경영전문석사학위 논문

Electronic Health Records을 이용한 치매 예측 및 주요 인자 조사

2019년 8월

서울과학종합대학원대학교

이 영 희

경영전문석사학위 논문

Electronic Health Records을 이용한 치매 예측 및 주요 인자 조사

2019년 8월

서울과학종합대학원대학교

이 영 희

Electronic Health Records을 이용한 치매 예측 및 주요 인자 조사

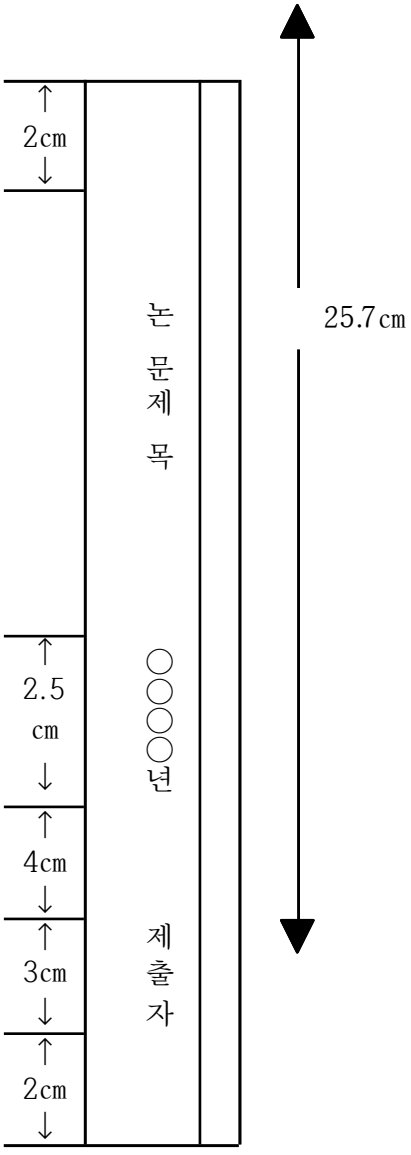
지도교수 김 진 호

이 논문을 경영학 석사 학위논문으로 제출함
2019년 7월
서울과학종합대학원대학교
이 영 희

이영희의 석사 학위논문을 인준함
2019년 7월

위 원 장 고 영 희 (인)

위 원 김 진 호 (인)


2cm
논 문 제 목
25.7cm
2.5 cm
○○○○년
4cm
제 출 자
3cm
2cm

학위논문 원문제공 서비스에 대한 동의서

본인의 학위논문에 대하여 서울과학종합대학원대학교가 아래와 같이 학위논문 저작물을 제공하는 것에 동의합니다.

1. 동의사항

①본인의 논문을 디지털화하여 국회도서관, 국립중앙도서관 등 인터넷 정보통신망을 통한 논문의 일부 또는 전부의 복제.배포.출력 및 전송 시 무료로 제공하는 것에 동의합니다.

②본인의 논문을 보존이나 인터넷 등을 통한 온라인 서비스 목적으로 이용할 경우 저작물의 내용을 변경하지 않는 범위 내에서의 편집·형식상의 변경을 허용합니다.

2. 개인(저작자)의 의무

본 논문의 저작권을 타인에게 양도하거나 또는 출판을 허락하는 등 동의 내용을 변경하고자 할 때는 소속대학(원)에 공개의 유보 또는 해지를 즉시 통보하겠습니다.

3. 서울과학종합대학원대학교의 의무

①서울과학종합대학원대학교는 본 논문을 외부에 제공할 경우 저작권 보호장치(DRM)를 사용하여야 합니다.

②서울과학종합대학원대학교는 본 논문에 대한 공개의 유보나 해지 신청 시 즉시 처리해야 합니다.

논문제목 : Electronic Health Records를 이용한 치매 예측 및 주요 인자 조사

학위구분 : **석사 ▣.박사 □**

학 번 : 1854061015

연 락 처 : 010-4425-0186

저 작 자 : 이 영 희 (인)

제 출 일 : 2019년 8월 일

서울과학종합대학원대학교 총장 귀하

초 록

인공지능은 인간이 가진 인지, 학습, 추론 등의 지적능력을 컴퓨터로 구현하는 기술로 기계학습(머신러닝), 자연어 처리, 음성인식, 영상인식 등의 방향으로 발전되고 있으며, 의료, 금융, 제조 등 다양한 분야로 활용이 급속히 확산되고 있다. 특히 의료계에서는 이러한 기술이 질병의 예측, 진단 및 처방에 적용되어 의료의 질이 개선되기를 기대한다.

대한민국은 세계에서 유례를 찾기 힘들 정도로 고령화가 빠른 속도로 진행되고 있다. 이에 따라 노화로 인한 질병이 점차 심각해지고, 이러한 질병 중 하나인 치매의 높은 발병율과 이로 인한 의료비의 부담이 가중됨에 따라 사회적 문제로 발전하였다. 여러 연구조사에 따르면 치매의 향후 발생률에 대한 정확한 예측은 질병 발병을 지연시키기 위한 조기 개입을 용이하게 하는데 기여할 것으로 본다.

따라서 본 연구에는 전자건강기록(EHR) 데이터를 이용하여 기계학습을 통한 치매 예측 모델의 가능성을 검토하고 예측에 영향을 주는 주요 요인을 분석하는 것을 목적으로 한다. 사용된 데이터는 국민건강보험공단의 '국민건강정보 DB'의 표본 코호트 DB를 사용한다. 표본 코호트 DB의 2002년부터 2013년까지 100만 명의 한국인 데이터 중 2009년부터 2013년의 기간 동안 저장된 데이터를 사용한다.

기존의 치매나 알츠하이머 관련 연구는 65세 이상 노인을 대상으로 하지만 본 연구에서는 치매 발병의 연령이 낮아지는 추세를 반영하여 대상 연령을 기존보다 낮추어 50세 이상을 대상자로 선정하여 데이터 분석을 진행한다. 표본 코호트 DB 중 자격 DB, 건강 검진 DB, 진료명세 DB에서 기본적인 건강검진 정보, 개인 및 가족 병력, 사회 인구 통계 등 50여 개의 고유한 임상 특징을 추출하여 사용한다. 이 중에서 치매에 영향을 미치는 주요 변수들의 5년간의 수치를 각각 속성으로 추가하여 Random Forest, XGBoost, SVC 등을 포함한 알고리즘을 이용하여 학습 및 예측을 진행한다. 본 연구에서는 제안된 분석방법과 주요 인자들을 이용하여 치

매에 대한 의미 있는 예측 결과를 확인하고 주요 변수들의 영향성을 확인한다.

목 차

표 목 차

그 림 목차

제 1장 서 론

제 1절 연구 배경

4차 산업혁명의 핵심기술인 인공지능(Artificial Intelligence)이 주목을 받으며 다양한 산업영역에서 이에 대한 연구 및 적용이 활발하게 진행되고 있다. 그 중 의료 산업에서는 인공지능을 활용하여 영상, 음성, 텍스트 등 다양한 형식의 의료 데이터를 분석하여 질병 예측, 진단 및 처방을 통해 의료의 질을 높이는 것에 큰 관심을 갖고 있다(의학신문, 2018). 특히, ICT 융합 의료기기의 증가로 인해 대규모 의료용 빅데이터의 확보가 용이해졌고, 이를 이용한 AI 기반 비즈니스가 점차 확산되고 있는 추세이다(정원준, 2018).

특히 전자건강기록(EHR)과 전자의료기록(EMR) 데이터베이스를 이용하여 질병 예측, 진단 및 예후에 대한 연구가 활발하게 진행되고 있으며 많은 경우는 EHR 데이터베이스를 기본으로 MRI, CT, 유전자 정보 등의 다양한 정보를 결합하여 진행되고 있다. 기존 EHR을 이용한 치매(알츠하이머) 질병 예측의 경우도 기본적인 건강검진 데이터와 인지기능 선별검사 항목을 포함하여 치매(알츠하이머) 예측을 진행한다. 인지기능 선별검사 중에서 전 세계적으로 널리 사용되고 있는 검사가 간이정신상태검사(MMSE)[1]이다. 많은 선행연구들에서 MMSE의 신뢰도는 확인되었고(오은아 외 3, 2010), 이 검사의 결과점수가 치매 예측에 높은 영향을 주는 변수이지만, 국내에서는 일반적으로 만 60세 이상에 한해서 시행되는 검사이므로 대상이 제한적이다.

따라서 본 연구에서는 보다 더 범용적인 대상으로 치매 예측 가능성을 연구하기 위하여 전자건강기록(EHR) 데이터의 속성을 이용하여 기계학습

1) MMSE(Mini-Mental State Examination)는 간이정신상태평가로 인지기능의 손상을 선별하고 측정하는 검사이다. 30점으로 구성

을 통한 치매 예측이 가능한지를 연구하기 위해 수행되었다.

제 2절 연구 목적

본 연구에서는 보다 더 범용적인 대상으로 치매 예측 가능성을 연구하기 위하여 MMSE 항목을 포함하지 않고 전자건강기록(EHR) 데이터 중 건강검진 데이터, 문진 데이터와 진료 명세 데이터 중 주상병 속성을 이용하여 기계학습을 통한 치매 예측 모델의 가능성을 검토하고 예측에 영향을 주는 주요 요인을 분석한다. 기존의 치매나 알츠하이머 관련 연구는 65세 이상 노인을 대상으로 하지만 본 연구에서는 연령을 낮추어 50세 이상을 대상자로 선정하여 분석을 진행한다. 치매 예측 관련한 이전 연구(김형섭 외 3, 2018)에서 밝혀진 치매의 위험인자를 주요 변수로 추가하여 치매여부를 예측하고, 예측율을 높이기 위한 추가 요인에 대한 분석을 진행한다. 이에 따른 기대효과로는 EHR 데이터를 이용하여 치매 예측을 진행하여 그 결과가 유의미한 범위 안에 있으면 해당 대상자가 간이정신상태검사(MMSE), CDR 검사 등을 통해 정확한 진단을 받도록 제안을 할 수 있다.

제 3절 연구 절차

본 연구의 2장에서는 의료산업의 기술 현황을 정리하고, 전자건강기록(EHR), 전자의료기록(EMR)을 이용한 치매와 관련된 선행연구 및 치매에 영향을 미치는 주요 인자 관련한 선행연구를 고찰한다.

3장에서는 먼저 치매에 대한 정의를 하고, 연구 대상 및 자료분석 방법에 대하여 설명한다.

4장에서는 전자건강기록(EHR) 데이터를 이용하한 기계학습 모델을 통한 치매 예측 결과를 보여주고, 이에 영향을 주는 주요 인자 분석 결과를

보여준다.

5장에서는 연구에 대한 결론 및 향후 방향성을 설명한다.

제 2장 이론적 배경 및 선행연구의 고찰

제 1절 의료산업의 기술 현황

우리나라는 IT 인프라가 잘 구축되어 있고, 전 국민 대상 의료보험 체계를 근간으로 의료정보 빅데이터 활용 등에 높은 잠재력을 보유하고 있다. 의학영상정보시스템(PACS: Picture Archiving and Communication System)과 전자의무기록(EMR: Electronic Medical Record)의 보급률이 세계 1위로 알려져 있다(박순영, 2017). 2018년 11월 '보건의료 빅데이터 플랫폼'구축을 위한 시범사업이 승인되어 보건의료 빅데이터와 IT 헬스사업의 접목이 속도를 내고 있다.

또한 전자의무기록(EMR)은 병원 내부의 범위를 넘어서 의료기관간에 진료 정보를 교환하고 이를 통해서 공동 활용할 수 있는 EHR(Electronic Health Record)의 단계로 발전하고 있고, 궁극적으로는 여러 의료기관에서 저장 관리되고 있는 개개인의 진료 기록뿐만 아니라 흡연, 운동, 식이습관 등 개인의 건강기록을 포함한 모든 데이터들을 관리 대상으로 하는 PHR(Personal Health Record) 단계로 발전으로 나아가고 있다(전진옥, 2018).

이런 기술적, 정책적 방향에 힘입어 EHR 시스템에는 지속적으로 정보가 급격하게 증가하고 있다. 인구 통계, 병력, 연구실 결과, 절차 및 약물과 같은 구조화 된 환자 정보, 진료 메모 및 퇴원 메모와 같은 비정형 정보는 각 임상 과정 동안 수집된다. 기계 학습 방법은 방대한 양의 데이터에서 정보를 추출하고 새로운 사례로 일반화하는 데 적합하고(Jingshu Liu 외 2, 2018), 이를 통한 진료의 질의 향상시킬 수 있는 기회를 제공한다.

