the right strategy for their students. One model that can be used is a machine learning-based regression model. Therefore, this study aims to implement and select a machine learning-based regression model to predict a student's study period based on GPA. The models used are least-square regression, ridge regression, Huber regression, quantile regression, and quantile regression with l_2-regularization provided by Machine Learning in Julia (MLJ). The regression model was selected based on several statistics, such as maximum error, root mean squared error (RMSE), and mean absolute proportional error (MAPE). The computational process was carried out using Julia version 1.6.5. which provides an environment for fast and easy implementation of various machine learning methods. The resulting model can be used by supervisors to predict the study period of their supervised students so that supervisors can characterize their students and can design appropriate strategies.

B. METHODS

The data used in this study is GPA data from semesters 1 to 6 and the study period of students in the mathematics undergraduate program at IPB University, who entered in the years of 2013-2016. The length of the study for students in the mathematics undergraduate study programs is normally 8 semesters. The data is divided into training data for students who entered in 2013-2015 and testing data for students who entered in 2016. The total data obtained was 203 data, with 178 data (87.68%) as training data and the rest as testing data.

This research begins with the collecting and processing of data as mentioned above. Based on the training data, the regression model coefficient values were calculated using several approaches, such as least-square regression, regression with l_2 -regularization (ridge regression), regression with Huber loss (Huber regression), quantile regression, and quantile regression with l_2-regularization. Furthermore, the model is evaluated using data testing based on maximum error, RMSE, and MAPE. The fittest model is selected based on these criteria. At the end of the research, some errors generated by the predicted value of the study period are analyzed. The research flow chart is shown in Figure 1 below to effectively understand this study's steps.

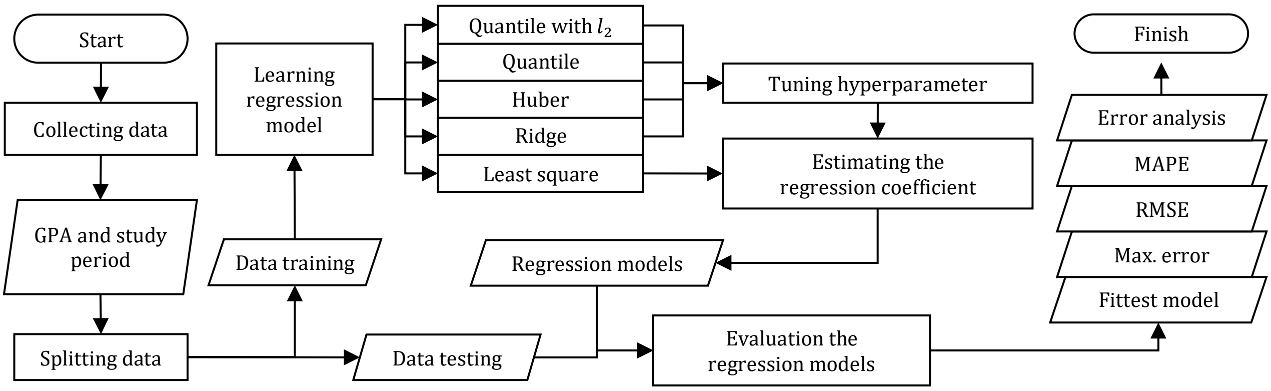

Figure 1. Research flow chart of this study

1. Multiple Linear Regression

This study uses six predictors, i.e., GPA semesters 1 to 6, with one response variable, i.e., student study period, so the regression model that will be used is

$$y = f_\beta(x) + \varepsilon = \beta_0 + \sum_{k=1}^{6} \beta_k x_k + \varepsilon \tag{1}$$

where y is the study period, x_k is the student's GPA in the k-th semester, β_k is the coefficient of x_k, β_0 is the intercept coefficient, and $f_\beta(x)$ is the predicted value of y. This study uses five machine learning approaches to estimate the regression coefficient, as follows.

a. Least-Square Regression

Least-square regression is the most popular and commonly used. The "best" coefficient value in the least-square regression model is obtained by minimizing the average value of the squared loss or mean squared error (MSE) (Heath, 2002; Johnson & Faunt, 1992), given by

$$\hat{\beta}_{ls} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{2}$$

where N is the number of training data, $\hat{\beta}_{ls}$ is the approximate value of the coefficient β based on least-square regression, $y^{(i)}$ and $f_\beta\left(x^{(i)}\right)$ are the actual and predicted values of the i-th student study period.

b. Ridge Regression

Ridge regression is a method for estimating the coefficients of a regression model with a scenario where each predictor is highly correlated (Jones, 1972). The "best" coefficient value for the ridge regression model is obtained by minimizing the mean squared error (MSE) added with l_2-regularization (Wang, 2019), given by

$$\hat{\beta}_r = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \|\beta\|_2^2 \tag{3}$$

where $\hat{\beta}_r$ is the approximate value of the coefficient β based on ridge regression, and λ is the hyper-parameter or tuning parameter of the model. Hyper-parameters can be tuned during the training process using the cross-validation method (An et al., 2007).

