경영전문석사학위 논문

로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트를 이용한 65세 이상 노인 골다공증 유병률 예측모형 개발

- 2016-2020년 국민건강영양조사 기반 -

2023 년 2 월

서울과학종합대학원대학교

손 다 인

로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트를 이용한 65세 이상 노인 골다공증 유병률 예측모형 개발

- 2016-2020년 국민건강영양조사 기반 -

지도교수 장 중 호

이 논문을 경영학 석사 학위논문으로 제출함

2023 년 2 월

서울과학종합대학원대학교

손 다 인

손다인의 석사 학위논문을 인준함

2023 년 1 월

위 원 장 _____ 오 태 연 _____ (인)

위 원 _____ 최 진 희 _____ (인)

위 원 _____ 장 중 호 _____ (인)

초 록

골다공증은 노년일수록 발병률이 높은 대표적인 질환이다. 따라서, 고령화 사회로 진입하는 한국에서 사회적 비용 비중이 높은 골다공증과 관련된 연구를 진행하는 것은 의의가 있다. 기존 연구를 살펴보면, 사회과학연구에서 사용하는 전통적 통계기법을 이용한 여성의 골다공증 요인 분석 논문이 주를 이루고 있다. 골다공증과 관련하여 빅데이터 분석을 접목한 폐경기 여성의 골다공증 예방행위 모형 개발이나 골다공증 예측 및 개인별 위험 요인 분석 모델의 구축을 주제로 하여 논문이 나오고 있지만 최근 연구를 종합적으로 보면 2-3개에 그칠 정도로 비중은 미미한 상황이다. 또한, 골다공증은 폐경 이후 여성에게서만 발생하는 질병으로 인식되어 상대적으로 남성에 대한 연구가 미진하다. 하지만, 남성 골다공증과 관련된 연구를 살펴보면, 골다공증은 남성에게서도 발생할 수 있으며, 여성에 비해 사망률이 높고, 인지율 및 치료율은 낮아 치명적인 질병이 될 수 있다는 점을 보여주었다. 이로 인해 65세 이상 전체 노년층을 대상으로 한 골다공증 유병률에 대한 빅데이터 분석 연구가 활발히 진행되어야 할 필요성이 제기되어, 본 연구에서는 국민건강영양조사데이터를 기반으로 65세 이상 남녀노인 골다공증 유병률을 예측할 수 있는 모형을 개발하였다.

2016-2020년까지 총 5개년간 국민건강영양 조사 자료에서 공통 변

수 372개를 선정하고, 이 중에 선행연구와 학회지 등 문헌조사를 기반으로 골다공증에 유관한 요인으로 지목하는 변수들을 선별했다. 종속변수는 골다공증 의사진단여부로 하였고, 앞서 선별한 변수들을 종속변수와의 관계에 있어 상관계수가 높은 순으로 46개를 선정하였다. 이 중에서 결측치가 높은 8개의 항목을 제거하고 총 36개를 독립변수로 선정하였다. 여기에 나이, 성별 변수도 추가하여 독립변수는 총 38개 변수로 최종 구성하였다. 데이터 전처리 진행시에, 소수형은 정수형으로 변경하고, 총 8,170명의 데이터 중 결측치를 제거한 5,365명을 대상으로 데이터 분석을 진행하였다.

해당 변수들에 대하여 머신러닝 분류모델인 로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트 분석 알고리즘을 적용하고, 각 알고리즘의 예측 성능을 확인하여 비교하기 위해서, 65세 이상 인구 5,365명의 데이터를 학습용 데이터와 테스트용 데이터로 구분하였다. 학습용 데이터는 전체 데이터의 80%로, 테스트용 데이터는 전체 데이터의 20%로 선택하였다. 종속변수인 y에는 골다공증 유병 유무를 target 변수로 정의하였고, 독립변수인 x에는 골다공증 유병 요인 변수로 최종 선정된 38개 변수에 대해서 전처리와 결측치 제거를 완료하고 더미변수화 시킨 변수들로 정의하였다. 평가지표는 이진분류에서 널리 사용되는 정확도, 정밀도, 재현율, F1 스코어, ROC 곡선을 구하여 도출된 AUC 값으로 하였다. 분류기준을 0.4, 0.45, 0.5, 0.55,

0.6으로 변경 했을 때의 혼동행렬을 기반으로 산출된 정확도, 정밀도, 재현율, F1 스코어에서 대부분 랜덤포레스트가 좋은 결과를 보였다. 특히, AUC 값은 랜덤포레스트(0.8068), XG 부스트(0.8059), 로지스틱 회귀분석(0.7800), 의사결정나무(0.6277) 순으로 낮아진다. 따라서, 65세 이상 남녀노인 대상으로 골다공증 유병률을 예측시에 해당 모델을 적용하는 것으로 최종 채택하였다. 또한, 랜덤포레스트로 골다공증 유병률을 예측할 때, 어떠한 변인들이 골다공증에 영향을 미치는지 파악해보았다. 보유질병 요소 관련해서는 골관절염과 이상지질혈증 보유가 골다공증 예측에 영향력이 있는 것으로 나타났다. 또한, 인구학적 요소로는 남성, 만나이, 소득 4분위수(개인), 가구원수 2-3명, 결혼여부, 결혼상태가 사별이거나, 민간의료보험 미가입인 경우, 기본건강 요소로는 신장, 체중, 허리둘레, 체질량지수, 총콜레스테롤 수치가 골다공증 예측에 영향력이 있는 것으로 나타났다. 마지막으로, 건강행태적 요소로 평생 5갑이상 흡연 또는 비흡연일 경우와 비타민D 섭취량이 영향력이 높게 나타났다.

본 연구를 통해 65세 이상 노인인구의 골다공증 유병률 예측 모델을 개발하면서 이 결과가 의료계와 보건연구에 도움이 될 것으로 보인다. 또한, 골다공증과 관련된 막대한 사회적비용을 절감하는 데에 기여하며, 고령화 사회에서 노인의 건강 관리 및 삶의 질 개선하는 데에 큰 역할을 할 것으로 기대된다. 침묵의 질환으로 불리는 골다공증

질환의 미인지 환자에게 골밀도 검사를 추천하고, 위험요인을 알리는 등 사전 예방 활동이 가능해 환자의 인지율 상승과 사전 관리, 그리고 골다공증 진행이 악화되는 것을 막는 효과가 클 것으로 기대된다. 이번 연구를 통해, 기존연구와는 차별화된 65세 이상 노인인구를 대상으로 머신러닝을 적용한 골다공증 유병률 예측모형 개발 연구의 초문을 열었다. 추후에 골다공증 분야에서도 머신러닝을 적용한 예측모형 개발이 활발히 진행 될 수 있는 단초를 제공하였으며, 이는 학계에 긍정적인 영향을 미칠 것이라고 본다.

목 차

표 목 차

그 림 목 차

제 I 장 서 론

제 1 절 연구의 배경 및 목적

골다공증은 노년일수록 발병률이 높은 대표적인 질환이다(정윤석, 2010). 이는 골다공증이 고령화로 진입한 현대 사회에서 주요하게 관리되어야 되는 질환이라는 의미로 해석된다.

하버드 의과대학에서 일반인에게 배포하는 자료를 살펴보면 미국에서 골다공증 환자의 약 800만명이 여성이고, 200만명이 남성이라고 한다. 골다공증에는 원발성 골다공증과 골다공증이 있는데, 골다공증 여성 환자의 95%, 골다공증 남성 환자의 80%가 원발성 골다공증이다. 골다공증은 주로 폐경 후 여성호르몬 수치가 낮아지는 여성에게 발생하고, 50대 이상의 남성이 가진 여성호르몬 수치가 노화됨에 따라 감소하면서 발생한다(Bolster, 2021).

이러한 골다공증에 대해서 세계보건기구(World Health Organization, WHO)는 골다공증을 "골량 감소와 미세 구조의 이상을 특징으로 하는 전신적인 골격계 질환으로, 결과적으로 뼈가 약해져서 부러지기 쉬운 상태가 되는 질환"으로 규정했다(WHO, 1994). 최근 미국 국립보건원(National Institutes of Health, NIH)에서는 "골강도의 약화로 골절의 위험성이 증가하게 되는 골격계질환"으로 명시하고 있다. 대한내분비 학회지에서는 골강도는 골량(quantity)과 골질(quality)에 의해 결정되며, 골량은 주로 골밀도(BMD)에 의해 표현되고 골질은 구조, 골교체율, 무기질화, 미세 손상 축적 등으로 구성

된다고 보고 있다(NIH Consensus Development Panel, 2001). 골다공증을 진단할 때는 골밀도를 측정하여 그 유무를 진단하고 있는 것이다(정호연, 2008).

한국 인구는 2017년부터 65세 이상 노인 인구가 증가하고 있다. <표 I>을 살펴보면, 65세 이상 인구 구성비율이 1960년 2.9%, 2020년 15.7%, 2067년에는 46.5%에 이르며, 80세 이상 인구 구성비율이 1960년 0.2%, 2020년 3.6%, 2067년 20.7%에 도달할 것임을 알 수 있다. 2067년에 중위연령은 무려 62.2세에 이를 것으로 예측되고 있다(김두섭, 2020).

<표 1> 고령화 추세와 전망

연도	인구 (100만명)	인구 구성률(%)		중위연령	노년 부양 인구비[1]	고령화 지수[2]
		65세 이상 인구	80세 이상 인구			
1960	25.0	2.9	0.2	19.0	5.3	6.9
1980	38.1	3.8	0.5	21.8	6.1	11.2
2000	47.0	7.2	1.0	31.8	10.1	34.3
2020	51.8	15.7	3.6	43.7	21.7	129.0
2040	50.9	33.9	10.2	54.4	60.1	345.7
2060	42.8	43.9	19.2	61.3	91.4	546.1
2067	39.3	46.5	20.7	62.2	102.4	574.5

출처: 통계청, 「장래인구추계: 2017-2067」, 2019.

[1] 노년부양인구비=(65세 이상 인구÷15-64세 인구)×100

[2] 고령화지수=(65세 이상 인구÷15세 미만 인구)×100

이렇게 고령화가 빠르게 진행되고 있는 우리나라에서 노인관련 주요 질병에 대한 연구를 진행하는 것은 사회적 비용 감소로도 연결될 수 있다. 이와 관련하여, 우리나라 골다공증으로 인한 사회경제적 질병 비용 측정 연구를 살펴보면 골다공증 질환으로 인한 사회경제적 비용은 크게 진단 종류와 성별기준 두가지로 산정되어 있다. 주진단 기준으로 약 3,201억, 부진단을 포함할 경우 1조 510억원의 규모이다.

성별로 나누어서 보면, 여성 2,972억-9,232억원, 남성 230억-1,279억원 수준이다. 골다공증의 사회경제적 질병비용은 우리나라 주요 질환 중 하나인 유방암과 약 70억원 차이로 비슷한 수준이고, 당뇨병 약 1조 2,473억원보다는 약 1,900억정도 낮다. 자궁경부암 약 3,602억원, 알레르기 비염 약 3,056억원보다는 3배 정도 높은 비용으로 우리나라 사회경제적 질병비용 비중에서 골다공증은 상당히 큰 부분을 차지한다. 골다공증 사회경제적 질병비용은 같은 해 우리나라 총 GDP(Gross Domestic Product) 1,730조 4천억원의 0.018-0.061% 수준에 해당된다(이은환 외, 2019). 추후, 고령화 사회가 가속화될수록 골다공증으로 인한 사회 경제적 질병 비용은 더욱더 증가할 것으로 판단된다. 따라서 사회경제적 질병 비용을 줄이기 위해서 골다공증 질병에 대한 연구가 활발히 진행될 필요가 있다.

현재, 골다공증과 관련하여 진행되는 선행연구를 살펴보면, 여성을 대상으로 한 연구에 초점이 맞추어져 있다. 특히, 폐경기 관련 요인이 여성 골다공증에 어떠한 요인을 미치는지에 대한 연구가 많다(신민호 외, 2022). 하지만, 골절 관련 사망률이 남성에게서 더 높게 나타나고 있고, 남성의 골밀도가 낮아

지는 추세를 보이고 있기 때문에 남성에게 영향을 미치는 인자에 대한 연구의 필요성도 제기된다(이혜상, 2016). 또한, 국내 남성 노인의 모든 골격 부위의 골밀도는 미국보다 현저히 낮았고, 일본보다도 낮은 것으로 연구 결과가 보고되었다(Park et al, 2014). 골다공증 유병률은 여성 37.3%, 남성 7.5%로 나타났지만, 골감소증의 비중은 여성과 유사하였다. 특히, 70대 이후 남성의 25%가 골다공증 유병자임에도 불구하고, 해당 질환을 여성 고유의 병으로 인식하다 보니, 인지율과 치료율이 더욱 낮았다(김윤아, 2014). 골다공증으로 인한 고관절 골절 후 1년 이내 사망률은 여성에 비해 남성이 높다는 점, 노인 환자의 고관절 골절 후 사망이 일반적이라는 점, 합병증으로 인한 사망률이 높다는 점, 그리고 이들의 삶의 질을 저하시킬 수 있다는 점에서 남성 노인의 골감소증 및 골다공증과 관련된 부분도 주요하게 다루어져야 한다(Jiang et al, 2005). 한국 성인 남성 중 특히 60대 이상에서 골다공증 유병률이 증가하였는데, 이를 자세히 살펴보면, 연령이 높고, 체중이 낮을수록, 근력이 감소할수록 유병률이 증가하였다. 또한, 신체활동 및 칼슘섭취 부족 여부, 만성 신장 질환을 앓은 여부 및 낮은 사회 경제적 수준 등의 위험 요인과 유병률이 관련 있음을 알 수 있었다(유정은, 2018). 즉, 골다공증이 폐경 이후 여성에게만 발생하는 질병이 아니라, 남성에게서도 발생하는 질병이며 때로는 더 치명적일 수 있다는 점을 보여준다. 또한, 65세 이상 노인 남녀 골다공증 관련 연구를 동시에 진행한 가장 최근 연구는 2017년도로 현시대를 반영한 새로운 분석이 필요할 것으로 보이며, 노인 남성 골다공증 연구에 관련된 비중도 적은 것으로 나타났다. 최근 골다공증과 관련하여 빅데이터 분석을 접목한 폐경기 여성의 골다공증 예방행위 모형 개발이나 골다공증 예측 및 개인별 위험 요인 분석 모델의 구축을 주

제로 하여 논문이 나오고 있지만 최근 연구를 종합적으로 보면 2-3개에 그칠 정도로 비중은 미미한 상황이다. 따라서, 65세 이상 전체 노년층을 대상으로 하여 골다공증 유병률에 대한 빅데이터 분석 연구가 활발히 진행되어야 할 필요성이 제기된다(김지영, 양영란, 2020). 이에 본 연구의 목적은 국민건강영양조사 자료를 기반으로 65세 이상 남녀노인 골다공증 유병률 예측할 수 있는 모형을 개발하는 것이다.

