ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://dx.doi.org/10.24225/jkaia.2023.1.1.17

머신러닝 데이터의 우울증에 대한 예측
Prediction of Depression from Machine Learning Data

Jeong Hee KIM[1], Kyung-A KIM[2]

Received: April 19, 2023. Revised: May 28, 2023. Accepted: June 29, 2023.

Abstract

The primary objective of this research is to utilize machine learning models to analyze factors tailored to each dataset for predicting mental health conditions. The study aims to develop appropriate models based on specific datasets, with the goal of accurately predicting mental health states through the analysis of distinct factors present in each dataset. This approach seeks to design more effective strategies for the prevention and intervention of depression, enhancing the quality of mental health services by providing personalized services tailored to individual circumstances. Overall, the research endeavors to advance the development of personalized mental health prediction models through data-driven factor analysis, contributing to the improvement of mental health services on an individualized basis.

Keywords : Depression#1, Machine Learning #2, Algorithm #3, Prediction#4

Major Classification Code : Machine Learning, Artificial Intelligence, Analysis

1. Introduction

현재 대한민국은 해마다 우울증 환자가 증가하는 추세이며 작년에는 코로나 19 감염병의 확산으로 코로나 블루로 인한 우울증이 더해져 우울증을 겪는 사람들이 더 늘었다.이러한 우울증 문제는 국내 뿐만 아니라 전 세계적으로 급증하고 있다.

또 다른 사회적 문제는 고령화이다. 우리나라 고령화 속도는 OECD 평균에 비해서 2 배 이상 빠르게 진행되고 있다. 통계청 자료에 따르면 22 년 65 세 고령인구는 우리나라 인구의 17.5%로 향후 계속 증가하여 25 년에는 20.6%로 우리나라가 초고령 사회에 진입할 것으로 전망된다. 급증하는 급격하게 변하는 도시에 적응하지 못하고 농촌으로 이동하게 되고 사회적 배제에 따른 우울 수준 또한 높아진다.

2. Body

이에 본 연구는 도시민들이 아닌 농촌에서 거주하는 사람들을 중점적으로 우울증 분석을 할 예정이다. 이를 통해 심층적인 이해를 도모하고 우울증의 조기 개입을 향상시키는 방안을 모색하며, 머신 러닝 모델을 활용해

1 First Author. Undergraduate student, Bigdata medical convergence, Bio-convergence, Eulji University, South Korea. Email: 2021158009@g.eulji.ac.kr
2 Corresponding Author. Professor, Bigdata medical convergence, Bio-convergence, Eulji University, South Korea. Email: kyungakim@naver.com

우울증에 영향을 미치는 다양한 요인을 조사하고 우울증 예방과 지원에 기여할 것이다. 또한 우울증과 관련된 중요한 문제에 대한 인사이트를 제공하며 비슷한 지역사회의 정신 건강 서비스 향상을 지원하고자 한다.

2.1. Variable Selection

해당 연구 자료는 케냐 서부 빅토리아 호수 근처 시아야 카운티 시골 지역의 부사라 센터에서 실시한 2015 년 자료 조사 데이터이다.

기존 데이터 파일에서 높은 정확도 예측을 위해 정확하게 변수들이 정의된 것과 우울증에 주요 원인으로 다뤄지는 변수들만 선택한 결과는 Table 1 과 같다. 총 23 개와 1432 개의 값으로 막 변수 depressed 로 우울증 진단 여부를 판단하며 0 일 경우 우울증을 진단받지 않았다고 판단하는 반면 셀 값이 1 인 경우 우울증을 진단 받았다고 판단한다.

