E-ISSN: 2962-6110
P-ISSN: 2964-7843
JSRET
jsret.knpub.com
Journal of Scientific Research, Education, and Technology

Predicting Student's Cumulative Grade (IPK) using Linear Regression with Variation in Testing Data Size

Munzir Absa
Malikussaleh University, Lhokseumawe, Indonesia

ABSTRACT

Higher education is a crucial part in the education of Indonesian youngsters. Some of the most important part of higher education for students as well as for institutions are cumulative grade (IPK) and study time. The purpose of this study is to find the relation between study time and cumulative grade, or more precisely to predict cumulative grade based on study time. This prediction is done using the linear regression formula. The data are collected from the academic database of Faculty of Teacher Training and Education, Malikussaleh University. The data are then preprocessed and split into training data and testing data. The size of the testing data is varied to find the most optimal one. After training and testing, it was found that the most optimal size of testing data is 6%, which resulted in Root Mean Square Error (RMSE) of 0.1481 and R-squared (R^2) of 0.6109. Though the result of these metrics are relatively worse compared to previous studies using linear regression, it can still be used to help find the optimal test size. To achieve better metrics, more data need to be collected from longer range of years.

Keywords: *Cumulative Grade, Study Time, Linear Regression, Testing Data Size*

Corresponding author
Name: Munzir Absa
Email: munzir.absa@unimal.ac.id

INTRODUCTION

Participation of Indonesian in higher education is increasing every year. According to a report of higher education statistic published in 2020 by Directorate General of Higher Education, more than eight million students are enrolled in higher education, be it in state university, private university, polytechnic, or vocational schools. The National Gross Enrolment Ratio (GER), a comparison between the number of undergraduate students (diploma and bachelor) with the population aged 19-23, is increasing every year from 31.61 in 2016 to 36.16 2020. This shows that there is a high level of enthusiasm for higher education in Indonesia.

In line with that, higher education institutions are also improving its quality to entice more and more students. One means to improve the quality of the institutions is through accreditation. National accreditation of higher education institutions through various scientific community accreditation bodies are very important, because it can affect

the decision of potential students in choosing a particular university / college / vocational school.

Cumulative grade (IPK) of students is one of the most important indicator of an institution's quality, according to some accreditation bodies. The average cumulative grade of students in a department of an institution is usually one of the first point to be reviewed for accreditation. Hence, it is important that the institution make an effort to improve student's mastery of each subject, thereby improving the average cumulative grade.

Study time is also an important indicator for an institution's accreditation. It is imperative that the majority of the students finish their study, and finish their study on time. There are ranges of time when students are expected to be graduated, different for different type of institution, where if the average study time exceed that, the grade for accreditation of the institution is reduced. Because of that, the institution also needs to make an effort to ensure students are not left behind as to repeat some subjects several time, or be late in the completion of their final project.

Prediction of cumulative grade and study time thus are very important for higher education institutions, because it can help them map the expected accreditation they will get. Recent research on predicting academic performance in Indonesian colleges has explored various machine learning approaches. Naive Bayes algorithms have been applied to predict student graduation status using cumulative grade point average (GPA) and social parameters, achieving accuracies of 75% and 85%, respectively (Hartatik, 2021). The Markov Chain Monte Carlo (MCMC) algorithm has been utilized to predict semester and cumulative GPAs based on student characteristics, with program of study and admission pathway emerging as significant factors (Malik, Muhammad, & Kusnawi, 2024). Decision tree models, particularly the C4.5 algorithm, have been employed to identify patterns in student graduation data and determine influential factors such as GPA and semester performance (Wibowo, Manongga, & Purnomo, 2020). These studies demonstrate the potential of data mining techniques in predicting academic outcomes, with reported accuracies ranging from 75% to 82.79%, providing valuable insights for educational institutions to improve student success rates and inform academic management decisions.

