Global Journal of Engineering Education

Predicting the probability of students' final passing results using the multinomial regression method

Julianti Kasih† & Sani Susanto‡

Maranatha Christian University, Bandung, Indonesia†
Parahyangan Catholic University, Bandung, Indonesia‡

ABSTRACT: This research has an aim similar to that of three earlier studies [1-3], that is, to facilitate lecturer in helping students to predict their final results (high distinction, distinction or pass) based on their performance in several subjects in the first four semesters of their study period. The main difference is that instead of predicting the students' final pass results, in this research, *the probability* of a student getting those results is to be predicted. This is done through a data mining and multivariate technique called the multinomial regression method. Three cases are presented and discussed in this article.

Keywords: Data mining, multivariate technique, the multinomial regression method

INTRODUCTION

This article extends earlier research by identifying several common goals with three previous research programmes [1-3] to assist academic supervisors in:

- predicting students' final results after pursuing their undergraduate programme based on their first four semesters' academic achievements in several subjects;
- helping lecturers to assist their students in setting up their study plans each semester in order for them to perform to their full potential.

While the three previous research programmes offered an explicit prediction of students' final results, this research aims to provide implicit prediction by computing the probability of getting each type of result (high distinction or *cum laude*, distinction or very satisfactory and pass or satisfactory predicate).

This will be done through a data mining or multivariate statistics tool called the multinomial logistic regression. The data set used is the same as that used in [1-3] and for confidentiality reasons, the arena is called the Faculty of Information Technology, University X in Bandung, West Java, Indonesia.

OVERVIEW OF BACKGROUND THEORY

As a modelling tool, linear regression has been widely used. However, this tool is inappropriate when the required model parameters need to be non-linear. For example, this is the case when modelling the probability that a case will experience the event of interest or that a case is in a particular category of the binary response. As a probability must fall between 0 and 1, the linear regression model cannot accommodate it. In this case, the logistic regression model can serve as an alternative [4].

The multinomial or polytomous logistic regression model is a regression model, in which the dependent (response) variable has more than two categories. Like other univariate and multivariate data analysis methods, this technique has been considered as instrumental in the medical, engineering and the manufacturing industries [5]. The basic concept of the multinomial logistic regression model was generalised from binary logistic regression [4][6].

The general outline in applying the multinomial logistic regression is adapted from Benoit's note [7] and is as follows. Firstly, one has data for n observations (in this case, n = 146 alumni as observations). Secondly, Y is a categorical (polytomous) dependent (or response) variable with C categories, taking on values 0,1,..., (C-1). Thirdly, one has k explanatory or independent variables $X_1, X_2, \ldots, X_k$.

The multinomial logistic regression model is based on the following assumptions [7]:

- Observations Yi are statistically independent of each other;
- Observations Yi are a random sample from a population where Yi has a multinomial distribution with probability parameters $\pi_i^{(0)}, \pi_i^{(1)}, \ldots, \pi_i^{(C-1)}$, and
- One has to set aside one category for a base category.

The logit for each non-reference category j = 1, …, (C-1) against the *reference category* 0 depends on the values of the independent (explanatory variables) through the following equation:

$$\ln\left(\frac{\pi_j^{(j)}}{\pi_0^{(j)}}\right) = \alpha^{(j)} + \beta_1^{(j)} X_{1i} + \cdots + \beta_{ki}^{(j)} X_{ki} \quad (1)$$

for each j = 1, 2, …, (C-1), where $\alpha^{(j)}, \beta_1^{(j)}, \ldots, \beta_{ki}^{(j)}$ are unknown population parameters to be estimated.

DATA PROCESSING, RESULTS AND INTERPRETATION

The data processing stage consists of the following steps:

- Determining the dependent (response) and possible independent (explanatory) variables;
- Collecting data;
- Selecting significant independent (explanatory) variables by running the model step wisely;
- Interpreting the results.

