광역시·도민의 우울경험에 대한 Random Forest 비교분석

이동수*, 김유정O
*세한대학교 인공지능빅데이터학과,
O조선간호대학교 간호학과
e-mail: dslee@sehan.ac.kr*, fight1004@cnc.ac.krO

Comparative analysis of random forest on depression experiences of metropolitan and provincial residents

Dong Su Lee*, Yu Jeong KimO
*Dept. of Artificial Intelligence & Bigdata, Sehan University,
ODept. of Nursing Science, Chosun Nursing College

● 요 약 ●

본 연구는 광역시와 광역도 간의 개인적 요인과 건강수준 정도가 우울경험 여부에 영향을 미치는 변수의 중요도를 파악하고자 시도되었다. 본 연구의 자료는 질병관리청의 2021년 지역사회건강조사 데이터를 활용하였다. 광역시의 데이터는 4,602건을 이용하였고, 광역도는 19,545건의 데이터를 이용하였다. 자료 분석에 활용된 빅데이터는 R 4.3.0 for Windows를 활용하여 단어 빈도 분석과 machine learning기법인 **Random Forest**분석을 실시하였다. 연구결과, train 데이터와 test 데이터의 과적합(overfitting)의 문제는 발생하지 않았으며, machine learning 기법의 분류모델은 약 94% 수준으로 나타났다. 분석 결과 광역시와 광역도 간의 우울경험여부에 미치는 중요도가 각각 다르게 나타났다. 두 지역의 시민에게 미치는 우울경험의 원인을 다르게 접근함으로써 보다 더 효율적인 정책수립이 가능 할 것으로 판단된다.

키워드: 우울(depression), 머신러닝(machine learning), 랜덤포레스트(random forest)

I. Introduction

1. Purpose

국가차원에서 보건의료를 통한 국민 건강을 예방하는 것은 매우 중요하다. 특히 지역에 따라 사회 환경과, 보건의료 환경이 각각 다르기 때문에 지역사회 건강은 또 다른 접근이 필요하다. 지역사회건강조사는 「지역보건법」 제4조(지역사회 건강실태조사)에 의거하여 실시하는 통계청 승인 일반통계이다. 지역보건의료계획을 수립 및 평가하고, 조사수행 체계를 표준화하여 비교 가능한 지역건강 통계를 생산하고자 2008년부터 매년 전국 보건소에서 실시하고 있다.

본 연구는 지역에 따라 우울경험 여부를 개인적 및 건강수준 특성에 따라 정확하게 진단 과 예측하기 위하여 기계학습(Machine Learning)을 이용한다. 대표적인 기계학습 분석 중에서 랜덤포레스트를 적용하여 광역시와 광역도 간의 우울경험 여부 예측모형을 개발과 이를 바탕으로 예방교육프로그램 개발을 위한 기초자료를 제공하고자 한다.

2. Research question

본 연구는 광역도시와 광역도 간의 지역사회 건강 정도의 비교분석을 통해 정책반영의 기초자료로 활용하기 위함이며 구체적인 연구문제는 다음과 같다.

연구문제 1: 광역도에 거주하는 도민의 우울경험여부에 미치는 주요 변수는 어떻게 다른가?

연구문제 2: 광역시에 거주하는 시민의 우울경험여부에 미치는 주요 변수는 어떻게 다른가?

연구문제 3: 광역시와 광역도민의 우울경험여부에 미치는 주요 변수는 차이가 있는가?

II. Research Method

1. Research Model

본 연구에서는 개인적 및 건강준의 특성변수에 따라 우울경험 여부가 분류되는지를 규명하기 위해 여러 변수를 설정하였다. 여러

원인변수들의 특성이 우울경험 정도에 따라 다르게 분류되는지를 분석하였다. 분류분석 위해 설정한 13가지 원인변수는 연령, 교육수준, 가구연소득, 하루평균수면시간, 걷기실천일수, 행복감지수, BMI지수, 운동능력, 자기관리, 일상행동, 통증불편, 주관적 스트레스, 주관적 건강이었다. 종속변수는 우울경험여부이다. 원인변수가 결과변수에 어느 정도 영향관계가 있는지 정도를 분석하고, 광역도와 광역시 간의 원인변수들의 영향력의 차이가 있는지 알아보고자 Fig. 1. 과 같이 연구모형을 설정하였다.