c. Huber Regression

Huber regression is one robust regression, a type of regression model that is insensitive to data outliers (Wager et al., 2005). The "best" coefficient value for the Huber regression model is obtained by minimizing the average value of the Huber loss (Huber, 1964), given by

$$\hat{\beta}_h = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right) \tag{4}$$

with

$$\ell(r) = \begin{cases} \frac{1}{2} r^2, & |r| < \delta \\ \delta\left(|r| - \frac{1}{2}\delta\right), & |r| \geq \delta \end{cases} \tag{5}$$

where $\hat{\beta}_h$ is the approximate value of the coefficient β based on Huber regression. The function $\ell\left(f_\beta(x), y\right)$ is called a Huber loss or smooth absolute loss function, while δ is a hyper-parameter of this model.

d. Quantile Regression

Quantile regression is an extension of the least-square regression used when least-square conditions or assumptions are not met (Koenker & Bassett, 1978). In contrast to least-square, which estimates the conditional mean, quantile regression estimates the conditional median or other quantile values of the response variable across values of the predictor variables (Davino et al., 2022). The "best" coefficient value for the quantile regression model is obtained by minimizing the average pinball loss (P_δ) a.k.a. linear loss, given by

$$\hat{\beta}_q = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} P_\delta\left(f_\beta\left(x^{(i)}\right) - y^{(i)}\right) \tag{6}$$

with

$$P_\delta(r) = \begin{cases} \delta r, & r > 0 \\ (1-\delta) r, & r \leq 0 \end{cases} \tag{7}$$

where $\hat{\beta}_q$ is the approximate value of the coefficient β based on quantile regression (Cahyani et al., 2016). The value of δ is the hyperparameter of this model, referred to as the quantile. If δ is equal to 0.5, then quantile regression estimates the conditional median of y across values of x.

e. Quantile Regression with l_2-Regularization

The last regression model used is the quantile regression model with l_2-regularization. The "best" coefficient value for this regression model is obtained by minimizing the linear loss added with l_2-regularization, given by

$$\hat{\beta}_{ql} = \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} P_{\delta}\left(f_{\beta}\left(x^{(i)}\right) - y^{(i)}\right) + \lambda \|\beta\|_2^2 \qquad (8)$$

where $\hat{\beta}_{ql}$ is the approximate value of the coefficient β based on quantile regression with

l_2-regularization. The value of δ used is the value obtained in the quantile regression in the previous model. Thus, the hyperparameter that needs to be tuned in this regression model is λ, which is the coefficient of l_2-regularization.

2. *k*-Fold Cross-Validation

Four of the five models have hyper-parameters that must be tuned using cross-validation. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample (Wainer & Cawley, 2017). One of the most popular procedures is k-fold cross-validation (Lyu et al., 2022). This procedure has a single parameter called k, which refers to the number of groups to be divided. This study chose the value of k, which is 10, so this procedure can be referred to as 10-fold cross-validation. An illustration of 10-fold cross-validation is shown in Figure 2 below

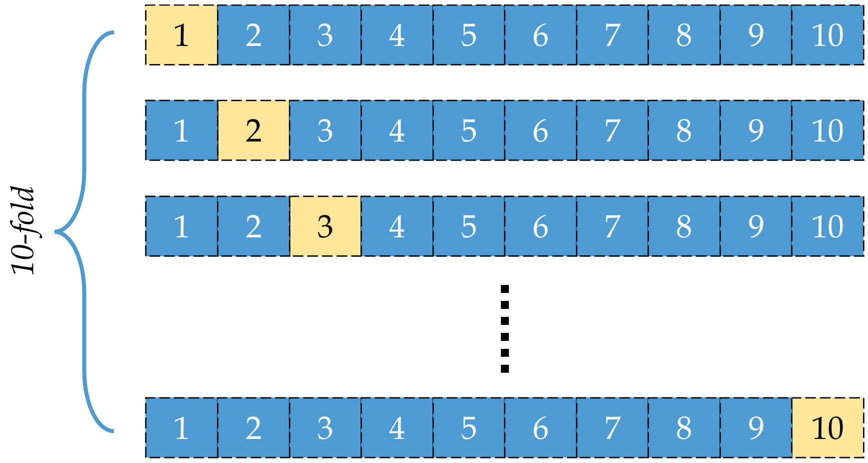

Figure 2. Illustration of 10-fold cross-validation

Suppose there is a model with hyper-parameter δ to be tested as many as M, namely $\delta_1, \delta_2, \dots, \delta_M$. The 10-fold cross-validation procedure divides the data into 10 equal parts of data (sub-data). For any hyper-parameter δ_1 the regression model's coefficient is estimated using training data with different sub-data. First, the regression model's coefficient is estimated using the 2nd to 10th sub-data, then the model is evaluated in the 1st sub-data. Second, the evaluation data used is the second sub-data, with the regression model's coefficient estimated using other sub-data. The process was repeated 10 times so that each sub-data was used as evaluation data. The evaluation result of δ_1 is the average of the evaluations of 10 meta-models. The process is carried out for each other hyper-parameter, $\delta_2, \delta_3 \dots, \delta_M$ and the hyperparameter δ is selected based on the best evaluation value. This study's evaluation at the cross-validation stage was based on MAPE.