Determining the Dependent (Response) and Possible Independent (Explanatory) Variables

The variables to be determined are divided into two types. Firstly, the grade of 16 informatics related subjects in the first four semesters are taken, those subjects are:

- In semester 1: IF 102, IF 103, IF 104, IF 105, IF 106;
- In semester 2: IF 201, IF 202, IF 203, IF 205;
- In semester 3: IF 302, IF 305;
- In semester 4: IF 401, IF 402, IF 403, IF 404, IF 405.

Those 16 subjects are numerical and represent the independent (explanatory) variables in the model to be established. The final marks of a subject were classified into five groups, as follows: A (4 = high distinction), B (3 = distinction), C (2 = credit), D (1 = pass) and E (0 = fail) with some intermediates, such as B+ (3.50) and C+ (2.50).

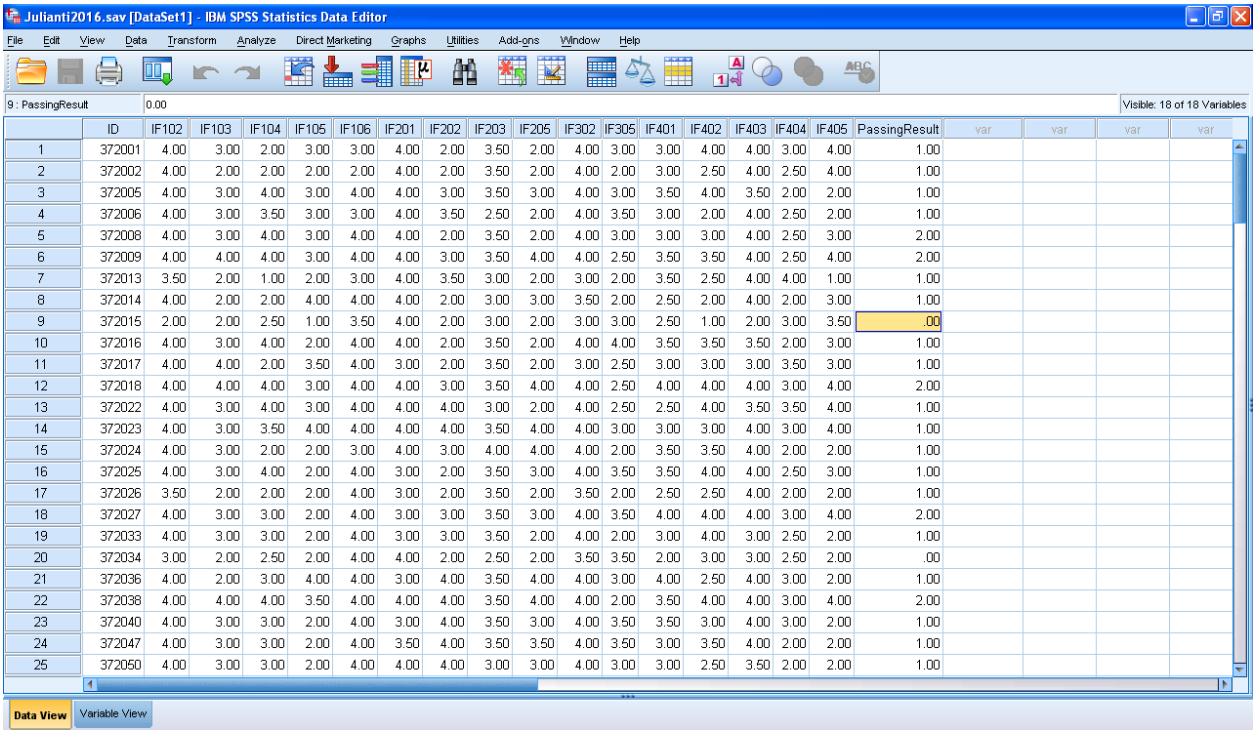
Secondly, the alumni's passing result at the end of their undergraduate programme consists of three possibilities, high distinction (0), distinction (1) and satisfactory (2). This variable is categorical and serves as the dependent (response) variable. There are 43, 99 and four alumni, with the passing result, respectively, high distinction, distinction and satisfactory.

