Print ISSN: 2233-4165 / Online ISSN 2233-5382
JIDB website: http://www.jidb.or.kr
Doi:http://dx.doi.org/10.13106/jidb.2020.vol11.no7.29

A Study on the Prediction Model of the Elderly Depression

Beom-Seok SEO[1], Eung-Kyo SUH[2], Tae-Hyeong KIM[3]

Received: May 16, 2020. Revised: July 01, 2020. Accepted: July 05, 2020

Abstract

Purpose: In modern society, many urban problems are occurring, such as aging, hollowing out old city centers and polarization within cities. In this study, we intend to apply big data and machine learning methodologies to predict depression symptoms in the elderly population early on, thus contributing to solving the problem of elderly depression. **Research design, data and methodology:** Machine learning techniques used random forest and analyzed the correlation between CES-D10 and other variables, which are widely used worldwide, to estimate important variables. Dependent variables were set up as two variables that distinguish normal/depression from moderate/severe depression, and a total of 106 independent variables were included, including subjective health conditions, cognitive abilities, and daily life quality surveys, as well as the objective characteristics of the elderly as well as the subjective health, health, employment, household background, income, consumption, assets, subjective expectations, and quality of life surveys. **Results:** Studies have shown that satisfaction with residential areas and quality of life and cognitive ability scores have important effects in classifying elderly depression, satisfaction with living quality and economic conditions, and number of outpatient care in living areas and clinics have been important variables. In addition, the results of a random forest performance evaluation, the accuracy of classification model that classify whether elderly depression or not was 86.3%, the sensitivity 79.5%, and the specificity 93.3%. And the accuracy of classification model the degree of elderly depression was 86.1%, sensitivity 93.9% and specificity 74.7%. **Conclusions:** In this study, the important variables of the estimated predictive model were identified using the random forest technique and the study was conducted with a focus on the predictive performance itself. Although there are limitations in research, such as the lack of clear criteria for the classification of depression levels and the failure to reflect variables other than KLoSA data, it is expected that if additional variables are secured in the future and high-performance predictive models are estimated and utilized through various machine learning techniques, it will be able to consider ways to improve the quality of life of senior citizens through early detection of depression and thus help them make public policy decisions.

Keywords : Machine Learning, Classification, Elderly Depresiion, Random Forest, Decision Tree.

JEL Classification Code : C38, H51, I18.

1. Introduction

우울증은 현대사회에서 가장 흔하면서도 심각한 정신건강 문제 중 하나이다(Regier, Jeffrey, Burke, & Rae, 1988). 세계보건기구 WHO는 2030년 이 우울증이 인류에게 가장 큰 부담을 야기하는 질병 1위가 될 것으로 보았으며(Park & Kim, 2011), 우리나라의 경우에도 보건복지부의 2006년 정신질환실태 역학조사에 따르면 주요 우울장애 유병률은 연간 2.5%로 2001년 연간 유병률 1.8%와 비교할 때 빠르게 증가하고 있는 추세다(MOHW, 2006). 특히 노년층의 우울증은 보다 심각한 수준이다. 2000년 노인인구가 전체인구의 7.3%, 2018년에는 전체인구의 14.3%를 차지하여 고령화 사회로 진입한 우리나라는 2030년에는 24.5%, 2050년 38.1%로 초고령 사회로 진입할 것으로 예측되고 있는데(Statistics Korea, 2018), 건강보험심사평가원 국민관심질병통계 자료에 따르면 우울증 진단을 받은 노인환자가 2014년 약 20만명에서 2018년 약 25만명으로 21.7% 증가하였고, 2018년 우리나라의 전체인구 중 우울증을 앓고 있는 환자 중 33.1%가 65세 이상으로 나타났다(HIRA, 2019).

1 First Author. Master's degree graduate, Department of Data Knowledge Service Engineering, Dankook University, Korea.
2 Professor, Graduate School of Business, Dankook University, Korea. Email: eungkyosuh@dankook.ac.kr
3 Corresponding Author. Professor, Department of Data Knowledge Service Engineering, Graduate School. Dankook University, Korea. Tel: +82-32-8005-3888. Email: kimtoja@dankook.ac.kr

노인의 우울증은 전반적인 삶의 질 저하와 인지저하를 수반하며 특히 독거노인과 같이 삶의 질 여건이 좋지 못한 노인들의 경우는 자살이나 사망에까지 이르게 되는 경우도 있으며 최근 연구에 따르면 노인 사망 원인의 25%가 정신적 장애로 인한 것이라는 연구도 있다(Kim, 2020). 노인의 우울증은 삶의 질과 밀접한 연관이 있으므로 이에 대한 개선이 필요하며 노인의 우울증을 유발하는 요인에 대하여 장소별, 경제적환경별, 심리상태별 체계적인 맞춤형 접근이 필요하다. 특히 노인이 처한 지지의 결핍으로 인하여 삶의 질이 낮아지게 되고 이러한 상황은 독거노인들의 외로움과 우울감을 키우는 것으로 알려져 있다(Russel & Taylor, 2009). 따라서 일괄적인 문제해결 방법보다는 노인 개개인의 특성에 맞추어 해결할 필요가 있다. 4 차 산업혁명 기술은 이러한 문제 해결에 도움을 줄 수 있다. Yoo, Suh, and Kim(2020)은 인공지능 스피커를 통하여 노인의 지속적 돌봄 서비스를 통하여 삶의 질 수준을 높이는 것을 통하여 우울증을 개선할 수 있음을 검증하였다. 하지만 노인의 우울증 증상을 빅데이터를 적용하여 적용한 사례는 없었다.

본 연구는 노인의 우울증 증상을 빅데이터와 기계학습 방법론을 적용해 노인 인구의 우울증상을 조기에 예측할 수 있는 최적의 모델을 선별하는데 그 목적이 있다. 노인의 우울증 증상을 예측할 수 있는 108 개 변수들의 데이터를 기반으로 의사결정나무 기법과 랜덤포레스트 기법을 적용하여 우울증 여부를 예측하고, 더 나아가 경증과 중증을 예측할 수 있는 모형을 검증해 보려고 한다. 이에 따라 노인 우울증을 예측할 수 있는 모델을 수립하여 노인들의 우울증 여부를 조기에 예측하여 선제적으로 대응할 수 있으며 또한 우울증이 경증, 중증 여부도 예측하여 그에 맞는 대처를 정확히 할 수 있을 것으로 파악된다. 특히 관리의 사각지대에 놓인 독거노인의 경우 자신이 우울증에 걸린 상황조차도 인지하지 못하고 있는 경우가 많아 본 연구에서 제시하는 우울증 예측모델을 기반으로 노인 우울증 문제를 해결하는데 기여하고자 한다.

2. 선행연구

2.1. 노인우울증

연구배경에서 밝힌 바와 같이 노인인구가 증가함에 따라 노인 우울증과 관련된 다양한 연구가 이루어져 왔으며 크게 연구의 흐름은 환경적 요인, 정신적 요인으로 나누어져 연구가 진행되어왔다. 그러나 최근 공공 빅데이터 기반의 패널데이터를 활용하여 복합적 모형을 연구한 사례가 나타나고 있다.