Table 1: data variable

변수	변수 설명
Survey_id	Individual Identifier
Ville_id	Village Identifier
sex	Men 1 / women 0
age	Age(respondent)
married	Marital status
Number_children	Number of children
Education_level	Years of education completed
Total_member	Household size
Gained_asset	Value of gained goods
Durable_asset	Value of durable goods
Save_asset	Value of savings
Living_expenses	Total expenses, monthly
other_expenses	Medical, food, etc
Incoming_salary	Wage labor primary income
incoming_own_farm	Own farm primary income
Incoming_business	Non-ag business primary income
Incoming_no_business	Non-agricultural business owner
Incoming_agricultural	agricultural revenue, monthly
Farm_expenses	Farm flow expenses, monthly
labor_primary	Casual or Wage Labor Primary Source of Income
lasting_investment	continuous investment
no_lasting_investment	nonessential
depressed	Meets epidemiological threshold for moderate depression

2.2. Data Preprocessing

모델링 분석에 앞서 각 데이터의 밀도를 분석해보자면 자산은 정규분포 형태를 보여주고 있는 반면 연령, 아동, 지속적인 투자 변수 등은 특정 수치만 집중적으로 높은 것으로 보인다. 그 중 자산의 경우 기본적으로 가지고 있는 자산보다 지속적으로 수익을 얻는 내구적인 경우에 더 우울증 지수가 높게 나타난다. 또한 이외의 농업을 하면서 들어오거나 나가는 지출이나 수입의 경우 전체적으로 2.5 에서 3.5 정도 되는 비율의 수치를 가지면 상대적으로 우울증에 걸리는 비율이 낮은 것으로 확인된다.

이어서 앞의 결과를 바탕으로 모델링을 진행하기 전 모델의 성능과 정확도를 높이기 위해서 전처리 과정에서 min-max 스케일링과 K-fold 교차검증을 활용한다.

Min-max 스케일링은 데이터를 일정 범위로 조절해서 서로 다른 특성과 스케일을 가진 데이터를 표준화 하여 모든 특성을 동일한 가중치로 고려해 모델 학습과 수렴을 개선한다.

K-fold 교차검증은 데이터를 k 개 부분으로 나누어 모델을 k 번 학습하고 평가하는 것으로 해당 연구에서는 k 를 5 로 설정했다.

3. Research Methods

3.1. Models

앞서 전처리를 한 결과를 바탕으로 다양한 머신 러닝 및 딥러닝 모델을 고려해서 주어진 데이터와 연구 목표에 적합한 모델을 선정과 해당 모델에서 변수의 중요도를 파악한다.

이를 위해 학습에 사용될 모델은 DT(Decision Tree)모델, RF(Random Forest)모델, NB(naïve bayes), SVM(Support Vector Machine)모델, KNN(K-Nearest Neighbors)모델, DL(Deep Learning)모델, RF(Random Forest)모델로 총 7 가지의 여러 모델의 성능평가를 진행한다. 각 모델 분석 방식에 대한 설명은 아래와 같다.

먼저 Decision tree 모델의 경우 각 노드의 특성을 기반으로 데이터를 분할하는 것으로 각 분할은 정보 이득 또는 지니 불순도를 최소화하도록 선택해 분류 모델을

구축한다. Random Forest 모델은 의사결정 트리의 앙상블 모델로 데이터를 분할하고 판단하며 과적합을 줄이고 안정적인 예측을 한다. 또한 Navie Bayes 모델의 경우 특성들이 서로 독립적이라고 가정해서 클래스에 대한 확률을 계산해 분류를 수행한다. SVM 모델은 두 클래스를 분리하는 초평면을 찾아 결정 경계 사이의 거리를 최대화하기 위해 가중치와 편향을 조절하고 Deep Learning 모델은 입력층, 은닉층, 출력층이 있고 각 뉴런은 가중치와 활성화 함수를 통해 입력 신호를 처리한다. 이어 K-Nearest Neighbors 모델은 새로운 데이터 포인트가 주어지면 가장 가까운 K 개의 이웃을 찾고 이 이웃들의 다수결 또는 평균을 통해 예측한다. 마지막으로 Logistic Regression 모델은 입력과 가중치를 조합해서 값 계산한 후 로지스틱 함수 또는 시그모이드 함수를 사용해서 선형 조합을 확률 값으로 변환한다.