제 2절 연구 문제

본 연구에서는 65세 이상 남녀노인인구 전체를 대상으로 하여, 골다공증 유병률을 예측할 수 있는 가장 적합한 모형이 무엇인지 확인하고자 한다. 이를 위해 국민건강영양조사의 최근 5개년(2016-2020년) 자료를 기반으로 하여, 65세 이상 노인의 골다공증 발병에 주요하게 영향을 미치는 변수를 선정하고, 이 변수들에 대한 머신러닝 분류모델인 로지스틱 회귀분석(Logistic regression analysis), XG 부스트(XG boost), 의사결정나무(Decision Tree), 랜덤포레스트(Random Forest) 알고리즘을 적용하여 골다공증 유병률과 관련한 예측률이 가장 높은 모형을 채택하고자 한다. 즉, 이 연구를 통해 65세 이상 노인 골다공증 유병률 예측모형을 개발하고자 한다.

제 3절 논문 구성

본 논문의 다음 제2장에서 이론적 배경을 통해 남녀 골다공증 유병률 관련 연구를 확인하고, 국민건강영양조사 기반 남녀 골다공증 유병률 연구 및 딥러닝 적용 질병 유병률 예측모형 개발과 관련하여 골다공증과 기타질병으로 나누어서 기존 연구에 대해 살펴보고자 한다. 이를 바탕으로 본 논문의 필요성과 차별점에 대해서도 제시하고자 한다.

제3장에서는 연구방법을 통해 제1절 연구대상에서 국민건강영양조사 자료에 대한 전반적인 설명과 연구대상에 대한 특징을 기술하고자 한다. 제2절 측정도구에서는 데이터 전처리 관련 내용에 대해 제시하고, 종속변수와 독립변수에 대한 전반적인 설명과, 더미변수로 처리한 과정에 대해 기술하고자 한다. 제 3절 분석방법 및 절차에 대해서는 예측모형 개발 도구로 활용할 분류분석 도구(로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트)에 대해 설명을 진행하고 분석방법을 제시하며, 예측력이 좀 더 높은 최종 모형을 선택하기 위해 채택한 모델 평가 지표 (혼동행렬(Confusion Matrix), 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1 스코어(F1 Score), ROC 곡선(Receiver Operating Characteristic Curve), AUC(Area Under the Curve)값)에 대한 이론적 내용을 명시하고자 한다.

제4장 연구결과에서는 제1절 기술통계분석에서 각 변수별로 빈도와 비중을 확인하여 기술하고자 하며, 제2절 골다공증 유병률 예측 결과 비교 분석에서는 로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트 모델을 적용한 뒤 최종 모형의 예측력을 비교하기 위해서 모델평가 지표를 활용하고자 한다.

또한 최종 채택된 모형의 예측력에 영향을 미치는 변인들을 확인하고자 한다.

제5장 논의 및 결론에서는 제1절 요약 및 논의를 통해 결과에 대한 요약 및 관련 내용을 논하고, 제2절 결론 및 제언을 통해 해당 결과에 대한 결론 및 향후 연구방향과 시사점에 대해 이야기하고, 제3절 연구의 제한점을 통해 연구를 진행하면서 발생한 문제와 향후 추가 연구 방향을 설정하고자 한다.

제 Ⅱ 장 이론적 배경

제 1 절 선행 연구 정리

1) 검색전략

본 연구에서는 구글스칼라와 RISS, KISS, DB pia, 페이퍼 서치를 검색원으로 사용하였다.

골다공증 관련 기존 연구에 대해서는 ‘골다공증’, ‘골다공증 유병률’, ‘골밀도’, ‘노인 여성’, ‘노인 남성’, ‘60세’를 이용하여 문헌을 검색하였다.

국민건강영양조사 기반 질병 관련 기존 연구와 관련된 문헌은 ‘국민건강영양조사 질병’, ‘국민건강영양조사 유병’, ‘국민건강영양조사’ 그리고 질병 유병률 예측 모형 기존 연구와 관련된 문헌은 ‘기계학습기반’, ‘예측모형 비교’, ‘로지스틱회귀분석’, ‘XG 부스트’, ‘의사결정나무’, ‘랜덤포레스트’을 이용하여 검색하였다. 전체기간을 대상으로 하여, 여성 골다공증 유병률 관련 연구, 남성 골다공증 유병률 관련 연구, 국민건강영양조사 기반 전체 골다공증 유병률 연구, 골다공증 유병률 딥러닝 적용 예측모형 개발연구, 기타질병 유병률 딥러닝 적용 예측모형 개발연구 중심으로 주요 연구를 집중적으로 살펴보았다.

2) 여성 골다공증 유병률 관련 연구

C도에 소재하는 일개 병원에서 외래치료를 받고 있는 60세 이상 노인여성 골다공증 환자 155명을 대상으로 운동기대감과 운동 자기효능감, 골다공증 지식의 정도와 관계를 파악하였다. 자료는 SPSS/WIN 21.0 프로그램을 이용하여 분석하였고, 연구결과 60세 이상 골다공증 여성노인의 골다공증 지식 정도는 중간 수준이었고, 골다공증 지식과 운동기대감, 운동기대감과 운동자기효능감 유의한 양의 상관관계가 있으나, 골다공증 지식과 운동자기 효능감의 관계는 유의하지 않았다. 소수의 환자를 대상으로 진행한 연구로 보편성을 적용하는데에 어려움이 있어, 향후 연구에서는 대상자 수를 확대하여 연구할 필요가 있는 한계점이 있다. 선행연구결과와는 다르게 골다공증 지식과 운동자기 효능감은 관계가 없는 것으로 나타나 반복적 연구가 필요하다고 보았다. 또한, 골다공증 여성노인의 운동능력을 높이기 위해, 운동을 하는데 있어 영향을 미치는 요인을 분석하는 추가 연구의 필요성이 있다고 보았다(박애화, 박형란, 2019).

2008-2009년도까지 총 2개년간 국민건강영양조사 자료에서 만 20세이상의 골다공증 유병환자 여성 1,156명을 대상으로 골다공증 인지여부에 따른 건강증진행위에 차이가 있는지를 연구하였다. Chi-square test, t-test, 다중로지스틱 회귀분석을 시행한 결과, 인지군이 비인지군에 비해 흡연자가 적었으며, 주관적 건강상태가 더 나쁘다고 생각하고, 고지혈증, 류마티스성 관절염 유병률이 더 높았으며, 지병으로 인한 활동제한도 더 많았다. 골다공증 인지와 골다공증 건강증진행위와의 관계는 근력운동시행과 비흡연에 유의한 차이

를 보였고, 비타민D, 칼슘 섭취, 신체활동, 음주유무에서는 유의한 차이를 보이지 않는 것으로 나타났다. 이는, 골다공증에 걸렸다는 사실을 인지한다는 사실이 건강증진행위를 증진시킨다고 볼 수 없는 것으로 해석된다. 예외적으로, 65세 이상 환자에서는 골다공증 인지군이 술과 흡연을 적게 하는 경향을 보였다. 이 연구에서 저자는 단면 연구이기 때문에 인과관계를 설명할 수 없다는 단점이 있다고 하였다(서정민, 2011).

2008-2009년까지 총 2개년간 국민건강영양조사 자료를 기반으로 자연폐경된 여성 2,827명 중 골다공증 검사를 완료한 2,043명을 대상으로 골다공증과의 관련성을 확인하였다. 연령을 기준으로 45세 미만 그룹, 45세 이상 그룹으로 나누고, 각 그룹의 빈도, 백분율 및 평균을 구하였다. 변수 중에 연속형은 t-test, 범주형은 x^2-test를 통하여 유의성을 검정을 실시하였다. 골다공증 여성 중 172명이 조기 폐경 여성이고, 635명이 비조기 폐경 여성이다. 폐경 이후 기간, 폐경 나이, 수면시간, 교육수준, 흡연, 음주, 수면시간, 칼슘과 커피 섭취, 공복혈당, 임신횟수, 여성호르몬 복용, B형 간염 유병이 조기 폐경 여성의 골다공증 관련요인으로 확인되었다. 또한, 교육수준, 폐경 이후 기간, 흡연, 커피 섭취 빈도, 수면, 칼슘섭취량, B형 간염 유병률, 임신횟수, 여성호르몬 복용이 비조기 폐경 여성의 골다공증 관련요인으로 확인되었다. 본 연구에서는 조사 시점보다 폐경시점이 앞서 발생하므로, 차이가 큰 경우 변수들의 상태가 폐경시점과 조사시점에서 상이할 수 있다는 것을 한계점으로 보았다(임지선, 2012).

2008-2009년까지 총 2개년간 국민건강영양조사 응답자 여성 11,064명

중 골다공증 검사를 한 폐경 여성 2,300명을 65세 이상 노인여성 1,158명과 65세 미만 중년여성 1,142명으로 그룹을 나누어서 골다공증 관련 요인을 연구하였다. 65세 미만 중년여성군은 나이, 체질량, 교육수준, 조기 폐경, 단백질 섭취량, 여성 호르몬제 복용여부가 골다공증과 유관한 요인으로 확인되었다. 65세 이상 노인여성군에서는 연령, 체질량, 허리둘레, 흡연여부가 골다공증과 유관한 요인으로 확인되었다. 이를 통해, 연령에 따른 골다공증 요인이 다르므로, 그에 맞는 건강관리와 검진이 중요함을 알 수 있다. 해당 연구의 한계점은 폐경 여성만을 대상으로 하였기 때문에 생리주기에 따른 조사는 이루어지지 못한 점이었다. 유전적 요소, 영양 섭취 기준, 기타 보유 질병, 여성의 신체적 특성이 골밀도에 미치는 영향력에 대해 확인하지 못한 부분이 이 연구의 한계점으로 언급되었다(유인영, 2011).

2015-2018년까지 총 4개년간 국민건강영양조사 대상자인 30세 이상 성인 여성 1만 1,427명을 대상으로 골다공증 위험도 평가점수 모형을 개발하였다. Rao-Scott 카이-제곱 검정, 가중 다중 로지스틱 회귀 분석을 이용해 분석하였으며, 골다공증 위험 점수 카드 모델은 오즈(PDO) 방법을 두 배로 늘리기 위해 포인트를 사용하여 가중 다중 로지스틱 회귀 분석을 통해 개발되었다. 골다공증 관련 요인은 나이, 소득 및 교육수준, 체질량, 지병(고혈압, 뇌졸중, 류마티스 관절염, 고콜레스테롤혈증, 근감소증, 갱년기) 등으로 나타났다. 스코어카드 결과에서는 폐경, 16-17세 연령, 교육수준(초등학교 이하), 류마티스 관절염 BMI(<18.5kg/m2)등에서 가장 높은 점수 범위가 관찰되어, 이들이 가장 중요한 위험요인임을 시사했다. 해당 연구의 의의는 우리나라 성인

여성의 골다공증과 다양한 요인의 인과관계를 종합적으로 파악한다는 데 있다. 또한 이러한 관계를 점수화하여 건강관련 기관에서 여성의 근골격계 건강관리에 유용한 측정도구로 활용할 수 있다. 골다공증의 유병 유무를 실제 진단기록이 아닌 국민건강영양조사를 이용한 부분을 연구의 한계점으로 언급했다(박일수, 2021).

3) 남성 골다공증 유병률 관련 연구

2008년-2011년까지 4개년간 국민건강영양정보조사 자료를 이용하여 골밀도 검사를 완료한 30세 이상 성인 남성을 대상으로 남성 골다공증의 유병률을 파악하고, 생활습관, 보유질병, 인구학적 특성 요인과 남성 골다공증 유병률의 관련성을 분석하였다. 로지스틱 회귀분석을 통해 골다공증과 관련된 요인에 대한 교차비를 살펴보았다. 그 결과, 60대 이상 성인남성에서 골다공증 유병률이 증가하였는데 연령이 높고, 체중이 낮을수록, 근력이 감소하고, 신체 활동 및 칼슘 섭취 부족, 만성 신장 질환을 앓은 여부 및 사회 경제적 수준이 낮을 경우 등의 요인이 유병률 증가와 관련이 있었다. 연구의 한계점은 단면 연구로서 골다공증과 여러 위험 요인들 사이의 인과 관계를 입증하기 어렵다는 것이고, 동반 질병력에 대한 자세한 분류와 스테로이드 제제 사용 등 이차성 골다공증 원인에 대한 조사 자료가 부족하다는 점이다. 또한 연구 자료에서 50세를 기준으로 골다공증의 정의를 다르게 하고 있어, 결과 해석에 제한이 있을 수 있다. 모든 연령층에 적용할 수 있는 공통된 골밀도 지표가 없어, 향후 이에 대한 연구 필요성을 이야기했다(유정은, 2018).

해당 연구에서는 대퇴경부 및 요추골밀도의 골밀도 수준을 60세 이상 남성노인 2,763명을 대상으로 파악하고, 골밀도의 영향요인을 병력, 신체 측정 정보, 생활습관 등으로 두고 연구를 진행하였다. 다중선형회귀분석을 실시하였고, 그 결과 남성 노인의 골다공증 유병률이 연령이 증가함에 따라 증가 추세임을 알 수 있었다(60대(6.7%)>70대(15.8%)>80대이상(31.4%)). 또한, 제지방량이 대퇴경부 및 요추 골밀도의 골밀도 수준에 영향을 가장 많이 미치

는 것으로 나타났다. 따라서 골다공증을 예방하기 위해서 남성 노인에게 제지방량을 증가할 수 있도록 권고하는게 좋을 것이라고 결론을 지었다. 골밀도 관련 인자는 파악했으나, 단면연구이기 때문에 인과관계를 설명하기 어렵다는 한계점이 있다고 보았다(김영란 외, 2013).

4) 국민건강영양조사 기반 전체 골다공증 유병률 연구

2008년-2011년까지 국민건강영양조사 자료에서 측정한 골밀도 자료를 기준으로 50세 이상 성인에 대한 골밀도 수준, 골다공증과 골감소증 유병률을 살펴보았다. 골다공증 유병률은 여성 37.3%, 남성 7.5%로 나타났지만, 골감소증의 비중은 여성과 유사하였다. 여성 골다공증 유병률은 50대 15.4%, 60대 36.6%, 70대 이상 68.5%, 남성 골다공증 유병률은 50대 3.5%, 60대 7.5%, 70대 이상 18.0%이었다. 연령이 증가할수록 남녀 모두에서 골다공증 유병률은 크게 증가하나 인지율과 치료율은 차이를 보이지 않았다. 특히, 70대 이상 남성의 25%가 골다공증 유병자임에도 불구하고 해당 질환을 여성 고유의 병으로 인식하다 보니, 인지율과 치료율이 더욱 낮았다. 이러한 골다골증은 노화로 인해 자연스럽게 생길 수 밖에 없는 질병이 아니라, 예방과 관리가 필요한 질병이라고 볼 수 있다. 규칙적 운동과 비타민 D, 칼슘, 인 섭취, 금연과 절주 습관, 생활 속에서 낙상으로 인한 골절을 방지하여 골다공증으로 인한 추가 합병증 발생을 줄일 수 있다고 보았다. 침묵의 병이라 불리는 골다공증은 당뇨병, 고혈압처럼 병이 많이 진행되기 전까지는 아무런 증상도 없고, 골절로 인해 뒤늦게 발견되는 경우가 많기 때문에, 조기진단과 치료가 매우 중요하다고 보았다(김윤아, 2014).

