2019학년도
석사학위논문

머신러닝 기법을 활용한 합병증 예측모형에 관한 연구

:국민건강데이터를 중심으로

A Analysis in Complication Prediction Using Machine Learning Prediction Algorithm

: Focusing on National Health Data

남서울대학교 복지경영대학원

빅데이터·산업보안학과 빅데이터전공

김 신 영

2019년 12월

머신러닝 기법을 활용한 합병증 예측모형에 관한 연구

:국민건강데이터를 중심으로

A Analysis in Complication Prediction Using Machine Learning Prediction Algorithm : Focusing on National Health Data

지도교수 김 성 준

이 논문을 석사학위논문으로 제출함

2019년 12월

남서울대학교 복지경영대학원

빅데이터·산업보안학과 빅데이터전공

김 신 영

김신영의 석사학위논문으로 인준함

심 사 위 원 장 _____ ㊞

심 사 위 원 _____ ㊞

심 사 위 원 _____ ㊞

남서울대학교 복지경영대학원

2019년 12월

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1. 연구의 배경 및 필요성

빠른 속도로 고령화와 만성질환의 증가는 노인의료비 및 만성질환 진료비로 이어지면서 의료비의 가파른 상승을 견인하고 있다. 특히 고혈압, 당뇨병과 같은 만성질환은 질환의 특성상 완치가 어렵고, 유병기간이 길기 때문에 이로 인한 의료비 부담은 앞으로도 더욱 가중될 것으로 보인다. 이에 따라 장기적인 관점에서 국가의 의료비 부담을 줄이고 국민들의 건강수준 향상을 도모하기 위하여 사전 예방적 서비스의 필요성이 대두하게 되었다.

질병예방(prevention)은 크게 1·2·3차로 나뉜다. 일차예방(Primary prevention)은 질병이 없는 건강한 사람들에게 질병이 발생하지 않도록 예방하는 것으로, 금연, 절주, 운동 등의 생활습관 개선에서부터 면역체계 증진을 위한 예방접종, 위생상태 개선, 유해한 환경 노출 방지 등을 포함한다. 이차예방(Secondary prevention)은 질병이 있으나 발견하지 못하는 사람들이 질병의 발생을 조기에 확인하여 악화되는 것을 예방하는 것으로 대표적으로 건강검진이 있다. 셋째는 삼차예방(Tertiary prevention)으로 질병이 발생한 사람들이 합병증 이환 등으로 이어지지 않도록 재활 등을 통해 후유증을 줄이는 노력을 말한다(Leavell & Clark, 1958; Wallace, 2019). 우리나라는 1995년 국민건강증진법과 2008년 건강검진기본법의 제정에 따라 건강증진 및 건강검진에 관한 중장기적 종합계획을 수립하면서 건강검진 및 각종 건강증진사업의 실시 등 1차 및 2차 예방에서 다양한 노력을 기울이고 있다.

건강보장의 패러다임이 사후적인 질병 치료중심에서 사전적인 질

병 예방으로 변화하면서, 보건의료체계 역시 질병중심에서 환자중심의 구조로 변화하였다. 과거에는 임상 정보에 근거하여 의사가 대부분의 치료를 결정했던 반면에, 환자 중심의 구조에서는 환자가 스스로의 치료과정에서 개인적인 요구와 선호도에 맞춘 서비스를 제공받게 되면서 보다 적극적인 환자로서 역할을 하게 된다(Stanton, 2002).

환자 중심에서 건강서비스를 제공함에 있어 효과를 증가시키기 위해서는 개개인의 건강상태에 최적화되어 관여도가 높은 “맞춤형 건강관리”가 필요하다. 소비자 행동이론에 따르면 소비자들의 관여도가 높은 제품일수록 의사결정에 필요한 정보처리의 수준이 높기 때문에, 소비자들은 보다 능동적이고 적극적인 정보탐색을 실시하며, 이는 의료서비스에도 마찬가지로 적용된다(조희숙, 2003). 따라서 관여도가 높은 능동적 소비자들에게 소구하기 위해서는 타인과는 차별화된 정보를 부각시켜야 한다. 의료서비스는 개인의 건강과 관련하여 영향을 미친다는 점에서 지속적인 관여도가 있으며, 지속적인 관여로 인해 나타나는 행동은 일시적 관여로 인해 나타나는 행동보다 더욱 변함이 없고 안정적이기 때문에 개인별로 차별화된 건강관리 서비스를 제공하는 것이 더 효과적일 것이다.

이런 점에서 개인별 맞춤형 건강·질병 위험 예측 모형은 사전 예방적 맞춤형 건강서비스의 일환으로서 질병 발생 이전에 개인의 건강의 위험도를 제공하여 생활습관 변화를 유도하고 질병 발생의 시기를 늦추는 도구적인 의의를 가진다고 할 수 있다.

고혈압의 유병률은 2000년 전 세계 성인인구 약 26.4%에서 2025년 약 29.2%로 증가할 것으로 예측된다. 우리나라에서는 2016년 조사 결과 고혈압 유병률이 29.1%로, 주의혈압과 고혈압 전 단계를 합한 혈압의 유병률은 25.9%이고 30세 이상 성인인구의 55%가 정상혈압보다 높은 것으로 나타났다.

우리나라는 1995년 제정된 국민건강증진법 제4조 ‘국민건강증진종합계획의 수립’에 근거하여 1차 계획(2002~2005)을 수립하면서 국가가

주요하게 관리해야 할 6개 영역 중 하나로 고혈압을 설정하였고, 고혈압에 대한 예방, 교육 및 홍보를 위해 '국민 고혈압 사업단'을 2001년 지정하였으며, 2007년에서 2009년까지 '고혈압 당뇨병 등록관리 시범사업'을 통해 얻을 결과를 바탕으로 2010년 '심뇌혈관질환종합대책'을 발표하였고, 보건소의 '만성질환관리사업'과 '맞춤형 방문 건강관리사업' 등의 정책을 통해 고혈압을 국가적으로 관리하고자 노력해왔다.

또한 당뇨병은 인슐린 분비 혹은 인슐린 작용의 결함으로 인한 고혈당으로 특징지어지는 대사 질환이다. 전 세계적으로 당뇨병은 중요한 문제이며 지난 10년간 꾸준히 증가하였고, 특히 아시아 국가에서 급증하고 있다. 당뇨병 인구 중 85~95%는 제2형 당뇨이며, 이 중 약 80%는 인도, 중국 등의 아시아 국가에 속한다. 서양인과 비교 시 아시아인의 당뇨병 발생 증가율은 높으며 경제성장이 진행되는 향후 20년 동안 증가율은 더욱 가속화 될 것으로 예상된다(Namachandran A. Ma RC. 2010). 실제로 우리나라의 당뇨병 유병률은 2013년 11.9%로 지난 30여년간 약 5배 증가하였으며(국민건강통계, 2014) 당뇨병으로 인한 사망률은 2014년 6위를 차지하였다.

당뇨병은 죽상경화증을 촉진하여 대혈관 합병증(Macrovascular complication)을 야기하며 신부전증, 망막병증, 신경병증 등과 같은 미세혈관 합병증(Micrivascular complications)을 일으킨다. 우리나라도 최근 국민들의 생활습관과 식습관이 서구와 유사하게 되면서 당뇨병의 발생이 증가하고 있으며 이에 다라 당뇨로 인한 합병증의 발생도 증가할 가능성이 커지고 있다(Park IB, Balk SH. 2009). 그럼에도 불구하고 당뇨병과 합병증의 관리의 중요성에 대한 인지도는 낮은 편이다(Lee DW, Park CY, Song SJ. 2011). 당뇨병의 미세혈관 합병증의 하나인 당뇨병성 망막병증은 성인에서 발생하는 저시력 및 실명의 주요한 원인이며, 특히 25~49세 사이의 생산층 인구에서 발생하는 실명의 제1원인으로 알려져 있다. 전 세계적으로는 제2형 당뇨병 환자 중 40%가 당뇨병성 망막병증을 동반하며, 우리나라는 2012년 기준

16.7%가 이환된 것으로 알려져 있다. 우리나라에서 당뇨병성 망막병증이 실명을 일으킬 수 있는 가장 중요한 원인으로 보고 되고 있음에도 불구하고(Lee DW, Park CY, Song SJ. 2011), 당뇨병성 망박병증을 인지하지 못하여 안과 검진을 제대로 받지 않고 있는 환자들이 상당수일 것으로 추정된다. 당뇨병성 망박병증의 진행속도를 늦추기 위해서는 혈당 및 혈압관리가 중요하다. 따라서 당뇨 환자에서 고혈압이 동반되었을 경우 적정 혈압을 유지하기 위하여 혈압강하제를 사용하는 것은 망막병증의 발생 위험을 낮출 수 있을 것으로 기대된다.