3.2. Model Performance Analysis

TP(True positive) = 모델이 실제로 양성 클래스를 양성으로 정확하게 예측한 경우의 수

FN(False Negative) = 모델이 실제로 음성 클래스를 정확하게 예측한 경우의 수

FP(False Positive) = 모델이 실제로 음성 클래스를 양성으로 잘못 예측한 경우의 수. 양성 클래스를 "오검출" 경우의 수

TN(True Negative)= 모델이 실제로 양성 클래스를 음성으로 잘못 예측한 경우의 수. 양성 클래스를 "미검출" 한 경우의 수

이를 활용하여 계산한 수식은 다음과 같다.

$$\text{Accuracy}=\frac{Sum\ of\ diagonals(TP)}{Total\ number\ of\ instnance} \quad (1)$$

$$\text{Percision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall=Sensitivity}y=\frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity}=\frac{TN}{TN+FP} \quad (4)$$

$$\text{F1 Score}=\frac{2*(Precision*Recall)}{Precision+Recall} \quad (5)$$

각 모델의 정확도 예측하기 위해 각 혼동행렬의 정확도 및 오류율, 정밀도, 재현율 및 특이성을 계산하는데 사용한다.

3.3. Model Comparison

일반적으로 F1-score 가 0.7 이상을 좋은 성능이라고 평가하고 0.5 에서 0.7 사이는 어느 정도 균형된 성능 그리고 0.5 미만일 경우 성능이 떨어진다고 판단한다. 이를 바탕으로 Table 2 확인해본 결과 Deep Learning 과 Logistic Regression 을 제외하고는 0.5 이하의 수치로 성능이 낮다고 판단된다.

Table 2: Model performance indicators

	depressed	precision	recall	F1-score	support	Confusion matrix
Decision Tree	0	0.85	0.85	0.85	241	[204 37 39 2]
	1	0.14	0.15	0.14	41	
Random forest	0	0.86	0.99	0.92	241	[239 2 40 1]
	1	0.33	0.02	0.05	41	
Naive Bayes	0	0.86	0.95	0.90	241	[229 12 37 4]
	1	0.25	0.10	0.14	41	
Svm	0	0.85	1.00	0.92	241	[241 0 41 0]
	1	0.00	0.00	0.00	41	
Deep Learning	0	0.86	1.00	0.92	241	[240 1 39 2]
	1	0.67	0.05	0.09	41	
K-Neighbors	0	0.86	0.98	0.91	241	[235 6 39 2]
	1	0.25	0.05	0.08	41	
LogisticRegression	0	0.85	1.00	0.92	241	[235 0 41 0]
	1	0.53	0.07	0.11	41	

이어 우울증 진단 문제 정확하게 해결하기 위해 데이터 파일과 모델과의 정확도도 추출한 결과는 Table 3 과 같다. F1-score 가 전체적으로 낮은 반면 정확도를 돌린 결과 Decision Tree 를 제외하고 나머지 모델은 0.8 이상의 값으로 해당 데이터 셋에 적합한 모델이다.

Table 3: Model accuracy

	Accuracy
Decision Tree	0.71
Random forest	0.85
Naive Bayes	0.83
Svm	0.85
Deep Learning	0.85
K-Neighbors	0.84
Logistic Regression	0.85

앞의 자료를 바탕으로 클래스 간 불균형이나 특정 클래스의 중요성 등을 고려하기 위해 ROC 커브를 돌려본 결과 정확도에서는 대부분 0.8 이상의 수치가 나온 반면 ROC 커브에서는 0.5 최댓값이 Decision Tree 가 0.53 으로 가장 낮고 Random Forest 를 제외하고는 대부분 0.4 의 수치를 보인다.

4. Conclusion

정확도 측면에서는 해당 연구에서 사용한 데이터는 정확도가 0.9 이하로 높은 정확도는 보이지 않고 0.85 중간 정도의 수치를 보이고 있지만 ROC 커브를 활용하여 분석해본 결과 0.5 보다 낮은 수치로 해당 모델들이 적합하지 않다고 판단된다.

F1-score 나 ROC 커브의 정확도 값은 낮게 나왔지만 본 연구의 목적은 7 개의 모델 중 해당 데이터 셋에 적합한 모델을 찾는 것이기 때문에 변수들 간의 자세한 관계 파악과 객관적인 평가를 위해 Learning Curve(학습 곡선)을 이용하여 추가적인 분석을 해본 결과는 다음과 같다.

Learning Curve 에 적용할 모델은 정확도 분석 당시 같은 수치를 보였던 Random Forest, SVM, Deep Learning, Logistic Regression 으로 총 4 개의 모델이다.