먼저 환경적 요인에 있어서는 Kim(2009)은 노인의 우울과 가족의 지지정도와의 관계에 대한 연구를 진행하였다. 연구 결과 가족의 지지정도와 우울은 서로 역상관관계가 있는 것으로 나타났고, 배우자가 없을 경우 대신할 지지자원 제공의 필요성에 대해 주장했다. 또한 가족의 지지는 노인의 강한 욕구를 충족시켜주고 스트레스 경험에 대한 가능성을 감소시켜주는 중요한 환경 변인이라고 이야기했다. Jeong(2015)은 서울시 복지패널 2 차년도 자료를 활용하여 노후 준비가 사회참여, 스트레스, 사회적지지와 우울을 통해 자살생각에 미치는 매개효과를 검증하였다. 대상은 50~59 세의 중·고령자 911 명으로 연구방법으로는 구조방정식 모형분석, Soble test 를 실시하였다. 연구 결과 우울증상은 사회참여, 사회적지지, 스트레스 등의 변수들과 함께 자살생각에 유의미한 수준에서 자살생각에 대한 매개효과가 검증되었다. 하지만 노후준비는 자살생각에 직접적인 영향을 미치지는 못하는 것으로 나타냈다.

다음으로 정신적 요인에 대한 연구도 진행되어 왔다. 정신적 요인에 대한 연구는 주로 삶의 질에 대한 인식, 불안정 등이 우울증에 미치는 영향에 대한 연구가 주를 이루었다. Ko(2012)는 한국고용정보원에서 45 세 이상 중고령자를 대상으로 실시하고 있는 패널조사 자료인 '고령화 연구 패널조사(Korean Longitudinal Study of Ageing, KLoSA)' 데이터를 기반으로 60 세 이상 노인들의 정신건강 실태를 파악하고, 우울 증상과 건강 위험행동 간의 연관성을 로지스틱 회귀분석을 이용하여 분석하였다. 연구 결과 우울 증상과 저하된 삶의 질, 불안 증상 그리고 위험 음주 여부는 통계적의로 유의미하지 않았으며, 흡연자가 우울 경험이 더 많은 것으로 보였다고 주장하였다. Kim(2017)은 2008 년부터 2015 년까지 45 세~80 세의 중·고령자를 대상으로 진행된 한국복지패널조사 자료를 활용하여 개인의 정신건강에 자산 및 소득 수준과 관련된 경제적 불안정이 미치는 영향에 대해 혼합모형으로 분석하였다. 연구 결과 중·고령층의 경제적 불안정과 우울증상과 자살생각은 정의 상관관계를 보였고, 이에 따라 중·고령층의 우울증상을 감소시키기 위해 기초노령연금 등의 소득을 보장해주는 복지정책과 개인의 소득 안정성을 증가시킬 수 있는 경제 분야의 정책으로 중·고령층의 정신건강 증진과 자살율 감소를 위한 통합적인 사회 안정망이 필요하다고 주장하였다.

가장 최근에는 Kim, Cho, Choi, Lee, Kang, and Kim (2019)에 의해 노인의 인지능력과 우울간의 연관성에 대한 연구도 진행되었다. 연구결과 알츠하이머성 치매환자군에서는 인지기능 변수들과 우울증과의 상관관계를 확인하였다.

현대사회에서 우울증은 심각한 질병으로 여겨지고 우울증 초기진단 및 치료를 위한 노력들이 지속적으로 이루어지고 있으며, 이를 위해 신뢰성과 타당성이 높은 우울증 조사도구를 통해 보다 빠르게 우울증을 측정하고자하는 연구 및 시도들이 진행되고 있다(Heo, Park, & Bae, 2015).

2.2. 기계학습

기계학습(Machine Learning)은 컴퓨터 학습 이론으로부터 시작된 연구 분야로, 컴퓨터가 인간의 학습 능력과 같은 기능을 알고리즘과 프로그램을 이용하여 데이터로부터 스스로 배워 새로운 정보를 발견하거나 의사결정 하는 것을 말한다 (Kim, 2014). 머신 러닝이라는 용어는 Arthur Samuel 에 의해 "명시적으로 프로그램을 작성하지 않고 컴퓨터에 학습할 수 있는 능력을 부여하기 위한 연구 분야"라고 최초 정의된 바 있으며, Mitchell(1997)은 "컴퓨터 프로그램이 어떤 작업 T 와 평가 척도 P 에 대해서 경험 E 로부터 학습한다는 것은, P 에 의해 평가되는 작업 T 에 있어서의 성능이 경험 E 에 의해 개선되는 경우를 말한다"라는 형식적인 정의를 내린 바 있다.

기계학습 알고리즘은 Figure 1(COGNUB, 2019)처럼 데이터 학습방법에 따라 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning), 강화학습(Reinforcement)으로 크게 3 가지로 나눌 수 있다.

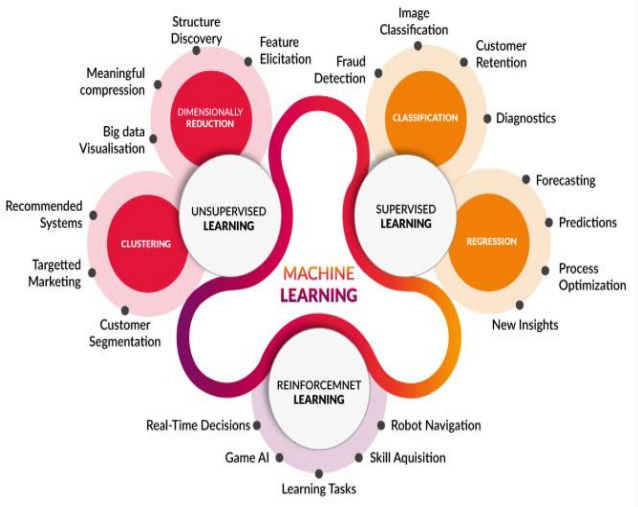

Figure 1: Machine Learning Method and Utilization Field from http://www.cognub.com/index.php/cognitive-platform/

지도학습이란 구성이 비교적 단순하며 널리 알려진 학습방법으로 입력 값 즉, 독립변수에 대응하는 결과 값(종속변수)이 존재하는 데이터를 활용한다. 목적은 학습된 모형을 이용하여 또 다른 입력 값에 대응하는 결과 값을 예측하는 것이다(Bae, 2019).

비지도학습은 지도학습과 달리 결과 값이 존재하지 않는 데이터를 학습하는 방법으로, 기능이 유사한 데이터를 분류하고 관계를 찾는다. 비지도학습은 인간이 찾아내지 못했던 데이터 속 구조, 패턴, 관계 등을 찾아낼 수 있다는 장점이 있지만, 도출된 추론의 타당성을 확인하기 어렵다는 단점도 있다(Hastie et al., 2009).