3. Evaluation of the Model

After the hyper-parameter values are tuned and the regression model coefficients are estimated, the following process evaluates the models used in the data testing. Data testing is not used at all during the learning process. Some of the measures used for evaluation include the maximum error ($\varepsilon_{\max}$), RMSE, and MAPE, given by

$$\varepsilon_{max} = \max_{i=1,2,\ldots n} \left| f_{\widehat{\beta}}\left(x^{(i)}\right) - y^{(i)} \right| \tag{9}$$

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} \left[f_{\widehat{\beta}}\left(x^{(i)}\right) - y^{(i)} \right]^2 \right)^{\frac{1}{2}} \tag{10}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{f_{\widehat{\beta}}\left(x^{(i)}\right) - y^{(i)}}{y^{(i)}} \right| \tag{11}$$

where n is the number of testing data, $y^{(i)}$ and $f_{\widehat{\beta}}\left(x^{(i)}\right)$ are the actual and predicted values of the study period based on the regression model with a coefficient of $\hat{\beta}$.

4. Julia Programming Language

All computational process in this study was carried out using Julia version 1.6.5. Julia is a new programming language with its primary target on technical computing (Bezanson et al., 2017). Julia claimed to have, speed like C, dynamic like Ruby, feel like Lisp, familiar with mathematical notation like MATLAB, easy to use like Python and R (Joshi & Lakhanpal, 2017). With a simple language, fast, and open source, Julia has quickly become a competitive language in data sciences and scientific computing (Ardhana et al., 2022; Gao et al., 2020). Julia provides many packages that can be used to help the computing process of its users. On the other hand, Julia users can also participate in providing packages and sharing them with other users. In 2022, Julia community has registered over 7,400 Julia packages for community use.

This study uses the MLJ (Machine Learning in Julia) package version 0.16.1 (Blaom et al., 2020). MLJ is a toolbox that provides interfaces and meta-algorithms for selecting, tuning, evaluating, compiling, and comparing more than 180 machine learning models written in Julia and several other languages. MLJ integrates with various other machine learning packages, such as scikit-learn, Flux, GLM, and many more, making model selection easier.

C. RESULT AND DISCUSSION

This section explains the results of hyper-parameter tuning and coefficient estimation of five machine learning-based regression models. After that, the results of the evaluation and comparison of each regression model are explained using the maximum error, RMSE, and MAPE.

1. Data Summary

Before we discuss the training and testing of the model, this section will show the characteristics of the data used. A brief summary of the data used can be seen in Figure 3 below.

