





AKADÉMIAI KIADÓ

# Machine learning-based analysis of adolescent gambling factors

WONJU SEO<sup>1</sup>, NAMHO KIM<sup>1</sup>, SANG-KYU LEE<sup>2\*</sup>  and  
SUNG-MIN PARK<sup>1\*</sup> 

Journal of Behavioral Addictions

9 (2020) 3, 734–743

DOI:  
10.1556/2006.2020.00063  
© 2020 The Author(s)

<sup>1</sup> Department of Creative IT Engineering, Pohang University of Science and Technology, 77 Cheongam-ro, Nam-gu, Pohang, 37673, Republic of Korea

<sup>2</sup> Department of Psychology, College of Medicine, Hallym University, 1 Hallymdaehak-gil, Chuncheon, 24252, Republic of Korea

Received: March 12, 2020 • Revised manuscript received: April 12, 2020 • Accepted: August 26, 2020  
Published online: October 3, 2020

## FULL-LENGTH REPORT



### ABSTRACT

*Background and aims:* Problem gambling among adolescents has recently attracted attention because of easy access to gambling in online environments and its serious effects on adolescent lives. We proposed a machine learning-based analysis method for predicting the degree of problem gambling. *Methods:* Of the 17,520 respondents in the 2018 National Survey on Youth Gambling Problems dataset (collected by the Korea Center on Gambling Problems), 5,045 students who had gambled in the past 3 months were included in this study. The Gambling Problem Severity Scale was used to provide the binary label information. After the random forest-based feature selection method, we trained four models: random forest (RF), support vector machine (SVM), extra trees (ETs), and ridge regression. *Results:* The online gambling behavior in the past 3 months, experience of winning money or goods, and gambling of personal relationship were three factors exhibiting the high feature importance. All four models demonstrated an area under the curve (AUC) of >0.7; ET showed the highest AUC (0.755), RF demonstrated the highest accuracy (71.8%), and SVM showed the highest F1 score (0.507) on a testing set. *Discussion:* The results indicate that machine learning models can convey meaningful information to support predictions regarding the degree of problem gambling. *Conclusion:* Machine learning models trained using important features showed moderate accuracy in a large-scale Korean adolescent dataset. These findings suggest that the method will help screen adolescents at risk of problem gambling. We believe that expandable machine learning-based approaches will become more powerful as more datasets are collected.

### KEYWORDS

adolescents, problem gambling, machine learning-based analysis method, feature engineering

## INTRODUCTION

Recent advances in information and entertainment technologies have provided not only significant benefits to everyday life but also have engendered potentially harmful activities. For example, the easy accessibility to online gambling activities through web browsers (Lavoie & Ladouceur, 2004) and smartphone apps (Calado, Alexandre, & Griffiths, 2017b) has created social issues (Griffiths, 2003). Adolescents can be highly vulnerable to video games with loot boxes, and conventional online gambling such as casino, lottery, card games, and sports betting via virtual platforms (UK Gambling Commission, 2018). Moreover, they are at higher risk than adults for gambling addiction and problem gambling because of their greater physical and psychological instability (Gupta & Derevensky, 2000). Previous studies on problematic gambling in children and adolescents have concluded that teenagers are vulnerable to gambling problems, and most of these studies have found that the rate of problem gambling of teenagers is 4-fold higher than that of adults (Jacobs, 2000). Particular, South Korean adolescents are exposed to various addictive materials via the Internet because

\*Corresponding author.  
Tel.: +82 10 7208 7740.  
E-mail: sungminpark@postech.ac.kr  
E-mail: skmind@hallym.ac.kr

of excessive academic stress, entrance exam-oriented education, and lack of recreational activities (Park & Kim, 2018). Therefore, a deep understanding of their gambling characteristics as well as efforts to identify adolescents at risk of problem gambling are essential.

Adolescent gambling behavior is often associated with various social and personal issues. First, to provide resources for more gambling or to pay off gambling debts (Kryszajtys et al., 2018; Magoon, Gupta, & Derevensky, 2005), adolescents with gambling addiction may undertake illegal activities. Second, these adolescents can experience physical (Giralt et al., 2018) and mental disorders, such as cognitive impairment, mental distress, poor academic achievement, suicidal tendencies, and low self-esteem (McCormick, Russo, Ramirez, & Taber, 1984; Rossen et al., 2016). Problem gambling can degrade family and social relationships of adolescents with gambling addiction (Gupta & Derevensky, 1997). Hence, the importance of early prevention has become increasingly evident (Kang, Ok, Kim, & Lee, 2019).

Gambling is a progressive problem behavior and, thus, can have a serious effect if it begins at a young age. Considering that gambling behaviors are generated and maintained by the interactions of cognitive, emotional, behavioral, and physiological factors, the individual's psychology is not considered the only factor to affect the onset of gambling problems. We therefore should examine the factors that affect adolescent gambling behaviors in an overall context and study the ecological system, considering the interaction and dynamics within the recent environment where the access to gambling activities became much easier (Derevensky & Gilbeau, 2015; Livazović & Bojčić, 2019). These efforts will significantly help in preventing problem gambling of adolescents.

Considering the issues related to adolescent problem gambling, better actions have to be developed using a predictive model to identify an individual at risk of problem gambling. Specifically, an effective predictive model has the following characteristics: (1) the capacity to identify an individual at risk of problem gambling by considering various gambling factors and, thereby, understanding the relationships between problem gambling and these factors, and (2) the ability to warn adolescents regarding the risk of problem gambling with high accuracy.

