ISSN: 2508-7894
KJAI website: http://acoms.kisti.re.kr/kjai
doi: http://dx.doi.org/10.24225/kjai.2023.11.3.23

Analysis of Adolescent Suicide Factors based on Random Forest Machine Learning Algorithm

Gi-Lim HA[1], In Seon EO[2], Dong Hun HAN[3], Min Soo KANG[4]

Received: January 20, 2023. Revised: March 20, 2023. Accepted: September 05, 2023.

Abstract

The purpose of this study is to identify and analyze suicide factors of adolescents using the Random Forest algorithm. According to statistics on the cause of death by the National Statistical Office in 2019, suicide was the highest cause of death in the 10-19 age group, which is a major social problem. Using machine learning algorithms, research can predict whether individual adolescents think of suicide without investigating suicidal ideation and can contribute to protecting adolescents and analyzing factors that affect suicide, establishing effective intervention measures. As a result of predicting with the random forest algorithm, it can be said that the possibility of identifying and predicting suicide factors of adolescents was confirmed. To increase the accuracy of the results, continuous research on the factors that induce youth suicide is necessary.

Keywords : Suicide Factors, Adolescent, Random Forest Algorithm, Machine learning

Major Classification Code : Artificial Intelligence

1. Introduction

1.1. Background and Purpose of Research

This study finds factors that influence adolescents' thoughts about suicide. Among the various factors of suicidal ideation, the characteristics of factors affecting adolescents will be identified. Based on this, it is intended to help reduce adolescents' suicidal ideation and provide data for suicide prevention. Since suicidal thoughts can only be identified as a subjective individual's experience, asking these questions to emotionally sensitive adolescents may increase their suicidal ideation and may be at high risk depending on any factor or situation. Therefore, it is necessary to consider asking and investigating adolescents about the experience, status, and degree of suicidal ideation. Suicide has a great impact and ripple effect not only on individual families but also on our society as a whole (Korea National Statistical Office, 2021). According to the National Statistical Office's announcement, 26.0 people per 100,000 people aged 0 to 24 committed suicide in 2021. In addition, according to the "2021 Youth Statistics" released by the National Statistical Office and the Ministry of Gender Equality and Family, suicide was the highest cause of death per 100,000 teenagers between 9 and 24 years old

* This work was supported by the research grant of the KODISA Scholarship Foundation in 2023.

1 First Author. Dept. of Medical IT, Eulji University, Korea. Email: rlfla00@g.eulji.ac.kr
2 Second Author. Cloud Manager, Sysone, Korea. Email: eis@sysone.co.kr
3 Third Author. Dept. of Medical Intelligence, Eulji University, Korea. Email: d555v@naver.com
4 Corresponding Author. Professor, Dept. of Medical IT, Eulji University, Korea. Email: mskang@eulji.ac.kr

as of 2019. In adolescence, physical, emotional, and social changes are rapid, so it is stressful and psychologically and emotionally unstable according to the changes. Adolescents are more vulnerable than other age groups and are known to easily commit impulsive behaviors such as suicide compared to other age groups due to their developmental characteristics.

However, research is needed to understand adolescents' suicidal ideation because they are intertwined with complex relationships, including social factors such as family and school, as well as individual emotional and psychological problems. Furthermore, the purpose is to use machine learning algorithms to predict whether individual adolescents are thinking about suicide without investigating suicide ideation. These studies can contribute to establishing effective intervention measures while protecting adolescents.

Figure 1: Korea National Statistical - suicide rate (2017~2021)

(단위 : 인구 10만 명당 명, %)

