JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 4, No. 1, June 2021, pp. 114~120
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 36/E/KPT/2019
DOI: 10.36378/jtos.v4i1.1401

BINARY LOGISTIC REGRESSION IN DETERMINING AFFECTING FACTORS STUDENT GRADUATION IN A SUBJECT

Shedriko
Department of Informatics Engineering, PGRI University of Indraprasta, Indonesia

Article Info

Article history:

Received 03 June, 2021
Revised 09 June, 2021
Accepted 09 June, 2021

Keywords:

logistic
regression
binary
graduation
quantitative

ABSTRACT

Good communication and coordination between lecturers are needed in delivering material by different lecturers to ensure the relatively uniform quality of education. Knowing the success information from several classes to predict other classes, should be completed by significant parameters used in the algorithm. This research is using a quantitative analysis method with binary logistic regression methodology in determining critical factors of train data on "Introduction to Information Technology" subject in the university of XYZ. Several statistical testing are conducted to give the expected results using software excel with Real Statistics add-ins and Orange Data Mining in testing the pass-prediction from the given data training. The successive model can also be used to classify graduation for the different subjects.

Corresponding Author:

Shedriko
Department of Informatics Engineering
PGRI University of Indraprasta
Jakarta, Indonesia
Email: shedriko@gmail.com

1. Introduction

The affordable cost of education at university is not factor for university is not qualified. However, the quality of educators constantly and continuously in improving knowledge is the factor to determine the quality of university, such as improving teaching discipline, conducting continuous research, providing good educational services and providing organized scholarship programs, etc. Increasing the student discipline in attendance, doing assignments and being active in class also become supporting factor for quality of university.

The research was conducted at XYZ University, Faculty of Technology and Computer Science, Department of Informatics Engineering in the Introduction to Information Technology *(Pengantar Teknologi Informasi)* Subject. This department has about 50 classes per semester and 11 lecturers for one subject. This condition makes good communication and coordination between lecturers needed in giving material from particular subject by several lecturers to ensure uniformity in the quality of education.

This research conducted several statistical tests to answer the questions about the factors of affect student graduation in a subject by using data mining software to obtain graduation predictions from testing data. Meanwhile, the purpose of this research to obtain model to predict or to classify graduation in other subject

2. Research Method

This research used binary logistic regression methodology as problem solving algorithm and quantitative analysis method [1] in completing this research. As part of simply statistical model towards complex and messy data [2], binary logistic regression is used to model binary variable (0, 1) based on one or more variables named predictor [2]. 2]. While quantitative analysis is research based on numbers or quantity calculations for all phenomena related to numeric [1]. So it can be concluded that this research was conducted by analyzing the numbers based on certain formulas in determining the effect factors for student graduation.

The total population was 175 students of Information Technology major in *pengantar teknologi informasi*'s subject. The input parameters taken are assignments (HomeWork), midterm exams (MidTest) and final exams (FinalTest). Information on the values obtained are categorized into:

1. 0 - <56, represented by binary value = 0
2. >=56, represented by binary value = 1

Calculations are carried out using "real statistics" tools in add-ins in the excel application, some test values can be accumulated automatically. The steps taken are [3], [4]:

1. Logit transformation to see pass/fail classification
2. Regression coefficient calculation
3. Determination of logistic regression model/formula
4. Perform Partial or Wald Tests
5. Perform Simultaneous Test or Deviance and Hosmer Test
6. Calculating Odds Ratio
7. Interpretation
8. Pass prediction test
9. Conclusions

Logistic regression is one of the most popular estimation methods, because it gives an estimation range between 0 and 1 [5]. The following is the logistic regression formula along with other formulas:

2.1 Logistics Regression

It is a statistical analysis method to describe the relationship between two or more dependent variables (pass = 1 and fail = 0) with independent variables (HomeWork, MidTest and FinalTest) [3]

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i)} \quad (1)$$

with:

