한국정보전자통신기술학회논문지(jkiiect)'24-12, Vol.17 No.6
논문 24-17-06-601 http://dx.doi.org/10.17661/jkiiect.2024.17.6.601

스트레스 수준 예측을 위한 분류 모델 및 회귀 모델 개발에 대한 연구

정유연*, 정진형**

Research on developing classification models and regression models to predict stress levels

YuYeon Jung*, Jin-Hyoung Jeong**

요 약 본 연구는 스트레스 수준을 예측하기 위한 두 가지 머신러닝 모델, 즉 이진 분류 모델과 회귀 모델을 개발하고, 그 성능을 평가한 연구이다. 연구의 주요 목적은 스트레스 수준을 보다 정확하게 예측할 수 있는 모델을 제시하는 것을 목표로 한다. 이를 위해 랜덤 포레스트 분류(Random Forest Classifier)와 랜덤 포레스트 회귀(Random Forest Regressor)를 각각 사용하여 두 모델을 훈련시키고, 예측 성능을 비교하였다. 이진 분류 모델에서는 스트레스 수준을 "높음"과 "낮음"으로 이진화하여 분류하였고, 정확도는 100%, 정밀도, 재현율, F1 점수 모두 1.0을 기록하였다. 이는 모델이 스트레스 수준을 명확하게 구분하는 데 매우 효과적임을 보여주었다. 회귀 모델에서는 스트레스 수준을 연속적인 값으로 예측하며, 평균 제곱 오차(MSE)는 0.00059, R^2 점수는 0.9999로 매우 높은 성능을 보였다.

Abstract This study developed two machine learning models for predicting stress levels, namely, a binary classification model and a regression model, and evaluated their performance. The main purpose of the study is to present a model that can predict stress levels more accurately. For this, the two models were trained using Random Forest Classifier and Random Forest Regressor, respectively, and the prediction performance was compared. In the binary classification model, the stress levels were classified by binarizing them into "high" and "low," and the accuracy was 100%, and the precision, reproducibility, and F1 scores were all 1.0. It showed that the model was very effective in clearly distinguishing stress levels. The regression model predicts stress levels as continuous values with a mean square error (MSE) of 0.00059 and an R^2 score of 0.9999, showing very high performance.

Key Words : Binary Classification Model, Machine learning, RandomForest, Regression Model, Stress Level

1. 서론

스트레스는 현대 사회에서 개인적, 정신적 건강에 중요한 영향을 미치는 요인으로, 다양한 질병과 정신 건강 문제의 주요 원인으로 알려져 있다. 스트레스는 일상적인 업무와 인간관계에서부터 경제적 불안정성, 사회적 압박 등 다양한 원인에 의해 발생하며, 만성적인 스트레스는 심혈관 질환, 당뇨병, 우울증, 불안 장애 등 여러 건강 문제를 유발할 수 있다. 세계보건기구(WHO)는 스트레스를 "지속적으로 직면한 외부 요인에 의해 정신적, 신체적으로 자극을 받는 상태"로 정의하며, 스트레스가 만성화될 경우 "정신 건강의 주요 위

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1G1A1012189122268212123010 2).
*Department of Dental Hygiene, Catholic Kwandong University
**Corresponding Author : Department of Biomedical Information, Catholic Kwandong University (wlsgud0201@cku..ac.kr)
Received November 25, 2024 Revised December 14, 2024 Accepted December 19, 2024

협"이 될 수 있다고 경고하고 있다[1].

최근 연구에 따르면, 약 70% 이상의 성인이 일상에서 경험하는 스트레스를 느끼고 있으며, 스트레스가 유발하는 건강 문제는 이미 경제적 부담을 초래하고 있다. 예를 들어, 2018년 한국보건사회연구원의 보고서에 따르면, 스트레스 관련 질환의 사회경제적 비용은 연간 수십조 원에 달하는 것으로 추정된다[2][3]. 이로 인해 스트레스 관리와 예측 시스템의 개발은 개인의 건강 증진뿐만 아니라, 사회 전체의 건강 보험 비용 절감에도 중요한 역할을 할 수 있다.

