Predicting the Probability of Student's Academic Abilities and Progress with EMIR and Data from Current and Graduated Students

○Kunihiko TAKAMATSU*
Faculty of Education,
Organization for the Advancement of Higher Education,
Center for the Promotion of Excellence in Research and Development of Higher Education,
Life Science Center
Kobe Tokiwa University,
Kobe, Japan
ktakamatu@gmail.com
*Corresponding Author

Katsuhiko MURAKAMI
Human Genome Center,
The Institute of Medical Science,
The University of Tokyo,
Tokyo, Japan
murakami.ktk@gmail.com

Tsugumi Oshiro
Faculty of Education,
Kobe Tokiwa University,
Kobe, Japan
t-ohshiro@kobe-tokiwa.ac.jp

Aoi SUGIURA
Kobe City Nishi-Kobe Medical Center,
Kobe, Japan
aoi.kishida@gmail.com

Kenya BANNAKA
Department of Oral Health
Center for the Promotion of Excellence in Research and Development of Higher Education
Kobe Tokiwa College,
Kobe, Japan
k-bannaka@kobe-tokiwa.ac.jp

Yasuo NAKATA
Faculty of Health Sciences,
Organization for the Advancement of Higher Education,
Center for the Promotion of Excellence in Research and Development of Higher Education
Kobe Tokiwa University,
Kobe, Japan
nakata0325@gmail.com

***Abstract*—In 2016, Kobe Tokiwa University constructed an office for institutional research (IR) promotion. The purpose of this office is to propose, manage, arrange, and collect information on students at the university not only as a general management strategy, but also to support enrollment management.**

Our database currently contains 3,495 points of data (i.e., headcounts), each containing 1,246 items of numerical value. Last year, we reported on an analysis that focused on the "student dropout" phenomenon by using these data from both current graduate and dropout students. This year, we formulated a research question that is centered on predicting the probability of students' progress and academic abilities through Enrollment Management / Institutional Research (EMIR).

We obtained results with these data by processing them through a machine learning technique using random forest, which yielded a correction rate of about 90%.

Keywords—academic ability, EMIR, machine learning, Random Forest

I. INTRODUCTION

A report prepared by the Central Council for Education in 2008, titled "Toward building bachelor degree education"

clarified that teaching management and professional faculty and staff development were subject to three policies (i.e., those involved in awarding degrees, curriculum implementation and organization, and student enrollment) [1]. After this report, scientific thinking styles involving statistical data and "evidence-based" research gradually came to be required at the university level.

We followed this trend at Kobe Tokiwa University by establishing an institutional research (IR) committee in preparation for the 2015 term before establishing an IR promotion office in 2016. At our university, IR consists of proposing, managing, arranging, and collecting information on students. This is not merely a general management strategy, but it is also one that is aimed at enrollment management. Operations at the IR promotion office were initiated in 2016, at which point we began collecting and arranging data.

An IR promotion unit was established in 2017. In contrast to the IR promotion practices in which the department is managed solely by an administrative staff, the IR promotion unit at Kobe Tokiwa University is a collaborative group consisting of both staff and faculty.

In this unit, we conducted an analysis that focused on the "student dropout" as part of our main inquiry in 2016, since "dropout of students" is a considerable problem in higher education [2] in Japan. We reported the results in a few articles [3][4][5]. The term "dropout" is not only related to negative connotations including "economic difficulty" and a decrease in the desire to learn but it also positive or spontaneous events such as "studying abroad" or gaining "admission to another university [2]. "

Since students drop out of higher education for many reasons, it has been difficult to predict the phenomenon itself. To address this problem, we conducted an analysis that focused on the "dropout of students" through evidence-based research using Enrollment Management / Institutional Research (EMIR) data with machine learning methods. We successfully investigated the probability of student dropouts through EMIR among graduate students in 2017 [3][4][5]. Similar research that investigates the probability of student dropout through EMIR with machine learning methods has already been reported [6][7][8][9].

In recent years, according to research conducted by the Tokyo University of Science, grades at the time of graduation are correlated with grades at the end of the first year. According to the Mainichi Shimbun article, "The results at the time of graduation from university almost coincided with the results at the end of one year, and there was no correlation with the results of the entrance examination." Makoto Yamamoto, the university's vice president said, "It became clear in a survey conducted for large students, and in particular, the attendance of the first week of the year affects the subsequent student life" [10].

A similar trend was observed in the analysis of our IR promotion office (data not shown). Therefore, we centered our research question on "examining the possibility of academic progress prediction using EMIR with machine learning" as evidence-based research.

In this study, we used data on 1,155 students who were involved in 118 courses. We tried to predict whether or not a student passed the national examination using 1-year course grade data with machine learning methods.

II. METHODS

A. Data Preparation

EMIR Data are undisclosed in our university. Therefore, we obtained anonymized data from the IR promotion office. These data were anonymized and excluded all personal information such as name, student ID number, etc.

We obtained Excel style data from the IR promotion office. The case number, as the number of students, was 1155, and there were 118 explanatory variables such as "achievement of freshman." The numeric range for each course subject was 0 to 100. We set the score as 0 for students who did not register for the particular subject, so these were not defective scores. The students who obtained a score higher than 0 had mostly scored between 60 and 100. It was extremely rare for a student to score less than 60.

B. The Machine Learning Technique

We used Python (3.6.0) on Mac OS X 10.11.6 to conduct our analysis and relied on the Python library. We used numPy [11], matplotlib [12], scikit-learn [13], and pandas [14]. We used 70% of the data from all cases as training data (assembly of a concretely predicting model), while using 30% as test data (these data were unknown and had never been used in the assessment of the predicting model). We then conducted the assignment phase using scikit-learn to reduce bias.

