J Inf Process Syst, Vol.17, No.3, pp.630~644, June 2021
https://doi.org/10.3745/JIPS.04.0218

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Evaluation of Predictive Models for Early Identification of Dropout Students

JongHyuk Lee*, Mihye Kim**, Daehak Kim*, and Joon-Min Gil**

Abstract

Educational data analysis is attracting increasing attention with the rise of the big data industry. The amounts and types of learning data available are increasing steadily, and the information technology required to analyze these data continues to develop. The early identification of potential dropout students is very important; education is important in terms of social movement and social achievement. Here, we analyze educational data and generate predictive models for student dropout using logistic regression, a decision tree, a naïve Bayes method, and a multilayer perceptron. The multilayer perceptron model using independent variables selected via the variance analysis showed better performance than the other models. In addition, we experimentally found that not only grades but also extracurricular activities were important in terms of preventing student dropout.

Keywords

Educational Data Analysis, Student Dropout, Predictive Model

1. Introduction

In recent years, the field of data analytics in education is attracting increasing attention with the rapid growth of educational data and the spread of big data technology. With the increasing use of online and software-based learning tools, the amounts and types of educational data have increased greatly, and new methods are required to analyze them. Thus, educational data mining [1] and learning analytics [2] are being expanded to explore large-scale data generated by universities and intelligent tutoring systems to better understand students; all of data mining, machine learning, and statistics are employed to this end.

The prediction and prevention of student dropout are very important in terms of continuing education. In other words, we seek not only to identify students at risk of leaving school but also to understand why dropout occurs, which will aid the design of future educational policies. Although many educators have statistically explored relationships between student lifestyle factors and dropout rates, they have focused on ensuring that students graduate, rather than an accurate prediction of dropout. As a result, there are many kinds of researches [3-12] on the features affecting the student dropout and the evaluation of predictive models. However, it is also required the study of the new features and the performance evaluation results when they are reflected in the predictive model. In this paper, we find the new features

Manuscript received October 1, 2018; first revision February 8, 2019; second revision July 10, 2019; third revision October 21, 2020; fourth revision December 21, 2020; accepted December 22, 2020.

Corresponding Author: Joon-Min Gil (jmgil@cu.ac.kr)
* Dept. of Artificial Intelligence and Big Data Engineering, Daegu Catholic University, Gyeongsan, Korea (jonghyuk@cu.ac.kr, dhkim@cu.ac.kr)
** School of Computer Software, Daegu Catholic University, Gyeongsan, Korea (mihyekim@cu.ac.kr, jmgil@cu.ac.kr)

that prevent the students from dropping out and evaluate the models that predict the dropout candidates so that they can get appropriate measures.

Data mining extracts important patterns or knowledge from large amounts of data and is used in retailing, finance, telecommunications, education, fraud detection, stock market analysis, and text mining. Here, to predict and prevent student dropout, we developed and evaluated predictive models using various data-mining methods, including logistic regression (LR), a decision tree (DT), a naïve Bayes (NB) method, and a multilayer perceptron (MP). In particular, we focused on features that exert major influences on dropout; we used analysis of variance to this end and created an optimized model via feature selection. As a result, we found that engagement in extracurricular activities is an important feature to prevent students from dropping out. The results of MP among the four data-mining methods were the best in terms of performance metrics such as F-score and area under the curve.

The remainder of the paper is organized as follows: in Section 2, we present related work on dropout prevention; in Section 3, we introduce the architecture that we used for model creation and evaluation. Section 4 describes the four methods used to generate predictive models, and Section 5 describes the evaluation of the models. Finally, Section 6 contains our conclusions and future plans.

2. Related Work

In data mining, classification is a problem of determining which category a new observation belongs to. Classification is generally regarded as supervised learning using predefined classes. On the other hand, clustering corresponds to unsupervised learning, which groups objects without prior knowledge of classes. Classification and clustering are data-mining techniques by which data are grouped into several classes, and it has many applications. Classification and clustering models can be used to identify tumors [13], fingerprint large numbers of people [14], evaluate employee performance [15], explore student dropout [16-19], monitor anti-phishing [20], and DoS attack detection [21]. Several studies have focused on the causes and features of dropout in terms of prevention, construction of predictive dropout models, and development of dropout prevention systems.

