Development of Machine Learning Models for Predicting Academic Stress in Adolescents

- A Survey on Mental Health Status of Adolescents in 2021 -

Jeong-suk So1 Tae-yeon Oh2

ABSTRACT

In the field of social science research, traditional statistical methods have been predominantly used to analyze the major causes of adolescent stress. Most existing studies focus on this approach. In relation to academic stress, this study aims to develop an analysis model based on artificial intelligence and machine learning-driven statistical methods, utilizing big data analysis. Academic stress has a significant impact on the mental health of adolescents, and predicting it can provide valuable insights into students' stress levels, ultimately suggesting effective ways to alleviate their stress. The collected data was sourced from the 2021 Korea Children and Youth Data Archive survey, where 225 common variables were selected. To identify specific variables influencing adolescent academic stress and to choose the most effective machine learning method, five regression models—Random Forest, Linear Regression, Ridge, Lasso, and Gradient Boosting—were applied to determine significant variables.

The performance of these regression models was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score as assessment metrics. Based on these evaluations, Gradient Boosting was selected as the most effective predictive model. Additionally, an analysis of variable importance was conducted to identify key factors affecting academic stress

☞ keyword : Academic Stress, Machine Learning, Gradient Boosting, Adolescent Mental Health, Big Data Analysis

1. Introduction

According to the Children and Youth Quality of Life 2022 report by Statistics Korea, the largest proportion of leisure time spent by Korean adolescents is on private academies and tutoring, accounting for 47.3%. In contrast, smartphone usage only accounts for 14.1%. This suggests that students do not have sufficient rest and dedicate a significant amount of time to academic activities.

Additionally, the PISA 2015 report by the OECD indicates that the academic-related anxiety index among Korean adolescents is 0.10, which is higher than the OECD average of 0.01. This demonstrates that Korean adolescents experience relatively higher academic stress compared to their peers in other major countries. Notably, 41.9% of students responded affirmatively to the statement, "I feel very tense when studying. Given that the percentage of OECD adolescents who agreed with the same statement is 36.6%, it can be inferred that academic anxiety and stress are particularly severe among Korean students.[1]

According to the 2023 Youth Statistics report published by the Korea Youth Policy Institute under the Ministry of Gender Equality and Family, a survey on depression and stress among adolescents revealed that 41.3% of middle and high school students reported feeling stressed in 2022. This figure represents a 2.5% increase compared to the previous year[2]. Similarly, the 2017 Survey on the Human Rights of Children and Adolescent reported that Korean adolescents suffer from sleep deprivation and academic stress, with their overall happiness levels being the lowest among OECD and major European countries. Among various stress factors, academic stress particularly highlights the significant difficulties adolescents face in their studies[3]. School serves as a crucial environment for adolescents' psychological, social, and cognitive development, helping them form proper values and enhance their social adaptation skills through interpersonal interactions[4]. According to a study by the Korea Institute for Health and Social Affairs, the academic stress level of Korean adolescents is significantly higher than that of their peers in other countries, recorded at 50.5%, compared to the average of 33.3% across 30 surveyed countries[5]. Korean adolescents spend a significant amount of time in school, and much of their academic stress originates from school-related factors. Many reports indicate that schools account for a substantial portion of the stress experienced by students.

This study aims to develop a machine learning model for predicting adolescent academic stress and to analyze specific variables that may influence academic stress. In particular, this research focuses on constructing a predictive model that identifies the key causes of academic stress and assesses their impact, with the goal of accurately forecasting academic stress levels.

The primary objective of this study is to enhance the predictability of academic stress. While many existing studies have analyzed the causes and effects of academic stress, relatively few have proposed quantitative predictive models. Therefore, this research aims to build an academic stress prediction model using machine learning techniques, with the expectation that this model will enable more accurate predictions of students' academic stress levels.

Additionally, this study seeks to provide educational implications by leveraging machine learning techniques. Utilizing machine learning for predicting academic stress facilitates data-driven decision-making and effectively processes a wide range of variables. By applying a machine learning model, this research aims to predict students' academic stress and, based on the findings, derive effective strategies for stress management.

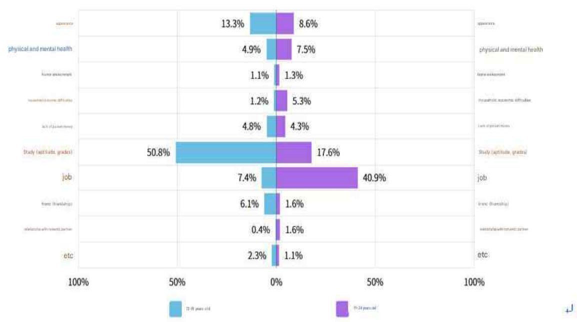
2. Literature Review

2.1 Academic Stress

Stress can be experienced severely due to sudden and unexpected life events. However, it can also have a positive effect by enhancing concentration and preventing tasks from stagnating[3]. Severe and chronic stress depletes mental and physical resources, leading to exhaustion(burnout). This occurs when individuals become excessively fixated on their goals and push themselves too hard, both physically and mentally. It is a condition in which anxiety and tension persist for several weeks or more. In contrast, general stress frequently arises in daily life due to both major and minor tensions or trivial events. Academic stress refers to the negative psychological states such as tension, anxiety, and worry that students experience in relation to grades and studying, among various stress factors encountered in school life[6].

Considering this, academics can be one of the major sources of stress for adolescents in South Korea. Previous studies have also found that Korean adolescents experience academic stress, and that the excessive academic burden is the primary cause of stress for students, aligning with these findings[7].

Especially during adolescence, academic stress has a significant impact on students' personal, social, and academic lives. The potential negative consequences of academic stress include emotional instability, depression, decreased self-esteem, and even physical illnesses, highlighting its seriousness. This type of stress is closely related to the psychological pressure that adolescents experience while engaging in academic activities, representing the mental burden that individual students bear in pursuit of academic achievement. According to a 2022 survey conducted by the Youth Policy Analysis and Evaluation Center targeting 200 male and female adolescents aged 13 to 18, the most concerning issue among teenagers was studies (aptitude and grades), accounting for 50.8%, the highest proportion Figure 1.



(Figure 1. Problems that teenagers are concerned about)
(Data: Youth Policy Analysis and Evaluation Center (by age as of 2022)

2.1.2 Review of Previous Studies

In this study, KISS, RISS, DBpia, and Google Scholar paper search were used as search sources. In previous studies related to academic stress, literature was searched using the keywords "adolescent academic stress," "adolescents," "adolescent stress," "adolescent mental health," and "mental health."

Literature related to previous studies based on the Adolescent Mental Health Survey was searched using the keywords "adolescent mental health," "mental health survey," "adolescent report," and "adolescent stress." For previous studies related to academic stress prediction models, literature was searched using the keywords "machine learning," "prediction model comparison," "machine learning," "random forest," "gradient boosting," and "logistic regression analysis." A search was conducted on adolescent-based topics, and previous studies on adolescent stress, academic-related research, and adolescent mental health were reviewed, referencing literature that addressed topics and issues related to adolescent stress.

Studies utilizing machine learning techniques remain relatively limited. Therefore, this study focuses on the development of a machine learning model for predicting academic stress, closely examining and referencing the arguments and content of previous studies.

