Universiti Teknologi MARA
Cawangan Perlis
Jurnal Intelek

Volume 18 Issue 1 (February) 2023

Logistic Regression in Determining Affecting Factors Student Success in an Introductory Statistics Subject

Nur Syuhada Muhammat Pazil[1], Norwaziah Mahmud[2*], Nuridawati Baharom[3], Siti Hafawati Jamaluddin[4]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Melaka, Kampus Jasin, 77300 Merlimau, Melaka, Malaysia
[2,3,4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Perlis, Kampus Arau, 02600 Arau, Perlis, Malaysia

*Authors' Email Address: [1]syuhada467@uitm.edu.my, [*2]norwaziah@uitm.edu.my, [3]nuridawati@uitm.edu.my, [4]hafawati832@uitm.edu.my*

Received Date: 18 October 2022
Accepted Date: 16 December 2022
Revised Date: 2 January 2023
Published Date: 31 January 2023

*Corresponding Author

ABSTRACT

This study aims to find the best model for predicting students' success based on a binary logistic regression. This analysis was also used to determine the factor that affects student success in Statistics Subjects. Five different data partitioning sets were used. The results indicate that the data with a partitioning set of 70% for the estimation set and 30% for the evaluation set is the best fit model using six independent variables. The predictors under investigation were assessment achievements such as test 1, test 2, quiz, assignment, group project, and final test marks. The outcome showed a significant difference in test 2 and the final test marks in determining the factor affecting the subject's result. Besides, the overall model explained further that 95.8% of the sample was classified correctly. This study was carried out using SPSS software and excel. In order to determine the significant variables, further research can be done using the linear regression analysis method.

***Keywords**: accuracy, best fit model, logistic regression, prediction, subject*

INTRODUCTION

The quality of lecturers is a factor that determines the quality of a university, such as strengthening teaching discipline, performing continuous research, and providing good educational services. Another element contributing to university excellence is increasing student behavior in attendance, assignment completion, performance in assessments, and class participation (Shedriko, 2021). Logistic regression analysis is widely and popularly used analysis similar to linear regression analysis to determine the relationship between dependent and dependent variables. One of the differences between these analyses is that the dependent variable of logistic regression is dichotomous (Alija, 2015).

Several studies in Logistic Regression have been successfully employed in educational research. For example, Hu and Hu (2021) employed logit regression to conduct an empirical analysis of students'

ISSN: 2231-7716 / E-ISSN 2682-9223
DOI: http://10.24191/ji.v18i1.20133

performance in order to investigate the impact of students' gender characteristics on their subjects taken. It has been discovered that student gender has a considerable impact on student performance in different subject. Wang et al. (2021) examine whether numerous parameters influence on students passed nuclear professional English using binary logistic regression. It was found that, curriculum arrangement, review of examination, attention in class, and online learning are related to test passing.

Manieri et al. (2015) studied logistic regression analysis of data to see which of three pre-admission assessments best predicts achievement in an associate degree in nursing programme. The admission assessment (A2) examination and the Test of Essential Academic Skill (TEAS) examinations have been demonstrated to be statistically significant in predicting nursing programme success. In the same way, Wanvarie and Sathapatayavongs (2007) used the demographic factors and academic records to predict students' performance in the Medical Licensing Examination of Thailand (MLET) examination.

The findings in Wambuguh and Yonn-Brown (2013) showed that the performance on the quizzes is related to the performance on the final examination by using logistic regression analysis. These findings highlight the importance of educators' engagement in providing frequent formative feedback to students; the importance of students' self-evaluation and staying on track with course material as the semester progresses; and the need for students to be proactive in seeking help early in the semester to understand course content and improve their grades.

Besides, Suliman et al. (2014) studied the predicting students' success at the medical Pre-University Studies by investigating their national high school exit examination-Sijil Pelajaran Malaysia (SPM) achievements, gender and high school background factors. In terms of the logistic model, it was found that the probability is nearly identical to the actual case. Similarly, Baars et al. (2017) developed a model for predicting students who will fail to pass the first year of the undergraduate medical curriculum within two years of starting based on pre-admission and post-admission variables.