The linear regression and similar algorithm is used in various cases for predicting various types of variable, from financial (Adiguno, Syahra, & Yetri, 2022; Kusuma & Hidayat, 2024), agricultural (Ihsani Raehan, Budiman Kusdinar, & Indrayana, 2024), manufacturing (Absa, Setiawan, Fatwa, & Hidayat, 2023), industrial (Hidayat, Darnis, & Hidayatussa'adah, 2024), healthcare (Harsiti, Muttaqin, & Srihartini, 2022), and also educational (Hariningrum, Yogatama, & Utomo, 2024). Linear regression analysis has been used to predict study duration based on Grade Point Average (GPA), revealing a strong negative correlation between GPA and study length. Other studies have applied linear regression to forecast student specialization choices in computer science programs (Destria, Nurlita, & Terttiaavini, 2023). Logistic regression has been employed to identify factors affecting cumulative GPA, with study program and living arrangements showing potential influence (Tampil, Komaliq, & Langi, 2017). Additionally, log-linear models have been utilized to examine relationships between study duration, admission pathways, and GPA,

demonstrating interactions between these factors. In this research, cumulative grade will be predicted using the method of linear regression with the input of study time. The relation between these two indicator can also be explored.

METHOD

Flowchart of this research is shown in Figure 1 below. First, cumulative grade (IPK) and study time data are collected from four study programs in the Faculty of Teacher Training and Education, Malikussaleh University. The data are from the year 2019 to 2024, with a total of 592 data collected. The variables of this study are divided into two categories: independent variable and dependent variable. The independent variable is a variable that affects the dependent variable, which in this study is assumed as the study time. Meanwhile the dependent variable is a variable affected, which is the cumulative grade (IPK) in this research.

The data are then preprocessed and split for training and evaluation. This is done through the Python programming language in Jupyter Notebook, with the help of libraries such as skicit-learn and pandas (Absa & Setiawan, 2023). The library pandas is used to read data from Ms. Excel file that will be used for training. Then, skicit-learn is used to preprocess the data using minmaxscaler preprocessing to normalize the data, so that it is in the range from 0 to 1. The data are then also split into two: data for training and data for testing (evaluation). Here the size of data for testing is varied with respect to data for training to find the optimal result. The data are then trained using the algorithm of simple linear regression. Mathematically, the algorithm can be written as a formula:

$$Y = a + bX \quad (1)$$

Where:

Y = dependent variable (cumulative grade)
X = independent variable (study time)
a = slope of the regression line
b = y-intercept of the regression line

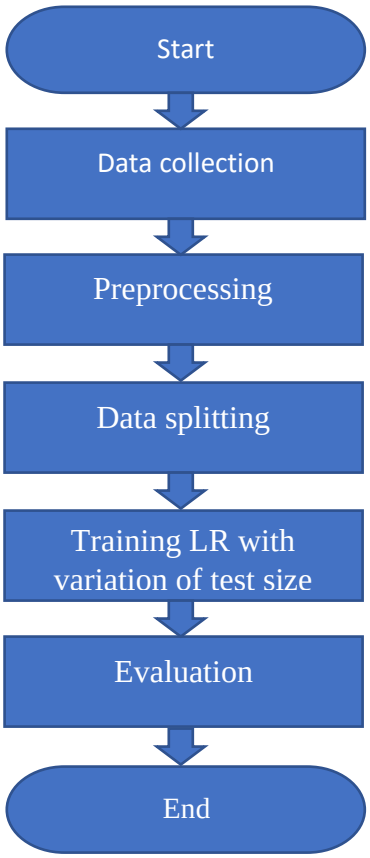

Figure 1: Flowchart of this research

The regression formula are then evaluated to find out which is the optimal one, or in other words which is the best to predict cumulative grade with the input of study time. The evaluation is done with the metric of Root Mean Square Error (RMSE) and R-squared (R^2), through the library skicit-learn.