고혈압과 당뇨 등 합병증 예측과 대응방안에 대한 연구는 주로 설문 기반의 한정된 자료만을 분석하였고, 대부분 전통적인 통계방법인 회귀분석, 구조모형분석 방법을 사용하였으며, 의사결정 나무(Decision Treee: 이하 DT)의 단일한 방법만을 적용하여 제한적으로 분석한 연구도 있었다. 합병증 발생 가능성이 높은 환자들을 대상으로 한 기계학습(Machine Learning: 이하 ML) 기법을 활용한 연구들을 살펴보면 인구 통계학적 정보와 의료기록 그리고 생활습관을 통한 합병증 예측 연구, 유전체, 영상기록, 의료기록을 통한 알즈하이머 치매 예측 연구 그리고 치매 검사 배터리 점수의 변화와 치매의 위험요소인 biomarker를 통한 예측 연구 등이 있어왔다. 이와 같은 연구들은 치매 노인들을 대상으로 치매의 유형 또는 기간 등 일부 내용만을 다루는 제한적인 연구들이었고, 또한 합병증에 관하여 기계학습 기법을 활용하여 예측 연구 논문은 없었다.

합병증에 관한 다각적인 분석을 위해서는 상당한 노력과 전문성이 요구되기에 이러한 복잡한 분석과정을 자동으로 지원해줄 수 있는 다양한 방법들을 적용할 필요가 있다. 특히 대용량 자료의 분석 및 예측에 용이한 ML 기법을 새롭게 적용하고자 한다. 본 연구를 통해 합병증 발생 영향요인 분석 및 예측을 위한 ML 기법의 적용 가능성과 그에 따른 문제점 및 해결방안을 함께 살펴보고자 한다.

2. 연구 범위 및 방법

연구에 활용된 데이터는 공공데이터자료로 2018년 1년치 데이터로 분석하였다. 해당 데이터들을 분석 예측하기 위해 빅데이터 분석방법을 사용하였다. 이러한 분석방법은 비교적 큰 데이터들을 다루기 쉽게 하고, 예측 및 분류에 다양하게 활용할 수 있는 것에 효과적으로 나타났으며, 분석에 활용된 도구로써 파이썬 프로그램이 활용되었다.

1992년~1995년 건강검진 수검자들을 대상으로 개발되었던 기존의 모형을 적용하는데는 한계가 있기 때문에, 새로운 건강검진 공공데이터를 활용하여 현 시점에도 여전히 한국인들의 당뇨합병증 위험도를 예측하는 예측모델을 구성하고자 한다. 이를 기반으로 개인정보, 건강정보를 활용한 머신러닝 모델을 웹서비스로 제공하여 실시간 건강관리에 활용할수 있으며 더 나아가 주요 질병 위험도 산출 및 합병증 예측을 통하여 근거리 위치의 병원 추천 시스템 개발하는데 있어서 근거 예측모형으로 제안하고자 한다. 이를 통하여 향후 효과적인 당뇨합병증 위험도 예측 서비스를 제공하기 위한 정책적 제언 등이 가능할 것이다. 또한 당뇨합병증 뿐만 아니라 향후 다양한 질환들을 대상으로 맞춤형 건강·질병 위험 예측 프로그램을 제공하는 과정에서도 시사점을 줄 것으로 기대한다.

Ⅱ. 관련 연구 및 이론적 배경

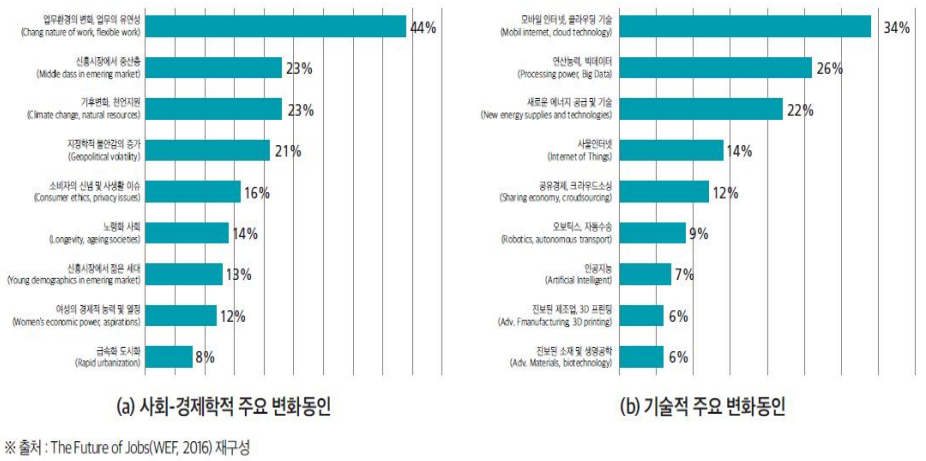
1. 4차산업혁명과 공공데이터

1) 4차 산업혁명

(1) 제4차 산업혁명의 주요 변화동인

우선 제4차 산업혁명의 특성을 찾기 위해 제4차 산업혁명을 일으키는 원인을 살펴보고자 한다. 많은 미래 전망 보고서들은 제4차 산업혁명과 미래사회 변화가 기술적 측면의 변화동인과 사회·경제적 측면의 변화동인으로 인해 야기될 것으로 전망하고 있다. 특히 「The Future of Jobs(WEF, 2016)」는 '업무환경 및 방식의 변화', '신흥시장에서의 중산층 등장' 및 '기후변화' 등이 사회·경제적 측면에서의 주요 변화동인이고, 과학기술적 측면에서는 '모바일 인터넷', '클라우드 기술', '빅데이터', '사물인터넷(IoT)' 및 '인공지능(A.I.)' 등의 기술이 주요 변화동인이 될 것으로 보고 있다.

보스톤 컨설팅(Boston Consulting Group), 옥스퍼드 대학(Oxford Univ.) 및 CEDA(Canadian Engineering Development Association) 등 주요 컨설팅 기업, 대학 및 연구기관들도 미래사회의 변화동인과 미래사회 변화에 대한 연구를 수행하여 다음와 같은 결과를 제시하고 있다. Boston Consulting은 「인더스트리 4.0(Industry 4.0)」에 기반하여 독일 제조업 분야에서 나타나는 노동시장의 변화를 연구하였는데, 기술적 측면의 변화동인들이 일자리 지형에 직접적인 영향을 미쳐 기술발전을 적용(adoption)함으로써 제조업 생산성이 크게 향상될 것으로 전망하고 있다. 그리고 이러한 변화의 중심에는 빅데이터, 로봇 및 자동화 등의 기술이 자리할 것으로 예측하고 있다.



※출처 : The Future of Jobs(WEF, 2016) 재구성

[그림 1] 제4차 산업혁명의 주요 변화동인

옥스퍼드 대학(Oxford Univ.)의 Martin School은 유럽에서의 미래 일자리 지형 변화를 연구하였는데, 유럽 노동시장이 '글로벌화'와 '기술적 혁신'으로 인해 변화될 것으로 전망하고 있다. 또한 과학기술의 발전이 단순 업무에서부터 복잡한 업무까지 자동화시켜 일자리뿐만 아니라 업무영역에서도 커다란 변화가 나타날 것으로 전망하고 있다. 특히 S/W 및 빅데이터 등 정보통신기술(ICT)의 발달로 업무영역이 자동화되고, 자율주행기술 및 3D 프린팅 기술 등의 등장으로 일자리 지형이 크게 변화할 것으로 예측하고 있다.

CEDA는 호주 노동시장의 미래 변화에 대한 연구를 수행하였는데, 과학기술적 측면과 과학기술 외적 측면에서의 변화동인을 제시하고 있다. 과학기술 외적으로는 글로벌화, 인구통계학적 변화, 사회변화 및 에너지 부족 등이 변화동인으로 제시되었고, 과학기술적 측면에서는 클라우드 서비스, 사물인터넷(IoT), 빅데이터, 인공지능 및 로봇기술 등이 변화동인으로 제시되고 있다. 또한 세계적 민간기업인 제너럴일렉트릭(GE, General Electronics Corp.)는 미래 공급체인의 발전과 고

객 니즈 충족과 관련된 기술을 연구하였는데 다양한 과학기술의 보고서는 다양한 과학기술의 발달이 기업의 공급체인을 더울 발전시키고 고객의 다양한 요구를 충족시켜 경제규모를 더욱 크게 만들 것으로 전망하고 있다. 특히 클라우드, 자동화 기술, 예측 분석 및 선행제어를 위한 스마트 시스템 등의 기술이 미래에 생산성을 높일 기술로 제시되고, 기계 센서와 커뮤니케이션 기술, 3D 프린팅 기술 등은 고객의 니즈를 충족시킬 수 있는 기술이 될 것으로 예측하고 있다.

이러한 다양한 미래 전망자료를 종합·분석해보면, 과학기술 측면에서 제4차 산업혁명과 미래사회 변화를 야기하는 주요 변화동인이 ICBM 등 정보통신기술(ICT) 기반의 기술임을 알 수가 있다. 이를 바탕으로 우리는 제4차 산업혁명이 가지고 있는 특성을 이해할 수 있을 것이다.

(2) 제4차 산업혁명의 특징

제4차 산업혁명은 '초연결성(Hyper-Connected)', '초지능화(Hyper-Intelligent)의 특성을 가지고 있고, 이를 통해 "모든 것이 상호 연결되고 보다 지능화된 사회로 변화"시킬 것이다.

