ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://dx.doi.org/10.24225/jkaia.2023.1.1.1

머신러닝 기반 한국 청소년의 자살 생각 예측 모델

Machine Learning-based Predictive Model of Suicidal Thoughts among Korean Adolescents.

YeaJu JIN[1], HyunKi KIM[2]

Received: April 18, 2023. Revised: May 28, 2023. Accepted: June 30, 2023.

Abstract

This study developed models using decision forest, support vector machine, and logistic regression methods to predict and prevent suicidal ideation among Korean adolescents. The study sample consisted of 51,407 individuals after removing missing data from the raw data of the 18th (2022) Youth Health Behavior Survey conducted by the Korea Centers for Disease Control and Prevention. Analysis was performed using the MS Azure program with Two-Class Decision Forest, Two-Class Support Vector Machine, and Two-Class Logistic Regression. The results of the study showed that the decision forest model achieved an accuracy of 84.8% and an F1-score of 36.7%. The support vector machine model achieved an accuracy of 86.3% and an F1-score of 24.5%. The logistic regression model achieved an accuracy of 87.2% and an F1-score of 40.1%. Applying the logistic regression model with SMOTE to address data imbalance resulted in an accuracy of 81.7% and an F1-score of 57.7%. Although the accuracy slightly decreased, the recall, precision, and F1-score improved, demonstrating excellent performance. These findings have significant implications for the development of prediction models for suicidal ideation among Korean adolescents and can contribute to the prevention and improvement of youth suicide.

Keywords : Machine Learning, Suicidal Thoughts, Logistic Regression Analysis, Korea Youth Risk Behavior Survey

Major Classification Code : Machine Learning, Artificial Intelligence, Bigdata

1. Introduction

1.1. Importance of Research

2023 년 'OECD 보건통계(Health Statistics)'에 따르면, 2020 년 대한민국은 인구 10 만 명당 24.1 명의 자살 사망률을 기록하여, 다른 OECD 국가들보다 가장 높은 수치를 보였다. 이는 OECD 평균인 11.0 명에 비해 약 2 배 이상 높은 수치이며, 자살 사망률이 6.3 명으로 가장 낮은 멕시코에 비해 약 3.8 배 정도 높은 수치를 보였다(Ministry of Health and Welfare, 2023). 2023 년 여성가족부의 '청소년통계'에 의거하면, 자살이 청소년 사망원인 중 1 위를 차지하였으며, 인구 10 만 명당 자살 사망률은 11.7 명으로 2011 년 이후 계속해서 1 위를 차지하고 있다. 2017 년에는 7.7 명에서 21 년에는 11.7 명을 기록하며, 지속적으로 증가하고 있는 추세이다(Ministry of Gender

1 First Author. Ph.D, Big Data Medical Convergence, Eulji University, Korea. Email: 2002dew@naver.com
2 Corresponding Author or Second Author, 신남정보통신 대표. Email: r48019@naver.com

Equality and Family, 2023). 이렇듯 우리나라의 자살 문제는 심각한 상태를 보이고 있다.

자살 생각은 자살 실행으로 진행될 수 있다(Hong, 2020). 이는 자살에 대한 생각 자체가 삶을 위협하는 위험 요인이 될 수 있다는 것을 의미한다(Han, 2022). 청소년기의 자살 생각은 30 대 자살의 주요 예측 원인이 될 수 있으며, 성인이 되어서도 심리.정서적인 문제와 더불어 사회 적응에도 다양한 어려움을 겪을 수 있는 것으로 나타났다(Oh & Kwon, 2019). 또한, 자살 생각이 높은 집단일수록 자살을 시도하는 가능성이 높은 것으로 나타났으며, 평소 자살을 생각하는 청소년의 경우 그렇지 않은 청소년에 비해 자살로 생을 마감할 가능성이 매우 높은 것으로 나타났다(Oh & Kwon, 2019).

