

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

Final Report

<성인 저근로군의 우울증 예측 모델>

발표 영상:

목 차

1. Introdduction 4
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 7
8. 연구 대상
9. Data Understanding
10. 연구에 사용한 변수 및 Pre-Processing
    1. Cluster Analysis에 사용된 변수 및 Pre-Processing
    2. Classification Analysis에 사용된 변수 및 Pre-Processing
11. 분석 방법
12. results 28
13. Cluster 기초 통계량 (평균 및 표준편차)
14. Decision Tree Model
15. discussion 32
16. cONCLUSION 34
17. limitation 33
18. Individual contribution 36
19. references 51

표 및 그림 목차

[[표 1] 우울증 관련 변수 8](#_Toc43462827)

[[표 2] 우울증 관련 변수 12](#_Toc43462828)

[[표 3] 만성질환 관련 변수 13](#_Toc43462829)

[[표 4] 소득 관련 변수 14](#_Toc43462830)

[[표 5] 신체 관련 변수 14](#_Toc43462831)

[[표 6] 의료 서비스 관련 변수 15](#_Toc43462832)

[[표 7] 업무 관련 변수 15](#_Toc43462833)

[[표 8] 여가 활동 변수 15](#_Toc43462834)

[[표 9] 의료시설 이용 관련 변수 16](#_Toc43462835)

[[표 10]음주 관련 변수 16](#_Toc43462836)

[[표 11] 기타 변수 17](#_Toc43462837)

[[표 12] 군집별 수면시간, 근로시간 평균 및 표준편차 28](#_Toc43462838)

[[그림 1] OECD 주요국 우울증 경험률, OECD 4](#_Toc43088043)

[[그림 2] Elbow Methods 10](#_Toc43088044)

[[그림 3] Clustering 결과 11](#_Toc43088045)

[[그림 4] Clustering 결과 table 12](#_Toc43088046)

[[그림 5] 국민건강영양조사 향후 원시자료 공개계획 26](#_Toc43088047)

[[그림 6] Cluster별 수면시간 boxplot 28](#_Toc43088048)

[[그림 7] Cluster별 근로시간 boxplot 29](#_Toc43088049)

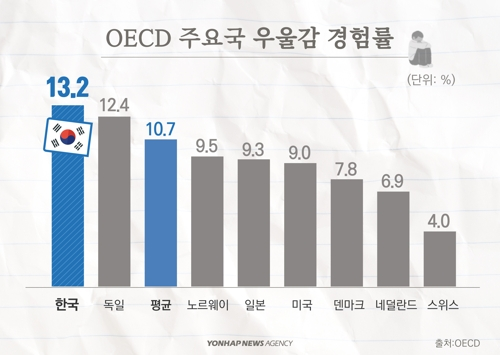
[[그림 8] 저근로군의 Decision Tree 30](#_Toc43088050)

[[그림 9] Decision Tree 결과 table 31](#_Toc43088051)

1. **Introduction**
2. **과제 수행 배경 및 동기**

현대 사회에서 우울증은 심각한 문제로 여겨지고 있다. 특히 우울증으로 인한 사회, 경제적 심각성은 고혈압, 당뇨병, 심장병 등 주요 만성질환과 비등하거나 더 심한 것으로 조사되고 있다. 2019년 기준 우리나라의 ‘우울감 경험률’은 13.2%이고, 우울증이 주요 원인 중 하나인 자살률 또한 10년 이상 OECD 중 1위를 기록하고 있는 만큼, 사회 전반적으로 우울증이 만연해 있다는 것을 알 수 있다[그림1].

이러한 우울증은 당사자가 느끼거나 판단하기 어렵기 때문에, 도구를 사용해 우울증은 판단하는 것이 중요하다. 우울증을 판단하는 대표적 척도로는 PHQ-9이 있는데, PHQ-9이 10점 이상인 경우 경도 우울증인 것으로 판단[[1]](#endnote-1)되고, PHQ-9이 유의미한 우울증 판단 척도인 것은 이미 검증된 바 있다[[2]](#endnote-2).



[그림 1] OECD 주요국 우울증 경험률, OECD

1. **문제 정의 및 도출**

 심각한 문제로 대두되는 우울증을 예측하기 위한 다양한 모델들이 개발되었지만, 우울증 발병의 확률이 높은 특정 집단의 우울증 예측 모델은 거의 존재하지 않는다. 따라서 본 연구에서는 우울증과 관련된 Attribute들을 선정하고 Cluster Analysis를 통해 해당 Attribute에 특정 Value를 가지는 Data Object들이 속한 집단을 발견하고, 해당 집단을 대상으로 하여 경도 우울증(PHQ-9의 점수가 10점 이상, 이후 우울증으로 언급)을 예측하는 Decision Tree 모델을 수립하고자 한다.

앞에서 언급한 바와 같이, PHQ-9 은 우울증 판단 척도로서의 우수성과 타당성은 이미 입증되었으므로 PHQ-9 점수를 Target Attribute로 삼기로 한다. 여기에 정량적이고 객관적인 Attribute들 및 객관적 지표를 이용해 특정 집단의 우울증을 예측하고, 이를 통해 해당 집단의 우울증 유발 인자들을 분석해 관련 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 파악해 우울증을 예측하는 것이다.

또한 기존 연구에서 우울증과 근로시간의 상관관계를 다룬 사례는 존재하지만, 이는 근로시간이 과도하게 많은(주당 근로시간 60시간 이상) 경우에서의 연구이며, 근로시간이 적은 경우에서의 우울증과의 상관관계에 대한 연구는 드물다.

1. **제안하는 방법의 개요**

기 등재된 논문을 참고해 부족한 Domain Knowledge를 보완해 국민건강영양조사 제7기 제3차년도(2018년) 원시 자료에 존재하는 736개의 Attribute들 중 우울증과 관련 있다고 검증된 연속형 Attribute들을 Select해 Cluster Analysis를 실시했다. 그 결과, 우울증은 수면시간, 근로시간과 연관이 큰 것으로 나타났다.

Cluster Analysis는 크게 계층적 방법과 비계층적 방법으로 나눌 수 있다. 본 연구에서는 사전에 군집 수 ‘K’를 정한 후, 각 개체를 K 중 하나의 군집에 배정하는 비계층적 방법을 사용하였다.

또한 데이터에 K개의 Centroid를 위치시키고, 각각 중심에 모인 데이터를 같은 군집으로 할당하는 대표적인 군집 분석 방법인 K-means 군집분석으로 군집 간의 유사성을 측정하였다. 이 때, 유사성의 척도로서는 유클리드 제곱거리(Euclidean square distance)를 사용하였다. 유클리드 거리의 계산 방식은 다음과 같다.



