기계학습 알고리즘 기반 청소년 대상 자살 생각 예측 모델에 관한 연구

이시언 [1], 강슬미 [2], 고명환 [1], 정진우 [3], 조석헌*
경기대학교 [1], 연세대학교 [2], 건국대학교 [3], *University of California, San Diego (UCSD)

sieonheaven@kyonggi.ac.kr[1], tmfal5023@yonsei.ac.kr[2], gmhgmh119@kyonggi.ac.kr[1],
jinwoo0302@konkuk.ac.kr[3], *justinshcho@gmail.com

Study on Machine Learning Algorithm-based Prediction Models for Adolescent Suicidal Ideation

Sieon Lee[1], Seulmi Kang[2], Myeonghwan Go[1], Jinwoo Jung[3], and Seokheon Cho*
Kyonggi University[1], Yonsei University[2], Konkuk University[3],
*University of California, San Diego (UCSD)

요 약

This study proposes a machine learning algorithm-based models for predicting adolescent suicidal ideation, which can be used for detecting youth at risk of suicide and facilitating proper support to prevent significant societal problem of adolescent suicide. We used Korea Youth Risk Behavior Web-based Survey (KYRBS) dataset and analyzed various variables including behavior patterns, emotional state, family, and social support system of adolescents. This study applied Logistic Regression and Random Forest algorithms and achieved the estimation of generalization performance of our proposed models by using 5-fold cross-validation. Furthermore, we improved performance by utilizing Synthetic Minority Over-sampling Technique (SMOTE), which is one of the over-sampling techniques to solve imbalanced data problems.

Ⅰ. 서 론

한국 사회의 심각한 문제 중의 하나인 청소년들의 자살률은 2015 년부터 꾸준히 증가하는 추세이다 [1]. 이 현상은 아직도 청소년의 건강 수준이 크게 개선되지 않고 있음을 시사하고 있다. 따라서, 본 연구에서는 청소년기의 자살률을 낮추는데 기여하고자 청소년들의 자살 생각 여부를 예측하는 모델을 제시하고자 한다.

T. Lee *et al.*은 한국 아동 및 청소년 패널조사 (KYPS) 데이터를 활용해 자살 생각의 변화 궤적을 도출했다 [2, 3]. eXtreme Gradient Boosting (XGBoost) 알고리즘을 통해 자살 생각의 원인을 도출한 결과, 우울, 자아 존중감 및 주말 취침 시간 등이 주요 원인으로 영향을 미쳤다. 여기에서 사용한 데이터는 중학교 1 학년 학생들의 데이터로, 전체 청소년을 대표한다고 보기 어렵다는 한계가 있다. D. Lekkas *et al.*은 소셜 네트워킹 데이터 (SN)에 Ensemble 알고리즘을 적용하여 독일 청소년의 급성 자살 생각을 예측했다 [4]. 계정을 팔로우하는 수, 사용자의 참여도와 부정적 감정 등이 자살 생각에 영향을 미치는 주요 원인으로 나타났다. 중학생부터 고등학생까지 SNS 를 사용하는 47 명의 소규모 청소년을 대상으로 한 연구로, 표본이 너무나 적어 일반화된 자살 생각 예측 모델로 간주하기 어렵다. C. Su *et al.*은 Connecticut Children's Medical Center (CCMC)에서 제공하는 Electronic Health Record (EHR) 데이터 세트 [5]를 사용하여 자살 시도를 예측했다 [6]. 이 EHR 데이터 세트는 동일한 아동에 대해서 장기간에 걸쳐 여러 차례 수집한 데이터로 구성되어 있다. 자살 시도를 예측할 때 사용한 과거 데이터의 기간을 변화시키면서 데이터 세트를 달리하여 학습 및 테스트를 진행하였다. 고려한 Logistic Regression Classifier 알고리즘 기반 모델의 ROC-AUC 의 값은 0.81 부터 0.86 까지의 분포를 보였다. 이 연구에서는 사용된 EHR 데이터 세트의 자살 시도라는 종속 변수가 가지는 불균형을 해소하기 위한 어떠한 기법도 사용하지 않아 자살 위험과 관련된 특정 요인들을 효과적으로 학습하지 못하는 한계점을 가진다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 원본 데이터 세트와 전처리 과정에 대해서 설명한다. 제 Ⅲ장은 청소년 자살 생각 예측 모델을 위해 사용되는 알고리즘 및 성능 지표에 대해 언급한다. 제 Ⅳ장에서는 본 연구에서 제시하는 청소년 자살 생각 예측 모델들의 성능을 분석한다. 마지막으로 제 Ⅴ장은 본 연구의 결과와 향후 과제에 대해 논의한다.