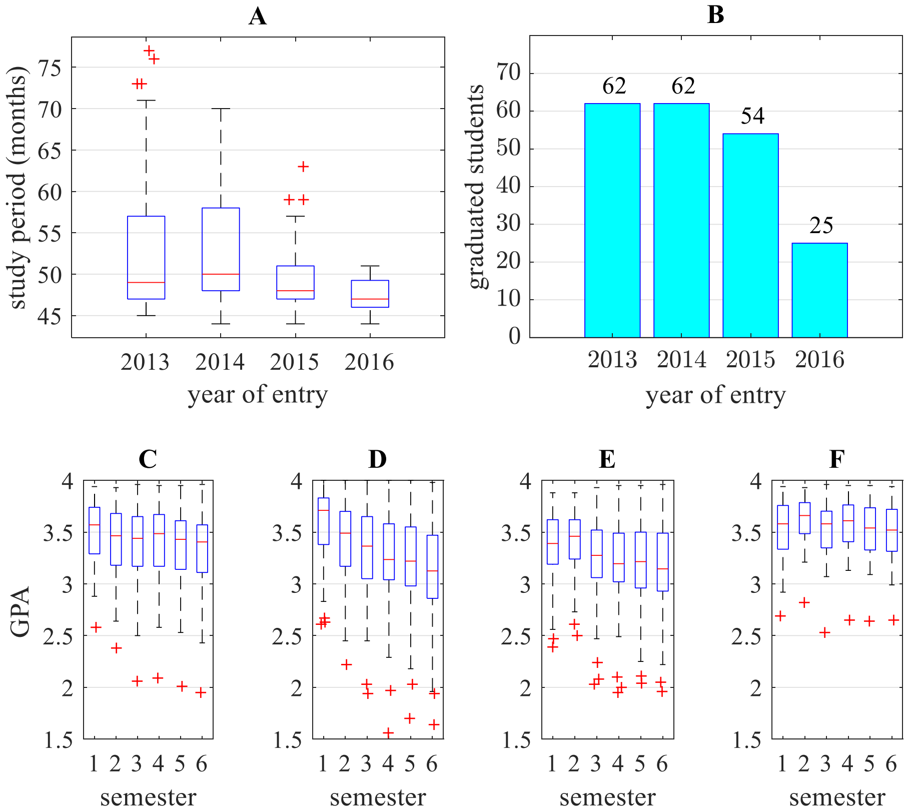

Figure 3. A brief summary of the data used: (A) study period and (B) number of graduated students by year of entry; GPA semester 1-6 of graduated students who entered in (C) 2013, (D) 2014, (E) 2015, and (F) 2016.

As previously mentioned, the data used as predictors are the GPA from semesters 1-6 and the study period as response variables. For students who entered in 2015 and 2016, the data used is only those who have graduated in 2020. The 2021 graduate data is not used because student graduation at that time was influenced by the Covid-19 outbreak.

2. Estimation of the regression coefficient

Using training data, the regression coefficient in Eq. 1 is estimated using five machine learning approaches. Following are the results of each of these approaches.

a. Least-Square Regression Model

The coefficients of the least-squares regression model can be found easily using a matrix formulation. Suppose there is a matrix X containing a set of predictors and β is a vector containing the coefficients of each corresponding predictor, i.e.,

$$X = [1\ \ X_1\ \ X_2\ \ X_3\ \ X_4\ \ X_5\ \ X_6], \qquad \beta = [\beta_0\ \ \beta_1\ \ \beta_2\ \ \beta_3\ \ \beta_4\ \ \beta_5\ \ \beta_6]^T \tag{12}$$

where X_k is a column vector containing student GPA in the k-th semester on the training data. If Y is a column vector containing the response variable, i.e., the study period on the training data, then the coefficient value of β_{ls} (estimator for β) that satisfies Eq. 2 is given by

$$\hat{\beta}_{ls} = (X^T X)^{-1}(X^T Y) \tag{13}$$

Based on the training data, the least-square regression model is obtained and is given by

$$f_{\hat{\beta}_{ls}}(x) = 66.47 + 1.84X_1 - 3.43X_2 + 5.03X_3 + 20.24X_4 - 20.76X_5 - 7.69X_6 \tag{14}$$

Because of the convenience provided by least squares in estimating the coefficients of the regression model, there are many packages in MLJ that provide this regression model, such as ScikitLearn, GLM, MLJLinearModels, and MultivariateStats.

b. Ridge Regression Model

Similar to least squares, ridge regression can also be solved using a matrix formulation. However, ridge regression requires the hyper-parameter value λ to be tuned. For any hyper-parameter value λ, the coefficient value β_r (estimator for β) that satisfies Equation (3) is given by

$$\hat{\beta}_r = (X^T X + \lambda I)^{-1}(X^T Y) \tag{15}$$

where I is a identity matrix, and $\lambda > 0$ is small.

Using 10-fold cross-validation, the hyper-parameter value λ is selected in the interval $[0, 0.2]$ with 51 points being tested, i.e., $0.004a$ for $a = 0,1,\dots,50$. The cross-validation results for the ridge regression are shown in Figure 3.A. Based on the MAPE value, $\lambda = 0.048$ gives the best accuracy. Thus, this value is used as a hyper-parameter for the ridge regression. Based on the training data and the hyper-parameter value $\lambda = 0.048$, the ridge regression model is given by

$$f_{\hat{\beta}_r}(x) = 63.85 + 2.32X_1 - 2.97X_2 + 4.69X_3 + 17.14X_4 - 15.77X_5 - 9.45X_6 \tag{16}$$

Some packages in MLJ that provide ridge regression models are ScikitLearn, MLJLinearModels, and MultivariateStats.

c. Huber Regression Model

Huber regression, also called robust regression with Huber loss, is a regression type that is not sensitive to data outliers. Although not sensitive to outliers, Huber regression does not ignore the effect of data outliers. Huber regression only assigns a lower weight to the outlier. Based on Equation (5), Huber regression will optimize the square loss when the absolute value of the residual between the actual and predicted values is less than a bound δ, which is called the hyperparameter. Meanwhile, the absolute loss will be optimized if the residual value is greater than the hyper-parameter value.

In contrast to least squares and ridge regression, the estimation of the Huber regression's coefficients cannot use a matrix formulation. One method used for Huber regression or other robust regression is M-estimation (Huber, 1964). The letter of M in M-estimation stands for "maximum likelihood type". Using MLJ packages, Huber regression coefficients can be estimated easily and quickly.

Using 10-fold cross-validation, the hyper-parameter value δ is selected in the interval $[0, 1]$ with 51 points being tested, i.e., $0.02a$ for $a = 0,1,\dots,50$. The cross-validation results for the Huber regression are shown in Figure 3.B. Based on the MAPE value, $\delta = 0.84$ gives the best accuracy. Thus, this value is used as a hyper-parameter for the Huber regression.