우리 사회는 이미 초연결 사회로 진입하고 있다. 사물인터넷(IoT), 클라우드 등 정보통신기술(ICT)의 급진적 발전과 확산은 인간과 인간, 인간과 사물, 사물과 사물 간의 연결성을 기하급수적으로 확대시키고 있고, 이를 통해 '초연결성'이 강화되고 있다. 2020년까지 인터넷 플랫폼 가입자가 30억 명에 이를 것이고 500억 개의 스마트 디바이스로 인해 상호 간 네워크킹이 강화될 것이라는 전망은 초연결사회로의 진입을 암시하고 있다. 또한 인터넷과 연결된 사물(Internet-connected objsects)의 수가 2015년 182억 개에서 2020년 501억 개로 증가하고, M2M(Machine to Machine, 사물-사물) 시장규모도 2015년 5조2000억 원에서 2020년 16조5000억 원 규모로 성장

할 것으로 전망되고 있다. 이러한 시장 전망은 '초연결성'이 제4차 산업혁명이 도래하는 미래사회에서 가장 중요한 특성임을 보여주고 있다.

또한 제4차 산업혁명은 '초지능화'라는 특성이 존재한다. 즉 제4차 산업혁명의 주요 변화동인인 인공지능(A.I.)과 빅데이터의 연계 및 융합으로 인해 기술 및 산업구조가 '초지능화' 된다는 것이다. 2016년 3월 이미 우리는 '초지능화' 사회로 진입하고 있음을 경험하였다. 인간 '이세돌'과 인공지능 컴퓨터 '알파고(Alphago)'와의 바둑 대결이 그것이다. 바둑판 위의 수많은 경우의 수 7 와 인간의 직관 등을 고려할 때 인간이 우세할 것이라는 전망과 달리 '알파고'의 승리는 사람들에게 충격으로 다가왔다. 이 대결은 '초지능화' 사회의 시작을 알리는 단초가 되었고, 많은 사람들이 인공지능과 미래사회 변화에 대해 관심을 갖기 시작했다. 사실 2011년에도 이미 인공지능과 인간과의 대결이 있었다. 미국 ABC 방송국의 인기 퀴즈쇼인 '제퍼디!(Jeopardy!)'에서 인간과 IBM의 인공지능 컴퓨터 왓슨(Watson)과의 퀴즈대결이 있었는데, 최종 라운드에서 왓슨은 인간을 압도적인 차이로 따돌리며 우승하였다. 이 대결은 인공지능 컴퓨터가 계산도구에서 벗어나 인간의 언어로 된 질문을 이해하고 해답을 도출하는 수준까지 도달했음을 보여주는 사례로 회자되고 있다.

산업시장에서도 딥 러닝(Deep Learning) 등 기계학습과 빅데이터에 기반한 인공지능과 관련된 시장이 급성장할 것으로 전망되고 있다. 트렉티카 보고서에 따르면 인공지능 시스템 시장은 2015년 2억 달러 수준에서 2024년 111억 달러 수준으로 급성장할 것으로 예측되고 있고, 인공지능이 탑재된 스마트 머신의 시장 규모가 2024년 412억 달러 규모가 될 것으로 보고 있다(BCC Research, 2014). 이러한 기술발전 속도와 시장성장 규모는 '초지능화'가 제4차 산업혁명 시대의 또 하나의 특성이라는 점을 말해주고 있다.

지금까지 우리는 제4차 산업혁명의 주요 변화 동인을 살펴보았고,

‘초연결성’과 ‘초지능화’라는 제4차 산업혁명의 특성을 이해하였다. 이제는 이러한 특성을 통해 미래사회가 어떻게 변화할 것인지에 대해 살펴볼 필요가 있다. 미래사회 변화의 방향에 대한 분석함으로써 우리는 보다 합리적이고 우리나라 현실에 맞는 대응 방안을 모색할 수 있을 것이다.

2) 공공데이터

(1) 빅데이터 정책 연구동향

공공데이터는 정부 또는 공공기관의 업무를 통해 생성, 축적, 보유하고 있는 데이터를 말한다. 이와 같은 공공데이터는 민간데이터와 비교할 때, 데이터의 목적성이 명확하고 정형화되어 있다는 점에서 상대적으로 가치가 높다. 또한 생성된 목적 이외에도 이종 데이터와의 교차, 연계를 통해 새로운 효용가치의 가능성을 가지고 있다. 미국, 영국 등 선진국에서는 공공 빅데이터의 활용을 정부 정책으로 추진하여 대민 서비스를 향상시키고자 노력하고 있다.

공공데이터 활용을 위해서는 데이터의 공개에 대한 부처 간 장벽을 극복해야 한다. 부처 간 장벽의 극복을 위해서는 최고 책임자의 정책에 대한 의지가 가장 중요하다. 미국은 2009년 비벡쿤드라를 미 연방정부의 CIO(Chief Information Officer)로 임명하여, 정부의 정보를 공개 하였다. 그 결과 정부의 정보를 공개하는 웹사이트인 ‘data.gov′를 구축 하였다. 영국 역시 같은 시기에 팀 버너스리를 책임자로 임명하여 ’data.gov.uk′를 개발했다. 영국은 미국보다 늦었지만 팀 버너스리의 웹 정보 공개에 대한 확고한 방향성으로 미국보다 앞선 수준의 공공정보를 제공하고 있다. 정부 보유의 데이터는 각종 정책을 개발하는데 활용 하는데 목적을 두고 있으며, 이를 통해 정부는 대민 서비스를 수행한다. 공공 빅데이터를 활용한 국외 서비스 사례

로는 미국의 'IT Dashboard', 영국의 'Where does my money go?', OECD의 'Better Life Initiative' 등이 있다.

2013년 우리 정부는 「공공데이터의 제공 및 이용 활성화에 관한 법률(이하 공공데이터법)」을 제정과 동시에 공공정보의 적극적 개방을 통한 신뢰 정부 구현 및 신성장동력 창출을 주 골자로 하는 정부 3.0 비전을 선포하였다.

공공데이터란 공익적 업무를 수행하는 기관들이 업무를 수행하면서 축적한 데이터를 의미하며, 공공데이터는 해당 업무에 관한 상세정보를 담고 있다. 또한 정부3.0의 개방은 정부기관이 사용자에게 정보를 재활용 가능한 형태로 제공하고, 사용자에게 정보를 영리적 및 비영리적으로 재가공 및 활용할 권한을 부여하는 것이다.

공공데이터법은 "공공기관이 이용자로 하여금 기계 판독이 가능한 형태의 공공데이터에 접금할 수 있게 하거나 이를 다양한 방식으로 전달하는 것"이라고 공공데이터의 제공에 대해 명시하고 있다. 이때 "기계 판독이 가능한 형태"란 "소프트웨어로 데이터의 개별 내용 또는 내부구조를 확인하거나 수정, 변환, 추출 등 가공할 수 있는 상태"를 의미한다. 즉, 불특정 다수에게 정보를 제공하여, 불법적 사용을 제외한 모든 사용 목적을 불문하고, 영리적 및 비영리적 목적은 물론 재이용에 관한 권한도 부여하는 것을 의미한다. 이때 제공되는 공고데이터의 종류는 국가안보 및 개인정보 등의 민감정보를 제외한 공공기관이 생성하고 관리하는 모든 공공데이터가 대상이 된다. 또한 정부 3.0의 개방은 정부기관이 사용자에게 정보를 재활용 가능한 형태로 제공하고, 사용자에게 정보를 영리적 및 비영리적으로 재가공 및 활용할 권한을 부여하는 것이다.

공공데이터법은 "공공기관이 이용자로 하여금 기계 판독이 가능한 형태의 공공데이터에 접근할 수 있게 하거나 이를 다양한 방식으로 전달하는 것"이라고 공공데이터의 제공에 대해 명시하고 있다. 이때 "기계 판독이 가능한 형태"란 "소프트웨어로 데이터의 개별내용

또는 내부구조를 확인하거나 수정, 변환, 추출 등 가공할 수 있는 상태"를 의미한다. 즉, 불특정 다수에게 정보를 세공하여, 불법적 사용을 제외한 모든 사용 목적을 불문하고, 영리적 및 비영리적 목적은 물론 재이용에 관한 권한도 부여하는 것을 의미한다. 이때 제공되는 공공데이터의 종류는 국가안보 및 개인정보 등의 민감정보를 제외한 공공기관이 생성 하고 관리하는 모든 공공데이터가 대상이 된다. 또한 이용자는 신의 성실의 의무에 따라 공공데이터를 이용하여야 하며, 법령과 이용조건에 따른 의무를 준수하여 국가안전보장 등 공익이나 타인의 권리를 침해 하지 않아야 한다. 정부는 관련법에 근거하여 공공데이터 전략위원회 및 분쟁조정위원회를 설치하였고, 공공데이터 개방을 확대하는 정책을 추진해오고 있다. 그 결과, 2017년 현재 데이터 개방 순위를 2014년 12위에서 5위로 끌어올렸고, 활용서비스 수도 30여개에서 2,000여개로 확장하였다. 그러나 양적확대에도 불구하고 산업적 활용도는 선진국 대비 낮은 수준으로 평가되고 있다.