전자건강기록(EHR)을 활용한 질병의 예측, 진단 및 처방 관련한 연구가 기업과 의료기관이 연계하여 활발하게 진행되고 있다. 구글은 전자건강기록(EHR)에 저장된 환자의 진료기록을 딥러닝으로 분석하여 입원한

환자의 치료결과를 정확히 예측하는 인공지능을 2018년 1월에 발표했다. 이 딥러닝을 이용하면 환자가 입원 중에 사망할 것인지, 장기간 입원할 것인지, 혹은 퇴원 후 30일 내에 재입원할 것인지, 그리고 퇴원 시의 진단명은 어떻게 될 것인지도 높은 정확도로 조기에 예측할 수 있다. (최윤섭, 2018). 이와 같이 의료계에서 활용되는 딥러닝 기술은 질병의 발병 예측 뿐 만 아니라 진행정도 및 예후까지 예측할 수 있는 단계까지 발전하였다.

제 2절 치매 예측 관련 선행 연구

2018년에 발표된 일산병원 보고서 '치매 특별 등급(장기요양 5등급) 자료 분석을 통한 치매 예측 모델 개발 및 조기 개입 효과 조사'에 따르면 "알츠하이머 치매 환자들과 정상군을 비교하였을 때, 고혈압, 당뇨, 고지혈증의 정도, 신부전 여부, 허혈성심질환, 운동력, 체질량지수에서 유의한 차이를 보이는 것으로 조사하였다. 이 중 고혈압, 운동력, 체질량지수는 오즈비가 1보다 작음을 확인할 수 있었다. 혈관 치매 환자들과 정상군을 비교하였을 때 고혈압, 당뇨, 고지혈증의 정도, 신부전 여부, 심장부정맥, 심혈관질환, 운동력, 체질량지수에서 유의한 차이가 있었으며, 운동력에서만 1미만 오즈비 값을 확인할 수 있었다. 즉 알츠하이머 및 혈관 치매 모두 내과적 기저질환이 심할수록 치매 발생과 깊은 관계가 있다."고 하였다(김형섭 외 3, 2018). 위의 연구는 기계학습이나 인공지능이 아닌 통계학적 방법으로 연구가 진행되었지만 치매와 주요질환과의 상관관계가 분석되었고, 이를 기계학습 방법을 통해서 분석하려고 한다.

특정 건강 관련 행태는 치매 위험 발생과 일관된 연관성을 지닌다는 연구결과가 있다. 신체활동, 흡연, 음주, 영양불균형, 정신적 요인 등이 치매 위험에 영향을 줄 수 있다(ADI, 2014). 다만 흡연, 음주 등의 요소는 건강검진의 문진데이터에 포함되어 있으나 정상적으로 입력되어 있는 건수가 많지 않아서 영향성 분석은 어려워 보인다.

고혈압, 비만, 콜레스테롤(이상지질혈증), 당뇨병, 흡연은 심뇌혈관 질환을 발생시키는 선행요인이다. 뇌졸중의 병력은 고령 인구에서 치매의 위험을 두 배로 증가시킨다는 연구 결과도 있다(Savva and Stephan, 2010). 중년기의 고혈압은 후기 생애에서 치매의 발생 위험을 증가시키고, 총콜레스테롤은 치매와 관련성이 있고, 당뇨병과 치매의 연관성은 다소 간접적이지만 당뇨병의 지속 기간은 연관이 있다. 또한 우울증은 치매 위험을 증가시킬 수 있다고 제시한다(ADI, 2014).

주요질환과 치매와의 연관성에 대해서 많은 선행연구에서 확인이 되었다. 다만 치매의 경우 알츠하이머성 치매, 혈관성 치매, 다른 질환에서 유래된 치매 및 원인불명의 치매 등으로 다양하고 연구마다 분류, 시점, 세부 항목이 조금씩 다르기는 하지만 일관된 연관성을 주는 요소는 확인된다. 즉, 고혈압, 당뇨, 고지혈증, 우울증이 치매와 연관성이 있다고 공통적으로 제시한다.

따라서 본 연구에서는 기계학습을 이용하여 치매를 예측하고 선행연구들에서 공통적으로 연관성이 제시된 질환들을 변수로 포함하여 학습 및 예측을 진행하여 그 결과의 비교를 통해 영향성을 검토한다.

제 3절 치매 조기 발견의 중요성

예방 가능한 질병의 조기 발견은 질병 관리 개선, 개선 된 개입 및 보다 효율적인 의료 자원 배분에 중요하다. 이 작업을 위해 전자 건강 기록(EHR)의 정보를 활용하기 위해 다양한 기계 학습 접근법이 개발되었다(Jingshu Liu 외 2, 2018).

치매의 진단 및 치료와 관련하여 전반적인 단계에서, 치매 발생 이전에 전체 인구집단을 대상으로 치매예방 사업을 수행하고 치매의 조기 발견을 위해 치매 의심환자를 선별하여 진단하기 위한 다양한 방법에 대한 연구가 진행되고 있다. 현재 지역사회와 임상에서 사용되고 있는 인지기능 선별검사 중에서 전 세계적으로 가장 널리 사용되고 있는 검사가

Mini-Mental Status Examination (MMSE)이다(오은아 외 3, 2010). 보건복지부, 질병관리본부 보도자료에 따르면 “대부분의 치매 임상진단은 신경심리검사도구 (MMSE, CDR, KGS 등)에 의존하고 있으며, 확진에는 뇌영상진단(MRI, amyloid-PET 등)이 이용되고는 있으나 뇌위축(brain atropy)이 상당히 진행된 상태에서 진단이 가능하여 치매 조기발견은 어려운 실정이나 치매 고위험군의 조기발견을 통해 치매의 발병을 2년 정도 지연시킬 경우 20년 후 치매 유병률이 80% 수준으로 감소할 것으로 예상한다”고 한다(질병관리본부 외 1, 2014).

선행연구 및 국가정책에서도 인구집단을 대상으로 치매환자를 조기에 검진하는 것이 매우 중요하다는 것이 확인된다. 또한 치매 발병이 65세 이하에서도 늘어나고 있으므로 이에 대한 스크리닝 방안도 필요하다. 앞서 언급한 것처럼 치매를 스크리닝하기 위한 다양한 선별검사 및 진단도구가 있다(고숙자 외 2, 2016). 다만, 만 60세 이상에 한해서 시행되는 검사이므로 연령에서 제한적이라는 단점이 있다.

치매의 조기 발견 관련한 다양한 선별검사 및 진단도구가 있고, 기계학습을 포함한 인공지능을 이용하여 조기 예측하려는 연구 및 시도가 지속적으로 진행되고 있다. 이런 측면에서 건강검진데이터는 접근성도 좋고 2년마다 정기적으로 시행하기 때문에 데이터의 최신성 및 대상에 제한도 상대적으로 적다. 따라서 전자건강기록(EHR)을 이용하여 치매 예측 모델에서 의미 있는 결과가 나온다면 향후 이를 이용한 조기 치매 예측도 가능하리라 본다.

제 3장 연구 방법

제 1절 연구 대상

(1) 치매 환자의 정의

치매는 성장기에는 정상적인 지적 수준을 유지하다가 후천적인 원인으로 인지기능의 손상되거나 파괴되고 인격의 변화가 발생하는 질환이다.(위키백과)(고숙자 외 2, 2016)

이 연구에서는 진료 명세서 DB의 주상병 변수의 상병코드를 기준으로 치매 환자를 정의한다. 상병코드는 한국표준질병사인분류표(KCD 코드[2])의 상병분류기호에 근거하고, 치매질환에 해당하는 상병코드는 'F00', 'F00.0', 'F00.1', 'F00.2', 'F00.9', 'F01', 'F010', 'F011', 'F012', 'F013', 'F018', 'F019', 'F02', 'F020', 'F021', 'F023', ''F028', 'F03', 'F030'이다. 치매로 분류된 상병코드에 대한 상세한 내용은 <표 1>상병코드에 따른 치매 질환의 분류와 같다.

(2) 연구 대상

본 연구에서는 국민건강보험공단(National Health Insurance Service, NHIS)의 '국민건강정보 DB'를 대상으로 표본 추출한 표본 코호트 DB를 사용한다. 이 데이터베이스에는 2002년부터 2013년까지 대표성을 띄는 한국인 100만 명에 대한 모든 건강관리와 관련된 서비스, 진단 및 처방의 개별적인 특성이 포함되어 있다. 임상 특징에는 Participant Insurance Eligibility 데이터베이스의 인구 통계 및 사회 경제 변수가 포함된다. 건강관리 이용 데이터베이스의 질병 및 약물코드(40세 이상의 성인들에게 요구되는 2년마다의 건강 검진)를 통해 국립 건강 검진 데이터베이스에서 개인 및 가족의 병력에 대한 정보를 취합 및 제공한다. 국민건강정보 DB

2) KCD 코드 : ICD-10(KCD-6). 국제질병분류, 현재 KCD(ICD-10 2009년 ~ 2014 업데이트) 사용 중

는 대한민국 2002년부터 2010년까지의 건강 및 보험청구 데이터가 있는 65세 이상 노인에 대한 무작위로 추출된 10% 표본이 포함되어 있다(국민건강보험공단, 2015). 이 연구에서는 건강검진기록을 이용하여 치매 예측 및 주요 변수의 영향을 연구하는 것을 목적으로 하므로 전체 데이터베이스 중 건강검진 데이터베이스를 기준으로 대상자를 선정한다. 건강검진 DB의 경우 2009년을 기준으로 건강검진에 포함되는 항목이 바뀌었다. 이 중 예측에 영향을 주는 변수 중 일부가 2009년 이전에는 포함되어 있지 않아서 대상연도를 2009년부터 2013년으로 제한하여 진행한다. 이 기간의 건강검진 대상자는 502,812명이다. 이 대상에는 해당 기간에 사망한 대상자는 제외된다. 기존 치매나 알츠하이머 관련 연구는 65세 이상 노인을 대상으로 하지만 이 연구의 경우 좀 더 연령을 낮추어 50세 이상을 대상자로 선정하였고, 기본 대상 중 선정된 최종 연구 대상자는 253,742명이다. 이 대상자 중 주상병을 기준으로 치매로 분류된 대상자는 6,316명이고, 이는 전체 대상자의 약2.5%를 차지한다.

(3) 케이스 컨트롤

앞에서 언급한 50세 이상의 전체 데이터를 가지고 학습 및 예측을 진행하는 경우 일반인 대상자의 특성이 강하게 학습되어 제대로 예측이 안되는 문제가 발생한다, 따라서 이를 방지하기 위해서 기계학습을 위한 대상자는 기본적으로 전체 데이터에서 치매로 레이블 된 대상자를 선정하고, 이와 같은 숫자의 일반 대상자를 선정하여 추출 후 이를 학습 표본으로 하여 모델링을 진행한다. 표본 추출 시에 치매 대상자의 나이그룹을 기준으로 동일한 숫자에 해당하는 일반인 숫자를 추출한다. 추가로 Sampling Unit은 1배수, 2배수 추출하여 진행하고 이의 결과를 비교한다. 치매예측에 영향을 주는 요소들을 확인하기 위하여, 남성, 여성, 건강검진 회수 등 다양한 특성으로 샘플을 추출하여 비교를 진행한다.

제 2절 변수 정의

이 연구에서는 앞서서 언급한 세 가지 DB의 다음 변수들을 사용한다. 자격 DB에서 연령과 성별 변수, 건강검진 DB의 기본 정보(신장, 체중, 허리둘레, 수축기 혈압, 이완기혈압, 식전혈당, 총콜레스테롤, 트리글리세라이드, HDL 콜레스테롤, LDL 콜레스테롤, 혈색소, 요단백, 혈청 크레아티닌, 혈청지오티, 혈청지피티, 감마지티피 등), 개인 및 가족 병력 이력을 사용한다. 신장과 체중 변수를 이용하여 체질량 지수를 구하고 이를 속성에 추가한다.

추가로 치매에 영향을 주는 질병으로 알려진 우울증, 당뇨, 고혈압, 고지혈증과의 연관성을 분석하기 위해서 이 질병을 속성에 추가한다. 해당 질병을 정의하기 위한 상병코드는 ICD-10의 상병분류기호에 근거하여 정의하였고, 해당 질병에 대한 상병 코드는 부록의 <표 2>관련 질병의 상병코드 표에 명시한다.