Jeslet et al. (2021) used linear regression and Support Vector Machine (SVM) to predict the results of students in Covid-19 lockdown based on their register number, the number of arrears, total mark of all the previous semester. As a result, it can be concluded that this technique prediction system is the most pressing need in this pandemic period. Furthermore, Adnan et al. (2021) found that teaching experience, number of family, enough number of devices, enough internet data and convenience are significance influenced the preparedness of open and distance learning (ODL) among the lecturers in Universiti Teknologi MARA (Pahang).

From the previous studies, it is clear that logistic regression can determine the relationship between two variables in education. Hence, this study aims to determine the factors that affect student success in an introductory statistics subject using Binary Logistic Regression analysis. Meanwhile, this study aims to develop a model for predicting or classifying students' success in other subjects.

METHODOLOGY

This study used secondary data obtained from all students who took the subject introduction to statistics in UiTM Cawangan Melaka Kampus Jasin. There were 236 undergraduate students in the Faculty of Plantation and Agrotechnology who took that subject in two semesters of 2021. The input parameters taken are test 1, test 2, quiz, assignment, group project, final test, and overall marks.

Common practice divides data into two parts: estimation and evaluation sets; for example, 70% of data is for estimation and 30% for evaluation. However, this study used several sets of estimation and evaluation with different percentage values since this paper does not make any comparison with other models. Partitioning data for estimation sets and evaluations sets are being used because it helps to build more robust and accurate models. Table 1 shows five sets of estimation and evaluation percentages to be tested.

Table 1: Partitioning data for Estimation Sets and Evaluation Sets

Set	Estimation Sets	Evaluation sets
1	65%	35%
2	70%	30%
3	75%	25%
4	80%	20%
5	85%	15%

Binary Logistic Regression

Binary logistic regression is used to model the relationship between the dichotomous dependent variables and multiple independent variables, either continuous or categorical. It estimates the probability of occurrence of an event by fitting data to a logistic curve. The dependent variable is the population proportion or probability that the resulting outcome is equal to 1. Parameters obtained for the independent variables can estimate odds ratios for each of the independent variables in the model. The dependent variable in this study is overall marks (results) and the independent variables are test 1, test 2, quiz, assignment, group project and final test. The specific form of the logistic regression model is:

$$P_i = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+....+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+....+\beta_n x_n}} \tag{1}$$

The transformation of the conditional mean P_i logistic function is known as the logit transformation:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2}$$

where P_i is the probability of the outcome of interest or event,
β_0 is the intercept, $\beta_1,\ldots,$ β_n are regression coefficients, x_1, x_2 $,\ldots x_n$ are independent variables. (Chatterjee & Hadi, 2006)

For this study, the dependent variable is success of failure and six independent variables, the binary logistic regression model is estimated to be:

$$\begin{aligned}\ln\left(\frac{P_i}{1-P_i}\right) &= \beta_0 + \beta_1 Test1 + \beta_2 Test2 + \beta_3 Quiz + \beta_4 Assignment \\ &+ \beta_5 Group\,\mathrm{Pr}\,oject + \beta_6 FinalTest\end{aligned} \tag{3}$$

All independent variables used in this study was a continuous variable.

The goodness of Fit of the Model

To determine the best-fit model, the model must be validated using a number of statistical criteria. There are four statistical criteria used:

i. *Omnibus Model Coefficient Test*
The test determines whether the model matches the data. If the p-value is less than 0.05, the model fits the data very well. If not, it indicates that the model does not fit the data. The model with the lowest p-value is the best model.

ii. *Nagelkerke's R Square*
Nagelkerke's R^2 is an adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1. The model with the largest R^2 statistic is the best model (Nagelkerke, 1991)

iii. *Hosmer and Lemeshow Test*
Lemeshow and Hosmer test can be used to test of fitness for the logistic regression model. If the p-value is less than 0.05, the model does not good fit for the data (Hosmer et al., 2013). The model with the highest p-value is the best model.

iv. *Sensitivity, Specificity Analysis and Classification Accuracy*
Sensitivity is the percentage of cases that had the observed characteristic which are correctly predicted by the model. Specificity is the percentage of cases that did not have the observed characteristic and were also correctly predicted as not having observed characteristic. Classification accuracy is used to evaluate the model's efficiency by comparing the observed variable response categories to the expected variable response categories. It discussed the model's predictive performance. The best model shows the highest percentage.