따라서 발전전략으로서 '국가중점개방데이터' 발굴을 추진하였다. '국가중점개방데이터'는 전국단위 행정데이터 중 지방행정, 교육 등 분야별 공통시스템을 통해 관리되고 있는 공공데이터로 일정수준의 품질을 갖춘 대용량데이터를 중심으로 일상생활과 밀접하고, 비즈니스에 즉시 활용 가능한 공공데이터가 선정되었다. 이중 보건복지부, 건강보험심사평가원 등으로부터 제공되는 보건, 복지 데이터군은 활용도를 높이기 위해 융·복합 서비스가 가능한 연관데이터로서 제시되고 있다. 이와 같은 국내외적 정책적 환경과 더불어 공공데이터의 효율적 활용을 위해 우리나라에서도 '공공데이터포털 (data. go.kr)'을 구축하였다. 우리나라 정부는 '공공데이터포털(data. go.kr)'의 개발, 운영을 통해 국가가 보유하고 있는 공공빅데이터를 국민에게 개방함은 물론 이를 편리하고 효율적으로 이용할 수 있도록 하고 있다. 이를 통해 대민 서비스와 대국민 신뢰를 향상시키고 공공빅데이터의 민간 활용 지원을 통한 일자리 창출을 하고자 한다. 모든 정보는 원자료 공개를

원칙으로 하고 비공개 범위를 최소화함으로써 민간에 개방하여 효용 가치를 극대화하고 있다. 우리 정부는 모바일 앱(application)과 소프트웨어를 개발하여 공공 빅데이터에 민간 접근성을 높였고, 공공 빅데이터를 개방하고 공유하는 공통적 기반을 조성하였다. 그리고 공공 빅데이터 산학연 협력 촉진을 위한 '국가오픈데이터포럼'을 출범하였으며, 공공 빅데이터를 일괄 관리, 등록, 제공하는 범정부 단일 정보 공유 플랫폼 구축을 위해 데이터베이스를 개방형 및 표준형으로 전환하였다. 또한 1인 창조기업, 사회적 기업 등을 대상으로 공공 빅데이터의 활용 지원과 공공 빅데이터간 융합으로 새로운 비즈니스 모델을 발굴하는 등의 공공빅데이터 민간 활용을 통해 양질의 일자리를 창출하고자 한다. 특히 새로운 부가가치를 창출하기 위한 사업의 일환으로 모바일 앱의 개발과 정보서머스 수집, 분석, 판매 및 가공 데이터의 전략적 활용과 서비스 융합을 수행 중이다. 공공데이터 포털에서 제공하고 있는 주요 서비스는 (1) 데이터 및 API (Application Program Interface) 공유자원의 검색 및 활용 지원, (2) 정부부 처 및 산하기관에서 발행하는 백서를 포괄적으로 제공하는 '공공백서'서비스, (3) 데마별 전문가가 선정한 최신 이슈 관련 지식모음인 데마정보, (4)지역별, 분류체계별, 제공기관별, 활용방법별로 분류된 공공데이터 개방현황 정보, (5)개방 가능한 공공정보에 대해 제공기관에 제공 신청 또는 중계신청, (6) 사용자들의 문의사항에 대해 전문적인 대응을 위한 전문 컨설팅 및 1:1 상담, 그리고 (7) 개발자들을 위한 개발가이드의 제공 등이 있다. 공공데이터 포털의 운영현황을 살펴보면, 중앙 및 지자체 등 8백여 개의 공공기관에 대해 약 2만여 종의 데이터가 개방되었는데, 약 1만4천여개가 넘는 공공데이터셋과 1천 9백여종이 넘는 API가 제공되고 있다. 그 중에서도 보건의료 관련 데이터는 약 1,559 종이 있다.

(2) 보건의료 정책에 대한 빅데이터 연구동향

보건복지부, 국민건강보험공단, 건강보협심사평가원은 2013년 정부3.0 성포에 따라 보건의료 빅데이터의 민간개방을 하였으며, 이를 통해 국민에게 개별적인 맞춤형 보건의료 서비스를 제공하고자 노력하고 있다. 특히 국민건강보험공단과 건강보협심사평가원은 지난 10여년이 넘는 기간 동안 다양한 국민건강정보가 축적되었으며, 빅데이터 활용과 관련한 자체 세미나를 진행하는 등 적극적으로 사업들을 추진하고 있다. 국민건강보협공단의 경우, 건강보험 정보시스템을 통해 1조 8천억건이 넘는 방대한 빅데이터를 국민건강정보 DB로 구축하고 있으며, 국내 법정의무보험자로서 전 국민의 자격, 보험료 부과·징수, 진료비 지급, 건강검진 및 노인요양보험 등의 다양한 사업을 최첨단 ICT기반 위에 수행하고 었다. 국민건강보험공단은 2014년 7월부터 정책 및 학술 연구를 위해 보건의료 분야 빅데이터인 표본코호트 DB를 제공하고 있으며 정부 부처 및 연구기관, 학회, 대학 등에 제공되어 연구가 진행되고 있다.

또한 식품의약품안전처, 기상청, 국립환경과학원과 협업을 통해'국민건강 알림서비스'를 제공하고 있다. '국민건강 알림서비스'는 빅데이터 분석기술을 기반으로 감기, 눈병, 피부염, 식중독 등 주요 유행성 질병에 대한 지역별 위험도와 행동수칙을 질병 위험 징후 시 단계별로 알람을 제공하며, 천식질환의 경우 일교차, 최저기온, 이산화황, 미세먼지, 오존 등의 측정값을 질병 예측모델에 반영하여 사전에 알람도 제공한다.

신뢰성 높은 알람서비스를 세공하기 위해 국민건강보협공단 빅데이 터를 활용한 한국인의 질병 현황· 예방·관리에 대한 공동 연구를 여러 보건의료 관련 학회들과 수행하여 정확도를 향상시키는 작업을 진행하고 있다. 보건의료 빅데이터개방시스템을 통해 건강보협심사평가원(이하 심평원)은 심평원이 보유하고 있는 다양한 공공빅데이터를

개방하고 있다. 보건의료 빅데이터개방시스템에서 제공하는 서비스는 공공데이터 (데이터셋, 오픈 API)의 제공, 의료빅데이터 분석 (빅데이터센터), 의료통계 분석(질병, 의약품, 의료기관)등이 있다. 이러한 공공빅데이터 개방을 위한 심평원의 노력을 통해 국민과 보건의료 산업분야 그리고 의료연구기관 등이 원하는 다양한 의료정보와 서비스를 제공하고 있으며, 심평원은 이와 동시에 다양한 자료원을 지속적으로 발굴하고 있다.

주요 서비스에 대해 자세히 살펴보면, 먼저 (1) 보건의료 관련 공공 데이터 파일 다운로드 및 신청, (2) 0pen API 검색 및 활용 신청 등이 있다. 또한 다양한 의료통계 분석 서비스를 세공하고 있는데 (3) 시각화 조희 서비스를 통해 다양한 보건의료 통계 및 OECD 보건통계를 쉽게 이해할 수 있게 해주며, (4) 통계분석 환경 지원을 통해 대학, 산업체, 전문 연구기관에게 국민 건강 증진, 의료정보 산업 활성화 등을 위한 연구를 도와주며, (5) 통계분석 서비스를 세공하여 대용량 데이터를 빠르게 탐색하고, 다양한 시각적인 방법으로 데이터의 숨어있는 패턴을 찾게 해준다. 그리고 (6) SAS를 활용하여 대용량 데이터 통계분석 수행, (7) 웹기반의 R Studio를 제공하여 심평원의 R분석 서버를 통한 분석환경 제공. 마지막으로 (8) 개선의견 및 이용자들의 문의사항에 대한 답변 등의 서비스가 세공되고 있다.

보건의료 빅데이터 개방시스템에서 제공하고 있는 공공데이터의 종류 및 주요 항목들은 다음과 같다. 대한민국 국민의 국내에서 수행된 진료 정보 및 의료기관, 제약희사, 유관기관 등 다양한 경로를 통해 수집된 정보를 분석 및 정제한 데이터로서, (1) 8만 7천여 의료기관에서 제출된 청구자료 기반의 자료인 의료기관 방문 환자들의 상병, 수술, 처치, 의약품 처방 및 조세 등의 데이터와 의약품에 대한 데이터들 (2) 의약 품 관련 생산, 수입, 사용 등 의약품 유통 정보, (3) 의약품 인·허가 정보, 부작용 둥 의약품 안전정보, (4) 마약류 등 집중관리 의약품 융합 데이터. 그리고 의료자원 데이터인 (4) 의료기관

의 인력, 시설, 장비 관련 정보, (5) 의료기기 정보, 의료처치용 치료재료 정보 등으로 구성되어 있다. 국민건강보협공단에서는 앞에서 소개한 국민건강정보 DB를 바탕으로 '국민건강보험자료 공유서비스' 홈페이지를 통해 전 국민 의무보험인 건강보험 자료를 연구자와 정책 기획자들에게 제공하고 있다. 앞서 살펴본 바와 같이 우리나라의 건강보협제도는 대한민국 국적을 가진 내국민을 대상으로 법에 의해 가입이 강제되는 의무보험으로 국민건강 보험공단은 전 국민의 의료기관 및 약국 사용에 관련된 모든 정보를 저장 및 관리하고 있다. 이와 같은 국민건강보협공단의 공공 빅데이터 개방은 국민의 의료이용 자료에 대한 효용성을 제고하게 되었고, 국민건강보험공단은 건강보험 자료를 가공하여 공개하는 서비스를 제공하고 있다.