스트레스 수준을 예측하기 위한 기존 연구들은 주로 생리학적 지표나 설문 조사를 기반으로 스트레스의 영향을 평가하고, 이를 진단하는 데 집중해 왔다. 예를 들어, 기존의 많은 연구에서는 심박수 변동성, 코르티솔 수치, 혈압 등 생리적 지표를 사용하여 스트레스 수준을 평가하고자 했다[4][5]. 그러나 이러한 방법들은 실시간 모니터링이 어렵고, 고정된 환경에서만 적용할 수 있어서 일상적인 생활에서의 스트레스를 정확히 평가하는 데 한계가 있었다.

최근에는 머신러닝 기법을 이용한 스트레스 예측 연구가 증가하고 있다. 예를 들어, 다양한 센서를 활용한 생체 신호 데이터를 분석하거나, 스마트폰 애플리케이션을 통해 수집된 데이터를 기반으로 스트레스를 예측하는 연구들이 있다[6]. 그러나 기존 연구들 중 상당수는 단일한 데이터를 분석하거나, 특정 실험 환경에 국한되어 연구가 진행되는 경우가 많았다. 또한, 예측 모델의 정확도나 실용성 측면에서도 개선의 여지가 많았다.

본 연구는 기존의 생리적 데이터 기반 예측을 넘어서, 나이, 키, 몸무게, 스트레스 요인 등을 포함한 다양한 개인적 특성을 종합적으로 고려한 예측 모델을 개발하고자 한다. 또한, 이진 분류 모델과 회귀 모델을 비교하여, 스트레스 수준을 정확하게 예측할 수 있는 방법을 제시하는 점에서 차별성을 가진다. 기존의 연구들은 주로 스트레스 수준을 이진 분류로 처리하거나, 생리적 데이터에 의존하는 경우가 많았으나, 본 연구는 이를 넘어서는 다양한 데이터를 기반으로 한 모델링을 통해 보다 실용적인 예측 시스템을 구축하는 데 목표를 두고 있다.

2. 연구 방법

2.1 기존 연구

김현숙 외의 연구(2021)에서는 지식근로자의 심박수와 피부 온도 등의 시계열 데이터를 랜덤 포레스트 분류(Random Forest Classifier)를 통해 스트레스 상태를 분류 예측하는 연구를 진행하였다. 정규화한 데이터를 사용한 랜덤 포레스트 모델의 정확도는 0.79로 낮은 성능을 보였다[7].

그리고 Anu Priya 외의 연구(2020)에서는 우울, 불안, 스트레스 척도 설문지(DASS 21) 및 인터뷰 등을 통해 수집된 데이터를 활용하여 로지스틱 회귀(Logistic Regression), 의사결정나무(Decision Tree), 랜덤 포레스트(Random Forest) 등의 머신러닝 알고리즘을 통해 학습했다. 그 결과, 스트레스에 대한 f1 score의 값이 랜덤 포레스트 모델에서 최대 0.711의 낮은 성능을 보였다[8].

또한, Sanchez, W 외의 연구(2023)에서는 직무 스트레스 인식을 위한 예측 모델 연구를 진행했다. 컴퓨터와 사용자의 상호작용 모니터링 앱을 개발하여 앱에서 수집된 상호작용 수와 설문지와 손목에 착용한 생체 신호 센서 등을 통해 수집한 데이터를 활용하여 머신러닝 알고리즘을 통해 예측 모델을 개발했다. 랜덤 포레스트(Random Forest)에서 F-Measurerk 0.76으로 사용한 머신러닝 모델 중 가장 좋은 성능을 보였지만 80%에 도달하지 않는 낮은 성능을 보였다[9].

이와 같은 기존 연구의 한계를 고려하여 본 연구에서는 스트레스 수준을 예측하는 두 가지 모델을 개발하는 것을 목적으로 한다. 첫 번째는 이진 분류 모델로, 스트레스 수준을 일정한 임계값을 기준으로 "높음"과 "낮음"으로 분류하여 예측하는 방법이다. 두 번째는 회귀 모델로, 스트레스 수준을 연속적인 값으로 예측하여 더 정밀한 분석을 가능하게 한다. 이 연구를 통해 스트레스 수준 예측에 대한 정확도를 향상시키고, 다양한 개인적 특성을 고려한 예측 모델을 제시하려고 한다. 또한, 두 모델의 성능을 비교하여, 특정 상황에 가장 적합한 모델을 선택할 수 있는 기준을 마련하고자 한다.