Collecting Data

The data used in this research were collected from the academic transcripts of 146 alumni. Each transcript contained the final marks of 31 subjects from the 1st until the 8th semester. As mentioned earlier, only data from the first four semesters were chosen as alumni or student attributes. These data are stored in the independent (explanatory) variables. The other data collected are the alumni's passing results, which are stored in the dependent (response) variable. Figure 1 illustrates a section of the data set, which is saved in the form of a SPSS worksheet file.

In relation to the Benoit's outline [7] presented in the *Overview of Background Theory* section, in this research case:

- The observations are data from n = 146 alumni;
- The categorical (polytomous) variable Y is the alumni's passing result at the end of their undergraduate programme, which consists of C = 3 possibilities high distinction (0), distinction (1) and satisfactory (2);
- The explanatory or independent variables are $X_1, X_2, \ldots, X_{16}$ which represent the final mark of the k = 16 subjects IF 102, IF 103, IF 104, IF 105, IF 106, IF 201, IF 202, IF 203, IF 205, IF302, IF305, IF401, IF402, IF403, IF404, and IF405, respectively.



	ID	IF102	IF103	IF104	IF105	IF106	IF201	IF202	IF203	IF205	IF302	IF305	IF401	IF402	IF403	IF404	IF405	PassingResult
1	372001	4.00	3.00	2.00	3.00	3.00	4.00	2.00	3.50	2.00	4.00	3.00	3.00	4.00	4.00	3.00	4.00	1.00
2	372002	4.00	2.00	2.00	2.00	2.00	4.00	2.00	3.50	2.00	4.00	2.00	3.00	2.50	4.00	2.50	4.00	1.00
3	372005	4.00	3.00	4.00	3.00	4.00	4.00	3.00	3.50	3.00	4.00	3.00	3.50	4.00	3.50	2.00	2.00	1.00
4	372006	4.00	3.00	3.50	3.00	3.00	4.00	3.50	2.50	2.00	4.00	3.50	3.00	2.00	4.00	2.50	2.00	1.00
5	372008	4.00	3.00	4.00	3.00	4.00	4.00	2.00	3.50	2.00	4.00	3.00	3.00	3.00	4.00	2.50	3.00	2.00
6	372009	4.00	4.00	4.00	3.00	4.00	4.00	3.00	3.50	4.00	4.00	2.50	3.50	3.50	4.00	2.50	4.00	2.00
7	372013	3.50	2.00	1.00	2.00	3.00	4.00	3.50	3.00	2.00	3.00	2.00	3.50	2.50	4.00	4.00	1.00	1.00
8	372014	4.00	2.00	2.00	4.00	4.00	4.00	2.00	3.00	3.00	3.50	2.00	2.50	2.00	4.00	2.00	3.00	1.00
9	372015	2.00	2.00	2.50	1.00	3.50	4.00	2.00	3.00	2.00	3.00	3.00	2.50	1.00	2.00	3.00	3.50	.00
10	372016	4.00	3.00	4.00	2.00	4.00	4.00	2.00	3.50	2.00	4.00	4.00	3.50	3.50	3.50	2.00	3.00	1.00
11	372017	4.00	4.00	2.00	3.50	4.00	3.00	2.00	3.50	2.00	3.00	2.50	3.00	3.00	3.00	3.50	3.00	1.00
12	372018	4.00	4.00	4.00	3.00	4.00	4.00	3.00	3.50	4.00	4.00	2.50	4.00	4.00	4.00	3.00	4.00	2.00
13	372022	4.00	3.00	4.00	3.00	4.00	4.00	4.00	3.00	2.00	4.00	2.50	2.50	4.00	3.50	3.50	4.00	1.00
14	372023	4.00	3.00	3.50	4.00	4.00	4.00	4.00	3.50	4.00	4.00	3.00	3.00	3.00	4.00	3.00	4.00	1.00
15	372024	4.00	3.00	2.00	2.00	3.00	4.00	3.00	4.00	4.00	4.00	2.00	3.50	3.50	4.00	2.00	2.00	1.00
16	372025	4.00	3.00	4.00	2.00	4.00	3.00	2.00	3.50	3.00	4.00	3.50	3.50	4.00	4.00	2.50	3.00	1.00
17	372026	3.50	2.00	2.00	2.00	4.00	3.00	2.00	3.50	2.00	3.50	2.00	2.50	2.50	4.00	2.00	2.00	1.00
18	372027	4.00	3.00	3.00	2.00	4.00	3.00	3.00	3.50	3.00	4.00	3.50	4.00	4.00	4.00	3.00	4.00	2.00
19	372033	4.00	3.00	3.00	2.00	4.00	3.00	3.00	3.50	2.00	4.00	2.00	3.00	4.00	3.00	2.50	2.00	1.00
20	372034	3.00	2.00	2.50	2.00	4.00	4.00	2.00	2.50	2.00	3.50	3.50	2.00	3.00	3.00	2.50	2.00	.00
21	372036	4.00	2.00	3.00	4.00	4.00	3.00	4.00	3.50	4.00	4.00	3.00	4.00	2.50	4.00	3.00	2.00	1.00
22	372038	4.00	4.00	4.00	3.50	4.00	4.00	4.00	3.50	4.00	4.00	2.00	3.50	4.00	4.00	3.00	4.00	2.00
23	372040	4.00	3.00	3.00	2.00	4.00	3.00	4.00	3.50	3.00	4.00	3.50	3.50	3.00	4.00	3.00	2.00	1.00
24	372047	4.00	3.00	3.00	2.00	4.00	3.50	4.00	3.50	3.50	4.00	3.50	3.00	3.50	4.00	2.00	2.00	1.00
25	372050	4.00	3.00	3.00	2.00	4.00	4.00	4.00	3.00	3.00	4.00	3.00	3.00	2.50	3.50	2.00	2.00	1.00