RESULTS AND DISCUSSIONS

Descriptive analysis

Tables 2 and 3 show the descriptive analysis for categorical and continuous variables, respectively. This represented 236 students who took the subject introduction to statistics and they came from the Faculty of Plantation and Agrotechnology in UiTM Cawangan Melaka Kampus Jasin. Majority of students were male (n=136, 57.6%). The highest mark for Test 1 was 97%, while for Test 2, Quiz and Assignment were 100%. The highest marks for the Group Project and Final Test were 94% and 92%, respectively. For the Overall Marks, most students get 55% marks (n=8, 3.4%). There were 26 students who failed and 210 who succeeded in this subject.

Table 2: Descriptive Analysis for Categorical variable

Variable		Frequency	Percentage
Gender	Male	136	57.6
	Female	100	42.4
Result	Failure	26	11.0
	Success	210	89.0

Table 3: Descriptive Analysis for Continuous variable

Variable	Minimum	Maximum	Mean	Mode		
	Marks	Marks	Marks	Marks	Frequency	Percentage
Test 1	2.00	97	41.5	20	8	3.4
Test 2	0	100	46.8	85	14	5.9
Quiz	0	100	46.3	90	31	13.1
Assignment	0	100	43.2	10	31	13.1
Group Project	0	94	76.5	80	34	14.3
Final Test	4	92	61.1	61	7	3
Overall marks	9	92.1	63.3	55	8	3.4

Binary Logistic Analysis

Binary Logistic Regression was applied for determining the factors student's success in subject introduction to statistics. All factors (test 1, test 2, quiz, assignment, group project and final test) were included in the model using five different sets of data partitioning. The dependent variable is overall marks (results) is categorized into 1= success and 0=failure. The results of performing the logistic regression from the evaluation sets of the five sets of data partitioning were tabulated and analysed in Table 4 ,5,6,7 and 8.

Table 4: Parameter Estimate from Binary Logistic Model Set 1 (65% Evaluation sets 35% Estimation sets)

Variables	Coefficient	Wald Statistics	Significant (p-value)	Exp(B) Odds Ratio
Constant	-12.073	3.257	0.071	0.000
Test 1	-0.030	0.112	0.737	0.971
Test 2	0.404	3.608	0.058	1.497
Quiz	-0.037	1.330	0.249	0.964
Assignment	0.041	0.236	0.627	1.042
Group Project	-0.009	0.024	0.876	0.991
Final test	0.252	6.650	0.010	1.286

Table 5: Parameter Estimate from Binary Logistic Model Set 2 (70% Evaluation sets 30% Estimation sets)

Variables	Coefficient	Wald Statistics	Significant (p-value)	Exp(B) Odds Ratio
Constant	-11.410	4.933	0.026	0.000
Test 1	0.025	0.151	0.698	1.026
Test 2	0.201	3.866	0.049	1.223
Quiz	-0.050	0.744	0.379	0.951
Assignment	-0.057	1.449	0.229	0.945
Group Project	0.017	0.142	0.707	1.017
Final test	0.221	12.177	0.000	1.247

Table 6: Parameter Estimate from Binary Logistic Model Set 3 (75% Evaluation sets 25% Estimation sets)

Variables	Coefficient	Wald Statistics	Significant (p-value)	Exp(B) Odds Ratio
Constant	-12.847	7.108	0.008	0.000
Test 1	0.035	0.707	0.400	1.036
Test 2	0.067	5.539	0.019	1.069
Quiz	-0.014	0.335	0.563	0.986
Assignment	-0.042	1.703	0.192	0.959
Group Project	0.0237	0.733	0.392	1.035
Final test	-13.660	15.300	0.000	1.268

Table 7: Parameter Estimate from Binary Logistic Model Set 4 (80% Evaluation sets 20% Estimation sets)

Variables	Coefficient	Wald Statistics	Significant (p-value)	Exp(B) Odds Ratio
Constant	-13.872	8.220	0.004	0.000
Test 1	0.047	1.598	0.206	1.048
Test 2	0.068	5.775	0.016	1.070
Quiz	-0.010	0.203	0.653	0.990
Assignment	-0.053	3.869	0.049	0.948
Group Project	0.039	0.908	0.341	1.040
Final test	0.249	16.475	0.000	1.283

Table 8: Parameter Estimate from Binary Logistic Model Set 5 (85% Evaluation sets 15% Estimation sets)

Variables	Coefficient	Wald Statistics	Significant (p-value)	Exp(B) Odds Ratio
Constant	-14.134	8.549	0.003	0.000
Test 1	0.050	1.878	0.171	1.051
Test 2	0.069	6.163	0.013	1.072
Quiz	-0.009	0.173	0.677	0.991
Assignment	-0.055	4.374	0.037	0.946

