

AI-Based Modeling for Predicting Open University Student Retention*

Sang Im Jung (Korea National Open University)
Eun Kyung Lee** (Korea National Open University)
Byeong Rae Lee (Korea National Open University)
Sanghong Kim (Korea National Open University)
Chan Hee Park (Korea National Open University)

<Abstract>

Open university students have more possibility of dropout than ordinary college students. K University has so far conducted an analysis of who continues or stops studying after the fact. However, artificial intelligence(AI)-based big data analysis has allowed K University to predict students whether to continue or stop studying during the semester.

Using the K University data warehouse, this research collected four-year data for the entire process from admission to graduation of about 100,000 students and made artificial intelligence learn them. The K University retention model which was developed using multiple algorithms found more than 220 influencing factors. The K University retention prediction system using the ensemble technique was established, and the prediction of students' retention attending the spring semester of 2023 was realized. The empirical results of the K University retention prediction system were statistically significant. K University began predicting dropout by calculating the possibility of individual students' retention every month during the semester.

* This article is a revised version of the content presented by researchers at the 36th annual conference of AAOU (Asian Association of Open University).

** Corresponding author: Eun Kyung Lee (hannah1222@gmail.com)

According to the results of the dropout prediction, students are divided into three groups (general, risk, and high risk). The prediction results of students' dropouts are provided to faculty members every month along with principal learning-related indicators.

Through the artificial intelligence-based retention-dropout prediction modeling, K University has become possible to provide data-based student support services and institutional administration. There is an active discussion on what kind of intervention and learning support will be possible for students identified as having a high risk or risk of dropout.

- **Key words:** artificial intelligence, big data, retention, dropout, prediction
-

I. Introduction

Unlike the face-to-face environment, distance education is characterized by education through educational media in a situation where instructors and learners are separated, so efforts have been made to explain academic continuity and discontinuation in consideration of the specificity of distance education. Distance education using these media has been spotlighted as an alternative higher education by enabling large-scale education, but a considerable number of students quit without continuing their studies. At the institutional level, students' dropout negatively affects administration and finance, and at the student level, the self-development process through study is interrupted. Preventing students from dropping out will enable open universities to realize economies of scale and gain stability in academic, administrative, and financial terms. Students will also be able to realize their academic objectives and demands to study in open universities, ultimately continuing circular lifelong learning that combines work and study.

Open universities have conducted continuous research and efforts to prevent their students from dropping out. Continuing research related to students' academic retention and dropout in open universities and trying to identify the cause is linked to the purpose of preventing students from dropping out. These open universities' interest in

dropout is ultimately aimed at reducing students' dropout and continuing their studies. If so, if dropout can be predicted using big data and artificial intelligence, it is significant in that there is room for students to intervene during the semester so that they do not drop out.

Artificial intelligence has opened a new way in higher education institutions for supporting students' learning based on numerous academic and learning data(Ouyang, Zheng, & Jiao, 2022). Recent technological development is very fast, and new technologies such as artificial intelligence have allowed the university to advance various functions. Universities have previously analyzed various data on events and results that occurred in schools to derive the causes and alternatives. However, artificial intelligence technology has enabled universities to not only interpret and apply data analysis, but also predict and respond. However, in order for such artificial intelligence analysis to be possible, data must be systematically collected, and in order to apply it at the institutional level beyond the level of pilot modeling, it must be equipped with a system that is easy to collect and utilize data at the institutional level.

K University went through the process of upgrading existing statistical information into a big data analysis environment, which enabled administrative work to be efficient based on data utilization or data analysis results. K University worked on building a data hub portal and then applied artificial intelligence technology to predict students' behavior by analyzing existing big data. artificial intelligence learning analysis examined learners' academic behavior, identified factors that affect students' academic continuity, and predicted and provided subsequent academic activities and results through current students' academic behavior. K University's student retention-dropout prediction modeling is meaningful as a practical example in that it has a system for collecting and utilizing such data at the university level. Also retention-dropout predictive modeling is significant in terms of research and education in that it predicts students' learning and supports their learning.

II. Literature Review

This study is about AI-based retention-dropout prediction modeling at open

universities. The theoretical background and prior studies related to these studies were reviewed by dividing them into research on retention-dropout, which is the subject of the study, and artificial intelligence-based research, which is the method of research.

1. Students' Retention and Dropout

Open university has contributed to the expansion of higher education and the improvement of excellence of adult learners. However, contrary to expectations, the rate of learners failing to continue their studies and stopping or dropping out of open universities is higher than that of traditional universities(Breslow et al., 2013; Tyler-Smith, 2006). The dropout rate of open universities is 3-4 times that of traditional universities(Lee, Jeong, & Kim, 2020). Students at open universities have characteristics that vary in age, occupation, and academic background compared to students at traditional universities. Unlike traditional students who concentrate on their studies in college, adult learners often play various social roles such as jobs and housework while performing their studies. In addition, difficulties such as a sense of isolation and lack of interaction in a distance education environment may affect continuing study (Kwon, 2010). In order to prevent academic interruption and support academic continuity of open university students, it is necessary to consider new alternatives through consideration of the learning experience of distance university learners(Marr, 2018).

Models of Tinto(1975), Kember(1989), Eaton and Bean(1995) and various empirical studies based on it have continued to identify variables of academic interruption of online university learners and explore the relationship between each variable to explain the phenomenon of academic interruption comprehensively and structurally. Researchers tried to explore various variables such as individual variables such as learners' environment, psychological variables such as motivation, and educational institutional variables that affect the curriculum and learning to explore the causes of adult learners' continuity or suspension of study and suggest alternatives(Jung & Lee, 2017).