	0세	1-9세	10-19세	20-29세	30-39세	40-49세	50-59세	60-69세	70-79세	80세이상
1위	출생전후기에 기원한 특정 병태	악성신생물	고의적 자해(자살)	고의적 자해(자살)	고의적 자해(자살)	악성신생물	악성신생물	악성신생물	악성신생물	악성신생물
	136.8	1.8	5.9	19.2	26.9	41.1	119.6	281.4	695.0	1402.6
	(51.0%)	(17.5%)	(37.5%)	(51.0%)	(38.0%)	(26.7%)	(37.3%)	(43.1%)	(35.7%)	(17.9%)
2위	선천 기형, 변형 및 염색체 이상	운수사고	악성신생물	악성신생물	악성신생물	고의적 자해(자살)	고의적 자해(자살)	심장 질환	심장 질환	심장 질환
	45.4	1.1	2.2	4.2	13.0	31.0	33.3	57.0	197.1	972.2
	(16.9%)	(10.2%)	(13.7%)	(11.1%)	(18.9%)	(21.7%)	(10.4%)	(8.7%)	(10.1%)	(12.4%)
3위	영아 돌연사 증후군	기왜(타살)	운수사고	운수사고	심장질환	간 질환	심장 질환	뇌혈관 질환	뇌혈관 질환	폐렴
	17.0	0.9	1.8	3.7	3.9	10.7	25.4	40.4	152.3	918.6
	(6.3%)	(8.8%)	(11.3%)	(9.9%)	(5.7%)	(7.5%)	(7.9%)	(6.2%)	(7.8%)	(11.7%)
4위	심장 질환	선천기형변형염색체이상	심장 질환	심장 질환	운수사고	심장 질환	간 질환	고의적 자해(자살)	폐렴	뇌혈관 질환
	5.2	0.7	0.7	1.4	3.8	10.3	23.4	33.7	137.2	636.1
	(1.9%)	(6.6%)	(4.7%)	(3.6%)	(5.5%)	(7.2%)	(7.3%)	(5.2%)	(7.0%)	(8.1%)
5위	기왜(타살)	심장 질환	익사 사고	뇌혈관 질환	간 질환	뇌혈관 질환	뇌혈관 질환	간 질환	당뇨병	알츠하이머병
	4.9	0.6	0.4	0.5	3.1	8.2	19.0	23.5	63.8	325.7
	(1.8%)	(5.8%)	(2.8%)	(1.4%)	(4.5%)	(5.7%)	(5.9%)	(3.6%)	(3.3%)	(4.2%)

Figure 2: Cause of Death Statistics Results

According to statistics on the cause of death by the National Statistical Office in 2019, intentional self-harm (suicide) was the highest cause of death among teenagers aged 10 to 19. As shown in the figure, suicide was the number one cause of death between the ages of 10 and 39, and unlike other age groups, the rate of suicide was high. In addition, except 2014, it has been found that youth suicide has become a steady problem as it has been maintained as the No. 1 cause of youth death in the past 10 years, indicating that attention should be paid to it.

1.2. Research Method

This study aims to provide data that can predict whether adolescents think of suicide without investigating suicide ideation using machine learning techniques, and to identify factors that affect suicide ideation. Machine learning techniques, a method for analyzing big data, learn by finding specific rules from large amounts of data. Since the process called learning is data analysis, the model derived through this analysis is evaluated by adapting it to data that was not included in the model establishment. Once the performance is recognized, the derived model can be used to predict new data. The random forest algorithm, which generates multiple trees and selects the most voted tree results, reflects randomness and interaction with variables, so it does not cause errors even in high-dimensional models, making it suitable for studies where many factors can be involved in addition to the specific factors set.

1.3. Machine learning - Random forest

The Random Forest algorithm is an ensemble technique, and the basic component is a decision tree algorithm. The Random Forest algorithm, an ensemble technique that collects multiple results to create one result, can maximize the performance of the algorithm by collecting multiple results from the decision tree to create one result. The decision tree algorithm is an analysis method that performs classification and prediction by charting decision rules into a tree structure. The division criteria are divided based on the attributes of the data, and are divided into trees according to the division criteria and modeled. The interpretation is simple as the data analysis results can be initially confirmed through the appearance of the decision tree. The first question, the first part of the classification, is called the Root Node, and the last node output is called the Terminal Node.

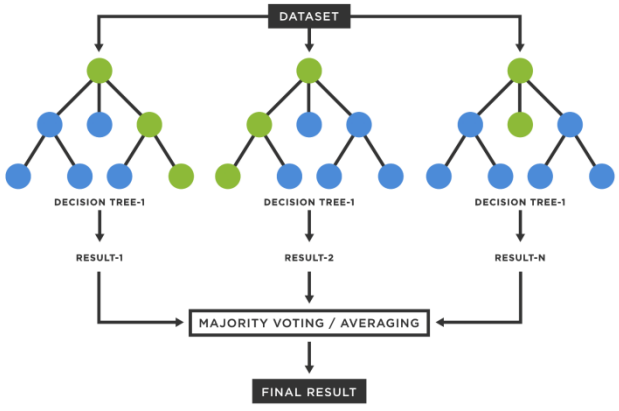


Figure 3: Random forest example

Decision trees have the advantage of being intuitively easy to understand, but they tend to be less predictive because only one specific factor is considered when branches are divided. In addition, even a slight change in data can change the composition of the tree. In other words, the bias is relatively low, but there is a disadvantage that it is difficult to generalize the model because it has a high variance error. To compensate for this, the Random Forest algorithm goes through the process of forming a forest with a number of randomly sampled decision trees when analyzing and aggregating several decision trees and making final predictions. Random Forest can reduce prediction errors by giving maximum randomness in sample selection and variable selection for each model. Previous studies also show that the Random Forest algorithm does not cause errors even in high-dimensional models by reflecting randomness and interaction with variables (Géron, 2017).