이 모델 학습에서 중요한 것은 치매환자와 일반인을 잘 분류하는 속성을 찾아내는 것이다. 따라서 이 연구에서는 치매에 영향을 주는 요인으로 알려진 8개 변수(수축기 혈압, 식전혈당, 총콜레스테롤, 트리글리세라이드, HDL 콜레스테롤, LDL 콜레스테롤, 혈색소, 요단백)와 추가 5개 변수(수축기 혈압, 혈청크레아틴, 혈청지오티, 혈청지피티, 감마지티피)에 대해서는 5년의 데이터를 속성으로 추가한다. 앞서 언급한 것처럼 치매에 영향을 주는 요인이 있다면, 이 요인의 값이 정확하게 들어가 있어야 정확한 학습 및 예측이 가능하므로 건강검진 회수도 변수에 추가한다. 건강 검진 회수는 편의상 주요 변수 중에 결측치가 가장 낮은 변수인 식전혈당(BLDS)를 기준으로 정의한다. 이 연구에서 사용되는 변수에 대한 자세한 정의는 <표 3>변수 정의와 같다.

제 3절 자료 분석 방법

(1) 데이터 전처리 방법

데이터에 대한 결측치는 아래와 같은 기준으로 처리한다. 나이와 성별에 따라 영향이 있는 변수(신장, 체중, 허리둘레 등)는 결측치에 해당하는 나이와 성별의 중간값으로 채우고, 개인 및 가족 병력의 결측치는 해당없음에 해당하는 값으로 채운다. 연도별 변수로 추가한 13개 변수의 결측치에 대해서는 5년의 평균을 구해서 그 값으로 해당 변수의 결측치를 채운다.

(2) 변수 선정을 위한 피처 엔지니어링

이 연구에서는 Random Forest, XGBoost, MLP, SVC 등을 포함하여 다양한 기계 학습 모델을 사용한다. 최초 독립변수의 경우 전체 변수를 대상으로 Random Forest와 XGBoost를 사전에 돌려서 Feature Importances를 찾고, 이를 기준으로 각 모델을 이용하여 학습 및 예측을 반복적으로 진행한다. 예측 결과 및 변수의 영향을 검토하기 위해서 주어진 조건에 따라서 변수들의 선택을 컨트롤 하여 해당 특성의 치매의 영향성을 분석하고 좋은 결과를 내는 조건을 찾는다. 독립변수들은 단위가 달라서 sklearn의 standardscaler를 이용하여 표준화 후 학습을 진행한다.

(3) 기계학습을 위한 하이퍼파라미터 튜닝

하이퍼파라미터(hyperparameter) 최적화는 모델별로 가능한 하이퍼파라미터 값들에 대한 선택 값을 직접 지정한 후 이를 sklearn의 모형 하이퍼파라미터도구인 GridSearchCV 클래스를 통해서 최적의 파라미터를 찾는다. GridSearchCV는 5-fold Cross Validation으로 진행한다.

각각의 모델에 대해서 최적의 파라미터를 찾고 이를 이용하여 각 모델별로 테스트를 진행한다. 모델 평가는 테스트 세트의 결과로 측정하고,

평가 기준은 ROC Curve(AUC, Area Under Curve)를 사용한다.

(4) 주요 인자 분석 방법

치매에 영향을 미치는 주요 변수들에 대한 영향성을 평가하기 위해서 5년간 1회의 건강검진을 진행한 대상자를 선택하여 학습 및 예측을 진행한다. 영향성의 검토는 기본적인 변수를 선택하여 진행한 결과를 기준으로 영향성이 있는 변수를 추가했을 때 예측율의 차이로 평가하고, 그 평가 기준은 AUC로 한다.

(5) 기계학습 알고리즘

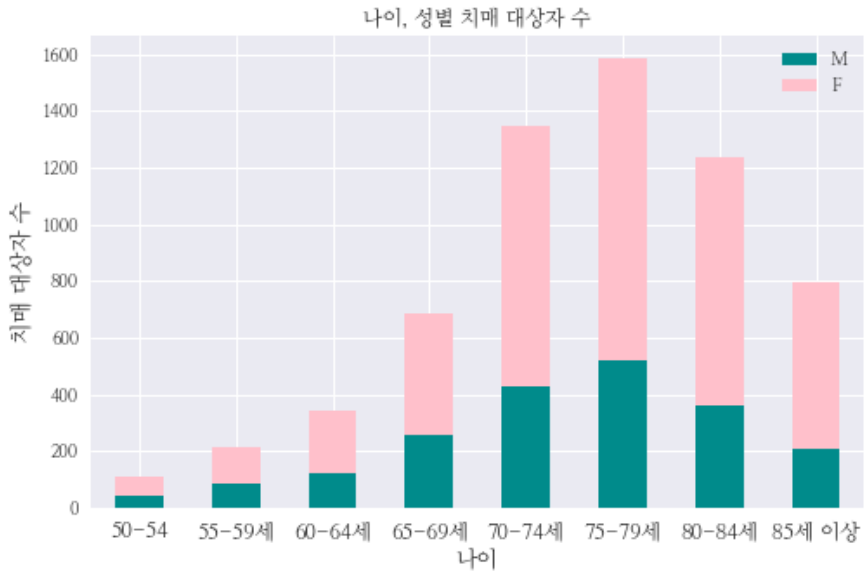
이 연구에서는 다양한 기계학습 알고리즘(LogisticRegression, k-NN, DecisonTree, ExtraTree, RandomForest, XGBoost, GradientBoosting, AdaBoost, SVC)을 사용하여 치매 학습 및 예측을 진행합니다

제 4장 연구결과

제 1절 연구 대상자의 특성

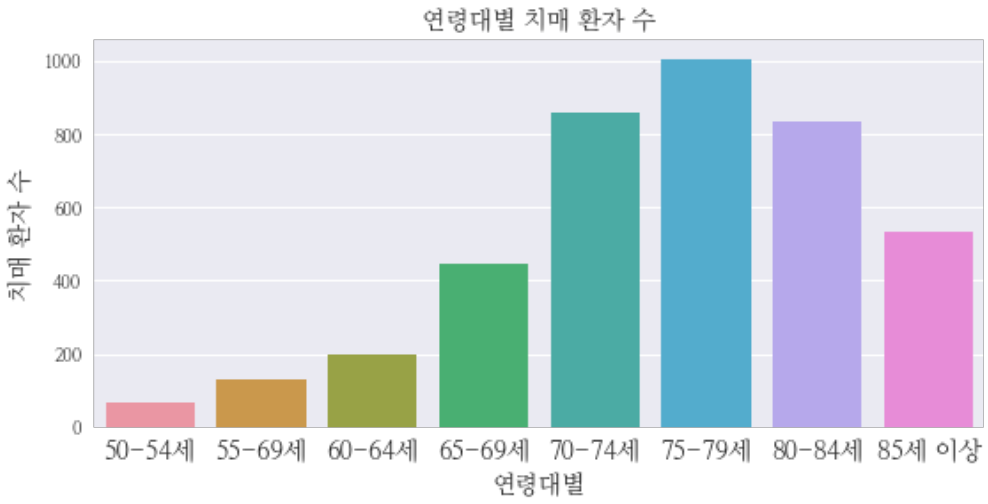
(1) 연구 대상의 주요 특징

치매는 노화에 따라 후천적으로 발생하는 질병이므로 고 연령대로 갈수로 대상자가 증가하고, 여성이 남성보다 2-3배가량 높은 비율을 차지한다. 연구 대상인 50세 이상 중 치매로 분류된 대상은 약 2.5%이고, 치매 대상자 중 65세 이하는 약 10.6%, 65세 이상은 89.4%에 가까운 비율을 차지한다. 각 연령대별, 성별 치매 환자 수는 <그림 1>과 같다.

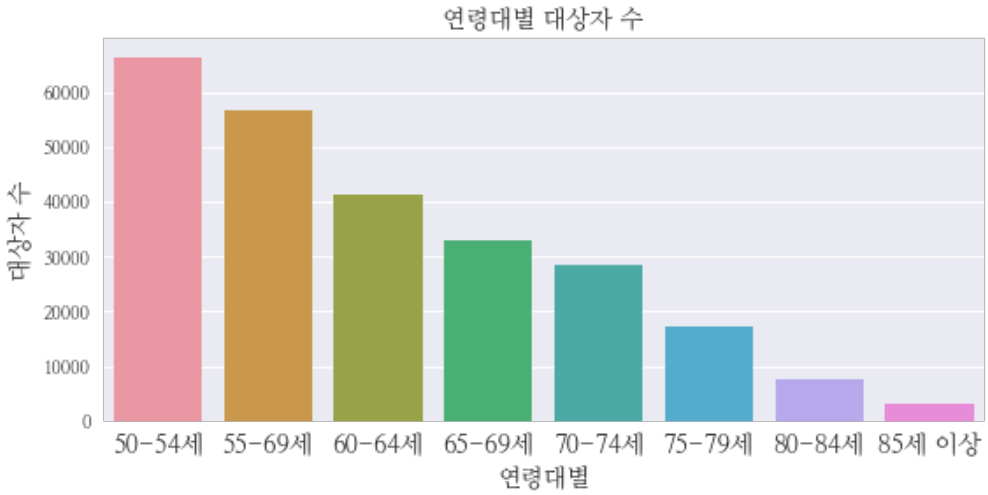


<그림 1> 연령대별, 성별 치매 환자 수

연령대별 치매 환자수 및 대상자 수는 각각 <그림 2>, <그림3>와 같다.



<그림 2> 연령대별 치매 환자 수



<그림 3> 연령대별 대상자 수

(2) 건강검진 회수에 따른 대상자

대상자 중 해당 기간 동안 건강검진을 3회 받은 대상자는 115,919명으로 가장 많고, 2회 받은 대상자는 44,950명으로 두 번째로 많다. 이 중 35,868명은 5년 동안 매년 건강검진을 받았다. 건강검진 회수 및 성별에 따른 전체 대상자 수와 치매 환자 수는 각각 <그림 4>, <그림 5>와 같다.

아래 그림에서 보여 지는 것과 같이 전체 대상자의 경우 남성과 여성의 비율이 비슷하나 치매 환자의 경우 여성의 건강검진 회수의 비율이 상대적으로 낮은 것을 확인할 수 있다.

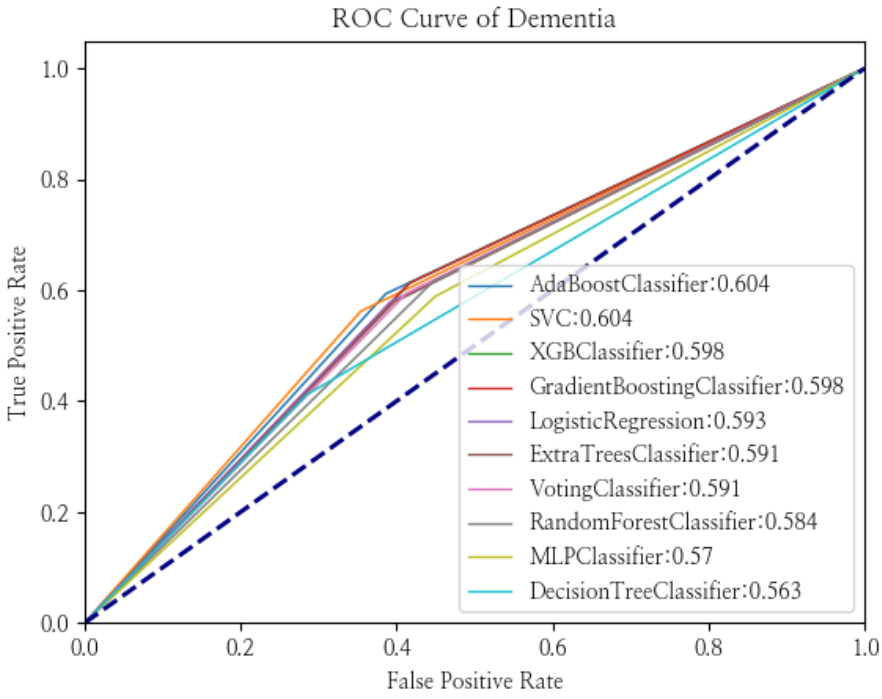
<그림 4> 건강검진 회수, 성별 대상자 수

<그림 5> 건강검진 회수, 성별 치매 환자 수

제 2절 예측 결과

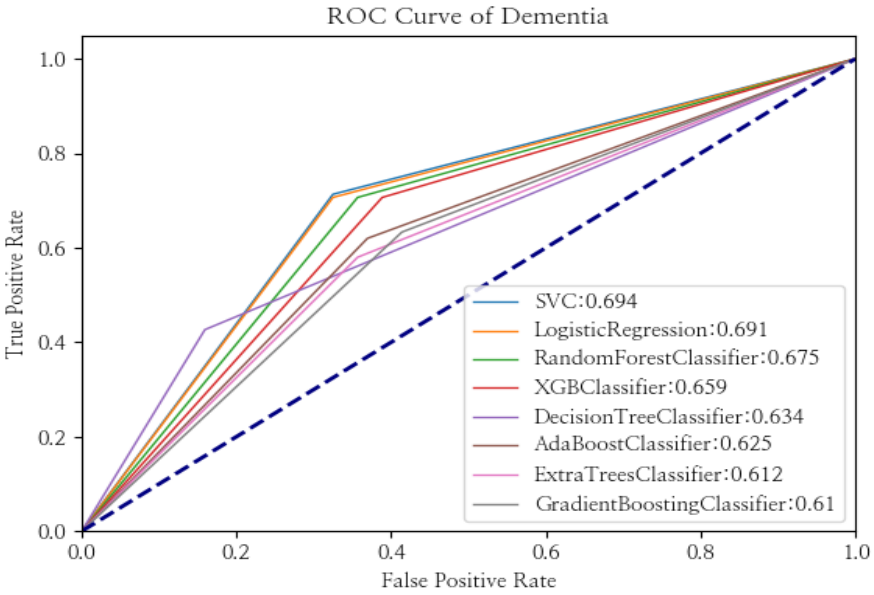
(1) 예측 정확도

치매환자 6,316명과 동일한 나이그룹에 해당하는 일반인 6,316명을 샘플링하여 12,632명을 대상으로 치매 모델링을 진행한 결과 AUC가 0.60으로 기대보다 저조한 예측 결과를 얻었다. 10개의 기계학습 알고리즘에 모델 각각에 대한 AUC는 <그림 6>과 같다.