With digital education platforms on the rise, open universities in many countries have been facing significant challenges related to student dropouts. Studies by Shou et al.(2024), Elibol and Bozkurt(2023), and Appavoo, Gungea, and Sohoraye(2023) indicated that dropout rates are higher in open and distance learning compared to traditional educational formats. These studies emphasized the complexity of accurately predicting

and addressing dropout due to the nuanced and multi-faceted nature of student engagement and retention in online environments.

The lack of a consistent definition of “dropout factors” complicates the efforts to address this issue across different educational settings, as highlighted by Elibol and Bozkurt(2023). This inconsistency hampers the ability to develop universally effective interventions, suggesting the need for a more standardized approach to how dropout is understood within the academic community.

Studies on dropouts of domestic cyber college students generally deal with learner variables as important(Im, 2007; Jeon, 2010; Kwon, 2010). Im(2007), who analyzed the causes of dropout by dividing the causes of dropout into maladjustment of learners and discontinuation of purchase of educational services in consideration of the specificity of cyber colleges, reveals that the causes of dropout were in the order of difficulty in securing learning time, financial difficulties, maladjustment to online education environments, and anxiety about cyber colleges’ awareness. Jeon(2010) also identifies variables that predict potential dropouts who are likely to drop out during the degree process based on variables related to dropout of cyber college learners, and suggests implications for them to successfully complete the course. It is revealed that variables directly related to the learning of the subject (total learning time, curriculum/content, and number of lecture accesses) have a relatively important influence on predicting the classification of dropouts of cyber college learners. Kwon(2010) also investigated the effects of personal variables, educational institution variables, and social variables on the decision to drop out intention of cyber universities, and as a result of examining the hierarchical effects of the three variables, only individual variables had a statistically significant effect. Among the individual variables, the variable that had the greatest influence on the intention to drop out of cyber college students was the learner’s interest variable.

Studies have also been conducted on students’ retention and dropout in the Korea National Open University, a remote university that conducts blended learning that combines online education and face-to-face education. Recently, studies have been conducted to reveal the relationship or structural impact between various variables in relation to the discontinuation of studies of adult learners at distant colleges(Kwon et al., 2020; Lee, Jeong, & Kim, 2020). Lee and Kim(2022) analyzed 88,000 data on K University students’ academic continuity and analyzed factors that affect K University

students' academic continuity among their personal characteristics (gender, age, entrance grade, entrance semester) and academic performance (number of subjects participating in mid-term evaluation, final evaluation, rating, acquisition, etc.). Among the variables of students' personal characteristics, age and admission grade influenced students' academic continuity. Among the academic performance variables, almost all variables, such as the number of subjects participating in the interim evaluation, the number of subjects participating in the final evaluation, ratings, and acquisition credits, had a relatively large impact on academic continuity. It is interpreted that the variables of academic performance have a great influence on academic continuity because the sincerity of academic performance is linked to academic continuity.

Recent data-based studies have attempted to identify variables related to academic continuation-interruption and reveal a comprehensive relationship, as well as to predict academic continuation or suspension of online university learners (Joung, 2020). Even if the meaning and background of learners' learning behavior cannot be fully explained, the product of a certain learning process was used to predict the behavior of learners to continue or stop their studies in the future.

2. AI-Based Research in Online Higher Education

Recently, research and practice on AI-based online higher education have been widely conducted (Ouyang, Zheng, & Jiao, 2022). Studies on AI-based online higher education have been applied to "prediction of learning status, performance or satisfaction, resource recommendation, automatic assessment, and improvement of learning experience" (Ouyang, Zheng, & Jiao, 2022, p. 7893). About two-thirds of cases of applying artificial intelligence to online higher education focus on predicting students' academic performance, and predicting students' dropout risks, academy performances, and sites effects in online courses (Ouyang, Zheng, & Jiao, 2022). Mubarak, Cao, and Zhang (2020) built a model to predict students at risk of dropout based on interaction logs in an online learning environment, with an accuracy of 84%. Aguiar et al. (2014) predicted the persistence of online classes by analyzing the portfolio of engineering students and showed results that predict them more accurately than models based on traditional academic data such as SAT scores, GPA, and demographics.

In an online higher education environment with a relatively high risk of dropout,

predicting the continuation or suspension of such learning can benefit both individual students and higher education institutions. Baneres, Rodriguez-Gonzalez, and Serra(2019) provided guidance and feedback by early identification of students with the possibility of academic interruption. In this way, the prediction system allows instructors and school administrators to identify problems with students' studies, support students' academic processes, and help students continue without stopping their studies(Moreno-Marcos et al., 2018). AI-based predictive modeling, as explored by Shou et al.(2024), provided a potent tool for identifying at-risk students before they leave their courses. By leveraging multidimensional time-series data that include demographic information, learning behaviors, and assessment outcomes, institutions can proactively implement support systems tailored to the needs of individual students. Elibol and Bozkurt(2023) discussed the application of data mining and machine learning techniques in analyzing dropout trends and patterns. These technologies are pivotal in unearthing underlying factors contributing to dropout, which can be obscured in traditional analyses. However, the success of these methods depends heavily on their integration with human-centered educational strategies that address the psychological and motivational needs of students.