강화학습은 행동의 주체가 되는 에이전트가 주어진 환경에서 상태를 인식하고 선택 가능한 행동 중 최대 보상을 가지는 행동을 선택하는 방법이다. 이는 인간의 지식 습득 과정처럼 에이전트가 관측, 행동, 보상의 일련의 상호작용을 경험하고 시행착오를 통해 학습해 나가는 방식으로 주로 Game AI, Robot Navigation 등에서 활용되고 있다(NA, 2019).

2.3. 의사결정나무 및 랜덤포레스트

랜덤포레스트(Random Forest)를 제대로 이해하기 위해서는 그 근간이 되는 의사결정나무(Decision Tree)에 대한 이해가 필수적이어야 한다. 의사결정나무는 의사결정규칙(decision rule)을 도식화하여 분석대상을 여러개의 소집단으로 분류 또는 예측을 수행하는 분석방법이다. 또한 의사결정나무는 분류나 예측을 원하는 모든 상황에서 이용될 수 있으며, 분석의 정확도보다는 분석의 과정의 설명이 중요할 경우 더 유용하다(Choi & Seo, 1999).

랜덤포레스트는 의사결정나무 알고리즘 중 하나인 분류회귀트리(classification and regression tree: CART)의 발전된 방법으로, 그룹 내 동질성(homogeneity)이 높은 집단을 판단하기 위해 불순도함수(Impurity)를 활용한다. 불순도함수는 데이터들 간의 흩어짐 정도를 의미하며, 동질성이란 데이터의 형태 및 특성이 비슷한 것을 의미한다(Kim, Lee, Hwang, & Won, 2008). 의사결정나무의 경우 학습 속도가 빠르고 이상치에 크게 영향을 받지 않는 장점이 있으나, 과대적합(overfitting)의 위험으로 정확도가 비교적 낮은 편이지만(Kim, Lee, & Hong, 2017) 가지치기와 Tree 성장을 제어하면서 정확도보다 일반화 성능을 향상시키기도 한다(Cho & Ye, 2014).

하지만 랜덤포레스트의 경우 하나의 의사결정나무를 여러 개의 의사결정나무로 확장한 것으로, 대수의 법칙(law of large numbers)으로 인해 나무의 수가 많아질수록 일반화의 오류가 특정한 값으로 수렴하여 과적합의 위험이 줄어들고, 본래 의사결정나무를 기본으로 하고 있어 이상치에 비교적 영향을 받지 않는다(Kim & Ahn, 2016).

랜덤포레스트에서 다수의 결정트리를 만들기 위해 예측인자와 관측치에 대한 무작위 샘플링을 반복하게 되며(Breiman, 2001), 샘플링에는 의사결정나무와 같이 부트스트래핑 기법이 사용된다(Choi & Man, 2018).

3. 연구방법

3.1. 데이터 분석 방법 및 절차

본 연구의 데이터 분석은 크게 4 단계로 구성하여 진행하였다. 먼저 연구에 필요한 데이터를 수집한 후 연구대상을 선정하였다. 다음으로 분석에 활용할 변수를 선택하고, 선택한 변수들의 특징을 설명하였다. 데이터 분석 및 예측은 R 을 이용하여 의사결정나무와 랜덤포레스트 기법을 사용하였으며, 마지막으로 추정된 예측 모형을 검증하였다.

3.1.1. 데이터 수집

한국고용정보원에서 실시하고 있는 고령화연구패널조사(KLoSA)의 데이터를 취득하여 활용하였다. 본 데이터는 효과적인 사회경제정책 수립과 학술 연구의 기초자료로 활용하고자 2006 년부터 제주도를 제외한 지역에 살고 있는 45 세 이상 중·고령자 중 일반 가구 거주지를 대상으로 표집 및 조사하여 수집된 데이터이다. 2006 년부터 격년 주기로 동일한 조사 항목의 기본조사를 실시하고 있으며, 현재는 2018 년 제 7 차 기본조사까지 진행되었다. 본 연구에서는 현재 조사 및 데이터 취합이 완료된 제 6 차 기본조사 데이터를 활용하여 연구를 진행하였다(KEIS, 2019).

3.1.2. 변수 선택 및 설명

노인우울증 예측 모형을 추정하기 위해 제 6 차 기본조사 대상자 6,618 명 중 65 세 이상 노년층 1,986 명을 대상으로 선정하였다. 우울 척도로는 많은 조사 도구들 중 CES-D10 을 활용하였다. CES-D10 은 일반인들이 자기보고형식으로 본인이 경험하는 우울을 보다 용이하게 측정하기 위해 개발된 CES-D 의 단축형으로 우울 증상 선별을 위한 척도로 사용되고 있다. CES-D10 은 10 점을 총점으로 하여 절단값(cut-off) 4 점을 기준으로 정상과 우울을 선별한다(Irwin, Artin, & Oxman, 1999). 선행연구 탐색결과 CES-D10 의 점수로 경도 우울과 중증 우울을 명확히 구분하는 절단값에 대한 연구를 찾기 어려워 본 연구에서는 4~6 점을 경도 우울, 7~10 점을 중증 우울의 선별 기준으로 구분하여 연구를 실시했다.

Table 1: Frequency of Dependent Variable

Normal/Depresion Classification	Normal	Depression	
	2,560 (56.3%)	1,986(43.7%)	
Moderate/Severe Classification	-	Moderate	Severe
		1,142(57.5%)	844(42.5%)

종속변수는 정상/우울을 구별하는 변수와, 경도 우울/중증 우울을 구별하는 변수 총 2 가지로 설정하였다. 두 가지 종속변수의 빈도는 Table 1 과 같다. 연구 대상 4,546 명 중 2,560 명(56.3%)이 정상으로, 1,986 명(43.7%)은 우울증상을 보이는 것으로 선별되었다. 우울증상을 보이는 대상 중 경도 우울은 1,142 명(57.5%), 중증 우울은 844 명(42.5%)으로 구분되었다. 또한 독립변수 데이터들 간의 계층불균형 문제가 있는 것으로 판단하여 SMOTE 기법을 통해 분석 데이터 셋을 다시 생성한 후 학습 및 예측을 진행하였다.

독립변수는 선행연구에 따라 개인 및 가구의 총소득과 같은 자산 및 소득 수준(Kim, 2017), 인지 기능(Kim et al, 2019), 삶의 질과 건강 상태, 일상생활 수행능력(Ko, 2012), 가족의 지지여부와 관계(Kim, 2009), 인적속성, 노동여부에 따른 사회 참여(Jeong, 2015) 등의 변수를 선택하여 분석 및 예측을 진행하였다.