<그림 6> 전체 대상자 치매 예측 결과

다만 5년 동안 건강검진을 1회만 진행한 739명과 동일한 나이그룹에 해당하는 739명을 샘플링하여 1,478명을 대상으로 치매 모델링을 진행한 결과 AUC가 0.69로 다소 유의한 예측 결과를 얻었다. 자세한 결과는 <그림 7>과 같다.



<그림 7> 건강검진 1회 대상자 치매 예측 결과

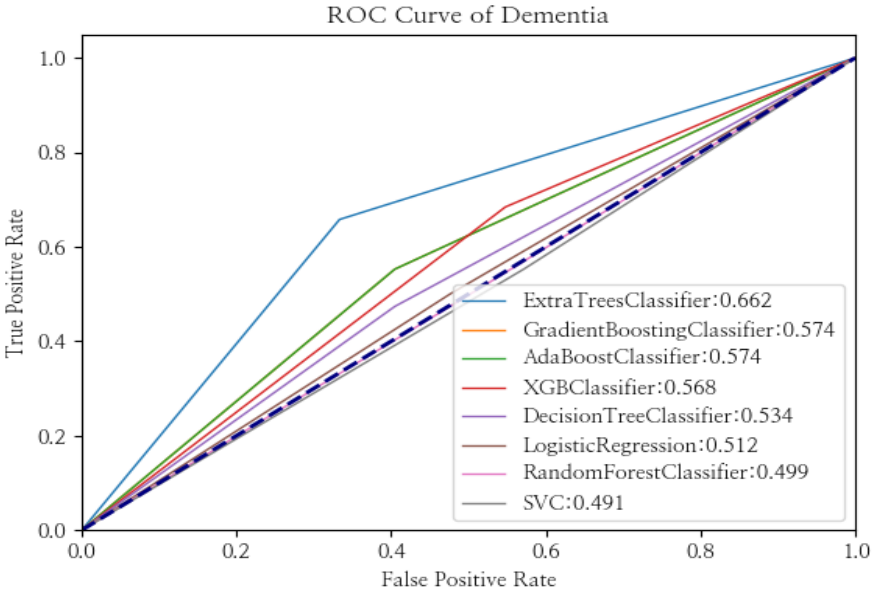
건강검진 회수에 따라서 결과가 다르고 전체적으로 예측이 되지 않은 원인이 무엇인지 파악하고 좀 더 예측 가능성을 높이기 위해서 대상 케이스를 다양하게 나누어서 모델링을 진행했다.

1) 건강검진 회수에 따른 결과

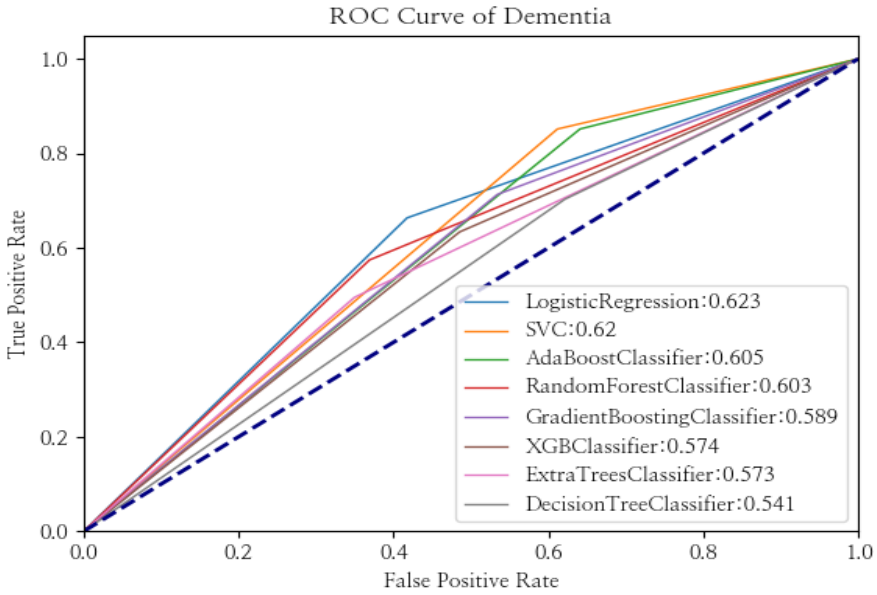
건강검진을 받은 회수에 따라서 학습 및 예측을 진행한 결과 건강검진 회수와 예측의 결과에서 차이가 있음이 확인된다. 즉, 건강검진 회수가 증가할수록 예측의 결과가 조금씩 낮아진다. 치매 비율 및 건강검진 회수의 비율이 남성과 여성이 큰 차이를 보이므로, 회수에 따른 예측은 남성과 여성을 분리하여 진행한다. 남성의 치매 예측율은 1회는 AUC가 0.67로 가장 높고, 5회는 0.62, 3회의 경우 0.60로 가장 낮은 예측 결과가 나왔고, 최고와 최저 예측율은 0.07 정도의 차이를 보인다. 여성의 치매 예측율은 1회는 0.65로 가장 높고, 3회는 0.59, 5회의 경우 0.55로 가장 낮아, 최고

와 최저 예측율의 차이는 0.1이다. 남성보다는 상대적으로 조금 더 큰 차이를 보이나 이 정도는 샘플링에 의한 차이로 해석된다.

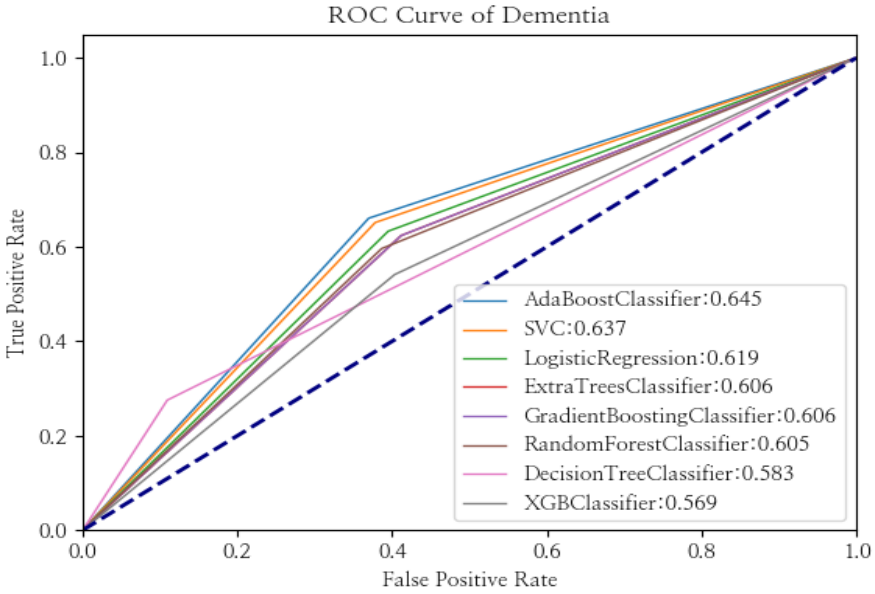
남성, 여성 각각에 대하여 건강검진 회수(1회, 5회)에 따른 모델별 상세한 AUC 결과는 <그림 8>, <그림 9>, <그림 10>, <그림 11>과 같다.