Learning Curve 는 모델의 과소적합 또는 과적합을 진단하며 훈련 데이터와 검증 데이터 간의 성능 차이를 확인할 수 있다. 또한 데이터 양이 증가함에 따라 모델의 성능이 어떻게 변하는지 확인함으로써 일반화 성능을 시각적으로 비교하여 어떤 모델이 더 나은 성능을 보이는 판단할 수 있다.

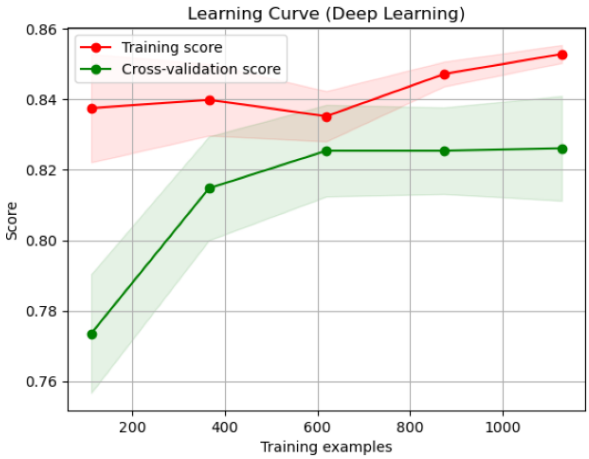

Figure 1: Deep Learing Curve

다음 Figure1 의 Deep Learning 의 Learning Curve 를 돌린 결과이다.

Training Score(훈련점수)의 경우 Cross-Validation(교차검증)보다 높으면서 두 그래프 사이의 간격이 큰 것을 확인할 수 있다. 이를 통해 모델이 훈련 데이터와 검증 데이터 간에 성능 차이가 크다는 것을 나타내는 것으로 판단할 수 있다. 특히 Training Score 의 성능은 높지만 Cross-Validation 가 검증 데이터에 대한 성능이 낮기에 훈련 데이터에 맞춰져서 새로운 데이터에 일반화하지 못하는 과대적합 상태라고 볼 수 있다.

이어 Random Forest 모델의 Learning Curve 를 시각화해 본 결과 Deep Learning 모델과 동일한 형태로 나왔으며 오히려 Random Forest 모델의 Training Score 과 Cross-Validation 의 간격이 큰 것을 확인할 수 있었다. 이에 Random Forest 모델도 과적합 상태라고 판단할 수 있다.

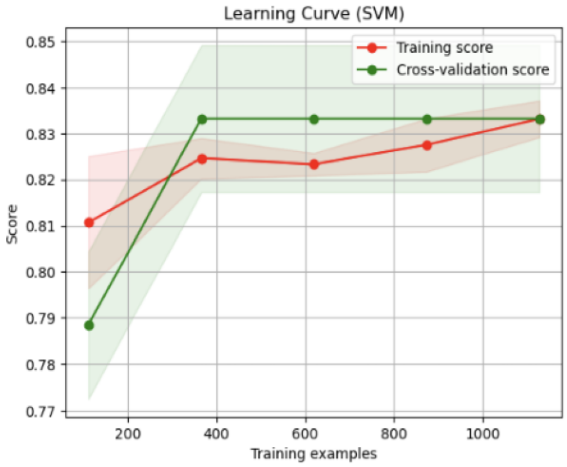

Figure 2: SVM Curve

세 번째 Learning Curve 는 SVM 모델로 시각화한 결과는 Figure 2 과 같다.

SVM 은 Deep Learning 과 Random Forest 과 동일하게 두 그래프의 간격이 멀어져 있는 것을 확인할 수 있는데 이전 모델과는 달리 Training Score 가 Cross-Validation Score 보다 낮다. 이는 모델의 훈련 및 검증 데이터의 성능이 낮아 성능이 수렴하지 않는 것으로 해당 모델이 간단하거나 복잡한 데이터 패턴을 잡아내지 못하는 과소적합인 상태로 볼 수 있다.

