과 목 명: 데이터 마이닝

담당교수: 진창호 교수

학 과: 산업경영공학과

학 번: 2016100937 김성수

2016100946 김효준

2016100966 윤태웅

2016100989 정일도

2016101000 홍성희

DATA MINING Term-Project

Intermediate Report

<성인 독신 여성의 우울증 예측 모델>



발표 영상: \

목 차

1. Introdduction 3
2. 과제 수행 배경 및 동기
3. 문제 정의 및 도출
4. 기존 연구 및 기술의 문제점
5. 제안하는 방법의 개요
6. 논문 주요 결과의 개요
7. methods 5
8. 연구 대상
9. Data Understanding
10. 연구에 사용한 변수 및 Pre-Processing
11. 분석 방법
12. results 24
13. Decision Tree Model
14. Confusion Matrix
15. discussion 27
16. limitation 29
17. 결측치 대체의 한계
18. 가중치 적용의 한계

표 및 그림 목차

[[표 1] 우울증 관련 변수 7](#_Toc40388269)

[[표 2] 만성질환 관련 변수 8](#_Toc40388270)

[[표 3] 소득 관련 변수 9](#_Toc40388271)

[[표 4] 신체 관련 변수 9](#_Toc40388272)

[[표 5] 의료 서비스 관련 변수 10](#_Toc40388273)

[[표 6] 업무 관련 변수 10](#_Toc40388274)

[[표 7] 여가 활동 변수 10](#_Toc40388275)

[[표 8] 의료시설 이용 관련 변수 11](#_Toc40388276)

[[표 9]음주 관련 변수 11](#_Toc40388277)

[[표 10] 기타 변수 12](#_Toc40388278)

[[그림 1] 한국 우울증 진료 환자 추이, 국민건강보험 4](#_Toc40400157)

[[그림 2] 국민건강영양조사 향후 원시자료 공개계획 14](#_Toc40400158)

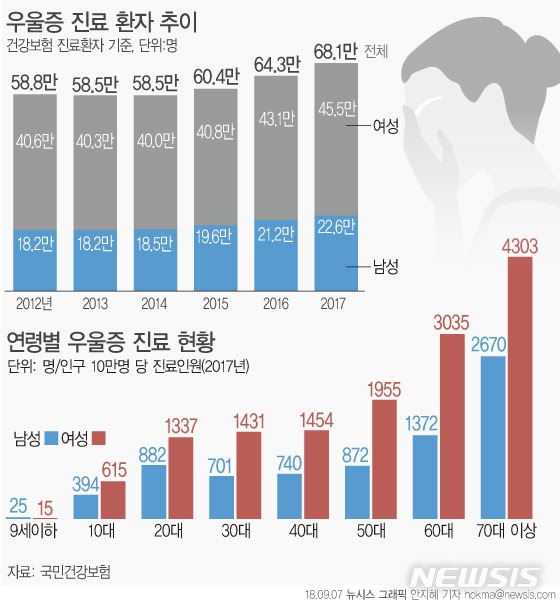
[[그림 3] Decision Tree Model 16](#_Toc40400159)

[[그림 4] Confusion Matrix 18](#_Toc40400160)

1. **Introduction**
2. 과제 수행 배경 및 동기

현대 사회에서 우울증은 심각한 문제로 여겨진다. 특히 우울증으로 인한 사회, 경제적 심각성은 고혈압, 당뇨병, 심장병 등 주요 만성질환과 비등하거나 더 심한 것으로 조사되고[[1]](#endnote-1) 있다.

국민건강영양조사 제7기 제3차년도(2018년)의 원시 자료를 확인해 보는 과정에서 여성 3334명 중 201명(6.02%)이 우울증 진단 척도인 PHQ-9의 점수가 10점 이상인 것을 확인하였고, 그 중에서도 특히 독신 성인 여성의 경우, 465명 중 50명(10.8%)에 달한다. 여기서 PHQ-9이 10점 이상인 경우 경도 우울증인 것으로 판단되고[[2]](#endnote-2), PHQ-9이 유의미한 우울증 판단 척도인 것은 이미 검증된 바 있다.[[3]](#endnote-3)



[그림 1] 한국 우울증 진료 환자 추이, 국민건강보험

1. **문제 정의 및 도출**

독신여성의 우울증은 높은 비율을 차지하고 있으나 해당 segment의 우울증 예측모델은 존재하지 않는다. 앞에서 언급한 바와 같이, PHQ-9 우울증 판단 척도의 우수성과 타당성은 이미 입증되었으므로 PHQ-9 점수를 target attribute로 삼기로 한다. 여기에 정량적이고 객관적인 attribute들 및 객관적 지표를 이용해 독신여성 우울증을 예측하고자 한다.

Classification Analysis 기법 중 Decision Tree model을 이용해 경도 우울증인(PHQ-9의 점수가 10점 이상) 독신 성인 여성의 객관적 Pattern을 파악하여 향후 경도 우울증으로 발전할 가능성이 높은 인원을 예측하는 모델을 만든다. 이를 이용해 성인 독신 여성의 삶 속 우울증 유발 인자들을 분석해 해당 요인들에 대한 정부 차원의 지원을 보다 효율적으로 제공하는데 도움을 주고자 한다.

1. **기존 연구 및 기술의 문제점**

기존 우울증을 측정하는 도구 중 널리 이용되는 것으로는 PHQ-9(Patient Health Questionnaire-9), BDI(Beck Depression Inventory), HDRS(Hamilton Depression Rating Scale) 등이 있다. 이러한 평가 척도들의 설문 항목은 주로 피검자의 주관적 기분 및 생각을 묻는 질문으로 구성되어 있다. 주관적 질문을 통해 우울증 유병자를 높은 정확도로 선별해낼 수는 있지만, 삶 속의 우울증 유발 요인들을 찾아내 우울증을 예측하기에는 적절치 않다. 즉, 해당 척도들에서 일정 수준 이상의 점수가 나왔다는 것은 이미 우울증세를 보이고 있다는 것을 의미하지만, 본 연구에서 진행하고자 하는 것은 Classification Analysis 방법 중 Decision Tree model을 이용해 경도 우울증으로 발전할 가능성이 높은 사람들의 Pattern들을 파악하고 예측하는 것이다.

1. **제안하는 방법의 개요**

연구 대상은 국민건강영양조사 제7기 제3차년도(2018년)에서 설문에 참여한 성인 독신 여성을 대상으로 하였다. 국민건강영양조사 제7기 제3차년도(2018년) 원시자료에 존재하는 736개의 Attribute들 중 기 등재된 논문을 이용해 부족한 Domain Knowledge를 보완해 관련 있다고 검증된 Attribute들을 뽑아낸 후, 가장 Information Gain이 높은 attribute들을 Feature Subset selection을 통해 선별했다. 이 과정에서 rpart를 이용한 wrapper approach를 시도했다. Feature subset selection을 통해 선별된 attribute들을 rpart로 classification analysis를 실시해 Decision Tree model을 수립했다.

1. **논문 주요 결과의 개요**

Classification analysis 기법 중 Decision Tree model을 이용해 모델링을 한 결과, root node로는 기초생활수급여부(allownc) attribute가 사용되었다. 기생활수급여부가 있을 때(1)와 없을 때(0) 모두 그 다음 parent node는 필요 의료서비스 충족 여부(medi\_ser\_unsat) attribute가 사용되었다.