Academic stress among adolescents in South Korea is known to be exceptionally high due to the highly competitive education system focused on college entrance exams. For example, 72.6% of South Korean adolescents identified academic issues as a source of stress, whereas in other countries, only 40–50% of adolescents cited academic problems as a cause of stress[4]. Using structural equation modeling(SEM), a study analyzed data from 2,444 respondents in the 5th-year panel dataset of the Korea Youth Panel Survey(KYPS) for the 4th-grade panel. The results indicated that academic stress and academic performance negatively impact psychological well-being[8]. Concerns and stress caused by declining academic performance can lead to severe frustration and feelings of inferiority, ultimately resulting in a decrease in academic efficiency. This, in turn, causes disappointment in oneself, leading to doubt about studying and a decline in motivation[10]. Additionally, as the intensity of academic stress increases, individuals become more psychologically unstable, leading to depression and even suicide. Given the severity of the problems caused by academic stress, research on this topic has been continuously conducted to date[11].

Interest in academic stress has been significantly increasing, establishing itself as a key research area for improving mental health during adolescence. In particular, understanding how academic performance, classroom concentration, and emotional symptoms such as anxiety and depression influence academic stress has emerged as a critical issue[6]. These factors are closely related to the stress adolescents experience at school. By clearly identifying the impact of each factor on academic stress, it is possible to propose better learning environments and mental health improvement strategies.

Therefore, academic performance significantly influences students' stress levels, and dissatisfaction with grades serves as a major cause of academic stress. Additionally, classroom concentration, which refers to students' ability to focus during lessons, directly affects the quality of learning and is also closely related to academic stress. Furthermore, emotional symptoms such as anxiety and depression experienced by adolescents have an even greater impact on academic stress. These factors often interact with one another, further exacerbating academic stress[12]. Therefore, studying how these various factors interact and contribute to academic stress among students is highly important. It has been stated that a significant portion of adolescent issues stem from stress, and a survey on stress experiences among middle and high school students revealed that academic performance issues were the most frequently reported cause of stress for both groups[9]. Academic-related stress has been classified into four categories: decline in academic performance, test anxiety, academic efficiency, and motivation for studying itself. Academic stress arises from the perception that academic activities are too difficult and do not bring happiness. It is a psychological state characterized by feelings of depression, discomfort, anxiety, worry, and mental distress, leading to an overall sense of unhappiness and unease[13].

The stress perception rate among South Korean adolescents is reported to be the highest compared to adolescents in other countries. According to statistics from the 20th(2024) Youth Health Behavior Survey, research findings indicate that stress perception rates increase from 9th grade (middle school) to 12th grade (high school). South Korean adolescents are known to experience the highest levels of academic-related stress, primarily due to a highly competitive, exam-focused education system that emphasizes tests, studying, and grades[14]. This aligns with the notion that stress in adolescent mental health is closely linked to academic-related factors.

3. Research Method

3.1 Study Participants

The Korea Youth Mental Health Survey is a national survey conducted by the Korea Institute for Youth Policy (a government-affiliated research institute under the Prime Minister's Office) to assess the mental health status of South Korean adolescents and utilize the findings for youth mental health policies. This 2021 dataset was collected through a web/mobile survey conducted nationwide with approximately 6,700 adolescents from elementary, middle, and high schools, as well as out-of-school institutions. The survey targeted adolescents who voluntarily agreed to participate, gathering data on their mental health status and policy needs.

Based on this dataset, this study aims to develop a

(Table 1. Question 7. Perceived Stress Scale Questionnaire)

division	There was none at all	There was almost none	There were times	It happened quite often	It happevery often
	0	1	2	3	4
1. How often have you felt upset because something unexpected happened?					
2. How often have you felt like you couldn't do important things well?					
3. How often have you felt anxious or stressed?					
4. How often have you felt confident in your ability to handle personal problems?					
5. How often have you felt that things went your way?					
6. How often have you felt like you couldn't handle everything you had to do?					
7. How often have you felt like you were handling everything well?					
8. How often have you been upset about something you couldn't do anything about?					
9. How often have you thought you were good at managing your time?					
10. How often have you felt like things were too difficult for you to overcome?					

predictive model for adolescent academic stress.

3.2 Measurement Tool

3.2.1 Data Preprocessing and Variable Setting

In this study, data preprocessing was performed on 313 variables collected from the 2021 Korea Youth Mental Health Survey dataset. The missing value analysis revealed that 61 out of the 313 variables contained missing data. Fifty-two variables with a missing rate exceeding 70% were removed, while the missing values in the remaining variables were imputed using the mean value. In machine learning, variables are often referred to as "features." Therefore, in this study, "variables" and "features" will be defined as having the same meaning[15].

In this study, SPSS (Statistical Package for the Social Sciences) sav format data was used. To enhance analysis efficiency, the data was converted into CSV format using the Pandas library. Additionally, variable names were modified into an interpretable format based on the information provided in the codebook.

3.2.2 Dependent Variable

The Korea Youth Mental Health dataset does not include a variable that directly measures academic stress. However, it was determined that academic stress could be estimated based on items related to "Question 7: Perceived Stress Scale." Based on this, 10 related variables from the scale were analyzed, and academic stress was selected as the dependent variable. The academic stress variable was created by calculating the average values of the related variables, and the detailed information is presented in Table 2.

Some items in the stress scale can have both positive and negative meanings, making it inappropriate to assign the same weight to each item. In the survey questions, some items interpreted "never" as a positive response, while others interpreted "very often" as a positive response.

As shown in Table 1, the items with a positive meaning include. Item 4: "How often have you felt confident in your ability to handle personal problems?" Item 5: "How often have you felt that things were going your way?" Item 7: "How often have you felt that you were coping well with everything?"

Item 9: "How often have you thought you were managing your time well?"

For these positively interpreted items, reverse coding was applied. Through this process, a high score (4) was converted to a low score (0), and a low score (0) was converted to a high score (4). After applying reverse coding, a new variable called "Academic Stress" was created. The stress score was determined by calculating the average value of the 10 items included in Question 7, and this score was set as the dependent variable.

3.2.3 Independent Variable

When performing machine learning modeling, there are advantages and disadvantages to using all variables versus selecting only certain variables. According to research, when the number of explanatory variables is large, including all variables in the model may lead to unexpected results that differ from the researcher's intentions[16]. When using all variables, the model can learn potential relationships between variables. However, if irrelevant variables or noise are included, the model may become overfitted to the training data, leading to poor performance on test data. On the other hand, using only a subset of variables has the advantage of simplifying the model and increasing training speed by selecting only important features. However, this approach also has a drawback, as it may omit some crucial information.

In this study, machine learning models such as Correlation, Linear Regression, and Random Forest were used to select meaningful variables. Important variables were extracted through each model, and among the duplicated variables, only those deemed significant were selected. As a result, a total of 225 independent variables were finalized, as shown in Table 2.

〈Table 2 Independent variable settings〉

Variable	value	datail
Question 1: related to ADHD CASS (27 questions)	0	Not at alL (Absolutely not)
	1	A little bit like that
	2	Quite so (It happens often)
	3	Very true (Very common)
Question 2: related to self-harm (5 questions)	1	Yes
	2	No
Question 3: related to child body sculpting (36 questions)	0	No experience
	1	A little difficult
	2	It's very difficult
	3	Very very difficult
Question 4: related to depression (12 questions)	0	Never
	1	A little bit like that
	2	Yes
	3	Quite so
	4	Very much so
Question 6: Suicidal Tendency	0	Never
	1	A little bit like that
	2	Yes
	3	Quite so
	4	Very much so
Question 11: Stress with people around you/ Whether or not you talk openly about your psychological discomfort	1	Not at all
	2	Not really
	3	It's like that
	4	Very much so
Question 12: The extent to which you discuss mental health issues with people around you (8 questions)	1	Not at all
	2	Not really
	3	It's like that
	4	Very much so
	5	No siblings

correlation may lead to the loss of crucial information.