Appavoo, Gungea, and Sohoraye(2023) argued for the necessity of induction sessions and enhanced tutorial support to improve student retention. They emphasize that personalized support and tailored intervention strategies are crucial in mitigating the risk of dropout, particularly for students who might feel isolated within the digital learning environment. The combination of AI-driven analytics with human-centered teaching practices offers a comprehensive approach to improving student retention. This integration helps in creating adaptive learning environments that not only predict potential dropouts but also engage students in a manner that is responsive to their individual learning preferences and challenges. The insights provided by these studies underscore the importance of continued research into both technological and pedagogical interventions in open university settings. Future research should focus on refining artificial intelligence models to better predict student behavior and developing frameworks that enhance the human elements of online learning, thereby fostering an educational atmosphere conducive to higher retention rates.

According to Ouyang, Zheng, and Jiao(2022), the most frequently used algorithms in research on AI-based online higher education are decision tree(DT), neural network(NNN), naive bays(NB), and support vector machine(SVM), and several

algorithms are used in one study. When these artificial intelligence-enabled algorithms are applied to multiple variables, the accuracy of the prediction model increases.

AI-based research in online higher education has been increasing rapidly, but the cases applied to open universities are still insignificant. In order to utilize the development of artificial intelligence technology to reduce K University students' academic dropout and increase academic retention, an artificial intelligence-based K University academic retention prediction system was developed.

III. Method

1. Data-Hub System

Ultimately, this study aimed to derive an analysis model that predicts retention-dropout of K University students by analyzing big data based on artificial intelligence and to provide it to professors and school administrators. In order to develop such a students' retention analysis model, it was necessary to systematically organize the data and specify the goal of what results they wanted to obtain. The tasks performed before the predictive analysis model can be summarized as follows.

First, the current data status of K University was investigated to provide additional data and consolidate overlapping data. Through this, data indicators were extracted through the expert meeting. Second, we investigated the data status of each administrative department and analyzed their needs to weight the dropout prediction indicators. Third, we conducted FGI with instructors to identify factors of students' retention.

In order to select indicators for K University's data status and academic persistence, we investigated all the current data status of K University. Currently, the data status provided by the statistical information system is divided into the following areas: Academic Affairs / Educational Environment / Administration/Major Disclosures / Student Life Cycle Analysis. It supports time series analysis by integrating distributed data within the K University and maintains a system for managing key data of the K University. We also want to provide an easily accessible information environment so

that users can utilize the statistics they want in the web environment. We organized the statistical information and confirmed the factors for academic persistence through expert meetings to form the basic variables of the artificial intelligence-based dropout prediction system.

This research investigated the data status of each department and conducted a needs analysis. In fact, the person in charge of each department conducted interviews with each department to investigate all data on campus to secure meaningful data and develop indicators. This research received various opinions and ideas on the areas and forms that need to be analyzed, and sought to explore ways to utilize the analysis results for 'teaching and learning support services' and 'induction of academic continuation'.

The faculty FGI was conducted on teaching and learning aspects, department/university policy-decision aspects, U-Campus data analysis, and survey questions. To select the instructors, we selected professors who are in charge of major policy decisions at the school, professors who have experience as the head of the headquarters of educational informatization, professors who have conducted research on mid- to long-term development strategies, and professors who have served as the director of the Distance Education of Institute. The above professors were provided with explanatory materials on the need and purpose of developing an artificial intelligence-based dropout prediction system, as well as the contents of the data hub portal and learning analytics, and their opinions were collected through in-depth interviews.

2. Retention-Dropout Prediction Model

After this data restructuring, specific steps were taken for retention-dropout prediction. Data on students' characteristics (gender, age, region, etc.), affiliation (department, grade, college, etc.), course (taking courses, enrollment, etc.), academic affairs (registration, grade, etc.), learning (online learning variables), and counseling support (counseling, tutoring, etc.) were collected. By synthesizing the data collected from various sources, unnecessary variables were removed and derivatives such as academic participation were generated. After applying the variable selection method, a total of more than 220 variables were finally put in, and these variables are divided into three areas: academic record,

registration for courses, and online learning record.

Academic record variables included acquisition credits, the number of F-credit subjects, final evaluation scores, interim evaluation scores, formative evaluation scores, and ratings. Variables of the registration for courses were, for examples, the year of admission, the number of semesters elapsed, the number of leave of absence, the number of unregistered courses, and the number of courses registered. The online learning record variable included learning progress rate, average weekly learning progress, online learning hours, and homepage login count.

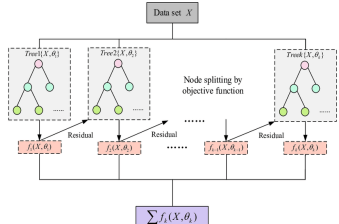
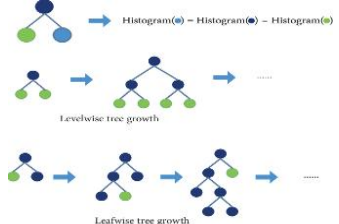
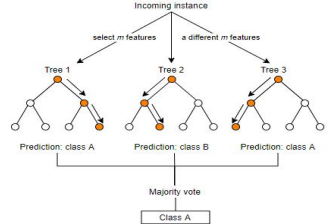
A total of 540 million cases were collected, and the data were split up according to the year so that artificial intelligence could train, validate, and test. Among the data from 2018 to 2022, 2020 data, which has different trends in students' learning from other years due to the impact of the COVID-19 pandemic, was excluded from the analysis. artificial intelligence learned data from 2021 to 2022 and verified it with data in 2019, and tested it with data in 2018. The prediction cycle for retention-dropout of K University students was set on a monthly basis and was designed to predict the likelihood of retention-dropout of students at the beginning of each month based on data up to the previous month.