첫 Decision Tree model을 수립하고, CP(complexity parameter)를 정해 Pruning을 하여 총 6개의 depth로 이루어진 Decision Tree model을 수립했다. 이 때의 Accuracy는 0.9135이다. 그 중에서 관심 있는 우울증 유병환자의 예측 정확도를 확인하기 위해 Confusion Matrix를 계산했다.

따라서 본 연구는 소득이 낮은 성인 1인 가구 여성을 대상으로 향상된 의료 서비스를 제공하고, 의료보험의 가입을 권장 및 장려하는 정부 차원 지원의 필요성을 시사한다. 뿐만 아니라 해당 연구를 통해 지역사회에서 성인 독신 여성의 우울증 예방을 위한 정책과 프로그램을 개발하는데 실질적 지표를 제시한다는데 의의를 가진다.

1. **METHODS**
2. **연구 대상**

본 연구는 2018년도에 시행된 제7기 3차년도(2018년) 국민건강영양조사에 참여한 만 19세 이상 성인 독신 여성 499명을 대상으로 하였다.

국민건강영양조사는 보건복지부와 질병관리본부가 전국 규모의 건강 및 영양을 조사한 통계자료이다. 해당 조사는 전 국민을 대상으로 하여 건강설문조사, 검진조사, 영양조사를 1년 단위로 진행한다.

1. **Data Understanding**

Strength: 원시자료 구성이 건강설문조사 ⋅ 검진조사 ⋅ 영양조사 별로 구분이 잘 되어 있어 필요한 데이터에 대한 접근성과 활용성이 뛰어나다. 또한 성별과 가구원 수 등의 분류가 잘 이루어져 있어 우리가 필요한 독신 여성에 대한 데이터를 추출해내기 용이하다. 또한 질병관리본부 건강영양조사과에서 공개한 공공 데이터이므로 해당 데이터 전반에 대한 신뢰도가 높다.

Limitation: 제7기로 지정된 16년, 17년, 18년도의 설문 항목들이 서로 동일하지 않아 부재하는 항목이 존재해 일부 데이터를 사용할 수 없다. 다른 기에 해당하는 데이터를 가져올 때는 가중치를 처리해야하는 어려움이 있다.

Benefit: 공공 데이터이므로 사용시 별도의 비용이 발생하지 않는다. 2018년도의 데이터뿐 아니라 같은 제7기로 지정한 2016년, 2017년도의 데이터도 무료로 가져와 가중치를 적용해 분석할 수 있다.

Cost: 736개의 column(2018년 기준) 수와 다수의 결측치를 처리하는 데에 많은 시간과 노력이 소요된다. 예를 들어 성인인 경우와 청소년인 경우 응답해야 하는 설문 항목이 다르기 때문에 각 Data Object마다 특정 Attribute value의 Missing Value가 상당수 존재한다.

1. **연구에 사용한 변수 및 Pre-Processing**

Data의 Dimensionality를 줄여 Curse of Dimensionality를 피하기 위해 총 2가지의 Feature Subset Selection 기법을 사용했다.

* Irrelevant features elimination

 연구에 사용할 feature를 selection 하는 과정에서 GIGO(Garbage In, Garbage Out) 현상을 방지하기 위해서 우리는 전처리에 많은 노력을 기울였다. 전체 Feature Set에서 Data Mining Technique에 more suitable 한 Feature를 선택하기 위해 736개의 모든 Feature를 분석하는 것이 가장 이상적이다(Ideal Approach). 하지만 본 연구를 진행하는데 있어 시간의 제약이 있기 때문에 최적의 방법을 사용해야 했다. 따라서 domain knowledge를 활용하여 본 연구의 목적인 “성인 독신 여성의 우울증”과 관련 있는 Feature들을 논문에 근거하여 선별함으로써Irrelevant feature를 제거하였다. 선별한 Feature들은 다음과 같다.

* 우울증 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| PHQ-9 점수 | mh\_PHQ\_S |
| 우울증 의사 진단 여부 | DF2\_dg |
| 우울증 유병 여부 | DF2\_pr |

[표 1] 우울증 관련 변수

김성래 외 6명의 연구 (2011) [[4]](#endnote-4)에 따르면 PHQ-9이 우울증 선별에 우수한 모델이기 때문에 우울증 예측 모델의 종속변수로 사용했다. 우울증 의사 진단 여부와 우울증 유병 여부는 우울증 예측 모델의 목적에 맞는 Feature이므로 선별했다.

* 만성질환 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 만성질환 의사진단 가족력 여부 | HE\_fh |
| 고혈압 의사진단 여부(부)  고혈압 의사진단 여부(모)  고혈압 의사진단 여부(형제자매) | HE\_HPfh1  HE\_HPfh2  HE\_HPfh3 |
| 고지혈증 의사진단 여부(부)  고지혈증 의사진단 여부(모)  고지혈증 의사진단 여부(형제자매) | HE\_HLfh1  HE\_HLfh2  HE\_HLfh3 |
| 허혈성심장질환 의사진단 여부(부)  허혈성심장질환 의사진단 여부(모)  허혈성심장질환 의사진단 여부(형제자매) | HD\_IHDfh1  HD\_IHDfh2  HD\_IHDfh3 |
| 뇌졸증 의사진단 여부(부)  뇌졸증 의사진단 여부(모)  뇌졸증 의사진단 여부(형제자매) | HE\_STRfh1  HE\_STRfh2  HE\_STRfh3 |
| 당뇨병 의사진단 여부(부)  당뇨병 의사진단 여부(모)  당뇨병 의사진단 여부(형제자매) | HE\_DMfh1  HE\_DMfh2  HE\_DMfh3 |
| 갑상선질환 의사진단 여부(부)  갑상선질환 의사진단 여부(모)  갑상선질환 의사진단 여부(형제자매) | HE\_THfh1  HE\_THfh2  HE\_THfh3 |
| B형 간염 의사진단 여부(부)  B형 간염 의사진단 여부(모)  B형 간염 의사진단 여부(형제자매) | HE\_HBfh1  HE\_HBfh2  HE\_HBfh3 |
| 2년간 암 검진 여부 | BH2\_61 |

[표 2] 만성질환 관련 변수

박종삼의 연구 (2018)[[5]](#endnote-5)에 따르면 만성질환의 발생이 우울증상에 영향을 미치는 요인임을 확인하였으며 만성질환의 발생을 줄이고 예방함으로써 미래에 발생할 수 있는 우울증을 예방할 수 있음을 확인하였다. 또한 만성질환 중에서도 당뇨, 고혈압, 심혈관질환, 뇌혈관 질환이 우울증상에 영향을 미치는 정도에 차이가 있으며 암과 뇌혈관질환의 신규 발생 시 우울증상에 유의미한, 영향을 미치고 있음을 확인하였다.

구정회 외 5명의 연구 (2004)[[6]](#endnote-6) 뇌졸중 후 우울증은 흔하게 발생하는 합병증으로 1999년 Gillen 등 1)은 그 발생 빈도에 대해 문헌에 따라 25∼79%로다 양하게 보고되고 있으며, 이러한 결과는 대상의 선별, 우울증을 평가하는 도구의 선택과 같은 연구 방법에 따른 차이로 설명할 수 있다고 한다.

* 소득 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 기초생활 수급 여부 | allownc |
| 소득 | ainc |
| 식생활형편 | LF\_SAFE |
| 경제활동 상태 | EC1\_1 |

[표 3] 소득 관련 변수

정은희의 연구(2015)[[7]](#endnote-7)에 따르면 전체 성인기 집단의 우울개선을 위해서는 소득 수준의 향상이 주요한 개입 전략으로 논의될 필요가 있을 뿐만 아니라, 동시에 소득 수준 개선을 위해서도 우울증상의 완화가 주요개입전략으로써 재고될 필요가 있다는 것을 알 수 있다. 따라서 소득 관련 Feature들을 선별했다.