'국민건강보험자료 공유서비스'는 보건의료 정책 및 학술연구를 지원하고, 연구자들에게 국민건강정보자료를 제공하고 연구 성과를 공유할 수 있도록 제반 업무의 편의를 제공하고 있다. '국민건강보험자료 공유서비스'를 통해 제공하는 공공 빅데이터의 종류는 표본코호트DB, 맞춤형DB, 건강검진코호트DB, 직장여성검진 코호트DB, 영유아김진코호트DB, 노인코 호트DB, 건강질병지표 등이 있으며, 보건의료 정책 및 학술연구에 근거자료로 사용팁은 물론, 사회, 경제, 환경, 산업 등의 다양한 분야에서 공공 빅데이터가 사용되고 있다.

국민건강보험자료 공유서비스에서 제공하고 있는 자료는 크게 두 가지로서 '표본DB'와 '맞춤형DB'로 구분할 수 있다. 먼저 표본DB는 국민건강보험공단에서 일정한 목적에 맞추어 연구자들에게 제공하기 위해 제작한 DB이다. 국민건강보험공단에 저장된 전 국민 의료이용 자료에는 전자의무기록의 특성상 개인정보에 대한 민감한 자료가 포함되어 있으며, 일반적인 사용자가 데이터를 분석하기에는 데이터의 크기때문에 물리적으로 접근 및 처리가 어려웠다. 이러한 사용자의 접근성을 가로막는 여러 가지 문제집들을 해결하기 위해 제작된 DB가 '표본DB'이다. '표본DB'는 개인정보 보호문제에서 자유롭고 연구

자들이 쉽게 사용할 수 있으며, 원자료(raw database)에 준하는 결과물들을 얻을 수 있도록 제작 되었다. 국민건강보협공단에서는 건강보험 및 의료 급여권자에 진료명세서와 진료내역, 상병내역, 처방전내역 등을 2002년부터 2010년 기간 동안 코호트 방식으로 추출하여 '표본DB'의 하나인 표본 코호트DB를 구축하였다.

'표본DB'는 2002년부터 2010년까지 우리나라의 건강보험 가입자 건 체에 대한 진료명세서와 진료내역, 상병내역, 처방전내역 등을 포함하 고 있다. 각각의 '표본DB'는 연구 대상자의 모집단에 대한 대표성을 확보하는 전국민 대상 표본코호트DB, 건강검진코호트DB, 노인코호트 DB, 직장여성코 호트DB, 영유아검 진코호트DB로 구성되어있다. 이중 표본코호트DB는 2002년을 기준으로 모집단에 대한 대표성을 유지하도룩 비례배분(proportional alocation)에 의한 층화무작위추출법을 통해 추출 되었으며, 전국민 의료이용 정보를 통해 성별, 연령 등의 인구학적 특성과 소득분위에 대해 고려하였다. 자격 대상자 약 100만명이 표본으로 추출된 자료로서 '표본DB' 자료 중 가장 대표적인 자료이다. 다음으로 국민건강보험자료 공유서비스는 '맞춤형DB'를 제공하고 있다. '맞춤형DB'란 정책 및 학술 연구목적으로 이용할 수 있도록 국민 건강보험공단이 수집, 보유, 관리하는 건강정보자료를 수요맞춤형 자료로 가공하여 DB형태로 제공하는 것이다.

'표본DB'와 제공되는 자료의 주요항목은 동일하지만, '표본DB'와의 차이점은 맞춤형 자료를 신청할 경우 신청 조건에 따라 자료의 활용 목적에 부합하는 자료만 제공한다는 점이다. 또한 이때 제공되는 자료는 신청 조건에 따른 전수 자료가 제공된다. 예를 들어, 2013년 기준 서울특별시의 치매 유병률 산출에 대한 연구를 수행한다면, 노인코호트 DB의 경우 2002년에 표본 추출된 55만명의 코호트 자료에서 2013년에 해당되는 치매 환자의 자료를 다시 추출하게 되어 대표성 문제와 표본 크기의 문제가 발생한다. 하지만 '맞춤형DB' 자료를 이용할 경우 2013년 서울특별시 인구의 모든 자료가 제공되기 때문에

노인코호트 DB 자료를 사용했을 때의 문제들이 해결될 수 있다. 또한 '표본DB' 자료와 '맞춤형DB' 자료의 중요한 차이점 중 하나는 '표본DB'의 경우 원자료를 전자 파일의 형태로 연구자에게 제공하는 반면, '맞춤형DB'는 원자료의 외부 반출이 금지되어 있어 연구자가 원자료의 열람 및 분석을 위에서 살펴본 보건의료 공공 빅데이터의 제공 환경에 도움으로 국민건강보험공단과 심평원에서 제공하는 보건의료 공공 빅데이터들로 질병에 관한 연구가 활발히 수행되고 있다.

2. 개인별 맞춤형 건강·질병 위험 예측 모형

1) 개인별 맞춤형 건강관리

개인별 맞춤형 건광관리는 보건의료의 패러다임이 질병치료에서 사전 예방적 서비스로, 공급자인 의료인 중심에서 수요자인 환자 중심으로 변화함에 따라 환자의 상태에 적합하게 차별화된 메시지를 제공하여 생활습관 개선을 유도하는 건강관리의 한 방식이다.

맞춤형 건강관리는 그 일차적인 목적이 질병의 발생을 사전에 예방하고 건강한 삶을 영위할 수 있도록 지원하는 것에 있기 때문에 건강증진활동의 일부로 여겨지기도 하고, 의학적인 처치의 효과를 높이기 위한 환자 교육의 수단으로 논의되기도 하는 등 그것이 사용되는 맥락과 배경, 목적에 따라 상당히 광범위하게 통용되고 있다. 또한 실실적인 서비스 제공의 관점에서 접근하는 경우가 많아 학문적으로 통일된 개념적 정의는 다소 부족하다.

애초에 건강관리 서비스란 그 안에 서비스 제공자, 대상자, 내용을 모두 포함하고 있기 때문에 정의를 내리기가 매루 어렵다는 지적이 있다(이정찬, 2010)는 개인 맞춤형 건강관리 서비스를 "전담 헬스매니저의 도움을 받아 상시 건강 상담을 받고, 개인의 의료정보를 관리하

며, 건강에 대한 정보를 받고, 개인별 맞춤 건강관리 지침을 제공받는 일체의 서비스"라고 정의하였으며, 조경희 등(2014)은 "환자의 과거력, 유전적 성향, 특이 생체지표에 따라 개별화된 진단 및 치료를 제공하는 것으로 개인의 의료정보를 바탕으로 건강에 대한 정보를 제공받고, 개인별 맞춤 건강관리 지침을 제공받는 일체의 서비스"라고 정의하였다. 이러한 점에 비추어 보았을 때 개인별 맞춤형 건강관리를 위해서는 먼저 개인의 건강정보에 기반하여 현재 건강 상태를 정확하게 진단하고, 향후 발생할 수 있는 건강관련 문제를 예방하기 위하여 적절한 노력을 유도하는 개입(처방, 처치, 생활습관 개선 등)이 필요하다. 개인별 맞춤형 건강·질병 위험 예측모형은 이러한 건강관리를 가능하게 해주는 기술적·역학적 방법론을 제공한다.

Chawla & Davis(2013)는 맞춤형 건강관리의 다음 단계가 다양하게 수집된 데이터를 통합하여 활용할 수 있는 기술적 구조를 개발하여, 단순히 환자 각자의 의무기록에서 얻어진 정보뿐만 아니라 다른 환자들과의 집단적 유사성, 관련성을 이해하여 도출된 정보들을 바탕으로 개별적 질병 위험의 프로파일을 제공하는 것이라고 주장하였다. 이처럼 개인별 건강·질병 위험 예측모형은 현재의 건강위험 요인을 바탕으로 질병 및 사망의 위험도를 예측하여 개인의 위험수준별 맞춤형 건강정보를 제공한다(조비룡 등, 2007). 실제로 해외에서는 개인별 건강수준에 따라 맞춰진 정보를 제공했을 때 질병이 더 효과적으로 관리된다는 연구가 있으며, 이렇게 개발된 심혈관질환 예측모형의 사용을 공식적으로 권고하는 등 정책적 실시를 위한 중요한 방법론으로 역할하고 있다(고민정 등, 2009).