Table 2: Independent Variables

NO	Domain	Variables	Variable Description	Data Type
1	Family	w06bp1	Living parents	factor
2		w06bb_adl1	Inconvenient family status	factor
3		w06bb_adl2	Receiving family status	factor
4		w06bb_adl3	Non-cooperative family status under care	factor
5		w06s_sum	the number of surviving brothers/sisters	num
6	Health status	w06C001	Health condition	factor
7		w06C152	Subjective health condition	factor
8		w06C003	Whether or not the doctor has been judged to be disabled	factor
9		w06C005	Restriction of activity due to health condition	factor
10		w06chronic_a	High blood pressure diagnosis status	factor
11		w06chronic_b	Diabetes/hyperglycemia diagnosis status	factor

12		w06chronic_c	Cancer and malignant tumor diagnosis status	factor
13		w06chronic_d	Chronic lung disease diagnosis status	factor
14		w06chronic_e	Diagnosis of liver disease	factor
15		w06chronic_f	Diagnosis of heart disease	factor
16		w06chronic_g	Diagnosis of cerebrovascular diseases	factor
17		w06chronic_h	Psychiatric disease-related diagnosis status	factor
18		w06chronic_i	Arthritis/Ryumatis Diagnosis Status	factor
19		w06C074n	Diagnosis of digestive diseases	factor
20		w06C078n	Disk diagnostic status	factor
21		w06C081	Difficulties in day-to-day activities due to vision	factor
22		w06C082	Using hearing aid normally	factor
23		w06C085	Wearing Dentures normally	factor
24		w06C106	Weight variation of more than 5 kilograms in a year	factor
25		w06C108	Regular exercise	factor
26		w06smoke	Smoking	factor
27		w06alc	Drinking	factor
28		w06chronic_sum	Chronic disease count	num
29		w06C105	Weight	num
30		w06C107	Kidney	num
31		w06bmi	Koup Index (BMI)	num
32	Economic activity	w06present_ecotype	Current state of economic activity	factor
33		w06present_labor	Current Labor Status	factor
34	Medical security and use of facilities	w06C301	National health insurance/medical benefits subscription status	factor
35		w06C302	Whether or not a person is a member of a plantation/region	factor
36		w06C305	health insurance premium payer	factor
37		w06C310	Private medica insurance coverage status	factor
38		w06C313	Whether to benefit from the first round of free medical checkups	factor
39		w06C318n	Recognition of long-term care insurance for the elderly	factor
40		w06C329n	Awareness of the Elderly Care Service Project	factor
41		w06C343	Regularly prescribed medication	factor
42		w06C346	1year experience in purchasing medical aids	factor
43		w06oopm2	Out-of-town expenses	num
44		w06C318	Number of hospitalizations since the last basic survey	num
45		w06C330	Number of visits to dental clinics in a year	num
46		w06C333	Number of visits to health centers in a year	num
47		w06C334	Number of visits to oriental medicine hospitals per year	num
48		w06C337	Number of outpatient visits to other hospitals/ clinics in one year	num
49		w06C340	Number of visits to medical personnel in one year	num
50	Personality	w06_fam1	Generational composition	factor
51		w06edu	Academic background	factor
52		w06gender1	Gender	factor

53		w06marital	Current state of marriage at the time of basic investigation	factor
54		w06A030	Religion	factor
55		w06A036	Furniture owner status	factor
56		w06enu_type	Dwelling type	factor
57		w06region1	Residence_City	factor
58		w06region2	Residence_Urban	factor
59		w06region3	Residential area_large cities/small and medium cities/ townships	factor
60		hhsize06	household head	num
61		w06A002_age	Age	num
62		w06A032	Number of times you meet close friends	num
63		w06Ba003	Current number of surviving children	num
64		w06Ba068	Number of grandchildren	num
65	Cognitive ability	w06mmseg	Cognitive function classification	factor
66		w06C401	Memory (date-year/month/day)	factor
67		w06C402	Memory (Days)	factor
68	Cognitive ability	w06C406	Memory test (3 words memorization)	factor
69		w06C407	Attention Concentration and Calculation (Exclude1)	factor
70		w06C408	Attention concentration and calculation (excluding)	factor
71		w06C409	Attention concentration and calculation (excluding 3)	factor
72		w06C410	Attention concentration and calculation (excluding 4)	factor
73		w06C411	Attention concentration and calculation (excluding5)	factor
74		w06C412	Memory test	factor
75		w06C413	Purpose of personal belongings (possession 1)	factor
76		w06C414	Purpose of personal belongings (possession 2)	factor
77		w06C419	Command Execution (draw)	factor
78		w06mmse	Cognitive score	num
79	Ability to perform daily life	w06C201	Changing clothes	factor
80		w06C202	Washing/brushing/washing hair	factor
81		w06C203	Taking a bath/shower	factor
82		w06C204	Dining with the food set up	factor
83		w06C205	To rise from the bed and leave the room	factor
84		w06C206	Toilet use	factor
85		w06C207	To adjust the urine volume	factor
86		w06C208	Grooming oneself	factor
87		w06C209	Daily household chores (cleaning, etc.)	factor
88		w06C210	Meal preparation	factor
89		w06C211	Laundry (laundry, hanging laundry, etc.)	factor
90		w06C212	Near-field outing (no means of transportation)	factor
91		w06C213	To go out by means of transportation	factor
92		w06C214	Shoplifting	factor
93		w06C215	Money management (money/bank account/property management, etc.)	factor

94		w06C216	Dial and receive telephone	factor
95		w06C217	Take your medicine	factor
96	Assets and income	w06f001type	Furniture type	factor
97		w06hhinc	Gross household income	num
98		w06hhassets	Total household assets	num
99		w06hhliabilities	Total household debt	num
100		w06hhnetassets	Net worth of households	num
101		w06pinc	personal gross income last year	num
102	subjective expectation	w06G031	Subjective hierarchical consciousness	factor
103		w06G026	sfaction with life_your health condition	num
104		w06G027	Satisfaction with life_the state of one's own economy	num
105		w06G030	Satisfaction with life_The overall quality of life	num
106		w06G032	monthly allowance	num
107	Depression	ces_d_2	Depression- yes/no	factor
108		ces_d_4_7	Depression – moderate/severe	factor

3.1.3. 분석 및 예측

먼저 예측 모형을 추정하기 전에 대상자의 성별, 나이, 학력수준을 기준으로 데이터 탐색을 실시하였다. 모형 추정을 위한 분석도구는 R을 사용하였고, 분석 방법은 R 의 "tree" 패키지와 "randomForest" 패키지를 활용하여 의사결정나무와 랜덤포레스트 기법으로 분석 및 예측을 시도하였다. 본 연구에서 진행한 노인우울증 예측 모형은 두 가지로, 우울증상 여부 예측모형과 우울증상을 보이는 집단을 대상으로 경도우울/중증우울을 예측하는 모형으로 구분된다. 두 모형 모두 데이터 중 70%는 학습에 사용하였고, 나머지 30%는 검증에 사용하였다(Yoo, 2015).

3.1.4. 성능평가

예측 모형의 성능평가 역시 R 을 이용하여 진행하였다. 성능평가 지표로는 정확도, 민감도, 특이도를 모델의 정확도 및 성능을 평가하였고, MDA(Mean Decrease Accuracy)와 MDG(Mean Decrease Gini)

지표를 통해 랜덤포레스트 모형이 결과를 예측하는데 있어 중요하게 작용했던 변수를 파악하고자 했다.

