STMIK
Amik Riau

JAIA – Journal Of Artificial Intelligence And Applications
Journal Homepage : http://jurnal.sar.ac.id/index.php/JAIA/

Prediction of Student Study Duration Using Multiple Linear Regression Method

Rahmawati [1], Triyani Arita Fitri [2], Lusiana [3], Rini Yanti [4]
[1,2,3,4]STMIK Amik Riau, Jl. Purwodadi Indah Km. 10 Pekanbaru, Riau, Indonesia

Article Info

Received	: October 23, 2023
Revised	: October 27, 2023
Accepted	: October 30, 2023

Keywords :

Prediction
Student Study Duration
Multiple Linear Regression

Abstract

Data mining is a process of extracting valuable and meaningful information from large or complex data sets. In the field of education, data mining can be used to predict the length of study of students by identifying factors that affect the length of study of students. This research aims to predict the length of study of students and to find out the most influential variables in completing the length of study. The method used in this research is the Multiple Linear Regression method. Training data as much as 292 data is taken from data on graduates from 2016 - 2018. While the testing data is taken from the active student data class of 2018 as much as 148 data. The model formed will be evaluated to determine the accuracy and RMSE values. The results showed that the Multiple Linear Regression method succeeded in carrying out the prediction process optimally with a percentage accuracy value of 85%, and an RMSE value of 0.76, which means that the error rate of this model is very low. Based on the resulting coefficient value, the SKS variable is the most influential variable in the length of study of students.

Rahmawati,
Email: rahmawatitsmn@gmail.com

1. Introduction

The length of study is the period required for students to complete their education.[1] The provisions for the length of study have been regulated in the provisions of the Ministry of Education and Culture, Directorate General of Higher Education regarding the higher education system which explains that the competence of graduates for undergraduate students can complete a mandatory load of at least 144 credits with a study period of 4 years and / or a maximum of 7 years of completion of the study programme.[2] . With the regulations that have been set, universities are required to have a competitive advantage by utilising all the resources they have. One of them is Human Resources (HR), in this case students. Therefore, it is necessary to monitor and evaluate student graduation rates regularly by finding knowledge or patterns of student graduation rate trends based on academic achievement. This knowledge can be used as a consideration for the Study Programme to determine policies or guidance related to student graduation rates in higher education.

.

Data Mining is a process that uses one or more computer learning techniques (Machine Learning) to analyse and extract knowledge or knowledge automatically [3]. Data mining is divided into several methods based on tasks, one of which is prediction. Prediction is the process of systematically estimating what is likely to happen in the future based on past and present information, so that the error of difference between something that happens and the predicted results can be minimised [4].

One of the algorithms used in prediction is linear regression. Linear regression is divided into two types, namely simple linear regression and multiple linear regression. Multiple linear regression is a linear relationship between two or more independent variables and the dependent variable which aims to determine the direction of the relationship between the independent variable and the dependent variable [5].

Several studies on predicting student graduation have been conducted using different algorithms and accuracy values, including the Fuzzy C-Means and K-Nearest Neighbors methods with an accuracy of 71% [6]. The same research using the C4.5 algorithm has an accuracy value of 90%, precision 91.38% and recall 98.15%. which was tested using Rapid Miner software [7]. Furthermore, the CRISP-DM (Cross Industry Standard Process for Data Mining) method using the Support Vector Machine algorithm. gives an accuracy value of 94.4% [8]. Other research uses the Decision Tree and Artificial Neural Network methods with an accuracy rate of 74.51% and an artificial neural network of 79.74% [9].

This study aims to determine the accuracy of the Multiple Linear Regression method in predicting the length of study of students, and determine the variables that have the most influence on the length of study of students in completing their studies in the Informatics Engineering study programme. This research is expected to help the study programme evaluate the study period of students. knowing the factors that have the most influence on the completion of student studies so that control and monitoring can be improved for students in completing their studies through the Academic Advisor (PA) lecturer.