A machine learning-based analysis method is well suited for building this model. The method includes a feature engineering technique, unlike the conventional statistical methods with limited feature engineering, such as feature extraction and feature selection (Mak, Lee, & Park, 2019). The feature selection process helps determine the factors (among the environmental, psychological, biological, and social institutional factors, the last of which includes policies, laws, regulations, and the relationships with family and friends) that are important to predict the degree of problem gambling. This feature selection process enables a more in-depth study and overcomes the limitations of previous research studies that examined these factors separately (Calado, Alexandre, & Griffiths, 2017a). Although the method of machine learning-based analysis has these

abilities as well as some studies have used the method for providing gambling risk information (Hassanniakalager & Newall, 2019) or gambling-related events such as setting limits for gambling (Auer & Griffiths, 2019), self-exclusion (Percy, França, Dragičević, & d'Avila Garcez, 2016; Philander, 2014), and identification of high-risk Internet gamblers (Braverman, LaPlante, Nelson, & Shaffer, 2013), few studies have predicted the degree of problem gambling based on the Gambling Problem Severity Scale (GPSS) and sought to understand what features are important (Mak et al., 2019).

In this study, we propose a machine learning-based analysis method to predict the degree of problem gambling of adolescents and discuss how the method can be applied to the field of a gambling addiction analysis. This method presents a new perspective to analyze gambling addiction and has potential to become a powerful tool to prevent gambling addiction of adolescents.

## METHODS

### Participants

This study was based on the 2018 National Survey on Youth Gambling Problems dataset (conducted by the Korea Center on Gambling Problems) of students from the first grade of middle school (13 years old in Korean education system) to the second grade of high school (17 years old in Korean education system).

### Subsampling

Considering that the adolescents who have recently gambled are more at risk of problem gambling, we excluded individuals who had not participated in any gambling activities in the past 3 months. The study also excluded individuals who did not answer several numerical questions in their self-report, such as those regarding age at gambling onset, money spent on the most frequent gambling behavior (KRW) in the past 3 months, money lost to the most frequent gambling behavior (KRW) in the past 3 months, and average monthly allowance (KRW).

### Measurements

**General questionnaire.** A self-report questionnaire was used to extract the participant demographic information, gambling behaviors, awareness of and attitudes toward gambling, and other information (e.g., family background and average allowance per month). For the machine learning-based analysis, the gambling factors were extracted from the self-reports. From the demographic information, sex, age, and region of residence extracted. From the gambling behavior information, gambling factors were extracted: (1) online gambling behavior in the past 3 months such as "yes" or "no"; (2) number of gambling behavior in the past 3 months; (3) most frequent gambling behavior in the past 3 months; (4) frequency of the most frequent gambling behavior in the past 3 months; (5) average time

(min) per day spent the most frequent gambling behavior in the past 3 months; (6) money spent on the most frequent gambling behavior (KRW) in the past 3 months; (7) money lost on the most frequent gambling behavior (KRW) in the past 3 months; (8) awareness of the amount money spent on most frequent gambling behavior such as “small” or “large”; (9) the first cognitive path to the most frequent gambling behavior in the past 3 months; (10) the main place for gambling behaviors in the past 3 months; (11) the form of money/stuff transaction in the past 3 months; (12) people who have been together for gambling behaviors in the past 3 months; (13) and the main reason for gambling in the past 3 months. From the awareness of and attitude toward gambling information, gambling factors were extracted: (1) experience of winning money and goods; (2) age at gambling onset; (3) the time of year when engaged in gambling such as on “vacation” or during “the school year”; (4) academic performance degradation due to gambling; (5) experience borrowing money from acquaintances due to gambling; (6) experience borrowing money from a facility due to gambling; (7) serious thoughts of suicide due to gambling; (8) experience of planning suicide due to gambling; (9) nearby presence of people engaged in online gambling or sports betting; (10) intention to participate in gambling as being an adult; (11) awareness of adolescent problem gambling such as “non-serious” or “serious”; (12) participation in gambling prevention education; (13) contact with a promotion or campaign to inform about the risk of gambling; and (14) what activities are needed to prevent problem gambling in adolescents. From other information, gambling factors were extracted: (1) probabilistic item purchase experience while playing online games; (2) gambling of personal relationships, defined as the presence of peer gambling, parental gambling, sibling gambling, or other contacts who gamble; (3) presence of nearby gambling facilities; (4) father’s country of origin either “Korea” or “non-Korea/do not know”; (5) mother’s country of origin either “Korea” or “non-Korea/do not know”; (6) living with parents; (7) honest communication with family such as “not at all/not enough” or “somewhat/frequent”, and (8) average monthly allowance (KRW).

**Gambling Problem Severity Scale.** To evaluate the degree of problem gambling of the adolescents, we used the GPSS, a subscale of the Canadian Adolescent Gambling Inventory (Tremblay, Stinchfield, Wiebe, & Wynne, 2010). The GPSS consists of nine questions with a 4-point Likert scale (Kang et al., 2019). Cronbach  $\alpha$ -value for the GPSS of sampled participants was 0.768, which is considered reliable.

### Machine learning-based analysis method

**Preprocessing.** The study participants were categorized into two classes based on their GPSS scores: those with low GPSS scores (0–1) were considered to have no problem gambling (Class 0) and those with high GPSS scores ( $\geq 2$ ) were considered to be at low to moderate risk of gambling harm or higher (Class 1). Adolescents with medium-to-high-

severity risk based on GPSS scores may be vulnerable to gambling addiction; therefore, the analysis entailed a binary classification.