2015-2017년까지 국민건강영양조사 자료에서 50세 이상의 남녀를 대상으로 영양소 섭취와 식이 다양성 점수가 골밀도에 미치는 효과(통제 변인:흡연, 음주, 사회경제적요인, 신체활동, BMI)를 파악하였다. 골다공증 위험도는 50-64세가 75세 이상에 비해 2.38배 정도 낮으며, 모든 연령군에서 남성의

골다공증위험도가 여성보다 8.85배 이상 낮았다. 50–64세에서는 성별, 연령, 소득수준, 교육수준, 65–74세에서는 성별, 교육수준, BMI, 음주, 75세 이상의 경우, 성별, 연령, 음주, 흡연, BMI, 에너지 섭취량이 골다공증 발병에 영향을 미치는 통제변인이다. 전 연령대에서 탄수화물, 단백질, 지방, 식이섬유, 칼슘, 인, 나트륨, 칼륨, 비타민 B1, 비타민 B2, 니아신, 콜레스테롤 섭취량이 증가할수록 골다공증 발생의 위험이 낮아졌으나, 연령별 통제 변인을 모두 통제하면, 각종 영양소 섭취량은 영향을 미치지 않는 것으로 나타났다. 해당 연구에서 선정된 통제 변인은 보정변인으로 활용될 것이라 기대된다고 언급하였다(권세혁, 이정숙, 2020).

2008−2011년 국민건강영양조사 자료에서 50세 이상 남성 3,071명, 여성 3,635명의 골다공증 유병률 및 요인 분석, 치료율과 인지율 분석을 진행했다. SAS 프로그램을 통해 Rao−Scott x^2 로 빈도분석을 진행하고, 로지스틱 다중회귀분석으로 유병률, 인지율, 치료율에 영향을 미치는 요인을 확인했다. 남성에게서 골다공증 인지율과 치료율이 특히 낮은 수준으로 파악되었으며, 2차 예방의 중요성을 알 수 있었다. 또한, 연령 증가, 낮은 사회적 경제수준일때, 골다공증 유병률이 증가함을 확인했다. 유병률은 체질량 지수가 낮을 경우 높은 수치를 보였고, 골다공증 인지율 및 치료율은 건강 상태를 나쁘다고 인지한 여성에게서 높은 수치를 보였다. 골절경험이 있는 남녀 대상자들의 골다공증 치료율 및 인지율이 높았으며, 건강검진을 받은 여성의 골다공증 인지율이 높았다. 해당 연구는 골다공증 예방프로그램을 남성과 여성 각각의 특징에 맞게 개발되는데에 도움이 될 것이라고 언급하였다(김윤미 외, 2015).

5) 딥러닝 적용 유병률 예측모형 개발연구 - 골다공증

2008 – 2011년까지 총 4개년간 국민건강영양조사 자료에서 대퇴골 및 대퇴 경부 골밀도 검사 수치가 있고, 폐경이거나 50세 이상인 8,680명을 연구대상으로 하였다. LIME 기법을 통해 골다공증 유병에 영향을 미치는 위험 요인을 도출하고, 각 변수들의 기여도를 산출하였다. 또한 개인별로 골다공증 유병 요인의 위험 분석이 가능하도록 하였다. 그리고 7가지의 머신러닝 기법(Non−Linear Support Vector Machine, 의사결정나무, Extra Trees, Light Gradient Boosting Machine classifier, 로지스틱 회귀분석, KNN, and Multi−Layer Perceptron)을 이용한 골다공증 진단 모델을 만들었다. 해당 모델은 ROC 곡선, OST, ORAI, OSIRIS을 이용하여 성능을 비교 평가하였다. 이를 통해, 딥러닝 모델을 이용하여 골다공증의 위험을 예측하고, 설명가능한 인공지능 기술을 이용하여 개인별 골다공증 위험 분석 모델을 구축하였다. 해당 연구에서는 골다공증 진단환자 비중이 적어 딥러닝 모델을 구축하는 데에 있어 class imbalance 문제가 발생할 수 있으며 이는 모델의 부정확성을 높일 수 있다고 보았다. 하지만 해당 연구가 딥러닝 모델을 골다공증의 위험 예측에 적용한 첫 연구임에 의의가 있다(김혜연, 2022).

2010~2011년 국민건강영양조사 응답자 중 골다공증검사를 완료한 폐경여성 2,135명에서 결측치를 제외한 1,995명을 연구대상으로 선정하였다. 트리기반 머신러닝 분류 알고리즘인 의사결정나무, 랜덤포레스트, GBM, XG 부스트 모델을 사용하여, 폐경여성의 골다공증 유병여부 유무를 예측하였다. 모델 평가로는 ROC 곡선과 AUC 값, 정확도, 정밀도, 재현율, F1 스코어를 사

용하였다. 모델 별 AUC 값은 의사결정나무(0.663), GBM(0.702), 랜덤포레스트(0.704), XG 부스트(0.710)로 XG 부스트의 AUC 값이 가장 높았다. 해당 연구는 모델별 예측율 차이가 거의 없다고 볼 수 있다. 머신러닝 모델은 하이퍼 파라미터(Hyper parameter) 조절을 통하여 성능을 향상시킬 수 있으므로, 차후 연구에서는 이러한 조정을 통해 연구가 진행되면 좋을 것으로 보인다고 언급하였다(이인자, 이준호, 2020).

6) 딥러닝 적용 유병률 예측모형 개발연구 - 기타질병

2015년-2019년 국민건강영양조사 자료에서 2015년 548명, 2016년 626명, 2017년 598명, 2018년 575명, 2019년 607명을 대상으로 당뇨병성 콩팥병 예측모형을 개발하였다. kNN, 의사결정나무, LGBM, Voting, XG 부스트의 5가지 분류기(Classification)알고리즘을 적용하여 당뇨병성 콩팥병 발생 영향요인 분석 및 예측을 위한 가장 적합한 기계분석 알고리즘을 찾고자 하였고, 지표로 평균제곱근오차(Root Mean Squre Erro, RMSE)와 결정 계수(R^2)를 활용하였다. 그 결과 'XG 부스트 Regression 모델'의 정확도가 가장 높았다. 본 연구에서는 당뇨병성 콩팥병이 다양한 원인에 의하여 발생할 수 있는 만큼 유의미한 변수들을 찾는데 어려움이 있는 것으로 보았다(박윤진, 강혜경, 2022).

2016-2020년 국민건강영양조사 자료를 기반으로 총 22,914명 대상자 중 만 19세 이하 224명을 제외하여 생애주기별 고혈압 발병 요인의 차이를 머신러닝을 적용하여 분석하였다. 분류 알고리즘으로 랜덤포레스트, XG 부스트, Light GBM을 사용하였다. 그 결과, XG 부스트가 중년과 노년 모두 예측 성능이 높은 모델로 나타났다. 개인특성, 유전요인, 영양섭취가 중년의 고혈압 위험요인으로, 영양섭취, 식생활, 생활습관이 노년의 고혈압 위험요인으로 도출되었다. 이러한 연구 결과는 생애주기별 고혈압 관리를 위한 기초 자료로 유용하게 사용될 것이다. 본 연구에서는 데이터의 불균형으로 인해 연구에 포함할 수 없었던 청년층을 언급하였고, 향후 다양한 샘플링 시도를 통해 불균형을 해결하고, 추가적인 데이터 확보도 필요한 것으로 보았다(강성안 외,

2022).

2002년 1월부터 2013년 12월까지 국민건강보험공단에서 제공하는 건강검진 코호트 데이터를 활용하여 60세 이상 당뇨병 환자를 대상으로 치매 발병 여부를 예측하는 모형을 개발하였다. 설명변수는 성별, 연령, 4가지 대표적인 동반 질환(고혈압, 뇌졸중, 심장질환, 고지혈증)발병 여부, 찰슨 동반 상병 지수, 당뇨병 약의 복용 여부로, 반응변수는 치매 발병 여부로 두었다. 전체 데이터의 80%는 훈련 자료, 20%는 시험 자료로 구분하였고, 생존분석에 보편적으로 사용되는 콕스 회귀모형과 기계학습 기반의 랜덤 생존 포레스트와 딥서브를 비교한 결과, 당뇨병 환자 치매 발생 여부에 대하여 훈련 자료에서는 랜덤생존 포레스트, 시험 자료에서는 딥서브가 가장 높은 예측 성능을 보였다. 또한, 뇌졸중의 동반질환발생 여부가 치매 발병 여부를 판가름할 수 있는 가장 주요한 요인이고, 심장질환, 마비질환, 뇌혈관질환, 백혈병, 림프종을 포함한 악성 종양이 중요 요인임을 알 수 있다. 해당 연구는 전체 국민을 대상으로 실시한 국민건강보험공단의 데이터를 활용함으로써 연구 분석 결과의 신뢰도를 높였다고 볼 수 있다(정보미 외, 2020).

골다공증 관련 기존 연구를 확인하고, 국민건강영양조사 기반 골다공증 제외 질병 관련 기존 연구와 질병 유병률 예측모형 개발과 관련 기존 연구에 대해 알아보았다. 이를 통해 본 논문의 필요성과 차별점을 알 수 있었다.

기존 연구에서는 여성 위주의 연구가 대다수를 차지하였으며, 남성에 대한 연구도 일부 있지만 미미한 수준이었다. 기존 골다공증이 여성에 국한된 질환으로 인식되어 남성의 골다공증 관리가 이루어지지 않는 부분이 있는 것으로

나타났다. 따라서 남녀 전체를 대상으로 한 연구는 의미가 있다. 특히, 골다공증은 대표적인 노인성 질환 중에 하나이기 때문에 그 대상을 65세 이상 노인인구로 정하는 것은 중요한 의미를 지닌다. 또한, 국민건강영양조사 데이터 기반 분석 시, 분석 질병에 대해서 가장 많이 분석되는 질병은 고혈압, 당뇨병으로 나타났다. 하지만, 이와 유사하게 사회적 비용이 많이 지출되는 골다공증에 대한 연구는 부족한 것으로 확인되었다. 또한, 기존 골다공증 연구의 경우에는 전통적 방식의 기술통계 분석에 그치는 경우가 많았다. 조사한 바에 따르면, 머신 러닝을 적용하여 남성, 여성 전체에 대한 골다공증 유병률 예측과 관련된 최근 연구가 논문 1건에 그칠 정도로 연구가 활발히 이루어지지 않고 있다는 점을 알 수 있다. 이에, 본 연구는 골다공증 예측모형 개발에 대해서는 두번째 연구가 되겠지만, 65세 이상 노인인구에 대상에 대한 골다공증 예측모형을 개발하는 것에 대해서는 첫번째 연구로써 의의가 있다.

제 Ⅲ 장 연구 방법

제 1절 연구 대상

국민건강영양조사는 국민건강증진법 제16조에 근거하여 시행되는 법정조사로, 국민의 건강행태와 만성질환 유병, 영양 섭취 실태에 관하여 대표성과 신뢰성을 갖춘 조사이다. 1998년부터 2005년까지는 3년 주기로 실시하였고, 이후부터는 매년 시행하고 있다. 본 논문에서는 최근 5개년 조사자료인 2016-2020년 국민건강영양 조사 자료에서 65세 이상 남녀노인 인구를 대상으로 골다공증 유병률 예측모형 연구를 진행하고자 한다.

제 2절 측정 도구

1) 데이터 전처리 및 변수 선정

2016-2020년까지 총 5개년간 국민건강영양 조사 자료에서 공통 변수 372개를 선정하고, 이 중에 선행연구와 학회지 등 문헌조사를 기반으로 골다공증에 유관한 요인으로 지목하는 변수들을 선별했다. 종속변수는 골다공증 의사진단여부로 하였고, 앞서 선별한 변수들을 종속변수와의 관계에 있어 상관계수가 높은 순으로 46개를 선정하였다. 이 중에서 결측치가 높은 8개의 항목을 제거하고 총 36개를 독립변수로 선정하였다. 여기에 나이, 성별 변수도 추가하여 독립변수는 총 38개 변수로 최종 구성하였다. 데이터 전처리 진행시에, 소수형은 정수형으로 변경하고, 총 8,170명의 데이터 중 결측치를 제거한 5,365명을 대상으로 데이터 분석을 진행하였다.

<표 2> 사용 변수 설명

변수명	변수설명	내용
DM4_dg (Target)	골다공증 의사진단 여부	0. 없음 1. 있음
Year	조사연도	
Sex	성별	1. 남성 2. 여성
Age	만 나이	1-79. 1-79세 80. 80세이상
Incm	소득 4분위수(개인) ※ 4분위수 구분 기준금액 참조	1. 하 2. 중하 3. 중상 4. 상
Cfam	가구원수	1. 1명 2. 2명 3. 3명 4. 4명 5. 5명 6. 6명 이상
marri_1	결혼여부	1. 기혼 2. 미혼

변수명	변수설명	내용
marri_2	결혼상태	1. 유배우자, 동거 2. 유배우자, 별거 3. 사별 4. 이혼 88. 비해당(문항10-②)
Npins	민간의료보험 가입여부	1. 예 2. 아니오 9. 모름, 무응답
D_1_1	주관적 건강인지	1. 매우 좋음 2. 좋음 3. 보통 4. 나쁨 5. 매우 나쁨 9. 모름, 무응답
DI2_dg	이상지질혈증 의사진단 여부	0. 없음 1. 있음
DM2_dg	골관절염 의사진단 여부	0. 없음 1. 있음
DM3_dg	류마티스성 관절염 의사진단 여부	0. 없음 1. 있음

변수명	변수설명	내용
DE2_dg	갑상선 질환 의사진단 여부	0. 없음 1. 있음
DF2_dg	우울증 의사진단 여부	0. 없음 1. 있음
DN1_dg	콩팥병(신장질환) 의사진단 여부	0. 없음 1. 있음
DK4_dg	간경변증 의사진단 여부	0. 없음 1. 있음
Educ	교육수준: 학력	1. 서당/한학 2. 무학 3. 초등학교 4. 중학교 5. 고등학교 6. 2년/3년제 대학 7. 4년제 대학 8. 대학원

변수명	변수설명	내용
Graduat	교육수준: 졸업여부	1. 졸업 2. 수료 3. 중퇴 4. 재학/휴학 중 8. 비해당(문항1-①②) 9. 모름, 무응답
EC1_1	경제활동 상태	1. 예(취업자) 2. 아니오(실업자, 비경제활동인구)
BD1_11	1년간 음주빈도	1. 최근 1년간 전혀 마시지 않았다 2. 월1회미만 3. 월1회정도 4. 월2-4회 5. 주2-3회정도 6. 주4회이상 8. 비해당(문항1-①⑧)
BD2_1	한 번에 마시는 음주량	1. 1-2잔 2. 3-4잔 3. 5-6잔 4. 7-9잔 5. 10잔 이상 8. 비해당(문항2-1-①⑧)
BD2_31	폭음 빈도	1. 전혀 없음 2. 월1회미만 3. 월1회정도 4. 주1회정도 5. 거의 매일 8. 비해당(문항2-1-①⑧) 9. 모름, 무응답

변수명	변수설명	내용
BD7_4	가족/의사의 절주 권고 여부	1. 없었다 2. 과거에는 있었지만 최근1년 동안에는 없었다 3. 최근 1년 동안에 그런 적이 있었다 8. 비해당(문항1-①, 청소년, 소아)
BD7_5	(성인) 1년간 음주문제 상담 여부	1. 예 2. 아니오
BS1_1	평생 일반담배(궐련) 흡연 여부	1. 5갑(100개비) 미만 2. 5갑(100개비) 이상 3. 피운 적 없음 9. 모름, 무응답
BE3_71	고강도 신체활동 여부: 일	1. 예 2. 아니오 9. 모름, 무응답
BE5_1	1주일간 근력운동 일수	1. 전혀 하지 않음 2. 1일 3. 2일 4. 3일 5. 4일 6. 5일이상 9. 모름, 무응답

변수명	변수설명	내용
HE_ht	신장	□ CM
HE_wt	체중	□ kg
HE_wc	허리둘레	□ CM
HE_BMI	체질량지수	□ kg/m2
HE_obe	비만 유병여부	1. 저체중 2. 정상 3. 비만전단계 4. 단계비만 5. 2단계비만 6. 3단계비만
HE_chol	총콜레스테롤	□ mg/Dl
HE_HDL_st2	HDL-콜레스테롤(보정값)	□ mg/Dl
HE_TG	중성지방	□ mg/Dl
HE_HCHOL	고콜레스테롤혈증 유병여부 (19세이상)	0. 없음 1. 있음
HE_HTG	고중성지방혈증 유병여부 (10세이상)	0. 없음 1. 있음
N_CA	칼슘 섭취량	1일 칼슘 섭취량(㎍)
N_VITD	비타민D 섭취량	1일 비타민D 섭취량(㎍)

출처: 국민건강영양조사 제7기-8기 원시자료 이용지침서(2016-2018, 2019-2020).