$\pi(x)$: *regresi logistik*
β_i *: koefisien regresi ke i*
X_i *: variabel bebas ke i*

2.2 Logit Transform

The equation is transformed into logistic regression logit form to interprete of the regression parameters [6]

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \quad (2)$$

with:

$g(x)$: logit $\pi(x)$

2.3 Pearson Chi Square

It is a type of non-parametric comparative test on two variables of the nominal data scale of two variables [4], [7]

$$X_p^2 = \sum_{i,j} \frac{(f_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

with:

X_p^2 : chi square
f_{ij} : *nilai observasi / pengamatan baris ke-i kolom ke-j*
E_{ij} *: nilai ekspektasi baris ke-i kolom ke-j*

2.4 Wald Test

It is used to test the presence or none of the influence of the independent variable to the dependent variable partially by comparing the Wald statistical value with the comparison value of Chi Square [3], [8]. The hypothesis used is as follows [6]:

2.4.1 Condition

H_0 : $\beta_i = 0$ (There is no significant effect between independent and dependent variables)
H_1 : $\beta_i \neq 0$ ((There is a significant effect between the independent and dependent variables)

2.4.2 Significance

The level of significance (α) = 5% is the level of confidence in the statistics of 95%, which is obtained from 100 - 95 = 5% [9]

2.4.3 Formula

It use the following formula [3]

$$W = \frac{\widehat{\beta_i}}{SE(\widehat{\beta_i})} \quad (4)$$

with:

W : *Uji Wald*
$\widehat{\beta_i}$ *: nilai dugaan untuk parameter* (β_i)
$SE(\widehat{\beta_i})$ *: dugaan galat baku (standar error) untuk koefisien* β_i

2.4.4 Decision

Reject decision H_0 if $W > x^2_{(0,05;1)}$ or W > *p-value* (on excel)

2.5 Simultaneous Test

It also known as Deviance Test, it is carried out to test whether the resulting model based on multivariate/simultaneous logistic regression is feasible or mutual accord [4]. The hypothesis is as follows:

2.5.1 Condition

H_0 : Model accord to data
H_1 : Model doesn't accord to data

2.5.2 Significance

Significance Level (α) = 5% [9]

2.5.3 Formula

It use the formula of Hosmer-Lemeshow Goodness of Fit as follow:

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1-\bar{\pi}_k)} \quad (5)$$

with:

O_k : *observasi pada grup ke-k* $(\sum_{j=1}^{C_k} y_j$ *dengan* c_k *: respon (0,1))*
$\bar{\pi}_k$: *rata-rata taksiran peluang* $(\sum_{j=1}^{C_k} \frac{m_j \hat{\pi}_j}{n'_k})$
g : *jumlah grup (kombinasi kategori dalam model serentak)*
n'_k : *banyak observasi pada grup ke-k*

2.5.4 Decision

Reject Decision H_0 if $\hat{C} > x^2_{(0,05;1)}$ or *p-value* > α (on excel)

2.6 Odds Ratio

It is compilation of divisible odds by other odds [3], [10]

$$\Psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \quad (6)$$

If $\Psi = 1$ so there is no relationship between two variables, and if $\Psi < 1$ so the relationship between two variables is negative, and if $\Psi > 1$ so the relationship is positive.

3. Result and Discussion

By using the "real statistics" tools in excel application software, it was obtained the classification of pass/fail as a result of the logit transformation as shown in table 1 below.

Table 1. Pass/Fail from Logit Transformation

Homework	*MidTest*	*FinalTest*	*Success*	*Failure*	*Total*
0	0	0	0	10	10
0	0	1	0	1	1
0	1	0	0	5	5
0	1	1	0	4	4
1	0	0	5	11	16
1	0	1	16	1	17
1	1	0	27	3	30
1	1	1	91	1	92
			139	36	175

The regression coefficient values obtained from the results of calculations in excel application with add-ins real statistics, it can be seen in table 2 below.