Several gambling factors were considered and converted to gambling features that served as inputs for the model. Of the 38 factors extracted from the general questionnaires (described in the Measurement section), 31 were categorical factors, which were converted into gambling features with dummy coding. The coding represents a categorical factor having  $N$  items, and  $N-1$  dummy variables were generated such that when the item is  $i$ , the  $i$ -th dummy variable sets to 1 and the others set to 0. In the reference group, all dummy variables are 0 (Hayes & Preacher, 2014). There are seven numerical gambling factors; that is, seven numerical gambling features. We converted the 38 gambling factors into 92 gambling features by including 85 categorical gambling features and 7 numerical gambling features.

**Feature engineering.** After extracting the features, we had to find a feature set that was most relevant to the degree of problem gambling. Adequate feature selection is an important step not only to prevent overfitting the model but also to accelerate the training and help us understand how the machine learning model makes a decision. First, to check a collinearity between features, variance inflation factors were calculated. Then, a random forest-based feature selection was used because it considers both numerical and categorical features (Wang, Yang, & Luo, 2016). In this case, we used permutation importance (Cutler, Cutler, & Stevens, 2012), which is determined by calculating the difference between the error rates before and after the permutation. The error rate before permutation is calculated by a trained RF on its out-of-bag data which is the data that the RF model does not use during training (Breiman, 2001). The  $k$ th feature is then selected, and the value of the feature is permuted. The error rate after permutation is then calculated by the RF on the permuted data. If the  $k$ th feature’s difference is relatively large after calculating the permutation error rate for all features, we can conclude that the  $k$ th feature is important. If the difference is not large, we can conclude that the feature is not important. In the feature selection process, we selected the top 10 features with high feature importance in a training set and ignored the rest.

**Machine learning-based predictive models.** We used four machine learning models (random forest [RF], support vector machine [SVM], extra trees [ETs], and ridge regression [RR]) to develop a predictive algorithm. These models are frequently applied in various medical fields (Goetz et al., 2014; Seo, Kim, Kim, Lee, & Park, 2019). RF (Breiman, 2001) is an ensemble method that uses several individual decision trees to make a decision. Although a single tree is highly intuitive and easy to understand, it shows high variance and is easily overfitted in a training dataset when it has a high volume of depth. With the ensemble approach, RF reduces the variance of the single tree and maintains its bias, thereby resolving the overfitting problem of the single tree. SVM (Cortes & Vapnik, 1995) is a linear classifier with



a maximized margin, which ensures greater generalization ability and thus lower variance than a simple linear classifier. With a kernel trick, SVM can classify data points on nonlinear dimensions, which enables SVM to learn nonlinear relationships. ET (Geurts, Ernst, & Wehenkel, 2006) is a variant of RF that increases randomness by randomly splitting each node with a candidate feature and choosing the best split. Last, RR is similar to a linear regression, except that the output is a logistic function (or sigmoid function) ranging from 0 to 1 and it uses L2 regularization for its weights.

**Model evaluation.** To evaluate the machine learning models, we used the area under the curve (AUC), accuracy, and F1 score, which are typically used in a binary classification problem. We used the following equations:

$$F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (1)$$

$$\text{Recall (or sensitivity)} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}. \quad (2)$$

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}. \quad (3)$$

$$\text{AUC} = \text{Area under a receiver operating characteristic curve}. \quad (4)$$

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{All cases}}. \quad (5)$$

In evaluation, the label of the positive class is Class 1 which means that an adolescent has a low or moderate risk of gambling or more. The F1 score is the harmonic mean of a precision and a recall, thereby calculating a balanced performance between precision and recall, which have a trade-off relationship. Accuracy was calculated by dividing all correctly predicted cases by all cases. AUC was calculated as the area under a receiver operating characteristic (ROC) curve (Greiner, Pfeiffer, & Smith, 2000). Models with an AUC of <0.5 are considered inferior to random prediction. An AUC of 0.5–0.7 is considered as a less accurate performance, 0.7–0.9 is considered as a moderately accurate performance, and 0.9–1.0 is considered as a highly accurate performance (Greiner et al., 2000). Thus, AUC scores should be >0.7 to ensure the accuracy of the trained model.

**Machine learning-based analysis frame.** Fig. 1 shows the proposed machine learning-based analysis method frame. First, we selected participants who had gambled in the past 3 months. In the second step, gambling factors were extracted from the self-reported questionnaires and were converted into gambling features to train four machine learning models. We split the sampled participants into a training set (70%) and a testing set (30%) with a stratified sampling method. Then, VIFs were calculated and the collinearity was checked. Next, the 10 features were selected by the RF-based feature selection method. Using 5-fold stratified cross-validation on the training set, the models' hyper-parameters (e.g., number of trees for RF, a type of a kernel for SVM) were

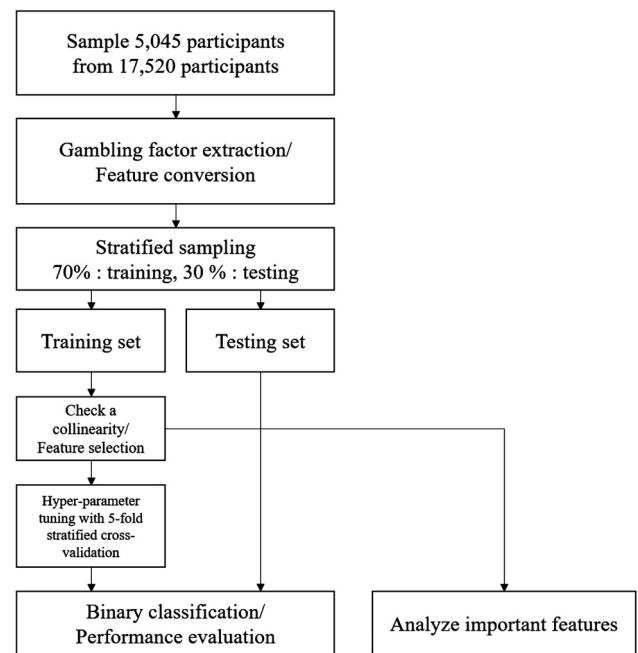


Fig. 1. The proposed machine learning-based analysis method frame. In training models, a grid search method with 5-fold stratified cross-validation was used to tune the hyper-parameters of each model