(Table 2 Independent variable settings)

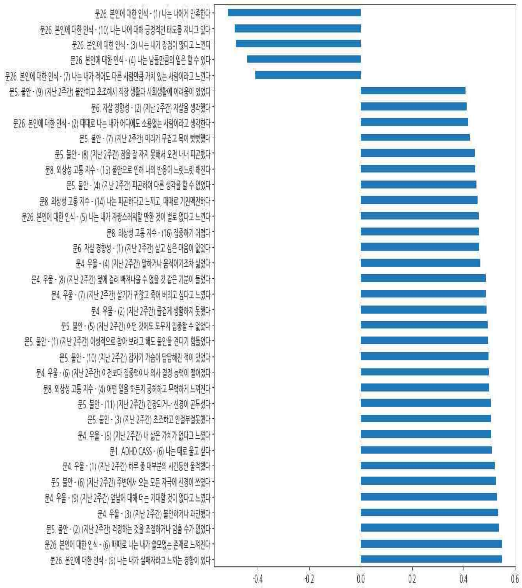
Variable	value	datail
Question 18: related to school life (24 questions)	1	Not at all
	2	Not really
	3	It's like that
	4	Very much so
Question 19: Academic performance	1	Very bad level
	2	Not good enough
	3	middle level
	4	Good level
Question 20: Satisfaction with academic performance	1	Not satisfied at all
	2	I'm not very satisfied
	3	I'm satisfied
	4	Very satisfied
Question 20: Regarding self-awareness (10 questions)	1	Not at all
	2	Not really
	3	It's like that
	4	Very much so

3.2.3.1 Correlation Variable Importance

Correlation is a concept that represents the degree of association between two variables and is commonly used to describe their relationship. In this study, the Pearson Correlation Coefficient, which is the most widely used method for measuring relationships, was utilized to assess the connection between academic stress and other variables.

The Pearson Correlation Coefficient measures the strength of the linear relationship between two numerical attributes. The coefficient ranges from -1 to +1. The Pearson correlation value (or coefficient) between two attributes is denoted by the letter r. If r = 0, it indicates that there is no correlation between the two attributes. If r = +1, it signifies a perfect positive correlation between the two attributes[17].

One advantage of correlation analysis is that it helps identify relationships between variables, allowing for the reduction of redundant information (multicollinearity). However, a drawback is that even if the correlation is low, a variable may still be important due to nonlinear relationships. Thus, excluding variables based solely on the items related to "Question 7: Perceived Stress Scale," which were used to derive the academic stress variable, were already included in its calculation. Therefore, the correlation analysis was conducted on the remaining variables. The analysis results identified 36 variables with correlation coefficients below -0.4 or above 0.4, as shown in Figure 2.

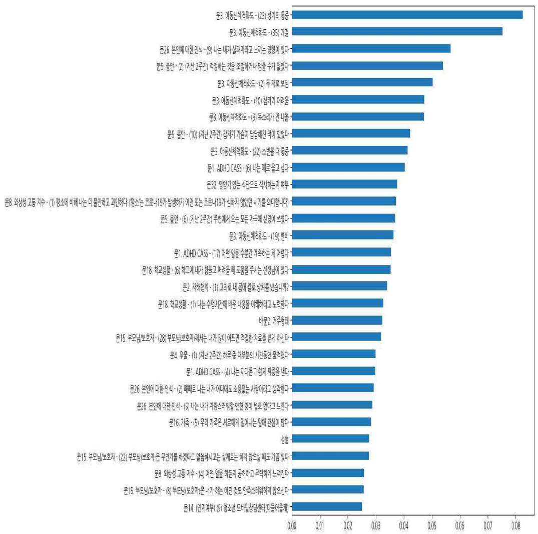


(Figure 2. Importance Criteria of 30 Variables in Correlation Analysis)

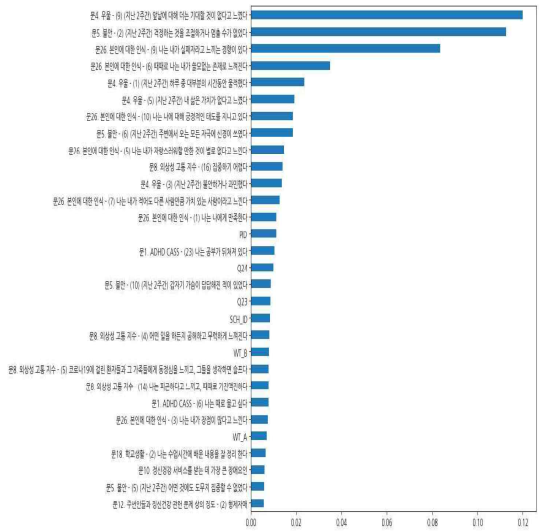
3.2.3.2 Linear Regression Variable Importance

The Linear Regression model has the advantage of intuitively identifying the contribution of each variable through regression coefficients. However, since Linear Regression cannot handle nonlinear relationships, it may have limitations when nonlinearity is an important factor.

In this study, to explain the independent variables affecting academic stress, a total of 251 variables (excluding 10 items related to "Question 7: Perceived Stress Scale" from the 261 preprocessed variables) were trained along with the academic stress variable. Subsequently, the importance of independent variables was evaluated based on their regression coefficients, and variables with significant influence were selected. The top 30 variables with the highest regression coefficients are presented in Figure 3.



(Figure 3. Importance Criteria of 30 Variables in Linear Regression)



(Figure 4. Importance Criteria of 30 Variables in
Random Forest)

3.2.3.3 Random Forest Variable Importance

Random Forest was utilized as a machine learning technique capable of detecting nonlinear relationships and assessing interaction effects between variables. This model has the advantage of effectively calculating variable importance and was deemed useful for capturing relationships not explained by Correlation and Linear Regression models.

Similar to the Linear Regression model, 251 variables (excluding 10 items related to "Question 7: Perceived Stress Scale" from the 261 preprocessed variables) were trained along with the academic stress variable to explain the independent variables affecting academic stress.

Subsequently, based on the variable importance scores generated by the Random Forest model, the top 30 variables with significant influence were selected, and the results are presented in Figure 4.

3.3 Analysis Methods and Procedures

All analyses in this study were conducted using Python in the Google Colab environment. The Pandas library was used for data processing and analysis, while the scikit-learn module was utilized for machine learning modeling to develop a predictive model for adolescent academic stress.

Machine learning, also known as automated learning, refers to the process of extracting knowledge from data [18]. It is particularly useful for problems that humans can solve but are too large in scale, as well as for problems that are mathematically definable but extremely complex, making them difficult to address[19].

Machine learning can be categorized based on the dependent variable of the data into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a learning method that utilizes given data and corresponding labels to make predictions and explore important variables in the process. In other words, the algorithm is provided with both inputs and expected outputs, and it learns to find a way to generate the desired output from the given input. Unsupervised learning, on the other hand, is a method of discovering useful patterns from data, where the algorithm is provided with input data but no corresponding output labels [18].