2) 종속 변수

골다공증의 의사진단 여부(0:없음, 1:있음), 즉, 골다공증 유병여부를 종속변수로 한다.

3) 독립 변수

각 독립변수에 대해서는 좀 더 정확한 분석 예측을 진행하기위해 파이썬 Pandas를 이용하여 더미변수 처리를 진행하였다. 선정된 독립변수는 크게 인구학적 요소, 기본건강요소, 건강행태적 요소, 보유질병 요소 4가지 군으로 나누어서 분류하였다. 인구학적 요소로 성별, 만나이, 소득 4분위수(개인), 가구원수, 결혼여부, 결혼상태, 민간의료보험 가입여부, 교육수준 학력과 졸업여부, 경제활동상태가 있다. 기본건강 요소로 신장, 체중, 허리둘레, 체질량지수, 총콜레스테롤 수치, HDL-콜레스테롤 수치, 중성지방 수치가 있다. 건강행태적 요소로 주관적 건강인지, 1년간 음주빈도, 한 번에 마시는 음주량, 폭음 빈도, 가족/의사의 절주 권고 여부, 1년간 음주문제 상담 여부, 평생 일반담배(궐련) 흡연 여부, 고강도 신체여부(일), 1주일간 근력운동 일수, 칼슘 섭취량, 비타민D 섭취량이 있다. 보유질병 요소로 이상지질혈증, 골관절염, 류마티스성 관절염, 갑상선 질환, 우울증, 콩팥병(신장 질환), 간경변증, 비만, 고콜레스테롤혈증, 고중성지방혈증 진단 여부가 있다.

제 3절 분석 방법 및 절차

본 연구의 모든 분석은 파이썬 3.8.16버전을 이용하여 진행하였다. 65세 이상 노인 골다공증 유병률 예측모형을 만들기 위해 머신러닝을 사용하였다. 머신러닝은 크게 지도학습, 비지도학습, 강화학습으로 분류할 수 있으며, 지도학습은 주어진 데이터와 답을 이용해서 값을 예측하는 학습 방법이고, 비지도학습은 데이터 자체에서 유용한 패턴을 찾아내는 학습 방법이다. 강화학습은 기계가 선택과 피드백을 반복하는 방법으로 장기적인 관점에서 얻는 이득을 최대화 시키는 학습 방법이다(김승연 외, 2017).

본 연구에서는 머신러닝의 지도학습방법을 채택하였다. 지도학습에서는 회귀, 분류, 랭킹/추천으로 나누어지는데, 분류문제는 답을 도출하는 데이터에 해당되는 변수 이산 값일 때 사용되며, 회귀문제는 숫자 값이 크고 작음에 의미를 부여해 예측하는 것이다(김승연 외, 2017),(아카바 신야 외, 2019). 본 연구에서는 유병여부를 0,1로 구분하여 값을 구하고자 하기 때문에 분류 알고리즘을 적용하기로 했다. 이러한 분류 알고리즘은 퍼셉트론, 서포트 백터 머신, 로지스틱 회귀, k-NN, 결정나무 알고리즘, 신경망 등으로 이루어져 있다(아리가 미치아키 외, 2018). 해당 분류 알고리즘 중에 선행연구상 질병예측에 많이 사용되면서 미국 데이터 사이언티스트가 집필한 책에서 TOP 10 알고리즘으로 손꼽힌 로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트를 본 연구에 사용하기로 하였다(권시현, 2022), 선행연구를 살펴보면 질병예측 모형에 가장 많이 사용되는 알고리즘은 로지스틱 회귀 분석이며, 실제 논문에 따르면 최근 10년간 대사증후군 예측에 사용된 분석방법으로는 로지

스틱 회귀분석이 63.6%에 달한다고 한다(성대경 외, 2021). 질병 예측모형에 사용시, 의사결정나무와 랜덤포레스트, XG 부스트 알고리즘도 장점이 있기 때문에, 다른 질병군 예측모형에서 우수한 예측성능을 가진 알고리즘으로 확인되기도 한다. 회귀모형은 변수들의 독립성과 정규성을 가정하는데, 질병처럼 분포의 정규성이 위배되는 데이터의 경우 정확한 결과를 도출해 내는 것이 어렵다. 회귀모형의 한계점을 해결하는 방법으로 최근 다양한 분야에서 랜덤포레스트와 XG 부스트가 널리 사용되고 있다고 언급했다. 실제 한 연구에서는 로지스틱 회귀분석, 랜덤포레스트, XG 부스트를 이용하여 우리나라 지역사회 노인들의 신체기능 장애를 예측하는 모형을 개발하고, 정확도 0.67, 민감도 0.81, 특이도 0.75로 도출된 SMOTE 기반 XG 부스트를 예측성능이 가장 우수한 모델로 선정하였다(변해원, 2021). 또 다른 연구에서는 유방암 종양 분류에 트리 기반 알고리즘인 랜덤포레스트와 부스팅 기반 알고리즘인 XG 부스트를 사용하였다. 랜덤포레스트, XG 부스트 모두 평가지표(정확도, 정밀도, 재현율, AUC 값) 점수가 95% 이상의 높은 결과를 나타내었다. 또한, 해당 연구에서는 머신러닝을 학습 및 테스트를 통해 타겟의 속성값을 추론하거나 분류할 수 있고, 상관성이 높은 속성을 파악할 수 있는 강력한 도구로 보았다. 머신러닝은 특히 의료분야에서 진단 또는 판단을 하는데 널리 사용되고 있다고 언급하였다(윤우진 외, 2021). 대사증후군 유병 예측 모형을 생성하기 위한 연구에서는 알고리즘으로 의사결정나무, 로지스틱 회귀분석, 인공신경망을 사용하였다. 최종 예측모형은 의사결정나무일 경우 90.32%의 가장 높은 예측력을 나타내고 있었다(김한결 외, 2016). 백내장 예측모형을 개발하기 위한 연구에서는 의사결정나무, 나이브 베이즈, 배깅, 아킹, 랜덤포레스트 알고리즘

의 예측성능을 비교 분석하였고, 그 결과 정확도 67.16%, 민감도 72.28%로 랜덤포레스트가 가장 높은 예측력을 가진 모형으로 확인되었다(한은정 외, 2009).

본 연구에서는 이러한 근거를 바탕으로 연구에 적용할 알고리즘을 선택하였다. 파이썬을 기반으로 한 머신러닝 라이브러리인 scikit-learn에 내장된 linear model인 로지스틱 회귀분석, XG 부스트 모듈에서 XGB Classifier, sklearn ensemble모듈에서 의사결정나무 Classifier, 랜덤포레스트 Classifier 알고리즘을 적용하여 학습 모델을 만들고 성능지표를 도출하였다.

1) 로지스틱 회귀분석

로지스틱 회귀분석은 독립 변수의 선형 결합을 바탕으로 사건 발생 가능성을 예측하는데 사용되는 통계 기법이다(Cox, 1958). 로지스틱 회귀 분석은 선형 회귀 분석에 바탕을 두고 있다. 하지만, 선형 회귀 분석은 연속된 변수 예측에 쓰이고, 로지스틱 회귀 분석은 Yes/No처럼 2가지로 나누어지는 분류 문제에 쓰인다. 그 중에서 이항 로지스틱 회귀모형은 연속형 또는 범주형 설명변수에 대해 반응 변수가 0,1로 나뉘는 범주형 변수인 경우 주로 사용한다. 그 예로, 질병발생여부(종속변수)와 나이, 병명(설명변수)와의 관련성을 알아보는 모형을 개발할 때 로지스틱 회귀분석을 사용할 수 있다(계묘진, 2013).

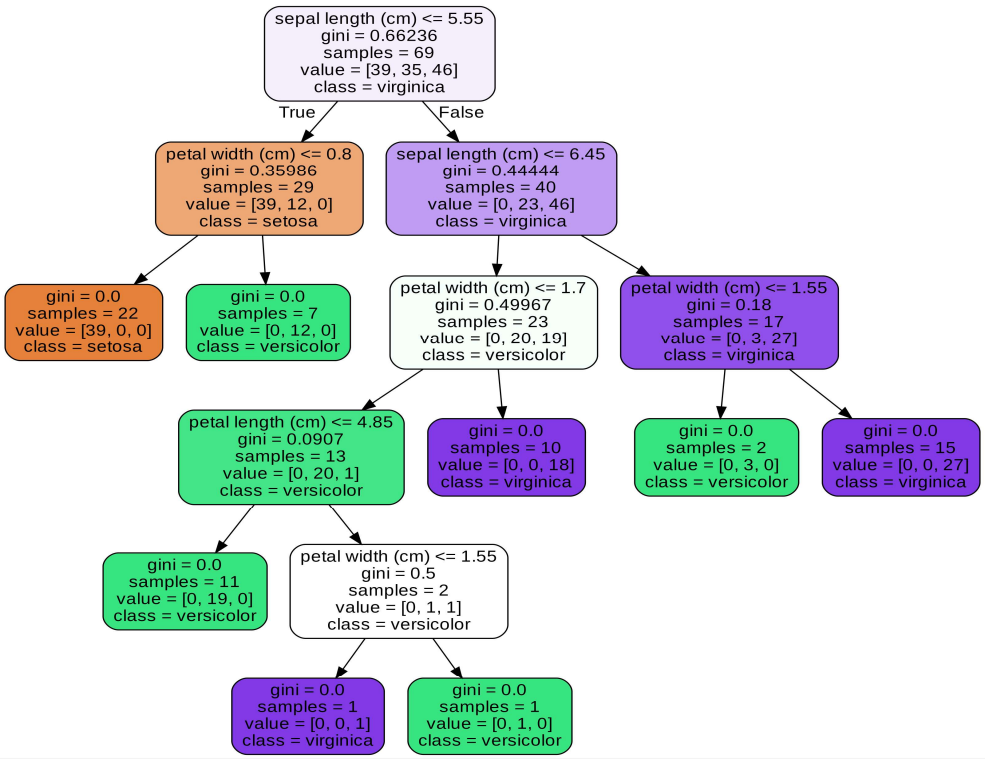
2) XG 부스트

트리 부스팅은 매우 효과적이고, 널리 사용되는 기계 학습 방법이다. 머신러닝 문제에서 최첨단 결과를 달성하기 위해 데이터 과학자들이 널리 사용하는 XG 부스트라는 확장 가능한 엔드 투 엔드 트리 부스팅 시스템이다.

XG 부스트는 많은 컴퓨터 언어 플랫폼에서 사용 가능하며, 머신러닝과 데이터 마이닝 챌린지에서 우수한 성적을 나타내었다. Kaggle에서는 우승자가 사용했던 알고리즘 중에 17개가 XG 부스트였다. XG 부스트의 장점은 확장성이다. 속도는 기존 모델보다 10배 빠르고, 메모리를 많이 차지하지 않아 제한된 자원에서도 분석이 가능하다. 또한, 분류문제에서 XG 부스트의 효과가 좋은 것으로 나타났다. 오픈소스 패키지로 자유롭게 사용 가능하다(Chen & Guestrin, 2016). 랜덤포레스트는 각 트리를 독립적으로 만드는 알고리즘이다. 반면, 부스팅은 순차적으로 트리를 만들어 이전 트리로부터 더 나은 트리를 만들어내는 알고리즘이다. 부스팅 알고리즘은 트리 모델을 기반으로 한 최신 알고리즘 중 하나로, 랜덤포레스트보다 훨씬 빠른 속도와 더 좋은 예측 능력을 보여준다. 이에 속하는 대표적인 알고리즘으로 XG 부스트, 라이트 GBM(LightGBM), 캣부스트(CatBoost) 등이 있다. 그 중, XG 부스트가 가장 먼저 개발되었고, 가장 널리 활용된다. XG 부스트는 손실 함수뿐만 아니라 모형 복잡도까지 고려한다. 해당 알고리즘은 지도학습의 알고리즘으로 회귀, 분류 문제 모두에서 사용된다(권시현, 2022).

3) 의사결정나무

<그림 1> 의사결정나무 모형



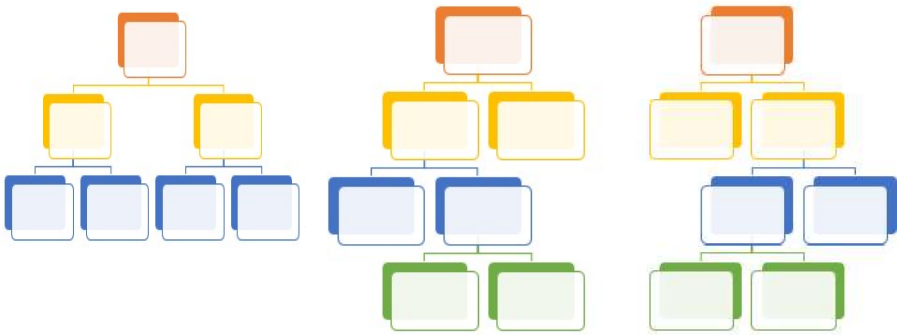
의사를 결정하거나 분류 및 예측하는데 사용하는 트리로 가장 큰 조건의 트리 뿌리를 만들고, 세부조건의 트리 가지를 만들며, 해결방안은 트리의 잎 노드로 의사결정 나무를 형성하여 분석하는 알고리즘이다(이근영, 2015).