Based on the training data and the hyper-parameter value $\delta = 0.84$, the Huber regression model is given by

$$f_{\hat{\beta}_h}(x) = 66.55 + 1.24X_1 - 4.36X_2 + 8.09X_3 - 1.75X_4 - 2.83X_5 - 5.64X_6 \tag{17}$$

Some packages in MLJ that provide Huber regression are ScikitLearn and MLJLinearModels.

d. Quantile Regression Model

Quantile regression is usually used when the conditions or assumptions in the least-square regression are not met. Similar to Huber regression, quantile regression has no sensitivity to data outliers. Quantile regression will choose the conditional median or other quantile value. This quantile value is called the hyper-parameter in the quantile regression model and is denoted by δ.

Using 10-fold cross-validation, the hyper-parameter value δ is selected in the interval $[0, 1]$ with 51 points being tested, i.e., $0.02a$ for $a = 0,1,\ldots,50$. The cross-validation results for the quantile regression are shown in Figure 3.C. Based on the MAPE value, $\delta = 0.6$ gives the best accuracy. Thus, this value is used as a hyper-parameter for the quantile regression.

Based on the training data and the hyper-parameter value $\delta = 0.6$, the quantile regression model is given by

$$f_{\widehat{\beta}_q}(x) = 63.19 + 0.83X_1 - 2.85X_2 + 7.18X_3 - 3.31X_4 - 0.44X_5 - 5.85X_6 \tag{18}$$

The package in MLJ that provides quantile regression is MLJLinearModels.

e. Quantile Regression with l_2-Regularization Model

The last regression model used is quantile regression with l_2-regularization. The basis of this regression model is the same as that of quantile regression. The difference is, the loss function in quantile regression is added with l_2-norm regularization. Thus, this model has two hyper-parameter values, i.e., δ as the quantile value and λ as the regularization weight. The value of δ used is derived from the quantile regression model, i.e., $\delta = 0.6$, so that only λ will be tuned using cross-validation.

Using 10-fold cross-validation, the hyper-parameter value λ is selected in the interval $[0.5, 1]$ with 51 points being tested, i.e., $0.5 + 0.01a$ for $a = 0,1,\ldots,50$. The cross-validation results for this regression model are shown in Figure 3.D. Based on the MAPE value, $\lambda = 0.9$ gives the best accuracy. Thus, this value is used as a hyper-parameter for the quantile regression with l_2-regularization.

Based on the training data and the hyper-parameter values δ=0.6, and λ=0.9, the quantile regression model with l_2-regularization is obtained and is given by

$$f_{\widehat{\beta}_{ql}}(x) = 63.19 + 0.83X_1 - 2.85X_2 + 7.18X_3 - 3.31X_4 - 0.44X_5 - 5.85X_6 \tag{19}$$

The results of the tuning process for the hyper-parameter values of ridge regression, Huber regression, quantile regression, and quantile regression with l_2-regularization models as shown in Figure 3.