* 신체 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 비만 유병여부(19세 이상) | HE\_obe |
| 체질량지수 | HE\_BMI |

[표 4] 신체 관련 변수

이혁 외 5명의 연구(2016)[[8]](#endnote-8)에 따르면 성인의 비만 및 BMI는 우울증에 유의미한 영향을 미친다는 것을 알 수 있다.  따라서 비만 및 BMI 관련 Feature들을 선별했다.

* 의료 서비스 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 필요 의료 서비스 미충족 여부 | M\_2\_yr |
| 민간의료 보험 가입 여부 | npins |

[표 5] 의료 서비스 관련 변수

박혜민의 연구(2016)[[9]](#endnote-9)에 따르면 건강보험 및 의료급여가 우울증에 영향을 미친다. 따라서 의료 서비스 미충족 여부와 민간의료 보험 가입 여부 Feature를 선별했다.

* 업무 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 일 중강도 신체활동 여부 | BE3\_81 |
| 일 고강도 신체활동 여부 | BE3\_71 |
| 주당 평균 근로시간 | EC\_wht\_23 |

[표 6] 업무 관련 변수

김소희 외 2명의 연구(2014)[[10]](#endnote-10)에 따르면 업무와 관련된 고강도, 중강도 신체활동은 모두 우울증 발생과 유의미한 연관성을 가진다. 따라서 강도별 신체활동 여부와 주당 평균 근로시간 Feature를 선별했다.

* 여가 활동 변수

|  |  |
| --- | --- |
| 장소활동 신체활동 여부 | BE3\_92 |
| 여가\_고강도 신체활동 여부 | BE3\_75 |
| 여가\_중강도 신체활동 여부 | BE3\_85 |
| 하루 60분 이상 신체활동 실천 일수 | BE9 |

[표 7] 여가 활동 변수

이호길의 연구(2016)[[11]](#endnote-11)에 따르면 여가활동 및 신체활동에 대한 지각이 높을수록 우울증에 대한 영향력이 낮다. 따라서 여가 및 운동 관련 신체활동과 관련된 변수들을 선별했다.

* 의료시설 이용 관련 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 1년간 입원이용 여부 | MH1\_yr |
| 입원이용 횟수 | MH1\_1 |
| 외래이용 횟수 | MO1\_1 |

[표 8] 의료시설 이용 관련 변수

서은원의 연구(2016)[[12]](#endnote-12)에 따르면 우울증은 만성질환자의 내원일수, 외래총진료비에 통계적으로 유의한 영향을 미쳤다. 따라서 입원 및 외래 관련 Feature들을 선별했다.

* 음주 관련 변수

|  |  |
| --- | --- |
| **변수명** | **음주 우울증코드** |
| 한 번에 마시는 음주량 | BD2\_1 |
| 폭음 빈도 | BD2\_31 |

[표 9]음주 관련 변수

김예리의 연구 (2018)[[13]](#endnote-13)에 따르면 부정적인 음주습관으로의 변화에 주목하며 우울증 예방에 힘써야 할 것으로 사료된다. 또한, 우울증 예방을 위해 만성질환이 없는 비교적 건강한 사람들의 음주습관이 악화되지 않도록 관리하는 것이 중요하다고 하였다. 따라서 음주 관련 Feature들을 선별했다.

기타 변수

|  |  |
| --- | --- |
| **변수명** | **코드** |
| 교육수준 재분류 코드 | edu |
| 활동제한 여부 | LQ4\_00 |
| 나이대 별 분류 | age |

[표 10] 기타 변수

이태성의 연구(2012)[[14]](#endnote-14)에 따르면 교육수준에 따른 우울증상을 보면 교육수준이 높아질수록 우울 증상은 낮아지는 경향을 보였다. 따라서 교육수준 관련 Feature를 선별했다. 또한, 이도연, 남승민의 연구(2019)[[15]](#endnote-15)에 따르면 활동제한을 받았을 때 그렇지 않은 대상자보다 우울증이 약 3.37배 높다. 따라서 활동제한 여부 Feature를 선별했다.

② Redundant features elimination

* 우울증 관련 변수

PHQ-9 [mh\_PHQ-S]는 앞서 말했듯이 우울증 유병자를 선별하는 우수한 모델이기 때문에 다른 우울증 관련 변수들은 redundant features로 판단해 제거했다.

* 소득 관련 변수

Numeric Attribute인 소득 [ainc]을 사용하지 않고도 기초생활 수급여부 [allonwc]를 가지고도 분석할 수 있다고 판단해 소득 관련 변수 중에서는 [ainc] Feature를 제거했다.

* 신체 관련 변수

Numeric Attribute인 체질량지수 [HE\_BMI]를 사용하지 않고도 비만 유병여부만으로 비만과 우울증의 관계를 설명할 수 있다고 판단해 체질량지수 Feature를 제거했다.

* 의료시설 이용 관련 변수

입원이용 횟수 [MH1\_1] 변수가 1년간 입원이용 여부 [MH1\_yr] 변수를 대변할 수 있고 우울증과의 관계를 설명하는데 더 Suitable한 Feature라고 판단해 1년간 입원이용 여부 Feature를 제거했다.

Redundant한 변수를 제거한 후 남은 변수들에 대한 전처리를 다음과 같이 실시하였다.

* 1. PHQ-9 [mh\_PHQ\_S]

 PHQ-9의 Attribute value는 0~27 사이인 것을 확인했고, PHQ-9의 점수가 10점 이상은 경도 우울증에 해당한다는기존의 연구를 기반으로, value가 10 이상인 값을 1로, 10 미만인 값을 0으로 할당하였다. 그 후 변수명을 알아보기 쉽도록 “phq\_score”로 변경하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Interval에서 Nominal로 변경하였다.

* 1. 만성질환 의사진단 가족력 여부 [HE\_fh]

 만성질환 의사진단 가족력 여부 Attribute의 value는 0. 아니오, 1. 예, 9. 모름으로 구성되어 있다. 9를 NA로 처리하였다. 그 후 변수명을 알아보기 쉽도록 “chr\_dis\_fh”로 변경하였다.

* 1. 고혈압 의사진단 여부(부) [HE\_HPfh1]

 고혈압 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 고혈압 의사진단 여부(모) [HE\_HPfh2]

 고혈압 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 고혈압 의사진단 여부(형제자매) [HE\_HPfh3]

 고혈압 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8. 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. 고지혈증 의사진단 여부(부) [HE\_HLfh1]

 고지혈증 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 고지혈증 의사진단 여부(모) [HE\_HLfh2]

 고지혈증 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 고지혈증 의사진단 여부(형제자매) [HE\_HLfh3]

 고지혈증 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8: 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. 허혈성심장질환 의사진단 여부(부) [HE\_IHDfh1]

 허혈성심장질환 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 허혈성심장질환 의사진단 여부(모) [HE\_IHDfh2]

 허혈성심장질환 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 허혈성심장질환 의사진단 여부(형제자매) [HE\_IHDfh3]

 허혈성심장질환 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8. 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. 뇌졸중 의사진단 여부(부) [HE\_STRfh1]

 뇌졸중 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 뇌졸중 의사진단 여부(모) [HE\_STRfh2]

 뇌졸중 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 뇌졸중 의사진단 여부(형제자매) [HE\_STRfh3]

 뇌졸중 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8. 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. 당뇨병 의사진단 여부(부) [HE\_DMfh1]

 당뇨병 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 당뇨병 의사진단 여부(모) [HE\_DMfh2]

 당뇨병 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 당뇨병 의사진단 여부(형제자매) [HE\_DMfh3]

 당뇨병 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8. 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. 갑상선질환 의사진단 여부(부) [HE\_THfh1]

 갑상선질환 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 갑상선질환 의사진단 여부(모) [HE\_THfh2]

 갑상선질환 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. 갑상선질환 의사진단 여부(형제자매) [HE\_THfh3]

 갑상선질환 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8. 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. B형간염 의사진단 여부(부) [HE\_HBfh1]

 B형간염 의사진단 여부(부)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. B형간염 의사진단 여부(모) [HE\_HBfh2]

 B형간염 의사진단 여부(모)의 Attribute value는 0. 아니오, 1. 예, 9. 모름, 무응답인 것을 확인했고 value가 9인 값들을 NA로 처리했다.