2.2 데이터 전처리

강원도특별자치도의 C 대학의 대학생 20명을 모집하여 본 연구의 목적과 과정에 대해 설명하고, 연구 참여에 대한 동의서를 받았다. 개인 정보는 익명으로 처리하고 연구 과정에서 발생할 수 있는 불편 사항은 최소화하도록 조치를 취했다. 이후 대상자들은 지각된 스트레스 척도 설문지(PSS)를 작성하였고, 수집된 데이터를 활용하여 학습 데이터로 활용했다. 측정된 데이터를 모델에 학습시키기 위해 데이터 전처리 단계에서 결측값 처리, 데이터 정규화, 그리고 스트레스 수준의 이진 분류와 회귀 모델을 위한 준비 작업을 하였다. scikit-learn의 Simple Imputer를 사용하여 결측값을 평균값으로 대체 하였으며, Standard Scalar를 사용하여 피처를 정규화하여 모델의 학습 성능을 향상시켰다. 회귀 모델과 이진 분류 모델을 위해 스트레스 수준을 각 연속 값과 이진값으로 변환하였다.

2.3 데이터 분석

아래 표 1에서는 스트레스 수준을 예측하기 위해 사용된 주요 피처와 타겟 변수 목록을 작성하였다. 피처 변수는 나이, 키, 몸무게 및 스트레스 요인들이 포함되었고, 타켓 변수는 스트레스 수준이다.

표 1. 주요 피처 및 타겟 변수
Table 1. Key Features and Target Variables

변수	설명
Age	나이(연령)
Height	키(cm)
Weight	몸무게(kg)
Stress Factor	스트레스 요인(다양한 용인에 의한 점수)
Stres s Level	스트레스 수준 (회귀 : 연속 값, 이진 분류: 높음/낮음)

2.4 모델 훈련

본 연구에서 사용한 두 가지 주요 모델은 랜덤 포레스트 분류(Random Forest Classifier)와 랜덤 포레스트 회귀(Random Forest Regressor)이다.

랜덤 포레스트 분류(Random Forest Classifier)는 여러 트리에서 따로 학습을 진행한 후, 트리의 결과들을 평균내어 하나의 결과로 출력하는 것을 말하며, 랜덤 포레스트 회귀(Random Forest Regressor)은 위와 비슷한 방법이지만, 평균적으로 계산하여 결과를 출력하는 것이 아닌 각 트리에 똑같은 질문을 넣어 결과를 확인하여 그 결과를 평균 내어 적당한 결과값을 도출하는 방법이다[10][11].

이진 분류 모델 랜덤 포레스트 분류(Random Forest Classifier)은 스트레스 수준을 높음과 낮음으로 이진화하여 예측한 스트레스 수준이 50 이상이면 '높음', 미만이면 '낮음'으로 분류한다. 데이터를 전처리 하는 과정에서 데이터정규화를 통해 값의 범위를 0~100 사이로 맞춘 뒤 중간 값인 50을 기준으로 설정했다.

회귀 모델인 랜덤 포레스트 회귀(Random Forest Regressor)은 연속적인 스트레스 수준을 예측한다. 모델을 훈련 시키기 위하여, 데이터를 훈련셋과 테스트셋으로 나누고, 각 모델을 학습시킨 후 성능을 평가한다.

3. 실험 결과

3.1 이진 분류 모델 결과

이진 분류 모델은 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 Score의 지표를 사용하여 평가하였다.

표 2. 이진 분류 모델 성능 평가 지표
Table 2. Binary Classification Model Performance Evaluation Indicators

지표	값
정확도	1.0
정밀도	1.0
재현율	1.0
F1 점수	1.0

스트레스 수준의 기준을 50으로 하여 이진화를 진행하였고, 모델이 상대적으로 명확한 두 범주로 데이터를 분류하였다. 이진 분류는 다중 클래스 분류 문제보다 예측이 간단하고, 임계값을 기준으로 분류하는 방식이 모델 학습을 용이하게 만든다. 모델에 사용된 데이터는 나이, 키, 몸무게, 스트레스 요인과 같은 개인적인 특성이 포함되어 있다. 이러한 특성은 스트레스 수준을 예측하는데 중요한 정보들을 제공하며, 모델이 각 특성에 대해 효과적으로 패턴을 학습할 수 있도록 도

와준다.