다양한 ‘K’값에 따른 K-means 군집분석을 진행할 때, 분석에 사용할 ‘K’ 값을 선택하기 위해 각 ‘K’값에 해당하는 두 Attribute 간의 SSE를 계산해 최적의 ‘K’값을 찾아주는 ‘엘보우 방법(Elbow Method)’를 사용하였다.

 Cluster Analysis의 결과로 군집은 총 4개로 나눠졌고, 해당 군집들은 다음과 같다.

‘고근로 수면부족군’, ‘고근로 수면과잉군’, ‘저근로 수면과잉군’, ‘저근로 수면부족군’

각 군집을 대상으로 PHQ-9의 점수를 확인해본 결과, 위 군집들 중 ‘저근로 수면부족군’ 과 ‘저근로 수면과잉군’ 이 PHQ-9의 점수가 10점 이상인 비율이 가장 높게 나왔으며 이 군집을 분석 대상으로 삼아 Classification Analysis를 실시했다. Classification Analysis를 실시할 때에도 논문을 참고해 Attribute들을 선별해 가장 Information Gain이 높은 Attribute들을 Feature Subset Selection을 통해 선별했다.

CART (rpart, ctree)는 각 마디에서 하나의 설명변수를 선택하고 이를 이용하 여 두 개로 분리하는 형태의 이진 나무구조를 생성하는 방법으로, 의사결정나무 중 가장 널리 쓰이고 있는 방법 중 하나이다. 결과에 대하여 저장하고 있는 정보에 대한 것으로 tree 함수는 나무구조의 결과에 대한 중요한 값만을 저장하고 있는 반면 rpart 함수는 나무구조의 결과물뿐 아니라 대리변수에 대한 정보 등 그 이외의 유용한 정보들을 모두 결과값으로 저장하고 있다.[[3]](#endnote-3) 또한, rpart에서는 모든 설명변수에 대한 정보를 가지고 있는 대신 범주형 변수만을 추려내어 그에 대 한 범주의 정보를 저장하고 있어 속도나 저장 공간 면에서 훨씬 효율적 인 관리를 할 수 있게 된다. 따라서 본 연구에서는 범주형 데이터가 대부분인 국민건강영양조사 데이터에서 rpart 패키지를 사용하는 것이 다른 패키지를 사용하는 것보다 더 효율적이라고 판단하였다. 이 과정에서 rpart를 이용한 Wrapper Approach를 시도했으며, Feature subset selection을 통해 선별된 Attribute들을 rpart로 Classification Analysis를 실시해 Decision Tree Model을 수립했다.

1. **논문 주요 결과의 개요**

 Cluster Analysis를 통해 우울증과 관련 있는 변수를 찾아본 결과, 우울증과 수면시간, 근로시간 간의 유의미한 상관관계를 발견하였다. 전체 280명 중 ‘고근로 수면 부족군’은 69명, ‘저근로 수면과잉군’은 86명, ‘고근로 수면 과다군’은 54명, ‘저근로 수면 부족군’은 71명으로 나타났다. 해당 Cluster들을 ‘수면 시간’을 기준으로 나누어 봤을때, ‘수면 과다’에 해당하는 인원은 140명, ‘수면 부족’에 해당하는 인원도 140명으로 같은 비율을 차지했고, ‘근로시간’을 기준으로 나누어 봤을때 ‘고근로’에 해당하는 인원은 123명(44%), ‘저근로’에 해당하는 인원은 157명(56%)으로 유의미한 차이(약 12%)를 보였다. 이에 따라 ‘저근로 수면부족군’과 ‘저근로 수면과잉군’을 대상으로 Classification Analysis를 진행하였다.

Classification Analysis의 기법 중 Decision Tree를 이용해 모델링을 한 결과, Informative Attribute로는 활동제한여부, 정규직 여부, 하루에 앉아서 보내는 시간, 소득, 한 번에 마시는 음주량이 선별되었다. 첫 모델을 수립한 후, cp(complexity parameter)를 정해 pruning을 하여 총 4개의 depth로 이루어진 Decision Tree를 수립했다. 이 때의 Accuracy는 0.9502로 계산되었고, 그 중에서 관심 있는 ‘저근로군’의 우울증 유병환자의 예측 정확도를 확인하기 위해 Confusion Matrix를 구했다.

1. **Methods**
2. **연구대상**

본 연구는 2018년도에 시행된 제7기 3차년도 국민건강영양조사에 참여한 만 19세 이상인 ‘저근로 수면부족군’ 71명과 ‘저근로 수면과잉군’ 86명 중에서 (만성질환 의사진단 가족력 여부, 기초 생활 수급 여부, 체중변환여부, 필요 의료서비스 충족, 활동제한 여부, 민간의료보험가입여부) 관련 설문이 미비한 경우(결측치가 존재하는 경우)를 제외하고 총 157명을 대상으로 하였다.

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

1. **Data Understanding**

Strength: 원시자료 구성이 건강설문조사 ⋅ 검진조사 ⋅ 영양조사 별로 구분이 잘 되어 있어 필요한 데이터에 대한 접근성과 활용성이 뛰어나다. 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 존재한다. 예컨대, PHQ-9 점수는 2016년, 2018년에만 조사되어 2017년의 데이터는 사용할 수 없다. 또한, 다른 기에 해당하는 데이터를 가져올 때는 가중치를 처리해야하는 어려움이 있다. 설문조사 자료의 특성상, Categorical Attribute가 대부분이라서 K-means Clustering에 Numeric attribute만 포함해야 한다는 한계가 있었다. 설문조사 자료의 특성상, Categorical Attribute가 대부분이다. 따라서 추후에 Numeric Attribute를 이용한 분석 활용에 제한이 있을 수 있다.

Benefit: 공공 데이터이므로 사용시 별도의 비용이 발생하지 않는다. 2018년도의 데이터 뿐 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 사용된 변수 및 Pre-Processing**

**3.1 Cluster Analysis에 사용된 변수 및 Pre-Processing**

Cluster Analysis를 실시할 때, 비용의 감소를 위해 우울증과 관련된 총 8가지 연속형(주중 하루 평균 수면 시간과 주말 하루 평균 수면 시간은 하나로 합쳤다.)Attribute들을 Select했다. Attribute들은 다음과 같다.

|  |  |
| --- | --- |
| 변수명 | 코드 |
| 입원 이용 횟수 | MH1\_1 |
| 1년간 와병 일 수 | LQ1\_mn |
| 하루 평균 수면 시간 | Total\_slp |
| BMI지수 | HE\_BMI |
| 최근 2주 간 불편감 일 수 | D\_2\_wk |
| 일주일 간 근력 운동 일 수 | BE5\_1 |
| 주당 평균 근로시간 | EC\_wht\_23 |
| 나이 | age |

[표 1] 우울증 관련 변수

Select한 각각의 Attribute들마다 다음과 같은 전처리를 실시했다.