〈Table 1〉 Data for students retention-dropout prediction model

Data		Case n	Note
Total Data Collecting		544,000,000	35 tables
Data	Training set	570,000	2021-2022 Y
	Validation set	310,000	2019 Y
	Test set	210,000	2018 Y
Predicting Cycle			Monthly(beginning)

The importance of variables was analyzed through basic statistical analysis and data preprocessing(abnormalization of outliers and missing values). A variable selection method that selects a variable through an algorithm was applied. Three algorithms (XGBoost, LightGBM, and CatBoost) were used, and the algorithm ensemble technique was applied using the average value of three algorithms.

<Table 2> Applied algorithms for students' retention-dropout prediction model

XGBoost	LightGBM	CatBoost
<ul style="list-style-type: none"> - Algorithm with parameters added to prevent overfitting - Learning to minimize residuals to reduce the difference between actual and predicted values 	<ul style="list-style-type: none"> - Perform residual minimization learning with vertical asymmetric tree model - Leaf-centric segmentation minimizes predictive error loss values 	<ul style="list-style-type: none"> - Direct use of categorical variables without proactive data preprocessing - Provides categorical loss functions to work effectively for classification problems
 <p>The diagram shows a data set Y being processed by three sequential trees. Each tree's output is added to the previous one to minimize residuals. The final output is the sum of all trees' outputs, $\sum f_i(x, \theta_i)$.</p>	 <p>The diagram illustrates histogram-based tree growth. It shows a histogram being converted into a tree structure through leafwise tree growth.</p>	 <p>The diagram shows an incoming instance being split into three trees based on different features. Each tree predicts a class (A or B). The final prediction is based on a majority vote.</p>

IV. Finding

1. Importance of Variables

The more than 220 variables used in the analysis are classified into three areas (academic record, registration for courses, and online learning record). <Table 3> shows the variables with high importance in predicting the retention-dropout of K University students in each area.

Among the input variables, the most important variable in predicting students' academic retention was 'the number of leave of absence' (importance .613). The second most important variable was 'total access count in week 6' (importance .592). In addition, the importance of 'number of mid-valuation', 'number of mid & final evaluation', and 'total number of registrations' were also high.

<Table 3> presents the top 10 variables of high importance according to the three areas of maintenance-dropping prediction modeling of K University students: academic records, course registration, and online learning records. In the prediction modeling of retention-dropout of K University students, 'number of leave of absence' in school academic records, 'number of md-evaluation' in courses registration, and 'total number

of access in week 6' in online learning records were the highest. When variables in all areas were input at the same time for prediction modelling, the importance of 'grade', 'total number of registrations', 'number of mid & final evaluation', 'number of web classes', and 'mid-evaluation score' was high. The importance of variables changed according to the prediction timing and target.

(Table 3) Variable importance by areas in predicting students retention-dropout

Academic record		Courses registration		Online learning	
Variables	Importance	Variables	Importance	Variables	Importance
Number of leave of absence	.613	Number of mid-evaluation	.335	Total access count in week 6	.592
Total number of registrations	.183	Number of mid & final evaluation	.294	Night access count in week 4	.090
Cumulative number of registrations	.051	Grade	.075	Total access count in week 5	.082
Registered student classification code	.043	Total score	.069	Weekend access count in week 5	.047
Final year of enrollment	.034	Number of evaluation	.059	Weekly access count in week 4	.047
Classification code of school origin	.012	Web class score	.046	Total access count in week 3	.046
Student classification code	.010	Number of F credits	.046	Night access count in week 7	.039
Cumulative number of admissions	.008	Number of P credits	.015	Weekday access count in week 5	.019
School year	.006	Number of final evaluation	.009	Total access count in week 2	.015
Last month of student access	.005	Number of web classes	.006	Total access count in week 1	.001

2. Retention Prediction Modeling

Using data built in consideration of K University students' admission to graduation, a

model was developed to predict students' retention-dropout by analyzing students' academic patterns with artificial intelligence. The number of data, accuracy, and recall rate of the prediction model for each algorithm and classification target are presented in <Table 4>. The average accuracy of the prediction model was very good at 97.3% as of June 2023. Among the models, the prediction model for enrolled students using the LGBM algorithm showed the highest accuracy(99.67). Among the various modeling, the modeling for new students using the XGB algorithm showed the lowest accuracy(92.09). Although there are differences between models, the accuracy and recall rate were very good in all models.

The characteristics of dropout between enrolled and new students were different. Students tended to give up their studies when learning activities were low, such as a low progress rate of formative evaluation without submitting assignments. New students were relatively often dropped out even when their learning activities were excellent, such as when submitting assignments and having a high progress rate of formative evaluation.