Root Mean Square Error (RMSE) is used to measure how large is the difference between dependent variable which have been predicted using linear regression to the dependent variable from the actual data. To find RMSE, first we square the difference of each data point (between the prediction and the actual data), then we take its average, and lastly we take a square root of the average. The formula for Root Mean Square Error are given below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Where:

n = sample count

y_i = data point (actual)

$\hat{y}_i$ = data point (predicted)

The R-squared (R^2), which is also called coefficient of determination, is an evaluation metric which gives an indication of the scale of variability in the dependent variable which is predicted by the algorithm. In other words, R-squared states the proportion of the variation in the dependent variable which can be predicted by the independent variable as shown in the formula of linear regression. The value of R-squared is in the range of 0 and 1, where the closer the number to 1, the better the algorithm / formula we used to predict is. Mathematically, the formula of R-squared is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \quad (3)$$

Where:

n = sample count

y_i = data point (actual)

$\hat{y_i}$ = data point (predicted)

$\bar{y_i}$ = average of dependent variable

FINDING AND DISCUSSION

RESEARCH RESULT

The results of Root Mean Square Error and R-squared metric with the variation of test size are shown in the table below. As the test size increase from 2% to around 6%, the R-squared also increase from 0.3631 to 0.6109. Above that, the R-squared started to decrease until a low point for test size of 30%. The Root Mean Square Error, meanwhile, does not follow a particular pattern. From 2% test size to 6% test size it is decreasing from 0.1948 to 0.1481, then it increases and stays around the same value except when the test size is 15%, where RMSE is 0.1455. From this result we can infer that the best variation of test size to make the most accurate prediction of cumulative grade is the test size of 6%.

Table 1: RMSE and R^2 of different test size

Test Size	Train Size	RMSE	R^2
2%	98%	0.1948	0.3631
4%	96%	0.1519	0.5571
6%	94%	0.1481	0.6109
8%	92%	0.1520	0.5696
10%	90%	0.1517	0.5573
15%	85%	0.1455	0.5462
20%	80%	0.1542	0.4634
25%	75%	0.1489	0.4709
30%	70%	0.1493	0.4558

In accordance with the aforementioned result, we then plot the difference between predicted data and actual data using linear regression with 6% test size. The result of that plot is shown in the Figure 2 below. We can see that the predicted value of

cumulative grade (IPK) is not that different from the actual value of cumulative grade (IPK). The largest difference between the predicted value and the actual value that we can see from the graph is around 0.25. This reassures our conclusion before that the linear regression algorithm with the test size of 6% is quite accurate to predict the possible cumulative grade (IPK) from the actual study time.