1. 입원 이용 횟수 [MH1\_1]

입원 이용 횟수 Attribute의 Value 중, 88. 비해당을 0, 99. 모름/무응답을 NA로 처리하였다.

1. 1년간 와병 일 수 [LQ1\_mn]

1년간 와병 일 수 Attribute의 Value 중, 888. 비해당을 0, 999. 모름/무응답을 NA로 처리하였다.

1. 주말 하루 평균 수면 시간 [Total\_slp\_wk]

주말 하루 평균 수면 시간 Attribute의 Value 중, 8888. 비해당을 0, 9999. 모름/무응답을 NA로 처리하였다.

1. 주중 하루 평균 수면 [Total\_slp\_wd]

주중 하루 평균 수면 시간 Attribute의 Value 중, 8888. 비해당을 0, 9999. 모름/무응답을 NA로 처리하였다.

1. 하루 평균 수면 시간 [Total\_slp]

‘주말 하루 평균 수면 시간’과 ‘주중 하루 평균 수면 시간’ Attribute를 합쳐 새로운 Attribute, 하루 평균 수면 시간 [Total\_slp]을 생성했다.

[‘주말 하루 평균 수면 시간’ \* 2 + ‘주중 하루 평균 수면 시간’ \* 5] / 7 = 하루 평균 수면 시간

1. 최근 2주 간 불편감 일 수 [D\_2\_wk]

최근 2주 간 불편감 일 수 Attribute 중, 88. 비해당을 0, 99. 모름/무응답을 NA로 처리하였다.

1. 일주일 간 근력 운동 일 수 전처리 [BE5\_1]

일주일 간 근력 운동 일 수 전처리 Attribute 중, 0일~7일이 1~8의 Value로 할당되어 있어 일 수에 맞게 0~7로 Value를 변경했다. 추가적으로, 8. 비해당을 0, 9. 모름/무응답을 NA로 처리하였다.

1. 주당 평균 근로시간 [EC\_wht\_23]

주당 평균 근로시간 Attribute 중, 888. 비해당을 0, 999. 모름/무응답을 NA로 처리하였다.

1. BMI 지수 [HE\_BMI]

BMI 지수 Attribute 중, 사분 범위 \* 1.5의 범위에서 벗어난 Outlier를 없앴다.

다음으로는, 변수마다 발생한 결측치를 대체했다. 결측치 대체 기법은 크게 단일 대체 방법과 다중 대체 방법으로 나뉜다. 단일 대체 방법은 모든 결측치를 하나의 값으로 대체하기 때문에 이후 분석 과정에서 추정정치에 편차가 생길 수 있으며, 표본 오차 및 분산 역시 과소 추정한다는 단점이 있다. 따라서, 하나의 결측치에 대해 여러개의 대쳇값을 생성하는 다중 대체 방법을 사용하기로 했다.

이 때 MICE 패키지를 이용하여 각 변수의 결측치를 대체했다. 다중 대체 방법 중 하나인 MICE 패키지는Multiple Imputation by Chained Equations의 약자로, 완전히 새로운 값 대신 결측치의 예측치와 유사한 값들로 결측치를 대체한다. 또한, MICE는 변수들이 Numeric 변수와 Categorical 변수들이 혼재할   때 잘 작동한다. (ex.월 평균 가구 총 소득 [ainc]).[[4]](#endnote-4) 또한, MICE는 다수의 대체 데이터에 대해 각각 분석을 진행한 후, 이를 하나로 통합하여 마치 하나의 완전한 데이터에 대하여 분석을 진행한 것과 같은 결과를 제공한다.[[5]](#endnote-5)

따라서 본 연구에서는 이 조건을 만족하는 국민건강영양조사 데이터 결측치 처리에 MICE를 사용하기에 적절하다고 판단하였다.

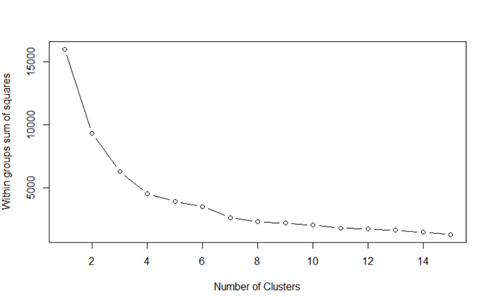
[Normalize the data]

 분석에 이용되는 변수별 단위가 다르므로 이는 단위가 큰 변수에 대해 영향력이 매우 크게 나타날 수 있는 문제를 방지하기 위해서 변수들에 대해 다음과 같은 표준화 작업을 실시하였다. 예를 들어 x라는 벡터 변수에 대해 아래와 같은 코드를 통해 평균은 0, 표준 편차는 1인 표준 정규 분포를 따르도록 변환시켜줄 수 있다.

ex) x = (x-mean(x))/sd(x)   혹은 x = scale(x)

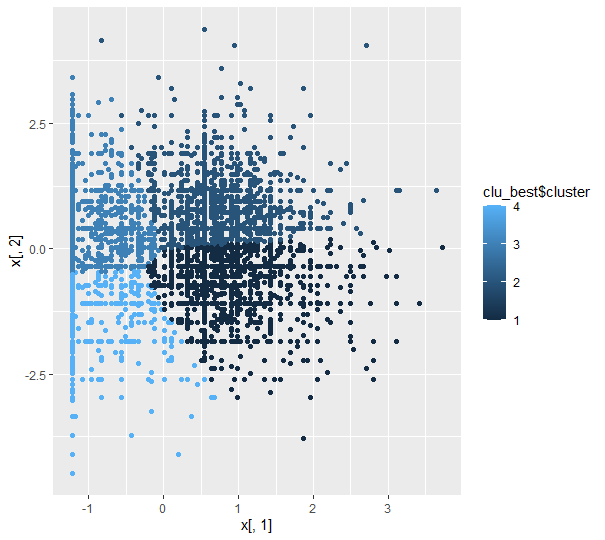
[Evaluating K-means Clusters]

k-means Clustering은 군집 내 오차제곱합(SSE)의 값이 최소가 되도록 군집의 Centroid를 결정해나가는 방법이다. 따라서 초기 군집 개수를 1개로 설정하며 이를 늘려갈 때마다, SSE를 계산한다. SSE의 값이 점점 줄어들다가 어느 순간 줄어드는 비율이 급격하게 작아지는 부분을 최적의 클러스터 개수로 설정한다. 즉, 그래프 모양상 팔꿈치에 해당하는 부분이라 일명 엘보우(elbow) 기법이라 불린다. 엘보우 기법은 최적의 클러스터 개수 ‘K’를 찾기 위해 2~15의 값을 대입해 각각의 SSE를 계산한다. 이에 대한 결과는 [그림 2]와 같다. 이를 통해 최적의 K (=4)를 직관적으로 판단하여 Empty Cluster를 방지하고 Post-Processing에 대한 시간적 비용을 감축하는 효과를 얻을 수 있다.