<Table 4> Results of students' retention-dropout prediction modeling

Training Date	Model-Period-Object	Period of Data	Number of Data	Accuracy	Recall: Enrolled	Recall: Expelled
20230612	CATB-Af/Mid-New	2018~2023	307,532	98.36	96.9009	99.8168
20230612	XGB-Af/Mid-New	2018~2023	307,532	92.09	99.8994	84.2801
20230612	LGBM-Af/Mid-New	2018~2023	307,532	94.51	99.7916	89.2296
20230612	CATB-Af/Mid-Enrolled	2018~2023	868,718	99.06	99.6254	99.5653
20230612	XGB-Af/Mid-Enrolled	2018~2023	868,718	99.68	99.7334	99.6324
20230612	LGBM-Af/Mid-Enrolled	2018~2023	868,718	99.67	99.7108	99.6236
	Average		3,528,750	97.32	99.2769	95.3579

The developed retention-dropout prediction model for K University students was actually applied to students attending in the spring of 2023, confirming how consistent it is with the actual retention-dropout results and prediction results of students. Based on last month's data, we operated the prediction model at the beginning of each month, and accordingly, we were able to see how well the predictions of enrolled and expelled

students were correct each month.

As of the spring semester of 2023, the actual student prediction rate of enrolled students was 70.33%, and the expulsion prediction rate of expelled students was 89.50%. In addition to the accuracy of the prediction model itself, the retention-dropout prediction for actual K University students in the prediction model as of the end of the spring semester in 2023 showed an accuracy of 79.92% (see <Table 5>). Although the accuracy was very high at the beginning of the semester and the prediction accuracy decreased toward the end of the semester, the overall prediction accuracy was good.

<Table 5> Results of students' retention-dropout prediction modeling

Target Semester / Month	Number of Students	Real Status		Prediction Result		Prediction Accuracy		
		Enrolled [A]	Expelled [B]	Enrolled [C]	Expelled [D]	Total	Enrolled [C/A]	Expelled [D/B]
Spring 2023	89,305	86,971	2,334	61,170	2,089	79.92	70.33	89.50
July 2023	87,139	86,971	168	74,044	81	66.68	85.14	48.21
June 2023	87,347	87,140	207	74,343	112	69.71	85.31	54.11
May 2023	87,731	87,397	334	74,965	250	80.32	85.78	74.85
April 2023	88,535	87,879	656	75,580	615	89.88	86.00	93.75
March 2023	89,305	88,336	969	77,761	942	92.62	88.03	97.21

In the retention-dropout prediction, students were classified into high-risk groups (over 70), risk groups (50-69), and general groups (less than 50) according to the possibility of dropout. It was as important to identify students at risk of dropping out as accurately predicting dropout, so high-risk groups and risk groups were set up by applying the actual dropout rate of K University. In 2018-2202, the size of K University students expulsion reached 15-17% every year. Based on information disclosure in 2023, K University's dropout rate was 20.76%. These statistics of expulsion and subsequent

dropout were reflected in the setting of risk and high-risk groups. The student distribution that tracked the predictions of the model considering this average expulsion scale is presented in <Table 6>.

(Table 6) Students distribution in the results of retention-dropout prediction model

Variables	Enrolled	Expelled	Group in risk (50~69)	Group in high-risk (more than 70)
Total students	85%	15%	10%	5%
Expelled students	10%	90%	15%	75%

3. Screen of Students' Retention-Dropout Prediction Results

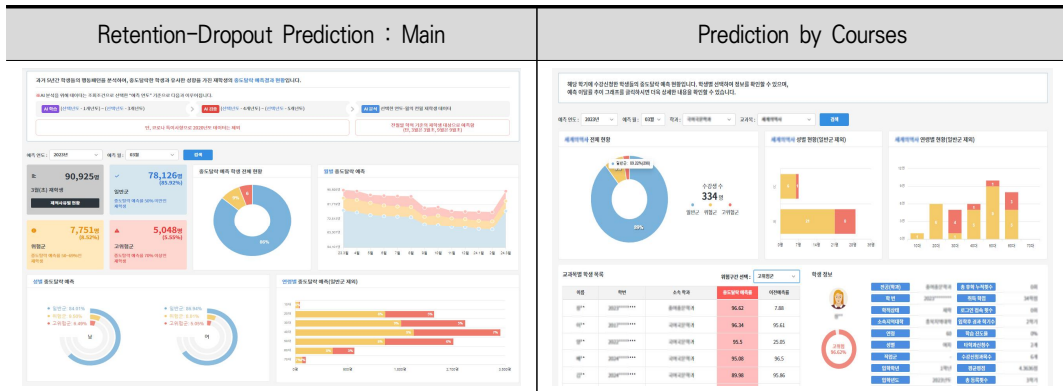
As well as K University students' retention-dropout prediction models, the interactive dashboards have been developed to provide real-time data on student performance and risk levels. These dashboards are designed to be user-friendly for administrators and faculty, enabling quick identification of at-risk students and facilitating immediate intervention. This tool helps in managing student data more efficiently and allows for swift action to support students in need.

The main page provides the overall monthly retention-dropout prediction results, and the page for each subject provided the retention-dropout prediction results of students taking the course. The prediction status is presented by operating a system that predicts retention-dropout on a monthly basis to instructors, and major influencing factors are identified. It is provided to inquire student prediction information by department and subject.

Risk Group Visualization is another critical feature of our system. Each category (general, at-risk, and high-risk) is clearly displayed with corresponding predictive scores and key indicators that influence their risk status. This visual method helps users quickly understand the risk distribution and focus resources more effectively. Trend analysis over time is conducted through graphical representations of data trends over time. This analysis assists in understanding the effectiveness of interventions and monitoring changes in student behavior or performance throughout the academic cycle. By observing these trends, K University can adjust strategies to better meet the needs

of their students. By providing sophisticated visualization tools, K University helps officials make data-based decisions directly that improve student performance and minimize dropout rates. The ability to visualize complex datasets simplifies the decision-making process and supports active educational practices.