Ⅱ. 원본 데이터 세트 설명 및 전처리 과정

2.1 원본 데이터 세트 설명

본 연구에서는 대한민국 질병관리청과 교육부가 2021 년 공동으로 실시한 '대한민국 청소년 건강행태조사'의 원본 데이터를 활용하여 분석을 진행했다 [7]. 원본 데이터 세트의 경우, 항목별 응답 중 평균에서 많이 벗어난 데이터는 이상치로 간주하여 결측 처리된 상태로 제공되었다. 이 데이터 세트에는 196 개의 독립 변수들과, 자살 생각이라는 종속 변수를 포함하고 있다.

2.2 데이터 전처리 및 분석용 데이터 세트 생성 과정

그림 1 은 Random Forest (RF) 알고리즘을 사용하여 자살 생각 여부를 예측할 때, 청소년 건강행태조사 데이터 세트에 포함된 197 개의 특성에 대한 특성 중요도 (feature importance)를 계산하고 상위 6 개의 특성과 특성마다 가지는 특성 중요도를 나타내고 있다. 이 방법을 통해서 자살 생각이라는 종속 변수에 많은 영향을 미치는 주요 특성들을 추출할 수 있다. 이 결과를 얻기 위해서 RF 알고리즘에서 Gini impurity 를 최소화하도록 설정하고 모델 수를 1,000 개로 구성했으며, 5 번의 K 겹 교차 검증 (K-fold cross validation) 기법을 사용하였다.

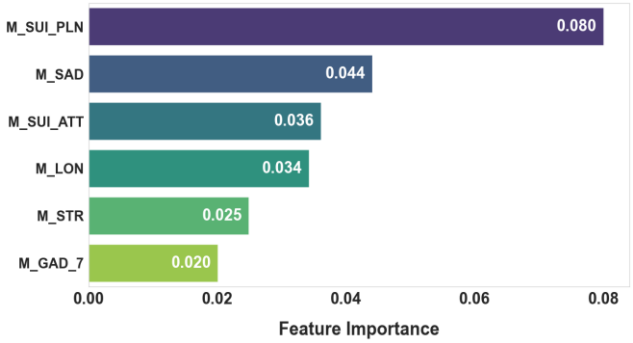

그림 1. Random Forest 알고리즘을 통해 선택된 상위 6 개 특성들에 대한 특성 중요도

그림 1 에 도시한 상위 6 개의 특성들의 의미는 다음과 같다. M_SUI_PLN 과 M_SUI_ATT 는 각각 지난 12 개월간 자살 계획 및 시도의 여부를 나타낸다. M_SAD 는 지난 12 개월간 슬픔 또는 절망감을 느꼈는지에 대한 여부이다. M_LON 은 지난 12 개월간 외로움을 느낀 빈도이며, M_STR 은 느끼는 스트레스의 정도를 나타낸 척도이다. M_GAD_7 은 지난 2 주간 끔찍한 일이 생길 것 같은 두려움으로 인해 고통받은 날들의 빈도를 나타낸다.

이때, 특성 중요도의 누적 합이 73.6%인 상위 60 개의 독립 변수와 종속 변수인 자살 생각을 포함하여 총 61 개의 변수를 가지는 Selected Feature Dataset (SFD)을 구성했다. 그림 2 는 SFD 의 종속 변수인 자살 생각 (M_SUI_CON)의 분포를 보여주고 있다. 전체 데이터 샘플 수는 총 54,848 개이며, 이 중 다수 클래스인 자살 생각을 하지 않은 청소년들 (M_SUI_CON = 1)과 소수 클래스인 자살 생각을 한 청소년들 (M_SUI_CON = 2)은 각각 47,892 명과 6,956 명으로 구성되어 있다. 부정 클래스와 긍정 클래스의 비율은 각각 87.3%와 12.7%으로 불균형한 데이터 세트를 형성하고 있다. 이때, 소수 클래스를 양성 클래스 (positive class)로 설정하였다.