2. Research Methods

The phase in this research is only through a few stages, which can be described in image 1 of the research course. This is then solved with algorithm performance testing

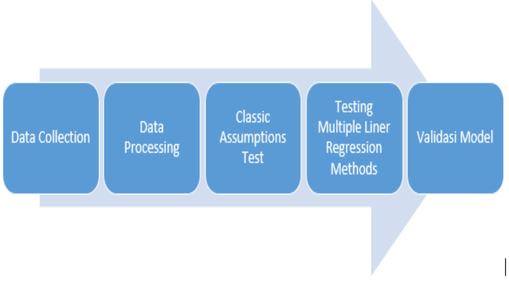


Figure 1. Research Stages

2.1. Data Collection

The data used in this study are data on graduates and students of the Informatics Engineering study programme of STMIK AMIK Riau from the PDPT section. 292 training data were taken from the data of graduates from 2016 - 2018. While the testing data is taken from the data of active students in 2018 as much as 148 data.

2.2. *Data Preprocessing*

The data used in this study are data on graduates and students of the Informatics Engineering study programme of STMIK AMIK Riau from the PDPT section. 292 training data were taken from the data of graduates from 2016 - 2018. While the testing data is taken from the data of active students in 2018 as much as 148 data.

3.3. Classical Assumption Test

Multiple linear regression is one of the hypothesis tests to determine the effect between independent and dependent variables [10]. The multiple linear regression equation model is as follows:

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

Description :

Y	: Variabel dependent
a	: Konstanta
β_n	: Koefisien regresi
X	: Variabel independent

Multiple linear regression algorithms have conditions that must be met. The resulting linear regression model must fulfil the BLUE (Best Linear Unbiased Estimation) criteria. To fulfil the BLUE criteria, a classical assumption test must be carried out before making a linear regression model. The classical assumption test consists of several processes, including normality test, heteroscedasticity test, multicollinearity test, and autocorrelation test.

3.4. Testing Method

After the classical assumption test is carried out, the data and variables used in this study are suitable to proceed to the Multiple Linear Regression method testing stage. The process of testing this method is done manually using Microsoft Excel 2016 and using an open source tool, Jupyter Notebook with the Python programming language. The method used in this manual calculation of multiple linear regression uses the Matrix method.

3.5. Model Validation

The best results of the regression equation obtained to explain the dependent variable, will depend on how small the error is in the analysis, to be able to minimise the error rate, it is necessary to validate the model, there are various forms of model validation. However, in this study, model validation uses R-Squared, and Root Mean Squared Error (RMSE). *R-Squared is the proportion of variation in the dependent variable that can be explained by the independent variables. R-Squared has a range of 0 ≤ R-Squared ≤ 1. If R-Squared is 1 then 100% of the variation in the dependent variable can be explained by the independent variables. Meanwhile, if R-Squared is 0, the variation in the dependent variable cannot be explained by the independent variables [12]. The equation for calculating R-Squared is like equation (2).*

$$R^2 = \frac{SSR}{SST} \quad (2)$$

1. *Root Mean Squared Error* (RMSE)

Root Mean Squared Error (RMSE) is a quadratic assessment rule that also measures the average magnitude of the error. RMSE is the square root of the average squared difference between predictions and actual observations of data, which measures the average error made by the model in predicting the outcome of an observation. Mathematically, RMSE is the square root of the Mean Squared Error (MSE), which is the average squared difference between the actual observed outcome value and the value predicted by the model. Thus, MSE = mean ((observed - predicted) ^ 2) and RMSE = sqrt (MSE). The lower the RMSE, the better the model [12]. The RMSE equation is like equation (3).

$$RMSE = \left(\frac{\sum(yi-\widehat{yi})}{n}\right)^{1/2} \qquad (3)$$

A low RMSE value indicates that the variation in values produced by a forecast model is close to the variation in observed values. RMSE calculates how different a set of values are. The smaller the RMSE value, the closer the predicted values are.