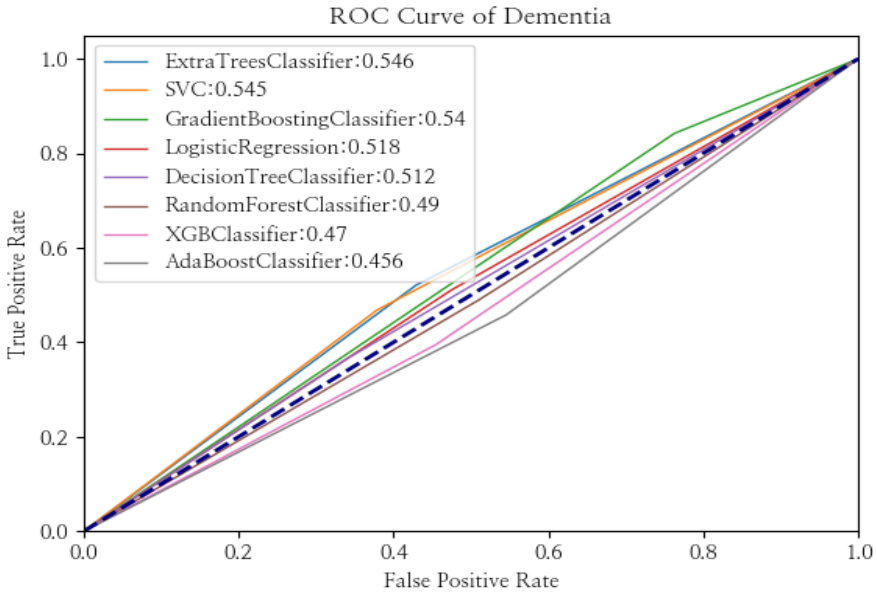
<그림 8> 남성, 건강검진 1회 대상자의 예측 결과



<그림 9> 남성, 건강검진 5회 대상자의 예측 결과



<그림 10> 여성, 건강검진 1회 대상자의 예측 결과

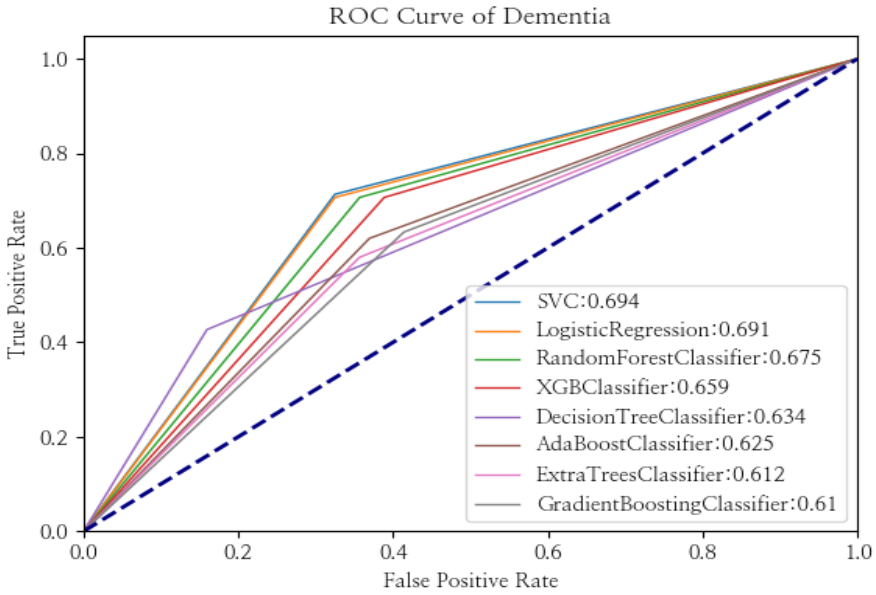


<그림 11> 여성, 건강검진 1회 대상자의 예측 결과

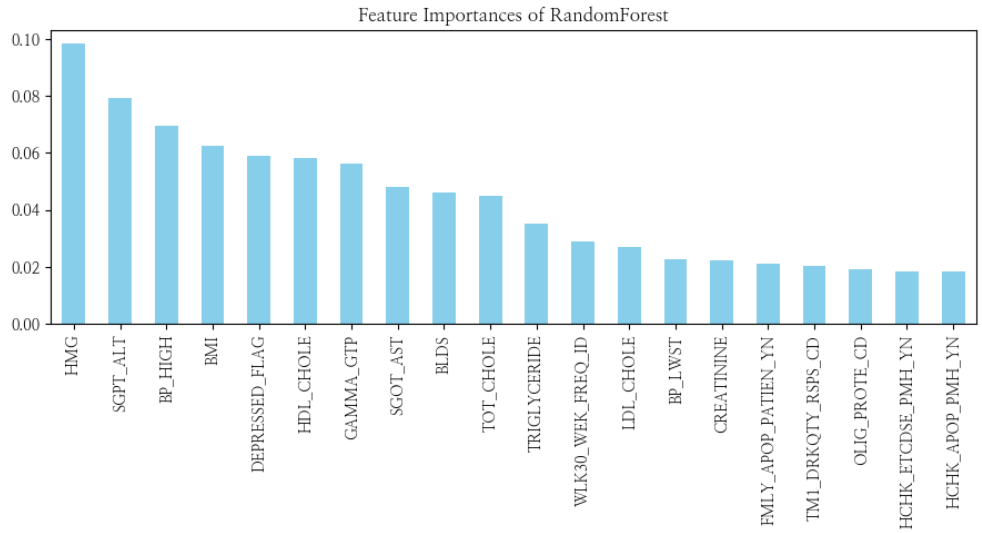
2) 건강검진 회수에 따른 예측율 차이 분석

건강검진 회수에 따라 예측율이 크게 달라지는 정확한 원인을 파악하기 위하여 대상자를 건강검진 1회와 5회로 나누어서 2013년 속성에 해당하는 변수만 추가하여 예측을 진행하였다. 앞서 언급한 것처럼 1회의 결과는 0.69로 다소 의미 있는 결과를 보였으나, 5회의 결과는 AUC 0.60으로 크게 낮아지는 것이 확인된다.

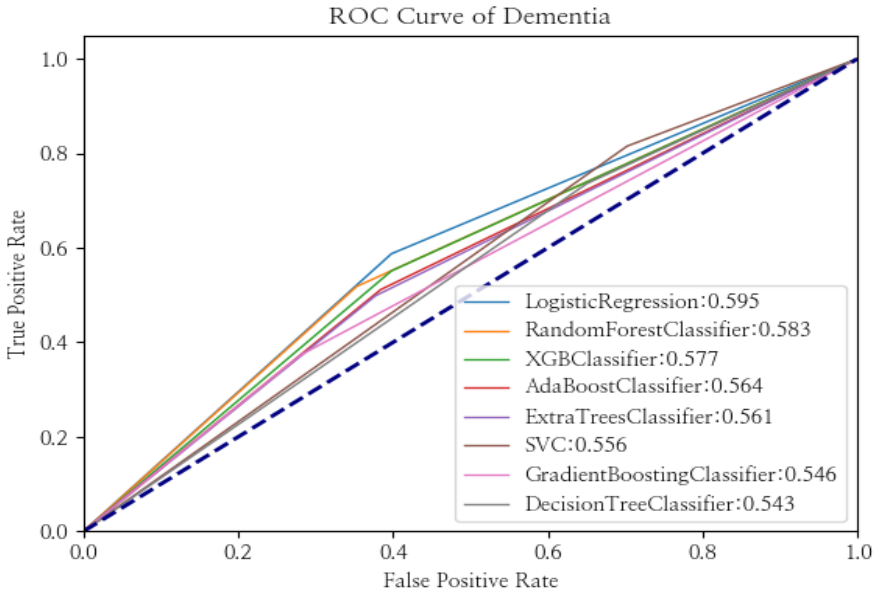
<그림 14>과 <그림 15>의 feature importances에서 보여 지는 것처럼, 1회 대상자와는 다르게 5회 대상자에서는 운동, 음주, 흡연 관련한 변수가 상대적으로 중요도가 있는 것이 확인된다. 이 변수들이 예측결과에 영향을 준 것으로 생각되나 정확한 원인에 대해서는 좀 더 연구가 필요하다. 건강검진 1회 대상자와 5회 대상자 각각에 대한 AUC 결과 및 feature importances는 각각 <그림 12>, <그림 13>, <그림 14>, <그림 15>와 같다.



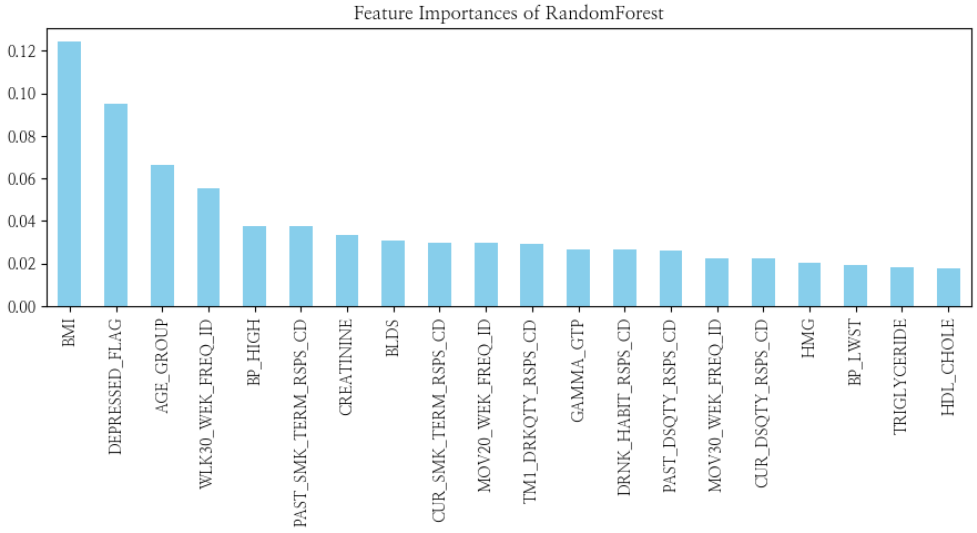
<그림 12> 건강검진 1회 대상자 중 2013년 변수만 포함 예측 결과



<그림 13> 건강검진 1회 대상자 중 2013년 변수만 포함의 feature importances



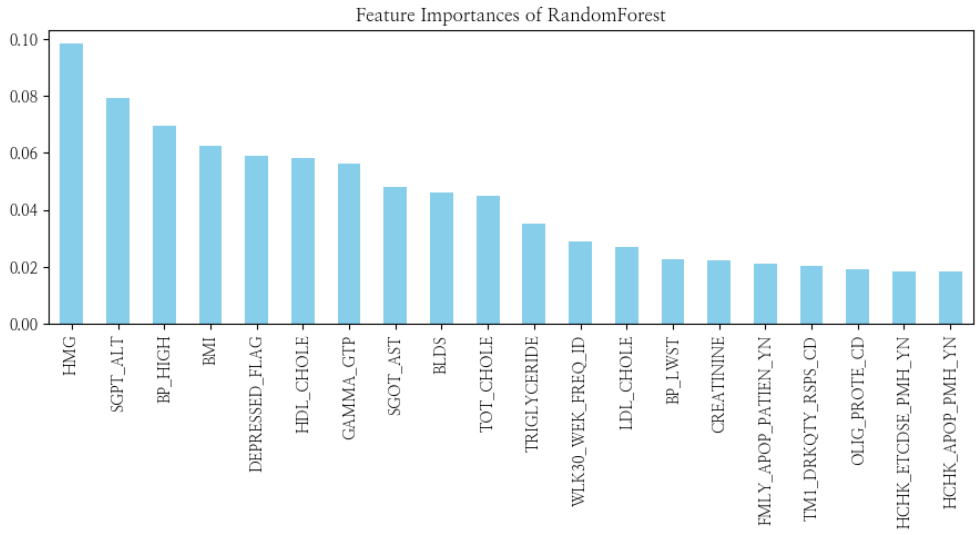
<그림 14> 건강검진 5회 대상자 중 2013년 변수만 포함한 예측 결과



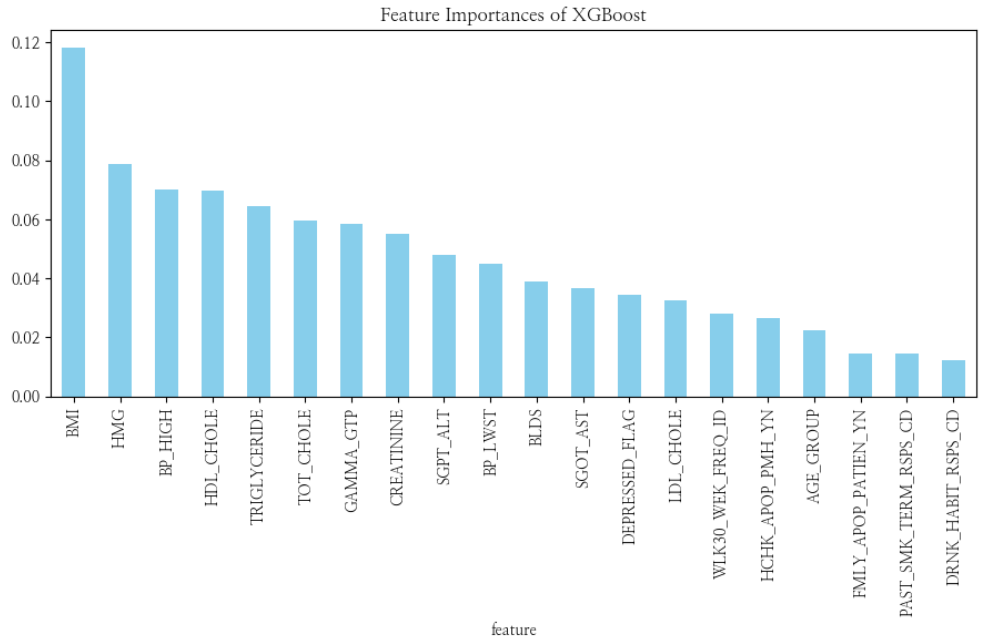
<그림 15> 건강검진 5회 대상자 중 2013년 변수만feature importances

(2) Feature Importances

Random Forest와 XGBoost를 이용하여 치매 예측에 영향을 주는 특성을 파악한 결과 조금 다른 결과를 보였다. 상대적으로 높은 영향을 주는 변수는 HMG(혈색소), BMI(체질량지수), BP_HIGH(식전혈당), HDL_CHOLE(고밀도 콜레스테롤), GAMMA_GTP(감마지티피), TOT_CHOLE(총 콜레스테롤)등의 순서를 보인다. 각 요소가 작은 차이를 보이지만 상대적으로 높은 영향도를 갖는 변수를 보면 체질량, 콜레스테롤 수치, 당뇨 등이 치매에 영향을 주는 요소라는 것을 간접적으로 설명해준다. 두 방법 모두에서 상대적으로 높은 feature importances 값을 가지는 변수들을 연구의 주요 인자로 정의하고 사용한다. Random Forest와 XGBoost의 feature importances는 각각 <그림 16>, <그림 17>와 같다.