Hoff et al. [3] considered various variables affecting dropout, discussed procedures and tools for the prevention of dropout, and showed examples of early warning systems termed EWIMS (Early Warning Intervention and Monitoring System) [4], DEWS (Dropout Early Warning System) [5], and NDPC-SD (National Dropout Prevention Center for Students with Disabilities) [6]. The variables used to identify dropouts and trigger preventative procedures were attendance, behavior, course performance, race, ethnicity, socioeconomic status, disability status, grade retention, school climate, engagement, and mobility. In this paper, we discovered new variables such extra-curricular activities as to identify dropouts. Yukselturk et al. [7] investigated the applicability of data-mining techniques (k-nearest neighbors, DT, NB, and neural networks) in predicting dropout among online students. Although the differences did not attain statistical significance, the k-nearest neighbors and DT classifiers were somewhat more sensitive than the other models. Manhaes et al. [8] developed an architecture using EDM techniques (an NB model, an MP, a support vector machine, and DT) to predict student dropout. The use of time-varying data aided the prediction of student achievement. The true-positive rate of the NB model was the highest among the four techniques. Guarin et al. [9] evaluated the data-mining models featuring an NB approach and a DT to predict dropout among students with low academic achievement. Dropout

prediction performance was improved when academic data were included. Omoto et al. [10] reviewed institutional research (IR) and Fujitsu trends. IR involves the measurement of many activities via on-campus data collection and analysis, planning of appropriate measures, and implementation and verification of management improvements, student support, and higher-quality educational techniques [10]. Fujitsu developed several statistical methods for the analysis of trends in quality improvement and dropout prevention. Support vector machines were used in data mining. Kuznar and Gams [11] developed a Metis system for the prediction of student dropout and prevention of associated negative consequences. The Metis system uses machine learning algorithms to analyze data from school information systems, identifies students who are likely to drop out, and triggers appropriate action from educational experts. Costa et al. [12] compared the effectiveness of four data-mining techniques (an NB method, a DT, a neural network, and a support vector machine) in predicting the likelihood that students would leave the Brazilian Public University introductory curriculum on programming. The support vector machine performed better than the other techniques.

This study is similar to related work in that we evaluated predictive models using data-mining methods as shown in Table 1. We also chose methods generally used for classification in the related work. However, it differs from related work in that we used new features such as extra-curricular activities and generated a model after selecting the optimal features thereof via LR.

Table 1. Data-mining methods used for identifying dropouts

	k-nearest neighbors	DT	NB	MP (neural network)	Support vector machine	LR
Yukselturk et al. [7]	O	O	O	O		
Manhaes et al. [8]		O	O	O	O	
Guarin et al. [9]		O	O			
Omoto et al. [10]					O	
Kuznar and Gams [11]		O	O		O	O
Costa et al. [12]		O	O	O	O	
Present study		O	O	O		O

3. System Architecture

Our dropout prevention system architecture is divided into five layers (collection, storage, processing, analysis, and visualization), as shown in Fig. 1.

- Collection layer: In this layer, data are collected from various systems inside and outside the organization. Our collector uses various interfaces (e.g., REST, HTTPS, SFTP) to connect different databases and files within the school affairs, learning, and library systems; and logs and documents.
- Storage layer: In this layer, the collected data are held permanently or temporarily in distributed storage. The HDFS and NoSQL approaches are used to permanently store large files and large amounts of messaging data, respectively. Personal data are anonymized.
- Process layer: In this layer, the data are formalized and normalized to render them suitable for analysis. For example, the learning management system processes unstructured log files to generate structured data, such as the numbers and times of logins per student.

- Analysis layer: In this layer, new patterns in large datasets are sought and interpreted to provide novel insights. For example, a predictive model is created to detect students who are likely to drop out.
- Visualization layer: In this layer, the big data results are presented in an easy-to-understand manner. For example, new student information is entered into a predictive model to explore whether particular students dropped out.

Fig. 2 illustrates the model creation and deployment process. Here, we selected significant features with the aid of the chi-squared test and analysis of variance (ANOVA) and generated models using four data-mining methods. ANOVA [22] is a method used when comparing two or more groups in statistics. In this paper, we try to find features that affect the dropout group through the ANOVA. This process is necessary for improving the prediction performance and reducing the learning time. When model performance exceeded the desired threshold, the best model was deployed, and new student data were input to identify students who might drop out.

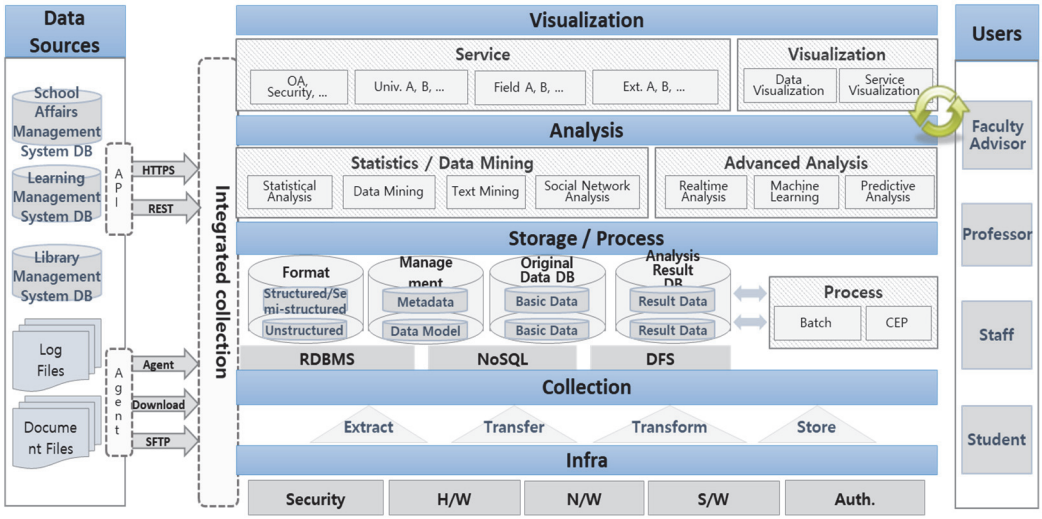


Fig. 1. The system architecture.