3. Results and Discussion

3.1. Classical Assumption Test

A low RMSE value indicates that the variation in values produced by a forecast model is close to the variation in observed values. RMSE calculates how different a set of values are. The smaller the RMSE value, the closer the predicted values are.Uji Normalitas

The normality test aims to test whether the independent variable regression model, the dependent variable or both are normally distributed or not using the Kolmogorov-Smirnov statistical test. The normality test aims to test whether the independent variable regression model, the dependent variable or both are normally distributed or not using the Kolmogorov-Smirnov statistical test.

1. Heteroscedasticity Test

   The results of the Heteroscedasticity test using the Park Test show that the X1 variable has a Sig value. 0.384> 0.05 and X2 has a Sig value. 0.96 > 0.05. Therefore, it can be interpreted that the variables X1 and X2 do not occur heteroscedasticity.

2. Multicollinearity Test

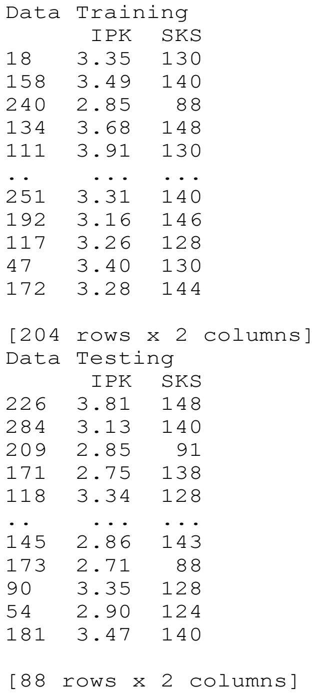
   Secondary research data is said to be free from multicollinearity problems if the Colinearity Statistics column shows Tolerance results above 0.1 and the Variation Inflation Factor (VIF) value is not more than ten. The resulting Tolerance result is 0.513> 0.1 and the total VIF result is 1.948 < 10. This shows that the research data is free from multicollinearity problems or no multicollinearity occurs.

3. Autocorrelation Test

   Secondary research data is said to be free from multicollinearity problems if the Colinearity Statistics column shows Tolerance results above 0.1 and the Variation Inflation Factor (VIF) value is not more than ten. The resulting Tolerance result is 0.513> 0.1 and the total VIF result is 1.948 < 10. This shows that the research data is free from multicollinearity problems or no multicollinearity occurs.

3.2 Multiple Linear Regression Method Testing

Testing with the multiple linear regression method uses Jupyter Notebook with the Python programming language. The data that has been obtained will be inputted with several libraries. The data will be split into 70%: 30%. The results of the dataset division, which consists of training data and testing data. The results of the division of testing data and training data are divided into 70:30 which explains 70% of training data and 30% of testing data for the test value of the sample data on the length of study of students. The total training data is 204 data with 2 columns, namely the GPA column and the SKS column. While the testing data is 88 data with 2 columns, namely the GPA column and the SKS column as shown in Figure 2.beberapa *Library*.



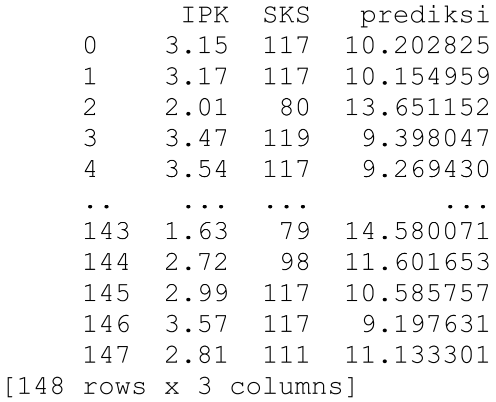
```
Data Training
       IPK  SKS
18     3.35  130
158    3.49  140
240    2.85   88
134    3.68  148
111    3.91  130
..      ...  ...
251    3.31  140
192    3.16  146
117    3.26  128
47     3.40  130
172    3.28  144

[204 rows x 2 columns]
Data Testing
       IPK  SKS
226    3.81  148
284    3.13  140
209    2.85   91
171    2.75  138
118    3.34  128
..      ...  ...
145    2.86  143
173    2.71   88
90     3.35  128
54     2.90  124
181    3.47  140

[88 rows x 2 columns]
```