청소년의 자살 생각은 다양한 요인에 의해 발생할 수 있다. 선행 연구에서 우울이 자살 생각에 가장 큰 영향을 미친다는 결론이 나왔다(Hong & Jung, 1999; Jeong et al., 2021). 자살 생각에 직접적인 영향을 미치는 요인은 우울이며, 이외에도 불안, 충동성, 학업 스트레스, 진로 스트레스 등의 요인들이 자살 생각에 영향을 미치는 것으로 나타났다(Oh & Kwon, 2019; Lee, 2018, Jeong & Seo, 2014). 특히, 청소년기는 우울증이 발현되기 쉽기 때문에 자살과 같은 심각한 결과로 이어지기 쉽다(Lim, 2004). 따라서 청소년들의 자살 생각을 사전에 예측하고 이에 대한 조기 지원과 개입을 제공하는 모델의 개발이 필요하다.

이에 본 연구는 청소년의 자살 생각과 연관 있는 다양한 변수들을 고려하여 MS Azure 를 통해 한국 청소년의 자살생각 예측 모델을 개발하고, 청소년들의 자살 예방에 기여하고자 한다.

1.2. Research Objective

본 연구의 목적은 MS Azure 프로그램을 활용하여 한국 청소년의 자살 예방을 위한 모델을 개발하는 것이다. 의사결정 숲, 서포트 벡터 머신, 로지스틱 회귀를 활용하여 자살 생각을 예측하는 모델을 개발하고, 각 모델의 정확성을 평가하여 최적의 모델을 선정한다. 이를 통해 한국 청소년의 자살 예방을 위한 유용한 지표를 제공하고, 건강하고 안정적인 성장과 행복한 미래를 창출하는 데 기여하고자 한다.

2. Research Methodology

2.1. Research Subjects and Data Collection

본 연구는 한국 청소년의 자살 생각을 예측하기 위해 제 18 차(2022 년) 청소년건강행태조사의 원시자료를 이용한 2 차 분석 연구이다(Korea Disease Control and Prevention Agency, 2022).

청소년건강행태조사는 중학교 1 학년 학생들부터 고등학교 3 학년까지의 학생들을 대상으로 실시하는 익명성을 유지한 자기기입식 온라인 조사이다. 본 조사는 한국 청소년들의 건강행태 통계를 산출하는 것을 목적으로 하여, 2005 년부터 매년 수행되고 있는 정부승인 통계조사(승인번호 제 117058 호)이다. 산출된 통계자료는 청소년 건강정책과 건강증진사업을 기획하고 평가하는데 필요한 기초적 자료로 활용된다. 제 18 차(2022 년) 조사는 총 800 개교를 대상으로 실시되었으나, 당사 사정으로 인해 중학교 398 개교와 고등학교 400 개교만 참여하게 되어, 총 51,850 명으로, 학생 인원수를 기준으로 조사 참여율은 92.2% 였다. 조사 대상자 수는 조사 당일의 출석부에 기재된 학생의 인원수를 나타낸다. 표본학급의 학생 중에 장기결석과 자체적인 조사 참여가 불가능한 특수아동 및 문자 해독장애가 있는 학생은 조사 대상자에서 제외되었다.

2.3. Research Tool

제 18 차 청소년건강행태조사의 조사 항목은 정신건강, 신체활동, 흡연 등 총 15 가지의 항목으로 구성되어 있다. 본 연구에서는 대상자의 정신건강, 일반적 특성, 신체활동, 흡연, 음주, 성행태와 관련된 항목을 사용하였다.