* 1. B형간염 의사진단 여부(형제자매) [HE\_HBfh3]

 B형간염 의사진단 여부(형제자매)의 Attribute value는 0. 아니오, 1. 예, 8. 비해당, 9. 모름, 무응답인 것을 확인했고 value가 8, 9인 값들을 NA로 처리했다.

* 1. 기초 생활 수급 여부 [allownc]

 기초 생활 수급 여부 Attribute의 value는 10: 수급 경험 있음, 20: 수급 경험 없음, 99: 모름/무응답으로 구성되어 있다. 10은 1로 할당하였고, 20은 0에 할당하였다. 또한 모름/무응답의 경우(99)는 NA로 처리하였다.

* 1. 식생활 형편 [LF\_SAFE]

 식생활 형편 Attribute value는 1. 충분한 양과 다양한 음식을 먹을 수 있었다. 2. 충분한 양의 음식을 먹을 수 있었으나, 다양한 음식은 먹지 못했다. 3. 경제적으로 어려워 가끔 먹을 것이 부족했다. 4. 경제적으로 어려워 자주 먹을 것이 부족했다. 9. 모름/무응답인 것을 확인하였고 value가 1과 2인 값들을 0으로, value가 3과 4인 값들을 1로 할당하였고 value가 9인 값들을 NA로 처리하였다.

* 1. 경제활동 상태 [EC1\_1]

 경제활동 상태의 Attribute value는 1. 예(취업자), 2. 아니오 (실업자, 비경제활동인구), 8. 비해당 (만 15세 미만), 9. 모름, 무응답인 것을 확인했고, 경제활동을 하는 사람인 value가 1인 값을 1로, 경제활동을 하지 않는 사람인 value가 2인 값을 0으로 할당하였다.

* 1. 비만 유병 여부[HE\_obe]

 비만 유병 여부 attribute의 value는 1: 저체중, 2: 정상, 3: 비만 전단계, 4: 1단계비만, 5: 2단계비만, 6: 3단계 비만으로 이루어져 있다. 이 중 value가 1, 3, 4, 5, 6인 값들을 하나로 묶어 비정상(1), value가 2인 값을 정상(0)으로 할당하면서 변수명을 정상체중여부로 해석하였다. 그 후 변수명을 알아보기 쉽도록 “norm\_weight”로 변경하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ordinal에서 Nominal로 변경하였다.

* 1. 필요 의료서비스 미충족 여부 [M\_2\_yr]

 필요 의료서비스 미충족 여부 attribute의 value는 1: 예, 2: 아니오, 3: 병의원 진료가 필요한 적이 없었다, 9: 모름/무응답로 구성되어있다. value가 2인 값을 0으로 할당하였다. 또한 value가 3, 9인 값을 NA로 처리하였다. 그 후 변수명을 알아보기 쉽도록 “medi\_ser\_unsat”로 변경하였다.

* 1. 민간의료보험가입여부 [npins]

 민간의료보험가입여부 attribute의 value는 1: 예, 2: 아니오, 9: 모름/무응답으로 구성되어 있다.value가 1인 값은 0으로, value가 2인 값은 1에 할당하였고, value가 9인 값은 NA로 처리하였다. 그 후 변수명을 알아보기 쉽도록 “pri\_medi\_ins”로 변경하였다.

* 1. 변형근로시간 [EC\_wht\_5]

 변형근로시간 Attribute value는 1. 주간 근무, 2. 저녁 근무 (pm14:00~24:00), 3. 밤 근무(pm21:00~익일 am8:00), 4. 주야간 규칙적 교대근무, 5. 24시간 교대근무, 6. 분할근무 (하루 근무시간대 2개 이상), 7. 불규칙 교대근무, 8. 기타, 88. 비해당(최근 1년 동안 일을 하지 않음), 99. 모름 무응답인 것을 확인하였고 value가 99인 값을 NA로 처리하였다.

* 1. 주당 평균 근로시간 [EC\_wht\_23]

 주당 평균 근로시간 Attribute value는 1~112 값과, 888. 비해당(최근 1년 동안 일을 하지 않음), 999. 모름, 무응답인 것을 확인하였고 value가 52이하인 값을 0으로 처리하였고, value가 52초과인 값을 1로 처리하였으며 value가 888, 999인 값들을 NA로 처리하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ratio에서 Nominal로 변경하였다.

* 1. 장소이동 신체활동 여부 [BE3\_91]

 장소이동 신체활동 여부 Attribute value는 1. 예, 2. 아니오, 8. 비해당(소아, 청소년), 9. 모름, 무응답인 것을 확인하였고 value가 2인 값들을 0으로 할당하였고 value가 8, 9인 값들을 NA로 처리하였다.

* 1. 여가\_고강도 신체활동 여부 [BE3\_75]

 여가\_고강도 신체활동 여부 Attribute value는 1. 예, 2. 아니오, 8. 비해당(소아, 청소년), 9. 모름, 무응답인 것을 확인하였고 value가 2인 값들을 0으로 할당하였고 value가 8, 9인 값들을 NA로 처리하였다.

* 1. 여가\_중강도 신체활동 여부 [BE3\_85]

 여가\_중강도 신체활동 여부 Attribute Value는 1. 예, 2. 아니오, 8. 비해당(소아, 청소년), 9. 모름, 무응답인 것을 확인하였고 value가 2인 값들을 0으로 할당하였고 value가 8, 9인 값들을 NA로 처리하였다.

* 1. 하루 60분 이상 신체활동 실천 일수 [BE9]

 하루 60분 이상 신체활동 실천 일수 Attribute value는 1. 최근 7일 동안 없다, 2. 주 1일, 3. 주 2일, 4. 주 3일, 5. 주 4일, 6. 주 5일, 7. 주 6일, 8. 주 7일, 88. 비해당(소아, 성인), 9. 모름, 무응답인 것을 확인하였고 value가 1인 값들을 0으로, value가 2, 3인 값들을 1로, value가 4, 5인 값들을 2로, value가 6, 7인 값들을 3으로, value가 8인 값들을 4로 할당하였고 value가 88, 99인 값들을 NA로 처리하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ordinal에서 Nominal로 변경하였다.

\*국민건강영양조사 원시자료 이용지침서 제7기 (2016-2018)에는 8. 비해당 (소아, 성인)이라고 잘못 기재되어 있다.