TABLE 1. CORRECT ANSWER RATE FOR THE TRAINING AND TEST SETS OF MACHINE LEARNING METHODS (RANDOM FOREST).

Data set for evaluation	Correct answer rate
Training accuracy	0.966
Test accuracy	0.913

TABLE 2. TWO KINDS OF CORRECT ANSWER RATE AND THE HARMONIC MEAN.

Precision	Recall rate	F-Value (Harmonic Mean)
0.967	0.817	0.885

III. RESULTS AND DISCUSSION

The number of students as data, which were obtained from the IR promotion office, was 1,155. Our university has a set medical technology department and a nursing department as part of the health science faculty, a child education department as part of the education faculty, and an oral health care department as part of the junior college department.

The data that we used to conduct our analysis were gathered from the above-listed four departments. All data had 118 items (subjects) from students in the freshman grade. Upon summing all departments, we removed the duplicate items.

Our research question was to "investigate the possibility of the learning progress prediction on EMIR data using machine learning." We defined the question as follows: "Whether the passing of the national examination can be predicted by using the results data by subject (1,155 people and 118 subjects) at the freshman level." Thus, we set the category as pass or fail under the prediction questions by using 118 explanatory variables as X and explained or objective variables Y (0 or 1) as the prediction target.

We predicted whether one would pass or fail in the national examination using the random forest method, which is known to be a highly accurate prediction method that works stably with various types of data [15]. In short, this method helps by creating many decision trees that are mutually correlated weakly by repeating "creating a decision tree from a part of learning data" and finally predicting a majority by relying on the majority decision.

We present the machine learning results in Table 1. In the machine learning method, the accuracy rate of learning is 0.966, and the accuracy rate of the test is 0.913. Both accuracy rates for learning and testing are high, and the possibility of overlearning seems to be small.

Next, for test data, we compared the accuracy rate (accuracy, precision, P) with the number of predicted passes as the denominator, and the accuracy rate (recall, Recall, R) with the number of actual passes as the denominator (Table 2). We calculated the F value, which is the harmonic mean (Table 2). The definition of the F value is $1 / F = (1 / P + 1 / R) / 2$. In this model, it is expected to be 96.7% correct when it is expected to pass. The recall rate is a very high accuracy rate, which can be inferred from the fact that 81.7% of the passes was correctly predicted.

The research question presented in the beginning was, "Can we predict whether students will pass the national examination using 1-year course data (1,155 people and 118 subjects)?" The answer is that it is likely to be predictable. Since the data used for the prediction model are the results of each subject at the freshman level, it can be expected that teaching based on accurate judgment criteria will be possible from as early as the first grade.

REFERENCES

[1] MEXT, "Construction of education for university," *The Central Council for Education*, 2008. [Online]. Available: http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/houkoku/080410.htm. [Accessed: 15-Mar-2018].

[2] K. Anegawa, "The Impact of Learning and Living Environments of Colleges on Dropout Rates," *Kyushu Daigaku Daigakuin Keizaigakukai*, vol. 149, pp. 1–16, 2014.

[3] K. Takamatsu, K. Murakami, K. Takao, J. Asahi, T. Kirimura, K. Bannaka, I. Noda, K. Mitsunari, T. Nakamura, and Y. Nakata, "Probability prediction of dropout of students on EMIR," *Proceeding 7th Meet. Japanese Institutional Res. MJIR2017*, pp. 60–65, 2017.

[4] K. Takamatsu, K. Murakami, K. Takao, J. Asahi, T. Kirimura, K. Bannaka, I. Noda, K. Mitsunari, T. Nakamura, and Y. Nakata, "Probability prediction of dropout of students on EMIR," in *Meeting on System and Information,The Society of Instrument and Control Engineers, SSI2017*, 2017, pp. 60–65.

[5] K. Murakami, K. Takamatsu, Y. Kozaki, A. Kishida, K. Bannaka, I. Noda, J. Asahi, K. Takao, K. Mitsunari, T. Nakamura, and Y. Nakata, "Predicting the Probability of Student Dropout through EMIR Using Data from Current and Graduate Students," in *Advanced Applied Informatics (IIAI-AAI), 2018 7th International Institute of Applied Informatics (IIAI) International Congress on. Institute of Electrical and Electronics Engineers (IEEE)*, 2018, pp. 478–481.

[6] N. Kondo, M. Okubo, and T. Hatanaka, "Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2017.

[7] N. Kondo and T. Matsuda, "Prediction Model of EMIR," in *Proceeding of 7th Meeting on Japanese Institutional Research, MJIR2017*, 2017, pp. 42–47.

[8] N. Kondo and T. Hatanaka, "Utilization Log data of LMS in EMIR," in *Meeting on System and Information,The Society of Instrument and Control Engineers, SSI2017*, 2017, p. 752.

[9] G. Hori, "Identifying Factors Contributing to University Dropout with Sparse Logistic Regression," in *2018 7th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2018, pp. 430–433.

[10] "Grades at graduation are correlated with grades at one year," *Mainichi Shinbun*, 2016.

[11] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011.

[12] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 99–104, 2007.

[13] F. Pedregosa and G. Varoquaux, *Scikit-learn: Machine learning in Python*, vol. 12. 2011.

[14] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, vol. 1697900, no. Scipy, pp. 51–56, 2010.

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 2, pp. 5–32, 2001.