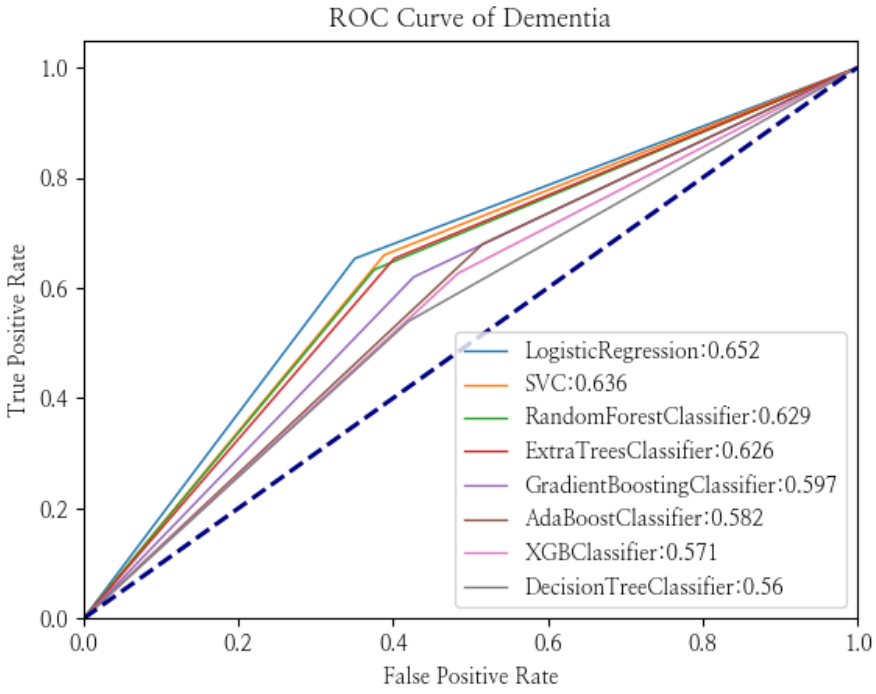
<그림 16> feature importances of Random Forest



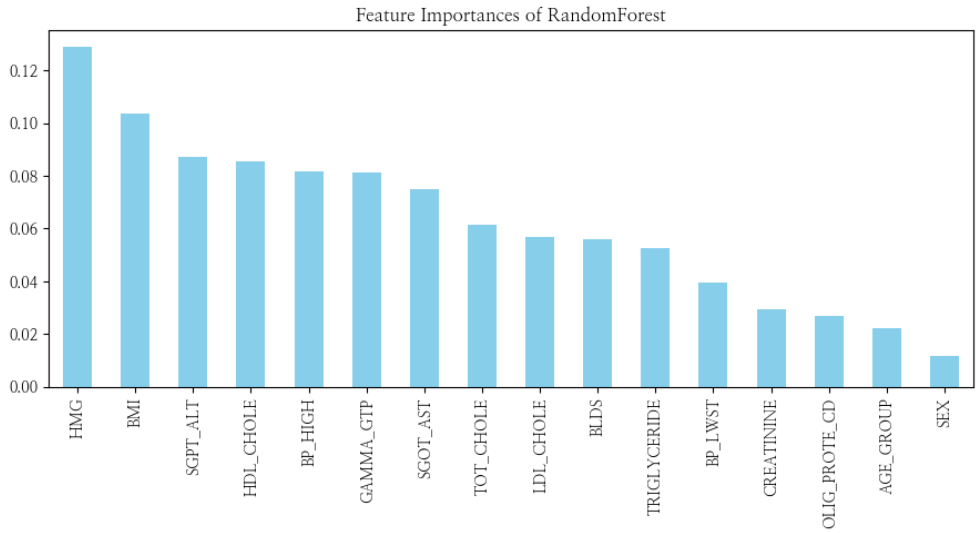
<그림 17> feature importances of XGBoost

제 3절 주요 인자 분석 결과

치매에 영향을 미치는 주요 변수들에 대한 영향성을 평가하기 위해서 5년간 1회의 건강검진을 진행한 대상자를 선택하여 학습 및 예측을 진행했다. 앞서 결과에서 보여 진 것처럼 1회 건강검진을 진행한 대상자들에 대한 전체 변수를 포함한 치매 예측율은 AUC 0.69로 다소 의미 있는 결과가 나왔다. 영향성의 검토는 기본적인 변수를 선택하여 진행한 결과를 기준으로 해당 변수를 추가했을 때 예측율의 차이로 평가하고 그 평가 기준은 AUC로 한다. 기본적인 변수를 선택하여 진행한 결과 AUC는 0.65로 다소 떨어짐을 확인 할 수 있다. 해당 결과에 대한 feature importances는 좀 더 높은 예측 결과를 보인 Random Forest를 기준으로 한다. 자세한 내용은 <그림 18>, <그림 19>과 같다.



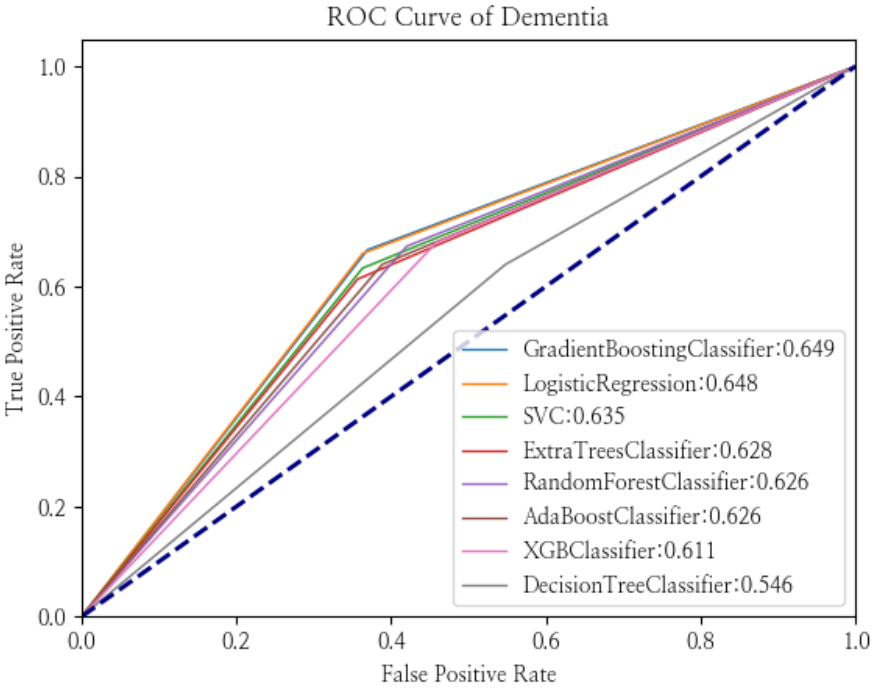
<그림 18> 건강검진 1회 대상자, 기본 변수만 포함 예측 결과



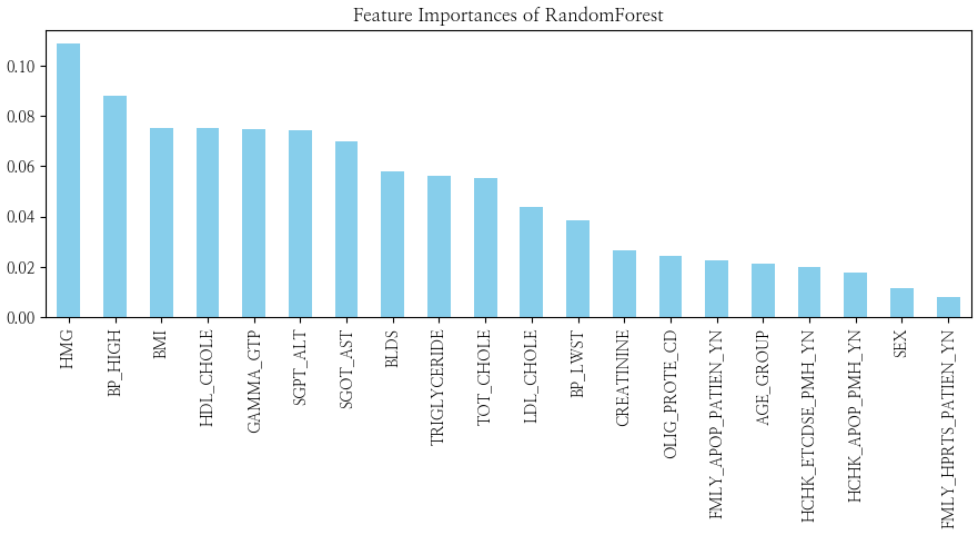
<그림 19> 건강검진 1회 대상자, 기본 변수만 포함의 feature importances

(1) 본인 과거 병력과 가족의 병력의 영향성 분석

본인의 과거 병력과 가족력이 치매에 미치는 영향을 살펴보기 위해 건강검진 항목 중 본인의 과거 병력과 가족력을 변수에 포함하여 학습 및 예측을 진행하였다. 그 결과 족의 뇌졸중과 고혈압 및 본인의 뇌졸증 과거 병력이 다른 요소에 상대적인 중요도는 높으나 예측 결과에서는 차이가 없다. 문진 데이터의 경우 입력된 내용이 많지 않아서 영향성을 파악하기에는 한계가 있다. 자세한 사항은 <그림 20>, <그림 21>과 같다.



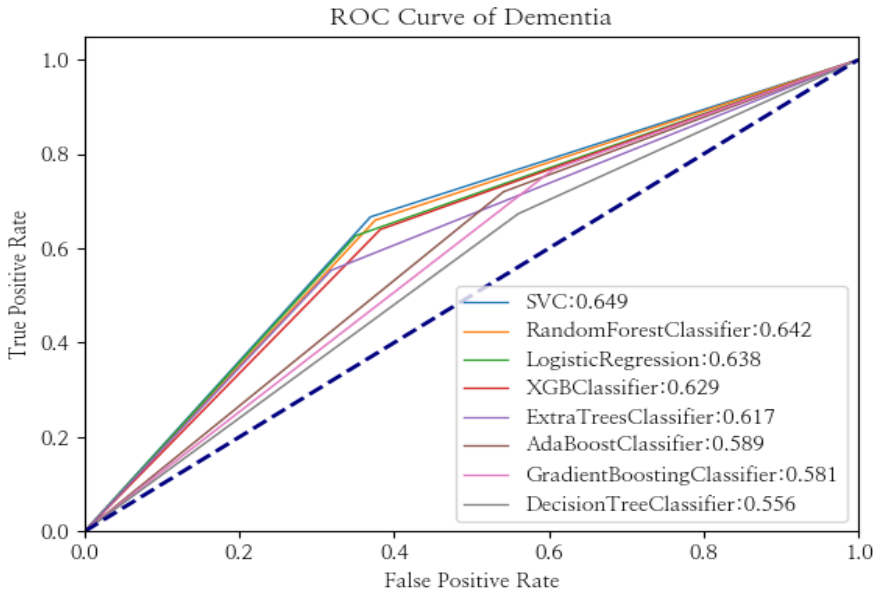
<그림 20> 본인 과거병력, 가족병력 포함 예측 결과



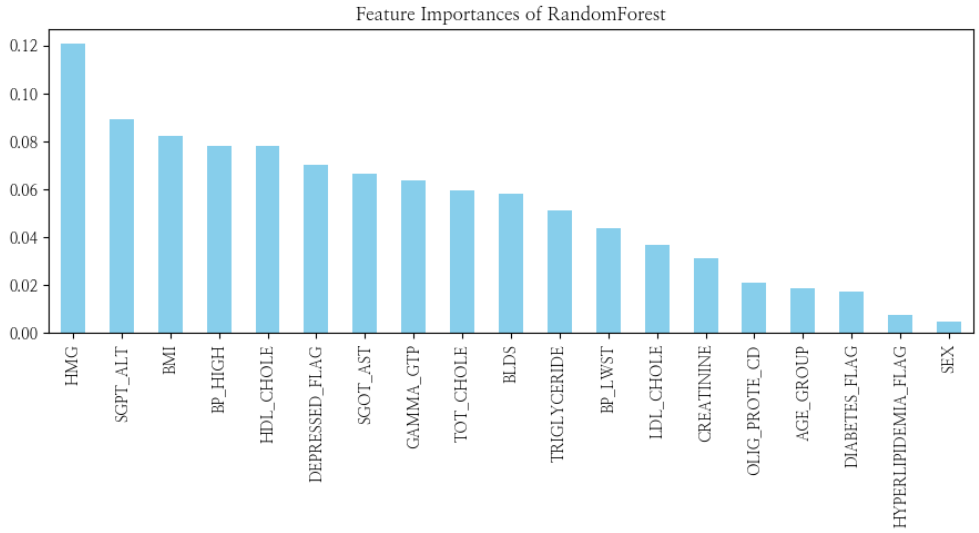
<그림 21> 본인 과거병력, 가족병력 포함의 feature importances

(2) 우울증, 당뇨, 고지혈증 영향성 분석

우울증은 치매의 위험인자로 알려져 있다. 우울증을 앓은 적이 있는 노인은 그렇지 않은 노인에 비해 치매 발병 가능성을 증가시킨다(ADI, 2014). 관련 질병이 치매에 미치는 영향을 분석하기 위하여 우울증을 포함한 당뇨, 고지혈증을 속성으로 추가하여 학습 및 예측을 진행했다. 그 결과 우울증과 고지혈증이 다른 질병보다 상대적으로 영향을 주는 것으로 보이나 AUC는 0.65으로 큰 영향은 보이지 않는다. 예측 결과 및 feature importance는 <그림 22>, <그림 23>과 같다.