의사 결정 트리 모델의 가장 중요한 단계 모델 구축은 분할, 중지 및 가지치기이다. 분할은 Target 변수와 관련된 input 변수로 여러 카테고리를 만들어내는 것을 의미한다. 중지는 복잡성이 커져서 오버피팅 가능성이 생기면 멈추는 것이며, 정지규칙을 만족하면 중단한다. 가지치기는 사전에 유의하지 않

은 것을 제거하거나 사후에 정확도를 높이기 위해 가지치기를 한다(Song & Lu, 2015). 의사결정나무는 실시간 적용이 가능하고 분류과정이 트리구조에 의한 추론 규칙(If- then구조)으로 표현되기 때문에 쉽게 이해하고 설명 가능하다(이근영, 2015). 또한, 전형적인 노이즈가 있거나 불완전한(빈 값) 데이터가 잘 처리된다는 장점이 있다. 의사결정나무는 매우 빠르고 정확하지만 보유한 지식에 따라 해석이 달라진다는 단점이 있다. 특정 분야에 대해 의사결정나무의 해석 결과가 전문가가 알고 있는 전문지식과 다른 경우가 종종 발생한다. 예를 들어, 알고리즘 해석 상 기준치 이하이면 위험하다고 판단되지만, 엔지니어 관점에서는 기준치 이상일 때 위험하다고 판단하여 관리하는 경우가 이에 해당된다(Quinlan, 1986). 의사결정나무는 수많은 트리 기반 모델의 기본 모델이 되는 중요 모델로서, 각 변수에 대한 기울기 값들을 최적화하여 모델을 만들어가는 선형 모델과는 달리, 각 변수의 특정 지점을 기준으로 데이터를 분류해가며 예측 모델을 만든다. 이 알고리즘은 지도학습의 알고리즘으로 회귀, 분류 문제 모두에서 사용된다(권시현, 2022).

4) 랜덤포레스트

<그림 2> 랜덤포레스트 모형



랜덤포레스트는 많은 수의 의사결정나무를 생성한 뒤, 각 트리들의 예측결과를 종합하고 가장 유력한 클래스를 선택하여 예측 정확도를 높이는 방법이다. 대수의 법칙 때문에 과적합 문제를 예방하고, 분류와 회귀에서 모두 높은 성능을 보이는 모델이다(Breiman, 2001). 빠른 학습 속도를 가지므로 많은 양의 데이터 처리 능력을 가지며, 단계별 노드의 수를 조절하여 멀티클래스로 쉽게 확정이 가능하다(이근영, 2015). 랜덤포레스트 모델은 결정 트리의 단점인 오버피팅 문제를 완화시켜주는 발전된 형태의 트리 모델이다. 랜덤으로 생성된 무수히 많은 트리를 생성하여 예측을 하기 때문에 랜덤포레스트라 불린다. 지도학습의 알고리즘으로 회귀, 분류 문제 모두에서 사용된다(권시현, 2022).

이러한 분석 알고리즘을 적용하고, 예측 성능을 확인하여 비교하기 위해서 65세 이상 노인인구 5,365명의 데이터를 학습용 데이터와 테스트용 데이터로

구분하였다. 학습용 데이터는 전체 데이터의 80%로, 테스트용 데이터는 전체 데이터의 20%로 선택하였다. 종속변수인 y 에는 골다공증 유병 유무를 Target 변수로 정의하였고, 독립변수인 x에는 골다공증 유관 변수로 최종 선정된 38개 변수에 대해서 전처리와 결측지 제거를 완료하고 더미변수화 시킨 변수들로 정의하였다.

머신러닝 분류모델인 로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤 포레스트 모델의 평가지표는 이진분류에서 널리 사용되는 정확도, 정밀도, 재현율, F1 스코어, ROC 곡선을 기반으로 도출된 AUC값으로 하였다.

분류문제는 AUC 값, 정밀도 및 ROC 곡선을 기반으로 모델을 평가하여 수행된다. 이러한 지표를 통해 데이터베이스의 성능을 평가하고 완전한 정확성을 위한 유용한 결과를 얻을 수 있다(Sharma et al, 2022). 그리고, 분류모형은 혼동행렬로 모형을 비교할 수 있다(김은하, 2015). 실제 다양한 선행연구에서 모델평가 지표로 ROC 곡선, AUC 값(성대경 외, 2021), 정확도, 정밀도, 재현율, AUC 값(윤우진 외, 2021), 정확도(김한결 외, 2016), 민감도, 특이도, 정확도(한은정 외, 2009), ROC 곡선, AUC 값 정확도, 민감도, 특이도(변해원, 2021)를 채택하는 등 본 연구에서 채택한 성능 평가 방법을 적용하여 진행하였다. 이에 본 연구에서는 혼동행렬을 기반으로 산출되는 정확도, 정밀도, 재현율, F1 스코어, ROC 곡선, AUC 값을 통해 성능 평가를 진행했다.

5) 혼동행렬

혼동 행렬은 일련의 테스트 데이터에 대한 분류 모델의 성능을 설명하는 행렬로 표시된다(<그림3> 참고). 혼동행렬은 예측된 값과 실제 값의 개수를 나타낸다. " True Negatives(TN)"은 정확하게 분류된 부정 사례의 수, " True Positives(TP)"는 정확하게 분류된 양성의 사례 수, "False Positives(FP)"는 양성으로 분류된 실제 음성의 사례 수, "False Negatives(FN)"은 음성으로 분류된 실제 양성의 사례 수를 의미한다(Sharma et al, 2022).

<그림 3> 혼동행렬 모형

	예측	
실제	True Positives	False Positives
	False Negatives	True Negatives

혼동행렬을 통해서 계산할 수 있는 몇 가지 방정식이 있는데, 정확도, 정밀도, 재현율이 그것이다.

6) 정확도

정확도는 올바른 총 예측 수의 비율을 제공한다.

$$\frac{\text{TruePositives} + \text{TrueNegatives}}{\text{TruePositives} + \text{TrueNegatives} + \text{FalsePositives} + \text{FalseNegatives}}$$

7) 정밀도

정밀도 또는 양성 예측 값은, 전체 예측된 양성 사례 중에서 양수 값의 비율이다. 즉, 정밀도는 올바르게 식별된 양성 값의 비율이다.

$$\frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

8) 재현율

전체 실제 양성 사례 중 양성 값의 비율이다. 즉, 올바르게 식별된 실제 양성 사례의 비율이며, 민감도라고도 불린다.

$$\frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}$$

9) F1 스코어

F1 스코어는 정밀도와 재현율의 조화 평균이다.

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

10) 특이도

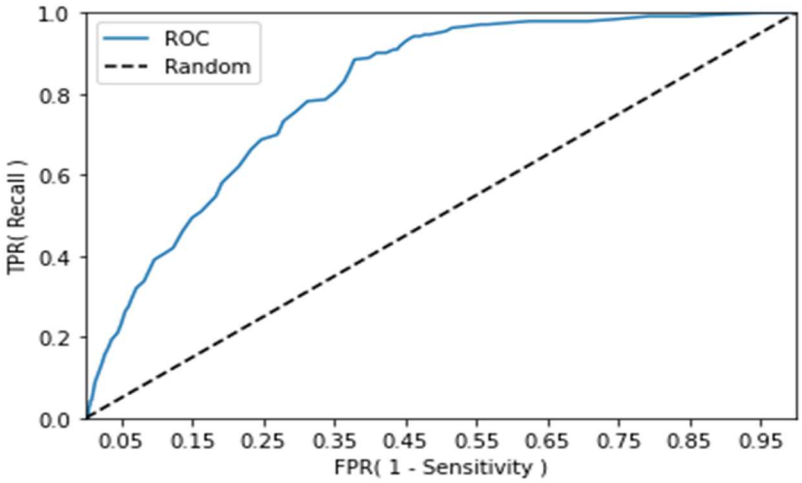
전체 실제 음성 사례 중 음성 값의 비율이다. 즉, 올바르게 식별된 실제 음성 사례의 비율이다. 특이도에서 파생된 FP rate는 (1 − 특이도)이다.

$$\frac{\text{TrueNegatives}}{\text{TrueNegatives} + \text{FalsePositives}}$$

11) ROC 곡선, AUC값

ROC 곡선은 다양한 임계값에 대한 재현율과 FP rate로 그려지는 곡선이다. AUC 값은 이 ROC 곡선 아래의 면적이다. 분류 모델의 성능을 측정하는데 사용되며(Vexler et al, 2011), AUC 값이 클수록 성능이 향상된다(Sharma et al, 2022).

<그림 4> ROC 곡선 모형



scikit-learn metrics 모듈에 내장된 accuracy_score, precision_score, recall_score, f1_score, confusion_matrix를 이용하여 분류모델별 정확도, 정밀도, 재현율, F1 스코어, AUC 값과 혼동행렬을 구하였다.

정확도, 정밀도, 재현율 분류문제에서 분류기준을 어떤 값으로 설정하는지에 따라 정확도, 정밀도, 재현율은 달라진다. 이에 따라, 본 연구에서는 scikit-learn preprocessing 모듈에 Binarizer을 사용하여 분류기준인 임계값을 0.4, 0.45, 0.50, 0.55, 0.60으로 했을 때의 분류모델별 정확도, 정밀도, 재현율, F1 스코어의 결과도 비교했다. 재현율과 FP rate의 관계를 나타내는 ROC 곡선을 도출하고, AUC 값도 구하여서 모델성능평가를 진행했다. scikit-learn metrics 모듈에 내장된 ROC curve를 적용하여, ROC곡선을 구하였고, 이때 matplotlib를 이용하여 시각화를 진행하였다. ROC AUC score를 이용하여, AUC 값도 구하였다. 해당 모델 성능 지표들을 통해서, 65세 이상 남녀노인 인구의 골다공증 유병률 예측을 좀더 정확하게 할 수 있는 모델을 찾고자 하였다.

또한, 채택된 모델이 실제 골다공증 유병률을 예측할 때, 어떠한 변인들이 영향을 미치는지 파악하였다. 이때, Model.feature_importances로 변수 중요도를 보았고, plotly 패키지의 express모듈을 이용하여 시각화를 진행하였다. 이를 통해 65세 이상 남녀노인 인구의 골다공증 유병률 예측 모델을 다각적으로 검토하고 분석할 수 있었다.

제 Ⅳ장 연구 결과

제 1절 기술통계 분석

성별, 소득4분위수(개인), 가구원수, 결혼여부, 결혼 상태, 민간의료보험 가입여부에 대하여 분석을 진행하였다. 그 결과, 남성은 2,335명(43.52%), 여성은 3,030명(56.48%)이었으며, 평균 연령대는 72.8세(표준편차: 5.02)였다. 소득 4분위수(개인)기준으로 상, 중상, 중하, 하로 인원수와 비중을 살펴보면, 상은 1,362명(25.39%), 중상은 1,352명(25.2%), 중하는 1,358명(25.31%), 하는 1,293명(24.1%)이었다. 각 기준별로 소득기준은 성별 및 금액에 따라 상이한 것으로 나타났다. 가구원수 기준으로 1-2명은 4,133명(77.04%), 3-5명은 1,140명(21.25%), 6명 이상 92명(1.71%)이었다. 결혼여부 기준에서 기혼자는 5,322명(99.2%), 미혼자는 43명(0.8%)이었으며, 기혼자 중에 유배우자:동거상태인 경우는 3,589명(66.9%), 유배우자:별거상태인 경우는 54명(1.01%), 사별 상태인 경우는 1,455명(27.12%), 이혼 상태인 경우는 224명(4.18%), 해당 없음 응답은 43명(0.8%)이었다. 민간의료보험 가입기준과 관련하여, 민간의료보험 가입자는 2,484명(46.3%), 미가입자는 2,844명(53.01%), 가입상태를 모르는 경우는 37명(0.69%)으로 나타났다. 경제활동상태가 취업인 경우는 1,775명(33.08%), 실업, 비경제활동인구인 경우는 3,590명(50.52%)이었다. 교육수준은 서당 9명(0.17%), 무학 611명(11.39%), 초등학교 2,074명(38.66%), 중학교 1,039명(19.37%), 고등학교 993명(18.51%), 2-3년제 대학 113명(2.11%), 4년제 대학 409명

(7.62%), 대학원 117명(2.18%)이었으며, 졸업여부를 살펴보면, 졸업 3,516명(65.54%), 수료 17명(0.32%), 중퇴 1,198명(22.33%), 재학/휴학 중 13명(0.24%), 비해당 620명(11.56%)이었다(<표 3> 참고).

<표 3> 인구학적 요소 기술통계

구분		인원수(명)	비율(%)	구분		인원수(명)	비율(%)
성별	남	2,335	43.52	결혼여부	기혼	5,322	99.20
	여	3,030	56.48		미혼	43	0.80
소득 4분위수 (개인)	하	1,293	24.10	결혼상태	유배우자, 동거	3,589	66.90
	중하	1,358	25.31		유배우자, 별거	54	1.01
	중상	1,352	25.20		사별	1,455	27.12
	상	1,362	25.39		이혼	224	4.18
					해당 없음	43	0.80
가구원수	1-2명	4,133	77.04	민간의료보험 가입여부	가입함	2,484	46.30
	3-5명	1,140	21.25		가입 안함	2,844	53.01
	6명 이상	92	1.71		모름	37	0.69
교육수준	서당	9	0.17	경제활동상태	취업	1,775	33.08
	무학	611	11.39		실업	3,590	50.52
	초등학교	2,074	38.66	졸업여부	졸업	3,516	65.54
	중학교	1,039	19.37		수료	17	0.32
	고등학교	993	18.51		중퇴	1,198	22.33
	2-3년제 대학	113	2.11		재학/휴학 중	13	0.24
	4년제 대학	409	7.62		비해당	620	11.56
	대학원	117	2.18		무응답	1	0.02

제 2절 골다공증 유병률 예측 결과 비교 분석

1) 혼동행렬

분류 기준 0.5로 했을 때 각 모형별 혼동행렬은 다음과 같다.

로지스틱 회귀분석의 경우, 1,073개의 데이터 중에 777개 데이터를 골다공증에 걸리지 않은 인원을 걸리지 않았다고 올바르게 예측하였고, 1,073개의 데이터 중에 54개 데이터를 골다공증에 걸린 인원을 걸렸다고 올바르게 예측하였다(<표4> 참고).

XG 부스트의 경우, 1,073개의 데이터 중에 765개 데이터를 골다공증에 걸리지 않은 인원을 걸리지 않았다고 올바르게 예측하였고, 1,073개의 데이터 중에 68개 데이터를 골다공증에 걸린 인원을 걸렸다고 올바르게 예측하였다(<표5> 참고).