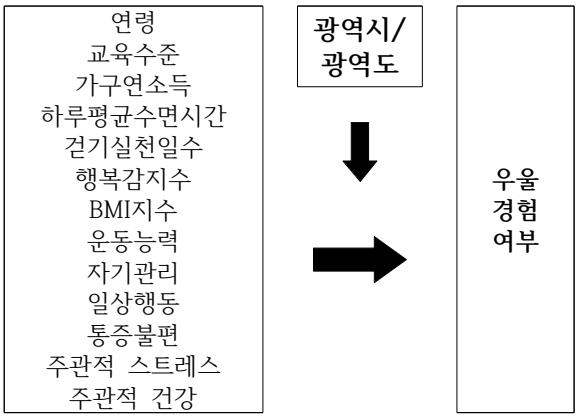


Fig. 1. Research Model

2. Data Anaysis

랜덤포레스트(random forest, RT)는 의사결정트리의 기법의 모델을 반복하는 앙상블(ensemble) 학습기법의 일종이다. RT는 일반화에 문제가 발생할 수 있는 과적합(overfitting) 문제를 해결할 수 있다. Fig. 1과 같이 주어진 데이터로부터 여러 개의 모델을 학습한 다음, 여러 모델의 예측 결과들을 종합해서 예측하여 정확도를 높이는 기법이다.

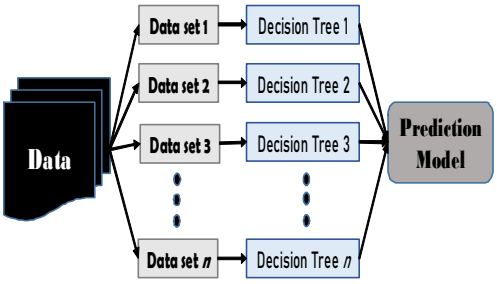


Fig. 2. Random Forest

본 연구는 질병관리청에서 제공한'2021 지역사회건강조사' 데이터를 활용하였다. 본 연구에 사용된 데이터는 질병관리청의 지역사회건강조사 2021년 자료이다. 전라남도 자료는 19,545건이며, 광주광역시의 자료는 4,602건이다. 수집된 자료는 빅데이터 분석도구인 R 통계패키지(ver. 4.3.0)를 사용하여 분석하였다.