* 1. 일 중강도 신체활동 여부 [BE3\_81]

 일 중강도 신체활동 여부 Attribute value는 1. 예, 2. 아니오, 8. 비해당(소아, 청소년), 9. 모름, 무응답인 것을 확인하였고 value가 2인 값들을 0으로 할당하였고 value가 8, 9인 값들을 NA로 처리하였다.

* 1. 일 고강도 신체활동 여부 [BE3\_71]

 일 고강도 신체활동 여부 Attribute value는 1. 예, 2. 아니오, 8. 비해당(소아, 청소년), 9. 모름, 무응답인 것을 확인하였고 value가 2인 값들을 0으로 할당하였고 value가 8, 9인 값들을 NA로 처리하였다.

* 1. 입원이용 횟수 [MH1\_1]

 입원이용 횟수 Attribute value는 1~20의 값과 88. 비해당, 99. 모름, 무응답인 것을 확인하였다. value가 5미만인 값들을 0으로 할당하였고 value가 5이상인 값들을 1로 할당하였다. value가 88, 99인 사람을 NA로 처리하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ratio에서 Nominal로 변경하였다.

* 1. 외래이용 횟수 [MO1\_1]

 외래이용 횟수 Attribute value는 1~12의 값과 88. 비해당, 99. 모름, 무응답인 것을 확인하였다. value가 5미만인 값들을 0으로 할당하였고 value가 5이상인 값들을 1로 할당하였다. value가 88, 99인 값을 NA로 처리하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ratio에서 Nominal로 변경하였다.

* 1. 한 번에 마시는 음주량 [BD2\_1]

 한 번에 마시는 음주량 Attribute value는 1. 1-2잔, 2. 3-4잔, 3. 5-6잔, 4. 7-9잔, 5. 10잔 이상, 8. 비해당, 9. 모름, 무응답인 것을 확인하였다. value가 1, 2인 값들을 0으로 할당하였고 value가 3, 4, 5인 값들을 1로 할당하였다. value가 8, 9인 값을 NA로 처리하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ordinal에서 Nominal로 변경하였다.

* 1. 폭음빈도[BD2\_31]

 폭음빈도 Attribute value는 1. 전혀 없음, 2. 월1회미만, 3. 월1회정도, 4. 주 1회정도, 5. 거의 매일, 8. 비해당, 9. 모름, 무응답인 것을 확인하였다. Value가 1, 2인 사람들을 0으로 할당하였고, value가 3인 사람을 1로, value가 4인 사람을 2로, value가 5인 사람을 3으로 할당하였다. value가 8, 9인 사람을 NA로 처리하였다.

* 1. 교육수준 재분류 코드 [edu]

 교육수준의 Attribute value는 1. 초졸 이하, 2. 중졸, 3. 고졸, 4. 대졸이상인 것을 확인했고, value가 고졸 이하를 의미하는 값인 1,2,3을 1로, 대졸을 의미하는 값인 4를 0으로 할당하였다. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Ordinal에서 Nominal로 변경하였다.

* 1. 활동 제한 여부 [LQ4\_00]

 활동 제한 여부 attribute의 value는 1. 예, 2. 아니오, 9. 모름/무응답으로 구성되어 있다. value가 1인 값을 1로, value가 2인 값을 0으로 할당했으며, value가 9인 값을 NA로 처리하였다. 그 후 변수명을 알아보기 쉽도록 “act\_rest”로 변경하였다.

* 1. 만나이 [age]

 만나이의 Attribute value는 1~80으로 구성된 것을 확인했고, Value가 9 이하인 값을 “0s”로, value가 10이상 19 이하인 값을 “10s”로, value가 20이상 29이하인 값을 “20s”로, value가 30이상 39이하인 값을 “30s”로, value가 40이상 49이하인 값을 “40s”로, value가 50이상 59이하인 값을 “50s”로, value가 60이상인 값을 “60s”로 할당하였다. 이들을 Factor함수와 levels, ordered 파라미터를 사용하여 Attribute Type을 Ratio에서 Ordinal로 변경하였다.

 위의 변수 중 최적의 모델에 사용된 Pre-processing된 attribute들은 다음과 같다.

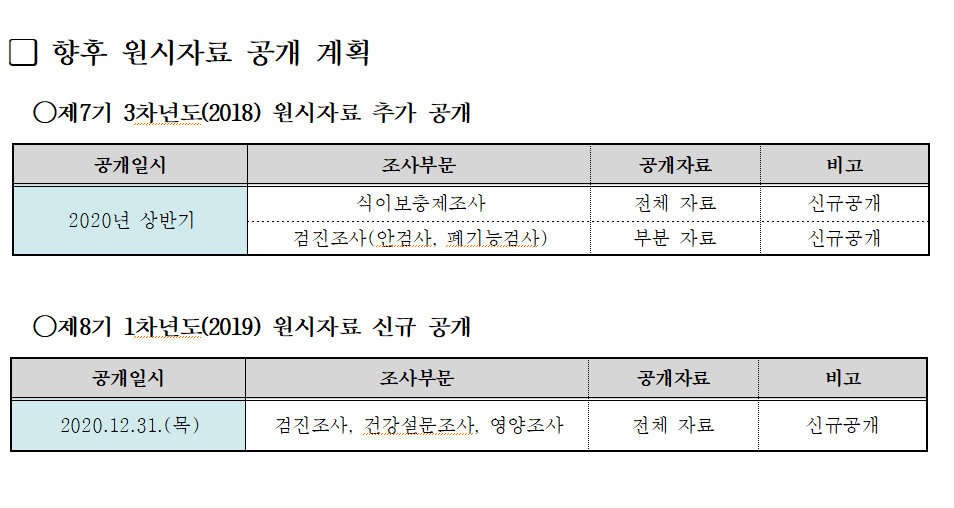
1. PHQ-9 [mh\_PHQ\_S]
2. 만성질환 의사진단 가족력 여부 [HE\_fh]
3. 기초 생활 수급 여부 [allownc]
4. 비만 유병 여부[HE\_obe]
5. 필요 의료서비스 미충족 여부 [M\_2\_yr]
6. 민간의료보험가입여부 [npins]
7. 활동 제한 여부 [LQ4\_00]

1. **분석 방법**

본 연구에서는 ILAI Framework와 CRISP-DM Process를 사용하여 Data Mining을 수행했다. ILAI Framework란 갖고 있는 Data를 면밀히 파악(Identify)하고, 그로부터 달성 가능한 분석 목적들의 목록을 만든 후에(List up), 그 중 선택된 목적을 위해 Data를 분석(Analyze)하여 발견된 패턴을 해석(Interpret)하는 것이다.  먼저, 국민건강영양조사 데이터를 면밀히 파악하고 그로부터 달성 가능한 분석 목적들의 목록을 만들었다.

① 부족한 영양소를 파악하는 분류 모델

R studio를 이용하여 식이 보충제 관련 변수를 분석해보았지만, 최근 1년 동안 2주 이상 식이보충제 복용여부[LS\_1YR] 관련 데이터만 존재하였다. 질병관리본부 홈페이지에서 찾아본 결과 [그림2]와 같이 종류나 구체적인 정보는 2020년 상반기에 공개한다고 명시되어 있었다.



[그림 2] 국민건강영양조사 향후 원시자료 공개계획

② 청소년 흡연 예방을 위한 예측 시스템

국민건강영양조사 데이터에서 청소년 흡연에 대한 현상을 잘 대변하는 모델을 수립할 만큼 표본의 수가 충분치 않고, 논문을 바탕으로 찾아낸 청소년 흡연과 유의미한 관계를 가진 변수들이 부재한 경우가 많았다. 예를 들어 흡연과 관련된 소득, 음주 여부 등에 대한 attribute들이 청소년이 아닌 성인을 대상으로 하기 때문에 타당한 모델을 도출하기 어렵다.