2. Research Methods and Materials

2.1. Research data

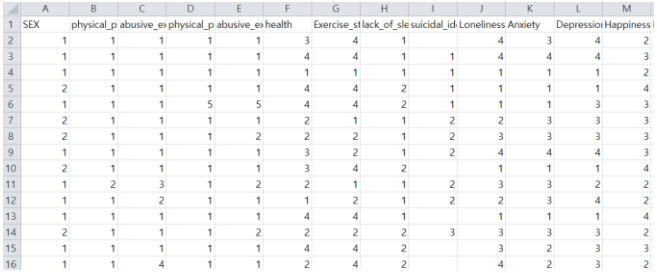
Table 1: Data

	Content
population	elementary, middle and high school students
sampling frame	2018 Annual Report of Education Statistics
sample count	9000 people sampling -> 9270 people final survey completed
Number of extracted schools	362 schools
time of investigation	From May to July 2019
investigation method	Interviews, Self-Entry Survey Respons

This study used data from the 2019 Human Rights Survey conducted by the Korea Youth Policy Institute to explore the predictors of adolescents' suicidal ideation, and analyzed suicide factors and suicide rates using Kaggle's Dataset. The 2019 Children and Youth Human Rights Survey was used by the Korea Youth Policy Institute to understand the overall situation of the human rights of children and adolescents in Korea and to come up with measures to protect and promote human rights. 9,000 students in the 4th, 5th, and 6th grades of elementary school, 1st, 2nd, and 3rd grades of middle school, and 1st, 2nd, and 3rd grades of high school nationwide were surveyed, and finally 9.270 students were surveyed.

2.2. Measuring Tools

The purpose of this report's empirical analysis is to distinguish between adolescents who think about suicide and adolescents who don't, and to identify factors that affect the prediction of suicide ideation. The data were analyzed using Machine learning studio Azure. Since the random forest algorithm allows analysis consisting of a number of explanatory variables, it was intended to understand the factors in detail by using the detailed items of the human rights survey without reducing them separately. The categories of school violence and cyber violence include abuse and insult, assault, bullying, extortion, intimidation, and sexual harassment, and the categories of parent and teacher violence include corporal punishment, abuse and insult. The eight discriminations categories include gender, academic, or family circumstances, residential areas, appearance, religion, and absence of parents, and the categories of depression include loneliness, anxiety, sadness, and depression. Self-esteem is divided into valuable people, good qualities, pride, and positive attitude. Other factors include academic, family discord, peer relationships, economic difficulties, appearance, future career and academic performance, economic conditions, health status, lack of sleep, sexual violence, and the presence of people to discuss. The answer to the question "Have you ever thought about dying in the last year?" was divided into three types: "I have never thought about it," "I think sometimes," and "I think often," and the answer except "I have never thought about it" was classified as having thoughts of suicide. School violence, cyber violence, parent violence, teacher violence, discrimination, and academic performance are on a 5-point scale, stress, depression, self-esteem, and health status are on a 4-point scale, and economic situation is on a 7-point scale.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SEX	physical_p	abusive_ex	physical_p	abusive_ex	health	Exercise_st	lack_of_sle	suicidal_id	Loneliness	Anxiety	Depressio	Happiness
2	1	1	1	1	1	3	4	1		4	3	4	2
3	1	1	1	1	1	4	4	1	1	4	4	4	3
4	1	1	1	1	1	1	1	1	1	1	1	1	2
5	2	1	1	1	1	4	4	2	1	1	1	1	4
6	1	1	1	5	5	4	4	2	1	1	1	3	3
7	2	1	1	1	1	2	1	1	2	2	3	3	3
8	2	1	1	1	2	2	2	1	2	3	3	3	3
9	1	1	1	1	1	3	2	1	2	4	4	4	3
10	2	1	1	1	1	3	4	2		1	1	1	4
11	1	2	3	1	2	2	1	1	2	3	3	2	2
12	1	1	2	1	1	1	2	1	2	2	3	4	2
13	1	1	1	1	1	4	4	1		1	1	1	4
14	2	1	1	1	2	2	2	2	3	3	3	3	2
15	1	1	1	1	1	4	4	2		3	2	3	3
16	1	1	4	1	1	2	4	2		4	2	3	2

Figure 4: data preprocessing

2.3. Data Analysis

The data used the Two-Class Decision Forest algorithm of Machine learning studio Azure. The 2019 Survey on Human Rights of Children and Youth by the Korea Youth

Policy Institute was used. In addition, using Kaggle's dataaset, suicide factors were predicted by dividing gender, suicide rate, population, HDI for year, GDP for year, and generation. The figure below shows the process of using Azure to clean up the data, then split the organized data and create data values through a trained model.