2.3.1. Mental Health

대상자의 정신건강과 관련된 문항은 '자살 생각', '평상시 스트레스 인지', '슬픔과 절망감 경험', '외로움 경험', 'GAD-7'을 사용하였다. 자살생각 문항은 주요 종속변수로, 최근 12 개월 동안 없다, 있다로 조사되었다(Korea Disease Control and Prevention Agency, 2022). 평상시의 스트레스 인지 문항은 '대단히 많이 느낀다'가 1 번, '많이 느낀다'가 2 번, '조금 느낀다'가 3 번, '별로 느끼지 않는다'가 4 번, '전혀 느끼지 않는다'가 5 번 으로 조사되었다(Korea Disease Control and Prevention Agency, 2022). 슬픔과 절망감 경험에 대한 문항은 최근 12 개월 동안 '없다'가 1 번,

'있다'가 2 번으로 조사되었다(Korea Disease Control and Prevention Agency, 2022). 외로움 경험은 '전혀 외로움을 느끼지 않았다'가 1 번, '거의 외로움을 느끼지 않았다'가 2 번, '가끔 외로움을 느꼈다'가 3 번, '자주 외로움을 느꼈다'가 4 번, '항상 외로움을 느꼈다'가 5 번 으로 조사되었다(Korea Disease Control and Prevention Agency, 2022). GAD-7 문항은 7 가지 문항으로 구성되어 있으며, 지난 2 주 동안 불안이나 걱정 등의 감정들로 인해 얼마나 자주 방해를 받았는지에 대해 '전혀 방해받지 않았다'는 0 점, '며칠 동안 방해 받았다'는 1 점, '7 일 이상 방해 받았다'는 2 점, '거의 매일 방해 받았다'는 3 점으로 나누어 점수를 부여하여 각 문항에 대한 점수를 합산한다(Korea Disease Control and Prevention Agency, 2022). 합산된 점수가 0 에서 4 점 사이이면 범불안장애에 대해 '정상', 5 에서 9 점 사이이면 '경도 불안', 10 에서 14 점 사이이면 '중등도 불안', 15 에서 21 점 사이이면 '심한 불안'으로 판단한다(Korea Disease Control and Prevention Agency, 2022). 이에 따라 0~4 점이면 이상 없음으로, 5~21 점이면 범불안장애가 있는 것으로 재분류하였다.

2.3.2. General Characteristics

대상자의 일반적 특성과 관련된 문항은 '학년', '학업성적', '경제상태', '거주형태' 를 사용하였다. 학년 문항은 중학교 1 학년부터, 고등학교 3 학년까지의 현재 학년으로 조사되었다(Korea Disease Control and Prevention Agency, 2022). 학업성적 문항은 상 부터 하까지 5 가지로 조사되었다. 경제상태 문항은 상 부터 하까지 5 가지로 조사되었다(Korea Disease Control and Prevention Agency, 2022). 거주형태 문항은 '가족과 함께 살고 있다'가 1 번, '친척집에서 살고 있다'가 2 번, '하숙, 자취'가 3 번, '기숙사'가 4 번, '보육시설'이 5 번 으로 조사되었다(Korea Disease Control and Prevention Agency, 2022).

2.3.3. Physical Activity

서성익과 김근국의 연구(2022)에서 청소년의 신체활동 수행 빈도가 높아질수록 스트레스는 적게 받는다고 나타났다. 이를 바탕으로 자살 생각에 대한 청소년 신체활동의 영향을 파악하고자 하였다. WHO 에서 제시한 제 3 차 신체활동 지침서를 개정한 한국건강증진개발원의 '한국인을 위한 신체활동 지침서(2022)'에 따르면, 만 6~18 세의 경우 매일 60 분 이상씩 중강도 이상의 유산소 신체활동을 수행해야 한다. 이 60 분의 신체활동 중에서 고강도 유산소 신체활동을 일주일에 3 번 이상 해야 한다고 권장하고 있다(Korea Health Promotion Institute, 2022). 이를 바탕으로, 2 일 동안 120 분의 신체활동만 한다면 권장 기준에 미달할 수 있기에, 신체활동 문항은 하루에 60 분 이상씩 신체활동을 하는 일수에 따라서 최근 7 일 동안 '없다', '주 1 일', '주 2 일'로 응답한 사람은 신체활동을 '안한다', 주 3 일~7 일은 '한다'로 재분류하였다.