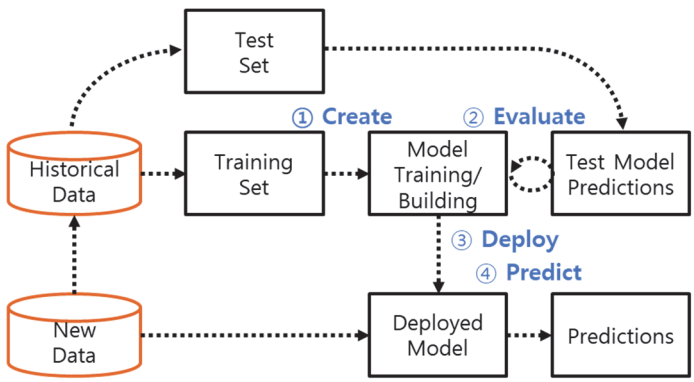


Fig. 2. Model creation and deployment.

4. Classification Methods

We predicted student dropout by analyzing existing data to create a predictive model and then entered new student data into this model to predict dropout. Such data analysis creates a classifier predicting whether dropout occurs (e.g., yes or no). We used LR, a DT, an NB method, and an MP to this end. Although examples for using student data in each method are briefly described in this section, experiments, where actual student data is used on a variety of cases, are shown in the next section.

4.1 Logistic Regression

LR is a well-known classification method for the derivation of relationships between dependent and independent variables, like linear regression (which explains a dependent variable as a linear combination of independent variables), but differing in that LR uses categorical (discrete) dependent variables, rather than numerical (continuous) variables. For example, assuming that student dropout can be categorized into two states (yes or no) by grade point average (GPA), the independent variable is the GPA and the dependent variable is the dropout. When we analyzed the relationship between the GPA and the probability of dropout (Fig. 3), that probability decreased gradually as the GPA rose from zero to a certain point, then rapidly decreased, and later gradually decreased further. Thus, we performed LR analysis employing the logit (or log-odds ratio) to describe the curve mathematically. The following equation is a logistic function created by the logit:

$$logistic\ function = \frac{e^{\beta_i x_i}}{1 + e^{\beta_i x_i}} \tag{1}$$

where the β values are regression coefficients (β_0, …, β_i) and x values are independent variables (x_0, …, x_i).

The odds ratios generated by LR can be used to determine the extent to which independent variables affect dependent variables. Generally, LR analysis is used when the dependent variables fall into two categories; multinomial LR is employed when two or more categories are in play, and ordinal LR is used when the dependent variable is sequential.

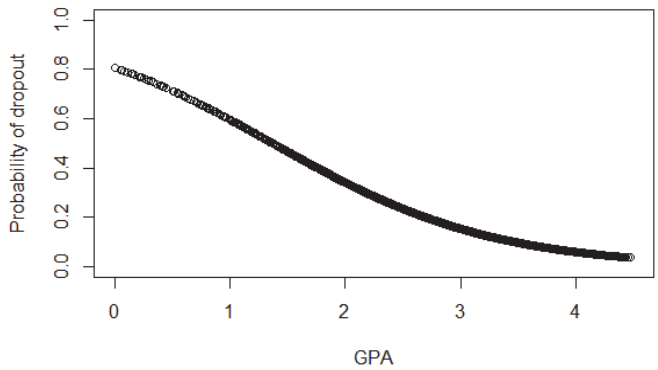


Fig. 3. The relationship between the GPA and the probability of dropout.

4.2 Decision Tree

A DT is a predictive model linking attributes (independent variables) to a class label (the dependent variable). In a DT, an internal node (that is not a leaf node) is used to test the value of an attribute, and this information is used to branch to another internal or leaf node, ultimately determining a class. A DT learns by induction, employing training data with class labels analyzed with the aid of algorithms such as ID3, C4.5, or CART. The algorithms differ in terms of the attribute selection methods (e.g., information gain, gain ratio, and Gini index) used to identify criteria that divide the training data well. A DT is constructed as follows.