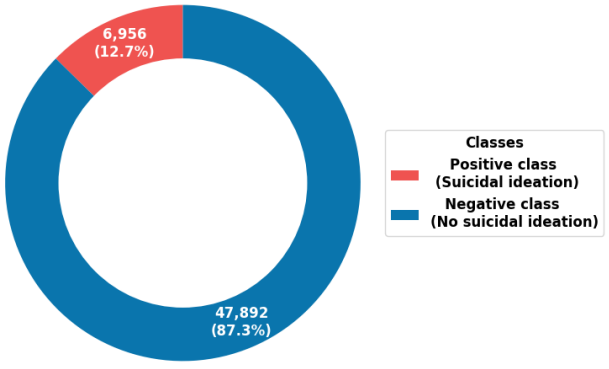

그림 2. SFD 의 클래스 분포

Ⅲ. 기계학습 알고리즘 및 성능 지표

3.1 알고리즘

Logistic Regression (LR)은 독립 변수의 선형 결합을 이용하여 범주형 종속 변수를 예측하는 통계 기반 분류 알고리즘이다. 본 연구에서 사용된 LR 알고리즘의 하이퍼 파라미터로 Epoch 는 1700 그리고 Variance 는 0.001 로 설정하였다. Random Forest (RF)는 여러 결정 트리를 결합한 앙상블 학습 알고리즘이다. 이 모델의 하이퍼 파라미터로 트리의 개수는 1500, 트리 깊이는 8 로 설정하였다. 이 두 개의 알고리즘을 활용한 모델 모두에서 5 번의 K-겹 교차 검증 (K-fold cross validation)기법을 사용하였다.

3.2 성능 평가 지표

자살 생각 예측 모델에 대한 성능 평가 지표로는 재현율 (Recall), 균형 정확도 (Balanced accuracy) 및 F2 score 를 사용했다. Recall 은 실제 양성 데이터 중 정확하게 양성으로 예측된 데이터의 비율이다. Balanced accuracy 는 민감도 (sensitivity)와 특이도 (specificity)의 산술 평균으로 얻어지고 모든 클래스를 동일한 비중으로 반영하여 불균형 데이터에서 더욱 공정한 성능 평가를 가능하게 한다. F2 score 는 정밀도 (Precision)과 Recall 에 1:2 의 가중치를 부여함으로써, 실제 양성 데이터를 음성으로 잘못 분류하게 되면 동일한 상황에서 얻어지는 F1 score 보다 낮은 F2 score 를 갖게 된다. 즉, F2 score 가 F1 score 보다 Recall 값에 더욱 밀접하게 반응하는 성능 지표이다.

Ⅳ. 알고리즘 성능 비교 및 분석

그림 3 은 본 연구에서 고려하는 다양한 상황에서의 청소년 자살 생각 예측 모델들의 성능 지표 값들을 보여준다. SFD 에 샘플링을 적용하지 않은 경우와 SFD 전체에 오버 샘플링 (over-sampling) 기법들 중의 하나인 'Synthetic Minority Over-sampling Technique (SMOTE)'를 적용한 경우로 나누어 예측 모델들의 성능을 비교하였다. 또한, LR 과 RF 의 2 개의 알고리즘 기반 모델들을 분석함으로써 총 4 가지 경우의 예측 모델들의 Recall, F2 score 및 Balanced accuracy 값들을 비교하였다.