4. 데이터 분석 및 예측 모형 추정 결과

4.1. 데이터 탐색

Table 3 는 노인우울증 예측 모형 추정에 앞서 연구 대상자들의 인구통계학적 특성을 기준으로 데이터 탐색을 실시한 결과이다. 성비는 남성 약 42.1%, 여성 57.9%로 여성이 보다 더 많았다. 나이의 산술평균은 75.7 세, 중앙값은 75 세로 큰 차이를 보이진 않았고, 65 세~74 세 구간과 75~84 세 구간이 거의 대부분을 차지했다(86.8%). 학력수준은 초등학교 졸업 이하가 57.1%로 가장 많았고, 대학교 졸업 이상이 6.7%로 가장 적었다.

Table 3: Characteristic of Population

Characteristic		personnel	Ratio	Remark
Total		4,546	100%	
Sex	Male	1,916	42.1%	
	Female	2,630	57.9%	
Age	65~74	2,116	46.5%	**Mean: 75.7**
	75~84	1,831	40.3%	
	85~	599	13.2%	**Median: 75**
Education	Elementary School Graduation or Less	2,597	57.1%	
	Middle School Graduation	739	16.3%	
	High School Graduation	905	19.9%	
	University graduation or higher	305	6.7%	

4.2. 모형 추정 결과

4.2.1. 정상/우울 분류 예측 모형

Figure 2 는 우울증상 여부를 분류하는 가지치기가 완료된 의사결정나무 모형이다. 뿌리 노드는 전반적인 삶의 질의 만족도(w06G030)로 나타났고, 해당 변수의 값이 59.9912 미만이면서 인지기능 점수(w06mmse)가 27.9943 미만일 경우와 전반적인 삶의 질 변수가 60.0349 초과, 69.9967 미만일 경우도 우울증으로 예측하였다.

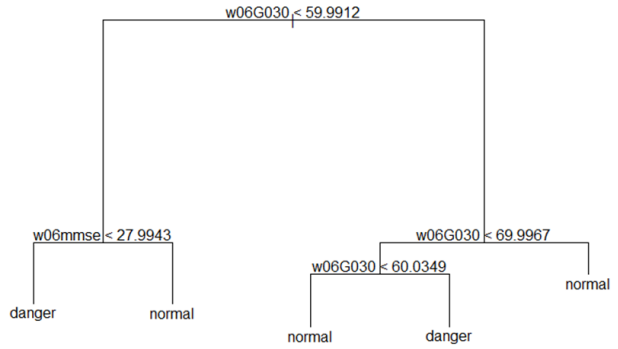

Figure 2: Final Decision Tree graph(Normal/Depression)

Table 4: Decision Tree Evaluation_Normal/Depression

Accuracy	**76.2%**
Sensitivity	72.3%
Specificity	82.1%

Table 4 는 모형의 성능을 평가한 내용으로, 모형 전체의 예측 성능인 정확도는 76.2%, 우울증을 예측할 확률인 민감도는 72.3%로 나타났다. 또한 우울증이 아니라고 예측할 확률인 특이도는 82.1%로 민감도보다 비교적 높게 나타났다. 위에서 언급했던 대로 의사결정나무를 활용한 분석을 여러번 시도해본 결과 하위 노드가 나누어질 때 특정 예측인자에 따라 정확도가 최소 69.7%에서 최대 77.1%까지 다른 결과를 보였다.

Table 5: Random Forest Evaluation_Normal/Depression

Accuracy	**86.3%**
Sensitivity	79.5%
Specificity	93.3%

Table 5 는 우울증증상 여부를 분류하는 랜덤포레스트 모형의 성능을 평가한 내용으로 정확도, 민감도, 특이도 모두 의사결정나무 모형보다 높게 나타났다. 정확도는 86.3%로 의사결정나무 모형의 결과보다 10.1% 높았다. 민감도는 79.5%로 의사결정나무 모형보다 7.2% 높았고 특이도는 93.3%로 12.2% 더 높았다.

Figure 3 는 우울증증상 여부를 분류하는 랜덤포레스트 모형의 중요변수를 나타내는 그래프이다. MDA 는 해당변수가 모형에서 제외되었을 때 성능에 얼마나 많은 영향을 미치는가를 뜻하며, MDG 는 해당변수가 모형에 적용되는 것이 분류 모형의 불순도를 얼마나 감소시키는지를 뜻한다. 즉 두 지표 모두 변수들의 중요도를 측정한 값이다.

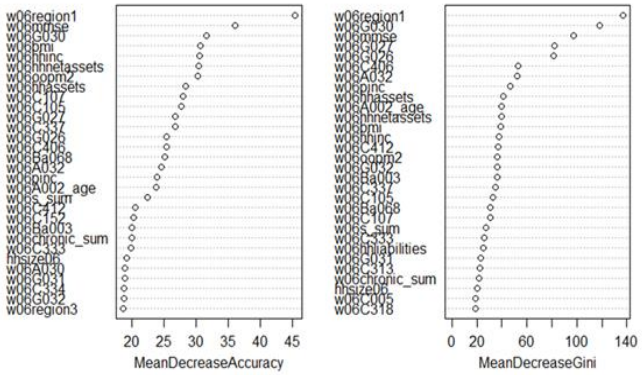

Figure 3: Important Variable Graph_Normal/Depression

MDA 값과 MDG 값 모두 기준으로는 조사대상이 거주하는 지역을 시, 도 기준으로 구분한 변수(w06region1)가 가장 중요한 것으로 나타났으며, MDA 의 경우 인지기능점수(w06mmse)와 전반적인 삶의 질의 만족도(w06G030)가 뒤를 이었다. MDG 값 기준으로는 삶의 질의 만족도(w06G030)와 인지기능점수(w06mmse) 순으로 중요도를 구분하였고, 두 지표 모두 상위 3 개의 중요 변수는 모두 동일한 것으로 나타났다. 또한 MDA 에서는 첫 번째로 중요한 변수와 두 번째로 중요한 변수 간의 측정 값 차이가 컸으나, MDG 의 경우 첫 번째, 두 번째, 세 번째 변수 간의 측정 값 차이가 비슷하고 상대적으로 크지 않았다.

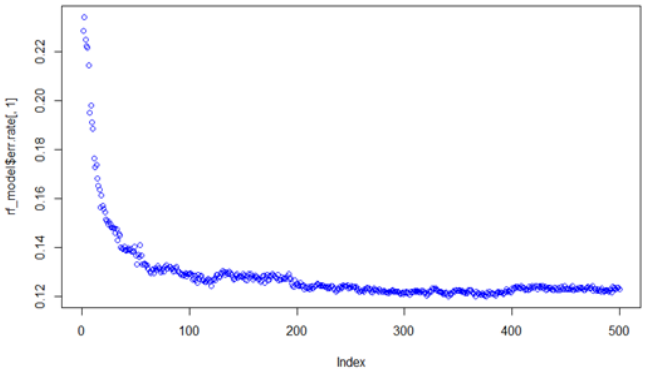

Figure 4: Error Rate Graph_ Normal/Depression

Figure 4 는 우울증증상 여부를 분류하는 랜덤포레스트 모형에서 의사결정나무 모형의 개수에 따른 에러율을 나타내는 그래프로, 해당 모형에서는 341 개의 의사결정나무 모형을 사용하여 분석하였을 때 에러율이 가장 낮았다.