- First, an appropriate split criterion and a stopping rule are defined, depending on the purpose of the analysis and the data structure. Here, the split criterion was a classification tree emphasizing the purity of the child node at the expense of that of the parental node. Parameters sharing high-level purity are more likely to belong to the same category. The split criterion varies depending on the type of dependent variable. For example, when the dependent variable is discrete, the *p*-value is used for the chi-squared test, the Gini index, and the entropy index. When the dependent variable is continuous, the F statistic is used for analysis and reduction of variance. Here, we employed the Gini index because the dependent variable (dropout) was discrete. The Gini index is based on a binary split of all attributes. In terms of the discrete value attribute, the subset with the lowest Gini index is selected. For continuous attributes, all possible separable values are considered. The independent variable with the smallest Gini index forms a branch of the DT. The following equation is used to obtain the Gini index:

$$Gini(T) = 1 - \sum_{i=1}^{k} p_i^2, \tag{2}$$

  where T is the dataset, and p_i is the probability that a tuple in T belongs to the i-th class. For example, if the two categories are in a ratio of 0.8:0.2, the Gini index is $1 - (0.8^2 + 0.2^2) = 0.32$. Fig. 4 shows an example of a DT considering the only GPA as an independent variable.
- The tree is pruned because non-standard learning data (noise and outliers) may be present and overfitting is possible. The pruned tree is smaller, less complex, and simpler than the original tree.
- The DT is evaluated against the test data using a cross-validation method.

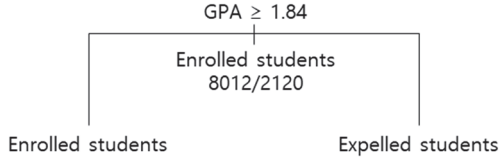


Fig. 4. An example of a DT.

4.3 The NB Approach

The NB approach is based on Bayes' theorem, as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{3}$$

where H and X are events, $P(H|X)$ is the posterior probability of H given that X is true, $P(H)$ is the prior probability of H, $P(X|H)$ is the posterior probability of X given that H is true, and $P(X)$ is the prior

probability of X. The NB approach assumes that the value of any variable is independent of the values of other variables (i.e., lack of dependency among features). For example, independent variables (terms, or the GPA) that affect student dropout are independent; each independent variable is assumed to contribute uniquely to the probability that a student will drop out. An NB classifier is generated as follows:

- If a tuple X with a class label (a dependent variable indicating whether or not dropout occurs) is composed of n attributes (independent variables), that tuple is expressed as a vector $X = (x_1, x_2, ..., x_n)$. We need to find the class label with maximum $P(X|C_i)P(C_i)(i = 1, 2)$ for two labels: C_1 (enrolled students) and C_2 (expelled students).
- When the elements of the vector X are continuous, the continuous value attribute is assumed to follow a standard Gaussian distribution with a mean μ and a standard deviation σ. The probability of the standard normal distribution is as follows:

$$gx, \mu, \sigma=12\pi\sigma 2e-(x-\mu)22\sigma 2 \tag{4}$$

Therefore, $P(x_k|C_i)$ can be written as:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \tag{5}$$

- When the tuple X satisfies the following expression condition, it is considered to a class label C_1:

$$P(X|C_1)P(C_1) > P(X|C_2)P(C_2) \tag{6}$$

4.4 The MP

An MP is used to model the neurons of the human brain involved in learning. Several hidden layers lie between an input and an output layer and identify data that are not linearly separable. Recently, artificial neural networks featuring MPs have been termed deep neural networks, and the algorithm used to study such networks is said to evaluate engagement in "deep learning." The value of an independent variable is input in an input layer node. Usually, the numbers of independent variables and input nodes are identical. The outputs of the input layer are used as inputs in hidden layer nodes. To calculate outputs of the hidden layer, a weighted sum of inputs is calculated and delivered to an activation function that may be linear, exponential, or sigmoid. Here, we used a sigmoid logistic function. The MP is constructed as follows:

- Initialization is performed by assigning connection weights to arbitrary values, calculating the inputs to each layer for a set of learning data (e.g., GPA, terms, etc.), and, finally, calculating outputs employing the activation function.
- After comparing the outputs to the expected values, the connection weight is adjusted via backpropagation to ensure that outputs lie within the error limit.
- This procedure is repeated using other learning data and terminated when the differences between the outputs and the target values are within the acceptable error range.

Here, we used an input layer in which the numbers of nodes and independent variables were equal; thus, 14 nodes with two hidden layers, and one output layer node (Fig. 5).

Although the MP usually shows better predictive performance than LR, the extent to which the input affects the output is difficult to determine. Therefore, we used LR to identify variables that should be preferentially considered. In addition, we selected, from the four methods, that method with the best real predictive performance.

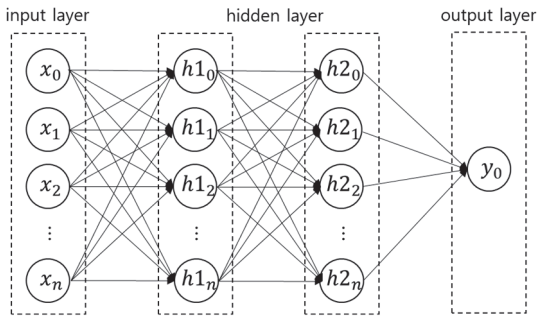


Fig. 5. Architecture of MP.