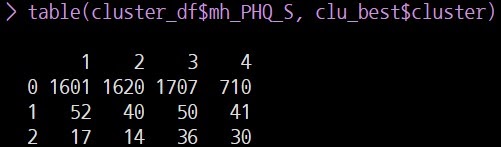
[그림 2] Elbow Methods

**3.2 Classification Analysis에 사용된 변수 및 Pre-Processing**

앞서 언급한 최적 군집의 개수 K(=4)일 때 두 가지 변수 근로시간과 수면시간을 사용하여 군집분석을 한 결과, 다음 [그림 3] 과 같이 4가지 군집이 발견되었다. x축은 정규화된 근로시간, y축은 정규화된 수면시간을 나타낸다. 이를 토대로 Cluster1은 ‘고근로 수면부족군’, Cluster2는‘고근로 수면과잉군’, Cluster3는 ‘저근로 수면과잉군’, Cluster 4는 ‘저근로 수면부족군’으로 각각 정의하였다. 

[그림 3] Clustering 결과

Cluster Analysis의 결과로 도출된 군집들의 PHQ-9 점수를 Table로 살펴본 결과는 [그림3]과 같다. 근로 시간이 적은 군집들 (‘저근로 수면과잉군’, ‘저근로 수면부족군’)에서 PHQ-9 점수가 10점 이상인 Data Object들의 비율이 높다는 것을 알아냈다. 따라서 Classification Analysis를 통해 Decision Tree를 수립할 때, 근로 시간이 적은 3번과 4번 군집에 속한 Data Object들을 합친 Data Set을 사용했다. 이 때 Data의 Dimensionality를 줄여 Curse of Dimensionality를 피하기 위해 총 2가지의 Feature Subset Selection 기법을 사용했다.



[그림 4] Clustering 결과 table

① Irrelevant features elimination

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 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 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 2] 우울증 관련 변수

김성래 외 6명의 연구 (2011) [[6]](#endnote-6)에 따르면 PHQ-9이 우울증 선별에 우수한 모델이기 때문에 우울증 예측 모델의 종속변수로 사용했다. 우울증 의사 진단 여부와 우울증 유병 여부 또한 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 3] 만성질환 관련 변수

박종삼의 연구 (2018)[[7]](#endnote-7)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)[[8]](#endnote-8) 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 4] 소득 관련 변수

정은희의 연구(2015)[[9]](#endnote-9)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 |
| 체중변화여부 | BO1\_1 |

[표 5] 신체 관련 변수

이혁 외 5명의 연구(2016)[[10]](#endnote-10)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 6] 의료 서비스 관련 변수

박혜민의 연구(2016)[[11]](#endnote-11)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 |
| 정규직여부 | EC\_wht\_0 |

[표 7] 업무 관련 변수

김소희 외 2명의 연구(2014)[[12]](#endnote-12)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 8] 여가 활동 변수

이호길의 연구(2016)[[13]](#endnote-13)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 9] 의료시설 이용 관련 변수

서은원의 연구(2016)[[14]](#endnote-14)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 10]음주 관련 변수

김예리의 연구 (2018)[[15]](#endnote-15)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 |
| 하루에 앉아서 보내는 시간 | BE8\_1 |

[표 11] 기타 변수

이태성의 연구(2012)[[16]](#endnote-16)에 따르면 교육수준에 따른 우울증상을 보면 교육수준이 높아질수록 우울 증상은 낮아지는 경향을 보였다. 따라서 교육수준 관련 Feature를 선별했다. 또한, 이도연, 남승민의 연구(2019)[[17]](#endnote-17)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 월 평균 가구별 소득 [ainc]를 가지고도 분석할 수 있다고 판단해 소득 관련 변수 중에서는 기초생활 수급여부 [allownc] Feature를 제거했다.

* 신체 관련 변수

Numeric Attribute인 체질량지수 [HE\_BMI]를 사용하지 않고도 체중변화여부[BO1\_1 ] 로 체중과 우울증의 관계를 설명할 수 있다고 체질량지수 [HE\_BMI] 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.

③ 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. 이들을 Factor함수와 levels 파라미터를 사용하여 Attribute Type을 Interval에서 Nominal로 변경하였다.

* 1. 만성질환 의사진단 가족력 여부 [HE\_fh]

 만성질환 의사진단 가족력 여부 Attribute value는 0. 아니오, 1. 예, 9. 모름인 것을 확인하였고 value가 9인 값들을 NA로 처리하였다. 이들을 Factor함수를 사용하여 Attribute Type을 Factor타입으로 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. 평소 하루 앉아서 보내는 시간 [BE8\_1]

평소 하루 앉아서 보내는 시간 Attribute value는 연속형 type과 88. 비해당(소아), 99. 모름, 무응답인 것을 확인하였고 value가 88, 99인 값들을 NA로 처리하였다.

* 1. 1년간 체중 변화 여부 [BO1\_1]

1년간 체중 변화 여부 Attribute value는 1. 변화없음, 2. 체중감소, 3. 체중증가, 8. 비해당(청소년, 소아), 9. 모름, 무응답인 것을 확인하였고 value가 1인 값들을 0으로 할당하였고, value가 2, 3인 값들을 1로 할당하였으며 value가 8, 9인 값들을 NA로 처리하였다. 이들을 Factor함수를 사용하여 Attribute type을 Factor타입으로 변경하였다.

* 1. 월평균 가구 총 소득 [ainc]

월평균 가구 총 소득 Attribute value는 연속형 type과 17. 월 17만원 이하, 1500. 월 1,500만원 이상인 것을 확인하였다.

* 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.

* 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.