랜덤 포레스트 분류(Random Forest Classifier)는 피처 간의 상호 작용을 잘 파악할 수 있는 특징을 가지고 있다[12]. 나이, 키, 스트레스 요인 등이 상호 작용하면서 스트레스 수준을 결정하는 복합적인 관계를 잘 학습할 수 있었으며, 높은 성능을 보이는 이유 중 하나이다.

3.2 회귀모델 결과

회귀 모델에서는 평균제곱오차(MSE)와 R2 Score를 사용하여 성능을 평가하였다.

회귀 모델의 성능 평가 결과, MSE가 매우 낮고, R2 Score가 매우 높은 정확도를 보였다. 회귀 모델은 스트레스 수준을 연속적인 값으로 예측한다. 이로 인해 모델은 더 세밀한 예측이 가능하며, 각 데이터에 대한 정확한 예측값을 제공한다. 랜덤 포레스트 회귀(Random Forest Regressor)는 비선형 관계를 잘 처리할 수 있다[13]. 이는 스트레스 수준과 같은 복잡한 연속적 데이터를 예측할 때 매우 유리하다. 회귀 모델은 여러 개의 결정 트리를 결합하여 예측을 수행하며, 다양한 입력 변수들의 복잡한 관계를 잘 모델링한다[14]. 또한, 과적합 방지와 예측 정확도 향상에 유리한 특성을 가지고 있어 높은 성능을 달성할 수 있었다.

데이터 전처리 과정에서 정규화와 결측값 처리를 통해 모델의 예측 정확도를 높일 수 있었다. Standard Scaler를 이용한 피처 정규화는 모델이 다양한 크기의 데이터를 동일한 기준으로 학습할 수 있게 하였으며, 결측값을 평균값으로 대체하여 모델이 결측값에 의해 영향을 받지 않도록 하였다. 이로 인해 데이터의 품질이 향상되어 회귀 모델이 더 정확한 예측을 할 수 있었던 것으로 판단된다.

4. 결론

본 연구에서는 스트레스 수준을 예측하기 위해 두 가지 머신러닝 모델인 이진 분류 모델과 회귀 모델을 개발하고 평가하였다. 두 모델 모두 뛰어난 성능을 보였으며, 스트레스 수준 예측의 정확도를 높일 수 있는 기법을 제시하였다.

이진 분류 모델에서는 랜덤 포레스트 분류(Random Forest Classifier)를 사용하여 스트레스 수준을 두 가지 범주("높음", "낮음")로 분류하였다. 이 모델은 정확도 100%, 정밀도, 재현율, F1 점수 모두 1.0을 기록하며, 스트레스 수준을 완벽하게 분류하였으며, 스트레스 수준을 이진화하여 예측하는 데 매우 효과적이었다. 이와 같은 정확한 예측 결과는 스트레스 수준을 빠르고 명확하게 파악해야 하는 실제 환경에서 매우 유용할 것으로 판단된다.

회귀 모델에서는 랜덤 포레스트 회귀(Random Forest Regressor)를 사용하여 스트레스 수준을 연속적인 값으로 예측하였다. 이 모델은 평균 제곱 오차(MSE)가 0.00059로 매우 낮았으며, R^2 Score가 0.999로 나타나 거의 완벽한 예측 성능을 보였다.

본 연구의 의의는 스트레스 수준 예측을 위한 두 가지 모델을 비교하고, 각 모델의 성능을 실험적으로 평가하여, 스트레스 수준 예측 시스템을 개발할 수 있는 기초 자료를 제공했다는 점에 있다. 특히, 본 연구에서는 이진 분류와 회귀라는 서로 다른 예측 접근법을 사용하여, 스트레스 수준을 다양한 관점에서 예측할 수 있다는 가능성을 확인하였다. 이진 분류 모델은 간단한 환경에서 빠르고 효율적으로 스트레스 수준을 분류할 수 있는 장점을 가지며, 회귀 모델은 스트레스 수준의 미세한 변화까지 예측할 수 있어 더 세밀한 분석이 가능하다. 이 두 가지 모델을 적절히 조합하면, 실시간 스트레스 모니터링 시스템을 구축하는 데 있어 더 나은 성능을 기대할 수 있을 것이다.