5. Experiments

5.1 Experimental Environment

We implemented a collector based on Sqoop (https://sqoop.apache.org) for the collection layer in our system architecture. Using the collector, we collected data from our university system and stored it in HBase (https://hbase.apache.org) for the storage layer. The data size is several hundred MB, and the number of samples is about fourteen thousand. Two categories for enrolled students and expelled students are in a ratio of 8:2. Table 2 shows variables affecting dropout in our experiment. We selected thirteen independent variables out of a total of 155 variables, Next, we used Spark (https://spark.apache.org) for the process layer to clean the dependent and independent variables. The cleaned training and test data were randomly divided at a ratio of 7:3. Finally, using the Spark for the analysis layer, we created and evaluated LR, DT, NB, and MP models.

Table 2. Variables and data cleaning

Variable	Data cleaning	Cases used
Dependent variable		
Registration	Enrolled and expelled students.	All cases
Independent variable		
GPA	Calculated grade point average for each student	All cases
Age		Cases #2–#5
Semester		Cases #2–#5
Sex		Cases #2–#5
Engagement in club activities		Cases #3–#5
Nationality	Koreans and foreigners	Case #4
Parental address	Nearby and distant	Case #4
Number of consultations		Case #4
Number of volunteer activities engaged in		Case #4
Number of surveys completed evaluating satisfaction with extra-curricular activities		Cases #4, #5
Number of surveys completed evaluating satisfaction with the department		Cases #4, #5
Extracurricular activities score		Cases #4, #5
Engaged in freshman camp activities		Cases #4, #5

As shown in Table 2, we used the independent variables from five cases to explore how the evaluations changed according to the characteristics and numbers of independent variables employed for dropout prediction.

- Case #1: GPA
- Case #2: Case #1 + {age, semester, sex}
- Case #3: Case #2 + {engagement in club activities}
- Case #4: Case #3 + {nationality, parental address, extracurricular activity score, number of volunteer activities, number of surveys completed evaluating satisfaction with extracurricular activities, number of surveys completed evaluating satisfaction with the department, number of consultations, and engagement in freshman camp activities}
- Case #5: Case #3 + {extracurricular activity score, number of surveys completed evaluating satisfaction with extracurricular activities, number of surveys completed evaluating satisfaction with the department, number of consultations, and engagement in freshman camp activities}

Case #5 was derived from Case #4 (i.e., full feature set) by excluding independent variables (i.e., nationality, parental address, and number of volunteer activities) that did not significantly affect dropout, as revealed by ANOVA (Table 3). ANOVA can be used to improve model performance when an independent variable is included, depending on the difference between the null and residual deviance. We found that engagement in extracurricular activities significantly reduced dropouts.

Table 3. Analysis of variance

	df	Deviance	Residual		Pr(>Chi)
			df	Deviance	
NULL			10,052	10,204.7	
Age	1	1,367.89	10,051	8,836.8	<2.2e-16
Semester	1	2,568.04	10,050	6,268.7	<2.2e-16
Sex	1	25.08	10,049	6,243.7	5.503e-07
GPA	1	456.07	10,048	5,787.6	<2.2e-16
Engagement in club activities	1	567.49	10,046	5,189.4	<2.2e-16
Nationality	1	0.12	10,045	5,189.3	0.72881
Parental address	1	3.41	10,044	5,185.9	0.06478
Extracurricular activities score	1	643.46	10,043	4,542.4	<2.2e-16
Number of volunteer activities engaged in	1	3.40	10,042	4,539.0	0.06518
Number of surveys completed evaluating satisfaction with extra-curricular activities	1	249.53	10,041	4,289.5	<2.2e-16
Number of surveys completed evaluating satisfaction with the department	1	72.42	10,040	4,217.1	<2.2e-16
Number of consultations	1	3.42	10,039	4,213.6	0.06447
Engaged in freshman camp activities	1	38.62	10,038	4,175.0	5.148e-10

5.2 Experimental Results

We used the accuracy, precision, recall, F-score, and area under ROC curve (AUC) parameters to evaluate the four models, as follows. In particular, we selected the F-score and AUC because the class ratio of the experimental data is imbalanced.

$$Accuracy = \frac{TP + TN + FP + FN}{TP + TN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

$$AUC = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT \tag{11}$$

To facilitate comprehension of the above equations, Table 4 shows the confusion matrix. The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are defined as follows.

- TP: The model predicted that students dropped out, and they did in fact drop out.
- TN: The model predicted that students did not drop out, and this was in fact the case.
- FP: The model predicted that students dropped out, but they did not drop out (type I error).
- FN: The model predicted that students would not drop out, but they did drop out (type II error).

Table 4. The confusion matrix

		Actual truth	
		True	**False**
Prediction result	True	TP	FP (type I error)
	False	FN (type II error)	TN

Both TP and TN (only) must be in play if a predictive model is to match all real truths, associated with an accuracy of unity. In this sense, precision is a measure of the extent of type I error, and recall is a measure of the extent of type II error.