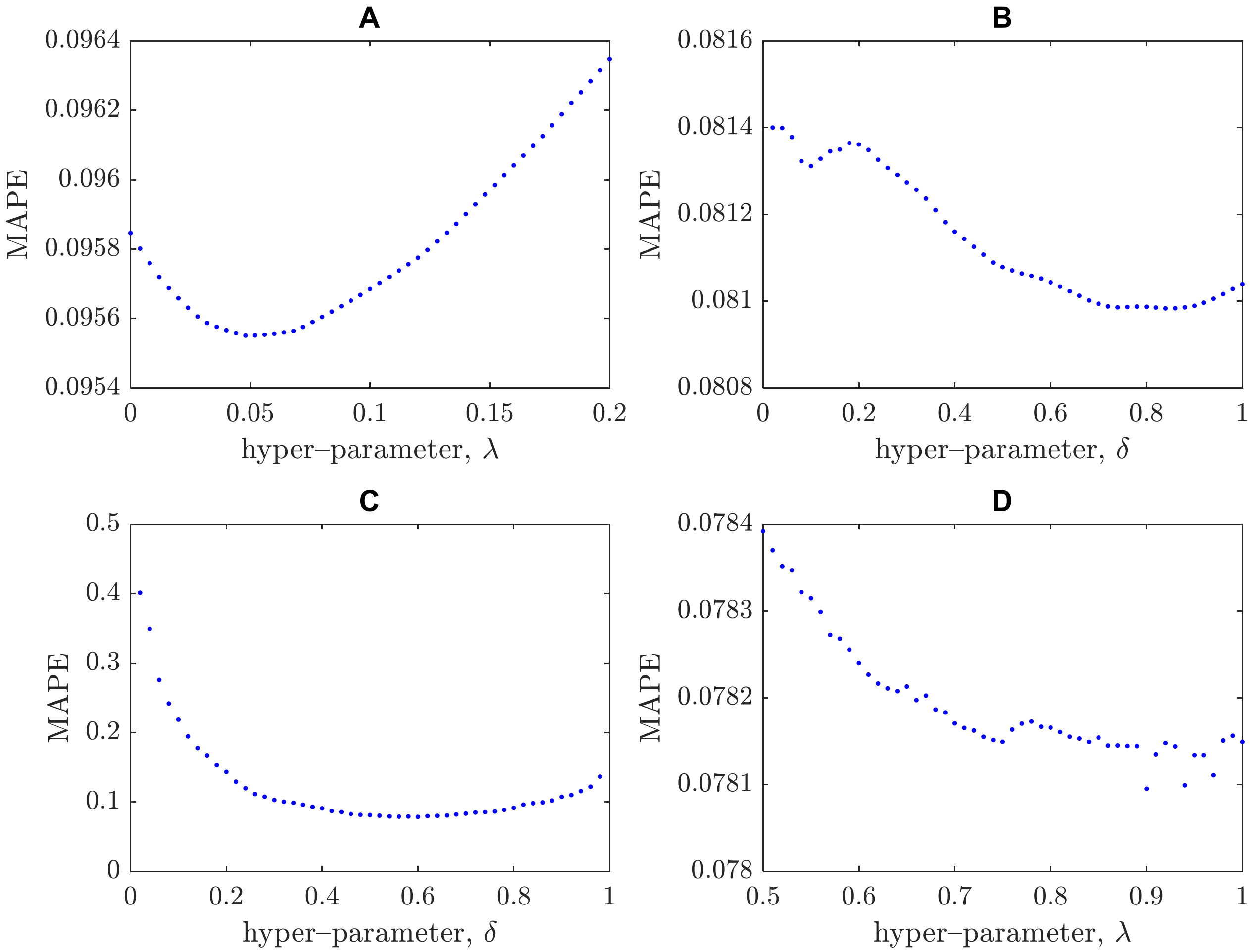

Figure 3. The results of cross-validation for the selection of hyper-parameter values in (A) ridge regression, (B) Huber regression, (C) quantile regression, and (D) quantile regression with l_2-regularization

3. Evaluation of the Regression Model

After the five models are obtained, the next step is to evaluate these models using data testing. There are three criteria used in this evaluation step, i.e., the maximum error value ($\varepsilon_{\max}$), the root mean squared error (RMSE), and the mean absolute proportional error (MAPE). The evaluation results of the five models. The evaluation value in bold indicates the best evaluation value among other models, as shown in Table 1.

Table 1. Evaluation results of machine learning-based regression models

Regression Models	λ	δ	ε_{max}	**RMSE**	**MAPE**
Least square	-	-	9.2712	3.9055	7.17%
Ridge	0.048	-	9.1667	3.7814	6.95%
Huber	-	0.84	5.4043	2.4019	3.90%
Quantile	-	0.60	5.2159	**2.3094**	**3.56%**
Quantile with l_2	0.90	0.60	**4.8959**	2.3673	4.03%

Based on the evaluation results, the least-square regression model produces the worst accuracy among the others. From Table 1, it can be seen that the maximum error values, RMSE, and MAPE of the least-square regression model are 9.2712, 3.9055, and 7.17%, respectively. The error value is higher than the other four models. The maximum prediction error in this model is more than 9 months, meaning that the student's study period can be 9 months faster or longer than the predicted value. Although the ridge regression model can refine the error of the least-square regression model, the refinement is not very significant. Based on Table 1, the values of ε_{max}, RMSE, and MAPE of the ridge regression model are not much different from the least-square regression model.

Meanwhile, Huber regression was able to significantly improve the error of the least-square regression model. In this model, the maximum error obtained is 5.4 months, much better than the least-square regression model. Likewise, the RMSE and MAPE values have improved significantly. However, the regression model that gives the best evaluation results is given by the quantile regression model. Based on RMSE and MAPE, the quantile regression model without regularization gives the best evaluation results, while the quantile regression model with l_2-regularization is better at the maximum error criterion. Thus, the fittest model to predict the study period based on GPA is the quantile regression model, as shown in Table 2.