* 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. 정규직 여부 [EC\_wht\_0]

정규직 여부 Attribute value는 1. 정규직, 2. 비정규직, 8. 비해당, 9. 모름, 무응답인 것을 확인하였고 value가 1인 값들을 0으로, value가 2인 값들은 1로 할당하였으며, value가 8, 9인 값들을 NA로 처리하였다. 이들을 Factor함수를 사용하여 Attribute Type을 Factor타입으로 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 확인하였고 value가 1, 2인 값들을 0으로 할당하였고, value가 3, 4, 5인 값들을 1로 할당하였으며, value가 8, 9인 값들을 NA로 처리하였다. 이들을 Factor함수를 사용하여 Attribute Type을 Factor타입으로 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]

  활동제한 여부 Attribute value는 1. 예, 2. 아니오, 9. 모름, 무응답인 것을 확인하였고 value가 2인 값들을 0으로 할당하였고 value가 9인 값들을 NA로 처리하였다. 이들을 Factor함수를 사용하여 Attribute Type을 Factor타입으로 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.

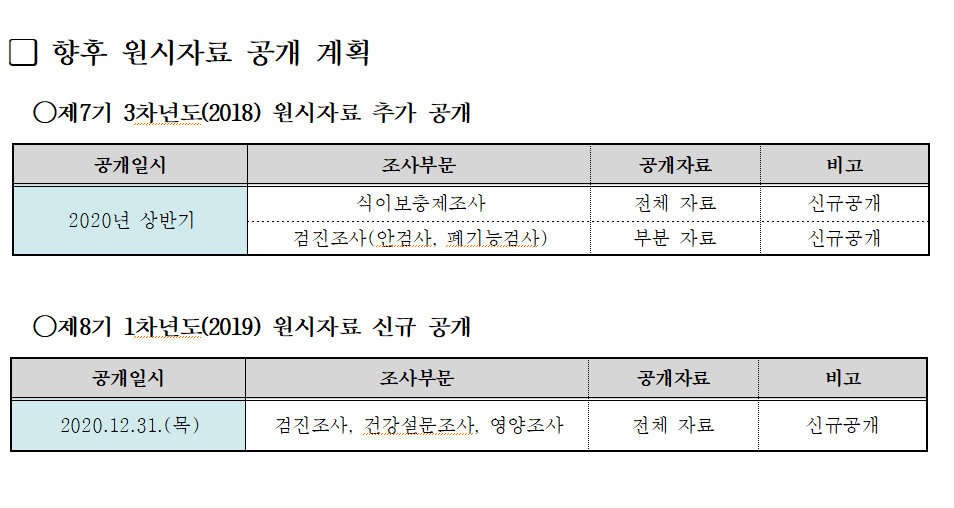
 위의 변수 중 최적의 모델에 사용된 Pre-processing된 attribute들은 다음과 같다.

1. PHQ-9점수[mh\_PHQ\_S] 🡪 phq\_score
2. 활동 제한 여부[LQ4\_00] 🡪 act\_rest
3. 정규직여부[EC\_wht\_0] 🡪 perma\_posit
4. 평소 하루 앉아서 보내는 시간[BE8\_1] 🡪 sit\_time
5. 월평균 가구 소득 [ainc]
6. 한 번에 마시는 음주량[BD2\_1] -> liquor\_amount
7. **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 식이 보충제 복용여부[LS\_1YR] 관련 데이터만 존재하였다. 질병관리본부 홈페이지에서 찾아본 결과 [그림2]와 같이 종류나 구체적인 정보는 2020년 상반기에 공개한다고 명시되어 있었다.



[그림 5] 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 Attribute들을 사용해 현상을 잘 대변하는 모델을 수립해 패턴을 분석한다.

④ 저근로군 우울증 예측 모델

독신 여성 우울증 예측 모델 수립 후, 국민건강영양제사 제7기 제3차년도(2018)에 Cluster Analysis를 적용하여 새로운 군집을 발견하고자 하였다. 군집별 PHQ-9 점수 분포를 확인하여 타 Cluster에 비해 더 많은 비율이 분포한 성인 저근로군을 대상으로 우울증 유병 여부의 패턴을 분석한다.

따라서 본 연구에서는 rpart 패키지를 이용해 Decision Tree model을 수립해, 저근로군 중 경도 우울증을 가진 사람들의 Pattern을 Describe 하였다. 또한 이를 통해 향후 경도 우울증을 겪을 가능성이 높은 근로 시간이 적은 사람을 Predict한다.

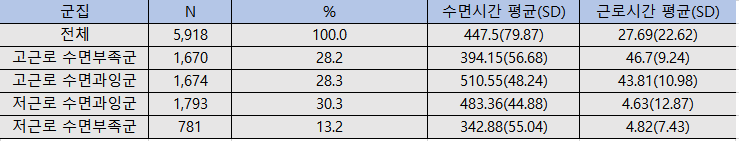
이 과정에서 추출된 data는 대표성을 가져야 한다. Sampling의 방법에는 복원 추출, 비복원 추출, 층화추출법 등이 있다. 복원 추출과 비복원 추출은 R의 Sampling 함수에 replace parameter를 T로 지정시 복원추출, Sampling함수에 replace parameter를 F로 지정 시 비복원 추출을 할 수 있다. 본 연구에서는 Stratified Sampling (층화 추출법)을 “caret” 라이브러리의 createDataPartition함수를 통해 구현했다.  Target Attribute인 phq\_score를 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. 하지만 복원 추출과 비복원 추출은 단순 임의 추출법으로, 오차가 커지게 되어 결과가 어긋날 수 있다. 층화추출법은 이러한 단순임의 추출의 단점을 보완하여 적은 비용으로 추정치를 정확히 구하는 것이 가능하다. 층화추출법은 크기에 비례하여 나누는 동일배분법과 각 층의 크기에 비례하여 표본을 배분하는 비례배분법 등 다양한 방법으로 표본을 각 층별로 배분한다. 우리는 caret패키지의 createpartition함수를 활용하여 우울증 증상이 있는 사람과 없는 사람을 비례배분법을 토대로 훈련 데이터와 테스트 데이터를 3:1으로 나누면서 신뢰성이 높은 추정치를 구할 수 있었을 뿐만 아니라 각 층에 대한 추정치를 부수적으로 구할 수 있었다.