then tuned in a grid search method. Numeric features were scaled using a MinMax scaler. Because of data imbalance between the number of adolescents in Class 1 versus Class 0, the models were trained with a cost-sensitive learning that gives more weight to the minor class than the major class (Longadge & Dongre, 2013). For the evaluation, we calculated the AUC, accuracy, and F1 score on the testing set.

## Statistical analysis

The IBM SPSS Statistics software was used to read the original set. All important tasks, including data pre-processing and training machine learning models, were performed in Python with sklearn, numpy, and pandas libraries. We used chi-square test to analyze the gambling features. The statistical significance was set at <0.05.

## Ethics

The study analyzed the secondary data based on the original data from the 2018 National Survey on Youth Gambling Problems dataset, which was conducted by the Korean Center on Gambling Problems. All participants' personal information was removed from the original data, and the study was approved by the Institutional Review Board of Hallym University (No. CHUNCHEON 2020-03-004-001).

## RESULTS

### Demographic characteristics

The general characteristics of the study participants are presented in Table 1. Following the exclusion criteria, 56

Table 1. General characteristics of the study population

Variables	Sampled set Frequency(%) / Mean $\pm$ 1 SD	Class 0 (GPSS/CAGI $\leq$ 1) Frequency(%) / Mean $\pm$ 1 SD	Class 1 (GPSS/CAGI $\geq$ 2) Frequency(%) / Mean $\pm$ 1 SD
Total	5,045	3,920	1,125
Sex			
Female	2,467(48.9)	2008(51.2)	459(40.8)
Male	2,578(51.1)	1912(48.8)	666(59.2)
Age	15.0 $\pm$ 1.4	14.9 $\pm$ 1.4	15.1 $\pm$ 1.5
School year			
Middle school 1	944(18.7)	758(19.3)	186(16.5)
Middle school 2	991(19.6)	787(20.1)	204(18.1)
Middle school 3	1,034(20.5)	820(20.9)	214(19.0)
High school 1	999(19.8)	785(20.0)	214(19.0)
High school 2	1,077(21.3)	770(19.6)	307(27.3)
Region of residence			
Capital (Seoul)	504(10.0)	421(10.7)	83(7.4)
Metropolitan area	1,656(32.8)	1,349(34.4)	307(27.3)
Provinces	2,885(57.2)	2,150(54.8)	735(65.3)
Age at gambling onset, years	12.7 $\pm$ 2.4	12.8 $\pm$ 2.4	12.6 $\pm$ 2.7
Number of gambling behaviors in the past 3 months	2.3 $\pm$ 1.8	2.0 $\pm$ 1.5	3.3 $\pm$ 2.2
GPSS/CAGI score			
$\leq$ 1 (i.e. Class 0)	3,920(77.7)	3,920(100)	0(0)
$\geq$ 2 (i.e. Class 1)	1,125(22.3)	0(0)	1,125(100)
Gambling of personal relationships			
No or do not know	3,883(77.0)	3,248(82.9)	635(56.4)
Yes	1,162(23.0)	672(17.1)	490(43.6)
Nearby gambling facilities			
No or do not know	4,382(86.9)	3,462(88.3)	920(81.8)
Yes	663(13.1)	458(11.7)	205(18.2)
Average monthly allowance (\$) <sup>a</sup>			
None	149(3.0)	110(2.8)	39(3.5)
Less than \$40	2,103(41.7)	1,695(43.2)	408(36.3)
Less than \$80	1,548(30.7)	1,227(31.3)	321(28.5)
Approximately \$80–\$240	1,077(21.3)	791(20.2)	286(25.4)
Approximately \$240–\$400	132(2.6)	80(2.0)	52(4.6)
Approximately \$400–\$800	26(0.5)	10(0.3)	16(1.4)
Greater than or equal to \$800	10(0.2)	7(0.2)	3(0.3)

<sup>a</sup> Changed KRW to US dollar (\$). Abbreviations: GPSS, Gambling Problem Severity Scale; CAGI, Canadian Adolescent Gambling Inventory.

subjects from 5,101 subjects who participated in at least one gambling behavior in the past 3 months were excluded.

Of the 5,045 participants, 2,578 were men (51.1%) and 2,467 were women (48.9%); the mean age of the participants was 15.0  $\pm$  1.4 years; 10.0% lived in Seoul, 32.8% lived in the metropolitan area, and 57.2% lived in other provinces. The mean age at gambling onset was 12.7  $\pm$  2.4 years, and the mean number of gambling behaviors in the past 3 months was 2.3  $\pm$  1.8. Moreover, the mean number of gambling behaviors of Class 1 was 3.3  $\pm$  2.2. Playing >2 games on average at an early age might indicate that the participant is at risk of problem gambling (Erens, Mitchell, Orford, Sproston, & White, 2004). Of the 5,045 participants, 3,920 (77.7%) were included in Class 0 and 1,125 (22.3%) were included in Class 1, suggesting that many adolescents are at risk of gambling addiction. Both the gambling of personal relationships and the presence of nearby gambling facilities

were considered. Regarding the gambling of personal relationships, 77.0% of the adolescents overall had no such relationships. The percentage of the gambling of personal relationships of Class 1 (43.6%) was much larger than that of Class 0 (17.1%). Regarding nearby gambling facilities (such as racecourses, bullfighting stadiums, casinos, lotteries, sports, toto shops, adult game rooms, and others), 86.9% of the adolescents responded that there were no nearby facilities. For average monthly allowance (\$), 41.7% of the adolescents received an allowance of less than \$40, followed by less than \$80 (30.7%). Only 0.7% received an allowance greater than or equal to \$400 per month.