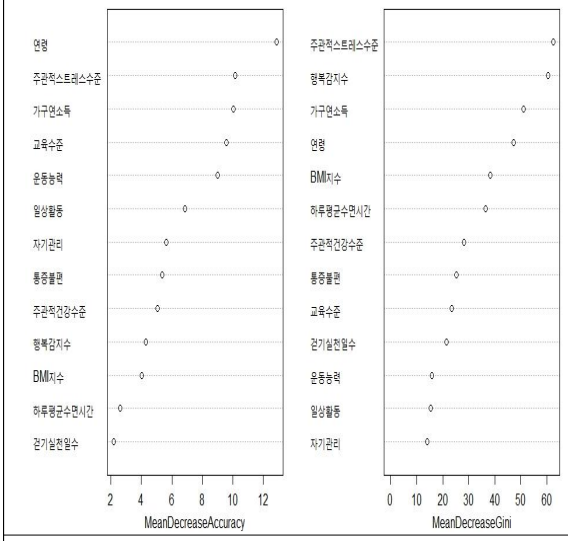
III. Results

1. Analysis of provincial residents

Table 1. Analysis Result of provincial residents

Train data

```
> randomForest::importance(train.rf1)
                  있음        없음      MDA      MDG
연령          -8.4124534 13.2344941 12.864980 47.27724
가구연소득     1.2428521  9.3250822 10.036851 50.99359
주관적건강수준 4.0864268  3.4963899  5.046540 28.08268
걷기실천일수  -0.6039594  2.7904070  2.173416 21.42132
BMI지수       -0.5992560  4.5469160  4.012784 38.24142
하루평균수면시간 3.8235409 0.9005307 2.596393 36.42294
주관적스트레스 16.0750409 2.3067933 10.153127 62.54274
행복감지수    11.4481009 -2.6745578  4.294817 60.50081
교육수준      -6.6531769 10.2326862  9.571783 23.45216
운동능력      -1.1313241  8.7656200  9.003319 15.92090
자기관리       0.4098199  5.2189432  5.644905 14.07372
일상활동       0.1024941  6.4623120  6.870817 15.35509
통증불편       3.3066497  3.9006902  5.373405 25.23191
```




```
Reference
Prediction  있음  없음
      있음     2     1
      없음   858 12939
Accuracy : 0.9378
95% CI : (0.9336, 0.9417)
```

Test data

```
Reference
Prediction  있음 없음
      있음    1    0
      없음  369 5375
Accuracy : 0.9358
95% CI : (0.9291, 0.942)
```

MDA: MeanDecreaseAccuracy
MDG: MeanDecreaseGini

광역도의 시민의 우울경험 여부에 대한 RF기법 분석결과 Table 1.과 같다. 랜덤포레스트(RF)로 우울경험 여부에 대해 분류를 하기 위하여 전체 데이터를 학습용 데이터와 검증용 데이터로 구분하였다. 결측값을 제외한 전체 데이터 개수는 19,545개였으며, 이중 학습용 데이터는 전체 데이터의 2/3 (13,800개), 검증용 데이터는 전체 데이터의 1/3(5,745개)로 구분하였다.

우울경험 여부에 대한 학습용 데이터의 정분류율은 93.78%(0.9378)로 나타났다. 우울경험 여부에 대한 검증용 데이터의 정분류율은 93.58%(0.9358)로 나타났다. 모델에 대한 분류율의 차이는 0.2%로 과적합의 문제는 없는 것으로 판단할 수 있다.

우울경험 여부에 대한 변수들의 분류율에 영향을 미치는 중요도는 정확도계수와 지니계수를 분석하였다. 학습용 데이터의 정확도계수에서 중요도는 연령(12.86), 주관적 스트레스(10.15), 가구연소득(10.03), 교육수준(9.57) 순으로 나타났다. 학습용 데이터의 지니계수에서 중요도는 주관적 스트레스(62.54), 행복감지수(60.50), 가구연소득(50.99), 연령(47.27) 순으로 나타났다. 우울경험 여부에 대한 분류율에 영향을 미치는 중요 변수는 연령, 주관적 스트레스, 가구연소득, 교육수준, 행복감지수 등으로 분석되었다.

2. Analysis of metropolitan residents

광역시의 시민의 우울경험 여부에 대한 RF기법 분석결과 Table 2.와 같다. 랜덤포레스트(RF)로 우울경험 여부에 대해 분류를 하기 위하여 전체 데이터를 학습용 데이터와 검증용 데이터로 구분하였다. 결측값을 제외한 전체 데이터 개수는 4,602개였으며, 이중 학습용 데이터는 전체 데이터의 2/3 (3,261개), 검증용 데이터는 전체 데이터의 1/3(1,341개)로 구분하였다.

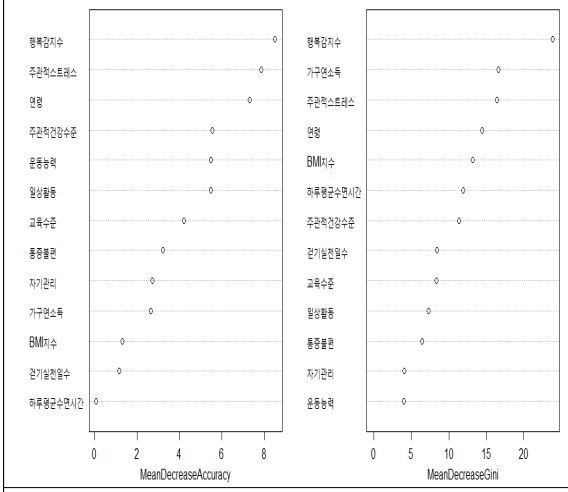
우울경험 여부에 대한 학습용 데이터의 정분류율은 93.78%(0.9374)로 나타났다. 우울경험 여부에 대한 검증용 데이터의 정분류율은 94.03%(0.9403)로 나타났다. 모델에 대한 분류율의 차이는 0.29%로 과적합의 문제는 없는 것으로 판단할 수 있다.