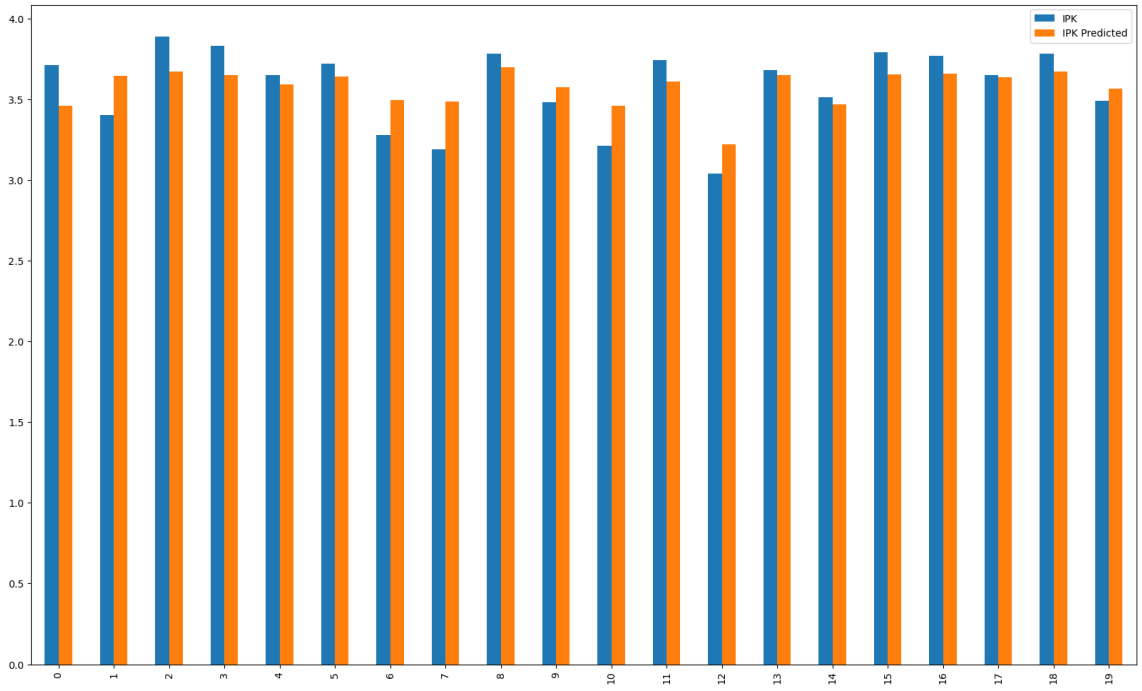

Figure 2: Difference between the actual cumulative grade with the predicted cumulative grade with the test size of 6%.

DISCUSSION

From the results above, we can see that the optimal test size to use for linear regression on this particular dataset is 6%. This means that with 6% test size and 94% training size, we will get the most accurate linear regression formula. There is no general rule for choosing the size of test data or training data to get the best linear regression formula. However, with a relatively low data size (592), we should expect the proportion of the training size to be quite high (Dubbs, 2024).

From these results as well, we see that the Root Mean Square Error metric and R-squared error metric does not always agree. The best (highest) R-squared metric we get is with 6% test size, while the best (lowest) Root Mean Square value we get is when the test size is 15%. This means that the two metric are effected by different variables. Even though the Root Mean Square value at 15% test size is lower than the one at 6%, the difference between them is not too pronounced. In contrast, the R-squared for the 6% test size is

comfortably higher than the 15% test size. This is why we can still consider the 6% test size to be the better one.

The data size that is relatively small also contributed to the metric R-squared which is also relatively low (Rahmawati, Kristanto, Setya Pratama, & Abiansa, 2022). A larger data size might be able to amend this problem and result in a more accurate prediction (Prasetyo, Salahuddin, & Amirullah, 2021). Because of that, the data will have to appended for some more years for the linear regression algorithm to be more accurate.

CONCLUSION

From this research, we found that the most optimal linear regression algorithm for predicting cumulative grade (IPK) from study time is when the test size is 6% and the training size is 94%. In this case, the Root Mean Square Error was found to be 0.1481 and the R-squared was 0.6109. Though the value of these metrics are relatively worse than previous studies, these results can still help determine how to find the optimal test size for this particular type of case. To give a better prediction and to achieve better metrics, data collection had to been done with a longer range of years to the future.

REFERENCES

Absa, M., & Setiawan, T. (2023). Comparison of Different Weight Optimization Algorithm in Neural Network to Predict Mechanical Properties of AAC Lightweight Brick. *Journal of Scientific Research, Education, and Technology (JSRET)*, *2*(1), 235–241.

Absa, M., Setiawan, T., Fatwa, I., & Hidayat, A. T. (2023). MLP neural network in google colaboratory to predict mechanical properties of manufactured-sand concrete. *Jurnal Mantik*, *6*(4), 3679–3687.