### Feature importance analysis

After preprocessing and feature extraction, 92 features were obtained and used to train the four models by following the



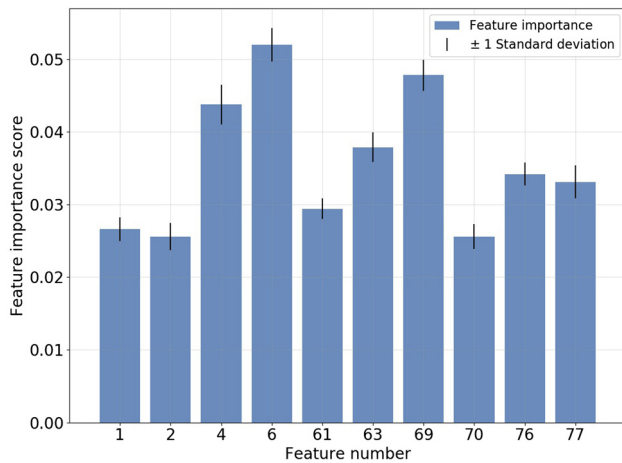


Fig. 2. Calculated permutation importance of selected features with the random forest-based method. Y axis = feature importance. X axis = selected features' number. A blue box indicates the mean feature importance. A black line indicates  $\pm 1$  standard deviation

machine learning-based analysis frame. The MinMax scaler was fitted on seven numerical features in the training set, and then the scaler transformed the numerical features in the testing set to prevent information leakage. When calculating VIF of all features, there was no feature having high VIF ( $>5$ ); hence, all features were utilized. Then, the RF-based feature selection was performed. Their mean importance and standard deviation are detailed in Fig. 2.

In the feature selection process, 10 features were selected as the top 10 features in the training set. In order, each feature was included in “sex,” “region of residence,” “gambling of personal relationships,” “online gambling behavior in the past 3 months,” “the main reason for gambling in the past 3 months,” “awareness of the amount money spent on the most frequent gambling behavior,” “experience of winning money or goods,” “the time of year when engaged in gambling,” “nearby presence of people engaged in online gambling or sports betting,” or “probabilistic item purchase experience while playing online games.” Among the various gambling factors, three gambling factors having the high feature importance were “online gambling behavior in the past 3 months,” “experience of winning money or goods,” and “gambling of personal relationships.”

When Class 1 was analyzed according to each factor, 517 (42.3%) of the 1,223 who had experience of online gambling in the past 3 months and 608 (15.9%) of the 3,822 who did not gamble online in the past 3 months were included in this class. Of the 3,198 participants who had an experience of winning money or goods, 858 (26.8%) were included, whereas of the 1,847 who had no experience of winning money or goods, 267 (14.5%) were included. Of the 1,162 participants who had the gambling of personal relationships, 490 (42.2%) were included, whereas of the 3,883 who did not have the gambling of personal relationships or replied “do not know”, 635 (16.4%) were included. When chi-square test

Table 2. Metrics on the testing set of each model

Model	AUC	Accuracy (%)	F1 score
RF	0.752	71.8	0.504
SVM	0.747	71.4	0.507
ET	0.755	71.5	0.502
RR	0.753	69.9	0.495

Abbreviations: AUC, area under the curve; ET, extra trees; RR, ridge regression; RF, random forest; SVM, support vector machine.

was performed on the two groups for each factor, results showed significant differences ( $p < 0.05$ ).

## Model evaluation

After feature engineering and training, we calculated the AUC, accuracy, and F1 score for the four models (RF, SVM, ET, and RR) (Table 2).

Of the four models, ET showed the highest AUC (0.755), RF demonstrated the highest accuracy (71.8%), and SVM demonstrated the highest F1 score (0.507) on the testing set. Considering that all models demonstrated an AUC of  $>0.7$ , we concluded that they exhibited moderately accurate performance (Greiner et al., 2000). When only considering AUC, we found that ET was the best among the four models. Fig. 3 shows the plots for the ROC curves for all models.

## DISCUSSION

This study is the first to apply a machine learning-based analysis method to analyze a large-scale Korean adolescent

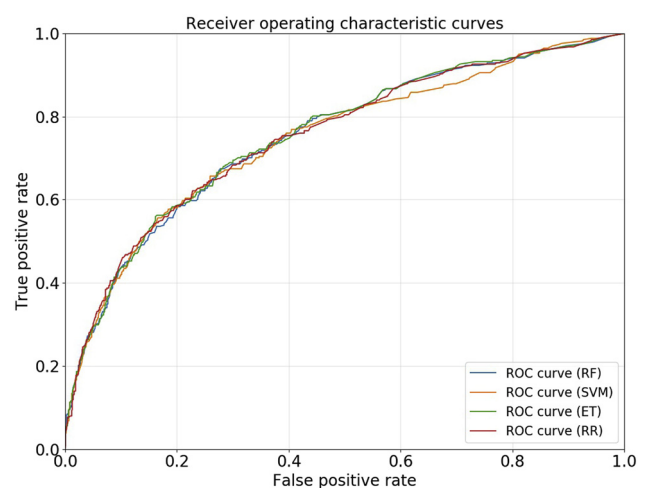


Fig. 3. ROC curves for all models. Each color line indicates each model's ROC curve. The blue line indicates the ROC curve of RF. The orange line indicates the ROC curve of SVM. The green line indicates the ROC curve of ET. The red line indicates the ROC curve of RR