Table 2. Regression Coefficient

	coeff b
Intercept	-25,0528
Homewor	24,38874
MidTest	2,708807
FinalTest	3,046053

From the values of these coefficients was obtained the logistic regression equation as follows below.

$$\pi(\mathrm{x}) = \frac{\exp\left(-25{,}0528 + 24{,}38874\,X_1 + 2{,}708807\,X_2 + 3{,}046053\,X_3\right)}{1 + \exp\left(-25{,}0528 + 24{,}38874\,X_1 + 2{,}708807\,X_2 + 3{,}046053\,X_3\right)}$$

Then for Partial Test or Wald test, the calculation results from the application can be seen in table 3 below

Tabel 3. Partial Test Result

	coeff b	*s.e.*	*Wald*	*p-value*
Intercept	-25,0528	7446,282	1,13E-05	0,997316
Homewor	24,38874	7446,282	1,07E-05	0,997387
MidTest	2,708807	0,700234	14,96475	0,00011
FinalTest	3,046053	0,834707	13,317	0,000263

p-value is the result of the chi-square calculation of the value of Wald test. In the "Homework" row, it appears that Wald Test value is < from p-value, while the other rows show Wald Test > from p-value. It means that the value of homework variable (X_1) is important or decisive factor for student graduation from the training data in this study.

Furthermore, the Simultaneous Test is calculated automatically by real-statistics from excel, itcan be seen in Table 4 below.

Table 4. Simultaneous Test result

Chi-Sq	119,3008
df	3
p-value	1,09E-25
alpha	0,05
sig	yes

The p-value is the result of the chi-square calculation of all the data obtained from the calculations on the Chi-Sq row with df, so the p-value has the value of the significance coefficient. If the p-value < alpha value, it is said that the model match the data. To strengthen the hypothesis, it can be seen in the results of the Hosmer test in table 5 below.

Table 5. Hosmer Test Result

Hosmer	0,603708
df	6
p-value	0,996338
alpha	0,05
sig	yes

The opposite from the Simultaneous Test, in Hosmer Test if p-value > alpha value (or Hosmer value < p-value) the significance is good or the model match the data. From the two tests, it appears that the model made match with given train data.
Then, from the results of Wald Test, the EXP (β) value wasused to interpret the Odds Ratio value. These values can be seen in table 6 below.

Table 6. *Odds Ratio*

	exp(b)
Intercept	1,32E-11
Homewor	3,91E+10
MidTest	15,01136
FinalTest	21,03216

From table 6, it can be seen the value of the odds ratio for all variables. In Homework variable,it was obtained large value and largest among the 3 variables, it is 3.91 x 1010. It means that for students always do assignments or have good grades, the probability of passing is $3{,}91 \times 10^{10}$ times than the students do not do assignments or have less grades. This large value is possible because there are no students who pass if the assignment score is bad or equal to 0 (zero). Homework variable ranks first for the chances of passing above. The odds ratio value for the Mid-test variable is 15.01, it means that the chance of students was taking the mid-test and getting a good score is 15.01 times than the students was taking the exam but get a bad score. While the odds ratio for Final Test variable is 21.03. It means that the probability of students who pass the exam and get a good score is 21.03 times than the students was taking the final test but get a bad score. Final Test odds rank second from three passing variables above.

From the existing train data, a passing test was carried out by using Orange Data Mining. This software is powerfull free data mining software. Various reviews about this software show high value than the other free software. Figure 1 below shows the test results from data model of the train data in this research.