Reinforcement learning is a learning method in which a defined agent interacts with a specific environment, takes actions based on its current state, and analyzes the results to learn the optimal decision-making strategy that maximizes future rewards. Reinforcement learning involves an iterative process where a machine repeatedly makes choices and receives feedback, aiming to maximize the long-term rewards [20].

Predictions using supervised learning can be categorized into classification and regression, depending on the use of input and output. In this study, regression was chosen as the predictive approach to estimate the continuous values that influence academic stress. The classification of machine learning methods based on learning approaches is presented in Table 3[19].

(Table. 3 Classification of Machine Learning Based on Learning Methods)
Source: Lee Jung-won, 2022, Machine Learning-Based Big Data Analysis for Frost Occurrence Prediction Modeling

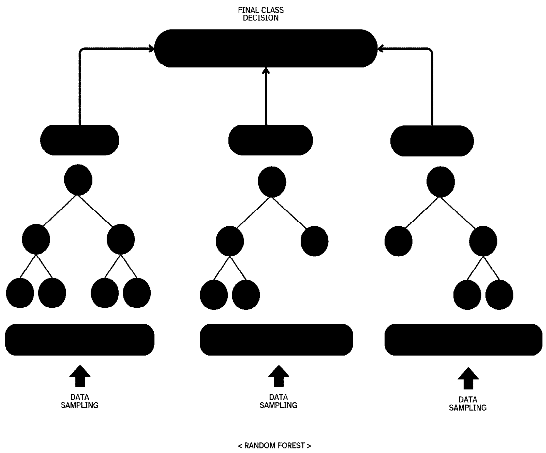
Category	Method	Prediction Approach	
Supervised Learning	Using Input and Output	Classification	Output is Yes/No
		Regression	Predicts continuous data such as weight, height, or numerical values (Value)
Unsupervised Learning	Using Input Only	Clustering	Uses clustering models such as k-means to predict output values based on input

In this study, the supervised learning method of machine learning was applied, and a regression algorithm was chosen to identify the key variables influencing stress. To train the model, machine learning analysis methods were selected based on previous studies and regression methodologies that demonstrated high performance. Compared to traditional statistical analysis methods, machine learning modeling exhibits higher predictive accuracy and enables the simultaneous analysis of hundreds of variables to explore factors influencing the dependent variable. Research in this area has been actively conducted[21]. This study aimed to apply various machine learning techniques to explore the factors influencing academic stress in adolescents. Among the selected five machine learning models, Linear Regression was used as the baseline for regression models.

To address cases where the data does not exhibit a linear relationship, Lasso and Ridge models were applied. Additionally, Random Forest and Gradient Boosting were chosen for comparison, as they differ significantly from linear models in terms of characteristics and structure. In this study, machine learning modeling was conducted using 6,689 data points and 225 features, excluding missing values from the analysis. The dataset was split into training data and test data, with 20% allocated for testing. To ensure reproducibility, the random_state was set to 42.

3.3.1 Random Forest

Random Forest is a model that constructs multiple decision trees and connects them, forming an ensemble model using the bagging method[22]. Since it combines multiple tree models, it generally exhibits improved predictive power and high stability. Since multiple tree models are combined, Random Forest generally exhibits enhanced predictive power and high stability. To address the limitations of individual decision trees, Random Forest employs the bagging method, which stands for Bootstrap Aggregation. Bagging involves bootstrap sampling from the entire dataset, splitting the data, training multiple models, and aggregating their predictions [15]. In this study, Random Forest was chosen because it effectively leverages the rich information contained in the dataset. Additionally, it allows for a comprehensive assessment of the influence of all predictor variables, making it suitable for identifying relative importance, discovering new predictive variables, and considering interactions between features. This capability makes Random Forest an effective tool for understanding the significance of different features.

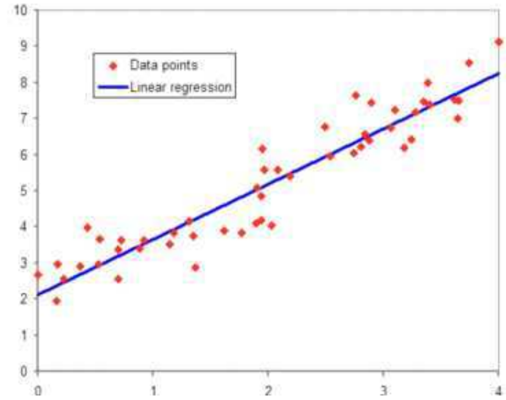


(Figure 5. Random Forest Model)
Source: Kwon Chul-min, Python Machine Learning Perfect Guide, Wikibooks.

3.3.2 Linear Regression

Among regression models, Linear Regression is the most fundamental and commonly used predictive model, representing a supervised learning algorithm. It is an analytical method that models the relationship between independent and dependent variables using a linear function that best represents their relationship[23].

Linear Regression is the most basic regression technique for modeling the linear relationship between a dependent variable and one or more independent variables. It assumes that the data is linearly distributed and expresses the influence of predictive variables on the target variable using a linear equation. However, if the data does not satisfy linearity, multicollinearity exists, or outliers are present, the model's performance may degrade. Therefore, preprocessing and the use of advanced techniques are crucial.

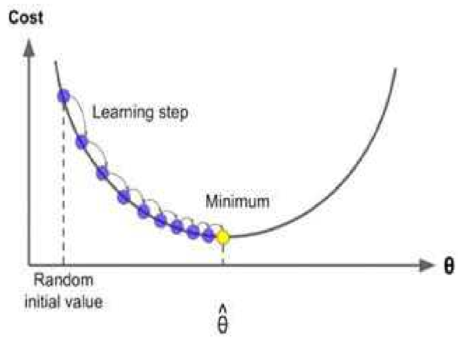


(Figure 6. Linear Regression Model)
Source: Lee Jun-ho, Shin Jae-woo (2024), Comparative Study of Machine Learning Models for Fine Dust Prediction, Journal of the Korea IT Policy & Management Society, 16(4), 3687-3693.

3.3.3 Gradient Boosting

Gradient Boosting is one of the ensemble techniques used in machine learning, where weak predictive models are sequentially combined and iteratively trained to ultimately build a strong predictive model. At each stage, learning progresses in a direction that minimizes errors to reduce the prediction errors of the previous model.

Unlike other boosting techniques, Gradient Boosting gradually reduces the errors of previous models using gradient descent as an optimization method[23].



(Figure 7. Gradient Boosting Model)
Source: Lee Jun-ho and Shin Jae-woo. (2024). Comparative Study of Machine Learning Models for Fine Dust Prediction. Journal of the Korea IT Policy & Management Society.

3.3.4 Ridge

Ridge regression uses the same prediction function as a linear regression model, but it is designed to impose an additional constraint when selecting weights (w). This constraint encourages the absolute values of the weights to be as small as possible, driving all weights closer to zero. As a result, Ridge regression reduces the magnitude of regression coefficients, lowering the model complexity and preventing overfitting.

The regularization method applied in Ridge regression is known as L2 regularization, which forces the model to be constrained, helping to prevent excessive fitting to the training data. L2 regularization introduces a penalty term based on the sum of the squared weights (w^2), effectively reducing the gradient and simplifying the model. This approach helps maintain predictive performance on training data while improving the generalization ability of the model to avoid overfitting[18].