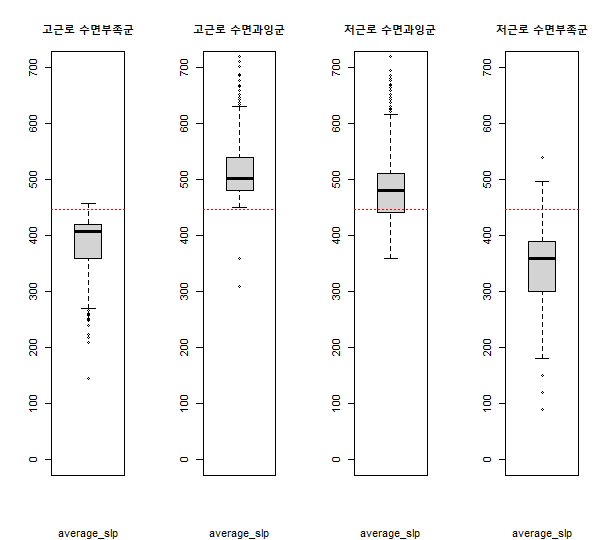
1. **Results**

4가지 군집의 구성 비율과 군집별 근로시간, 수면시간의 평균과 표준편차는 표1에서 확인할 수 있다. 또한, <그림\*>, <그림\*> 를 통해 각 군집별 Boxplot으로 평균과의 거리를 파악할 수 있으며 Boxplot 위의 빨간색 점선은 전체의 평균을 의미한다.

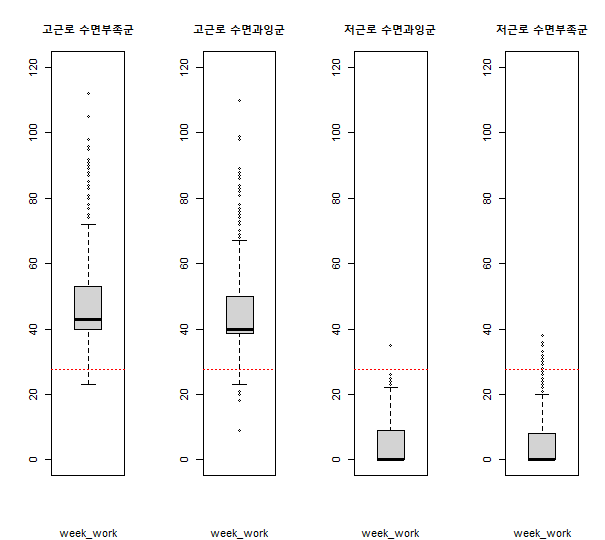
K-means에 의한 최종 군집 분류의 결과, 군집 1은 1,670명(28.22%), 군집 2는 1,674명(19.2%), 군집 3은 1,793명(16.9%), 군집 4는 781명(39.6%)으로 나타났다.

4개의 군집을 각각 ‘저근로 수면부족군’, ‘고근로 수면과잉군’, ‘고근로 수면부족군’, ‘저근로 수면과잉군’으로 할당하였다. 이를 통해 ‘저근로 수면부족군’은 근로시간과 수면시간이 모두 적으며, ‘고근로 수면과잉군’은 근로시간과 수면시간이 모두 많은 특성을 보였다. 또한 ‘고근로 수면부족군’의 경우 근로시간은 평균보다 높은 반면 수면시간은 평균보다 적은 특성을 보였다. 마지막으로 ‘저근로 수면과잉군’은 근로시간은 적은 반면 수면시간은 평균보다 높은 특성을 보였다. 

[표 12] 군집별 수면시간, 근로시간 평균 및 표준편차



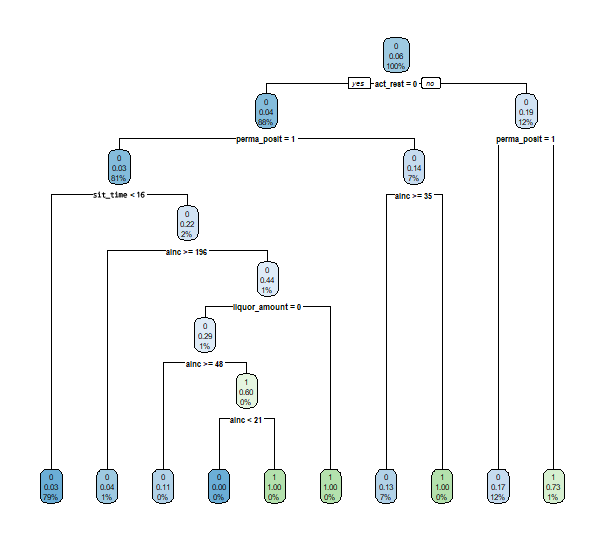
[그림 6] Cluster별 수면시간 boxplot



[그림 7] Cluster별 근로시간 boxplot

이 네 가지 군집 중 저근로 수면부족군, 저근로 수면과잉군을 합친 ‘저근로군’ 대해 Decision Tree를 통한 Classification Analysis를 진행한 결과는 다음과 같다. [그림8]

이 때, 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 저근로 수면 부족군과 저근로 수면 과잉군의 예측 모델을 수립한다.



[그림 8] 저근로군의 Decision Tree

  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 활동제한여부, 정규직여부, 하루에 앉아서 보내는 시간, 소득, 한 번에 마시는 음주량이 선별되었다.

 구체적으로 살펴보면, 우울증 여부를 가장 잘 변별해주는 Feature는 활동제한여부[act\_rest] 였다.

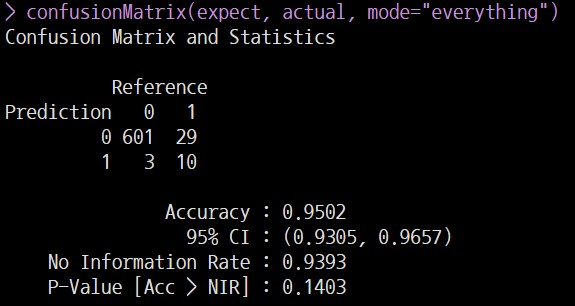
 즉, 수립된 Decision Tree를 봤을 때, 전체 Training set의 100%가 속해 있는 Root node는 ‘활동제한여부’로, 활동제한 경험이 없는 성인은 Training set의 88%, 활동제한 경험이 있는 성인은 12%에 해당된다.

Root node에서 Yes(경험 없음)가 나온 Data object들은 다음 Parent node의 Informative attribute인 ‘정규직 여부[perma\_posit]’로 분류되었다. 정규직 여부가 No인 Data object들은 Training set의 7%에 해당하였고, 이는 다시 ‘월평균 가구 총 소득[ainc]’으로 나뉘었고, value가 35 이상인 Data object들은 ‘우울증이 없다’, 이하인 Data object들은 ‘우울증이 있다’에 할당되었다.

정규직 여부가 Yes인 Data object들은 ‘평소 하루 앉아서 보내는 시간[sit\_time]’으로 분류되었다. 이 중, Yes에 해당하는 Data object들은 Training set의 79%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했다. ‘활동제한 여부’에서 No인 Data object들은 ‘월평균 가구 총 소득[ainc]’으로 분류되었다. 이 중, Yes에 해당하는 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 없다’로 할당했다. No인 Data object들은 ‘한 번에 마시는 음주량 [liquor\_amount]’로 분류되었다.