<Table 7> Screen with sample data of students' retention-dropout prediction results



Variables mainly related to students' retention-dropout were selected, and information on these variables was visualized and presented. It shows how the main variables of the retention-dropout of students vary depending on the student group(new first grade, transferred second grade, transferred third grade). The differences between the levels of major variables and groups can be identified, and necessary measures can be taken accordingly. The The major influencing factors of K University students on their retention-dropout prediction cover 'number of post-admission semesters', 'percentage of total leave of absence', 'total unregistered percentage', 'number of courses registered(current semester)', 'number of subjects retaken', 'years of admission'. The values of these factors were presented to the admission grade group and the prediction group(general, risk, high-risk), and the average value of enrolled students was presented so that the levels could be compared together (see <Table 8>).

<Table 8> Major influencing factors in students' retention-dropout prediction



V. Conclusion and Discussion

Students in open universities are more likely to drop out than regular college students. Open universities have carried out continuous research and efforts to prevent their dropouts. K University has so far conducted analysis of who continues or stops

studying after they stopped or graduated. However, through AI-based big data analysis, K University was able to predict students whether to continue or stop studying during the semester. Artificial intelligence has opened a new way in higher education institutions that support students' learning based on numerous academic and learning data(Ouyang, Zheng, & Jiao, 2022).

Using the K University data warehouse, this research collected four-year data on the entire process from admission to graduation of about 100,000 students, allowing artificial intelligence to learn. K University students' retention model, developed using multiple algorithms, found more than 220 influencing factors suitable for K University situation in various areas such as academic registration, online learning, and academic achievement. For K University students' retention prediction modeling, the ensemble technique was used. The students' retention prediction modeling was applied on the students enrolled in the spring 2023 semester. The empirical results of K University students' retention prediction system were statistically significant. K University began predicting dropouts by calculating the retention possibilities of individual students every month for a semester. Based on the dropout prediction results, students are divided into three groups (general, at-risk, and high-risk). The dropout prediction results of students are provided to faculty and staff every month along with indicators related to their learning. By department and subject, instructors can not only check the prediction of dropouts of students during the semester, but also check early recognition and support at-risk or high-risk students.

K University's AI-based dropout prediction system has enabled the provision of data-based student support services and system administration. Tailored interventions based on risk categorization incorporate specific characteristics that differentiate the general, at-risk, and high-risk groups. This study proposes targeted interventions that are anticipated to significantly reduce dropout rates.

For the general group, it is proposed that students receive general support measures such as regular academic advising and access to learning resources. These students would benefit from proactive engagement strategies designed to maintain their low-risk status. This approach aims to ensure that students who are currently performing well continue to receive the necessary support to sustain their success.

Students identified in the at-risk group could potentially receive additional support through customized learning plans and frequent monitoring. It would be effective to

consider proposed interventions such as peer tutoring, skill workshops, and more frequent counseling sessions to address their specific needs and challenges. Implementing this targeted support could help prevent potential dropouts by proactively addressing issues as they arise and by adapting the educational approach to meet the evolving needs of these students.

For those in the high-risk group, it is suggested that intensive support measures could be highly beneficial. Regular meetings with a student success coach, personalized academic and psychological support, and potential alterations to their academic workload could effectively manage stress and improve performance. These intensive interventions might be crucial for students who are at significant risk, providing them with the comprehensive support needed to remain in school and succeed. Implementing such measures could greatly enhance the retention rates at K University by providing targeted support where it is most needed.

Each intervention strategy is developed with the understanding that the needs of students vary significantly across different risk levels. By aligning the support services with the characteristics and needs of each group, K University aims to effectively prevent dropouts and enhance student retention. This tailored approach ensures that each student receives the appropriate level of support to thrive in their educational environment.

Leveraging AI-based modeling in conjunction with empathetic and personalized educational practices presents a promising pathway toward reducing dropout rates in open universities. This dual approach not only enhances the predictive accuracy of academic analytics but also ensures that interventions are meaningful and effectively tailored to student needs, ultimately leading to improved educational outcomes and student satisfaction.

K University has established a new framework for comprehensively collecting and providing a large amount of data in the institution, and has established a modeling that predicts retention-dropout of K University students based on artificial intelligence. There are several areas to be discussed in the results of student retention-dropout prediction modeling. In the modeling process, researchers continuously discussed the balance between accuracy and utilization. It was important to increase the accuracy of the prediction, but it was also important to secure the utilization of the modeling. Even if the accuracy was lowered a little, it was necessary to identify students who were likely

to drop out. The degree to which it will be allowed on the spectrum of accuracy and utilization can be determined by whether it is inclined toward research or practice.

Some of the variables identified as important predictors in the modeling results are in line with those related to dropout in previous studies. The variables in the course registration such as ‘number of mid-evaluation’ and ‘number of final evaluation’ were also identified as important factors explaining dropout in previous studies(Kwon et al., 2020; Lee & Kim, 2022). In this study, ‘total access count of week 6’, which was an important variable in predicting retention-dropout, was also discussed as a predictor of dropout in Nam et al.(2022)’s study. In the online learning process, several points can be discussed about the critical period of 4-6 weeks, which is the beginning and middle of the semester. One is that students’ dropouts are concentrated at the beginning of the semester, and if learning is being conducted until the middle of the semester, there is a high possibility of retention. The second is that the learning content of 4-6 weeks is often included as the content of the mid-evaluation. Since students tend to participate a lot in learning about the content for mid-evaluation, the learning behavior of the corresponding week plays an important role in predicting students’ retention-dropout. Finally, online learning data for a specific period may be more important. In Taylor, Veeramachaneni, and O’Reilly(2014)’s study, one week’s data had a stopout prediction accuracy of 0.7 in the MOOC program, and as such, online learning data in the critical period is an important factor in predicting students’ retention-dropout.