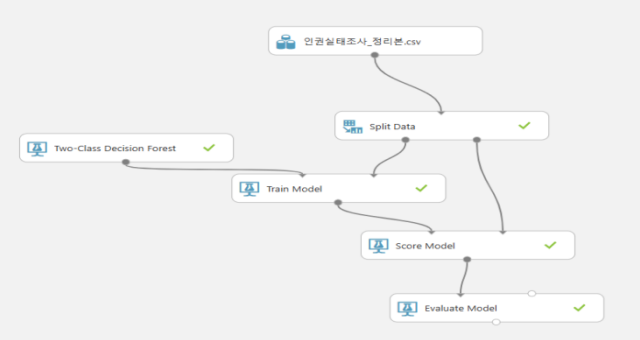


Figure 5: Data Modeling Process 1

3. Results and Discussion

3.1. Analysis Results

Performance indicators that evaluate the random forest algorithm include accuracy, sensitivity, precision, F1-score, and ROC curve, and the larger the value, the higher the predictive power of the model. The ROC curve showed better performance as it was attached to the upper left. Accuracy is an indicator of how much the prediction data is the same in my data, and it is the ratio of correctly predicting adolescents who think about suicide and adolescents who do not think about suicide. Sensitivity is the prediction that adolescents who are thinking about suicide are adolescents who are thinking about suicide. Precision represents the proportion of adolescents who actually think about suicide among adolescents who predict that they are thinking about suicide. Accuracy was found to be 78.7%, and sensitivity was found to be a slightly lower rate of 54.7%. Low sensitivity may mean that the prediction performance is not good. In other words, it can be interpreted that there is a high possibility that there are teenagers who are thinking of suicide in an unknown place or in a neglected state, and that it is difficult to find them. The specificity was high, which means that 89.7% correctly predicted that adolescents who do not think about suicide are adolescents who will not think about suicide. Using this, adolescents who do not think about suicide can be classified as adolescents who think about suicide and focus on adolescents who need help with ideation. In addition, F1-Score, the harmonic average of sensitivity and precision, was 61.9%. Increasing the value of F1-Score will require more data accumulation.

It is the variables divided using the human rights survey and the Scored Probabilities predicted using those variables.

Figure 6: Descriptive Statistics Information

Table 2: Predicted performance

	Predicted performance
Accuracy	78.7
Recall	54.7
Precision	71.1
Specificity	89.7
F1 score	61.9

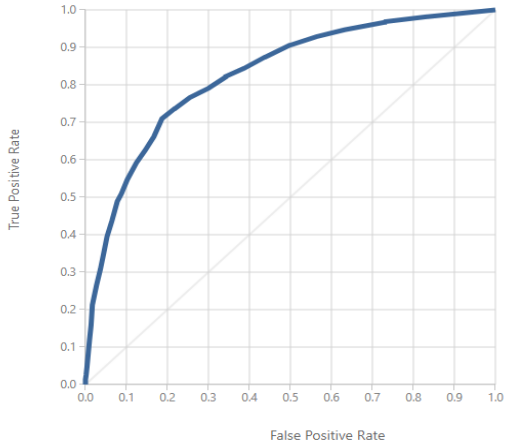


Figure 7: ROC curve

Azure's Two-Class Decision Forest (Random Forest) algorithm was used to predict suicide factors by dividing suicide rate, population, HDI for year, GDP for year, and generation based on gender, and accuracy was 84.5%, sensitivity was 81%, precision was 86%, and F1-Score was 83.5%. This data from kaggle is statistics on suicides according to various factors of all ages in Korea from 1985 to 2015. Among them, data from ages 5-14 and 15-24 were extracted to predict suicide rules with factors such as HDI, GDP, and generation.