현재까지 의학·역학 등 다양한 분야에서 개인의 건강정보를 바탕으로 현재의 건강상태를 진단하고 향후의 질병 발생의 위험도를 예측하려는 연구가 진행되어 왔다. 사망률에 기초하여 개인의 건강위험을 평가하는 전통적인 방식인 건강위험평가에서부터, 수십 년간의 코호트자료를 기반으로 다양한 질병의 위험인자를 찾아내어 발생위험을

예측하는프레이밍햄 연구(Framingham study)까지, 현재 가지고 있는 개인의 건강 위험 요인이 미래의 질병 발생에 영향을 미치는 긩위험 정도를 파악하는 질환 예측모형으로 확대·연구되고 있다(신호철, 2004).

2) 건강위험평가(Health Risk Appraisal)

건강위험평가(Health risk appraisal, HRA)는 “개인이 현재 갖고 있는 건강위험요인을 파악하고, 이를 바탕으로 각 개인의 미래 질병 발생 및 사망위험도를 예측 또는 평가하는 방법”, 또는 “특정한 위험 요인들을 갖고 있는 사람이 일정한 기간 이내에 특정한 질환으로 사망할 위험의 정도를 비교하는 기법” 등으로 정의할 수 있다(강임옥 등, 2006). 건강위험평가는 1940년대 후반 자궁경부암과 심장질환의 위험을 예방하기 위하여 Lewis C. Robbins이 환자별 건강위험을 기록하는 것에서 시작되었다. 이후 1968년 Sadusk와 Robbins는 생활습관, 가족력, 환경인자 등을 기초로 개인의 사망 위험도를 평가하는 개념을 제안하였고(신호철, 2004), 1970년 Jackson Hall과 함께 “How to practice rospective medicine’이란 책에서 건강위험평가의 구체적인 방법론을 최초로 제시하면서 본격적으로 연구가 시작되었다(강임옥 등, 2006).

통상적으로 건강위험평가는 설문지 등을 통해 인구사회학적 특성, 생활습관, 개인과 가족의 질병력 등에 대한 정보 및 생리적 특성에 대한 정보를 수집하고, 이를 기반으로 위험요인을 규명하여, 개인별 피드백을 제공하는 세 가지 요소로 구성된다(Anderson & Staufacker, 1996). 이 방법은 먼저 성별, 연령군 별 주요 사망원인 질환에 의한 10년 이내 사망확률 추정자료(평균위험도)가 필요하다. 다음으로 개인의 유전적, 환경적 요인 및 생활습관을 바탕으로 건강행태를 평가하여 위험요인에 대한 위험도를 계산하고 이를 평균위험도와 비교하여

향후 10년 이내 해당 질환으로 사망확률을 산출한다. 이 방식으로 주요 사망원인 질환들의 향후 10년인 사망확률을 모두 구한 다음 합산하여 총 사망 확률을 구한다(신호철, 2004 재인용).

우리나라에서는 2002년 대한가정의학회에서 건강나이와 질병으로 인한 사망확률을 제공하는 '한국형 건강위험 평가도구'를 개발하였고, 이를 바탕으로 2005년에 각 질병에 대한 위험요인들의 상대위험도를 합산하여 개인별 질병으로 인한 사망위험도를 산출하는 건강위험평가 도구가 고안되었다. 이 도구와 건강검진결과를 활용하여 건강iN 사이트에서 '건강나이 알아보기'라는 서비스 명칭으로 개인별 건강위험평가를 제공하고 있으며, 국가건강검진 후 제공되는 검진결과통보서 상에 뇌졸중(뇌경색), 협심증/심근경색, 혈관성치매, 위암, 대장암, 유방암에 대한 발생위험도를 제공하고 있다(조비룡 등, 2007). 또한 2009년부터 개편된 일반건강검진 및 생애전환기 건강진단 건강검진 결과통보서에서도 활용되면서 국내에서 개발된 건강·질병 위험 예측모형 중 가장 폭넓게 이용되고 있다(고민정 등, 2009).

3. 국내외 질병 예측 사례 및 한계

해외의 미국, 영국, 캐나다, 일본 등에서는 감염병 예방 연구에 있어서 빅데이터 분석을 기반으로 하여 많은 노력을 해왔다.

구 분	내 용
미국 국립보건원(NIH), 필박스(pill box) 프로젝트	의약품 정보를 제공하고 검색을 통해 얻어지는 사용자 데이터를 분석해 유행하는 질병, 전염 속도, 질병의 지역별 분포에 대한 통계를 수집·예측
미국 구글, 독감트렌드 (flu trend)	사람들이 독감에 걸렸을 때 검색하는 약 40가지의 단어를 바탕으로 독감의 발병을 예측하여, 독감 환자의 분포 및 확산 정보 제공
영국 NHS(National Health Service), 처방 데이터 수집·분석	전국 약국, 병원의 처방 데이터를 수집·분석하여 질병을 예측하며, CPRD(Clinical Practice Research Datalink)를 통해 다양한 데이터를 연구자에게 제공
영국 Horizon Scanning Center, 전염병 대응책 마련	동식물 및 인간의 전염병 확산에 대한 데이터 분석으로 말라리아 등 다양한 전염병에 대한 전망과 대응방안을 모색
캐나다 온타리오 공과대병원, 미숙아 모니터링	혈압, 체온, 심전도 등 미숙아 모니터링 장비에서 생성되는 데이터를 실시간 분석하여 이상 징후를 사전에 파악하는 등 미숙아 감염 예방 및 예측
싱가포르 RAHS(Risk Assessment & Horizon Scanning), 조류독감 시뮬레이션	전염병 등 국가 위험 정보를 수집·분석하여 사전 예측, 조류 독감이 싱가포르에서 발생할 경우와 지역에서 벌어지는 위협 수준 등을 분석
일본 IIJ 혁신 연구소(IIJ Innovation Institute), 전염병 데이터 랭킹 서비스	국립감염증연구소에서 매주 공개하는 전염병 발생 동향 조사 보고서의 데이터를 활용·분석하여 전염병 유행 상황과 정보를 제공
한국 국민건강보험공단 & 다음소프트, 국민건강 주의 알람 서비스	국민건강보험공단의 건강보험 DB와 SNS 정보를 연계하여 홍역, 조류독감, SAS 등 감염병 발생 예측 모델 개발하고 주의예보
한국 ㈜케이티 & 농림축산검역본부, AI 확산 대응 프로젝트	국가동물방역통합시스템(KAHIS) 내 축산차량 농장방문 기록, AI발병 농장정보 등(약 15만건), KT 통화로그(CDR) 데이터(약 224억건)를 분석하여 AI 질병 확산 징후 포착

[그림 2] 감염병 예방연구 해외사례

캐나다 온타리오 공과대에 있는 병원에서는 감염 예측을 하는데 있어서 미숙아를 모니터링하여서 연구를 진행한 경우가 있다. 미숙아는 병원균 등의 가염에 취약하여 진찰 후 감염 사실을 파악하게 되면 치료시기를 놓쳐 위험한 상황에 이를 수 있기에 사전 판단과 예방을 위한 연구를 추진하였다. 주요 내용으로는 신생아를 집중 치료하는 환자실의 모니터링 장비의 센서에서 측정되고 수집되는 많은 양의 빅데이터를 실시간 분석하여 질환이 언제 발병할지 예측하는 시스템을 도입하고 데이터를 축적하고 분석하는 통합 시스템을 구축하였다. 그리고 데이터는 의사 및 간호사에게 전달되어 신생아 치료 및 임상연구에 활용되어 사전에 질병을 감지하여 예방할 수가 있어서 의료진보다 최대 24시간 전에 감염 사실을 감지하여 신생아 질환 발병 예측을 할

수 있게 되었다.

[그림 3] 온타리오 공과대의 미숙아 모니터링 감염예측

일본에서는 IIJ혁신 연구소가 있는데 이 연구소에서는 전염병 데이터 랭킹서비스를 하고 있다. 전염병으로부터 국민의 건강을 보호하기 위해 주요 질병의 유행 상황에 사전 감지 및 대응 필요성이 높게 된 이유에서 시작된 서비스이다. 국립감염연구소의 감염병 발생 빅데이터를 활용하여 웹 기반의 전염병 데이터 랭킹 서비스를 제공하여 국민 누구나 쉽게 파악할 수 있도록 서비스 하고 있다. 1999년 4월, 시행된 ”감염 법“에 따라 주요 질환들에 대한 환자가 전국적으로 얼마나 발병했는지를 조사하고 집계한 결과를 국민에게 과거 2년의 추이 비교 및 분석 결과를 도식화하여 제공함으로써 향후 발생 가능한 전염병을 유추할 수 있도록 서비스를 하였다. 또한 각종 감염 랭킹을 제시함으로써 일반 국민들이 전염병 유행 상황을 간편하게 실시간으로 파악할 수 있도록 하는 서비스를 하여 전염병 발생에 대한 국민의 이해도를 한층 높였다.