Through this study, it was found that the ‘leave of absences’ of existing students most predicted retention-dropout. There have been studies related to ‘leave of absence’, but it has not been examined how students’ leave of absence experiences affect students’ return-dropout. The more leave of absence experience there is, the higher the risk of dropout, so continuous attention and support for students with leave of absence are required.

This study is meaningful in that it expanded AI-based predictive research, which was centered on online higher education, to adult learners studying at open universities. This retention-dropout prediction modeling is technically meaningful in that it has created an environment where significant dropout factors can be derived and analyzed based on artificial intelligence. In addition, as an open university, it is valuable in that it has taken a step toward new practice. K University has established a system to predict all students’ retention-dropout at the institutional level and built a framework to share

these prediction results with instructors and administrative staffs. It provided a service that enables data-based decision-making through institutional and administrative tasks and preemptively responds to students' dropouts. Without stopping here, K University is developing APIs so that they can be used as basic data in other operating systems to support academic recommendations and achievement for students in at-risk groups with a pattern similar to those of dropout students.

These artificial intelligence-based research and practices have issues with the scope of personal information management and utilization. It has the task of agreeing on how to use students' personal and academic data to what extent and how much security setting is appropriate. In addition, as much as accurately predicting the retention-dropout of students, it is as important how to intervene and support students' studies with this prediction system. In the fall semester of 2023, K University launched a new learning counselling service to provide students with the necessary information and support their studies by utilizing the results of their retention-dropout predictions. It is expected that this AI-based prediction system will greatly contribute to the continuation of adult learners' studies at open universities without being alienated or dropped out of the academic process.

References

- Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014). Engagement vs performance: Using electronic portfolios to predict first semester engineering student retention. *Journal of Learning Analytics, 1*(3), 7-33. <https://doi.org/10.1145/2567574.2567583>
- Appavoo, P., Gungea, M., & Sohoraye, M. (2023). Drop-out among ODL learners: A case study at the Open University of Mauritius. *Journal of Educational Technology and Online Learning, 6*(3), 665-682. <https://doi.org/10.31681/jetol.1273563>
- Baneres, D., Rodriguez-Gonzalez, M. E., & Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies, 12*(2), 249-263. <https://doi.org/10.1109/TLT.2019.2912167>
- Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment, 8*, 13-25. <https://eric.ed.gov/?id=ej1062850>
- Eaton, S. B., & Bean, J. P. (1995) An approach/avoidance behavioral model of college student attrition. *Research in Higher Education, 36*, 617-645. <https://doi.org/10.1007/BF02208248>
- Elibol, S., & Bozkurt, A. (2023). Student dropout as a never-ending evergreen phenomenon of online distance education. *European Journal of Investigation in Health, Psychology and Education, 13*(5), 906-918. <https://doi.org/10.3390/ejihpe13050069>
- Im, Y. (2007). A substantial study on the relationship between students' variables and dropout in cyber university. *Journal of the Society for Information Education, 11*(2), 205-219.
- Jeon, J. S. (2010). Identifying at-risk learners at a cyber university. *Andragogy Today, 13*(1), 121-139.
- Joung, Y. R. (2020). A prediction analysis on the dropout of cyber university based on learning analytics. *The Korean Journal of Educational Methodology Studies, 32*(2), 205-232.
- Jung, J. Y., & Lee, J. (2017). An exploratory study on dropout intention of cyber university students. *Korean Education Inquiry, 35*(4), 149-168. <https://doi.org/10.22327/kei.2017.35.4.149>
- Kember, D. (1989). A longitudinal-process model of drop-out from distance education. *The Journal of Higher Education, 60*(3), 278-301. <https://doi.org/10.1080/00221546.1989.1177>

5036

- Kwon, H. J. (2010). The effects of personal, institutional, social variables on determination of the cyber university students' dropout intention. *The Journal of the Korea Contents Association*, 10(3), 404–412. <https://doi.org/10.5392/JKCA.2010.10.3.404>
- Kwon, S., Kim, M., Seo, H. J., & Kim, M. J. (2020). Logistic regressions analysis of the dropout of adult-learners in higher distance education. *Journal of Lifelong Learning Society*, 16(4), 149–169. <https://doi.org/10.26857/JLLS.2020.11.16.4.149>
- Lee, E. K., Jeong, Y., & Kim, M. (2020). Distance learners' motivation clusters and their impact on suspension of study. *Journal of Lifelong Learning Society*, 16(3), 63–91. <https://doi.org/10.26857/JLLS.2020.8.16.3.63>
- Lee, E. K., & Kim, M. (2022). Understanding coursework and college retention in distance education during the COVID-19 crisis. *Journal of Lifelong Learning Society*, 18(1), 147–169. <https://doi.org/10.26857/JLLS.2022.2.18.1.147>
- Marr, L. (2018). The transformation of distance learning at open university: The need for a new pedagogy for online learning? In A. Zorn, J. Haywood, & J. Glachant (Eds.), *Higher education in the digital age* (pp. 23–34). Northampton, MA: Edward Elgar. <https://doi.org/10.4337/9781788970167.00008>
- Moreno-Marcos, P. M., Alario-Hoyos, C., Munoz-Merino, P. J., & Kloos, C. D. (2019). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies*, 12(3), 384–401. <https://doi.org/10.1109/TLT.2018.2856808>
- Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environment*, 30(8), 1414–1433. <https://doi.org/10.1080/10494820.2020.1727529>
- Nam, N., Kim, M., Kim, H., Song, S., & Jang, J. (2022). *Exploring predictors for academic persistence of KNOU students using machine learning techniques*. Seoul: Korea National Open University.
- Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*, 27, 7893–7925. <https://doi.org/10.1007/s10639-022-10925-9>
- Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting student performance in online learning: A multidimensional time-series data analysis approach. *Applied Sciences*, 14(6), 2522. <https://doi.org/10.3390/app14062522>
- Taylor, C., Veeramachaneni, K., & O'Reilly, U. (2014). Likely to stop? Predicting stopout in massive open online courses. *arXiv:1408.3382*. <https://doi.org/10.48550/arXiv.1408.3382>
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125. <https://doi.org/10.3102/003465430450010>