4.2.2. 경도/중증 우울 분류

우울증상이 있다고 판단되는 조사대상들 중 증상을 경도와 중증으로 분류한 가지치기가 완료된 의사결정나무 모형결과는 Figure5 에 제시되어 있다. 총 11 개의 노드로 구성되어 있으며, 총 4 개의 노드로 구성되어 있는 우울증상 여부를 분류하는 의사결정나무 모형과 비교했을 때 상대적으로 복잡한 형태를 보이고 있다. 뿌리 노드는 조사대상자의 거주지역을 시,도 기준으로 구분한 변수(w06region1)로 나타났고, 주관적 기대감 영역인 자신의 경제상태와 삶의 질에 대한 만족도(w06G027, w06G030), 주관적 계층의식(w06G031)과 인적속성인 세대구성(w06_fam1), 거주지역을 대도시/중소도시/읍면으로 구분하는 변수(w06region3), 건강상태 영역의 카우프지수(w06bmi), 인지력 영역의 주의집중 및 계산_뺄셈 5(w06C411)과 약 챙겨먹기(w06C217), 가족 영역의 생존해 있는 형제/자매 수(w06s_sum), 의료보장 및 시설이용 영역의 지난 1 년간 보건소 방문횟수(w06C333) 변수가 하위 노드로 나타났다. 이 중 조사대상자의 거주지역을 시, 도 기준으로 구분한 변수(w06region1)는 두 모형 모두 공통적으로 선택되었다.

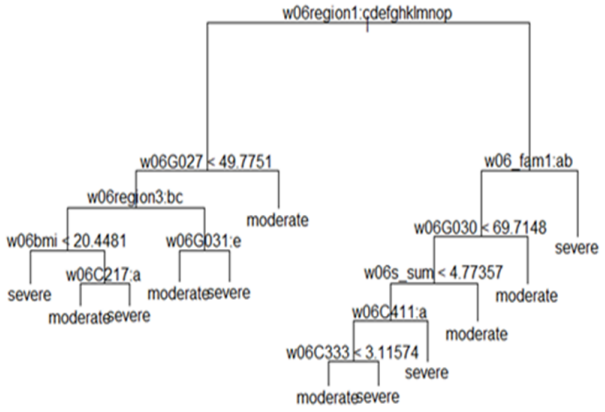

Figure 5: Final Decision Tree Graph (Moderate/Severe)

Table 6: Decision Tree Evaluation _ Moderate/Severe

Accuracy	73.1%
Sensitivity	74.9%
Specificity	71.9%

Table 6 는 의사결정 나무 모형의 성능을 평가한 내용으로, 모형 전체의 예측 성능인 정확도는 73.1%, 경도 우울을 예측할 확률인 민감도는 72.3%로 나타났으며. 중증 우울을 예측할 확률인 특이도는 71.9%로 민감도보다 비교적 낮게 나타났지만, 세가지 지표 모두 비슷한 수준을 보였다. 우울증 여부를 예측하는 모형의 성능과 비교했을 때 정확도는 3.1%, 특이도는 10.2% 차이를 나타내며 상대적으로 성능이 떨어졌으며, 민감도는 2.6% 차이로 성능이 비교적 높았다.

Table 7: Random Forest Evaluation_Moderate/Severe

Accuracy	86.1%
Sensitivity	93.9%
Specificity	74.7%

Table 7 는 우울증상이 있다고 판단되는 조사대상들 중 증상을 경도와 중증으로 분류한 랜덤포레스트 모형의 성능을 평가한 내용으로 정확도는 86.1%로 나타났으며 민감도는 93.9%, 특이도는 74.7%로 나타났다. 이 모형 또한 우울증 여부를 예측한 모형과 마찬가지로 정확도, 민감도, 특이도 모두 의사결정나무 모형보다 높게 나타났고, 특히 민감도의 경우 19% 차이로 예측성과를 크게 개선하였다. 또한 민감도와 특이도의 차이가 19.2%로 의사결정 나무 모형과 비교했을 때 큰 차이를 보였다.

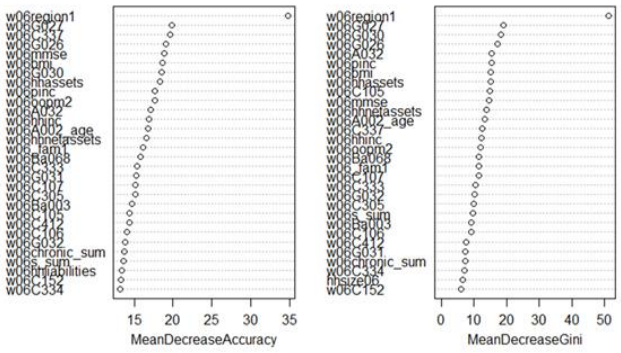

Figure 5: Important Variable Graph_ Moderate/Severe

Figure 5 는 우울증 증상을 경도와 중증으로 분류하는 랜덤포레스트 모형의 중요변수를 나타내는 그래프이다. MDA 값과 MDG 값 모두 조사대상이 거주하는 지역을 시, 도 기준으로 구분한 변수(w06region1)가 가장 중요한 것으로 나타났으며, 자신의 경제상태에 대한 만족도(w06G027) 또한 두 지표 모두 동일하게 두 번째로 중요한 변수로 나타났다. 우울증 여부를 분류하는 모형에서는 MDA 그래프와 MDG 그래프의 모형과 구성이 다른 형태를 띄고 있었으나, 해당 모형에서는 그래프의 모형과 구성 모두 비슷한 형태를 보이고 있다.

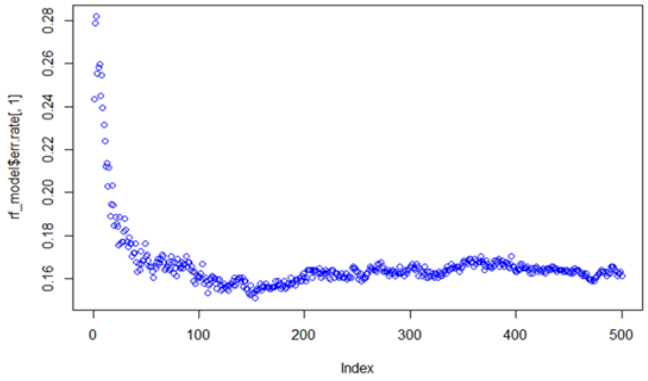

Figure 6: Error Rate Graph_ Moderate/Severe

Figure 6 는 우울증 증상을 경도와 중증으로 분류하는 랜덤포레스트 모형에서 의사결정나무 모형의 개수에 따른 에러율을 나타내는 그래프로, 해당 모형에서는 153 개의 의사결정나무 모형을 사용하여 분석하였을 때 에러율이 가장 낮았다.