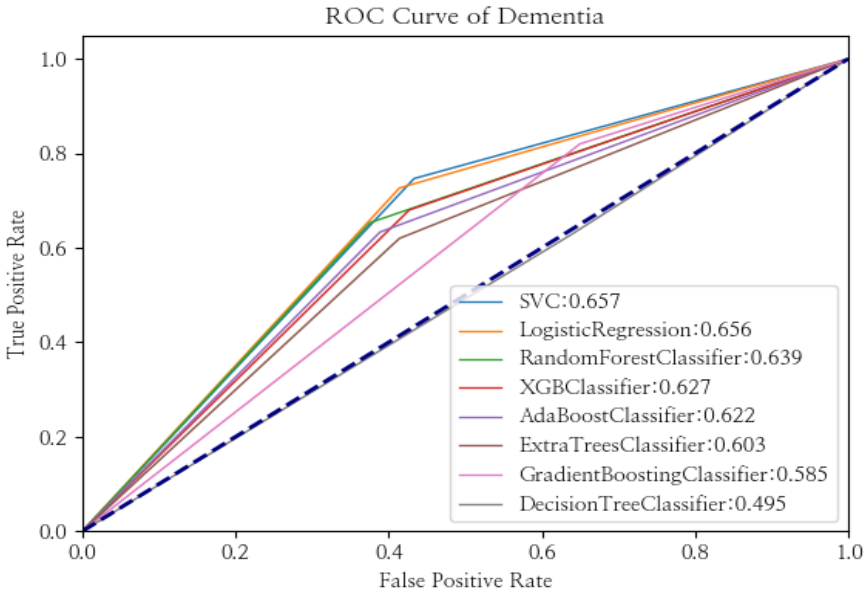
<그림 22> 관련 질병 포함 예측 결과



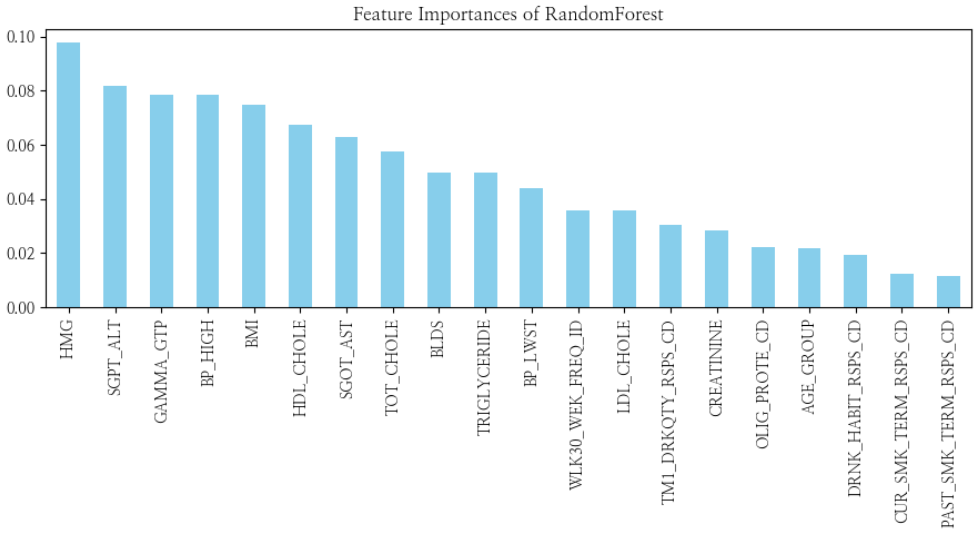
<그림 23> 관련 질병 포함의 feature importances

(3) 운동, 흡연, 음주 등 건강 관련 생활습관의 영향성 분석

운동이 치매와 음의 상관관계가 있다는 선행연구가 있다. 운동, 흡연, 음주 등의 생활습관이 치매에 영향을 주는 지를 확인하기 위해서 운동, 흡연, 음주관련 변수를 포함하여 예측을 진행했다. 그 결과 미미하지만 예측율이 약간 올라 간 것이 확인된다. 예측율 및 feature importances 관련한 자세한 사항은 각각 <그림 24>, <그림 25>와 같다.



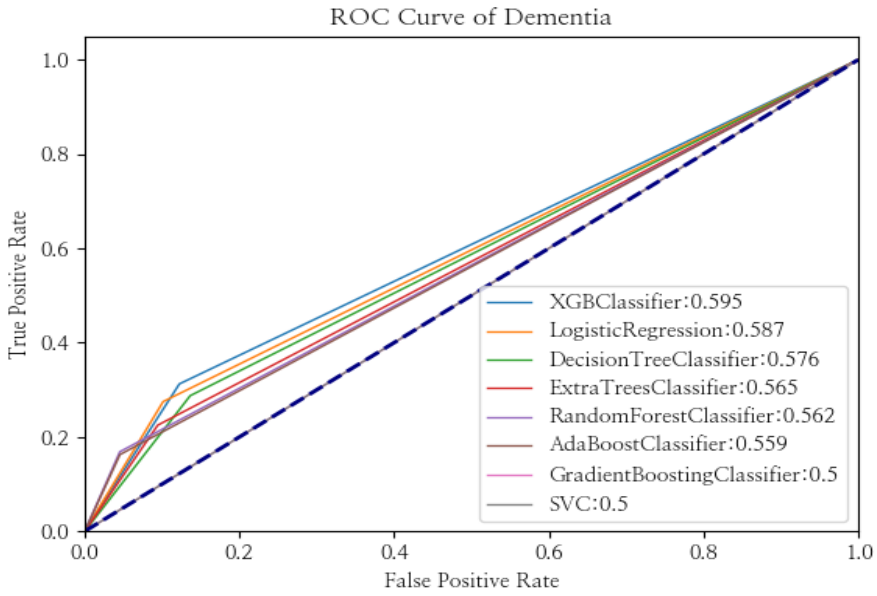
<그림 24> 운동, 흡연, 음주 변수를 포함 예측 결과



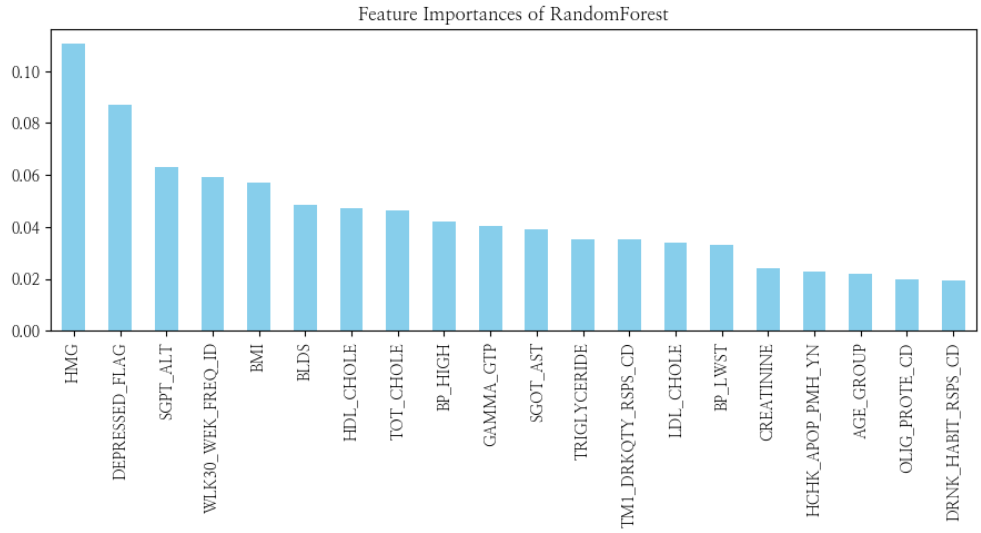
<그림 25> 운동, 흡연, 음주 변수를 포함의 feature importances

(4) 2배수 샘플링의 예측 결과

일반인 대상자 수가 많은 경우 이의 특성이 강하게 학습되어 치매 환자에 대한 정확한 예측이 되지 않는다. 이를 확인하기 위해서 치매 환자수의 2배수에 해당하는 일반인 대상자를 샘플링하여 학습 및 예측을 진행했다. 예측결과 AUC가 0.6으로 많이 떨어진 것이 확인된다. 이를 통해 질병 예측의 경우 케이스 컨트롤이 매우 중요함을 확인하였다. 이에 대한 모델 별 예측 결과 및 feature importance는 각각 <그림 26>, <그림 27>와 같다.



<그림 26> 치매 환자수의 2배수 일반인 대상자의 예측 결과



<그림 27> 치매 환자수의 2배수 일반인 대상자의 feature importances

제 5장 결론

제 1절 연구의 시사점

본 연구에서는 건강검진데이터를 이용하여 선행 연구들에서 밝혀진 치매의 위험인자를 기반으로 다양한 머신러닝 알고리즘을 이용하여 치매를 예측하고, 예측에 영향을 주는 요인들에 대해서 분석을 진행했다. 의료지식의 한계로 다양한 컨트롤 케이스에 대한 명확한 기준 및 정의가 부족하여 치매 예측율이 안정적이지 못하고, 기대했던 수준까지는 높이지 못했다. 다만 다양한 속성 추가 및 컨트롤 케이스에 대한 작은 기준을 정한 후 예측을 진행하니 제한된 케이스(건강검진 1회 대상자)에서는 어느 정도 의미 있는 결과가 나오는 것을 확인했고, 영향을 주는 요소들에 대해서도 확인이 되었다.

건강검진의 기본항목 이외에 문진정보(개인병력, 가족병력 및 음주, 흡연, 운동의 생활습관)가 치매 예측율에 영향을 주는 것은 확인되었으나 해당 데이터의 경우 제대로 입력이 되지 않아서 이 부분이 보완된다면 좀 더 의미 있는 결과가 나올 것으로 예상한다. 향후 건강검진데이터와 문진데이터가 규칙적으로 정확하게 쌓이고, 질병간의 정확한 연관성 및 질병에 대한 약 처방 등의 데이터를 추가한다면, 이 데이터를 기반으로 치매뿐만 아니라 다른 질병에 대한 예측도 가능할 것으로 생각된다.

제 2절 연구의 한계

인공지능 기술을 활용한 의료기기 개발 및 이를 의료에 활용하려는 사례가 증가함에 따라 의료서비스의 질이 크게 향상되는 추세이다. 다만 사회적으로 민감한 부분인 의료분야의 데이터는 반드시 해결해야 할 법적,

윤리적 이슈들이 존재한다. 특히 치매의 경우 2015년부터 민감 상병으로 분류되어 진료 명세 데이터베이스의 주상병에서 마스킹 처리가 되기 때문에 이와 같은 연구 및 분석에 제한이 될 것으로 예상된다. 치매를 예측하기 위해서는 일정 기간의 연속적인 데이터가 필요하고 예측된 결과를 가지고 이후 데이터에 적용하여 예측정확도를 확인할 필요가 있으나 법적인 제약으로 일정기간 접근이 어려울 것으로 예상된다.

또한 좀 더 정확한 분석 및 결과를 끌어내기 위해서는 의학적인 지식 및 추가 데이터에 대한 확인이 필요하나 이 부분의 연계가 부족하여 아쉬움으로 남는다.

제 3절 향후 연구 방향

본 연구를 바탕으로 2014년 이후에 진행한 건강검진자료를 추가하여 예측을 진행하고, 기존에 분석된 요인, 약물처방 및 추가 요소를 결합하여 좀 더 정확하게 치매 예측을 할 수 있도록 연구가 필요하다. 이를 통해 치매 발병 초기에 빠르고 적극적인 의료진의 개입으로 건강한 삶을 유지하여 삶의 질이 개선되기를 기대한다. 최근 몇 년 사이 데이터의 중요성이 부각되면서 의료데이터를 포함한 다양한 영역의 공공데이터가 공공데이터포털 및 다양한 기관을 통해 다운로드 및 API 연동을 통해서 오픈이 되어 있다. 아직까지 의료데이터를 이용하는데 제약이 많지만 개인정보 및 민감정보에 대한 정책이 잘 개선되어 다양한 의료 데이터를 분석하여 치매 뿐 만 아니라 다양한 질병들의 예측이 개인이 활용할 수 있는 수준으로까지 가능하게 되기를 바란다.

참고문헌

고숙자, 정영호, 김동영(2016), “초고령사회 대응을 위한 치매의 사회적 부담과 예방 및 관리 방안,” 한국보건사회연구원 연구보고서, 2016-04

국경완(2019), “인공지능 기술 및 산업 분야별 적용 사례,”, 정보통신기획평가원, www.itfind.or.kr/WZIN/jugidong/1888/file61118014710062059 40-188802.pdf, 검색일 1029.6.29.

국민건강보험공단(2015), 표본연구DB참고자료

김형섭, 김종헌, 조한얼, 임현선(2018), “치매 특별 등급(장기요양등급) 자료 분석을 통한 치매 예측 모델 개발 및 조기 개입 효과 조사,” 국민건강보험 일산병원 연구소, HNIMC-2018-01-008.