Table 3: Kaggle – Suicide prediction

	Predicted performance
Accuracy	84.5
Recall	81.1
Precision	86.0
F1 score	83.5

3.2. Discussion

This study used the 2019 Survey on Human Rights Status of Children and Youth, which is data provided by the Korea Youth Policy Institute. In addition, Kaggle's Suicide dataset was used. The factors influencing adolescents' suicidal ideation were identified through the random forest algorithm. The accuracy of adolescents who think about suicide and adolescents who do not think about suicide was 78.7%, and the sensitivity, which is the value of predicting adolescents who think about suicide, was low at 54.7%. In contrast, the specificity of predicting that adolescents who do not think about suicide was high at 89.7%. By using the machine learning algorithm, it was intended to confirm which characteristics of the group with high or low suicidal ideation are combined. In addition, data should be continuously accumulated to increase the prediction of adolescents' suicidal ideation. Furthermore, it should be possible to provide personalized management that controls and manages suicidal ideation or to develop programs. Adolescent suicide is increasing in severity as a social problem, and there are various attempts to prevent youth suicide, but deaths from suicide are still high. Therefore, continuous attention will be required to identify the factors that cause adolescent suicidal thoughts. In addition, suicide is more effective when interpreting and analyzing both individuals and society because many social factors affect it, not on the individual level. Continuous research is needed considering this. In addition, previous studies have continuously reported that there is a gender difference in adolescents' suicidal ideation. It emphasizes that gender differences appear in the effects of human relationships (Woo et al, 2010) and depression (Lee et al., 2016; Simons &Murphy, 1985) on adolescents' suicidal ideation. Although this study does not show the difference by dividing gender, it is necessary to find out more about whether the factors affecting suicidal ideation vary depending on gender.

References

Cha, E. B., & Cho, Y. I. (2022). The effect of adolescents' emotional neglect experience on suicidal thoughts: Focusing on the mediating effect of self-esteem. *Criminal Justice Research*, *5*(1), 7-26.

Dangeti, P., (2017), Statistics for machine learning, Birmingham, UK: Packt Publishing Ltd.

Géron, A., (2017), Hands-On Machine Learning with Scikit-Learn and TensorFlow, Sebastopol, CA: O'Reilly Media

Hong, K. H. (2020), A Predictive Model for Suicidal Ideation of Adolescents Using Random Forests Machine Learning Algorithm. *Korean Journal of Social Welfare. 72*(3). 157-180

Han, Myeunghee (2022), Analysis of Predictive Factors for Suicidal Ideation of Adolescents Using Decision Tree Analysis. *Journal of Korean Public Health Nursing*, *36*(2), 157-169

IT Wiki. (2020). Decision Tree. Retrieved from https://itwiki.kr/w/%EC%9D%98%EC%82%AC%EA%B2%B0%EC%A0%95_%EB%82%98%EB%AC%B4*Journal of Korean Public Health Nursing*, *36*(2), 157-169, August 2022.

Jeon, K. S., Park, S. Y., & Cho, S. H. (2012). Gender differences in correlates of depression and suicidal ideation among Korean adolescents. *The Korean Journal of Health*, *6*(4), 295-308.

Kaggle. (2021). World_Suicide_Rates_2000-2019[Data set] https://www.kaggle.com/datasets/ankanhore545/world-suicide-rates-20002019

Kim, H.-J. (2008). Factors Influencing Youth Suicide Risk. *Korean Journal of Social Welfare*, *72*(3), 157-180.

Lee, E.-T., & Lee, E.-K. (2016). The controlled hawk of self-esteem in the relationship between middle school students' experiences of school violence, depression, and suicidal thoughts: "Gae Effect". *Youth Culture Forum*, 61-85.

Park, W. J., Kim, H. S., Park, K. H., & Kim, M. H. (2012). The factors affecting youth's risk of suicide. *Journal of the Korean Society for Suicide Prevention*, *16*(1), 1-10.

Seo, G.-W. (2021). Exploration of causes for adolescent suicide and preventive measures considering the developmental characteristics of adolescence. (Research Report 21-R21). Korea Youth Policy Research Institute. Retrieved from https://www.nypi.re.kr/brdrr/boardrrView.do?menu_nix=4o9771b7&brd_id=BDIDX_PJk7xvf7L096m1g7Phd3YC&cont_idx=732&seltab_idx=0&edomweivgp=R

Statistics Korea. (2019). Cause of Death Statistics. Retrieved from https://kostat.go.kr/board.es?mid=a10301060200&bid=218&act=view&list_no=385219

Lawyers for a Democratic Society. (2019). 2019 Korea Human Rights Report. Retrieved from http://minbyun.or.kr/wp-content/uploads/2019/12/%EB%B3%B4%EA%B4%80%EC%9A%A9_2019%EB%85%84_%ED%95%9C%EA%B5%AD%EC%9D%B8%EA%B6%8C%EB%B3%B4%EA%B3%A0%EC%84%9C.pdf

The Ministry of Health and Welfare. (2022). Results of the 2022 Lonely Death Survey. Retrieved From http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=1&CONT_SEQ=374084