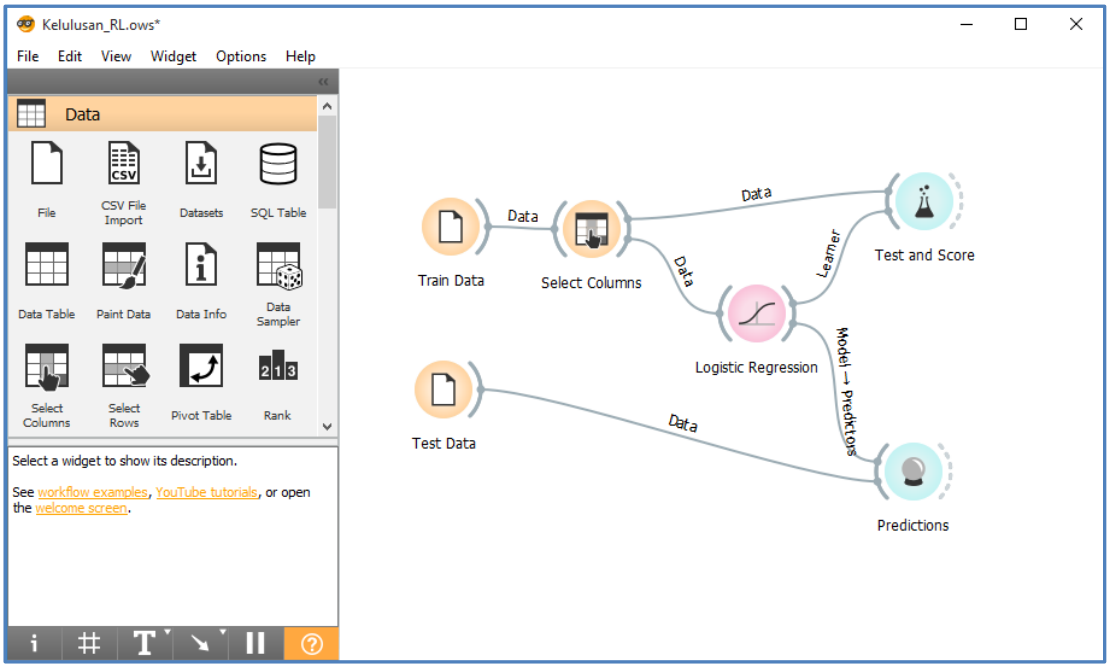

Figure 1. Data Model using Orange Data Mining

.

While the prediction results can be seen in Figure 2 below.

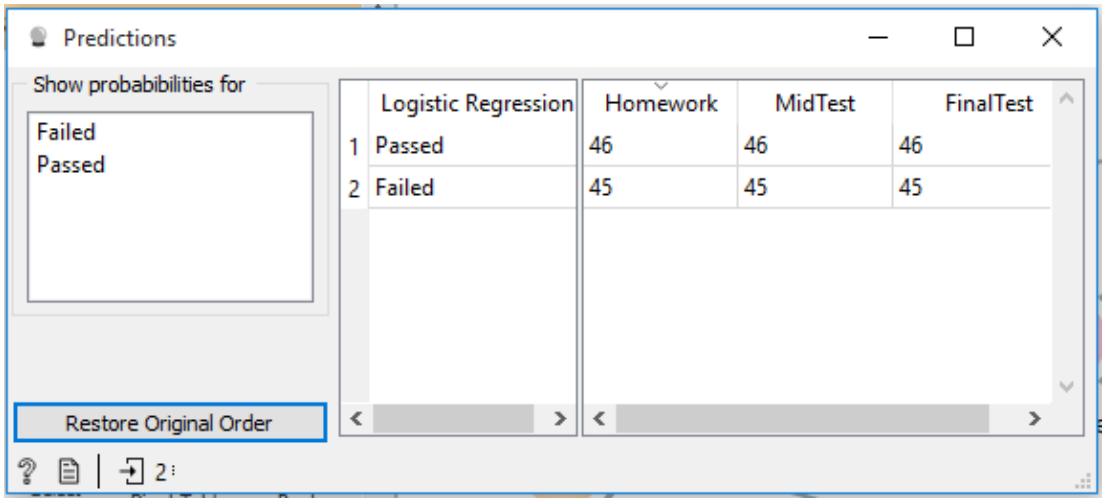

Figure 2. Prediction results with Orange Data Mining

From the results of testing by using Orange Data Mining, the graduation limit value is obtained, there are: the failed score at 45 and the pass score at 46. This relatively low passing score is very dependent on the train data being used basic processing. The train data in this research provides fairly large tolerance for student graduation. The logistic regression methodology provides fairly high level of precision, it is 0.892.