Table 2. Prediction of the study period using quantile regression with l_2-regularization and a maximum error of more than three months

No.	X_1	X_2	X_3	X_4	X_5	X_6	Y	$\widehat{Y}$	$\widehat{Y}-Y$
1	3.94	3.93	3.96	3.95	3.95	3.94	51.00	46.10	−4.90
2	2.69	2.82	2.53	2.65	2.64	2.65	47.00	51.48	4.48
3	3.56	3.55	3.58	3.58	3.48	3.35	44.00	48.30	4.30
4	2.92	3.21	3.16	3.24	3.23	3.22	45.00	49.07	4.07
5	3.78	3.82	3.64	3.70	3.66	3.58	51.00	47.38	−3.62
6	3.5	3.68	3.71	3.77	3.79	3.78	50.00	46.75	−3.25

Table 2 shows the prediction results of the study period using quantile regression with l_2-regularization on data testing with a maximum error of more than three months. The overestimated and underestimated prediction results are both three. The maximum error occurs when the prediction is less than the actual value (underestimated). In this data, students have a consistently high GPA of around 3.94, so the prediction of the study period is very fast, i.e., 46.10 months. However, these students take up to 51 months of study in reality. Meanwhile, predictions are overestimated in the second data. Because the GPA is relatively sufficient (less than 3), the regression model predicts that the student can graduate within 51.48 months. However, in reality, these students can graduate on time, which is within 47 months.

This study provides an alternative prediction model for student graduation based on GPA. While many models provide a predictive model in the form of a classification of whether students graduate on time or not (Purwanto et al., 2019; Risnawati, 2018; Thaniket et al., 2020), the model in this study provides an approach in the form of a regression model that estimates the number of months required for students to complete their undergraduate studies. However, based on the results, GPA is not the only factor that affects the study period. Other factors, such as the suitability of the field of interest, the ease of finding references, the regularity of the guidance process, and the presence or absence of other main activities, can affect the student's study period (Masykuri et al., 2021). This fact provides an opportunity for further research to construct a regression-based predictive model for the study period of undergraduate students with better accuracy.

The model in this study can be used to describe the characteristics of new guidance students. Thus, the supervisor can determine the right strategy for the supervision process. By knowing the estimated study period, the supervisor can better determine the appropriate topic for the student. For students who are estimated to have a fast study period, the topic of the final project for these students can be wider with several challenges and the supervision process can also run normally without special treatment. Meanwhile, for students who are estimated to have a long period of study, the topic of the final project must be adjusted to the ability of the student without compromising on quality. In addition, the mentoring process can also be carried out more rigorously. Thus, it is hoped that the student's study period can be completed quickly and have a good final project quality.

D. CONCLUSION AND SUGGESTIONS

This study models the length of the student's study period based on GPA using a machine learning-based regression model. The least-square regression model gives the worst evaluation results of the several regression models, although the calculation method is easy. Meanwhile, the quantile regression model gave the best results. Based on RMSE and MAPE, the quantile regression model without regularization gave the best evaluation result. Moreover, the quantile regression model with l_2-regularization had a better evaluation result on the maximum error criterion.

This study provides an alternative prediction model for student graduation based on GPA in the form of a regression model that estimates the number of months it takes students to complete their undergraduate studies, while other studies provide models that predict whether students graduate on time or not. However, this study can be developed by adding other supporting predictor variables, which can be obtained at the beginning of the supervision process, such as the suitability of field interests between students and supervisors, and many activities other than completing the final project. Comparison with more modern machine learning methods can also be applied to get better results.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Mathematics, IPB University for providing data and support for this research.

REFERENCES

Aggarwal, A., & Toshniwal, D. (2021). A hybrid deep learning framework for urban air quality forecasting. *Journal of Cleaner Production*, *329*. https://doi.org/10.1016/j.jclepro.2021.129660

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1). https://doi.org/10.1088/1742-6596/1142/1/012012

An, S., Liu, W., & Venkatesh, S. (2007). Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition*, *40*(8), 2154–2162. https://doi.org/10.1016/j.patcog.2006.12.015

Anugerah, A. S. P., Indriati, & Dewi, C. (2017). Implementasi Algoritme Fuzzy K-Nearest Neighbor untuk Penentuan Lulus Tepat Waktu (Studi Kasus: Fakultas Ilmu Komputer Universitas Brawijaya). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(4), 1726–1732.

Ardhana, N. K. K., Nurdiati, S., Najib, M. K., & Mukrim, S. A. (2022). Akurasi dan Efisiensi Solusi Persamaan Diferensial Biasa Dengan Masalah Nilai Batas Pada Julia dan Octave. *Jurnal Matematika UNAND*, *11*(1), 32–46. https://doi.org/10.25077/jmu.11.1.32-46.2022

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Blaom, A., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., & Vollmer, S. (2020). MLJ: A Julia package for composable machine learning. *Journal of Open Source Software*, *5*(55), 2704. https://doi.org/10.21105/joss.02704

Cahyani, T. B. N., Wigena, A. H., & Djuraidah, A. (2016). Quantile regression with elastic-net in statistical downscaling to predict extreme rainfall. *Global Journal of Pure and Applied Mathematics*, *12*(4), 3517–3524.

Campaigne, H. (1959). Some experiments in machine learning. *Proceedings of the Western Joint Computer Conference, IRE-AIEE-ACM 1959*, 173–175. https://doi.org/10.1145/1457838.1457868

Davino, C., Romano, R., & Vistocco, D. (2022). Handling multicollinearity in quantile regression through the use of principal component regression. *Metron*. https://doi.org/10.1007/s40300-022-00230-3

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179.