2.3.4. Smoking

흡연 청소년은 비흡연 청소년에 비해 우울장애, 불안장애, 행동장애 등의 정신과적 문제를 더 많이 보인다(Lim, 2004). 이를 바탕으로 자살 생각에 대한 청소년 흡연의 영향을 파악하고자 하였다. 흡연은 평생 흡연 경험 유무에 따라 흡연자, 비흡연자로 조사되었다.

2.3.5. Drinking

김미영의 연구(2017)에서 우울감이 높을수록 음주 수준도 높아졌다(p=.000). 이를 바탕으로 청소년의 음주가 자살 생각에 미치는 영향을 파악하고자 하였다. 대상자의 음주는 평생 음주 경험 유무에 따라 음주자, 비음주자로 조사되었다.

2.3.6. Sexual Behavior

염미정, 이경주와 이주영인의 연구(2022)에서, 성 경험이 있는 청소년들은 비교적으로 그와 반대되는 경우에 스트레스 수준, 우울감, 자살생각이 더 높았으며, 행복감은 낮았다. 이를 바탕으로 자살 생각에 대한 청소년 성 경험의 영향을 파악하고자 하였다. 대상자의 성관계 경험은 성관계를 해본 적이 '없다', '있다'로 조사되었다.

2.4. Data Preprocessing

MS Azure 의 'Clean Missing Data'를 사용하여 결측치가 있는 행을 제거하였다. 그 결과 전체 조사 대상자인 51,850 명에서 51,407 명만을 사용하게 되었다. 이후 '자살 생각', '평상시 스트레스 인지', '슬픔과 절망감 경험','외로움 경험', '범불안장애 여부', '학년', '학업성적', '경제상태', '거주형태', '신체활동', '흡연여부', '음주여부', ' 성관계경험 여부' 총 13 개의 항목을 선택하여 'Edit Meatadata'를 통해 범주화 하였다. 훈련데이터와 검증데이터는 무작위로 7:3 비율로 나누어 사용하였다. 추후 선정된 모델의 데이터 불균형을 처리하기 위해 'Smote'를 사용하였다.

2.5. Data Analysis

수집된 자료는 MS Azure 프로그램을 이용하여 분석하였으며, 대상자의 자살생각 예측모델을 개발하기 위해 의사결정 숲(Two-Class Decision Forest), 서포트 벡터 머신(Two-Class Support Vector Machine), 로지스틱 회귀(Two-Class Logistic Regression)을 사용하였다. 모델의 검증력은 ROC(Receiver Operating Characteristic) curve 를 사용하여 평가하였다.

3. Research Findings

3.1. Performance Comparison of Models

의사결정 숲을 사용한 자살 생각 예측 모델의 성능과 설계 과정은 [그림 1]과 같다. 의사결정 숲 모델을 사용하기 위해 'Two-Class Decision Forest'을 사용하였다. 의사결정 숲의 샘플링 방법으로는 배깅(Bagging)을 선택하였다. 트리의 개수는 사용된 변수의 개수인 13 개로 설정하였다. 트리의 최대 깊이는 32 로, 각 노드에서 선택할 후보 분할 수는 128 로 설정하였다. 그 결과 정확도는 84.8%, 정밀도는 44.0%, 재현율은 31.4%, F1-score 는 36.7%로 나타났다.

서포트 벡터 머신을 사용한 자살 생각 예측 모델의 성능과 설계 과정은 [그림 2]와 같다. 서포트 벡터 머신 모델을 사용하기 위해 'Two-Class Support Vector Machine'을 사용하였다. 반복 횟수는 10 으로 설정하였으며, Lambda 는 기본값인 0.001 로 설정하였다. 그 결과 정확도는 86.3%, 정밀도는 54.0%, 재현율은 15.9%, F1-score 는 24.5%로 나타났다.