의사결정 나무의 경우, 1,073개의 데이터 중에 639개 데이터를 골다공증에 걸리지 않은 인원을 걸리지 않았다고 올바르게 예측하였고, 1,073개의 데이터 중에 118개 데이터를 골다공증에 걸린 인원을 걸렸다고 올바르게 예측하였다(<표6> 참고).

랜덤포레스트의 경우, 1,073개의 데이터 중에 784개 데이터를 골다공증에 걸리지 않은 인원을 걸리지 않았다고 올바르게 예측하였고, 1,073개의 데이터 중에 64개 데이터를 골다공증에 걸린 인원을 걸렸다고 올바르게 예측하였다(<표7> 참고).

<표 4> 로지스틱 회귀분석 (분류기준: 0.5) 혼동행렬

구분		예측	
		0	1
실제	0	777	53
	1	189	54

<표 5> XG 부스트 (분류기준: 0.5) 혼동행렬

구분		예측	
		0	1
실제	0	765	65
	1	175	68

<표 6> 의사결정나무 (분류기준: 0.5) 혼동행렬

구분		예측	
		0	1
실제	0	639	191
	1	125	118

<표 7> 랜덤 포레스트 (분류기준: 0.5) 혼동행렬

구분		예측	
		0	1
실제	0	784	46
	1	179	64

2) 정확도, 정밀도, 재현율, F1 스코어

분류문제에서 분류기준을 어떤 값으로 설정하는지에 따라 정확도, 정밀도, 재현율, F1 스코어가 달라진다. 분류기준을 0.4, 0.45, 0.5, 0.55, 0.6으로 했을 때의 모형별 정확도, 정밀도, 재현율, F1 스코어 결과는 다음과 같다.

분류기준 0.4 결과에서 정확도는 랜덤포레스트(0.7633)가 가장 높았으며, 그 다음, 로지스틱 회귀분석(0.7586), XG 부스트(0.7558), 의사결정나무(0.7055) 순서이다. 정밀도는 랜덤포레스트(0.4788)가 가장 높았으며, 그 다음은 XG 부스트(0.4678), 로지스틱 회귀분석(0.4672), 의사결정나무(0.3819) 순서이다. 재현율은 XG 부스트(0.5679)가 가장 높았으며, 그 다음은 랜덤포레스트(0.5103), 의사결정나무(0.4856), 로지스틱 회귀분석(0.4691) 순서이다. F1 스코어는 XG 부스트(0.5130)가 가장 높았으며, 그 다음은 랜덤포레스트(0.4940), 로지스틱 회귀분석(0.4682), 의사결정나무(0.4275) 순서이다(<표8> 참고).

<표 8> 정확도, 정밀도, 재현율, F1 스코어(분류기준 0.4) 결과

Model Data	정확도	정밀도	재현율	F1 스코어
로지스틱 회귀분석	0.7586	0.4672	0.4691	0.4682
XG 부스트	0.7558	0.4678	0.5679	0.5130
의사결정나무	0.7055	0.3819	0.4856	0.4275
랜덤포레스트	0.7633	0.4788	0.5103	0.4940

분류기준 0.45 결과에서 정확도는 랜덤포레스트(0.7875)가 가장 높았으며, 그 다음은 로지스틱 회귀분석(0.7735), XG 부스트(0.7651), 의사결정나무(0.7055) 순서이다. 정밀도는 랜덤포레스트(0.5429)가 가장 높았으며, 그 다음은 로지스틱 회귀분석(0.5000), XG 부스트(0.4783), 의사결정나무(0.3819) 순서이다. 재현율은 의사결정나무(0.4856)가 가장 높았으며, 그 다음은 XG 부스트(0.4074), 랜덤포레스트(0.3909), 로지스틱 회귀분석(0.3539) 순서이다. F1 스코어는 랜덤포레스트(0.4545)가 가장 높았으며, 그 다음은 XG 부스트(0.4400), 의사결정나무(0.4275), 로지스틱 회귀분석(0.4145) 순서이다(<표9> 참고).

<표 9> 정확도, 정밀도, 재현율, F1 스코어(분류기준 0.45) 결과

Model Data	정확도	정밀도	재현율	F1 스코어
로지스틱 회귀분석	0.7735	0.5000	0.3539	0.4145
XG 부스트	0.7651	0.4783	0.4074	0.4400
의사결정나무	0.7055	0.3819	0.4856	0.4275
랜덤포레스트	0.7875	0.5429	0.3909	0.4545

분류기준 0.5 결과에서 정확도는 랜덤포레스트(0.7903)가 가장 높았으며, 그 다음은 XG 부스트(0.7763), 로지스틱 회귀분석(0.7745), 의사결정나무(0.7055) 순서이다. 정밀도는 랜덤포레스트(0.5818)가 가장 높았으며, 그 다음은 XG 부스트(0.5113), 로지스틱 회귀분석(0.5047), 의사결정나무(0.3819) 순서이다. 재현율은 의사결정나무(0.4856)가 가장 높았으며, 그 다음은 XG 부스트(0.2798), 랜덤포레스트(0.2634), 로지스틱 회귀분석

(0.2222) 순서이다. F1 스코어는 의사결정나무(0.4275)가 가장 높았으며, 그 다음은 랜덤포레스트(0.3626), XG 부스트(0.3617), 로지스틱 회귀분석 (0.3086) 순서이다(<표10> 참고).

<표 10> 정확도, 정밀도, 재현율, F1 스코어(분류기준 0.5) 결과

Model Data	정확도	정밀도	재현율	F1 스코어
로지스틱 회귀분석	0.7745	0.5047	0.2222	0.3086
XG 부스트	0.7763	0.5113	0.2798	0.3617
의사결정나무	0.7055	0.3819	0.4856	0.4275
랜덤포레스트	0.7903	0.5818	0.2634	0.3626

분류기준 0.55 결과에서 정확도는 랜덤포레스트(0.7884)가 가장 높았으며, 그 다음은 XG 부스트(0.7838), 로지스틱 회귀분석(0.7763), 의사결정나무(0.7055) 순서이다. 정밀도는 랜덤포레스트(0.6333)가 가장 높았으며, 그 다음은 XG 부스트(0.5632), 로지스틱 회귀분석(0.5205), 의사결정나무 (0.3819) 순서이다. 재현율은 의사결정나무(0.4856)가 가장 높았으며, 그 다음은 XG 부스트(0.2016), 랜덤포레스트(0.1564), 로지스틱 회귀분석 (0.1564) 순서이다. F1 스코어는 의사결정나무(0.4275)가 가장 높았으며, 그 다음은 XG 부스트(0.2970), 랜덤포레스트(0.2508), 로지스틱 회귀분석 (0.2405) 순서이다(<표11> 참고).

<표 11> 정확도, 정밀도, 재현율, F1 스코어(분류기준 0.55) 결과

Model Data	정확도	정밀도	재현율	F1 스코어
로지스틱 회귀분석	0.7763	0.5205	0.1564	0.2405
XG 부스트	0.7838	0.5632	0.2016	0.2970
의사결정나무	0.7055	0.3819	0.4856	0.4275
랜덤포레스트	0.7884	0.6333	0.1564	0.2508

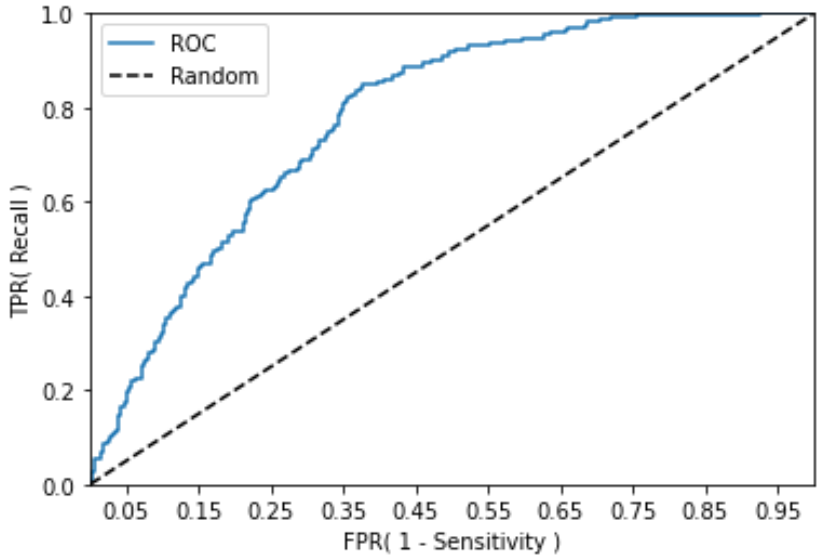
분류기준 0.6 결과에서 정확도는 XG 부스트(0.7819)가 가장 높았으며, 그 다음은 랜덤포레스트(0.7791), 로지스틱 회귀분석(0.7754), 의사결정나무(0.7055) 순서이다. 정밀도는 랜덤포레스트(0.6364)가 가장 높았으며, 그 다음은 XG 부스트(0.5882), 로지스틱 회귀분석(0.5238), 의사결정나무(0.3819) 순서이다. 재현율은 의사결정나무(0.4856)가 가장 높았으며, 그 다음은 XG 부스트(0.1235), 로지스틱 회귀분석(0.0905), 랜덤포레스트(0.0576) 순서이다. F1 스코어는 의사결정나무(0.4275)가 가장 높았으며, 그 다음은 XG 부스트(0.2041), 로지스틱 회귀분석(0.1544), 랜덤포레스트(0.1057) 순서이다(<표12> 참고).

<표 12> 정확도, 정밀도, 재현율, F1 스코어(분류기준 0.6) 결과

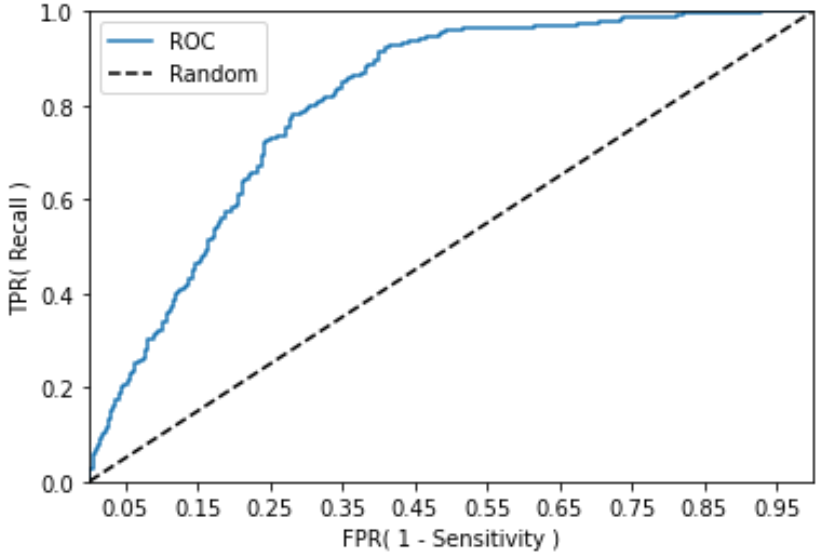
Model Data	정확도	정밀도	재현율	F1 스코어
로지스틱 회귀분석	0.7754	0.5238	0.0905	0.1544
XG 부스트	0.7819	0.5882	0.1235	0.2041
의사결정나무	0.7055	0.3819	0.4856	0.4275
랜덤포레스트	0.7791	0.6364	0.0576	0.1057