③ 독신 여성 우울증 예측 모델

성인 여성 우울증 유병자 중 독신 여성의 비율이 높게 나타났다. 독신 여성의 우울증 예측 모델의 부재, 증가하는 여성 우울증 유병자 수, 남성에 비해 월등히 높은 여성 우울증 유병자 수 등의 사실은 해당 분석의 타당성을 높여줬다. 가구원 수, 소득, BMI 지수 등의 attribute를 사용해 현상을 잘 대변하는 모델을 수립해 패턴을 분석한다.

따라서 본 연구에서는 Rpart 패키지를 이용해 Decision Tree model을 수립해, 독신인 여성 중 경도 우울증을 가진 사람들의 Pattern을 Describe 하였다. 또한 이를 통해 향후 경도 우울증을 겪을 가능성이 높은 독신 여성을 Predict한다.

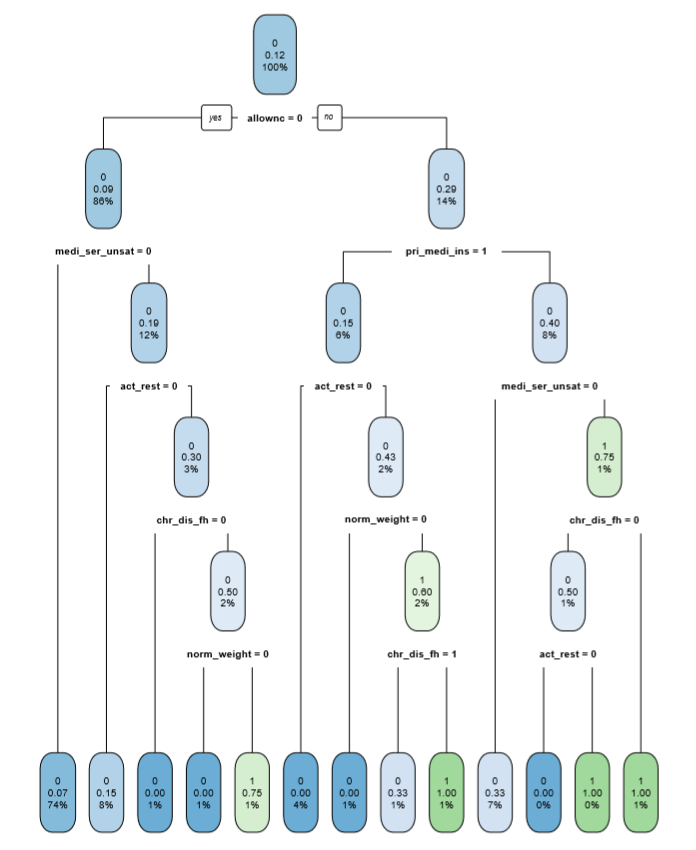
이 과정에서 추출된 data는 대표성을 가져야 한다. 이를 위해 Stratified Sampling (층화 추출법)을 “caret” 라이브러리의 createDataPartition함수를 통해 구현했다.  Target Attribute인 phq\_score를 기준으로 Train Data와 Test Data를 3대 1의 비율로 할당하였다.

본 연구에서는 target attribute인 phq\_score(경도 우울증 여부)에 영향을 미치는 변수들에 대한 Decision Tree model을 다음의 방법으로 구축하였다. “rpart” 패키지를 사용하여 위 단계에서 Sampling한 Training Data를 기반으로 Decision Tree model을 구축했으며 최소한 2개의 child node로 나눠지도록 파라미터를 설정하였다. “rpart” 패키지는 Hunt’s Algorithm을 이용하여 Decision Tree model을 그리는 방법으로, 동일한 변수를 사용하여 최적의 Decision Tree model을 찾아 “rpart.plot” 패키지를 이용하여 시각화 하였다. 이 때 사용되는 평가 지수로는 Gini 계수를 사용했다.

1. **RESULT**

Rpart 패키지를 통해 Decision Tree 모델을 수립할 때는 Hunt’s Algorithm을 사용하기 때문에 Greedy Strategy를 통해 Root node에서부터 Terminal nodes까지의 각 Depth마다 Local Optimum을 구하며 독신여성 우울증 예측 모델을 수립한다.

 분석결과는 다음 [그림 3]과 같다.



[그림 3] Decision Tree Model

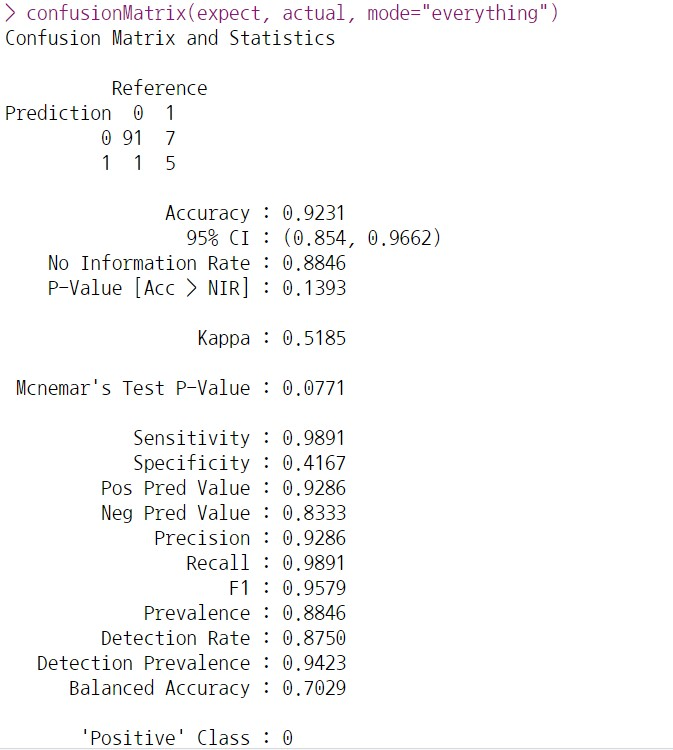
  Decision Tree 모델의 각 Node의 숫자가 나타내는 의미는 위에서부터 ‘Predicted Class’, ‘관심있는 Target Attribute Value의 Predicted Probability’, ‘해당 Node의 Data Object들이 Training Set에서 차지하는 비율’을 의미한다.

수립 결과, Informative Attributes로는 기초 생활 수급 여부, 필요 의료 서비스 미충족 여부, 민간 의료보험 가입여부, 활동제한여부, 정상체중 여부, 만성질환 의사진단 가족력 여부가 선별되었다.

 구체적으로 살펴보면, 독신여성의 우울증 여부를 가장 잘 변별해주는 Feature는 ‘기초생활수급여부 [allownc]’ 였다. 즉, 수립된 Decision Tree를 봤을 때, 전체 Training set의 100%가 속해 있는 Root node는 ‘기초생활수급 여부’로, 수급 경험이 없는 여성은 Training set의 86%, 수급 경험이 있는 여성은 14%에 해당된다.