Figure 1: A sample part of the data.

Selecting Significant Independent (Explanatory) Variables

One of the IBM SPSS 22.0 features, called multinomial logistic regression was applied in this data processing stage. Start with entering all of the 16 independent (explanatory) variables, then, remove through the Stepwise methods one independent (explanatory) variable at each step, based (by default) on the *p*-value. Finally, two independent (explanatory) variables, IF 102 and IF 103, remain that significantly influence the value of the dependent (response) variable; namely: the alumni's passing result at $\alpha = 0.10$. The significant influence of these two variables is indicated by the column *Sig.* in Table 1, the values of which are all less than the value of the chosen significance level $\alpha = 0.10$.

Passing result[a]		B	Std. error	Wald	df	Sig.	Exp (B)	95% confidence interval for Exp (B)	
								Lower bound	Upper bound
Pass	Intercept	20.964	5.550	14.268	1	0.000			
	IF102	-3.920	1.324	8.767	1	0.003	0.020	0.001	0.266
	IF103	-2.692	0.996	7.299	1	0.007	0.068	0.010	0.478
Distinction	Intercept	16.099	4.413	13.310	1	0.000			
	IF102	-1.884	1.040	3.280	1	0.070	0.152	0.020	1.168
	IF103	-2.320	0.453	26.194	1	0.000	0.098	0.040	0.239

a) The reference category is: high distinction

Tabel 1: Parameter estimates.

Interpreting the Results

As a result of the previous stages, the regression coefficients are presented in the *B* column of Table 1. From this table, one obtains the following two multinomial regression equations, both with the *high distinction* category as the reference or base category:

$$\ln\left(\frac{\text{P (pass)}}{\text{P (high distinction)}}\right) = 20.964 - 3.920(\text{IF102}) - 2.692(\text{IF103}) \quad (2)$$

$$\ln\left(\frac{\text{P (distinction)}}{\text{P (high distinction)}}\right) = 16.099 - 1.884(\text{IF102}) - 2.320(\text{IF103}) \quad (3)$$

To solve Equations (1) and (2), one needs an additional equation that is:

$$\text{P (pass)} + \text{P (distinction)} + \text{P (high distinction)} = 1 \quad (4)$$

The next discussion is how to interpret Equations (1) and (2) as a result of the data processing stage. This is done better by discussing the following three cases:

- Case 1 - a student getting D for IF 102 and C for IF 103;
- Case 2 - a student getting B for IF 102 and C for IF 103;
- Case 3 - a student getting A for IF 102 and A for IF 103.