반면 로지스틱 회귀의 경우 Figure 3 와 같이 두 그래프의 간격이 크지 않고 서로 유사하게 높은 값을 가지고 있다. 이를 통해 해당 연구 데이터에 적절한 일반화가 되어있으며 안정적인 모델 학습으로 훈련

데이터의 잡음이나 특이점이 크게 반응하지 않는 것을 알 수 있다. 따라서 과대적합 또는 과소적합 상태가 아니기에 다른 모델에 비해 해당 데이터에 Logistic Regression 이 적합한 모델이라고 판단된다.

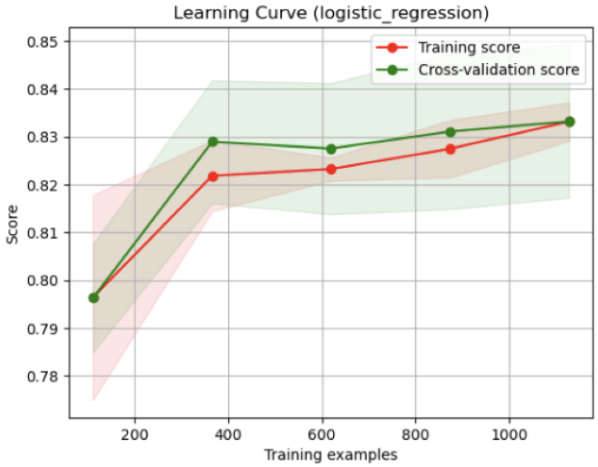

Figure 3: Logistic Reression Learning Curve

최종 연구 결과를 고려하여 7 개의 모델 중 Learning Curve 를 통해 적합한 모델을 선정했다. 하지만 기존에 MIN-MAX 스케일링과 K-Fold 교차 검증 등 전처리 작업을 진행하였음에도 ROC 커브나 F1-Score 가 낮은 정확도로 나타났다.

주된 이유로는 depressed 에서 우울증에 진단받지 않은 사람이 1174 명이고 우울증에 진단 받은 사람이 235 명으로 확률로 약 7 배 차이가 난다는 점이다.

따라서 향후 연구에서는 더 높은 품질의 데이터를 획득하고 이를 기반으로 연구를 수행할 경우 보다 정확하고 신뢰할 수 있는 예측 결과를 얻을 것으로 예상되며 이러한 노력을 통해 연구의 신뢰성을 높일 것으로 기대된다.

References

Borkowska, A., & Rybakowski, J. K. (2001). Neuropsychological frontal lobe tests indicate that bipolar depressed patients are more impaired than unipolar. *Bipolar disorders, 3*, 88-94.

Datahunt. (2023, November 10). “F1 Score and Machine Learning - Definition, Principles, Calculation, Limitations, and Mitigation Strategies”. Retrieved from https://www.thedatahunt.com/trend-insight/f1-score

Garga, S., Priyaa, A., & Tiggaa, N. P. (Year not provided). Predicting Anxiety, Depression, and Stress in Modern Life using Machine Learning Algorithms.

Kim, S., Kim, M., & Kim, H. (2019). “Impact of social exclusion on depression in rural elderly population”. *Rural Economy*, *2*, p106.

Kim, Y., Kim, J., Woo, G., Kim, H., & Park, H. (2020). Machine learning techniques for predicting depression based on lifestyle patterns using NHANES data. *Ming Ji University*, 0720.

Lim, J. H., & Lim, D. O. (2022). Prediction of depression in the elderly using a Wide & Deep Learning Model. Retrieved from https://e-jhis.org/m/journal/view.php?number=797

Robinson, L. J., & Ferrier, I. N. (2006). Evolution of cognitive impairment in bipolar disorder: a systemic review of cross-sectional evidence. *Bipolar disord. 8*, 103-116.

Statistics Korea. (2023, November 03). "Results of the 2021 Agricultural, Forestry and Fishery Census". Retrieved from https://kostat.go.kr/board.es?mid=a10301080500&bid=226&act=view&list_no=417699&tag=&nPage=1&ref_bid=

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490-499.

Simonsen, C., Sundet, K., & Vaskinn, A. (2008). Neurocognitive profile in bipolar I and bipolar.II disorder: differences in pattern and magnitude of dysfunction. *Bipolar Disord, 10*, 245-255.