Group Project	0.041	0.978	0.323	1.042
Final test	0.251	16.551	0.000	1.285

Table 4 shows only one variable was significant which is the final test since *p*-value 0.010 is less than alpha=0.05. From Tables 5 and 6, it can be noticed that two variables have a *p*-value less than alpha which are test 2 and final test. This indicates that these two variables are significant for set 2 and set 3 binary logistic models. Table 7 and table 8 show three significant variables: test 2, assignment and final test. The statistical criteria goodness of fit of the model was used to find the best model from the five sets of data partitioning.

Selection of the Best Model

The goodness of fit test results based on five sets of models are shown in Table 9.

Table 9: Best Fit Model Analysis

Best Fit Model	Significant Criteria	Estimation Sets				
		1	2	3	4	5
Omnibus	*p*-value close to 0	0	0	0	0	0
Hosmer and Lemeshow	*p*-value close to 1	0.937	1	0.999	1	1
Cox & Snell's	Highest R-square	0.501	0.578	0.44	0.378	0.437
Nagelkerke		0.841	1	1	1	1
Sensitivity	Highest percentage	92.90%	100.00%	100.00%	100.00%	100.00%
Specificity		97.10%	100.00%	100%	100.00%	100.00%
Accuracy		96.40%	100.00%	100%	100.00%	100.00%

In Table 9, all sets show the p-value less than 0.05 for the Omnibus test. For Hosmer and Lemeshow test, all sets computed that p-value is more than 0.05. This indicated that, all the models good fit for the data. At the same time, sets 2, 4, and 5 have the highest p-value=1. The R-square in Cox & Snell's of set 2 is 0.578 the highest value than other sets. All sets except for set 1 show the same values, which have the highest values for the Negelkerke, sensitivity, specificity and accuracy. Therefore, the model with 70% for the estimation set and 30% of the evaluation set (set 2) was selected as the best fit model for Binary Logistic Regression since this model fulfills all the best fit model criteria.

Final Model

Table 10 shows the final model using set 2 (70% estimation sets, 30% evaluation sets) after reducing the insignificant variable.

Table 10: Binary Logistic Regression Results for Final Model

Variables	Coefficient	Wald Statistics	Significant (p-value)	Exp(B) Odds Ratio
Constant	-10.655	23.151	0.000	0.000
Test 2	0.046	8.709	0.003	1.047
Final Test	0.245	25.067	0.000	1.278

All variables were significant because the *p*-value for test 2 and the final test were lower than the alpha value. From Table 10, it can be seen the value of the odds ratio for all variables. The odds ratio for test 2 is 1.047, which means that, for everyone's mark increase in test 2, the odds of result will increase by 1.047. In comparison, the odds ratio for the final test is 1.278, which means that, for everyone's mark increase in the final test, the odds of result will increase by 1.278. Table 10 also shows the result of wald statistics. The value of wald statistics for test 2 (8.709) and final test (25.067) are more than $\chi^2_{(0.05,1)}$ =3.841. This indicates that, there is a significant effect between independent variable (Test 2 and Final Test) and the Overall Marks.

Therefore, the best model using Binary Logistic Model is

$$\ln\left(\frac{P}{1-P}\right) = -10.656 + 0.046Test2 + 0.246FinalTest \quad (4)$$

Table 11: Binary Logistic regression Classification Table

		Predicted		Total	Percentage
		Failure	Success		
Observed	Failure	20	6	26	76.9
	success	4	206	210	98.1
Total		24	212	236	95.8

Note: The cut value is .500

Table 11 shows the result comparing the outcome predicted using the proposed Binary Logistic Regression Model with the actual data outcome. The data accuracy was 95.8%, which is more than the cut-off point of 50%, and then the model possesses good predictive efficiency.

CONCLUSION

This study is conducted using binary logistic regression to determine the factors affecting student success in the subject introduction to statistics. Five different sets of data partitioning were used and the results show that the data with a partitioning set of 70% for the estimation set and 30% of the evaluation set is the best fit model for Binary Logistic Regression since this model fulfils all the best fit model criteria. According to this model, it was determined that the test 2 and final test were significant variables in determining the factor that affect to the result of the subject. From the binary logistic classification table, the result shows that the model is estimated to give an accurate prediction of 95.8%. The analysis on this topic recommended that the study be conducted using other non-linear models or transformation techniques to produce a better prediction model with higher accuracy. The study could also be extended by using the linear regression analysis method to determine the significant variables.

