The Comparison of Ridge Regression Method and Lasso Regression Method to Predict The Graduation Time

Humaira[1] , Nikita Chairunnisa[2], Novi, Yulia Jihan Sy, Rika Idmayanti[3]

{humaira@pnp.ac.id[1],nikitachairunnisa01@gmail.com[2], rikaidmayanti@pnp.ac.id[3]}

Information Technology Department, Politeknik Negeri Padang, Padang, Indonesia

Abstract. Every year, graduates from the department of information technology are produced. There are not as many graduates as there are new students each year. This is as a result of the high rate of late graduations. The Department of Information Technology has a problem with this. Making an intelligent system is the answer to these issues. Intelligent system to estimate students' graduation times. Data from past years' graduations of students was gathered and trained. Utilizing regression to train on training data. The two types of regression are compared in this article: Ridge and Lasso. The prediction model's outputs had an accuracy of 92.22% and an MSE of 0.084, which is the best possible result. Ridge Regression, which produces the best prediction model, was used. Using its coefficients, Lasso Regression can identify the factors that have the greatest impact on the desired value.

Keywords: Prediction, Graduation Time, Ridge Regression, Lasso Regression, MSE.

1 Introduction

There are three study options available at the information technology department. That is, two study programs are Diploma three, whereas one is Diploma four. Students majoring in information technology graduate or graduate each year, and the number of students graduating is increasing year after year. The number of graduates, however, is not proportional to the number of incoming students. This is due to students graduating before completing their specified study period. A three-year diploma program should be finished in three years, and a four-year diploma program should be completed in four years. This is an issue that program management is concerned about. The growing number of students who do not graduate on time will have an influence on the study program's accreditation review. Of course, this will have an impact on the institutional level as well [1][2].

A solution to the above challenge must be found in the form of a system that can estimate the duration of students' study. The trend of Machine Learning in fixing the problem outlined above is undeniable. Past graduation data can be handled in such a way that program management can estimate the length of their students' studies. As a result, program managers can take preventive measures as soon as feasible if they detect forecasts of students surpassing the study duration.

ABEC 2023, September 21, Bengkalis, Indonesia

DOI 10.4108/eai.21-9-2023.2342995

Prediction cases are solved using a variety of ways. This article, on the other hand, employs the Regression method. Lasso regression was employed in the study by Mohammad Robbani and his team to identify the factors affecting Indonesian inflation. This study was able to choose

8 independent variables and decrease them to 6 independent variables by using Lasso regression [3]. The research conducted by Rahmi Susanti used Ridge regression to determine the variables that influence the fertility of fertile women in East Kalimantan, detect multicollinearity, and create a model using the MKT method and Ridge regression method. The VIF (Variance Inflation Factor) values and standard errors from the Ridge regression method are smaller than those of the MKT method, indicating that the Ridge regression method is superior to the MKT method [4].

The research conducted by Ahmad Maulana Malik Fattah and his colleagues aimed to predict car purchase prices using linear regression, ridge regression, lasso regression, random forest regressor, elastic- net, and support vector regressor (SVR) modeling. The research yielded the best-performing model, which was the lasso regression. The evaluation results using the R-squared (R2) test showed a value of 0.99958[5]. These two approaches' prediction models will be examined first before being used to a website-based system.

2 Research Method

All graduates from the information technology departments between 2017 and 2022 make up the population for this study, which includes 824 people. All current Information Technology department students enrolled in the Diploma Three program make up the sample for this study. For this study, the department of information technology itself has compiled secondary data from the graduation records of students from 2017 to 2022. Lasso and Ridge regression are the data analysis techniques used in this study. The Cross Industry Standard Process for Data Mining (CRISP-DM) technique is used as the methodology for this study.

3 Result and Discussion

3.1 Ridge Regression

Ridge Regression is certain technique developed to stabilize the value of regression coefficient

[6] [7]. Its equation (1) is

$$\Sigma^{n}{}_{i=1}\,(yi-\beta 0-\Sigma^{p}{}_{j=1}x_{ij}\beta_j)^2+\lambda\Sigma^{p}{}_{j=1}\beta j^2 \qquad \textbf{(1)}$$

Where $\lambda \geq 0$ is the depreciation parameters and $\Sigma\beta pj{=}1\beta 2j$ is the depreciation penalty. If λ=0, then the depreciation penalty does not give any influence. But, if $\lambda \rightarrow \infty$ then it will impact the depreciation penalty to be bigger and the estimation coefficient gets closer to zero. The optimal depreciation parameter (λ) is determined by cross validation method.

3.2 Lasso Regression

Lasso regression method is one depreciation method of some coefficients from independent variable and the coefficient will approach even determined to be zero [3] [8] for the independent variable which does not influence the dependent variable. Until with this method the unnecessary variable will be selected and it causes the usage of this lasso regression method tobe better compared to the ridge regression[3] [9]. Its equation (2) is

$$\Sigma^{n}{}_{i=1}(yi-\beta 0-\Sigma^{p}{}_{j=1}x_{ij}\beta_j)^2+\lambda\Sigma^{p}{}_{j=1}|\beta j| \quad \textbf{(2)}$$

With explanation, where yi states as the dependent variable of the-ith observation, $\beta\hat{}0$ is the constant, $\beta\hat{}_j$ is the coefficient from the ith independent variable, x_{ij} is the independent variable, N states the amount of observation and p is the am count of independent variable in certain model [10].