‘한 번에 마시는 음주량’에서 No, 즉 5잔 이상에 해당하는 Data object들은 Training set의 1%에 해당되고, Target attribute에 ‘우울증이 있다’로 할당했다. 또한 Yes, 즉 4잔 이하에 해당하는 Data object들은 ‘월평균 가구 총 소득[ainc]’의 value가 48과 21을 기준으로 분류되었는데, 이는 실질적으로는 Multi-way Split에 해당한다.

Root node에서 No(경험 있음)인 Data object들은 ‘정규직여부 [perma\_posit]’로 분류되었다. 정규직 여부가 Yes인 Data Object들은 Training set의 12%에 해당되고, 이 Combination을 가졌을 때, Target attribute에 ‘우울증이 없다’로 할당했다. 정규직 여부가 No인 Data object들은 Training set의 1%에 해당되고, 이 Combination을 가졌을 때, Target attribute에 ‘우울증이 있다‘로 할당했다.



[그림 9] Decision Tree 결과 table

Confusion Matrix 함수를 이용하여 예측한 결과를 Accuracy, Precision, Recall, F-measure 등의 척도로 Performance를 평가하였다. 이 때 Positive Class는 경도 우울증인 즉, 1로 설정하였다.

Accuracy : 0.9505, Precision = TP / TP + FP = 10/13 = 0.7692, Recall = TP / TP + FN = 10/39 = 0.2564, F-measure(2 \* {Precision\*Recall/(Precision + Recall)}) = 0.3846 를 통해 Precision과 Recall을 통합하여 Classifier가 얼마나 정확하게 분류하는지 판단할 수 있었다.

Classifier Performance 척도로 흔히 사용되는 F-measure를 개선시키기 위해서는, 논문의 결과를 활용하여 선정한 변수들 뿐만 아니라, 변수 간의 상관 분석 또는 회귀 분석 등을 통하여 우울증과 유의미한 관계를 지닌다고 판단되는 변수들을 선별해 의사결정나무에 사용할 수 있을 것이다.

  본 연구에서는 rpart 패키지를 통하여 각각의 변수들에 대해 Gini Index를 기준으로 사용하여 노드를 재귀적으로 분할하면서 의사결정나무를 만들어 모델을 수립하였다. 또한, 분석에 필요한 Attribute들을 선정할 때 기존 논문들을 바탕으로 선별하였으므로 이 모델을 타당하다고 말할 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 받은 적이 있는 응답자는 60명에 불과했다. 이는 우울증이라는 질환의 인지적 취약성 때문이기도 하며, 우울증 환자들에 대한 부정적인 사회적 감정 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]에서 알 수 있듯이, 2019년 기준 우리나라의 ‘우울감 경험률’은 13.2%이고, 우울증이 주요 원인 중 하나인 자살률 또한 10년 이상 OECD 중 1위를 기록하고 있다. Cluster Analysis를 실시한 결과, 국민건강영양조사에 응답한 피설문자들 중 ‘저근로군’ 군집에 해당하는 피설문자들의 경도 우울증 비율이 높다는 사실을 발견했다(280명 중 157명, 56%). 따라서 본 연구는 ‘저근로군’에 속하는 피설문자들의 우울증 예측 모델을 수립해 PHQ-9의 점수가 10점 이상이 될 것으로 예측되는 사람에게 정부 차원의 효율적 지원을 돕는 것을 목적으로 한다.

  수립된 Decision Tree모델을 통해서 ‘저근로군’의 일상생활 속 우울증 유발 요인들을 분석할 수 있었다. Informative한 Attribute는 총 5개로 ‘활동제한여부’, ‘정규직여부’, ‘하루에 앉아서 보내는 시간’, ‘월평균 가구 총 소득’, ‘한 번에 마시는 음주량’이 이에 해당된다. Decision Tree 모델에서 가장 Information Gain이 높은 Attribute는 ‘활동제한여부’ 이고, 그 다음으로는 ‘정규직여부’가, 다음으로는 ‘하루에 앉아서 보내는 시간’, ‘월평균 가구 총 소득’, ‘한 번에 마시는 음주량’이 선별되었다.  해당 분석을 통해 ‘저근로군’에서 신체활동도와 우울증은 음의 상관관계를 가진다는 것을 발견했다. 실제로 신체활동도와 우울증의 관계는 김영희의 연구(2012)[[18]](#endnote-18)를 통해서 증명된 바 있다. 또한 비정규직에 종사하는 ‘저근로군’에서 우울증 유병률이 높게 나타난다는 분석 결과를 얻었고, 마찬가지로 황주희의 연구(2013)[[19]](#endnote-19)를 통해서도 비정규직에 종사하는 사람들이 더 많은 우울감을 느낀다고 증명된 바 있다.

 국민건강영양조사의 설문 항목 중 ‘활동제한여부’는 심장질환, 류머티스, 치아질환, 등 질환으로 야기된 활동제한에 대한 설문이었다. 이는 ‘저근로군’의 불우한 의료 환경에서 연장된 문제로 여길 수 있다. 서영원의 연구(2019)를 살펴보면 임시, 비정규직 종사자가 상용, 정규직 종사자에 비해 미충족의료를 경험할 확률이 28%많다고 증명된 바 있다.

따라서 본 연구는 근로 시간이 적은 비정규직 종사자를 대상으로 개선된 의료 서비스를 제공하고, 의료보험의 가입을 권장하는 정부 차원 지원의 필요성을 시사한다. 뿐만 아니라 해당 연구를 통해 지역사회에서 ‘저근로 비정규직 종사자’의 우울증 예방을 위한 정책과 프로그램을 개발하는데 실질적 지표를 제시한다는데 의의를 가진다[[20]](#endnote-20).

1. **Conclusion**

현대 사회에서 우울증은 심각한 문제로 대두된다. OECD 국가 중 우울감 경험률이 10년 이상 1위에 머무는 현주소에서, PHQ-9 점수를 우울증 판단 척도로 하는 우울증 예측 모델을 수립하고자 연구를 진행했다. 기존 연구에서 수립된 우울증 예측 모델의 우울증 평가 척도 설문 항목은 주로 피설문자의 주관적 기분 및 생각을 묻는 질문이 많았다. 이를 개선하고자 본 연구에서는 예측 정확도를 높이기 위해 우울증으로 발전할 가능성이 높은 군집을 Cluster Analysis를 통해 구하고, 해당 군집의 우울증을 객관적인 평가 척도로 예측하는 Decision Tree를 수립했다.