Gao, K., Mei, G., Piccialli, F., Cuomo, S., Tu, J., & Huo, Z. (2020). Julia language in machine learning: Algorithms, applications, and open issues. *Computer Science Review*, *37*. https://doi.org/10.1016/j.cosrev.2020.100254

Heath, M. T. (2002). *Scientific Computing: An Introduction to Survey* (2nd ed.). McGraw-Hill.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Jalal, D., & Ezzedine, T. (2019). Performance analysis of machine learning algorithms for water quality monitoring system. *2019 International Conference on Internet of Things, Embedded Systems and Communications, IINTEC 2019 - Proceedings*, 86–89. https://doi.org/10.1109/IINTEC48298.2019.9112096

Johnson, M. L., & Faunt, L. M. (1992). Parameter estimation by least-squares methods. *Methods in Enzymology*, *210*, 1–37. https://doi.org/10.1016/0076-6879(92)10003-V

Jones, T. A. (1972). Multiple regression with correlated independent variables. *Journal of the International Association for Mathematical Geology*, *4*(3), 203–218. https://doi.org/10.1007/bf02311718

Joshi, A., & Lakhanpal, R. (2017). *Learning Julia: Build high-performance applications for scientic computing*. Packt Publishing Ltd.

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, *46*(1), 33. https://doi.org/10.2307/1913643

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica (Ljubljana)*, *31*(3), 249–268.

Lejeune, M. G., & Sarda, P. (1988). Quantile regression: a nonparametric approach. *Computational Statistics and Data Analysis*, *6*(3), 229–239. https://doi.org/10.1016/0167-9473(88)90003-5

Li, Z. (2021). Research on signal modulation based on machine learning intelligent algorithm and computer automatic identification. *Journal of Physics: Conference Series*, *2083*(4). https://doi.org/10.1088/1742-6596/2083/4/042092

Liu, T., Wang, Z., Zeng, J., & Wang, J. (2021). Machine-learning-based models to predict shear transfer strength of concrete joints. *Engineering Structures*, *249*. https://doi.org/10.1016/j.engstruct.2021.113253

Lyu, Z., Yu, Y., Samali, B., Rashidi, M., Mohammadi, M., Nguyen, T. N., & Nguyen, A. (2022). Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam. *Materials*, *15*(4). https://doi.org/10.3390/ma15041477

Marquardt, D. W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, *12*(3), 591. https://doi.org/10.2307/1267205

Martens, H. H. (1959). Two notes on machine "Learning." *Information and Control*, *2*(4), 364–379. https://doi.org/10.1016/S0019-9958(59)80014-0

Masykuri, W. S., Khatizah, E., & Bukhari, F. (2021). The application of perceptron method in predicting student graduation based on several identified key factors. *IOP Conference Series: Earth and Environmental Science*, *1796*(1). https://doi.org/10.1088/1742-6596/1796/1/012060

Purwanto, E., Kusrini, K., & Sudarmawan, S. (2019). Prediksi Kelulusan Tepat Waktu Menggunakan Metode C4.5 dan K-NN (Studi Kasus : Mahasiswa Program Studi S1 Ilmu Farmasi, Fakultas Farmasi, Universitas Muhammadiyah Purwokerto). *Techno (Jurnal Fakultas Teknik, Universitas Muhammadiyah Purwokerto)*, *20*(2), 131. https://doi.org/10.30595/techno.v20i2.5160

Risnawati. (2018). Analisis Kelulusan Mahasiswa Menggunakan Algoritma C.45. *Jurnal Mantik Penusa*, *2*(1), 71–76.

Rohmawan, E. P. (2018). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree dan Artificial Neural Network. *Jurnal Ilmiah MATRIK*, *20*(1), 21–30.

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. *Machine Learning*, *8*(3–4), 225–227. https://doi.org/10.1007/BF00992695

Thaniket, R., Kusrini, K., & Luthf, E. T. (2020). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma Support Vector Machine. *Jurnal FATEKSA : Jurnal Teknologi Dan Rekayasa*, *5*(2), 20–29.

Wager, T. D., Keller, M. C., Lacey, S. C., & Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage*, *26*(1), 99–113. https://doi.org/10.1016/j.neuroimage.2005.01.011

Wainer, J., & Cawley, G. (2017). Empirical evaluation of resampling procedures for optimising SVM hyperparameters. *Journal of Machine Learning Research*, *18*(13), 1–35.

Wang, S. (2019). A sharper generalization bound for divide-and-conquer ridge regression. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 5305–5312. https://doi.org/10.1609/aaai.v33i01.33015305

Zhou, J. (2021). Research on Time Series Anomaly Detection: Based on Deep Learning Methods. *Journal of Physics: Conference Series*, *2132*(1). https://doi.org/10.1088/1742-6596/2132/1/012012