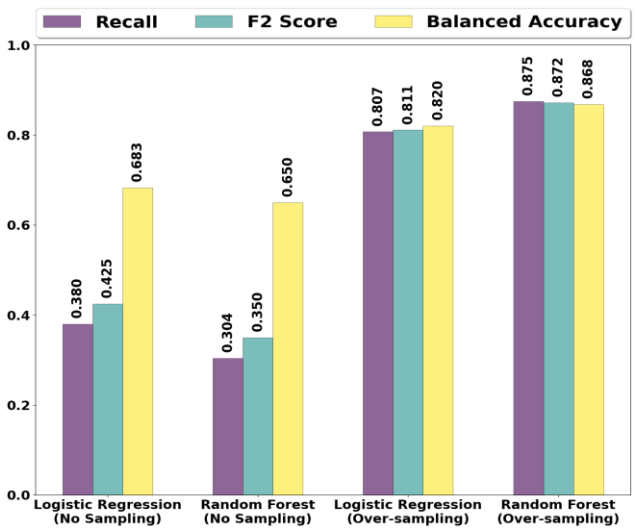

그림 3. 청소년 자살 생각 예측 모델 성능 비교

샘플링을 적용하지 않은 경우에 비해서 SMOTE 기법을 적용한 경우에 2 개의 알고리즘 기반 모델들의 모든 성능 지표 값들이 향상됨을 확인할 수 있다. 특히, 샘플링을 적용하지 않은 경우에, 2 개의 알고리즘 모두 Balanced accuracy 가 Recall 보다 높음을 알 수 있다. 이는 데이터의 불균형이 심한 데이터 경우 소수 클래스에 대한 학습이 제대로 진행되지 않았기 때문이다. 반면, SFD 전체에 over-sampling 을 적용하여 불균형을 해소함으로써, Recall, F2 score 및 Balanced accuracy 등 모든 성능 지표의 값들이 크게 향상됨을 볼 수 있다. 특히, 샘플링을 적용한 경우에 있어서 RF 의 성능이 LR 의 성능보다 우수했으며, Recall, F2 score 그리고 Balanced accuracy 가 각각 0.875, 0.872 와 0.868 의 값을 가진다.

V. 결론

본 연구에서는 최근 커다란 사회적 이슈 중의 하나인 청소년들의 자살 문제를 적극적으로 예방하기 위해 청소년 건강행태 설문조사 데이터를 활용하여 자살 생각 여부를 분류 예측하는 모델을 제시했다. 고려한 기계학습 알고리즘으로는 Logistic Regression 과 Random Forest 이다. 또한, 전체 데이터 세트에 Synthetic Minority Over-sampling Technique 기법을 적용하여 데이터의 불균형을 해소함으로써 모델의 성능을 향상시켰다. 본 연구의 결과로서 over-sampling 을 적용한 데이터 세트에 Random Forest 기반 예측 모델이 0.875 의 Recall 과 0.868 의 Balanced accuracy 를 가짐으로써 가장 우수한 성능을 보였다

본 연구의 결과는 청소년들의 자살 생각을 조기에 인식하고 적절한 예방 조치를 취해 궁극적으로 청소년들의 자살 예방에 기여할 수 있다는 데 의의가 있다. 그러나 본 연구는 횡단적 설문조사 (cross-sectional survey) 데이터에 기반하고 있어, 청소년의 실제 자살 시도 여부를 평가하는 데는 한계가 있다. 이에 따라, 청소년들의 자살 시도 여부에 대한 종단적 연구 (longitudinal study)를 기반으로 한 예측 모델 개발을 향후 과제로 남기는 바이다.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation). (No.2021-0-01393) & (No.2023-0-00054)

참 고 문 헌

[1] Su-Jin Sim and Eun-Ah Kim, “The Quality of Life for Children and Adolescents (2022),” Statistics Korea, pp. 42-44, Dec. 2022.

[2] Kantar Public, "Korean Youth Panel Survey 2018," National Youth Policy Institute, Mar. 2019, Available: https://www.nypi.re.kr/archive/board?menuId=MENU00220&siteId=null.

[3] Taek-Ho Lee and Kwang-Hyun Kim, “Analysis of Predictors of Suicidal Ideation Trajectory among Adolescence: Using LCGA and XGBoost,” Korean Journal of Youth Studies, vol. 30, no. 10, pp. 31-59, Oct. 2023.

[4] Damien Lekkas, Robert J. Klein, and Nicholas C. Jacobson, “Predicting acute suicidal ideation on Instagram using ensemble machine learning models,” Internet Interventions, vol. 25, Jul. 2021.

[5] "Electronic Health Record (EHR)", Connecticut Children’s Medical Center, 2020, Available: https://www.connecticutchildrens.org/patients-families/medical-records

[6] Chang Su, Robert Aseltine, Riddhi Doshi, Kun Chen, Steven C. Rogers, and Fei Wang, “Machine learning for suicide risk prediction in children and adolescents with electronic health records,” Translational Psychiatry, vol. 10, no. 1. Nov. 2020.

[7] “Korean Youth Risk Behavior Web-based Survey (2021),” Korean Disease Control and Prevention Agency, Apr. 2022, Available:http://www.kdca.go.kr/yhs/home.jsp