Fig. 6 compares the four methods in terms of accuracy. The NB model was less accurate than the other models; the application of the MP method to Case #5 yielded the greatest accuracy (i.e., 0.95). The accuracies of the LR and MP methods increased as more independent variables were added, but the accuracies of the DT and NB methods did not.

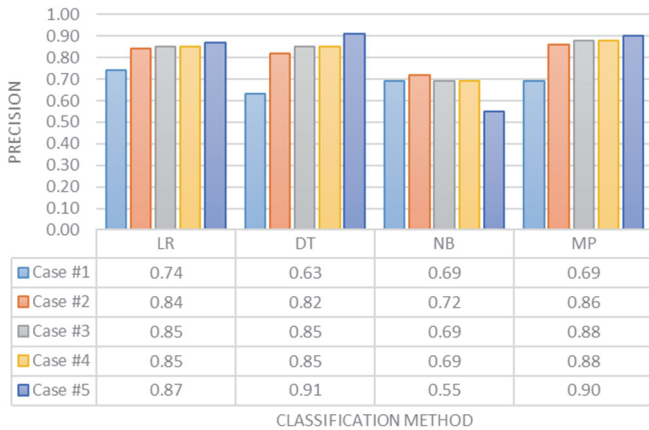
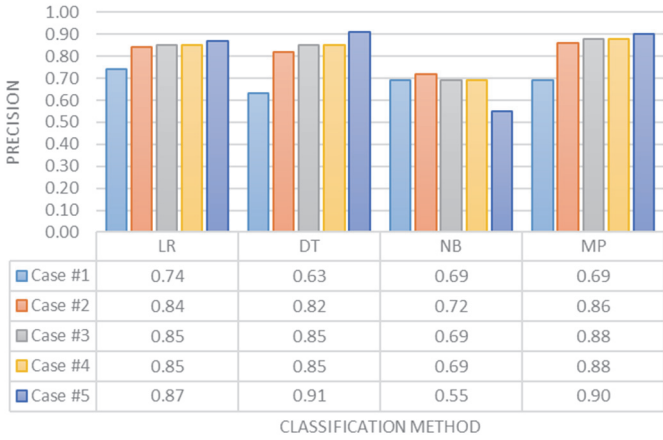


Fig. 6. Comparison of the four methods in terms of accuracy.

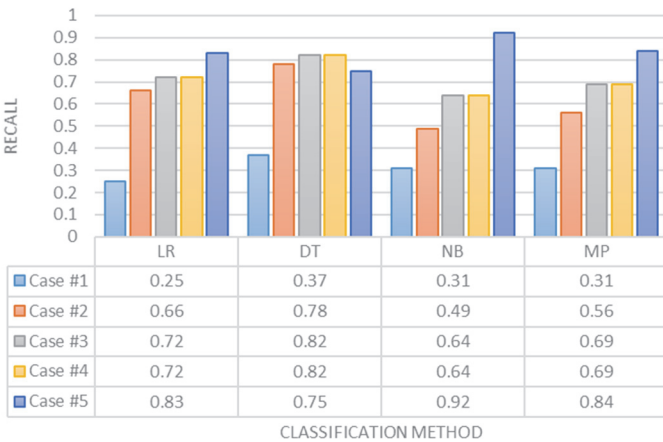
Fig. 7 compares the four methods in terms of precision. The NB method was less precise than the other methods; the application of the DT method to Case #5 yielded the greatest precision (i.e., 0.91). Thus, the use of the NB method resulted in more type I errors than did the use of other methods. The precisions of the LR, DT, and MP methods increased as more independent variables were added, whereas the precision of the NB method did not. The weakness of the NB method is that the dependence between independent variables has a relatively poor predictive performance. As a result of this experiment, the dependency between independent variables such as (age, semester) and (engagement in club activities, extracurricular activities score) seemed to have influenced the performance of the NB method.



	LR	DT	NB	MP
Case #1	0.74	0.63	0.69	0.69
Case #2	0.84	0.82	0.72	0.86
Case #3	0.85	0.85	0.69	0.88
Case #4	0.85	0.85	0.69	0.88
Case #5	0.87	0.91	0.55	0.90

Fig. 7. Comparison of the four methods in terms of precision.

Fig. 8 compares the four methods in terms of recall. The degree of recall was lower for the NB method than for the other methods for Cases #1, #2, #3, and #4; however, the degree of recall for the NB method for Case #5 was the highest (i.e., 0.92). Thus, type II errors created using the NB method were significantly reduced by optimizing variable selection (i.e., from Case #4 to Case #5). The degrees of recall of the LR, NB, and MP methods increased as more independent variables were added, whereas the degree of recall of the DT method did not.



	LR	DT	NB	MP
Case #1	0.25	0.37	0.31	0.31
Case #2	0.66	0.78	0.49	0.56
Case #3	0.72	0.82	0.64	0.69
Case #4	0.72	0.82	0.64	0.69
Case #5	0.83	0.75	0.92	0.84

Fig. 8. Comparison of the four methods in terms of recall.