(Table 4. Advantages and Disadvantages by Model)

Model	Advantages	Disadvantages
Random Forest	- Effectively models nonlinear relationships - Robust to outliers and noise - Provides variable importance - Low risk of overfitting	- Difficult to interpret - High computational cost - May require a large amount of data
Linear Regression	- Easy to interpret by understanding the relationship between variables through coefficients - Computationally efficient - Suitable for small datasets	- Susceptible to multicollinearity issues - Poor at modeling nonlinear relationships - Model performance is significantly affected if assumptions are not met
Gradient Boosting	- High predictive performance - Effectively models nonlinear relationships - Can significantly improve performance through boosting - Supports feature importance evaluation	- High computational cost and requires more resources - Prone to overfitting - Slower training speed
Ridge	- Addresses multicollinearity issues - Reduces overfitting - Considers feature importance	- All variable coefficients shrink toward zero, but none are completely eliminated, which may require additional feature selection
Lasso	- Enables feature selection - Enhances model interpretability - Reduces overfitting	- If there are many important features, performance may decrease due to excessive regularization

3.3.5 Lasso

To apply regularization to linear regression, Lasso is an alternative to Ridge. Both methods reduce model complexity by shrinking regression coefficients toward zero, but Lasso applies L1 regularization, producing different results from Ridge. Lasso regression can shrink some coefficients completely to zero, making the model more interpretable and revealing the most important features (Andreas Müller & Sarah Guido, 2017). In this study, Ridge and Lasso models were used to compensate for cases where the data did not exhibit a linear relationship in Linear Regression.

4. Research Results

4.1 Descriptive Statistics Analysis

An analysis was conducted on respondent characteristics, including gender, school level, multicultural family status, living arrangements, parental separation/divorce/death, economic status, father's education level, mother's education level, and family members living together. The results indicated that 5,937 students (88%) were enrolled in school, while 752 students (11%) were out-of-school youth. In terms of gender distribution, 3,104 students (46%) were male, and 3,585 students (53%) were female. Regarding school level, 2,039 students (30%) were in elementary school, 1,948 students (29%) were in middle school, and 1,950 students (29%) were in high school. Additionally, 752 students (0.3%) did not respond. For multicultural family status, 263 students (0.3%) belonged to multicultural families, while 6,426 students (96%) were from non-multicultural families.

Regarding living arrangements, the majority of students, 6,423 (96%), lived with their families. Meanwhile, 194 students (0.02%) resided in dormitories, boarding houses, or lived independently, and 72 students (0.01%) had other living arrangements.

In terms of economic status, 29 students (0.004%) reported being in the lowest economic category (very poor), while 168 students (0.025%) fell into the second-lowest category. Additionally, 516 students (0.07%) were in the third category, and 2,989 students (0.44%) considered themselves to be at an average economic level. The fifth category included 1,626 students (0.24%), while 899 students (0.13%) belonged to the sixth category. Finally, 462 students (0.69%) reported being in the highest economic category (very wealthy).

Regarding father's education level, 74 students (0.1%) reported that their father was not present, while 2 students indicated that their father did not attend school. Additionally, 28 students stated that their father completed elementary school, 73 students (0.1%) completed middle school, 1,271 students (19%) graduated from high school, and 3,173 students (47%) obtained a university degree. Furthermore, 634 students (0.9%) reported that their father completed graduate school, while 1,434 students (21%) were uncertain about their father's education level.

For mother's education level, 50 students reported that their mother was not present, and 11 students stated that their mother did not attend school. Additionally, 55 students indicated that their mother completed middle school, 1,443 students (21%) graduated from high school, and 456 students (0.6%) obtained a graduate degree. Furthermore, 1,353 students (20%) were uncertain about their mother's education level.

Regarding family members living together, 430 students (32%) lived with their grandfather or maternal grandfather, while 772 students (16%) lived with their grandmother or maternal grandmother. Additionally, 5,717 students (87%) resided with their father or stepfather, and 6,004 students (42%) lived with their mother or stepmother. Moreover, 5,195 students (0.5%) lived with their siblings, 246 students (32%) lived with relatives, and 77 students (41%) reported living with other family members. Lastly, 77 students (41%)stated that they lived alone. Refer to Table 5.

(Table 5. Descriptive Statistics of Demographic Factors)

Category		Number of Respondents (People)	Percentage (%)
Respondent Type	Students	5,937	88.7577
	Out-of-School Youth	752	11.2423
Gemder	Male	3,104	46.4045
	Female	3,585	53.5955
School Level	Elementary School	2,039	30.4829
	Middle School	1,948	29.1224
	High School	1,950	29.1523
	No Response	752	11.2423
Multicultural Family Status	Yes	263	3.9318
	No	6,426	96.0682
Living Arrangement	Living with Family	6,423	96.0233
	Dormitory/Boarding/Independent Living	194	2.9003
	Others	72	1.0764
Whether parents are separated, divorced or widowed	separation	188	2.8106
	Divorce	565	8.4467
	bereavement	110	1.6445
	Not applicable	5,826	87.0982
Household Economic Status	1(Very Poor)	29	0.4335
	2	168	2.5116
	3	516	7.7142
	4(Average Level)	2,989	44.6853
	5	1,626	24.3086
	6	899	13.44
	7(Average Level)	462	6.9069

(Table 5. Descriptive Statistics of Demographic Factors)

Category		Number of Respondents (People)	Percentage (%)
Father's Education Level	Father Absent	74	1.1063
	Did Not Attend School	2	0.0299
	Elementary School Graduate	28	0.4186
	Middle School Graduate	73	1.0913
	High School Graduate	1,271	19.0013
	University Graduate	3,173	47.4361
	Graduate School Graduate	634	9.4782
	Unknown	1,434	21.4382
Mother's Education Level	Mother Absent	50	0.7475
	Did Not Attend School	11	0.1644
	Elementary School Graduate	22	0.3289
	Middle School Graduate	55	0.8222
	High School Graduate	1,443	21.5727
	University Graduate	3,299	49.3198
	Graduate School Graduate	456	6.8172
	Unknown	1,353	20.2272
Family Members Living Together	Grandfather / Grandmother	430	2.322065
	Grandmother / Maternal Grandmother	772	4.168917
	Father / Stepfather	5,717	30.872664
	Mother/ Stepfather	6,004	32.422508
	Sibling	5,195	28.053786
	Relatives	246	1.328437
	Others	77	0.415812
	None	77	0.415812

4.2 Final Model Selection and Performance Comparison

4.2.1 Performance Evaluation of Regression Models(MAE, RMSE, R^2 Score)

There are various evaluation metrics used to assess regression performance, and in this study, MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 Score were utilized to measure how well the model explains and predicts actual data. These evaluation metrics primarily focus on the difference between actual values and predicted values in regression models. Simply summing the differences can cause positive and negative errors to cancel each other out, leading to misleading results. Therefore, metrics such as the mean of absolute errors, squared errors, or the square root of the mean squared errors are commonly used to provide a more reliable assessment[24].