Figure 2. Split Data 70% : 30%

The coefficient value and intercept are obtained based on the regression model that has been done before. the coefficient value is obtained which is [-2.39332038 -0.01945786] and the intercept is 20.018354727965352. based on the coefficient value that has been obtained, it can be determined that the most influential variable in the length of study of students, namely the SKS variable. Figure 3 shows the prediction test carried out with new data, namely the 2018 class student data.



```
      IPK  SKS   prediksi
0    3.15  117  10.202825
1    3.17  117  10.154959
2    2.01   80  13.651152
3    3.47  119   9.398047
4    3.54  117   9.269430
..    ...  ...        ...
143  1.63   79  14.580071
144  2.72   98  11.601653
145  2.99  117  10.585757
146  3.57  117   9.197631
147  2.81  111  11.133301
[148 rows x 3 columns]
```

Figure 3. New Data Prediction Result

In the fourth data record, it can be explained that the student has a GPA of 3.54 with 117 credits. Based on the prediction model that has been made using the multiple linear regression algorithm, the student is predicted to be able to complete the study for 9 semesters and 2 months..

Boxplot is a data visualisation method that is very useful and concise in describing the distribution of data. The following are the results of the boxplot visualisation shown in Figure 4. It can be explained that the lower quartile value of the GPA variable is 2.85, the SKS variable is 124, and the length of study variable is 8. Then for the middle quartile of the GPA variable is 3.285, the SKS variable is 130, and the length of study variable is 9. And for the upper quartile value of the GPA variable is 3.51, the SKS variable is 144, and the length of study variable is 14.

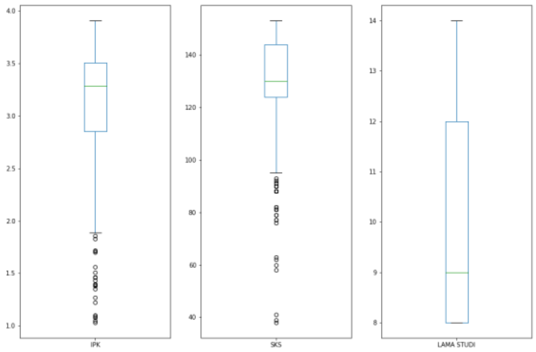


Figure 4. Boxplot Visualisation

Figure 5 presents a graph of the length of study, where line X is the length of study data, and line Y is the number of students based on the length of study. The largest number is in the length of study of 8 semesters with 78 students who are able to complete their studies for 8 semesters, while the least is in the length of study of 11 semesters with approximately 14 students who are able to complete their studies for 11 semesters.

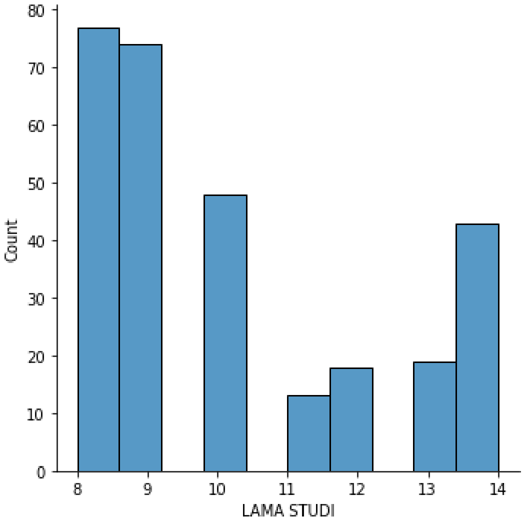


Figure 5. Duration of Study Chart

3.3 Model Validation

Based on the results of using multiple linear regression methods carried out using jupyter notebook (python). Get the results of R-Square and Root Mean Square Error. With the dataset process obtained from student data, the R Square value is 0.8524805070521941 or 85% which will be used as the accuracy value. By obtaining an accuracy value of 85%, it can be interpreted that the prediction model functions optimally. While the RMSE value obtained is 0.7937290270842988. This RMSE value can be interpreted that the error rate in the prediction model is relatively low.*Square Error*.