로지스틱 회귀를 사용한 자살 생각 예측 모델의 성능과 설계 과정은 [그림 3]과 같다. 로지스틱 회귀분석 모델을 사용하기 위해 'Two-Class Logistic Regression'을 사용하였다. L1 과 L2 의 가중치는 1 로 설정하였다. L-BFGS 의 메모리 크기는 20 으로 설정하였다. 그 결과 정확도는 87.2%, 정밀도는 58.4%, 재현율은 30.5%, F1-score 는 40.1%로 나타났다.

최종적으로 의사결정 숲, 서포트 벡터 머신, 로지스틱 회귀를 사용한 모델에 대해 비교 분석 결과, 로지스틱 회귀를 사용한 모델의 성능이 87.2%로 가장 높은 성능을 보였다. 또한, 정밀도와 재현율의 조화 평균인 F1-score 가 40.1%로, 세가지의 모델 중에 가장 높은 성능을 보였다. 이에 따라 본 데이터셋을 통해 예측 모델을 개발하기 위해서는 로지스틱 회귀를 사용하는 것이 가장 적합하다고 판단하였다.

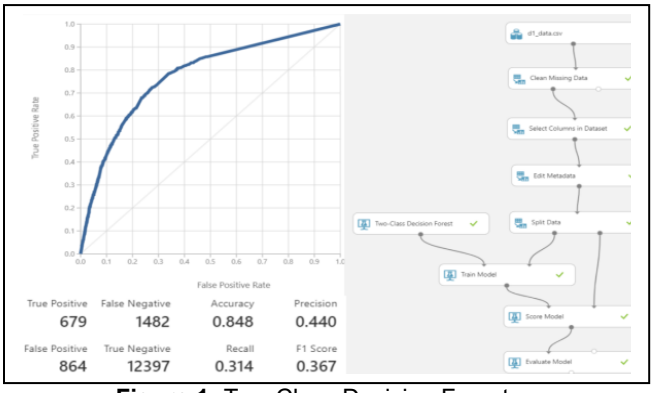

Figure 1: Two-Class Decision Forest

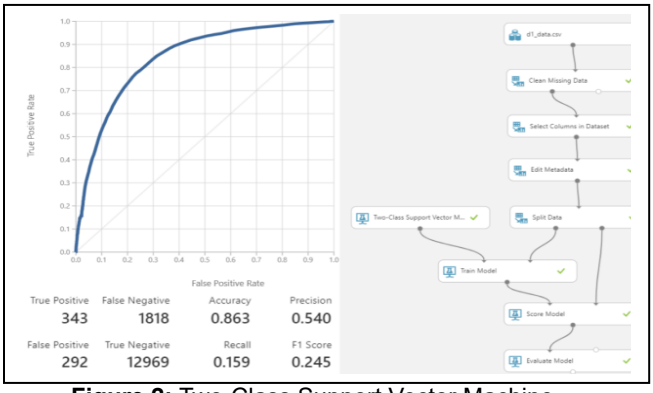

Figure 2: Two-Class Support Vector Machine

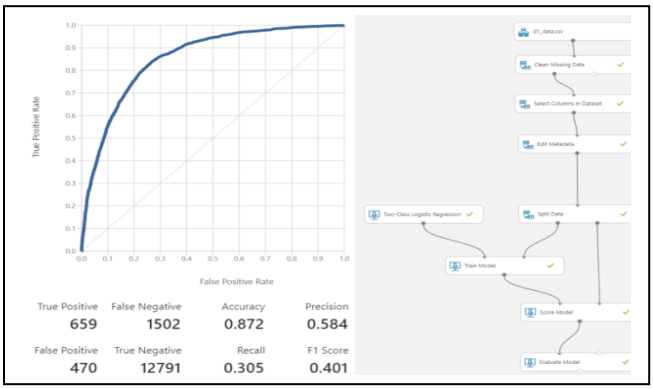

Figure 3: Two-Class Logistic Regression

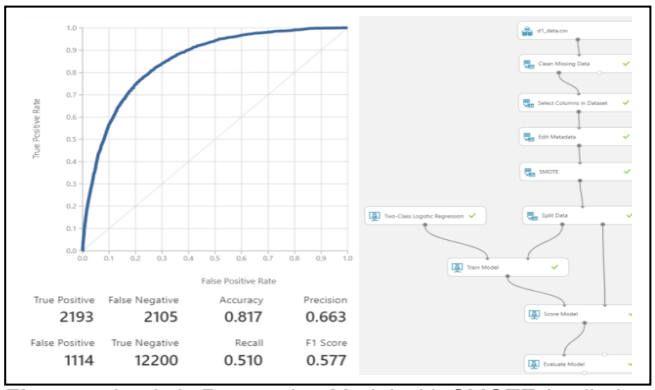

Figure 4: Logistic Regression Model with SMOTE Applied

3.2. Performance Improvement of the Model with SMOTE Applied

SMOTE 를 통해 로지스틱 회귀 모델을 사용한 자살 생각 예측 모델의 성능을 개선한 결과와 설계 과정은 [그림 4]과 같다. 이전의 설계 과정에 'Edit Meatadata'뒤에 'SMOTE'를 추가하여 예측 변수인 '자살 생각'을 기준으로 데이터셋의 불균형을 조절하였다. SMOTE 의 percentage 는 100 으로, 최근접 이웃 수는 1 로 설정하였다. 그 결과 정확도는 81.7%, 정밀도는 66.3%, 재현율은 51.0%, F1-score 는 57.7%로 나타났다.