In Case 1, the value for variable IF 102 is 1 and 2 for IF 103, so the following three equations are obtained:

$$\ln\left(\frac{\text{P (pass)}}{\text{P (high distinction)}}\right) = 20.964 - 3.920(1) - 2.692(2) = 11.66 \quad (5.1)$$

$$\ln\left(\frac{\text{P (distinction)}}{\text{P (high distinction)}}\right) = 16.099 - 1.884(1) - 2.320(2) = 9.575 \quad (5.2)$$

$$\text{P (pass)} + \text{P (distinction)} + \text{P (high distinction)} = 1 \quad (5.3)$$

Solving Equations (5.1) to (5.3) simultaneously will result in:

P (pass) = 0.889 P (distinction) = 0.111 and P (high distinction) = 0.000,

which means that it is highly likely that this student will obtain either a pass or satisfactory predicate at the end of his undergraduate programme.

In Case 2, the value for variable IF 102 is 3 and 2 for IF 103, so the following three equations are obtained:

$$\ln\left(\frac{\text{P (pass)}}{\text{P (high distinction)}}\right) = 20.964 - 3.920(3) - 2.692(2) = 3.82 \quad (6.1)$$

$$\ln\left(\frac{\text{P (distinction)}}{\text{P (high distinction)}}\right) = 16.099 - 1.884(3) - 2.320(2) = 5.807 \quad (6.2)$$

$$\text{P (pass)} + \text{P (distinction)} + \text{P (high distinction)} = 1 \quad (6.3)$$

Solving Equations (6.1) to (6.3) simultaneously will result in:

P (pass) = 0.120 P (distinction) = 0.887 and P (high distinction) = 0.003,

which means that it is highly likely that this student will obtain a distinction or satisfactory predicate at the end of his undergraduate programme.

In Case 3, the value for variable IF 102 is 4 and 4 for IF 103, so the following three equations are obtained:

$$\ln\left(\frac{\text{P (pass)}}{\text{P (high distinction)}}\right) = 20.964 - 3.920(4) - 2.692(4) = -5.484 \quad (7.1)$$

$$\ln\left(\frac{\text{P (distinction)}}{\text{P (high distinction)}}\right) = 16.099 - 1.884(4) - 2.320(4) = -0.717 \quad (7.2)$$

$$\text{P (pass)} + \text{P (distinction)} + \text{P (high distinction)} = 1 \quad (7.3)$$

Solving Equations (7.1) to (7.3) simultaneously will result in:

P (pass) = 0.03 P (distinction) = 0.327 and P (high distinction) = 0.670,

which means that it is highly unlikely that this student will obtain just a pass or satisfactory predicate at the end of his undergraduate programme.

CONCLUSIONS AND SUGGESTION FOR FURTHER RESEARCH

This research shows that the probability of students' final passing result can be predicted just by knowing their final mark for a few subjects they took in early semesters during their undergraduate programme. The use of the multinomial logistic regression as a data mining and multivariate technique has shown its effectiveness for the prediction.

Since out of 146 alumni only four of them received the satisfactory passing result, and the rest received high distinction or distinction, then, the dependent (response) variable Y is very closely or nearly to binary. When facing a similar case, it is suggested, instead of the multinomial logistic regression, to apply the binomial logistic regression, by combining the satisfactory passing result with distinction result into one category, say, the distinction-satisfactory category, so the model is more fine-tuned, and the precision of the prediction is increased.

REFERENCES

1. Kasih, J. and Susanto, S., Predicting students' final results through discriminant analysis. *World Trans. on Engng. and Technol. Educ.*, 10, **2**, 144-147 (2012).
2. Kasih, J., Ayub, M. and Susanto, S., Predicting students' final passing results using the Classification and Regression Trees (CART) algorithm. *World Trans. on Engng. and Technol. Educ.*, 11, **1**, 46-49 (2013).