본 연구에서 사용한 데이터셋은 소규모의 대상자를 대상으로 했으며, 나이와 특성 등의 제한적인 데이터셋이다. 이로 인한 이진 분류 모델과 회귀 모델의 과적합이 되었을 가능성을 배제할 수 없다. 따라서 향후 연구에서는 심리적 요인과 외부 환경 변수를 추가하여 스트레스 수준을 더 정밀하게 예측할 수 있는 방법을 모색하고 폭 넓은 표본을 모집하여 추가적인 연구를 진행할 필요가 있다. 또한, 실시간 데이터를 기반으로 한 스트레스 모니터링 시스템을 개발하여, 개인의 스트레스 수준을 지속적으로 추적하고 실시간으로 대응할 수 있는 시스템을 구축하는 연구가 이루어질 수 있다. 이를 통해 스트레스 관리의 효율성을 높이고, 정신 건강

을 개선하는 데 기여할 수 있을 것이다.

REFERENCES

[1] Pruessner, J. C., Kirschbaum, C., Meinlschmid, G., & Hellhammer, D. H. (2003). Two formulas for computing the area under the curve represent measures of total hormone concentration versus time-dependent change. Psychoneuroendocrinology, 28(7), 1167-1174.

[2] Stress-related factors in middle-aged and older adults in Korea: The 8th National Health and Nutrition Examination Survey, 2020

[3] Korea Institute for Health and Social Affairs. (2018). A study to estimate the economic burden of stress-related diseases. Korea Institute for Health and Social Affairs.

[4] Dongsu Kim, Yeonsu Jeong and Sekwon Park. (2004). Relationship between the stress hormone salivary cortisol and self-reported stress scale scores. Journal of the Korean Psychological Association: Health, 9(3), 633-646.

[5] Thayer, J. F., & Lane, R. D. (2009). Claude Bernard and the heart-brain connection: Further elaboration of a model of neurovisceral integration. Neuroscience & Biobehavioral Reviews, 33(2), 81-88.

[6] Koch, T., O'Hara, R., & Lanza, S. (2018). Predicting stress using wearable sensors and smartphone data. Journal of Medical Internet Research, 20(9), e10904.

[7] HyunSuk Kim, et al. (2021). A study on the Stress State Classification Method Using Random Forest Method. Proceedings of Symposium of the Korean Institute of communications and Information Sciences,

[8] Priya Anu,Garg Shruti,Tigga Neha Prerna. (2020). Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. Procedia Computer Science : 1258-1267.

[9] Sanchez, W., Martinez, A., Hernandez, Y. et al. (2023). A predictive model for stress recognition in desk jobs. J Ambient Intell Human Comput 14, 17-29

[10] Joon-Young Park, Myung-so Chae, & Sungkwan Jung. (2016). Proposal of Random Forest Algorithm-Based Crime Type Classification Model and Monitoring Interface Design Elements for Real-Time Crime Prediction. Journal of the Practical Journal of Computing of the Society for Information Science, 22(9), 455-460.

[11] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818.

[12] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[13] Louppe, G., & Wehenkel, L. (2013). Random forests for classification and regression. In Proceedings of the 3rd European Symposium on Computer and Communications Engineering. Springer.

[14] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

[15] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101.

저자약력

정 진 형 (Jin-hyoung Jeong) **[정회원]**

- 2012년 2월 : 가톨릭관동대학교 의료공학과 졸업(학사)
- 2014년 2월 : 가톨릭관동대학교 일반대학원 졸업(공학석사)
- 2017년 8월 : 가톨릭관동대학교 일반대학원 졸업(공학박사)
- 2017년 9월 ~ 2021년 2월 : 가톨릭관동대학교 초빙교수
- 2021년 3월 ~ 현재 : 가톨릭관동대학교 의료IT학과 조교수
- 2024년 3월 ~ 현재 : 가톨릭관동대학교 의료경영학과 조교수

〈관심분야〉 의료 시스템, 데이터 분석, 통신, 인공지능

정 유 연 (Jung YuYeon) **[정회원]**

- 2014년 충북대학교 이학박사
- 2018년 단국대학교 구강보건학박사
- 2020년~ 현재 가톨릭관동대학교 치위생학과 교수, 학과장 치위생 석사과정 주임교수

〈관심분야〉 예방치과학