최종적으로, SMOTE 를 사용하지 않은 로지스틱 회귀 모델과 비교한 결과는 [표 1]과 같다. SMOTE 를 사용한 로지스틱 회귀 모델의 정확도는 81.7%로 SMOTE 를 사용하지 않은 로지스틱 회귀 모델보다 5.5% 정도 낮았다. 그러나 정밀도와 재현율, F1-score 에서는 SMOTE 를 사용하지 않은 로지스틱 회귀 모델보다 7.9%, 20.5%, 17.6% 정도 향상된 우수한 성능을 보여주었다. 따라서 SMOTE 를 사용하지 않은 로지스틱 회귀 모델은 전반적인 성능 면에서는 우수하지만, 정밀도와 재현율, F1-score 를 고려하였을 때는 SMOTE 를 사용한 로지스틱 회귀 모델이 적합할 것으로 보인다.

Table 1: Comparison of Logistic Regression Models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression Model	0.872	0.584	0.305	0.401
Logistic Regression Model with SMOTE	0.817	0.663	0.510	0.577

4. Conclusion

본 연구에서는 한국 청소년의 자살 생각 예측 모델을 개발하기 위해 MS Azure 프로그램을 활용하여 의사결정 숲, 서포트 벡터 머신, 로지스틱 회귀 모델을 사용하였다. 최종적으로, 각 모델별 비교 분석을 통해 최적의 모델을 선정 후 성능을 개선하고 평가하였다.

연구 결과, 로지스틱 회귀 모델이 정확도, 정밀도, F1-score 에서 가장 우수한 성능을 보였다. 의사결정 숲은 재현율에서 가장 우수한 성능을 보였지만, 다른 측면에서는 상대적으로 낮은 성능을 보였다. 서포트 벡터 머신 또한 전반적으로 성능이 중간 수준이었으며, 특히 재현율이 가장 낮은 성능을 보였다. 따라서 로지스틱 회귀 모델이 가장 우수한 성능을 보여 자살 생각 예측에 적합한 모델임을 확인하고 최종 모델로 선택되었다.