Fig. 9 compares the four methods in terms of the F-score. The F-score was lower for the NB method than for the other methods. As shown in Fig. 9, the highest F-score was 0.87, obtained by using the MP method to analyze Case #5. Fig. 10 compares the four methods in terms of AUCs. As shown in Fig. 10, the highest AUC was 0.98, obtained when the MP method was used to analyze Case #5. Thus, the predictive model generated by analyzing the independent variables of Case #5 via the MP method showed the best performance. However, the results of the MP method do not differ much from those of the LR method. The LR method is similar to the one-layer neural network and divides the pattern space linearly into two regions. On the other hand, the MP method, which is a two-layer neural network used in this paper, divides the pattern space into convex regions, which is theoretically better than the LR method. We leave it to future studies that the MP method yields much better results than the LR method.

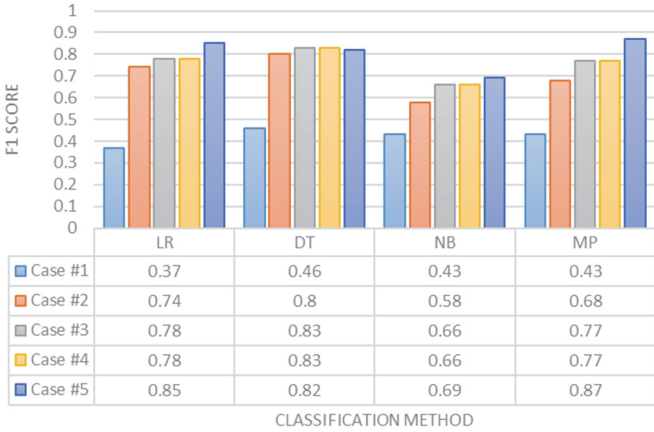


Fig. 9. Comparison of the four methods in terms of the F-score.

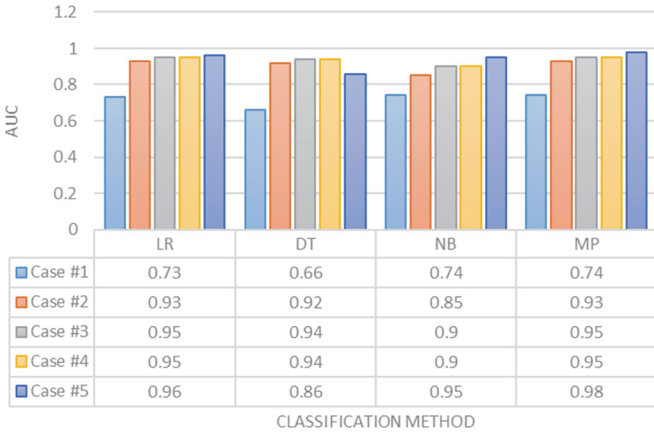


Fig. 10. Comparison of the four methods in terms of the AUC.

6. Conclusions and Future Work

Here, we used LR, a DT, an NB model, and an MP to create predictive models that might provide information for the prevention of student dropout. Predictive models using independent variables selected

with the aid of variance analysis and the MP method showed the best performance (the F-score and AUC were 0.87 and 0.98, respectively).

We will improve the performance of the MP model and apply the optimized model to our school management system to better prevent dropout. We will counsel students who are at risk (as revealed by data analysis), and establish a data-driven campus management plan embracing student guidance, the living environment, and campus activities.

Acknowledgement

This work was supported by research grants from Daegu Catholic University in 2017.

References

[1] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991-16005, 2017.

[2] D. Gasevic, V. Kovanovic, and S. Joksimovic, "Piecing the learning analytics puzzle: a consolidated model of a field of research and practice," *Learning: Research and Practice*, vol. 3, no. 1, pp. 63-78, 2017.

[3] N. Hoff, A. Olson, and R. L. Peterson, "Dropout screening and early warning," University of Nebraska-Lincoln, NE, USA, 2015.

[4] American Institutes for Research, "Early Warning Systems in Education," 2019 [Online]. Available: http://www.earlywarningsystems.org/

[5] Wisconsin Department of Public Instruction, "Dropout Early Warning System," c2021 [Online]. Available: https://dpi.wi.gov/ews/dropout

[6] National Dropout Prevention Center for Students with Disabilities, https://dropoutprevention.org/

[7] E. Yukselturk, S. Ozekes, and Y. K. Turel, "Predicting dropout student: an application of data mining methods in an online education program," *European Journal of Open, Distance and e-learning*, vol. 17, no. 1, pp. 118-133, 2014.

[8] L. M. B Manhaes, S. M. S. Cruz, and G. Zimbrão, "WAVE: an architecture for predicting dropout in undergraduate courses using EDM," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing,* Gyeongju, South Korea, 2014, pp. 243-247.

[9] C. E. L. Guarin, E. L. Guzman, and F. A. Gonzalez, "A model to predict low academic performance at a specific enrollment using data mining," *IEEE Revista Iberoamericana de tecnologias del Aprendizaje*, vol. 10, no. 3, pp. 119-125, 2015.