ACKNOWLEDGEMENTS

The authors would like to thank to the Universiti Teknologi MARA Cawangan Melaka and Universiti Teknologi MARA Cawangan Perlis who participated in this study. The author also like to express gratitude to the reviewer for their insightful comments and ideas.

AUTHORS’ CONTRIBUTION

All authors provided critical feedback and helped shape the research, analysis and manuscript.

CONFLICT OF INTEREST DECLARATION

We certify that the article is the original work. The article has not received prior publication and is not under consideration for publication elsewhere. This manuscript has not been submitted for publication nor has it been published in whole or in part elsewhere. We testify to the fact that all Authors have contributed significantly to the work, validity and legitimacy of the data and its interpretation for submission to Jurnal Intelek.

REFERENCES

Adnan, N. I. M., Wahid, S. N. S., Ujang, S., Yacob, N. A., & Zaini, A. A. (2021). Open And Distance Learning Preparedness Factors Among Academicians In Uitm (Pahang) Using Logistic Regression. *AIP Conference Proceedings*, *2355*(May). https://doi.org/10.1063/5.0053194

Alija, S. (2015). Application of ordinal logistic regression in the study of students ' achievement in external testing. *Bulletin of the Transilvania University of Brasov*, *8*(57), 1–6.

Baars, G. J. A., Stijnen, T., & Splinter, T. A. W. (2017). A model to predict student failure in the first year of the undergraduate medical curriculum. *Health Professions Education*, *3*(1), 5–14. https://doi.org/10.1016/j.hpe.2017.01.001

Chatterjee, S., & Hadi, A. S. (2006). Regression analysis bye example: fourth edition. In *Regression Analysis by Example: Fourth Edition*. https://doi.org/10.1002/0470055464

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic regression. In *Wiley Series in Probability and Statistics Wiley series in probability and statistics: Vol. 3rd ed.* http://search.lib.virginia.edu/catalog/ocn830163779

Hu, S., & Hu, Y. (2021). Research on the influence of students gender on students examination scores based on logistic regression model. *Journal of Physics: Conference Series*, *1955*(1). https://doi.org/10.1088/1742-6596/1955/1/012112

Jeslet, D. S., Komarasamy, D., & Hermina, J. J. (2021). Student result prediction in Covid-19 lockdown using machine learning techniques. *Journal of Physics: Conference Series*, *1911*(1). https://doi.org/10.1088/1742-6596/1911/1/012008

Manieri, E., de Lima, M., & Ghosal, N. (2015). Testing for success: A logistic regression analysis to determine which pre-admission exam best predicts success in an associate degree in nursing program. *Teaching and Learning in Nursing*, *10*(1), 25–29. https://doi.org/10.1016/j.teln.2014.08.001

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Shedriko. (2021). Binary logistic regression in determining affecting factors student graduation in a subject. *Jurnal Teknologi dan Open Source,* *4*(1), 114–120. https://doi.org/10.36378/jtos.v4i1.1401

Suliman, N. A., Abidin, B., Manan, N. A., & Razali, A. M. (2014). Predicting students' success at pre-university studies using linear and logistic regressions. *AIP Conference Proceedings*, *1613*(Soric 2013), 306–316. https://doi.org/10.1063/1.4894355

Wambuguh, O., & Yonn-Brown, T. (2013). Regular lecture quizzes scores as predictors of final examination performance: A test of hypothesis using logistic regression analysis. *International Journal for the Scholarship of Teaching and Learning*, *7*(1), 1-10. https://doi.org/10.20429/ijsotl.2013.070107

Wang, J., Ge, L., Li, F., Liu, X., Zeng, G., & He, X. (2021). Analysis of influencing factors and teaching reform of nuclear professional English based on logistic regression. *Journal of Physics: Conference Series*, *1774*(1). https://doi.org/10.1088/1742-6596/1774/1/012023

Wanvarie, S., & Sathapatayavongs, B. (2007). Logistic regression analysis to predict Medical Licensing Examination of Thailand (MLET) Step1 success or failure. *Annals of the Academy of Medicine Singapore*, *36*(9), 770–773.