3. Conclusion

Based on research on predicting the length of study of students using the Multiple Linear Regression method, it can be concluded that based on training data taken from graduates of the 2016 and 2017 batches, and testing data used by 2018 batch student data successfully classified 440 data out of a total of 464 data used. The application of the prediction method with the Multiple Linear Regression Algorithm successfully predicts the length of study of students in the Informatics Engineering study programme of STMIK AMIK Riau with an accuracy percentage of 85%, and an RMSE value of 0.76 which means that the error rate of this model is classified as very low. Based on the resulting coefficient value, the SKS variable is the most influential variable in the length of student study. Future research can add a combination of prediction methods to increase the accuracy value..

4. Reference

[1] E. Etriyanti, “Perbandingan Tingkat Akurasi Metode Knn Dan Decision Tree Dalam Memprediksi Lama Studi Mahasiswa,” *J. Ilm. Bin. STMIK Bina Nusant. Jaya Lubuklinggau*, vol. 3, no. 1, pp. 6–14, 2021, doi: 10.52303/jb.v3i1.40.

[2] R. Dwi, Pambudi, A. Afif, Supianto, and N. Y. Setiawan, “Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 2196*, vol. 3, no. 3. pp. 2194–2200, 2019.

[3] T. Syahputra, J. Halim, and K. Perangin-angin, “Penerapan Data Mining Dalam Memprediksi Tingkat Kelulusan Uji Kompetensi (UKOM) Bidan Pada STIKes Senior Medan Dengan Menggunakan Metode Regresi Linier Berganda,” *Sains dan Komput.*, vol. 17, no. 1, pp. 1–7, 2018.

[4] J. Adhiva, S. A. Putri, and S. G. Setyorini, "Prediksi Hasil Produksi Kelapa Sawit Menggunakan Model Regresi Pada PT . Perkebunan Nusantara V," *Semin. Nas. Teknol. Informasi, Komun. dan Ind.*, pp. 155–162, 2020.

[5] A. Fitri Boy, "Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara)," *J. Sci. Soc. Res.*, vol. 4307, no. 2, pp. 78–85, 2020.

[6] S. P. Nabila, N. Ulinnuha, and A. Yusuf, "Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy C-Means Dan K-Nearest Neighbors Menggunakan Data Registrasi Mahasiswa," *Netw. Eng. Res. Oper.*, vol. 6, no. 1, p. 39, 2021, doi: 10.21107/nero.v6i1.199.

[7] L. Y. Lumban Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5," *Brahmana J. Penerapan Kecerdasan Buatan*, vol. 2, no. 2, pp. 97–106, 2021, doi: 10.30645/brahmana.v2i2.71.

[8] E. Haryatmi and S. P. Hervianti, "Penerapan Algoritma Support Vector Machine Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu," vol. 1, no. 10, pp. 386–392, 2021.

[9] E. P. Rohmawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree dan Artificial Neural Network," *J. Ilm. MATRIK Vol.20 No.1, April 201821-30*, pp. 21–30, 2018.

[10] K. Puteri and A. Silvanie, "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda," *Jurnal Nasional Informatika*, vol. 1, no. 2. pp. 82–94, 2020.

[11] A. Nurdany, "Rentabilitas Terhadap Pendapatan Margin MurabahahBank Syariah (Studi Kasus pada PT. Mega Bank Mega Syariah Periode 2005-2012)," *Khazanah*, vol. 5, no. 2, pp. 13–24, 2012.

[12] S. H. Saputro, R. F. B. Atmaja, and Hengki, "Analisis Pengaruh Growth Dan Variabel Makroekonomi Terhadap," vol. 01, pp. 1–13, 2021.