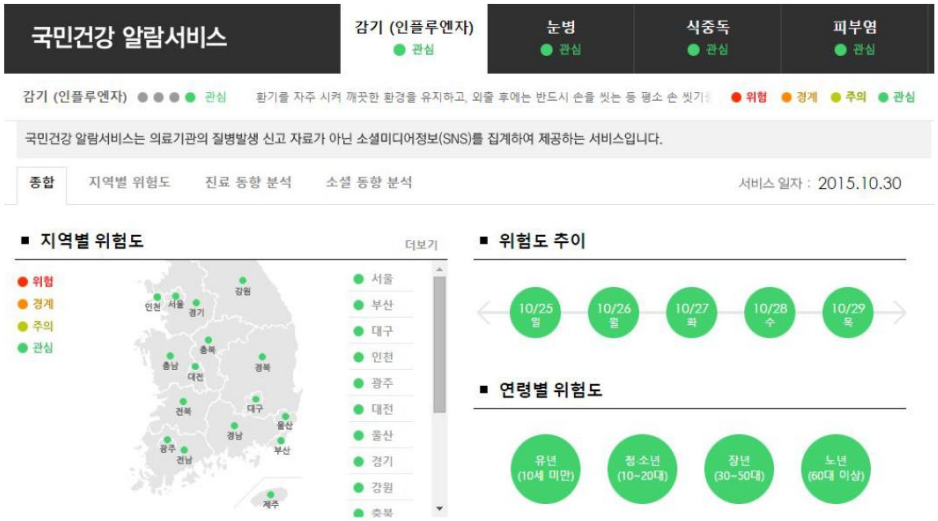
* 지료 : http://www.gryfon.iij-ii.co.jp/infection_ranking

[그림 4] 일본 IIj혁신연구소의 전염병 데이터 랭킹 서비스

싱가포르와 미국은 보안 및 위험관리 분야에 빅데이터를 적극 활용하고 있다. 싱가포르는 전염병 확산을 예방하기 위해 2004년부터 다양한 국가의 위험 데이터를 수집하고 분석하여 사전에 예측할 수 있는 빅데이터 시스템을 구축하여 분석해 오고 있다. 미국은 Pillbox 라는 프로젝트를 통해서 국립보건원 사이트 기반으로 의약품에 대한 정보서비스를 제공하고 제조사와 사용자 간의 정보 공유를 가능하게 했다. 이를 통해서 질병의 분포도와 증감 현황 데이터를 수집하고 분석할 수 있다.

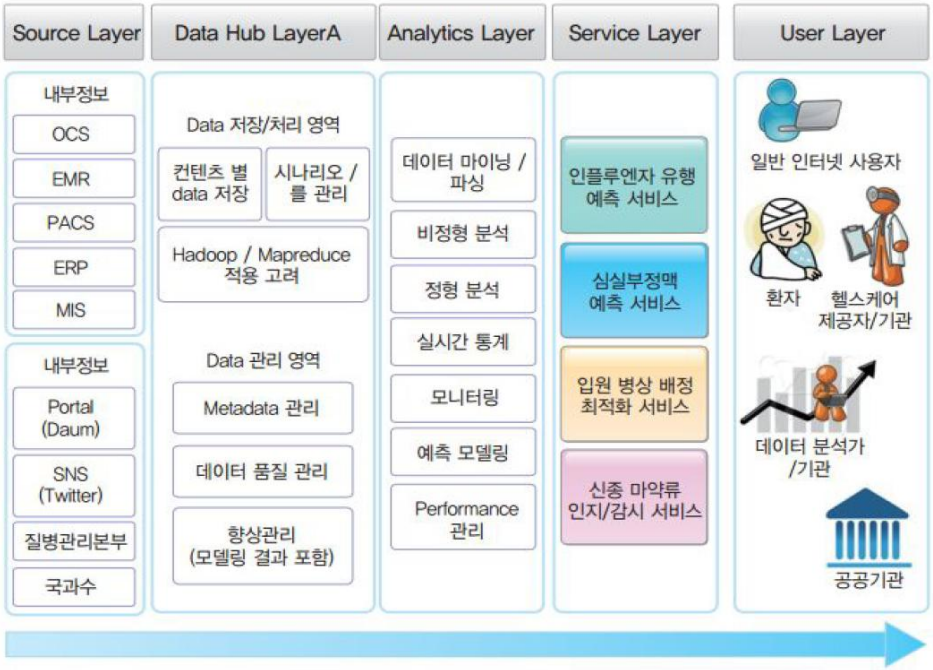
국내 사례로는 국민건강보험공간에서 구축한 ”국민 건강 주의 알람 서비스“가 있다. 전염병으로 인해 발생하는 사회비용 손실의 최소화와 빅데이터 분석을 기반으로 한 예방 및 예측 중심의 의료서비스를 제공하고자 하는 사전 예보 체계이다. 국민건강보험공간이 보유한 국민 건강정보 데이터베이스와 ㈜다음소프트가 보유한 SNS 정보를 융합하여 주요 질병의 동향과 위험도 알람을 제공한다. 2008년부터 2012년까지 5년 동안의 진료데이터를 분서하고 다빈도 감염병에 대한 SNS 빈도수와 월평균 등락률 등을 분석하여 알람 대상 질병을 선정하였다.

선정된 눈병, 식중독, 감기 등의 질병을 발생시기, 원인, 증상과 관련된 어휘들을 ㈜다음소프트의 SNS데이터와 연계하여 분석하고 질병 위험도 예측 모델을 개발하였다. 2013년 빅데이터 시범서비스로 선정된 이후 현재 '국민 건강 주의 알람서비스(http//forecast.nhis.or.kr)를 운영 중이며 서비스를 통해 유행성 질병에 대한 연령별, 지역별 위험도 및 관련 동향 등을 파악 가능하다.



[그림 5] 국민건강 알람 서비스

하지만 즉시 방역 대책 수립이 필요한 제1군, 국가 예방접종 대상인 제2군, 간헐적 유해 가능성이 있는 3군 감염병에 대한 대체적인 정보는 제공하지 않으며 의료기관의 질병 발생 신고 자료가 아닌 비교적 최신 기간(최근 90일)의 SNS만을 집계하여 제공함으로써 동향 파악 정도만 파악하는데 그칠 수 있는 한계가 있다. 서울아산병원이 주관하고 한국전자통신연구원 등이 참여한 보건의료 빅데이터 서비스도 보건의료 질을 향상하고 비용 절감을 위한 보건의료 서비스가 있다.



[그림 6] 아산병원의 보건의료 빅데이터 모델

SNS데이터, 검색데이터, 병원 경영자료, 국과수 마약류 관련 데이터 등의 빅데이터를 활용하여 인플루엔자 예측 동향을 웹서비스로 제공하고 심실부정맥 예측 모델을 개발하여 대시보드 형태로 병원에서 활용하는 시스템을 구축하였지만 감염병 예측보다는 질병 예측에 가까운 측면이 있다.

4. 당뇨합병증 발생 위험요인과 예측모형

1) 당뇨합병증 발생 예측의 위험요인

당뇨합병증 발생에 영향을 미치는 요인은 기본적으로 성, 연령, 출생시 저체중, 유전적 요인 등과 같은 조절이 불가능한 위험요인부터

고혈압, 당뇨병, 콜레스테롤, 체질량지수, 심방세동, 흡연, 운동, 식품 영양 등 조절이 가능한 위험요인, 대사증후군, 음주, 약물 남용, 경구용 피임제, 수면 중 호흡장애 등 조절이 가능한 잠재적 위험요인까지 다양하다. 이러한 위험요인은 국내외에서 많은 학자들의 노력으로 이미 20세기 말에 대부분 규명되었고, 조절이 가능한 위험요인들을 관리할 경우 실제로 당뇨합병증 발생의 감소로 이어진다는 사실이 임상시험을 통해 입증되었다.

그러나 당뇨합병증 발생의 위험요인이 많은 부분 규명되었음에도 불구하고 각각의 위험요인을 정의하는 방식이나 질병의 발생으로 이어지는 병인학적 방법론에는 여전히 다양한 이견이 존재한다. 대표적으로 살펴보면, 미국 프레이밍햄 연구에서 개발된 발생 위험도 예측 모형에서는 성, 연령, 수축기혈압, 항고혈압 치료제, 당뇨, 흡연, 뇌졸중 과거력, 심장잡음 등의 위험요인이 사용되었고, 중국 코호트에서는 비슷하게 연령, 수축기혈압, 이완기혈압, 총콜레스테롤, 흡연상태, 하루 흡연량이 사용되었다. 프레이밍햄 방법론을 적용하여 국내에서도 지선하 등이 성, 연령, 수축기혈압, 당뇨병, 총콜레스테롤, 운동, 1일 음주량을 사용하였으며, 이지성 등은 프레이밍햄 연구에서 사용했던 심방세동, 뇌졸중 과거력 등의 위험요인을 추가하여 성, 연령, 고혈압, 당뇨병, 고지혈증, 심방세동, 뇌졸중 과거력, 비만 흡연여부를 사용하였다. 조경희 등(2015)의 연구에서는 남자의 경우 나이, 혈당, 수축기혈압, 혈색소, 콜레스테롤, 흡연력, 음주력, 고혈압, 당뇨병, 고지혈증, 심장질환 등을 고려하였고, 여자의 경우 나이, 혈당, 수축기혈압, 단백뇨, 흡연력, 고혈압, 당뇨병, 심장질환 등을 위험요인으로 사용하였다.