Root node에서 Yes(경험 없음)가 나온 Data object들은 다음 Parent node의 Informative attribute인 ‘필요 의료서비스 미충족 여부 [medi\_ser\_sat]’로 분류되었다. 미충족 여부가 Yes인 Data object들은 Training set의 74%에 해당하고, 이 Combination을 가졌을 때, Target attribute에 ‘우울증이 없다’로 할당했다. 미충족 여부가 No인 Data object들은 ‘활동제한 여부 [act\_rest]’ 로 분류되었다. 이 중, No에 해당하는 Data object들은 Training set의 8%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했다. ‘활동제한 여부’에서 Yes인 Data object들은 ‘만성질환 의사진단 가족력 여부 [chr\_dis\_fh]’로 분류되었다. 이 중, No에 해당하는 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했다. ‘만성질환 의사진단 가족력 여부 [chr\_dis\_fh]’에서 Yes인 Data object들은 ‘정상체중여부 [norm\_weight]’로 분류되었다. ‘정상체중여부’에서 No, 즉 정상체중인 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했다. Yes, 즉 정상체중이 아닌 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 있다’로 할당했다.

  Root node에서 No(경험 있음)가 나온 Data object들은 ‘민간의료보험 가입여부 [pri\_meid\_ins]’로 분류된다. ‘민간의료보험 가입여부’에서 Yes인 Data object들은 Training set의 6%에 해당되며, 다시 ‘활동제한여부’로 분류되었다. ‘활동제한여부’에서 No인 Data object들은 Target attribute에 ‘우울증이 없다’로 할당했다. ‘활동제한여부’에서 Yes인 Data object들은 ‘정상체중여부’로 분류되었고, 정상체중인 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했다. 정상체중이 아닌 Data object들은 다시 ‘만성질환 의사진단 가족력 여부’로 분류되었다. 가족력 여부에서 Yes인 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했으며, No인 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 있다’로 할당했다. ‘민간의료보험 가입여부’에서 No인 Data object들은 Training set의 8%에 해당하며, ‘필요 의료 서비스 미충족 여부’로 분류되었다. 의료 서비스가 충족된 Data object들은 Training set의 7%에 해당되며, Target attribute에 ‘우울증이 없다’로 할당했다. 의료 서비스가 충족되지 않은 Data object들은 ‘만성질환 의사진단 가족력 여부’로 분류되었는데 가족력이 있는 Data object들은 Training set의 1%에 해당되며, Target attribute에 ‘우울증이 있다’로 할당했다. 가족력이 없는 Data object들은 다시 ‘활동제한여부’로 분류되는데, No일 경우에는 Target attribute에 ‘우울증이 없다’를, Yes 일 경우에는 Target attribute에 ‘우울증이 없다’로 할당했고, 두 Combination은 Training set의 1%에 해당한다.



[그림 4] Confusion Matrix

Confusion Matrix 함수를 이용하여 예측한 결과를 Accuracy, Precision, Recall, F-measure 등의 척도로 평가하였다. 평가한 결과는 구체적으로 다음과 같다.

예측력의 객관적 평가를 위해 의사결정나무분석의 정확도를 확인한 결과 F-measure(2 \* {Precision\*Recall/(Precision + Recall)})를 통해 Precision과 Recall을 통합하여 classifier가 얼마나 정확하게 분류하는지 판단할 수 있었다. 또한 Accuracy(Number of correct predictions/Total number of predictions)를 통해 정확성을 한 번에 나타낼 수 있었다. 결과적으로 Accuracy는 0.9231, F-measure는 0.9579로 매우 높게 평가되었으므로 가장 Performance가 좋은 모델이라고 판단하였다.

  본 연구에서는 rpart 패키지를 통하여 각각의 변수들에 대해 지니 불순도의 기준을 사용하여 노드를 재귀적으로 분할하면서 의사결정나무를 만들어 모델을 수립하였으므로 이 모델을 Valid하다고 말할 수 있다. 또한, 분석에 필요한 attribute를 선정하는데 있어서 논문을 바탕으로 선별하였고, BP와 관련 있는 변수들에 대한 데이터를 분석에 사용하였으므로 Reliable하다고 할 수 있다.

1. **DISCUSSION**

PHQ-9은 앞서 언급했듯이 우울증 유병자를 선별하는 데 우수한 모델이다. 국민건강영양조사 제7기 제3차년도(2018년) 원시자료를 이용한 분석 결과, PHQ-9 점수가 10점 이상인, 즉 경도 우울증에 해당되는 280명의 응답자 중 의사로부터 우울증 진단을 받아 본 응답자는 60명에 불과했다. 이는 우울증이라는 질환의 인지적 취약성[[16]](#endnote-16) 때문이기도 하며, 우울증 환자들에 대한 부정적인 사회적 감정[[17]](#endnote-17) 때문이기도 하다. 이러한 우울증의 속성을 인식해 PHQ-9을 통한 우울증 선별에 그치지 않고, 우울증 유발에 영향을 주는 일상의 요소들을 찾아내 우울증을 예측하는 모델을 수립하고자 했다.

 [그림1]에 제시된 바와 같이 증가하는 우울증 유병자 수, 특히 높은 비율을 차지하는 여성 우울증 유병자에 초점을 맞추었다. 국민건강영양조사에 응답한 여성 우울증 유병자 중, 독신 여성의 비율이 높다는 사실(465명 중 50명, 10.8%)을 데이터 분석을 통해 발견했다. 따라서 본 연구는 독신 여성의 우울증 예측 모델을 수립해 PHQ-9의 점수가 10점 이상이 될 것으로 예측되는 독신 여성에게 정부 차원의 효율적 지원을 돕기 위한 목적으로 실시되었다.

  수립된 Decision Tree 모델을 통해서 성인 독신 여성의 일상생활 속 우울증 유발 요인들을 분석할 수 있었다. Pruning한 결과 총 6개의 Informative Attribute가 선별되었고, ‘정상 체중 여부’, ‘기초 생활 수급 여부’, ‘만성질환 의사진단 가족력 여부’, ‘필요 의료서비스 미충족 여부’, ‘활동 제한 여부’, ‘민간의료보험가입여부’가 이에 해당된다.

  Decision Tree 모델에서 가장 Information Gain이 높은 Attribute는 ‘기초 생활 수급 여부’이고, 그 다음으로는 ‘필요 의료서비스 미충족 여부’와 ‘민간 의료보험 가입여부’, ‘활동제한여부’가 선별되었다. 해당 분석을 통해 소득수준과 우울증은 음의 상관 관계를 가지고 있다는 것을 알 수 있다. 또한 독신 여성의 의료 서비스 및 의료보험 환경도 우울증과 음의 상관 관계를 가진다는 것을 알 수 있다. 실제로 다인 가구 성인 여성에 비해 1인 가구 성인 여성의 소득 및 미충족 의료, 즉 필요한 진료를 받지 못하는 경우가 더 많다는 연구 결과가 있다[[18]](#endnote-18). 국민건강영양조사의 설문 항목 중 ‘활동제한여부’는 심장질환, 류머티스, 치아질환, 등 질환으로 야기된 활동제한에 대한 설문이었다. 이는 성인 독신 여성의 불우한 의료 환경에서 연장된 문제로 여길 수 있다.

 따라서 본 연구는 소득이 낮은 성인 1인 가구 여성을 대상으로 향상된 의료 서비스를 제공하고, 의료보험의 가입을 권장 및 장려하는 정부 차원 지원의 필요성을 시사한다. 뿐만 아니라 해당 연구를 통해 지역사회에서 성인 독신 여성의 우울증 예방을 위한 정책과 프로그램을 개발하는데 실질적 지표를 제시한다는데 의의를 가진다.