MAE (Mean Absolute Error) is a metric used to evaluate the performance of a model in regression analysis. It calculates the absolute value of errors, allowing for an intuitive interpretation of how much, on average, the predicted values deviate from the actual values. A lower MAE value indicates better model performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Yi - \hat{Yi} |$$

$Yi : Actual\ Value,\ \hat{Yi} : \mathrm{Pr}edicted\ Value,$
$n : Nmber of Data$

MSE (Mean Squared Error) is a fundamental metric for evaluating the performance of a regression model. It calculates the average of the squared differences between the actual and predicted values. MSE is particularly important for models that focus on minimizing prediction errors

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Yi - \hat{Yi})^2$$

$Yi : Actual\ Value,\ \hat{Yi} : \mathrm{Pr}edicted\ Value,$
$n : Nmber of Data$

RMSE (Root Mean Squared Error) is a metric derived from MSE (Mean Squared Error). Since MSE squares the errors, it tends to produce values larger than the actual average error. RMSE is obtained by taking the square root of MSE, making it a more interpretable metric. A smaller value indicates that the model's predicted values are closer to the actual values, making it a useful metric for evaluating the accuracy and effectiveness of the model's predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}$$

$Yi : Actual\ Value,\ \hat{Yi} : Predicted\ Value,$
$n : Nmber of Data$

R^2 Score evaluates predictive performance based on variance, using the ratio of the variance of predicted values to the variance of actual values as a metric. A value closer to 1 indicates higher prediction accuracy

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Yi - \hat{Yi})^2}{\sum_{i=1}^{n}(Yi - \overline{Y})^2}$$

$\underline{Yi} : Actual\ Value,\ \hat{Yi} : Predicted\ Value,$
$\overline{Y}: Mean\ Value$

These metrics differ in the approach and perspective used to evaluate model performance, and the choice of metric may vary depending on the research objective and the characteristics of the data. First, R^2 Score is a metric that evaluates how well the model explains the data, making it useful for assessing the model's explanatory power.

RMSE (Root Mean Squared Error) measures the magnitude of prediction errors by squaring the differences between the actual and predicted values, averaging them, and then taking the square root. This is particularly suitable when prediction performance is crucial, especially in cases where emphasizing larger errors is important. On the other hand, MAE (Mean Absolute Error) is calculated by averaging the absolute values of prediction errors. Since all errors are treated equally, it provides an intuitive assessment of overall prediction accuracy. In this study, survey data was used, and since survey responses are based on individuals' subjective evaluations, there may be significant variability between respondents and a lower consistency in responses. Due to these characteristics, data volatility may increase, and R^2 Score may not fully capture this variability. On the other hand, RMSE and MAE are metrics that directly measure the difference between predicted and actual values, making them more suitable for evaluating prediction accuracy in survey data, which is based on subjective responses. In particular, RMSE is sensitive to large prediction errors, making it advantageous when emphasizing the impact of significant errors that may occur in survey responses on model performance. MAE, on the other hand, is not overly sensitive to large errors while still effectively evaluating overall prediction accuracy.

However, given the nature of survey response data, where large errors can have a significant impact on model evaluation, this study selected RMSE as the primary performance evaluation metric.

To develop a machine learning model for predicting adolescent academic stress, five different models were trained to select the optimal model. The five machine learning algorithms compared in this study were Linear Regression, Random Forest, Ridge, Lasso, and Gradient Boosting. For modeling, the scikit-learn library was utilized. A total of 6,689 data points were trained using the 225 selected feature variables from Table 5, along with the stress score generated based on the Perceived Stress Scale items, which was set as the target variable. The performance evaluation results of the five regression models applied for predicting academic stress are presented in Table 6. The evaluation metrics used for analysis included MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 Score

4.2.1.1 Performance Evaluation of Linear Regression

The Linear Regression model achieved a Mean Absolute Error (MAE) of 0.31611, a Root Mean Squared Error (RMSE) of 0.39428, and an R^2 Score of 0.59685.

While the MAE and RMSE values indicate reasonable performance, the R^2 Score of 0.59685 suggests that the model can explain approximately 59.7% of the variance in the data.

4.2.1.2 Performance Evaluation of Ridge

The Ridge Regression model achieved a Mean Absolute Error (MAE) of 0.31608, a Root Mean Squared Error (RMSE) of 0.39425, and an R^2 Score of 0.59692.

While the results are similar to those of Linear Regression, the MAE and RMSE values are slightly lower, and the R^2 Score is marginally higher at 0.59692.

By applying regularization, the Ridge model helps prevent overfitting, ensuring that the model does not become excessively complex. This allows it to fit well to the training data while maintaining generalization performance on test data, thereby reducing the risk of performance degradation due to overfitting.

4.2.1.3 Performance Evaluation of Lasso

The Lasso Regression model recorded a Mean Absolute Error (MAE) of 0.46708, a Root Mean Squared Error (RMSE) of 0.62161, and an R^2 Score of -0.00204.

Among all models, Lasso showed the lowest performance, with significantly higher MAE and RMSE values compared to the other models. These results indicate that Lasso Regression is not well-suited for predicting academic stress, as its predictive accuracy is notably lower than that of the other models. Notably, the R^2 Score is negative, indicating that the model fails to explain the data at all.

Lasso regression has a characteristic tendency to simplify the model by reducing the coefficients of less important variables to zero. However, since academic stress data involves complex interactions among multiple variables, the Lasso model may have excessively simplified the data, leading to the loss of critical information.

4.2.1.4 Performance Evaluation of Random Forest

Random Forest is an ensemble method capable of effectively handling nonlinear data and interactions between variables, demonstrating high performance across all metrics.

The model recorded the lowest values for both MAE (0.31075) and RMSE (0.39021), indicating superior predictive accuracy. Additionally, the R^2 Score was 0.60513, reflecting a high level of explanatory power for the data

4.2.1.5 Performance Evaluation of Gradient Boosting

Gradient Boosting demonstrated high performance, similar to Random Forest, and notably achieved the highest R^2 Score of 0.60623 among all models. This indicates that the model effectively learned the nonlinear relationships between academic stress and independent variables.

MAE (Mean Absolute Error), which represents the average difference between predicted and actual values, recorded the highest value among all models, highlighting the model's sensitivity to prediction errors. RMSE (Root Mean Squared Error) represents the mean squared error and serves as a metric that indicates how sensitive the model is to large errors. A lower RMSE value signifies a higher-performing model. Additionally, both MAE and RMSE values were at a similar level to those of Random Forest, suggesting that Gradient Boosting is one of the most suitable models for predicting academic stress.

(Table 6: Comparison of Regression Model Performance)

Model	MAE	RMSE	R^2 Score
Linear Regression	0.31611	0.39428	0.59685
Ridge	0.31608	0.39425	0.59692
Lasso	0.46708	0.62161	-0.00204
Random Forest	0.31075	0.39021	0.60513
Gradient Boosting	0.31179	0.38967	0.60623

Based on MAE, the Random Forest algorithm demonstrated the best performance with a value of 0.31075. For RMSE, the Gradient Boosting algorithm achieved the highest performance with a value of 0.38967. Regarding R^2 Score, the Gradient Boosting algorithm also showed the best performance, recording 0.60623.

To optimize the performance of each predictive model, hyperparameter tuning was performed using the GridSearchCV API. GridSearchCV is a technique designed to automatically search for the optimal hyperparameters to enhance the performance of machine learning models. Since model training performance can vary depending on hyperparameters (e.g., learning rate, number of trees, etc.), it is essential to fine-tune them appropriately. GridSearchCV systematically tests the user-specified hyperparameter combinations, evaluates their performance through cross-validation, and identifies the best-performing configuration. In this study, GridSearchCV was applied to each algorithm to search for the optimal hyperparameters. The models were then retrained based on these optimal parameters, and their performance was compared. The optimization results and the performance comparison of each model are presented in Table 7.