연구를 통해 ‘활동제한’과 ‘정규직 여부’가 순서대로 가장 Information Gain이 높은 Attribute들로 선정되었고, 수립된 Decision Tree의 Accuracy는 0.9505로 높게 나왔다. 해당 모델을 미루어 봤을 때, 본 연구는 근로 시간이 적은 비정규직 종사자를 대상으로 개선된 의료 서비스를 제공하고, 의료보험의 가입을 권장하는 정부 차원 지원의 필요성을 시사한다.

본 연구에서 개선이 요구되는 부분은 F-measure의 값이다. 도출된 F-measure의 값은 0.3846으로, 해당 F-measure값을 개선하기 위해서 논문을 통해 선정한 변수들뿐만 아니라 변수 간의 상관 분석 혹은 회귀 분석 등을 통하여 우울증과 유의미한 관계를 지닌다고 판단되는 변수들을 선별해 의사결정나무에 사용할 수 있을 것이다.

1. **Limitation**

 본 연구의 Limitation은 다음과 같다.

[Clustering에 두가지 변수만 사용]

 본 연구에서는 근로 시간과 수면 시간의 두 가지 Feature만을 고려하여 Cluster Analysis가 이루어졌다. 이는 다양한 변수들을 해석하는데 어려움을 겪었기 때문인데, 향후 다양한 우울증과 관련된 요인을 예측하는 도구의 개발을 통해 다차원적인 우울증의 유형을 탐색하는 후속 연구를 기대할 수 있다.

[Clustering에 연속형 변수만 사용]

본 연구에서는 수업에서 다루지 않은 Jaccard를 이용한 Cluster analysis는 활용이 어렵다고 판단하여, 유클리드 거리를 이용한 Cluster Analysis, 즉 K-means를 수행하였다. 하지만, 대다수의 국민건강영양조사 원시자료의 변수는 Categorical Attribute이므로 유클리드 거리를 이용하여 유사도를 측정하여 Clustering을 하기에는 한계가 있었다. 이에 따라 소수의 Numeric Attribute만을 이용하여 Cluster analysis를 수행하였다.

이에 따라 '혼합형 데이터에 대한 군집분석 비교 연구'[[21]](#endnote-21)를 참고하여 범주형, 연속형 데이터가 함께 존재하는 데이터에 대한 군집분석 방법에 대해 고찰한 후, 이를 적용함을 통해 극복할 수 있다.

[가중치 적용의 한계]

부족한 Data set을 보완하려는 목적으로 제7기 자료들 중 2016년과 2017년 국민건강영양조사 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.

따라서 시간적 여유를 두고 원시 데이터에 대한 표본설계 및 가중치 계산을 분석하고 적용함을 통해 극복할 수 있다.

1. **References**

1. 곽병규, 박종철, 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. 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-2)
3. 신혜정, 이윤동, 이은경. (2016). 임상자료를 이용한 나무구조 분류모형의 성능 비교. J Health Info Stat, 41(1), 135-146 [↑](#endnote-ref-3)
4. 고길곤, 탁현우. 설문자료의 결측치 처리방법에 관한 연구: 다중대체법과 재조사법을 중심으로. (2016). 행정논총, 54(4), 291~319 [↑](#endnote-ref-4)
5. 이성영. 결측치 대체 알고리즘별 성능 비교. 석사학위논문, 한국외국어대학교, 서울 [↑](#endnote-ref-5)
6. 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-6)
7. 박종삼. (2018). 만성질환의 발생이 우울증상에 미치는 영향: 고령화연구패널 조사를 이용하여. 석사학위논문, 연세대학교, 서울 [↑](#endnote-ref-7)
8. 구정회, 이정우, 김성윤, 여진영, 유승돈, 전민호. (2004). 뇌졸중 입원 환자의 우울증과 기능적 회복의 상관관계. 보건연구정보센터. 8(1), 14-19 [↑](#endnote-ref-8)
9. 정은희. (2015). 한국사회 우울과 소득의 종단적 상호관계: 사회 원인가설과 선택가설 검증을 중심으로. 박사학위논문, 서울대학교, 서울 [↑](#endnote-ref-9)
10. 이혁, 제민지, 이상옥, 장철훈, 이창형, 김성수. (2016). 성인의 비만과 체중지각이 신체존중감, 자신감, 우울증상에 미치는 영향. 의료커뮤니케이션, 11(1), 17-26 [↑](#endnote-ref-10)
11. 박혜민. (2017). 건강보험자료를 활용한 인구학적, 도농 및 건강보험 종류에 따른 우울증 유병률 차이. 석사학위논문, 성균관대학교, 서울 [↑](#endnote-ref-11)
12. 김소희, 전무경, 강병두. (2017). 신체 활동과 우울증의 연관성: 국민건강영양조사 2014년 자료 이용. Korean J Fam Pract. 7(4), 610-613 [↑](#endnote-ref-12)
13. 이호길. (2016). 여가활동이 중년기 삶의 우울과 심리적 안녕감이 미치는 영향. 동북아관광연구. 12(2), 129-148 [↑](#endnote-ref-13)
14. 서은원. (2016). 우울증 유무에 따른 만성질환자의 의료 이용차이 분석: 고혈압, 당뇨를 중심으로. 석사학위논문, 연세대학교, 서울 [↑](#endnote-ref-14)
15. 김예리. (2018). 음주태도 변화가 우울증상에 미치는 영향: 중 중·노년 인구 패널 데이터 기반 연구. 석사학위논문, 숙명여자대학교, 서울 [↑](#endnote-ref-15)
16. 이태성. (2012). 지역사회 주민들의 사회경제적 수준과 우울증상과의 관련성. 석사학위논문, 전남대학교 대학원, 광주 [↑](#endnote-ref-16)
17. 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-17)
18. 김영희. (2012). 신체활동도와 우울증 및 자살생각의 관련성. 석사학위논문, 연세대학교 대학원, 서울 [↑](#endnote-ref-18)
19. 황주희. (2013). 한국 성인 임금근로자 정규직, 비정규직에서 대사증후군에 영향을 미치는 요인. 석사학위논문, 중앙대학교, 서울 [↑](#endnote-ref-19)
20. 서영원. (2019). 근로형태와 시간이 임금근로자의 소득수준과 미충족의료 경험관계에 미치는 영향. 석사학위논문. 서울대학교, 서울 [↑](#endnote-ref-20)
21. 한지수. (2017). 혼합형 데이터에 대한 군집분석 비교 연구. (2017). 석사학위논문. 고려대학교 대학원, 서울. [↑](#endnote-ref-21)