3) ROC 곡선, AUC 값

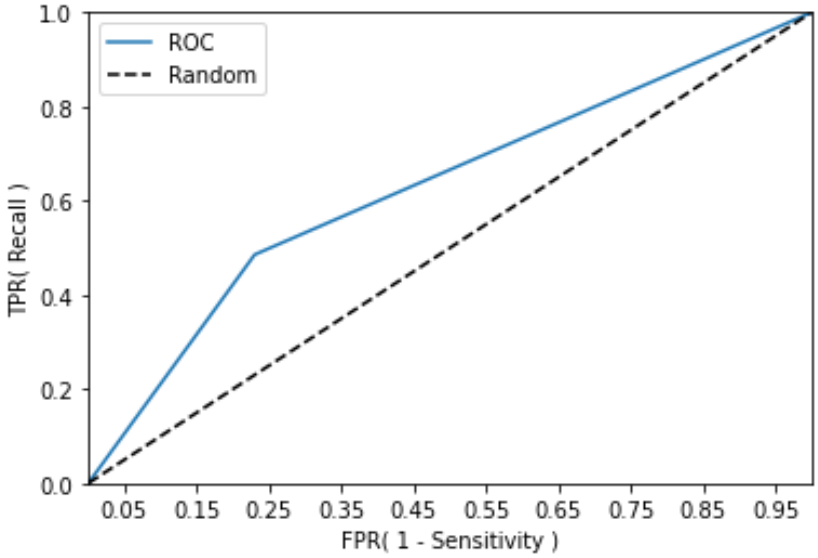
<그림 5> 로지스틱 회귀분석 ROC 곡선



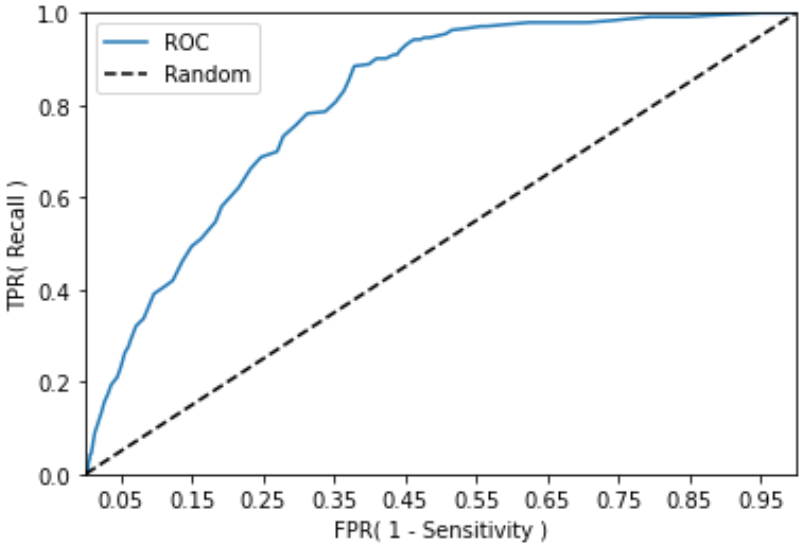
<그림 6> XG 부스트 ROC 곡선



<그림 7> 의사결정나무 ROC 곡선



<그림 8> 랜덤포레스트 ROC 곡선



<표 13> AUC 값 비교표

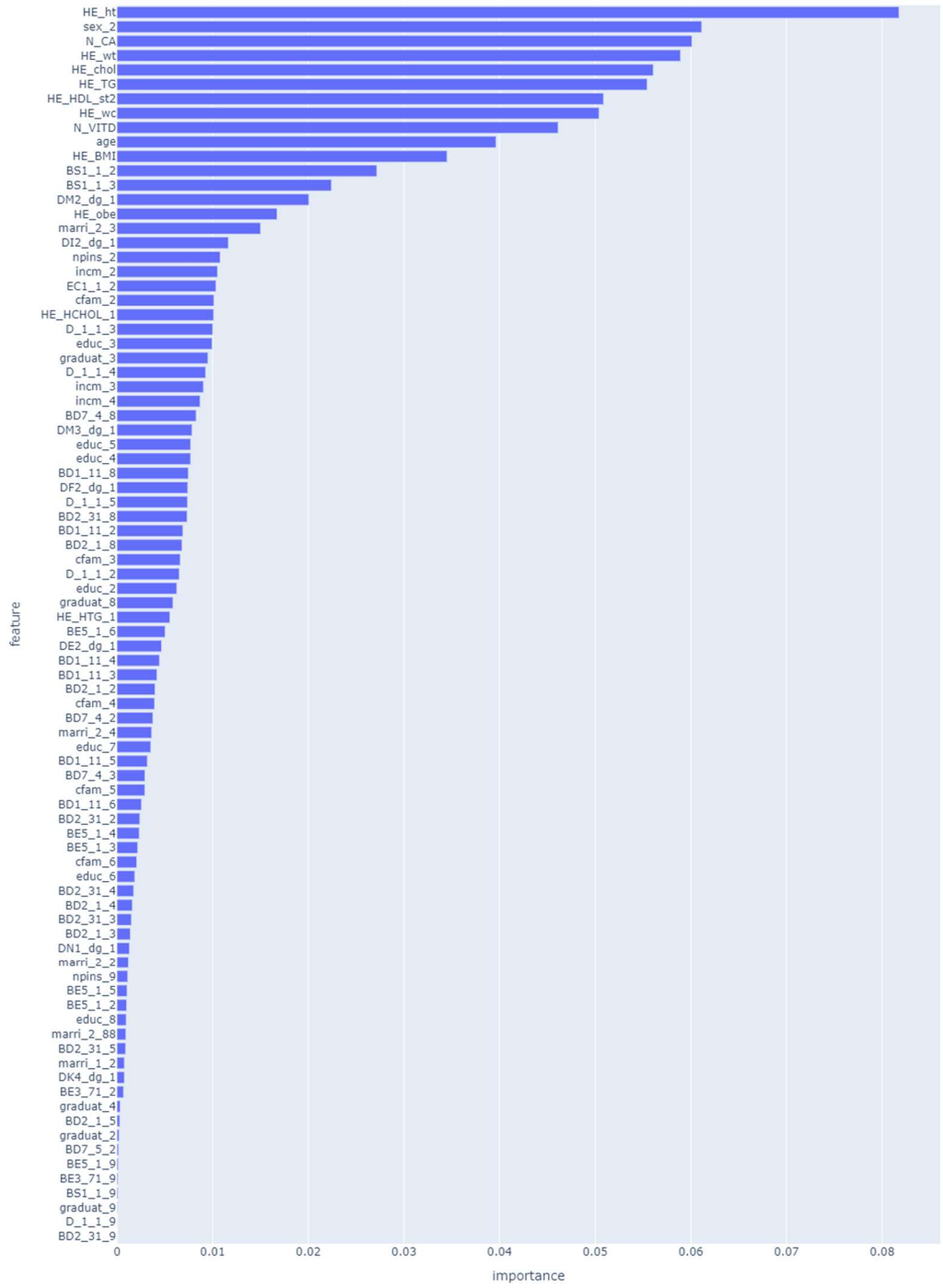
구분	로지스틱 회귀분석	XG 부스트	의사결정나무	랜덤포레스트
AUC	0.7800	0.8059	0.6277	0.8068

각 모형별 ROC곡선을 도출하고, 이에 따른 AUC 값을 확인한 결과, 모두 랜덤포레스트 결과가 가장 좋음을 알 수 있다(<그림 1-8>, <표13> 참고). AUC 값은 랜덤포레스트(0.8068), XG 부스트(0.8059), 로지스틱 회귀분석(0.7800), 의사결정나무(0.6277)순으로 낮아진다(<표13> 참고). 결론적으로 혼동행렬 기반 정확도, 정밀도, 재현율, F1 스코어, ROC 곡선, AUC 값과 같이 대부분의 지표에서 랜덤포레스트가 좋은 결과를 얻었다. 따라서, 65세 이상 남녀노인 대상으로 골다공증 유병률을 예측시에 해당 모델을 적용하는 것으로 최종 채택하였다.

4) 랜덤포레스트 예측 모형 변수별 중요도

랜덤포레스트로 골다공증 유병률을 예측할 때, 어떠한 변인들이 영향을 미치는지 파악해보았다. 보유질병 요소 관련해서는 골관절염과 이상지질혈증 보유가 골다공증 예측에 영향력이 있는 것으로 나타났다. 또한, 인구학적 요소로는 남성, 만나이, 소득 4분위수(개인), 가구원수 2-3명, 결혼여부, 결혼상태가 사별이거나, 민간의료보험 미가입인 경우, 기본건강 요소로는 신장, 체중, 허리둘레, 체질량지수, 총콜레스테롤 수치가 골다공증 예측에 영향력이 있는 것으로 나타났다. 마지막으로, 건강행태적 요소로 평생 5갑이상 흡연 또는 비흡연일 경우와 비타민D 섭취량이 영향력이 높게 나타났다(<표14> 참고).

<그림 9> 랜덤포레스트 예측 모형 변수별 중요도 도표



제 Ⅴ 장 논의 및 결론

제 1절 요약 및 논의

본 연구에서는 2016-2020년 국민건강영양조사를 기반으로 하여, 로지스틱 회귀분석, XG 부스트, 의사결정나무, 랜덤포레스트를 이용한 65세 이상 노인 골다공증 유병률 예측모형 개발하는 것을 연구문제로 잡고 연구를 진행하였다. 연구문제를 해결하기위해 파이썬을 기반으로 한 머신러닝 라이브러리인 scikit-learn에 내장된 linear model인 Logistic Regression, XG boost 모듈에서 XGB Classifier, sklearn ensemble 모듈에서 Decision Classifier, Random Forest Classifier 알고리즘을 적용하여 학습 모델을 만들고 성능지표를 도출하였다. 모델의 평가지표는 이진분류에서 널리 사용되는 정확도, 정밀도, 재현율, F1 스코어, ROC 곡선을 구하여 도출된 AUC 값으로 하였다. scikit-learn metrics 모듈에 내장된 accuracy_score, precision_score, recall_score, f1_score, confusion_matrix를 이용하여 분류모델별 정확도, 정밀도, 재현율, F1 스코어, AUC 값과 혼동행렬을 구하였다.

정확도, 정밀도, 재현율 분류문제에서 분류기준을 어떤 값으로 설정하는지에 따라 정확도, 정밀도, 재현율은 달라진다. 이에 따라, 본 연구에서는 scikit-learn preprocessing 모듈에 Binarizer을 사용하여 분류기준인 임계값을 0.4, 0.45, 0.50, 0.55, 0.60으로 했을 때의 분류모델별 정확도, 정밀도, 재현율, F1 스코어, 혼동행렬의 결과도 비교했다. 재현율과 FP rate의 관계를 나

타내는 ROC 곡선을 도출하고, 모델성능평가를 확인했다. 이때, scikit-learn metrics 모듈에 내장된 ROC curve를 적용하여, ROC 곡선을 구하였다. 또한, matplotlib를 이용하여 시각화를 진행하였다. ROC AUC score를 이용하여, AUC 값도 구하였다.

AUC 값은 랜덤포레스트(0.8068), XG 부스트(0.8059), 로지스틱 회귀분석(0.7800), 의사결정나무(0.6277) 순으로 낮아진다. 또한 이외 다른 지표들에서도 대부분 랜덤포레스트가 가장 높은 수치가 나왔다. 해당 모델 성능 지표들을 통해서, 65세 이상 남녀노인인구의 골다공증 유병률 예측에 가장 적합한 모델은 랜덤포레스트로 선정되었다. 유사한 연구로 영양소 성분 변수들을 기반으로 골다공증을 예측 및 식별하기 위해 생성된 모델들의 성능 비교를 진행한 논문에서의 AUC 값 (0.662)보다 본 연구의 AUC 값이 높았다(유정훈, 이범주, 2020).

랜덤포레스트로 골다공증 유병률을 예측할 때, 보유질병 요소 관련해서는 골관절염과 이상지질혈증 보유가 골다공증 예측에 영향력이 있는 것으로 나타났다. 또한, 인구학적 요소로는 남성, 만나이, 소득 4분위수(개인), 가구원수 2-3명, 결혼여부, 결혼상태가 사별이거나, 민간의료보험 미가입인 경우, 기본건강 요소로는 신장, 체중, 허리둘레, 체질량지수, 총콜레스테롤 수치가 골다공증 예측에 영향력이 있는 것으로 나타났다. 마지막으로, 건강행태적 요소로 평생 5갑이상 흡연 또는 비흡연일 경우와 비타민D 섭취량이 영향력이 높게 나타났다.

기존 골다공증 관련 연구에서는 전통적 방법을 사용하여 유병률에 대한 단

순 기술통계 분석을 진행한 것이 대부분이며, 일부 질병에 대해서는 유병률 예측 모형 연구가 활발히 진행되고 있으나, 골다공증에서는 총 10개 내외의 논문으로 확인될 만큼 그 연구 수준이 미미하다. 또한, 여성 중심의 골다공증 연구가 많이 진행되어, 65세 이상 노인인구를 대상으로 한 골다공증 유병 연구는 극히 드문 것으로 나타났다. 조사한 바에 따르면 남성, 여성 전체에 대한 골다공증 유병률 예측 모형과 관련된 최근 연구가 논문 1건에 그칠 정도로 연구가 활발히 이루어지지 않고 있다는 점을 알 수 있다. 이에, 본 연구는 골다공증 예측모형 개발에 대해서는 두번째 연구가 되겠지만, 65세 이상 노인인구를 대상으로 한 골다공증 예측모형을 개발하는 것에 대해서는 첫번째 연구로써 의의가 있다.

따라서, 이번 연구를 통해, 기존연구와는 연구대상이 차별화된 65세 이상 노인인구를 대상으로 머신러닝을 적용한 골다공증 유병률 예측모형 개발 연구를 진행하는 것에 대해 초문을 열었다. 추후에 골다공증 분야에서도 머신러닝을 적용한 예측모형 개발이 활발히 진행 될 수 있는 단초를 제공하였으며, 이는 학계에 긍정적인 영향을 미칠 것이라고 본다.

제 2절 결론 및 제언

본 연구를 통해 65세 이상 노인인구의 골다공증 유병률 예측 모델을 개발하면서, 이 결과가 의료계와 보건연구에 도움이 될 것으로 보인다. 또한, 골다공증과 관련된 막대한 사회적비용을 절감하는 데에 기여하며, 고령화 사회에서 노인의 건강 관리 및 삶의 질 개선하는 데에 큰 역할을 할 것으로 기대된다.

침묵의 질환으로 불리는 골다공증 질환의 미인지 환자에게 골밀도 검사를 추천하고, 위험요인을 알리는 등 사전 예방 활동이 가능해 환자의 인지율 상승과 사전 관리, 그리고 골다공증 진행이 악화되는 것을 막는 효과가 클 것으로 기대된다.

본 연구를 바탕으로 골다공증 유병률 예측하는 머신러닝 모델을 온라인상에 서비스로 제공하여 노인들의 지속적인 건강관리에 활용함으로써 골다공증과 관련된 위험 요인에 대해 예방할 수 있게 하는 것이 목표이다. 또한, 향후 위험요인에 대한 지속적인 미관리시 일어날 수 있는 추가 합병증에 대해서도 함께 예측가능 하도록 하여 총체적인 노인 건강관리프로그램을 마련하는 데에 초석이 되는 연구로 그 의의가 있다. 그리고, 이러한 연구의 결과는 아직 연구가 이루어지지 않은 다른 질병군들에게 적용하여 예측과 예방이 가능한 모델들을 만드는 데에 기초 자료가 될 것으로 기대된다.

제 3절 연구의 제한점

질병 발생에 관한 연구는 동일 대상에 대해 긴 추적관찰 기간이 필요하고, 이를 통해 질병의 인과관계를 명확히 밝힐 수 있다. 그러나 본 연구에서 사용된 국민건강영양조사는 횡단면연구 자료로써 전후관계를 명확히 할 수 없어, 인과관계 또한 명확치 않다. 즉, 시간의 흐름에 따른 질병 발생의 추이 및 건강위험 요인들 간의 인과관계에 대한 파악이 어렵다는 것을 의미하며, 이것이 본 연구의 한계이다.

또한, 실제 개인 진단 데이터를 기반으로 하는 것이 아니라, 응답자의 답변을 기준으로 진단 유무가 확정되기 때문에 정확성이 떨어질 수 있다. 국민건강영양조사의 경우 건강검진을 실시하는 응답자를 기준으로 하였기에, 전체 노인 인구를 대표하는 데에는 한계가 어느정도 있다. 국내 데이터를 기반으로 만들어진 골다공증 유병률 예측모형이 인종이나 국가에 따라 예측률이 상이한지를 확인하여, 범세계적인 예측모델을 개발할 필요성도 있다고 보여진다. 특히, 다른 질병군에서 인종이나 국가에 따라 동일한 질병 위험 예측 모델을 적용하는 것이 한계가 있는지 확인하는 연구가 되었고 그 결과 예측률이 떨어짐을 확인하였다. 이 부분을 골다공증 예측모델에도 적용해보는 것이 차후에 필요할 것으로 보인다. 또한, 추후 앙상블 기법을 적용하거나, 다른 새로운 모델에 의해 예측 성능이 향상될 수 있는지 확인이 필요할 것으로 보인다.

참고문헌

강성안, 김소희, 류민호 (2022), 머신러닝 기반 생애주기별 고혈압 위험 요인 분석, *한국산업정보학회논문지*, *27*(5), 73-82.

김두섭(2020), 인구 영역의 주요 동향, *한국의 사회동향 2020*, *15*, 24-36.

김승연, 정용주, 김태헌 (2017), 처음 배우는 머신러닝: 기초부터 모델링, 실전 예제, 문제 해결까지, 한빛미디어.

김영란, 남해성, 이태용 (2013), 60세 이상 노년 한국 남성들의 골밀도 수준 및 관련요인, *한국산학기술학회 논문지*, *14*(3), 1180-1190.

김윤아 (2014), 우리나라 50대 이후 성인에서 골다공증과 골감소증 유병률(2008-2011) , *주간건강과질병* , *7*(42), 939-942.

김윤미, 김정환, 조동숙 (2015), 골다공증 유병률, 인지율, 치료율 및 영향요인의 성별 비교: 국민건강영양조사 자료 (2008-2011 년) 활용. *Journal of Korean Academy of Nursing*, *45*(2), 293-305.

김은하(2015), 사회보장정보시스템을 활용한 복지 사각지대 발굴방안 연구, 보건복지부&사회보장정보원

김지영, 양영란(2020), 50세 이상 남성 골다공증 환자의 건강 관련 삶의 질 영향 요인, *Korean Journal of Adult Nursing*, *32*(2), 145-155.