Based on MAE, the Random Forest algorithm demonstrated the best performance with a value of 0.31028.

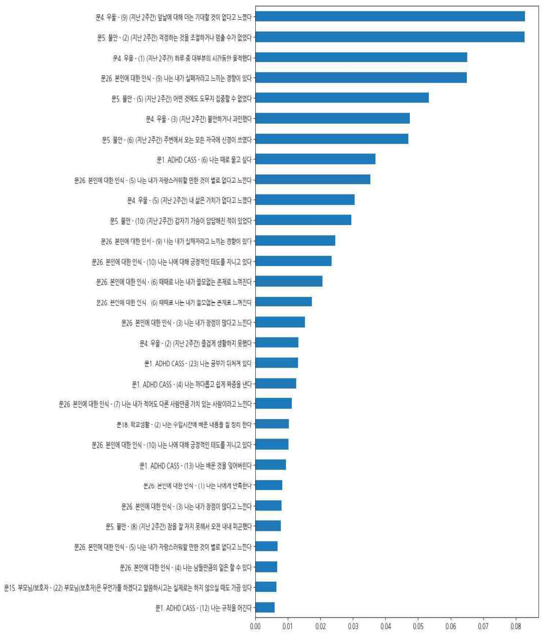
For RMSE, the Gradient Boosting algorithm achieved the highest performance with a value of 0.38770. Regarding R^2 Score, the Gradient Boosting algorithm also showed the best performance, recording 0.61019. Furthermore, it was confirmed that the R^2 Score of the Gradient Boosting model improved to 0.61019 due to hyperparameter tuning, enhancing its predictive performance.

(Table 7. Performance Comparison of Regression Models with Applied Hyperparameters)

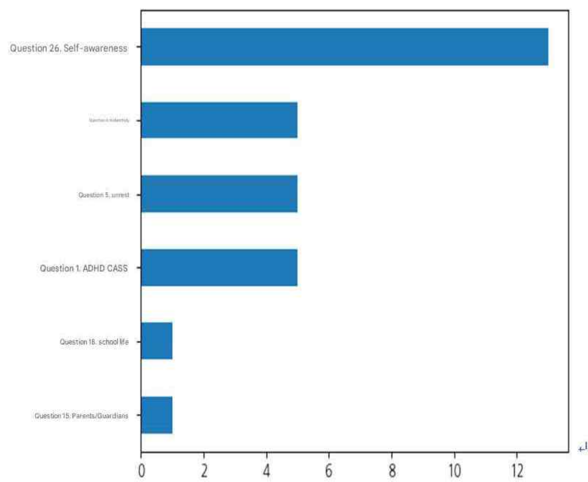
Model	Hyper parameter	MAE	RMSE	R^2 Score
Linear Regression		0.31611	0.39428	0.59685
Ridge	alpha=100	0.31447	0.39232	0.60085
Lasso	alpha=0.1	0.34060	0.44080	0.49612
Random Forest	n_estimators=200, random_state=42, max_depth=20, min_samples_split=2	0.31028	0.38928	0.60701
Gradient Boosting	n_estimators=200, random_state=42, learning_rate=0.1, max_depth=3	0.31058	0.38770	0.61019

4.2.2 Feature Importance Analysis of Gradient Boosting Predictive Model

In this study, the Gradient Boosting model, which showed the best performance based on the R^2 Score, was analyzed to identify the most influential features in predicting adolescent academic stress. Key variables included statements such as "I felt that there was nothing to look forward to in the future" and "I was unable to control or stop worrying." The results are presented in Figure 8. Out of the 225 feature variables, the top 36 variables were selected based on their importance scores in the Gradient Boosting model, and the findings were visualized in a graph.



(Figure 8. Top 36 Variables Based on Importance Scores in the Gradient Boosting Prediction Model)



(Figure 9. Frequency of Major Categories for Survey Items)

The frequency distribution of each survey item by major category is shown in Figure 9. The graph indicates that self-perception-related variables have the greatest impact on stress levels. Following this, the categories ranked in order of influence are depression, anxiety, ADHD CASS, school life, and parental/guardian factors.

5. Discussion and Conclusion

5.1 Summary and Discussion

The research was conducted on approximately 6,700 middle and high school students, including out-of-school youth, across South Korea in 2021. This study aimed to develop a predictive model for adolescent academic stress using Linear Regression, Random Forest, Ridge, Lasso, and Gradient Boosting models. To address the research question, machine learning models were developed and performance metrics were derived using Python-based machine learning libraries. Specifically, algorithms from the scikit-learn library were applied, including Logistic Regression from the linear model module, Gradient Boosting Classifier from the Gradient Boosting module, and Decision Classifier and Random Forest Classifier from the sklearn ensemble module.

The performance evaluation metrics for the models were based on MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 Score, which are commonly used in regression analysis. Among the models, Random Forest achieved the lowest MAE with a value of 0.31075, followed by Gradient Boosting (0.31179), Ridge (0.31608), Linear Regression (0.31611), and Lasso (0.46708). This ranking indicates that Random Forest and Gradient Boosting performed the best in terms of minimizing absolute errors, while Lasso exhibited the highest error, suggesting weaker predictive performance. The RMSE (Root Mean Squared Error) values for the models, in ascending order, were as follows: Gradient Boosting (0.38967), Random Forest (0.39021), Ridge (0.39425), Linear Regression (0.39428), and Lasso (0.62161). Similarly, the R^2 Score values, ranked in descending order, were Gradient Boosting (0.60623), Random Forest (0.60513), Ridge (0.59692), Linear Regression (0.59685), and Lasso (-0.00204). These results indicate that Gradient Boosting had the lowest RMSE, making it the best performer in terms of error minimization, while Lasso exhibited the highest RMSE and the lowest R^2 Score, suggesting poor predictive accuracy. Based on the RMSE evaluation metric, the Gradient Boosting model recorded the lowest RMSE value, demonstrating superior predictive performance. Additionally, even after applying the optimal hyperparameters, the Gradient Boosting model maintained consistent performance, showing more stable results compared to other models. Based on these findings, the Gradient Boosting model was ultimately selected as the most suitable model for predicting adolescent academic stress.

Analysis of Key Factors Influencing Adolescent Academic Stress Using the Gradient Boosting Model. The analysis of key factors influencing adolescent academic stress using the Gradient Boosting model revealed that the most significant variables were: Depression (Question 4-1): "Over the past two weeks, I felt down for most of the day." Anxiety (Question 5-2): "Over the past two weeks, I was unable to control or stop worrying." Depression (Question 4-9): "Over the past two weeks, I felt that there was nothing to look forward to in the future." These findings indicate that depression and anxiety are major contributors to adolescent academic stress.

Most previous studies on adolescent stress have employed statistical methods, primarily conducting descriptive statistical analyses on stress-related factors. In contrast, research using machine learning models to predict academic stress remains limited. Traditional statistical approaches focus on explaining data, emphasizing statistical significance testing and model-based inference to describe relationships within the dataset. However, machine learning techniques prioritize predicting dependent variables when new data is introduced rather than focusing solely on explanation[25].

In the study by [26], which applied machine learning techniques to explore influential variables, a greater variety of new factors were identified compared to studies using traditional research methods[27].