SMOTE 를 사용하여 로지스틱 회귀 모델의 성능을 개선한 결과, 정확도는 다소 낮아졌지만 재현율, 정밀도, F1-score 는 모두 향상되어 우수한 성능을 보여주었다. 따라서 SMOTE 를 활용하여 불균형한 데이터를 처리하였을 때 모델의 성능을 향상시킬 수 있음을 확인하였다.

결론적으로, 본 연구는 다양한 머신러닝 모델을 활용해 한국 청소년의 자살 생각 예측에 대한 연구를 수행하였고, 로지스틱 회귀 모델을 통해 높은 예측 성능을 얻을 수 있음을 확인하였다. 또한, SMOTE 를 적용하여 데이터 불균형을 처리하였을 때 전반적인 성능을 개선할 수 있음을 확인하였다. 이는 한국 청소년의 자살 생각 예측 모델 개발에 기여할 수 있는 중요한 의미를 가지고 있다. 이러한 결과를 통해, 청소년의 자살 예방과 개선에 기여할 수 있다.

Reference

Han, M. H., (2022), Analysis of Factors Predicting Adolescent Suicidal Ideation Using Decision Tree Analysis: Focused on the 2019 Survey on the Status of Children and Adolescents' Rights. *Korean Journal of Community Health Nursing*, *36*(2), 157-169.

Hong, K. H., (2020), Predicting and Analyzing Suicidal Ideation among Male and Female Adolescents Based on the Random Forest Machine Learning Algorithm. *Korean Journal of Social Welfare*, *72*(3), 157-180.

Hong, N. M., & Jung, Y. S., (1999), Analysis of Factors Influencing Adolescent Suicidal Ideation. *Korean Journal of Social Welfare*, *37*, 449-473.

Jeong, M. S., & Seo, S. K., (2014), Predictive Variables of

Adolescent Suicide Attempts. *Korean Journal of Youth Studies, 25*(2), 145-171.

Jeong, Y. M., Park, H. S., & Kim, S. M., (2021). Influencing Factors of Adolescent Suicidal Ideation: 2020 Youth Health Behavior Survey. *Journal of Healthcare Management Research*, *15*(3), 31-40.

Kim, M. Y., (2017), Analysis of Predictive Factors for Problem Drinking Using Logistic Regression Analysis. *Digital Convergence Research*, *15*(5), 487-494.

Lee, B. D., (2018), A Study on the Influencing Factors of Adolescent Suicidal Ideation: Focusing on the Vulnerability-Stress Model. *Korean Journal of Public Administration and Police Studies*, *27*(2), 193-216.

Lim, E. H., (2004), A Review of Adolescent Mental Health Issues. *Professor's Collection of Theses*, *8*, 303-324.

OECD Health Statistics, (2023), Ministry of Health and Welfare.

Oh, D. K., & Kwon, S. Y., (2019), Causal Relationships between Academic and Career Stress, Anxiety, Impulsivity, Depression, and Suicidal Ideation among Adolescents. *Journal of Future Youth Studies*, *16*(4), 29-45.

Revised Physical Activity Guidelines for Koreans, (2022), Korea Health Promotion Institute

Seo, S. I., & Kim, G. K., (2022), The Impact of Types and Frequency of Physical Activity on Happiness, Mental Health, and Subjective Health Perception in Korean Adolescents. *Journal of Humanities and Social Science*, *13*(2), 49-64.

Yeom, M. J., Lee, K. J., & Lee, J. Y., (2020), Mental Health and Influencing Factors in Korean Adolescent Boys and Girls - Focused on Sexual Experience with the Opposite Sex: Using Data from the 11th Adolescent Health Behavior Online Survey. *Journal of the Korean Academy of Psychiatric and Mental Health Nursing*, *29*(3), 195-206.

Youth Health Behavior Survey, (2022), Korea Disease Control and Prevention Agency.

Youth Statistics, (2023), Ministry of Gender Equality and Family.