dataset. The method includes the following: (1) data sampling, (2) factor extraction and feature conversion, (3) feature engineering and model development, and (4) performance evaluation and import feature analysis. After the feature extraction and conversion, we checked the collinearity between features and then the top 10 gambling features were selected with the RF-based feature selection method. As a result, three gambling factors having the high feature importance were obtained and we demonstrated that all models had a moderately accurate performance for prediction. In addition, the numbers of Class 1 adolescents were compared according to each factor. In our opinion, the proposed method helps in preventing problem gambling of adolescents and provides evidence for understanding the gambling factors that have high importance.

The proposed analysis method finds meaningful predictors from complex human behaviors, environmental factors, personal psychology, biological factors, policy factors, laws and regulations, and family and friend relationships. Among the selected gambling features, as is shown in Fig. 2, the main three gambling factors were “online gambling behavior in the past 3 months,” “experience of winning money or goods,” and “gambling of personal relationships.”

The online gambling behavior in the past 3 months is related to accessibility to an online gambling environment. The easy accessibility of online gambling to adolescents via web browsers and smartphone apps facilitates gaming addiction of adolescents (Parker, Taylor, Eastabrook, Schell, & Wood, 2008), leading to serious gambling behaviors (Griffiths & Parke, 2010). Therefore, a strict regulation system is needed to completely prevent adolescents from accessing gambling in online environments.

The experience of winning money or goods may be associated with positive thinking about gambling, which leads to a cognitive distortion of being able to control gambling (King, Delfabbro, & Griffiths, 2010). This distortion results from the assumption that the gambler can control the outcome of gambling; in turn, these errors may induce the development or maintenance of problem gambling behaviors (Yakovenko et al., 2016). Hence, it is necessary to correct any misconceptions related to the control of gambling outcomes (Turner, Zangeneh, & Littman-Sharp, 2006; Yakovenko et al., 2016).

Finally, the gambling of personal relationships might be related to the influence of family and peers (Delfabbro & Thrupp, 2003). If their parents or friends have experience with gambling, the adolescents are likely to engage in gambling (Delfabbro & Thrupp, 2003; Hardoon, Gupta, & Derevensky, 2004). Systematic education on preventing gambling, conveying the seriousness of gambling, and teaching how to use money wisely are needed for not only adolescents but also parents. Although the identification of these 3 gambling factors are not new, previous studies have considered them to be meaningful gambling-related information (Delfabbro & Thrupp, 2003; King et al., 2010; Lorenz & Yaffee, 1988; Parker et al., 2008; Potenza et al., 2011; Turner et al., 2006; Welte, Barnes, Tidwell, & Hoffman, 2009), which indicates that the feature selection in this study

was correct. Although this study addressed the meaning of and preventative actions for the three gambling factors, it did not mean that only these three factors should be used to train a machine learning model.

Several studies have used machine learning approaches to predict gambling-related events such as limit-setting, self-exclusion, and identification of high-risk Internet gamblers (Auer & Griffiths, 2019; Percy et al., 2016; Philander, 2014). The contributions of this current study compared with those studies are two-fold. The first is a classification target. The proposed model in this current study predicted the degree of problem gambling (either Class 0 or Class 1) with an AUC of 0.755. However, one other research group (Auer & Griffiths, 2019) predicted whether there was a limit-setting change with an AUC of 0.76, and other research groups (Percy et al., 2016; Philander, 2014) predicted whether an individual closed his or her account because of the self-exclusion and showed 0.551 of AUC and 0.76 of AUC, respectively; yet another research group (Braverman et al., 2013) predicted high-risk Internet gamblers with low sensitivity (19.8%). The second contribution of the present study is a type of dataset. In this study, the player's circumstances were addressed; however, all the other four studies focused on gambling behavior variables. These differences mean that gambling-related events or the degree of problem gambling can be predicted from the information present in the player. Although it is possible to build a more accurate model using all the information, depending on the situation, the number and type of data (e.g., player circumstance or player behaviors) may be limited.

The machine learning models we developed demonstrated moderately accurate performance, with an AUC of  $>0.7$ . To improve the generalization performance of a machine learning model, we can first collect additional datasets from other countries. This will help develop a model with good generalization performance that can be applied to all countries, which could contribute to prevent problem gambling of adolescents worldwide. However, international cooperation will be necessary to achieve this goal. Second, new gambling factors might be discovered that could explain the still unclear and complex gambling behavior patterns of adolescents, which will require deepening our understanding of adolescent behaviors and converting them into gambling factors. Finally, we could use the ensemble approach, which is based on the idea that models developed with different algorithms can complement each other to increase the generalization performance. Due to these various methods, the machine learning model will be highly likely to develop further.

This proposed machine learning-based analysis method frame can be expanded in various directions. Similar to the RF-based feature selection, for example, other feature selection methods (such as least absolute shrinkage and selection operator, recursive feature elimination, and the chi-squared test) could be used to find the best features with powerful relationships, which can increase the predictive efficiency. Alternatively, we can use a dimension reduction method such as a principal component analysis. From the

model viewpoint, several complex machine learning models (including artificial neural networks) can be applied to capture more complexities among gambling features and the degree of problem gambling. Next, a pretrained machine learning model can be easily installed and operated on mobile devices such as smartphones. The installed model can be used by adolescents to determine their degree of problem gambling on the basis of their input without restriction. Finally, a machine learning-based analysis targeting individuals or a homogenous group can be performed with an accumulated gambling-related behavior dataset of the individuals or groups. The analysis could suggest an effective gambling prevention and intervention, which will lead to strong gambling control for individuals or groups. Ultimately, this expansion can build an elaborate machine learning model for the early diagnosis of problem gambling of adolescents and provide effective warnings about problem gambling, leading to significant improvements in preventing gambling problems.