3.3 Cross Industry Standard Process for Data Mining (CRISP-DM)

The Cross Industry Standard Process for Data Mining (CRISP-DM) method, which seeks to assess problem-solving approaches from research or business, was utilized to create this final project. The stages of the CRISP-DM methodology are as follows:

Business Understanding
The purpose of this study is to forecast the length of graduation for students majoring in information technology. Furthermore, understanding what factors influence the length of student graduation.

3.4 Data Understanding

Secondary data, notably graduation data from the Information Technology department, was used in this study. The graduation data was gathered from students who graduated between 2017and 2021. This dataset has 824 rows and 12 columns. The 12 columns are as follows: No, gender, address, GPA (Grade Point Average), Program of Study, GPA for semester 1, GPA for semester 2, GPA for semester 3, GPA for semester 4, GPA for semester 5, GPA for semester 6, and the term of study.

3.5 Data Preparation

At this stage, we prepare the data used in the next phase. Of the 12 columns in the dataset, 11 columns are determined with 1 dependent variable. The dependent variable is length of study. Initially the dependent variable had a categorical value, while solving this problem was using a regression approach. Therefore, the dependent variable needs to be transformed into a continuous numerical value. Table 1 is an illustration of the transformation of study length values.

Table 1. Transformation of study length column values

Duration of Study	After Transformation
3 year 0 month	3,00

3 year 1 month	3,08
3 year 2 month	3,17
3 year 4 month	3,33
3 year 5 month	3,42
3 year 6 month	3,50
3 year 7 month	3,58
3 year 8 month	3,69
3 year 9 month	3,75
3 year 10 month	3,83
3 year 11 month	3,92
4 year 0 month	4,00
4 year 7 month	4,58
5 year 0 month	5,00

For the year grade, the length of study is fixed. While the month's value is calculated using the formula (3).

$$\text{Month value} = \frac{\text{number of month in study duration}}{12} \times 100 \quad \textbf{(3)}$$

3.6 Modeling

Two methods were used in the model creation stage to predict students' study duration: the Ridge Regression method and the Lasso Regression method. The simulation uses different amounts of data. Distribution of testing and training data in an 80:20 ratio. Eight experiments with various conditions were carried out in this study. The experiment findings are shown in Table 2.

Table 2. Modeling Result

No	Model	Total Data	Missing Value	Number of Independent Variable	Best of Alpha value	Model accuracy
1.	Ridge Regression	566	no	10	0,99	91,95%
2.	Lasso Regression	566	no	10	0,0	90,87%
3.	Ridge Regression with lasso selection feature	566	no	7	0,99	92,22%
4.	Lasso Regression with lasso selection feature	566	no	7	0,0	92,14%
5.	Ridge Regression	824	yes	10	0,99	87,18%

6.	Lasso Regression	824	yes	10	0,0	88,14%
7.	Ridge Regression with lasso selection feature	824	yes	9	0,99	88,17%
8.	Lasso Regression with lasso selection feature	824	yes	9	0,0	88,14%

The best model is obtained based on the description of the test results in table 2, which is the ridge regression model with the lasso regression selection feature. This model has the highest accuracy (92.22%).

3.7 Evaluation

The Mean Square Error (MSE) was used to analyze both models. Where a low MSE value indicates a good model. Table 3 shows the results of model evaluation using MSE.

Table 3. MSE Result

No	Model	MSE Value
1.	Ridge Regression (566 data)	0,087
2.	Lasso Regression (566 data)	0,100
3.	Ridge Regression with lasso selection feature (566 data)	0,084
4.	Lasso Regression with lasso selection feature (566 data)	0,084
5.	Ridge Regression (824 data)	0,133
6.	Lasso Regression (824 data)	0,133
7.	Ridge Regression with lasso selection feature (824 data)	0,133
8.	Lasso Regression with lasso selection feature (824 data)	0,133

MSE was used to analyze the model that had been created. The model with the lowest MSE value is the best. The model with the lowest MSE value in Table 3 was model number 7-8, which was Ridge Regression with lasso selection feature model and Lasso Regression with lasso selection feature model with MSE value 0,084, indicating that this model performed well.

3.8 Deployment

For integration into the web-based system, the model with the lowest MSE (Mean Squared Error) value was chosen. The model was integrated into the web-based system utilizing the Stream lit framework and the Python programming language. Figure 1 is an illustration of the system.