Tyler-Smith, K. (2006). Early attrition among first time eLearners: A review of factors that contribute to drop-out, withdrawal and non-completion rates of adult learners undertaking eLearning programmes. *Journal of Online Learning and Teaching*, 2(2), 73-85.

저 자 정 보

<p>정 상 임 Jung, Sang Im</p>	<p>소 속: 한국방송통신대학교 교육정보화본부 데이터허브팀장 연 락 처: jsn@mail.knou.ac.kr 연구분야: 교육공학, 컴퓨터교육, 컴퓨터공학, 빅데이터분석</p>
<p>이 은 경 Lee, Eun Kyung</p>	<p>소 속: 한국방송통신대학교 미래원격교육연구원 선임연구위원 연 락 처: hannah1222@gmail.com 연구분야: 교육행정, 고등교육, 원격교육, 다문화교육</p>
<p>이 병 래 Lee, Byeong Rae</p>	<p>소 속: 한국방송통신대학교 컴퓨터과학과 교수 연 락 처: brlee@knou.ac.kr 연구분야: 영상처리, 컴퓨터 시각, 머신러닝, 딥러닝</p>
<p>김 상 홍 Kim, Sanghong</p>	<p>소 속: 한국방송통신대학교 미래원격교육연구원 선임연구위원 연 락 처: bestteacher@knou.ac.kr 연구분야: 교육공학, 매체활용교육, 인공지능교육, 컴퓨팅사고력</p>
<p>박 찬 희 Park, Chan Hee</p>	<p>소 속: 한국방송통신대학교 교육정보화본부 데이터허브팀원 연 락 처: panhee@knou.ac.kr 연구분야: 평생교육, 원격교육, 컴퓨터공학, 빅데이터분석</p>

〈요 약〉

인공지능을 활용한 개방대학 학생의 학업유지 예측 모델링

정 상 임 (한국방송통신대학교)

이 은 경 (한국방송통신대학교)

이 병 래 (한국방송통신대학교)

김 상 홍 (한국방송통신대학교)

박 찬 희 (한국방송통신대학교)

개방대학 학생들은 일반대학 학생들보다 중도탈락의 가능성이 높게 나타나기 때문에, 개방대학에서는 학생들의 중도탈락을 막기 위해 지속적인 연구와 노력을 수행해 왔다. 개방대학인 K대학은 다년 간의 연구를 통해 사후에 누가 공부를 계속했는지 혹은 중단했는지에 대한 분석을 수행하였다. 인공지능 기반 빅데이터 분석을 통해, 본 연구는 사후가 아닌 학기 중에 K대학 학생들이 공부를 계속할지 혹은 중단할지를 예측하는 모델을 개발하였다. 대학들은 인공지능을 통해 수많은 학업 및 학습 데이터를 기반으로 학생들의 학습을 지원할 수 있게 되었다.

K대학 데이터 웨어하우스를 이용하여 약 10만 명의 학생들의 입학부터 졸업까지 전 과정에 대한 4년 간의 데이터를 수집하여 인공지능이 학습하도록 하였다. 다중 알고리즘을 이용하여 개발한 K대학 학업유지 모델은 220개 이상의 영향 요인을 확인하였다. K대학은 앙상블 기법을 활용한 학생들의 학업유지 예측 시스템을 구축하고 2023년 봄학기부터 재학생 학업유지 예측을 실현하였다. K대학의 예측 시스템의 실증적 결과는 통계적으로 유의하였다. K대학은 학기 중 매달 개별 학생의 학업유지 가능성을 계산하여 중도탈락을 예측할 수 있게 되었다. 중도탈락 예측 결과에 따라 학생들은 세 그룹(일반, 위험, 고위험)으로 구분되며, 학생들의 중도탈락 예측 결과와 관련 정보는 교수자들에게 제공된다.

인공지능 기반 중도탈락 예측 모델링을 통해, K대학은 데이터 기반의 학생 지원 서비스와 기관 운영을 실천하였다. 중도탈락 위험이 높거나 위험이 있는 것으로 파악된 학생들에게 효과적인 개입과 학습지원에 대하여 논의되었다.

- 주요어: 인공지능, 빅데이터, 학업유지, 학업중단, 예측

접 수 일: 2024. 3. 25.

심 사 일: 2024. 4. 17.

게재확정일: 2024. 4. 30.