5. 결론

본 연구는 기계학습의 랜덤포레스트 기법을 활용하여 노인우울증 여부와 우울의 정도를 분류하는 모형을 추정하고 각 모형의 성능평가 및 중요변수에 대해 분석하였다. 또한 선행연구와는 다르게 우울증상과 특정 변수와의 인과관계 추정이 아닌, 실제 관측값과 학습을 통해 추정된 모형의 예측값의 차이를 줄이는 것을 목표로 연구를 진행하였다. 본 연구에서 사용한 독립변수는 106 개로 조사대상자의 인적 속성, 의료보장 및 시설이용 현황, 가족관계, 건강상태, 고용 상태, 인지력, 일상생활 수행능력과 같은 객관적 특성들과 삶의 만족도와 같은 주관적 특성들까지 포함시켜 분석을 진행하였다.

Table 8: Performence Comparison_Normal/Depression

	Decision Tree	Random Forest
Accuracy	76.2%	86.3%
Sensitivity	72.3%	79.5%
Specificity	82.1%	93.3%

Table 9: Performence Comparison _Moderate/Severe

	Decision Tree	Random Forest
Accuracy	73.1%	86.1%
Sensitivity	74.9%	93.9%
Specificity	71.9%	74.7%

Table 8, 9 은 분류 기법에 따른 성능평가 비교한 표이다. 먼저 조사대상자의 우울증 여부를 분류하는 모형과 우울의 정도를 분류하는 모형 모두 의사결정나무 모형보다 랜덤포레스트 모형이 정확도가 10%이상 성능이 높게 나타났다. 예측성과 간의 차이를 자세히 살펴보면, 우울증 여부 분류 모형에서는 특히 특이도의 예측성과를 82.1%에서 93.3%로 가장 많이 개선하였으며, 반대로 우울 정도를 분류하는 모형에서는 민감도의 예측성과를 74.9%에서 93.9%로 크게 개선하였다.

Table 10: Random Forest performence Comparison

	Normal/Depression	Moderate/Severe
Accuracy	86.3%	86.1%
Sensitivity	79.5%	93.9%
Specificity	93.3%	74.7%

Table 10 은 분류기준에 따라 성능평가를 비교한 표이다. 랜덤포레스트 모형의 성능은 우울증 여부 분류 86.3%, 우울 정도 분류 86.1%로 거의 차이가 없었고, 각각 민감도는 79.5%와 93.9%, 특이도는 93.3%와 74.7%로 반대되는 큰 차이를 보였다.

Table 11: Random Forest performence Comparison(MDA)

	Normal/Depression	Moderate/Severe
1st	Residency City, Province	Residency City, Province
2nd	Mini-Mental State Examination Score	Life Satisfaction Your Economic Status
3rd	Life Satisfaction Overall Quality of Life	Number of Outpatient Treatments for Other Hospitals in Last Year

Table 11 은 우울증 여부와 우울 정도를 랜덤포레스트로 분류한 모형의 MDA 값의 순위를 나타낸 표이다. 우울증 여부 분류 모형과 우울 정도 분류 모형 모두 조사대상자의 거주 지역을 시, 도 기준으로 구분한 변수가 가장 중요한 변수로 나타났으며 우울증 여부 분류 모형에서는 인지기능점수와 전반적인 삶의 질, 우울 정도 분류 모형에서는 주관적으로 평가하는 자신의 경제상태가 2 순위로, 지난 1 년간 기타 병/의원 외래진료 횟수 3 순위로 분석되었다. 특히 우울 정도 분류 모형의 3 순위인 기타 병/의원 외래진료 횟수는 우울증상을 느끼고 있는 노인들을 대상으로 분석한 결과로 두 변수 간의 상관관계가 있을 것으로 유추할 수 있는데, 구체적인 인과관계는 향후 연구가 필요할 것으로 보인다. 또한 두 모형 모두 거주 지역과 마찬가지로 주관적인 삶의 만족도가 모형에 미치는 영향이 큰 것으로 나타났다.

Table 12: Random Forest performence Comparison(MDG)

	Normal/Depression	Moderate/Severe
1st	Residency_City, Province	Residency_City, Province
2nd	Life satisfaction_Overall Quality of Life	Life Satisfaction_Your Economic Status
3rd	Mini-Mental State Examination Score	Life satisfaction_Overall Quality of Life

Table 11 은 우울증 여부와 우울 정도를 랜덤포레스트로 분류한 모형의 MDG 값의 순위를 나타내고 있다. MDA 값과 마찬가지로 우울증 여부 분류 모형과 우울 정도 분류 모형 모두 조사대상자의 거주 지역을 시, 도 기준으로 구분한 변수가 가장 중요한 변수로 나타났으며, 두 모형 모두 주관적인 삶의 만족도가 모형에 큰 영향을 미치는 변수로 분석되었다.

MDA 와 MDG 값 모두 조사대상자의 거주지역이 가장 큰 영향을 미치는 변수로 분석되었다. 이는 특정 시, 도에 노인복지현황, 인구밀도, 생활환경 등과 연관이 있을 것으로 유추되는데, 이 또한 구체적인 인과관계는 향후 연구가 필요할 것으로 보인다.

연구목적에서도 언급했듯 대부분의 선행연구에서는 노인의 우울증상과 특정 변인 간 회귀분석을 통해 인과관계를 추정하고 계수 값을 도출하는 것에 목적을 두었지만, 본 연구에서는 랜덤포레스트 기법을 활용하여 추정된 예측 모형의 중요변수를 파악하고 예측성능 자체에 초점을 두고 연구를 진행하였다. 빠르게 변화하고 있는 현대사회에서는 현상과 현상 간의 인과관계도 중요하지만 예측성능 자체가 더욱 중요한 경우가 많아지고 있다. 본 연구에서 종속변수로 선택한 노인우울증의 경우에도 마찬가지로, 특정 변인과의 인과관계를 추정하여 근본적인 원인을 파악하는 것도 중요하지만, 노인우울증 예측에 많은 영향을 미치는 중요변수를 추정하고 이를 활용하여 노인우울증을 보다 빠르게 예측하는 것이 더 중요할 수도 있다. 연구결과 기존연구와 비슷하게 예측모형의 결과 우울증 여부에 있어서는 거주지역, 전반적인 삶에 있어 삶의 만족도, 인지기능 점수가 영향을 미치는 상위 3 개 요인으로 검증되었으며, 우울 정도에 있어서는 거주지역, 경제적 삶에 있어 삶의 만족도, 전반적인 삶에 있어 삶의 만족도가 영향을 미치는 상위 3 개 요인으로 검증되었다. 이러한 결과는 노인들에 관한 정책을 수립하는데 있어 보다 직접적인 도움을 줄 수 있을 것이다. 특히 거주지역이 우울증 여부와 우울정도에 가장 큰 영향을 나타났다는 점은 노인의 거주환경에 대한 현황 진단과 더불어 이에 대한 개선이 가장 시급한 부분임을 알아야 할 것이다. 또한 전반적인 삶에 있어서의 만족도 또한 중요한 것으로 나타났는데 이는 주관적 지표로 전반적 삶의 질 향상을 위한 방향성 또한 중요한 것으로 확인할 수 있었고, 이를 위하여 노인 맞춤형 케어 서비스 관련 정책이 수립되어야 할 것으로 보인다.