3. Kasih, J., Ayub, M. and Susanto, S., Predicting students' final passing results using the Apriori Algorithm. *World Trans. on Engng. and Technol. Educ.*, 11, **4**, 517-520 (2013).
4. DeMaris, A. and Selman, S.H., *Converting Data into Evidence: A Statistics Primer for the Medical Practitioner*. Springer Science + Business Media, New York (2013).
5. Bayaga, A., Multinomial logistic regression: usage and application in risk analysis. *J. of Applied Quantitative Methods*, 5, **2**, 288-297 (2000).
6. Hyun, W-Y., Using multinomial logistic regression analysis to understand anglers willingness to substitute other fishing locations, *Proc. 2006 Northeastern Recreation Research Symp.*, 248-255 (2006).
7. Benoit, K., Multinomial and Ordinal Logistic Regression, ME104: Linear Regression Analysis Lecture Note (2012), 17 April 2016, http://www.kenbenoit.net/courses/ME104/ME104_Day8_CatOrd.pdf

BIOGRAPHIES

Julianti Kasih is a Lecturer in the Faculty of Information Technology at Maranatha Christian University, Bandung, West Java, Indonesia. Currently, she is the Head of the Information System Expertise Group and leads a marketing team for this Faculty. She graduated with a Bachelor degree in economics from Diponegoro University, Semarang, Indonesia, in 1985. In 1997, she completed her Graduate Certificate in International Business from the Faculty of Business and Economics at Monash University, Melbourne, Australia. In 2008, she earned a Master degree in information systems for businesses from the Informatics Management and Computer Technological School LIKMI (STMIK LIKMI) in Bandung, Indonesia. Her research interests are in the area of behavioural information systems, marketing, consumer behaviour, engineering education and data mining-analytics-big data. Among other things, she has published three articles in the *World Transactions on Engineering and Technology Education* (WTE&TE), an international journal published by the World Institute for Engineering and Technology Education. Her article entitled *Predicting students' final results through discriminant analysis* (WTE&TE, Vol.10, No.2, 144-147, 2012) was awarded an incentive for international journal publication by the Directorate of Higher Education in the Republic of Indonesia, in 2012. This article has been cited in several international journal papers.

Sani Susanto is a Senior Lecturer in the Department of Industrial Engineering, Faculty of Industrial Technology at Parahyangan Catholic University, Bandung, West Java, Indonesia. Currently, he is the Secretary of the University Senate. He graduated with a Bachelor degree in mathematics from Bandung Institute of Technology, in 1987, and a Bachelor degree in Agro Socio Economics from the Faculty of Agriculture, Padjadjaran University in 1991, both campuses are in Bandung, Indonesia. In 1992, he completed his Master degree in industrial engineering at Bandung Institute of Technology. In 1994, he was awarded the John Crawford Scholarship (a Master degree leading to PhD), organised by the then Australia Development Cooperation Scholarship (ADCOS) to pursue a PhD degree in Industrial Engineering and Engineering Management in the Department of Mechanical Engineering at Monash University, Melbourne, Australia that was completed in mid-1998. After finishing his PhD, prior his return to Indonesia, he stayed in this Department under the Postgraduate Publication Award from this university. His research interests are in the area of operations research, multivariate analysis, systems modelling, engineering education and data mining-analytics-big data. He has had a long association with the WIETE (formerly UICEE) and has published five articles in the WTE&TE journal. In 2002, he was awarded the Silver Badge of Honour by the UICEE (now WIETE) on the occasion of the *6th Baltic Region Seminar on Engineering Education*. These awards were given to particular engineering educators to recognise their distinguished contributions to engineering education, outstanding achievements in the globalisation of engineering education through the activities of the Centre, and, in particular, for remarkable service to the UICEE. His article, written jointly with the first author, entitled *Predicting students' final results through discriminant analysis* (WTE&TE, Vol.10, No.2, 144-147, 2012) was awarded an incentive for the international journal publication by the Directorate of Higher Education in the Republic of Indonesia in 2012.