Adiguno, S., Syahra, Y., & Yetri, M. (2022). Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, *1*(4), 275. https://doi.org/10.53513/jursi.v1i4.5331

Destria, A., Nurlita, A., & Terttiaavini. (2023). Analisis Prediksi Pemilihan Mata Kuliah Peminatan pada Jurusan Teknik Informatika Universitas Indo Global Mandiri Menggunakan Metode Linier Regresi. *Journal Innovations Computer Science*, *2*(1), 1–6. https://doi.org/10.56347/jics.v2i1.119

Dubbs, A. (2024). Test Set Sizing via Random Matrix Theory. *Operations Research Forum*, *5*(1), 17. https://doi.org/10.1007/s43069-024-00292-1

Hariningrum, R.-, Yogatama, Y.-, & Utomo, S. B. (2024). Pemodelan Estimasi Kelulusan Mahasiswa Berbasis Data Akademik Melalui Regresi Linier Berganda. *INOVTEK Polbeng - Seri Informatika*, *9*(1), 192–202. https://doi.org/10.35314/isi.v9i1.4034

Harsiti, Muttaqin, Z., & Srihartini, E. (2022). Penerapan Metode Regresi Linier Sederhana Untuk Prediksi Persediaan Obat Jenis Tablet. *JSiI (Jurnal Sistem Informasi)*, *9*(1), 12–16. https://doi.org/10.30656/jsii.v9i1.4426

Hartatik, H. (2021). Optimasi Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes. *Indonesian Journal of Applied Informatics*, *5*(1), 32. https://doi.org/10.20961/ijai.v5i1.44379

Hidayat, T., Darnis, R., & Hidayatussa'adah, D. (2024). ALGORITMA REGRESI LINIER BERGANDA UNTUK ANALISIS EFISIENSI STOK PRODUK DI PT. MADU PRAMUKA BATANG. *Jurnal Informatika dan Teknik Elektro Terapan*, *12*(3). https://doi.org/10.23960/jitet.v12i3.4899

Ihsani Raehan, M. F., Budiman Kusdinar, A., & Indrayana, D. (2024). PENERAPAN REGRESI LINIER BERGANDA UNTUK MEMPREDIKSI HASIL PANEN KACANG KEDELAI. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(5), 10572–10579. https://doi.org/10.36040/jati.v8i5.11032

Kusuma, M. D. H., & Hidayat, S. (2024). Penerapan Model Regresi Linier dalam Prediksi Harga Mobil Bekas di India dan Visualisasi dengan Menggunakan Power BI. *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, *5*(2), 1097–1110. https://doi.org/10.35870/jimik.v5i2.629

Malik, H. H., Muhammad, A. H., & Kusnawi, K. (2024). PENERAPAN ALGORITMA MONTE CARLO UNTUK MEMPREDIKSI IPS DAN IPK BERDASARKAN KARAKTERISTIK MAHASISWA PERGURUAN TINGGI X DI KOTA CIREBON. *TECHNOVATAR Jurnal Teknologi, Industri, dan Informasi*, *2*(4), 81–96. https://doi.org/10.61434/technovatar.v2i4.225

Prasetyo, A., Salahuddin, S., & Amirullah, A. (2021). Prediksi Produksi Kelapa Sawit Menggunakan Metode Regresi Linier Berganda. *Jurnal Infomedia*, *6*(2), 76. https://doi.org/10.30811/jim.v6i2.2343

Rahmawati, D., Kristanto, T., Setya Pratama, B. F., & Abiansa, D. B. (2022). Prediksi Pelaku Perjalanan Luar Negeri Di Masa Pandemi COVID-19 Menggunakan Metode Regresi Linier Sederhana. *Journal of Information System Research (JOSH)*, *3*(3), 338–343. https://doi.org/10.47065/josh.v3i3.1507

Tampil, Y., Komaliq, H., & Langi, Y. (2017). Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado. *d'CARTESIAN*, *6*(2), 56. https://doi.org/10.35799/dc.6.2.2017.17023

Wibowo, A., Manongga, D., & Purnomo, H. D. (2020). The Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation. *Scientific Journal of Informatics*, *7*(1), 99–112. https://doi.org/10.15294/sji.v7i1.24241