박순영(2017), ICT 융합 의료산업 동향, 융합연구정책센터, 2017 APRIL Vol. 68

박운제(2017), “치매환자 특성의 추이 분석,” Journal of the Korea Academia-Industrial cooperation Society Vol. 18, No. 1 pp. 202-209.

류경희, 강연옥, 나덕렬, 이광호, 정진상(2000), 치매 환자의 우울 특성, Korean Journal of Clinical Psychology Vol. 19, No. 1, 117-129

보건복지부, 분당서울대학교병원(2013), 2012년 치매 유병률 조사.

질병관리본부, 질벼관리본부(2014), 치매 조기진단 중요성, 보도참고자료, 2014.03.20

오은아, 강연욱, 신준형, 연병길(2010), 치매선별검사로서 K-MMSE의 타당도 연구: 종합적인 신경심리학적 평가와의 비교, Dementia and Neurocognitive Disorders, 2010; 9: 8-12

의학신문(2018), “4차 산업 혁명시대, 의료계에 부는 AI 열풍,” http://www.bosa.co.kr/news/articleView.html?idxno=2075727, 2019.6.30.

전진아(2016), “한국 노인의 신체적 정신적 건강 및 건강행태: 사회적 지지를 중심으로,” 한국보건행정학회 학술대회논문집, 제2호 39.

정원준(2018), “국내 인공지능(AI) 의료기기 현황 및 규제 이슈”, 의료기기뉴스라인,

2018.01.30.

전진옥(2018), 우리나라 전자의무기록도입 현황 및 발전 과제, HIRA 정책동향 2018년 12권 3호

최성혜(2012), “치매의 임상적 진단,” J Korean Diabetes. 2012 Sep;13(3).

최윤섭(2018), “진료 기록으로 치료결과를 예측하는 구글의 인공지능,” http://www.yoonsupchoi.com/2018/03/14/google-emr-ai/, 검색일 2019.6.29.

ADI.((2014), *World Alzheimer's Report 2014 : Demential and Risk Reduction,* Alzheimer's Disease International.

Ji Hwan Park, Han Eoi Cho, Hyun Sun Lim, Jong-Hun Kim, Melanie Wall, Shinjae Yoo, Hyoung Seop Kim and Jiook Cha.(2019), Electronic Health Records Based Prediction of Future Incidence of Alzheimer's Disease Using Machine Learning. Psychiatry Neurology.

Jingshu Liu, Zachariah Zhang, Narges Razavian(2018), *Proceedings of the 3rd Machine Learning for Healthcare Conference*, PMLR 85:440-464.

SAVVA, G. & STEPHAN, B. (2010). *Epidemiological studies of the effect of stroke on incident dementia*: a systematic review. 4

Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi and Nadeem Qureshi(2017), *Can machine-learning improve cardiovascular risk prediction using routine clinical data?*, https://doi.org/10.1371/journal.pone.0174944

부록

[표 1] 상병코드에 따른 치매질환의 분류

상병코드	질환명
F00	알츠하이머병에서의 치매
F00.0	조기발병 알츠하이머병에서의 치매
F00.1	만기발병 알츠하이머병에서의 치매
F00.2	비정형/혼합형 알츠하이머병에서의 치매
F00.9	상세불명 알츠하이머병에서의 치매
F01	혈관성 치매
F010	급성 발병의 혈관성 치매
F011	다발 경색 치매
F012	피질하 혈관성 치매
F013	혼합형 피질 및 피질하 혈관성 치매
F018	기타 혈관성 치매
F019	상세불명의 혈관성 치매
F02	달리 분류된 기타 질환에서의 치매
F020	피크병에서의 치매
F021	크로이츠펠트.야콥병에서의 치매
F022	헌팅톤병에서의 치매
F023	파킨슨병에서의 치매
F024	인체면역결핍바이러스병에서의 치매
F028	달리 분류된 기타 명시된 질환에서의 치매
F03	상세불명의 치매

[표 2] 관련 질병의 상병코드 표

관련 질병	상병코드
우울증	'F329', 'F313', 'F412', 'H533','F328', 'F323', 'F323', 'F920', 'F341', 'F329', 'Z133', 'F412', 'F329', 'F328', 'F323'
당뇨병	'E741', 'E741', 'E741', 'R81', 'E723', 'E14', 'E1468', 'R730', 'E831', 'E1400', 'E1408', 'O249', 'E1428', 'N083', 'Z131', 'E1428', 'N083', 'O249', 'Z833', 'E785'
고지혈증	'E785', 'E780', 'H432', 'K824', 'E786', 'N084', 'J848', 'J84.8', 'E789', 'E789', 'E755', 'K8020'
고혈압	I109', 'I701', 'K766', 'O159', 'Z136', 'R030', 'I132', 'I119', 'I139', 'I101', 'R030', 'O15.2'

[표 3] 변수 정의

번호	변수명	설명
1	BLDS	식전혈당(공복혈당) 2013
2	BLDS_1	식전혈당(공복혈당) 2012
3	BLDS_2	식전혈당(공복혈당) 2011
4	BLDS_3	식전혈당(공복혈당) 2010
5	BLDS_4	식전혈당(공복혈당) 2009
6	BP_HIGH	수축기혈압 2013
7	BP_HIGH_1	수축기혈압 2012
8	BP_HIGH_2	수축기혈압 2011
9	BP_HIGH_3	수축기혈압 2010
10	BP_HIGH_4	수축기혈압 2009
11	BP_LWST	이완기혈압 2013

12	BP_LWST_1	이완기혈압 2012
13	BP_LWST_2	이완기혈압 2011
14	BP_LWST_3	이완기혈압 2010
15	BP_LWST_4	이완기혈압 2009
16	CREATININE	혈청크레아틴 2013
17	CREATININE_1	혈청크레아틴 2012
18	CREATININE_2	혈청크레아틴 2011
19	CREATININE_3	혈청크레아틴 2010
20	CREATININE_4	혈청크레아틴 2009
21	CUR_DSQTY_RSPS_CD	(현재)하루흡연양
22	CUR_SMK_TERM_RSPS_CD	(현재)하루흡연기간
23	DRNK_HABIT_RSPS_CD	음주습관
24	FMLY_APOP_PATIEN_YN	(가족력)뇌졸중환자유무
25	FMLY_CANCER_PATIEN_YN	(가족력)기타(함포함)환자유무
26	FMLY_DIABML_PATIEN_YN	(가족력)당뇨병환자유무
27	FMLY_HDISE_PATIEN_YN	(가족력)심장병환자유무
28	FMLY_HPRTS_PATIEN_YN	(가족력)고혈압환자유무
29	GAMMA_GTP	감마지티피 2013
30	GAMMA_GTP_1	감마지티피 2012
31	GAMMA_GTP_2	감마지티피 2011
32	GAMMA_GTP_3	감마지티피 2010
33	GAMMA_GTP_4	감마지티피 2009
30	HCHK_APOP_PMH_YN	(본인)뇌졸중관거병력유무
31	HCHK_DIABML_PMH_YN	(본인)당뇨병과거병력유무
32	HCHK_ETCDSE_PMH_YN	(본인)기타(암포함)질환관거병력
33	HCHK_HDISE_PMH_YN	(본인)심장병과거병력유무
34	HCHK_HPLPDM_PMH_YN	(본인)고지혈증(이상지혈증)과거병력
35	HCHK_HPRTS_PMH_YN	(본인)고혈압과거병력유무

36	HCHK_PHSS_PMH_YN	(본인)폐결핵과거병력유무
37	HDL_CHOLE	HDL 콜레스테롤 2013
38	HDL_CHOLE_1	HDL 콜레스테롤 2012
39	HDL_CHOLE_2	HDL 콜레스테롤 2011
40	HDL_CHOLE_3	HDL 콜레스테롤 2010
41	HDL_CHOLE_4	HDL 콜레스테롤 2009
42	HEIGHT	신장
43	HMG	혈색소 2013
44	HMG_1	혈색소 2012
45	HMG_2	혈색소 2011
46	HMG_3	혈색소 2010
47	HMG_4	혈색소 2009
48	LDL_CHOLE	LDL 콜레스테롤 2013
49	LDL_CHOLE_1	LDL 콜레스테롤 2012
50	LDL_CHOLE_2	LDL 콜레스테롤 2011
51	LDL_CHOLE_3	LDL 콜레스테롤 2010
52	LDL_CHOLE_4	LDL 콜레스테롤 2009
53	MOV20_WEK_FREQ_ID	1주 20분이상 격렬한 운동
54	MOV30_WEK_FREQ_ID	1주 30분이상 격렬한 운동
55	OLIG_PROTE_CD	요잠혈 2013
56	OLIG_PROTE_CD_1	요잠혈 2012
57	OLIG_PROTE_CD_2	요잠혈 2011
58	OLIG_PROTE_CD_3	요잠혈 2010
59	OLIG_PROTE_CD_4	요잠혈 2009
60	PAST_DSQTY_RSPS_CD	(과거)하루흡연양
61	PAST_SMK_TERM_RSPS_CD	(과거)흡연기간
62	PERSON_ID	개인식별번호
63	SGOT_AST	혈청지오티AST 2013

64	SGOT_AST_1	혈청지오티AST 2012
65	SGOT_AST_2	혈청지오티AST 2011
66	SGOT_AST_3	혈청지오티AST 2010
67	SGOT_AST_4	혈청지오티AST 2009
68	SGPT_ALT	혈청지피티ALT 2013
69	SGPT_ALT_1	혈청지피티ALT 2012
70	SGPT_ALT_2	혈청지피티ALT 2011
71	SGPT_ALT_3	혈청지피티ALT 2010
72	SGPT_ALT_4	혈청지피티ALT 2009
73	SMK_STAT_TYPE_RSPS_CD	흡연상태
74	TM1_DRKQTY_RSPS_CD	1회음주량
75	TOT_CHOLE	총콜레스테롤 2013
76	TOT_CHOLE_1	총콜레스테롤 2012
77	TOT_CHOLE_2	총콜레스테롤 2011
78	TOT_CHOLE_3	총콜레스테롤 2010
79	TOT_CHOLE_4	총콜레스테롤 2009
80	TRIGLYCERIDE	트리글리세라이드 2013
81	TRIGLYCERIDE_1	트리글리세라이드 2012
82	TRIGLYCERIDE_2	트리글리세라이드 2011
83	TRIGLYCERIDE_3	트리글리세라이드 2010
84	TRIGLYCERIDE_4	트리글리세라이드 2009
85	WAIST	허리둘레
86	BMI	체질량지수
87	WLK30_WEK_FREQ_ID	1주 총30분이상 걷기 운동
88	YKIHO_GUBUN_CD	검진기관종별코드
89	CFLAG	건강검진 회수(2009-2013)
90	GROUP	치매여부(0:정상, 1:치매)
91	DEPRESSED_FLAG	우울증
92	DIABETES_FLAG	당뇨

93	HIGH_BLOOD_PRESSURE_F	고혈압
94	HYPERLIPIDEMIA_FLAG	고지혈증
95	SEX	성별(1:남, 2:여)
96	AGE_GROUP	나이그룹(1~18), 5살 단위로

Abstract

The Dementia Prediction and Efficient Factor Analysis by Utilizing the Electronic Health Records

Lee, Younghee
Seoul School of Integrated Sciences and Technologies
Advisor:

Artificial Intelligence is a technology that implements and automates human abilities through executable programs including cognition, learning or inference. It has been developed into machine learning, natural languages processing, voice and video recognition, and its application has become more widespread in different fields such as medical, finance and manufacturing industries. Especially, in the medical industry, those technologies are applied to predict, diagnose diseases and prescribe medication. South Korea has rapidly become an aging society. Accordingly, the aging-related diseases have been aggravated, and among them, an outbreak of dementia and the weight of its health costs have developed into a social problem. Meanwhile, according to several researches, the accurate forecast on the possibility of dementia can facilitate the early engagement that can delay its outbreak.

Hence, this study explores the possibility of the prediction on de-

mentia based on the machine learning by employing EHR data, and analyzes critical factors that affect the prediction level. The study collates the sampling cohort database from the National Health Insurance Service of South Korea between 2002 and 2013 gathered from around 1 million people.

Existing researches on dementia and Alzheimer's disease have considered people older than 64 but the present study selects the targets older than 49 to investigate the early dementia prediction. Among EHR data, 50 intrinsic clinical traits are extracted including basic medical check-up information, individual or family medical history, demographics from DB of qualification, medical check-up and treatment bill. Then, the main factors that affect dementia are transformed into time-series data over 5 years, and data learning and prediction are processed through models such as Random Forest, XGBoost. With the suggested analysis methods and major factors, the result shows a high prediction level of early dementia.

Key words: EHR, dementia, machine learning, Random Forest, XGBoost

Student Number: 1854061015

감사의 글