This study has several limitations. First, the adolescents answered self-reporting questionnaires, which could have introduced a risk of inaccurate answers due to their recall of past gambling behaviors. Second, the cross-sectional study dataset analysis cannot be used to infer casual conclusions. Finally, other factors affecting the gambling behavior of adolescents, such as mental stress (Holub, Hodgins, & Peden, 2005), were not considered. To overcome these limitations, we will collect a new longitudinal dataset including more gambling factors in the future instead of a cross-sectional study and will expand our machine learning-based analysis method.

## CONCLUSION

This study is the first to propose a machine learning-based method for analyzing problem gambling in a large-scale Korean adolescent dataset. Results demonstrated that the machine learning models can predict the degree of problem gambling of adolescents with moderate levels of accuracy, which can provide useful information to support prediction of the degree of problem. With feature engineering, we trained several machine learning models and found that all models demonstrated moderately accurate performance for predicting the degree of problem gambling. “Online gambling behavior in the past 3 months,” “experience of winning money or goods,” and “gambling of personal relationships” were three gambling factors having the high feature importance and the numbers of Class 1 adolescents were analyzed based on these three factors. On the basis of the proposed machine learning-based analysis method, we discussed the future expansion of and the potential for the analysis. We believe that this method will provide new insights into problem gambling of adolescents and will ultimately help in preventing problem gambling.

*Funding sources:* This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the ICT

Consilience Creative program (IITP-2020-2011-1-00783) supervised by the Institute for Information and communications Technology Promotion (IITP), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A5A1015596), the Technology Innovation Program (or Industrial Strategic Technology Development Program, 20001841, Development of System for Intelligent ContextAware Wearable Service based on Machine Learning) funded By the Ministry of Trade, Industry and Energy (MOTIE, Korea), and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2005385). The funding sources had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication.

*Author’s contribution:* WJ conducted data preprocessing, and analyses and wrote the manuscript. SKL and SMP guided and supervised this manuscript. WJ developed a machine learning-based analysis. NK did the former preprocessing. All authors contributed editorial comments on the manuscript.

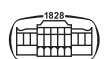
*Conflict of interest:* The authors declare no conflict of interest.

*Data availability:* 2018 National Survey on Youth Gambling Problems dataset is available from the Korea Center on Gambling Problems.

*Acknowledgments:* We would like to thank Korea Center on Gambling Problems which provided the original data sources and Lee Jae-Kyung of the Department of Prevention and Public Relations who gave various advice on presenting this research.

## REFERENCES

- Auer, M., & Griffiths, M. D. (2019). Predicting limit-setting behavior of gamblers using machine learning algorithms: A real-world study of Norwegian gamblers using account data. *International Journal of Mental Health and Addiction*, 1–18.
- Braverman, J., LaPlante, D. A., Nelson, S. E., & Shaffer, H. J. (2013). Using cross-game behavioral markers for early identification of high-risk internet gamblers. *Psychology of Addictive Behaviors*, 27(3), 868.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Calado, F., Alexandre, J., & Griffiths, M. D. (2017a). How coping styles, cognitive distortions, and attachment predict problem gambling among adolescents and young adults. *Journal of Behavioral Addictions*, 6(4), 648–657.
- Calado, F., Alexandre, J., & Griffiths, M. D. (2017b). Prevalence of adolescent problem gambling: A systematic review of recent research. *Journal of Gambling Studies*, 33(2), 397–424.





- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157–175): Springer.
- Delfabbro, P., & Thrupp, L. (2003). The social determinants of youth gambling in South Australian adolescents. *Journal of Adolescence*, 26(3), 313–330.
- Derevensky, J. L., & Gilbeau, L. (2015). Adolescent gambling: Twenty-five years of research. *Canadian Journal of Addiction*, 6(2), 4–12.
- Erens, B., Mitchell, L., Orford, J., Sproston, K., & White, C. (2004). *Gambling and problem gambling in Britain*. Routledge.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42.
- Giralt, S., Müller, K. W., Beutel, M. E., Dreier, M., Duven, E., & Wölfling, K. (2018). Prevalence, risk factors, and psychosocial adjustment of problematic gambling in adolescents: Results from two representative German samples. *Journal of behavioral addictions*, 7(2), 339–347.
- Goetz, M., Weber, C., Bloecher, J., Stieltjes, B., Meinzer, H.-P., & Maier-Hein, K. (2014). Extremely randomized trees based brain tumor segmentation. *Proceeding of BRATS challenge-MICCAI*, 006–011.
- Greiner, M., Pfeiffer, D., & Smith, R. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45(1–2), 23–41.
- Griffiths, M. (2003). Internet gambling: issues, concerns, and recommendations. *Cyberpsychol Behav*, 6(6), 557–568. <https://doi.org/10.1089/109493103322725333>. 14756922.
- Griffiths, M. D., & Parke, J. (2010). Adolescent gambling on the internet: A review. *International Journal of Adolescent Medicine and Health*, 22(1), 59–75.
- Gupta, R., & Derevensky, J. (1997). Familial and social influences on juvenile gambling behavior. *Journal of Gambling Studies*, 13(3), 179–192.
- Gupta, R., & Derevensky, J. L. (2000). Adolescents with gambling problems: From research to treatment. *Journal of Gambling Studies*, 16(2–3), 315–342.
- Hardoon, K. K., Gupta, R., & Derevensky, J. L. (2004). Psychosocial variables associated with adolescent gambling. *Psychology of Addictive Behaviors*, 18(2), 170.
- Hassaniakalager, A., & Newall, P. W. (2019). A machine learning perspective on responsible gambling. *Behavioural Public Policy*, 1–24.
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67(3), 451–470.
- Holub, A., Hodgins, D. C., & Peden, N. E. (2005). Development of the temptations for gambling questionnaire: A measure of temptation in recently quit gamblers. *Addiction Research & Theory*, 13(2), 179–191.
- Jacobs, D. F. (2000). Juvenile gambling in North America: An analysis of long term trends and future prospects. *Journal of Gambling Studies*, 16(2–3), 119–152.
- Kang, K., Ok, J. S., Kim, H., & Lee, K.-S. (2019). The gambling factors related with the level of adolescent problem gambler. *International Journal of Environmental Research and Public Health*, 16(12), 2110.
- King, D., Delfabbro, P., & Griffiths, M. (2010). The convergence of gambling and digital media: Implications for gambling in young people. *Journal of Gambling Studies*, 26(2), 175–187.
- Kryszajtys, D. T., Hahmann, T. E., Schuler, A., Hamilton-Wright, S., Ziegler, C. P., & Matheson, F. I. (2018). Problem gambling and delinquent behaviours among adolescents: A scoping review. *Journal of Gambling Studies*, 34(3), 893–914.
- Lavoie, M.-P., & Ladouceur, R. (2004). Prevention of gambling among youth: Increasing knowledge and modifying attitudes toward gambling. *Journal of Gambling Issues*, (10).
- Livazović, G., & Bojčić, K. (2019). Problem gambling in adolescents: What are the psychological, social and financial consequences? *BMC Psychiatry*, 19(1), 308.
- Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.
- Lorenz, V. C., & Yaffee, R. A. (1988). Pathological gambling: Psychosomatic, emotional and marital difficulties as reported by the spouse. *Journal of Gambling Behavior*, 4(1), 13–26.
- Magoon, M. E., Gupta, R., & Derevensky, J. (2005). Juvenile delinquency and adolescent gambling: Implications for the juvenile justice system. *Criminal Justice and Behavior*, 32(6), 690–713.
- Mak, K. K., Lee, K., & Park, C. (2019). Applications of machine learning in addiction studies: A systematic review. *Psychiatry Research*, 275, 53–60.
- McCormick, R. A., Russo, A. M., Ramirez, L. F., & Taber, J. I. (1984). Affective disorders among pathological gamblers seeking treatment. *American Journal of Psychiatry*, 141, 215–218.
- Park, S.-H., & Kim, Y. (2018). Ways of coping with excessive academic stress among Korean adolescents during leisure time. *International Journal of Qualitative Studies on Health and Well-Being*, 13(1), 1505397.
- Parker, J. D., Taylor, R. N., Eastabrook, J. M., Schell, S. L., & Wood, L. M. (2008). Problem gambling in adolescence: Relationships with internet misuse, gaming abuse and emotional intelligence. *Personality and Individual Differences*, 45(2), 174–180.
- Percy, C., França, M., Dragičević, S., & d'Avila Garcez, A. (2016). Predicting online gambling self-exclusion: An analysis of the performance of supervised machine learning models. *International Gambling Studies*, 16(2), 193–210.
- Philander, K. S. (2014). Identifying high-risk online gamblers: A comparison of data mining procedures. *International Gambling Studies*, 14(1), 53–63.
- Potenza, M. N., Wareham, J. D., Steinberg, M. A., Rugle, L., Cavallo, D. A., Krishnan-Sarin, S., et al. (2011). Correlates of at-risk/problem internet gambling in adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(2), 150–159.e153.
- Rossen, F. V., Clark, T., Denny, S. J., Fleming, T. M., Peiris-John, R., Robinson, E., et al. (2016). Unhealthy gambling amongst New Zealand secondary school students: An exploration of risk and protective factors. *International Journal of Mental Health and Addiction*, 14(1), 95–110.
- Seo, W., Kim, N., Kim, S., Lee, C., & Park, S.-M. (2019). Deep ECG-respiration network (DeepER net) for recognizing mental stress. *Sensors*, 19(13), 3021.



- Tremblay, J., Stinchfield, R., Wiebe, J., & Wynne, H. (2010). *Canadian adolescent gambling inventory (CAGI) phase III final report*. Retrieved from <https://prism.ucalgary.ca/handle/1880/48158>.
- Turner, N. E., Zangeneh, M., & Littman-Sharp, N. (2006). The experience of gambling and its role in problem gambling. *International Gambling Studies*, 6(2), 237–266.
- UK Gambling Commission. (2018). *Young people and gambling: 2018 report*. Retrieved from <https://www.gamblingcommission.gov.uk/PDF/survey-data/Young-People-and-Gambling-2018-Report.pdf>.
- Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*, 17(1), 60.
- Welte, J. W., Barnes, G. M., Tidwell, M.-C. O., & Hoffman, J. H. (2009). Legal gambling availability and problem gambling among adolescents and young adults. *International Gambling Studies*, 9(2), 89–99.
- Yakovenko, I., Hodgins, D. C., el-Guebaly, N., Casey, D. M., Currie, S. R., Smith, G. J., et al. (2016). Cognitive distortions predict future gambling involvement. *International Gambling Studies*, 16(2), 175–192.