4. Conclusion

The results of this research indicate high accuracy value can be obtained by using logistic regression methodology in predicting student graduation. It shown from evaluation of tests used Orange Data Mining. The training data has high tolerance for student graduation, it provides limitation for low pass value from the results of testing data set. The logistic regression formula is obtained through calculations being application so the coefficients of several independent variables, such as: homework, mid-test and final-test. Then tested the formulation by conducting Simultaneous and Hosmer Test, as the result is obtained the model match the data. And then conducted the Wald Test, it result the homework variable is very significant factor in student graduation. With good homework score provides large opportunity for graduation. It also shown in calculation

of Odds Ratio that large opportunity indicated by large odds value for the homework variable. Then the large odds value is followed by the final test variable and the last is the midtest variable.

There are many machine learning methodologies can be used to make various predictions of the problem. One of them have highest precision in these predictions, it is logistic regression. It is interesting to do comparison between these methodologies and logistic regression, to get results to help many parties in making decisions. For example in predicting student graduation as in this research. The research compares these methodologies is recommended to find out the highest precision, the possibility of knowing the factors affect the results, the difficulties and convenience experienced during making predictions from each of these methodologies, and so on. And the results of this research can be used as comparative information for further research.

Acknowledgement

It fully thanks to the University of Indraprasta PGRI and all its leaders for the support given completing this research.

References

[1] C. R. Kothari, *Research Methodology, Methods & Techniques*, Second Rev. New Age International (P) Ltd., 2004.

[2] J. M. Hilbe, "Practical Guide to Logistic Regression." CRC Press Taylor & Francis Group, p. 170, 2015.

[3] Y. A. Tampil, H. Komalig, and Y. Langi, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado," *d'CARTESIAN*, vol. 6, no. 2. Journal of Dedicators Community Vol. 6 No. 2 September 2017, pp. 57–62, 2017.

[4] Y. Anggraeni and I. Zain, "Pemodelan Regresi Logistik Biner Terhadap Peminat ITS di Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN) 2014." Jurnal Sains dan Seni ITS Vol. 4 No. 1 2015, 2015.

[5] D. G. Kleinbaum and M. Klein, "Statistics for biology and health: Logistical regression." pp. 1–709, 2010.

[6] A. Wulandari, F. M. Faruk, and F. S. Doven, "Penerapan Metode Regresi Logistik Biner Untuk Mengetahui Determinan Kesiapan Rumah Tangga Dalam Menghadapi Bencana Alam." Seminar Nasional Official Statistics 2019: Pengemabangan Official Statistics dalam Mendukung Implementasi SDG's, 2019.

[7] A. Hidayat, "Tutorial Rumus Chi Square Dan Metode Hitung," 2012. [Online]. Available: https://www.statistikian.com/2012/11/rumus-chi-square.html. [Accessed: 29-Dec-2020].

[8] A. Widarjono, *Analisis Statistika Multivariat Terapan, Edisi pertama*. Yogyakarta: UPP STIM YKPN, 2010.

[9] R. Stats, "Uji Z - Uji Hipotesis Rata-rata Satu Populasi," 2020. [Online]. Available: https://www.rumusstatistik.com/2017/01/uji-z-uji-hipotesis-rata-rata-satu-populasi.html. [Accessed: 27-Dec-2020].

[10] F. E. J. Harrell, "Regression Model Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Second Edition," *Statistical Methods in Medical Research*, vol. 13, no. 5. Springer, pp. 415–416, 2015.