우울경험 여부에 대한 변수들의 분류율에 영향을 미치는 중요도는 정확도계수와 지니계수를 분석하였다. 학습용 데이터의 정확도계수에서 중요도는 행복감지수(8.51), 주관적 스트레스(7.85), 연령(7.30), 주관적 건강 수준(5.56) 순으로 나타났다. 학습용 데이터의 지니계수에서 중요도는 행복감지수(23.80), 가구연소득(16.58), 주관적 스트레스(16.43), 연령(14.46) 순으로 나타났다. 우울경험 여부에 대한 분류율에 영향을 미치는 중요 변수는 행복감지수, 주관적 스트레스, 주관적 건강 수준, 가구연소득, 연령 등으로 분석되었다.

Table 2. Analysis Result of metropolitan residents

Train data

```
> randomForest::importance(train.rf1)
                  있음       없음       MDA       MDG
연령          -2.613077  7.4980652 7.30422568 14.464240
가구연소득     5.438322  0.4637365 2.64622677 16.588722
주관적건강수준 2.341310  5.0516120 5.56740832 11.384114
걷기실천일수   1.084194  0.6976238 1.16960957  8.408416
BMI지수        1.250439  0.7824001 1.31967606 13.161671
하루평균수면시간 4.192877 -2.2830282 0.08410997 11.89049
주관적스트레스 8.224622  4.9231188 7.85645040 16.430969
행복감지수    11.675157  2.7496196 8.51052373 23.802369
교육수준      -3.015657  4.7603614 4.21024114  8.347875
운동능력       1.166784  5.0356694 5.49057599  3.996727
자기관리       2.307931  1.8580670 2.72251605  4.085237
일상활동       4.537909  4.1593334 5.48853524  7.291933
통증불편       5.681018  0.1019608 3.21478744  6.490636
```




```
Reference
Prediction 있음 없음
      있음    5    2
      없음  202 3052
Accuracy : 0.9374
95% CI : (0.9286, 0.9455)
```

Test data

```
Reference
Prediction 있음 없음
      있음    1    0
      없음   80 1260
Accuracy : 0.9403
95% CI : (0.9263, 0.9524)
```

MDA: MeanDecreaseAccuracy
MDG: MeanDecreaseGini

IV. Conclusions

본 연구는 우울경험 여부를 진단하고 예측하기 위하여 인공지능 방법인 랜덤포레스트 기법 분석방법으로 광역도와 광역시의 차이를 파악하고자 시도되었다. 또한 개인적 및 건강수준의 특성변수 중에서 우울경험의 분류율에 영향을 미치는 중요 변수가 무엇인지를 규명하였다.

연구 결과, Table 3.과 같이 광역도와 광역시 간의 우울경험에 영향을 미치는 변수가 다르게 나타났다. 광역도의 경우 연령이 매우

중요한 변수로 도출이 되었고, 광역시의 경우는 행복감지수가 매우 중요한 변수로 나타났다.

특히, 지역사회 환경이 다른 상황에서 시민의 건강을 예방하기 위한 활동은 연구결과를 중심으로 판단할 때 접근이 다르게 하는 것이 중요하다고 본다.

Table 3. Comparison of Random Forest Analysis Results

구분	광역도		광역시	
계수	정확도	지니	정확도	지니
영향요인 우선순위	연령	주관적 스트레스	행복감지수	행복감지수
	주관적 스트레스	행복감지수	주관적 스트레스	가구연소득
	가구연소득	가구연소득	연령	주관적 스트레스
	교육수준	연령	주관적 건강	연령
	운동능력	BMI지수	운동능력	평균수면 시간

REFERENCES

[1] D. S. Lee, *"Comprehension of Bigdata and R"* Free Academy-Press, pp. 157-235. 2022.