[10] A. Omoto, Y. Lwayama, and T. Mohri, "On-campus data utilization: working on institutional research in universities," *Fujitsu Science Technology*, vol. 1, no. 51, pp. 42-49, 2015.

[11] D. Kuznar and M. Gams, "Metis: system for early detection and prevention of student failure," in *Proceedings of the 6th International Workshop on Combinations of Intelligent Methods and Applications (CIMA)*, Hague, Holland, 2016.

[12] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araujo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247-256, 2017.

[13] R. Gurusamy and V. Subramaniam, "A machine learning approach for MRI brain tumor classification," *Computers, Materials and Continua*, vol. 53, no. 2, pp. 91-109, 2017.

[14] C. Yuan, X. Li, Q. J. Wu, J. Li, and X. Sun, "Fingerprint liveness detection from different fingerprint materials using convolutional neural network and principal component analysis," *Computers, Materials & Continua*, vol. 53, no. 3, pp. 357-371, 2017.

[15] J. Kaur and K. Kaur, "A fuzzy approach for an IoT-based automated employee performance appraisal," *Computers, Materials and Continua*, vol. 53, no. 1, pp. 24-38, 2017.

[16] N. Iam-On and T. Boongoen, "Generating descriptive model for student dropout: a review of clustering approach," *Human-centric Computing and Information Sciences*, vol. 7, article no. 1, 2017. https://doi.org/10.1186/s13673-016-0083-0

[17] C. A. Christle, K. Jolivette, and C. M. Nelson, "School characteristics related to high school dropout rates," *Remedial and Special Education*, vol. 28, no. 6, pp. 325-339, 2007.

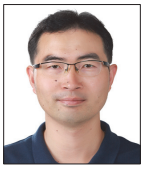
[18] J. Vasquez and J. Miranda, "Student desertion: What is and how can it be detected on time?," in *Data Science and Digital Business*. Cham, Switzerland: Springer, 2019, pp. 263-283.

[19] D. Olaya, J. Vasquez, S. Maldonado, J. Miranda, and W. Verbeke, "Uplift Modeling for preventing student dropout in higher education," *Decision Support Systems*, vol. 134, article no. 113320, 2020. https://doi.org/10.1016/j.dss.2020.113320

[20] D. Jampen, G. Gur, T. Sutter, and B. Tellenbach, "Don't click: towards an effective anti-phishing training: a comparative literature review," *Human-centric Computing and Information Sciences*, vol. 10, article no. 33, 2020. https://doi.org/10.1186/s13673-020-00237-7

[21] D. Tang, R. Dai, L. Tang, and X. Li, "Low-rate DoS attack detection based on two-step cluster analysis and UTR analysis," *Human-centric Computing and Information Sciences*, vol. 10, article no. 6, 2020. https://doi.org/10.1186/s13673-020-0210-9

[22] J. R. Turner and J. Thayer, *Introduction to Analysis of Variance: Design, Analysis & Interpretation*. Thousand Oaks, CA: Sage Publications, 2001.



JongHyuk Lee **https://orcid.org/0000-0002-8163-9388**

He received his Ph.D. of Computer Science Education from Korea University where he did research in distributed systems. He was previously a research professor at Korea University, a research scientist at University of Houston, and a senior engineer at Samsung Electronics. He is currently working as an assistant professor in the Department of Artificial Intelligence and Big Data Engineering at Daegu Catholic University. In the past, he authored and co-authored over publications covering research problems in distributed systems, computer architecture & systems, mobile computing, p2p computing, grid computing, cloud computing, computer security, and computer science education.

Mihye Kim **https://orcid.org/0000-0001-5581-6179**

She received her Ph.D. degree in Computer Science and Engineering from New South Wales University, Sydney, Australia in 2003. She is currently a professor in the School of Computer Software at Daegu Catholic University, South Korea. Her research interests include knowledge acquisition, machine learning, knowledge management and retrieval, computer science education, and Internet of Things.

Daehak Kim **https://orcid.org/0000-0003-4406-3233**

He received his Ph.D. degree in Statistics from Korea University, Korea in 1989 where he did research in kernel estimation of density function and regression function. He is currently working as a professor in the Department of Artificial Intelligence and Big Data Engineering at Daegu Catholic University since 1994. In the past, he authored and co-authored over publications covering research problems in various statistical computing and computational works based on statistical simulation, computer science and big data areas. His recent research interests include artificial intelligence, big data computing and data science.

Joon-Min Gil **https://orcid.org/0000-0001-6774-8476**

He received his Ph.D. degree in Computer Science and Engineering from Korea University, Korea in 2000. Before joining in School of Computer Software, Daegu Catholic University, he was a senior researcher in Supercomputing Center, Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea from October 2002 to February 2006. From June 2001 to May 2002, he was a visiting research associate in the Department of Computer Science at the University of Illinois at Chicago, USA. His recent research interests include artificial intelligence, big data computing, cloud computing, distributed and parallel computing, and wireless sensor networks.