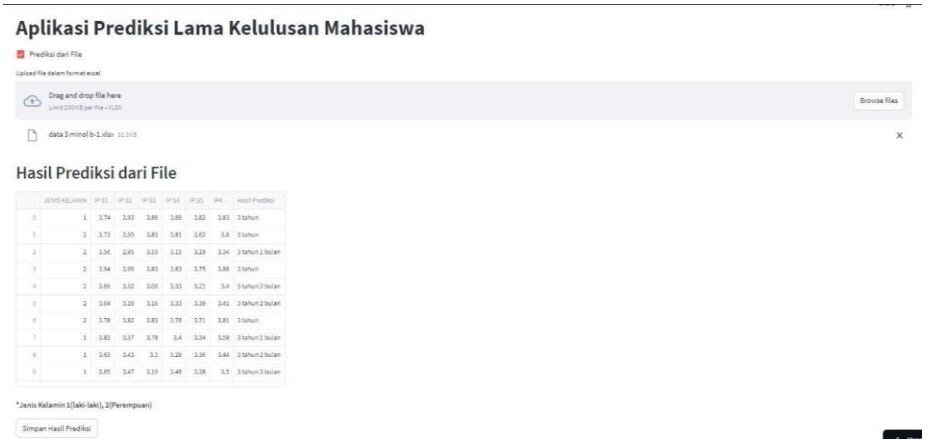

Figure 1. Deployment

4 Conclusion

The results of this study indicate that the Lasso regression model is capable of determining the most influential features on the prediction outcome. There are seven features that significantly affect the graduation duration of students, namely: gender, GPA for semester 1, GPA for semester 2, GPA for semester 3, GPA for semester 4, GPA for semester 5, and overall GPA . On the other hand, the Ridge regression model exhibits better accuracy than the Lasso regression. The accuracy of the Ridge regression model, with the application of Lasso feature selection, is 92.22%, with an MSE of 0.084.

The accuracy results obtained are already quite good. A suggestion for further research is to analyze this model for overfitting issues. Because high accuracy does not guarantee that this prediction system is reliable under all data conditions.

Acknowledgement

Thank you to P3M of Politeknik Negeri Padang that had facilitated the publication of this article.

References

[1] Wirawan, “Teknik Data Mining Menggunakan Algoritma Decision Tree C4.5 untuk Memprediksi Tingkat Kelulusan Tepat Waktu,” Appl. Inf. Syst. Manag., vol. 3,no. 1, pp. 47–52, 2020, doi: 10.15408/aism.v3i1.13033.

[2] I. N. Rudy Hendrawan, I. M. A. Budhi Saputra, G. A. P. Cahya Dewi, I.

[3] G. S. Adi Pranata, and N. L. N. Wedasari, “Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes,” J. Eksplora Inform., vol. 11, no. 1,pp. 50– 56, 2022, doi: 10.30864/eksplora.v11i1.606.
[4] M. Robbani, F. Agustiani, and N. Herrhyanto, “Regresi Least Absolute Shrinkage andSelection Operator (Lasso) Pada Kasus Inflasi Di Indonesia Tahun 2014-2017,” EurekaMatika, pp. 1–16, 2019.
[5] R. Susanti, C. D. Giyatri, and I. AB, “Penerapan Metode Regresi Ridge dalam Mengatasi Multikolinieritas pada Tingkat Fertilitas Wanita Usia Subur,” JI-KES(Jurnal Ilmu Kesehatan), vol. 5, no. 1,
[6] pp. 91–102, 2021, doi: 10.33006/ji- kes.v5i1.214.
[7] F. Rahmawati and R. Y. Suratman, “Performa Regresi Ridge Dan Regresi Lasso PadaData Dengan Multikolinearitas,” Leibniz J. Mat., vol. 2, no. 2,
[8] pp. 1–10, 2022.
[9] G. W. Kusuma and I. Y. Wulansari, “Analisis Kemiskinan Dan Kerentanan Kemiskinan Dengan Regresi Ridge, Lasso, Dan Elastic-Net Di Provinsi Jawa TengahTahun 2017,” Semin. Nas. Off. Stat., vol. 2019, no. 1, pp. 503–513, 2020, doi: 10.34123/semnasoffstat.v2019i1.189.
[10] M. Mravik, T. Vetriselvi, K. Venkatachalam, M. Sarac, N. Bacanin, and
[11] S. Adamovic,“Diabetes prediction algorithm using recursive ridge regression l2,” Computer. Contin., vol. 71, no. 1, 2022, doi: 10.32604/cmc.2022.020687.
[12] D. R. Ningsih, P. K. Intan, and D. Yuliati, “Pemodelan Tindak Pidana Kriminalitas di Kota Tangerang Menggunakan Metode Regresi Lasso,” ESTIMASI J. Stat. …, vol. 4, no. 1, pp. 64–77, 2023, doi: 10.20956/ejsa.vi.24853.
[13] Y. Wu, “Can’t Ridge Regression Perform Variable Selection?,”
[14] Technometrics, vol.63, no. 2, 2021, doi: 10.1080/00401706.2020.1791254.
[15] N. A. Bahmid, “Metode Least Absolute Shrinkage And Selection Operator Untuk Mengatasi Multikolinearitas Pada Regresi Logistik Ordinal,” no. November, 2018.