김진, 김수영(2019), 장래인구특별추계: 2017-2067년, 통계청

김한결, 최근호, 임성원, 이현실(2016), 국민건강영양조사를 활용한 대사증후군 유병 예측모형 개발을 위한 융복합 연구: 데이터마이닝을 활용하여, *Journal of Digital Convergence, 14*(2), 325-332.

김혜연 (2022), 인공지능을 이용한 골다공증 예측 및 개인별 위험 요인 분석 모델의 구축, 박사학위논문, 이화여자대학교

계묘진(2013), Lasso를 기반으로 한 로지스틱회귀모형 연구, 석사학위논문, 계명대학교.

권시현(2022), 데싸노트의 실전에서 통하는 머신러닝, 골든래빗.

권세혁, 이정숙 (2020), 2015–2017 년 국민건강영양조사 자료를 활용하여 영양소 섭취와 식이다양성이 중년 이후 성인과 노인의 골다공증에 미치는 영향, *Journal of Nutrition and Health*, *53*(2), 155-174.

박윤진, 강혜경 (2022), 당뇨병성 콩팥병 예측의 기계학습 적용을 위한 분류기 알고리즘별 성능 비교에 대한 연구, *한국산학기술학회 논문지*, *23*(7), 184-191.

박일수(2021), 우리나라 성인 여성의 골다공증 위험도 평가점수 모형 개발, *보건정보통계학회지*, 46(1), 44-53.

박애화, 박형란(2019), 60세 이상 골다공증 여성노인의 골다공증 지식, 운동기대감과 운동자기효능감과의 관계, *재활간호학회지*, *22*(2), 95-103.

변해원(2021), 로지스틱 회귀분석, 랜덤포레스트, XGBoost를 이용한 우리나라 노인의 신체기능장애의 예측 모형 개발, *한국웰니스학회 학술발표회*

서정민(2011), 한국 여성에서 골다공증 인지여부에 따른 건강행위의 차이: 제4기 국민건강영양조사 (2008 & 2009), 석사학위논문, 연세대학교.

성대경, 정경식, 이시우, 백영화 (2021), 최근 10년간 한국인 대상 대사증후군 예측 모델에 대한 체계적 문헌고찰, *한국콘텐츠학회논문지*, *21*(8), 662-674

신민호, 신희영, 정은경 & 이정애 (2002), 60 세 이상 노인여성에서 골다공증 유병률과 위험요인, *노인병*, *6*(2), 130-139.

오미애, 최현수, 김수현, 장준혁, 진재현, 천미경 (2017), 기계학습(Machine learning)기반 사회보장 빅데이터 분석 및 예측모형연구, *한국보건사회연구원.*

유인영(2011), 폐경 후 성인 여성의 연령에 따른 골다공증 관련 요인: 국민건강영양조사 제4기 2차년도 (2008), 3차년도 (2009) ,석사학위논문, 연세대학교

유정은(2018), 한국 성인 남성에서 골다공증의 유병률 및 관련된 위험 요인, 석사학위논문, 울산대학교

유정훈, 이범주(2020), 연관성 규칙 기반 영양소를 이용한 골다공증 예측 모델, *The Journal of the Convergence on Culture Technology (JCCT)*, *6*(3), 457-462.

윤우진, 서동호, 민세웅, 남해운 (2021), RandomForest 와 XGBoost 를 활용한 유방암 종양분류, *2021 년도한국통신학회 동계종합학술발표회.*

이근영(2015), 머신러닝을 활용한 스마트 서비스와 금융, *전자 금융과 금융보안*, *1*(창간호), 31-66.

이은환, 김욱, 남진영(2019), 우리나라 골다공증으로 인한 사회경제적 질병비용 측정, *경기연구원 기본연구*, *2019*(03),1-84.

이인자, 이준호 (2020), 폐경 여성에서 트리기반 머신러닝 모델로부터 골다공증 예측, *방사선기술과학*, *43*(6), 495-502.

이혜상(2016), “우리나라 50세 이상 남성의 골감소증・골다공증유병률과 관련

요인: 2010～2011 국민건강영양조사자료", *대한영양사협회학술지*, *22*(2), 106-117.

임지선(2012), 조기폐경과 골다공증의 관련성: 국민건강영양조사 제 4 기 2 차년도 (2008), 3 차년도 (2009) 자료를 활용하여, 석사학위논문, 연세대학교

정보미, 김재훈, 허태영 (2020), 치매 발병 여부 예측을 위한 통계적 모형 및 기계학습 기반 기법 적용에 관한 연구, *Journal of The Korean Data Analysis Society (JKDAS), 22*(5), 1819-1834.

정윤석(2010), 골다공증의 치료, *대한내과학회지*, 79(3), 250-253.

정호연(2008), 골다공증 진단 및 치료 지침 2007, *대한내분비학회지*, *23*(2), 76-108.

한은정, 송기준, 김동건 (2009), Random forests 기법을 이용한 백내장 예측모형: 일개 대학병원 건강검진 수검자료에서, *한국통계학회 응용통계연구*, *22*(4), 771-780

아카바 신야, 스기야마 아세이, 데라다 마나부(2019), 그림으로 공부하는 머신러닝 알고리즘 17 머신러닝 도감, 주식회사 제이펍.

아리가 미치아키, 나카야마 신타, 니시바야시 다카시(2018), 머신러닝 실무프로젝트: 실전에 필요한 머신러닝 시스템 설계, 데이터 수집, 효과 검증 노하우, 한빛미디어.

Breiman, L. (2001), Random forests., *Machine Learning*, *45*(1), 5-32.

Chen, T. and Guestrin, C.(2016), Xgboost: A scalable tree boosting system, *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and*

data mining .

Cox, D.R. (1958), The regression analysis of binary sequences., *Journal of the Royal Statistical Society: Series B*, *20*, 215-242.

Jiang, H.X., Majumdar, S.R., Dick, D.A., Moreau, M., Raso, J., Otto, D.D. and Johnston, D.W.C. (2005), Development and initial validation of a risk score for predicting in-hospital and 1-year mortality in patients with hip fractures., *Journal of Bone and Mineral Research*, *20*(3), 494-500.

Marcy B. Bolster. (2021), Osteoporosis- Bone, Joint, and Muscle Disorders - MSD Manual consumer version, https://www.msdmanuals.com/ko/%ED%99%88/%EB%BC%88,-%EA%B4%80%EC%A0%88,-%EA%B7%BC%EC%9C%A1-%EC%9E%A5%EC%95%A0/%EA%B3%A8%EB%8B%A4%EA%B3%B5%EC%A6%9D/%EA%B3%A8%EB%8B%A4%EA%B3%B5%EC%A6%9D, Accessed at Dec. 01, 2022

NIH Consensus Development Panel (2001), Osteoporosis prevention, diagnosis, and therapy, *Journal of the American Medical Association*, *285*(6), 785–795

Park, E.J., Joo, I. W., Jang, M. J., Kim, Y. T., Oh, K. and Oh, H. J. (2014), Prevalence of osteoporosis in the Korean population based on Korea National Health and Nutrition Examination Survey (KNHANES), 2008-2011., *Yonsei Medical Journal*, *55*(4), 1049-1057. doi.org/10.3349/ymj.2014.55.4.1049

Quinlan, J. R. (1986), Induction of decision trees., *Machine Learning*, *1*, 81-106.

Sharma, D. K., Chatterjee, M., Kaur, G. and Vavilala, S. (2022), Deep learning applications for disease diagnosis., *Deep Learning for Medical Applications with Unique Data ,1*, 31-51

Song ,Y.Y. and Lu, Y.(2015), Decision tree methods: applications for classification and

prediction., *Shanghai Arch Psychiatry*, *27*(2), 130-135.

Vexler. Albert, S.Liu. and E.F.Schisterman. (2011), Nonparametric-likelihood inference based on cost-effectively- sampled-data, *J. Appl.Stat*, *38* (4), 769-783.

WHO Study Group. (1994), ASSESSMENT OF FRACTUE RISK AND ITS APPLICATION TO SCREENING FOR POSTMENOPAUSAL OSTEOPOROSIS , *WHO technical Report Serises*, *843*, 1-136.

Abstract

Development of osteoporosis prevalence prediction model for the elderly aged 65 years or above using Logistic regression analysis, XG boost, Decision tree, and Random forest

- based on the Korea National Health and Nutrition Examination Survey 2016-2020 -

Son, Da IN

Seoul School of Integrated Sciences and Technologies

Advisor: Chang, Jung Ho

Osteoporosis is a representative disease with a higher incidence rate in the elderly. In Korea where it is entering an aging society, it is meaningful to conduct research related to osteoporosis, which entails high proportion of social costs. In previous studies, papers analyzing the factors of osteoporosis in women using traditional statistical techniques in social science research are mainly focused. Recently, papers have been published for postmenopausal women integrating big data analysis on the development of osteoporosis prevention behavior models or the establishment of osteoporosis prediction and individual risk factor analysis models, but the proportion is insignificant as there are only 2-3 studies. In addition, osteoporosis is recognized as a disease that occurs only in postmenopausal

women, and research on men is relatively insufficient. However, studies related to male osteoporosis showed that osteoporosis can also occur in men, having a higher mortality rate than women, and can be a fatal disease due to low recognition rate and treatment rate. Against these backgrounds, need for a big data analysis study on the prevalence of osteoporosis in all elderly people aged 65 or older was raised, and this study developed a model to predict the prevalence of osteoporosis in men and women aged 65 or older based on the Korea National Health and Nutrition Examination Survey data.

A total of 372 variables were selected from 2016-2020 National Health and Nutrition Examination Survey data that has been annually conducted in common, and among them, variables commonly pointed out as factors related to osteoporosis were selected through literature surveys from previous studies and academic journals. The dependent variable was whether osteoporosis was diagnosed, and 46 variables were again selected in the order of high correlation coefficients in the relationship with the dependent variable. Among them, 8 variables with high missing values were removed, and a total of 36 were selected as independent variables. Age and gender variables were also added to this, and total of 38 independent variables were consisted. Among them, the decimal type was changed to an integer type, and data analysis was conducted on 5,365 people by removing missing values from 8,170 people's data during data preprocessing.

Machine learning classification (Logistic regression analysis, XG boost, Decision tree, and Random forest) algorithms were applied to these variables, and data of 5,365 people aged 65 or older were divided into training data and test data to check and compare predictive performance. Of the total data, 80% was selected as training data and 20% as test data. Target variable was defined as the prevalence of osteoporosis for dependent variable y, and for independent variable x, osteoporosis prevalence factor variables were defined as

variables after pre-treatment, removal of missing values, and creation of dummy variables from 38 variables finally selected as osteoporosis-related variables. The evaluation indicators were determined as accuracy, precision, recall, F1 score, and AUC values derived from ROC curves widely used in binary classification. Most random forest results were good in accuracy, precision, reproduction rate, and F1 scores calculated based on the confusion matrix when the classification criteria were changed to 0.4, 0.45, 0.5, and 0.6. In particular, AUC values are lowered in the order of random forest (0.8068), XG boost (0.8059), logistic regression analysis (0.7800), and decision tree (0.6277). Therefore, it was finally adopted to apply the Random forest model when predicting the prevalence of osteoporosis for men and women aged 65 or older. In addition, when predicting the prevalence of osteoporosis with random forest, variables affecting osteoporosis were identified. Regarding the factors of the disease, it was found that osteoarthritis and dyslipidemia retention had an influence on predicting the prevalence of osteoporosis. In addition, following were found to be influential in predicting the prevalence of osteoporosis as demographic factors; height, weight, waist circumference, body mass index, and total cholesterol levels in terms of basic health factor, sex (male), age, income quartile (individual), number of household members, marital status, bereavement, or non-private health insurance. Lastly as a health behavior factor, smoking or non-smoking for more than 5 packs during lifetime and vitamin D intake were highly influential.

Through this study, it is expected to help the medical community and public health research from developing model for predicting the prevalence of osteoporosis in the elderly population aged 65 or older. In addition, it is expected to contribute to reducing the enormous social costs associated with osteoporosis, and play a major role in improving the health care and quality of life of the elderly in an aging society. To unrecognized patients

of osteoporosis disease, a silent disease, precautionary activities such as recommending bone density tests are expected to have a great effect on increasing recognition, pre-management, and reduction osteoporosis progression. Through this study, it opened the first study on the development of a model for predicting the prevalence of osteoporosis using machine learning for the elderly population aged 65 or older, which is differentiated from previous studies. In the future, it provided a starting point for the development of a predictive model applying machine learning in the field of osteoporosis, which will have a positive impact on academia.

Key words: Logistic regression analysis, XG boost, Decision tree, Random forest, accuracy, precision, recall, F1 score, AUC values, ROC curves

Student Number: 2125418004

감사의 글

우선, 입학시점부터 학업의 방향을 제시해주시고, 논문을 완성할 수 있게 용기와 힘을 주시며 저를 지도해주신 장중호 교수님께 감사드립니다. 또한 바쁘신 와중에서도 논문의 완성도를 높일 수 있도록 조언해주신 오태연 교수님, 최진희 교수님께 감사드립니다. 업무와 학업을 병행하면서 둘 다 잘 해낼 수 있도록 많은 도움을 준 동료 백인겸 선생님께 감사의 말을 전합니다. 대학원 공부를 응원해주신 안성모 실장님, 인생의 방향을 이끌어 주시고, 비전을 제시해주신 옥치국 대표님께도 감사드립니다. 학업의 기회를 만들어준 서울과학종합대학원대학교, 대학원 과정에 지원하고 공부를 시작하면서 격려해주신 최용주 부총장님, 새로운 것에 항상 도전하고 끊임없이 학습 할 수 있도록 원동력을 주신 김태현 총장님과 삶에 대한 열정과 열심히 해내는 것의 중요성을 깨닫게 해주신 김경성 원장님, 현상을 바라보는 새로운 시각과 이론을 발견할 수 있도록 많은 가르침을 주신 조동성 교수님께도 감사드립니다. 항상 모든 선택을 응원하고 지지해주시는 아버지 ‘손덕익’, 어머니 ‘박윤우’ 정말 많이 사랑합니다. 세상에 하나뿐인 내편들, 언니 ‘손수인’, 여동생 ‘손지호’, 남동생 ‘손호민’에게 고맙다는 말 전합니다. 항상 날 웃게 하고 힘을 주는 지인들 ‘J.H’‘J.M’‘J.Y’‘S.H’‘Y.J’‘K.H’‘H.N’‘S.J’‘Y.B’‘H.S’‘T.H’‘G.U’ 고맙습니다. 제 삶의 모토인 ‘심청사달’을 항상 가슴속에 새기며 사회에 선한 영향력을 줄 수 있는 사람이 될 수 있도록 끊임없이 배우고 성장해 나가겠습니다. 무한한 응원과 사랑을 주는 모든 주변 분들, 그리고 우리 빅데이터 동기들에게 감사드립니다.