5.2 Conclusion and Recommendations

This study holds significance in that it utilizes various machine learning techniques to identify key predictive variables for adolescent academic stress and, based on these findings, proposes academic stress management strategies, providing valuable direction for future researchers. In particular, the core contribution of this study lies in identifying the major factors influencing adolescent academic stress using the most predictive model and suggesting ways to leverage these findings from a preventive perspective. These findings open up possibilities for preventing risk factors associated with adolescent academic stress and further serving as foundational data for the development of comprehensive mental health management programs. In particular, this study makes a practical contribution by establishing a foundation for designing models that can be utilized not only for predicting and preventing academic stress but also for addressing other aspects of mental health. However, this study has several limitations that should be acknowledged.

First, this study relied on data collected at a single point in time, which presents a limitation in understanding how academic stress evolves over time and how various factors interact to influence stress levels. It remains unclear whether stress increased due to a specific event or whether stress itself led to other issues. To accurately determine these relationships, longitudinal studies that collect data over an extended period are necessary.

Second, the data used in this study is based on self-reported survey responses, which may introduce limitations in reliability and accuracy due to the subjective nature of the responses. To address this limitation, future research should incorporate diverse data types, such as biometric signals or behavioral data, enabling more objective and sophisticated analyses.

Third, the survey items used in this study may lack precision in design and evaluation, highlighting the need for new assessment methods that take respondents' psychological states into account. Additionally, since the machine learning models in this study were trained and analyzed on a limited dataset, future research should focus on expanding the dataset and applying various advanced algorithms, such as deep learning and ensemble techniques, to compare and improve predictive performance.

Despite these limitations, this study has introduced new possibilities in the field of adolescent academic stress research. It is expected that future studies will build upon these findings to conduct more in-depth and comprehensive analyses, further advancing the understanding and management of academic stress.

참고문헌(Reference)

[1] Cha Son. Min, "Korean Youth's Life Satisfaction 'Lowest' vs Academic Performance 'Top'", 2024.
https://www.newsis.com/view/NISX20240105_0002582920

[2] Min ju. Cha, "Korean Students Stuck In 'Study Hell', 41% Are 'Stressed'", 2023.
https://www.kmib.co.kr/article/view.asp?arcid=0924304673

[3] Hee jin. Han, "Factors affecting academic stress perceived by high school students. Personality education," Vol.16, No.3, pp.103-117, 2022.

[4] JH Kim, SB Kim, JI Jeong, "The Effects of Adolescents' Academic Stress on School Life Adaptation: Focusing on the Mediating Effects of Stress Coping," The Korean Youth Study, Vol.25, No.4, pp.241-269, 2014.

[5] SC Shin, SY Chol, "Academic stress perceived by adolescents to adapt to school life." The moderating effect of self-esteem on the effect", counseling psychological education welfare, Vol.6, No.3, pp.59-72, 2019.

[6] JI Jang, SB Kim, "The mediating effect of positive psychological capital and smartphone addiction in the relationship between adolescents' stress and school life adaptation." Youth Counseling Study, Vol.23, No.2, pp.447-465, 2015.

[7] DW Kim, HY Park, KI Lee, "The Effects of Parents' Academic Support and Emotional Support on Academic Stress and the Controlling Effect of Intrinsic Motivation," Human Development Research, Vol. 29 No.1, pp.21-41, 2022.

[8] CR Noh, SH Kim, "The Effects of Academic Stress and Academic Performance of Middle School Students on Psychological Well-being," Korean Department of Child Welfare, 39, 2021.

[9] KH Lee, "Study on the Stress of Adolescents", Ph.D. thesis, Sookmyung Women's University, 1995.

[10] GY Kim, "School life and counseling guidance", 560 Hakjisa, 2000.

[11] EH Choi, "The effect of academic stress in high school students on school life adaptation. Mediating effects of psychological state and moderating effects of parenting attitudes," 120, master's thesis, Daegu Oriental Medical University, 2016.

[12] HK Sang, YN Lee, "The relationship between academic achievement and academic stress: the moderating effect of academic self-concept," Educational Research, Vol.35, No.1, pp.53-71, 2012.

[13] MH Oh, SM Chun, "The Effects of Meditation Training to Analysis of Academic Stress Factors and Symptoms in Middle School Students and their Reduction", 78, master's thesis, Kyungpook National University, 1994.

[14] SM Mo, "Analyzing the moderating effects of adolescent academic stress-inducing variables: Focusing on self-esteem and decision-making autonomy", Journal of the Future Youth Society, Vol.7, No.2, pp.49-66, 2010.

[15] SM Kwon, "Study on the Predictive Model of Credit Ratings for Logistics Companies Using Machine Learning", 173, PhD thesis, Chung-Ang University, 2023.

[16] HS Seo, "Statistical Machine Learning for the Selection of Linear Models: A Comparative Study of Variable Selection Methods", Journal of The Korean Data Analysis Society, Vol.23, No.6, pp.2643-2653, 2021.

[17] John Kelleher, Brendan Tierney, OS Kwon, "Data Science as a Tool for Insight for Better Decision Making," 272, Kim Young-sa, 2019.

[18] A Muller, S Guido, "Introduction to Machine Learning with Python," 504, Hanbit Media, 2017.

[19] JW Lee, "Machine Learning-Based Big Data Analysis for Frost Occurrence Prediction Modeling," 113, Hanbat National University, 2022.

[20] SY Kim, YJ Jeong, TH Kim, "First Steps in Machine Learning: From Basics to Modeling, Practical Examples, and Problem Solving," 376, Hanbit Media, 2017.

[21] YH Son, HJ Park, MH Park, "Analysis of Group Determinants Based on Reading Literacy Levels Using Random Forest: Focusing on PISA 2018 Data," Asian Journal of Education Research, Vol.21, No.1, pp. 191-215, 2020.

[22] Breimanm L, "Random forests, Machine Learning", Vol.45, No.1, pp.5-32, 2020.

[23] JH Lee, JW Shin, "Comparative Study of Machine Learning Models for Fine Dust Prediction," Journal of the Korea IT Policy & Management Society, Vol.16, No.4, pp.3687-3693, 2024.

[24] CM Kwon, "Python Machine Learning Perfect Guide," 704, Wikibooks, 2022.

[25] HW Kim, JE Yoo, "Exploring Depression Prediction Variables for Second-Year Middle School Students Using Machine Learning," KCYPS 2018 5th Wave Data Analysis, Vol.25, No.3, pp.31-53, 2024.

[26] MJ Noh, JE Yoo, "Exploring Career Decision-Related Variables Using Adaptive LASSO," Open Education Research, Vol.27, No.4, pp.133-155, 2019.

[27] SY Park, JY Jung, "Analysis of Predictive Factors for High School Teachers' Instructional Improvement Activities Using Random Forest," Educational Research, Vol.62, No.5, pp.1-32, 2020.

◐ Author Biography ◑

Jeongsuk-suk So
Graduated from Seoul School of Integrated Sciencs & Technologies(aSSIST), Graduate School of AI, Department of AI Advanced Studies, Major in AI Big Data(Master of Engineering)
Research Interests : AI, Digital Transformation, Generative AI, Data Analysis etc.
E-mail : sojeongsuk@gmail.com

Tae-yeon Oh
Assistant Professor at Seoul School of Intergrated Sciences & Technologies(aSSIST)
Assistant Professor, University of Mississippi(2018-2022)
Ph.D. in Sports Management, Seoul National University
Research Interests : AI, Sports Analytics, Sports Marketing, Sports Media, Big Data Analytics etc.
E-mail : tyoh@assist.ac.kr