본 연구에서는 우울 정도를 분류하는 값의 기준이 명확하지 않은 점과 KLoSA 데이터 이외의 변수들을 반영하지 못한 점, 샘플링 방법에서 부트스트랩을 포함하지 못한점에 있어 연구의 한계가 존재한다. 하지만 본 연구에서 제시한 108 개의 변수 외에 우울증 증상 조기 예측을 위한 추가적인 변수를 확보하고 다양한 기계학습 기법을 통해 높은 성능의 예측 모형을 추정하여 활용한다면 첫째, 노인우울증 조기 예측을 통해 관리의 사각지대에 있는 노인들의 우울증에 대하여 능동적으로 대처할 수 있으며, 둘째, 위험군으로 분류된 노인의 우울증 증상을 경증, 중증으로 예측 분류하여 보다 상세한 대처가 가능할 것이다. 이에 따라 노인의 건강과 삶의 질을 향상시킬 수 있는 관리방안이 도출될 수 있으며, 공공 정책 의사결정에 더욱 큰 도움이` 될 것으로 기대한다.

References

Bae, S. W. (2019). *A Comparative Study on the Prediction of Housing Price Using Machine Learning* (Doctoral dissertation). Dankook University, Yongin, Korea.

Bae, S. W., & Shin W. S. (2005). The Center for Epidemiological Studies-Depression Scale (CES-D): Application of Verification Factor Analysis Method. *Health and Social Sciences, 18*, 166.

Breiman, L. (2001). Random forests, *Machine Learning, 45*, 5.

Cho, K. S., & Ye, W. J. (2014). P2P Traffic Classification using Advanced Heuristic Rules and Analysis of Decision Tree Algorithms. *Journal of The Korea Society of Computer and Information, 19*(3), 45-54.

Choi, J. H., & Seo, D. S. (1999). Application of Data Mining Decision Tree. *Statistical Analysis Study*, *4*(1), 62.

Choi, P. S., & Man, I. S. (2018). A Model for the Employment Prediction of College Graduates Using Machine Learning Technique. *Job Competency Development Research, 21*(1), 35-36.

COGNUB (2019). *Cognitive Computing and Machine Learning.* Retrieved May 14, 2020, from http://www.cognub.com/index.php/cognitive-platform/

Trevor, H., Tibshirani, R., & Friedman, J. (2009). *Unsupervised learning. The element of statistical learning* (pp.305-585). New York, NY: Springer.

Health Insurance Review & Assessment Service (2019). *Statistic of National Concerned Diseases-Depression*. Retrieved May 12, 2020, from http://opendata.hira.or.kr/op/opc/olapMfrnIntrsIlnsInfo.do

Heo, M. S., Park, B. S., & Bae, S. W. (2015). Verification of measurement invariant of the abbreviated CES-D scale in Korean. *Mental Health and Social Welfare, 43*(2), 314.

Jeong, I. Y. (2015). *A Study on the Causal Model of Preparing for the Elderly and Suicide of Middle and Middle-aged People: Focused on the Intermediary Effects of Social Participation, Social Support, Stress, and Depression* (Doctoral dissertation). Catholic University, Bucheon, Korea.

Myers, J. K., Morton, K., Robins, L. N., George, L. K., Karno, M., & Locke, B. Z. (1988). One-Month Prevalence of Mental

Disorders in the United States. *Arch Gen Psychiatry, 45*, 977-986.

KEIS (2016). *The 6th Basic Survey of the Aging Research Panel (KLoSA) in 2016*.

KEIS (2019). *2018 Aging Research Panel User Guide.*

Kim, B. J. (2020). Factors Influencing Depressive Symptoms in the Elderly: Using the 7th Korea National Health and Nutrition Examination Survey. *Journal of Health Informatics and Statistics*, *45*(2), 165-172.

Kim, D. H. (2009). A Study on the Relationship between Family Support, Self-respect, and Depression in which Older People are Late. *Elderly Welfare Research*, *13*, 136, 138.

Kim, D. J., Cho, S. Y., Choi, J. S., Lee, M. W., Kang, S. H. & Kim, S. W., (2019). A Study on the Relationship between cognitive function and senile depression and senile stress. *Journal of the Korean Society of Clinical Examination and Sciences*, *51*(1), 111.

Kim, I. J. (2014). Deep Learning: New Trends in Machine Learning. *Journal of the Korean Telecommunications Society*, *31*(11), 52.

Kim, J. K., Lee, K. B., & Hong, S. K. (2017). ECG-based biometric authentication using random forest. *Journal of Electronic Engineering Society*, *54*(6), 102.

Kim, J. W. (2017). *The Effects of Economic Stabilization and Retirement Income Security Policy on Mental Health* (Doctoral dissertation). Seoul National University, Seoul, Korea.

Ko, K. D. (2012). The Relationship between Health Risk Behavior and Mental Health in the Elderly in Korea: Korean Longitudinal Study of Ageing(KLoSA). *Journal of the Korean Geriatrics Society*, *16*(2), 66-73.

Kim, S. J., Ahn, S. J., & Ahn, H. C. (2016). Application of Random Forest to Predict Corporate Credit Ratings. *Industrial Innovation Research*, *32*(1), 187-191.

Kim, T. H., Lee, Y. T., Hwang, E. P., & Won, J. M. (2008). A Study on the Establishment of Subway Station Area in New Town Area Using CART Analysis. *Journal of the Korean Railway Society*, *11*(3), 217.

Mitchell, M. (1997). *Machine Learning.* New York, NY: McGraw-Hill.

Na, D. Y. (2019). *A Study on the Smart Farm Technology for Animal Welfare based on IoT using Machine Learning Algorithm* (Doctoral dissertation). Konkuk University, Seoul, Korea.

Irwin, M. R., Artin, K. H., & Oxman, M. N. (1999). Screening for Depression in the Older Adult: Criterion Validity of the 10Item Center for Epidemiological Studies Depression Scale (CES-D). *Archives of Internal Medicine*, *159*(15), 1701-1704.

Ministry of Health and Welfare. (2006). *Epidemiological Survey on the Actual Condition of Mental Diseases* (pp.83).

Park, J. H., & Kim, K. W. (2011). A Study on the Epidemiology of Depression in Korea. *Journal of the Korean Medical Association*, *54*(4), 362.

Russell, D., & Taylor, J. (2009). Living alone and depressive symptoms: The influence of gender, physical disability, and social support among Hispanic and non-Hispanic older adults. *Journal of Gerontology Series B: Psychological Science and Social Sciences*, *64*(1), 95-104.

Yoo, H. S., Suh, E. K., & Kim, T. H. (2020). A Study on Technology Acceptance of Elderly living Alone in Smart City Environment: Based on AI Speaker. *Journal of Industrial Distribution & Business*, *11*(2), 41-48.

Yoo, J. E. (2015). Random forests, an alternative data mining technique to decision tree. *Journal of Educational Evaluation*, *28*(2), 433-434.