1. **LIMITATION**
2. [결측치 대체의 한계]

결측치는 분석결과의 편향이나 통계적 검정력의 문제를 일으킨다. 이에 연구결과가 모집단의 특성을 대표하지 못하는 자료에 의존하는 문제점이 발생할 수 있다. 결측치를 보완하는 가장 완벽한 방법은 모름 및 무응답자들에 대한 재설문조사이지만 이는 시간과 비용이 많이 든다는 단점이 있다[[19]](#endnote-19). 따라서 우리는 대체 값을 평균값, 최빈값, 중앙값 등의 단일 대체 방법으로 처리하고자 했다. 그러나 이는 결측치를 해당 변수의 단일 값으로 대체하는 방법으로, 비교적 쉽게 적용하여 완전한 데이터를 얻을 수 있는 장점이 있지만 모든 결측치를 하나의 값으로 대체하기에 이후 분석 과정에서 추정치에 편차가 생길 수 있고, 표본 오차 및 분산 역시 과소 추정될 수 있다는 단점이 있다.

그 후 비율에 맞게 NA값을 대체하는 다중대체법을 사용하고자 하였다. 다중 대체법은 적은 비용과 노력으로 보다 많은 표본을 확보할 수 있는 방법으로 알려져 있다. 이는 가능한 대체 값의 분포에서 추출된 서로 다른 값으로 결측치를 처리한 복수의 데이터셋을 생성한 후 이를 데이터셋에 대해 각각 분석을 수행한다. 그 결과 얻은 모수의 추정량과 표본오차를 통합하여 하나의 분석결과를 제시하는 방법이다. 이에. 결론적으로 모름 및 무응답 등 값을 NA로 대체하고 결측치를 처리하는데 있어서는 다른 값으로 대체하지 않고 분석에 임했다. 그 결과 Decision Tree model을 그리는데 적용하는 Data set이 충분하지 않는다는 한계가 있었다.

1. [가중치 적용의 한계]

부족한 Data set을 보완하려는 목적으로 제7기 자료들 중 2016년과 2017년 국민건강영양조사의 원시자료를 가져와 분석에 포함하고자 하였다. 가중치는 표본설계 시점과 조사 시점 간 가구수 및 인구수 차이에 따른 포함오차, 불균등 추출률, 조사 미참여자의 무응답오차 등이 보정되어 목표 모집단인 우리나라 국민의 건강행태, 만성질환 유병, 식품 및 영양섭취 실태 관련 추정치의 대표성 및 정확성을 높이기 위하여 사용된다. 그러나 가중치를 적용함에 있어서 가중치 자체에 대한 이해와 수식적용에 한계에 부딪혔고 이들 데이터를 합치는 과정에서 명확한 해답을 찾지 못하여 결론적으로 적용하지 못했다.

1. 곽병규, 박중철, 이왕진, 장신해, 천동호. (2018). 한국 성인에서 우울증과 건강 관련 삶의 질: 2014년도 국민건강영양조사 자료를 이용하여. Korean J Fam Pract. 8(2), 236-243 [↑](#endnote-ref-1)
2. 곽병규, 박중철, 이왕진, 장신해, 천동호. (2018). 한국 성인에서 우울증과 건강 관련 삶의 질: 2014년도 국민건강영양조사 자료를 이용하여. Korean J Fam Pract. 8(2), 236-243 재인용 [↑](#endnote-ref-2)
3. 김성래, 신호철, 이덕철, 김철환, 성은주, 이계화, 김자영. (2011). 일차의료에서 주요우울장애 선별을 위한 PHQ-2/PHQ-9 연속선 별검사의 유용성. 스트레스, 19(4), 405-410 [↑](#endnote-ref-3)
4. 김성래, 신호철, 이덕철, 김철환, 성은주, 이계화, 김자영. (2011). 일차의료에서 주요우울장애 선별을 위한 PHQ-2/PHQ-9 연속선 별검사의 유용성. 스트레스, 19(4), 405-410에서 재인용 [↑](#endnote-ref-4)
5. 박종삼. (2018). 만성질환의 발생이 우울증상에 미치는 영향: 고령화연구패널 조사를 이용하여. 석사학위논문, 연세대학교, 서울 [↑](#endnote-ref-5)
6. 구정회, 이정우, 김성윤, 여진영, 유승돈, 전민호. (2004). 뇌졸중 입원 환자의 우울증과 기능적 회복의 상관관계. 보건연구정보센터. 8(1), 14-19 [↑](#endnote-ref-6)
7. 정은희. (2015). 한국사회 우울과 소득의 종단적 상호관계: 사회 원인가설과 선택가설 검증을 중심으로. 박사학위논문, 서울대학교, 서울 [↑](#endnote-ref-7)
8. 이혁, 제민지, 이상옥, 장철훈, 이창형, 김성수. (2016). 성인의 비만과 체중지각이 신체존중감, 자신감, 우울증상에 미치는 영향. 의료커뮤니케이션, 11(1), 17-26 [↑](#endnote-ref-8)
9. 박혜민. (2017). 건강보험자료를 활용한 인구학적, 도농 및 건강보험 종류에 따른 우울증 유병률 차이. 석사학위논문, 성균관대학교, 서울 [↑](#endnote-ref-9)
10. 김소희, 전무경, 강병두. (2017). 신체 활동과 우울증의 연관성: 국민건강영양조사 2014년 자료 이용. Korean J Fam Pract. 7(4), 610-613 [↑](#endnote-ref-10)
11. 이호길. (2016). 여가활동이 중년기 삶의 우울과 심리적 안녕감이 미치는 영향. 동북아관광연구. 12(2), 129-148 [↑](#endnote-ref-11)
12. 서은원. (2016). 우울증 유무에 따른 만성질환자의 의료 이용차이 분석: 고혈압, 당뇨를 중심으로. 석사학위논문, 연세대학교, 서울 [↑](#endnote-ref-12)
13. 김예리. (2018). 음주태도 변화가 우울증상에 미치는 영향: 중 중·노년 인구 패널 데이터 기반 연구. 석사학위논문, 숙명여자대학교, 서울 [↑](#endnote-ref-13)
14. 이태성. (2012). 지역사회 주민들의 사회경제적 수준과 우울증상과의 관련성. 석사학위논문, 전남대학교 대학원, 광주 [↑](#endnote-ref-14)
15. 이도연, 남승민. (2019). 로지스틱 회귀분석을 이용한 뇌졸중 환자의 활동제한이 우울증 및 삶의 질에 미치는 영향. J Korean Soc Phys Med. 14(4), 125-132 [↑](#endnote-ref-15)
16. 조지현, 조용래. (2020). 사회불안과 우울증상에 공통된 인지적 취약성과 고유한 인지적 취약성 검증. 인지행동치료. 20(1), 47-69 [↑](#endnote-ref-16)
17. 조현주, 현명호, 임현우, 조선진, 채정호, 전태연, 방명희, 기백석 (2007). 우울증에 대한 인식 질문지 개발을 위한 예비조사. 한국심리학회 학술대회 자료집, 2007(1), 240-24 [↑](#endnote-ref-17)
18. 최현주, 김미종. (2019). 1인 가구 성인 여성과 다인 가구 성인 여성의 건강행태, 의료서비스이용과 건강관련 삶의 질. Korean J Women Health Nurs. 25(3), 299-314 [↑](#endnote-ref-18)
19. 고길곤, 탁현우. (2016). 설문자료의 결측치 처리방법에 관한 연구: 다중대체법과 재조사법을 중심으로. 행정논총, 54(4), 291-319 [↑](#endnote-ref-19)