앞서 언급했듯이 당뇨합병증의 위험요인은 출생 시의 저체중 여부, 유전적 요인, 폐경 후 호르몬 치료, 약물남용, 수면 중 호흡장애, 과다응고증, 염증, 감염증, 무증상 열공성병변 및 백색질 변성 등 다양하게 나타나고 있지만, 실제 예측모형에서 사용되는 위험요인은 성, 연령, 수축기혈압, 흡연, BMI, 음주 등의 비슷한 요인으로 제한되는 경향이

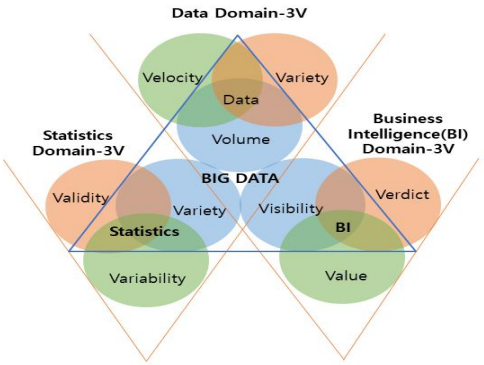
있다. 이는 각각의 위험요인별로 의학적 근거에 대한 연구의 수준에 차이가 있고 국내에 적용될만한 연구결과가 부족하거나, 무엇보다 해당 위험요인을 확인하거나 활용할 수 있는 분석자료 상의 한계로 인하여 예측모형에 다양하게 포함되지 못한 것으로 추측된다.

5. 빅데이터와 예측 분석기법

연구에 사용하게 될 빅데이터 예측분석 기법의 배경에 대해 알아보았다.

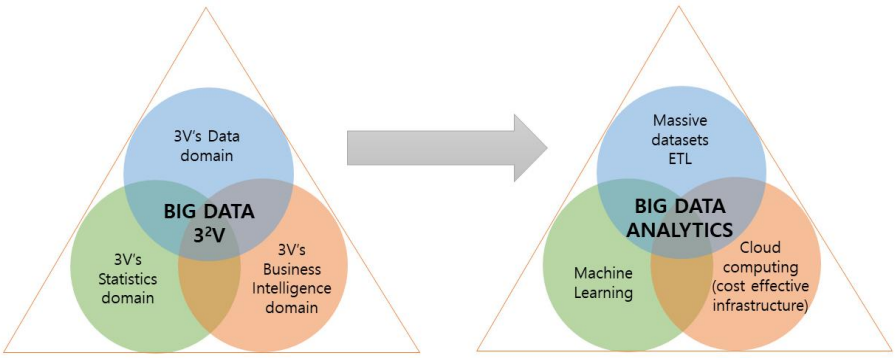
1) 빅데이터

빅데이터(Big Data)는 단순히 규모가 큰 데이터를 이야기 하는 것이 아니며, Gartner는 3V인 데이터의 크기(Volume), 속도(Velocity), 다양성(Variety)의 3가지 특징을 가진 것을 빅데이터라고 정의 하였다. Caser 등의 다른 학자들은 최근 기존의 빅데이터 정의를 바탕으로 새롭게 정리한 포괄적인 빅데이터에 대한 정의를 3^2Vs로 구분하여 다음 [그림 7]과 같이 나타내었다.



[그림 7] 계층모델에서의 3^2Vs 벤다이어그램
출처: Caser Wu et al., 2016

이를 통한 빅데이터 분석(Big Data Analytics)은 빅데이터의 실용적 의미라고 할 수 있는데, 이는 3^2Vs의 데이터, 통계, BI영역이 각각 Massive datasets, Machine learning, Cloud computing으로 [그림 8]과 같이 대응되었다.



[그림 8] 빅데이터 분석에서의 3^2Vs 벤다이어그램
출처: Caser Wu et al., 2016

빅데이터 분석을 통한 연구 및 활용이 활발해지며 과거에는 시도하지 못했던 데이터들을 통한 예측이 가능해졌고, 의사결정에서도 더 나은 해답을 찾을 수 있게 되었다. 현재는 대표적으로 큰 기업들인 구글, 아마존 등은 물론이고, 공공기관 등에서의 다양한 사례를 찾아볼 수 있게 되었다(이형탁, 2019).

2) 예측 분석기법

과거 컴퓨터에 한정된 의미와 달리 현재 머신러닝(Machine Learning)의 의미는 빅데이터 분석 기법 중 하나로써 데이터의 패턴을 인식하고, 이를 이용해 미래 예측이나 의사결정의 방법으로 정의되었다(Kevin P. Murphy, 2016).

머신러닝은 학습종류에 따라 3가지로 분류되는데 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning), 강화학습(Reinforcement Learning)으로 나뉘며 이에 따른 대표적인 알고리즘들을 정리하면 다음[표 8]에 나타내었다. 지도학습은 데이터를 통한 예측모델에 많이 사용되었다. 비지도학습은 입력 데이터에서 규칙 및 패턴을 도출해 데이터를 요약하거나 그룹화해 결과를 해석 및 분석하는 방법으로 알려졌다. 강화학습은 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식해, 선택 가능한 행동들 중 최대 보상을 얻게 되는 행동 혹은 행동순서를 선택하는 방법으로 정의되었다(박인근 외, 2019).

머신러닝 알고리즘 중 특히 지도학습 알고리즘이 미래예측에 용이하고, 이를 통한 대비가 가능하여 여러 산업분야에서 이를 이용한 연구들이 활발하게 진행되었다. 지도학습 알고리즘들을 통해 기업부실의 예측모형 성과를 비교하거나, 조선 생산 분야에서의 리드타임 예측, 인공신경망과 다중회귀분석을 이용한 서울의 다세대 실거래가격을 예측하는 등 다양하게 존재하나 아직 주류의 생산 및 소비에 대해서는 경제학적인 수요함수를 도출하는 데에 그쳤다. 따라서 본 연구에서는 기존의 분석 및 예측 방법과 차별화를 두어 빅데이터 예측 분석 알고리즘 가장 기본적으로 사용되는 회귀 알고리즘을 다양하게 사용하였다.

[표 1] 머신러닝 분석 알고리즘

Learning	Analysis	Algorithm
Supervised Learning (지도학습)	Classificaation (분류)	KNN (K-Nearest Neighbors)
		NBC (Naive Bayes Classification)
		SVM (Surpport Vector Machine)
		Decison Tree
		Random Forest
		Cross-Validation
		Ensemble
		Bagging, Boosting
		Hidden Markov Model
	Regression (회귀)	Linear Regression
		Logistic Regression
		Ridge Regression
		Lasso Regression
		ElasticNet Regression
		Jackknife Regression
	Artificial Neural Network (인공신경망)	CNN (Convolutional Neural Network)
		RNN (Recurrent Neural Network)
		DQN (Deep Q-Network)
Unsupervised Learning (비지도학습)	Clustering (군집화)	K-means Clustering Density-based
		Clustering Fuzzy Clustering
		SOM, Self-Organization Map
	Association rules (연관성 규칙)	Association Rules
		Naive Bayes
	Dimensionality Reduction (차원 축소)	PCA (Principle Component Analysis)
		ICA (Independent Component Analysis)
		NNMF (Non-negative matrix factorization)
		SVD (Singular Value Decomposition)
Reinforcement Learning (강화학습)	Monte Carlo Method	
	TD-Larning, Temporal Difference Learning	
	Policy Gradient Algorithm	

Ⅲ. 분석 연구

합병증에 연관이 있다고 예상되는 공공데이터들을 수집 후 탐색적 자료 분석을 실시하여 모형 형성에 맞게 전처리 작업을 하였다. 전처리된 데이터들을 학습데이터와 예측데이터로 분리한 후 의사결정나무(Decision Tree), 랜덤포레스트(Random Forest), 그래디언트부스팅(Gradient Boosting), 에이다 부스팅(Ada Boosting), 로지스틱 회귀분석(Logistic Regression) 분석방법들을 적용하여 그 중에서 가장 좋은 예측률과 실제 및 예측 일치확률값이 높은 모형을 선택하였다.

1. 데이터 수집 및 데이터 설명

데이터 수집은 국민건강보험공단에서 제공하는 공공데이터들을 활용하였으며, 2018년 1년 단위의 데이터로 수집 및 정리하였다. 가장 중심이 되는 데이터는 합병증에 직접적으로 연관되어 있는 측정 공복혈당, 측정 수축 및 이완기 혈압, 측정 콜레스테롤이었고 그 외에 합병증에 영향을 미칠 수 있다고 여겨지는 연령, 신장, 체중, 허리둘레 등을 함께 수집 및 가공하여 분석하였다.

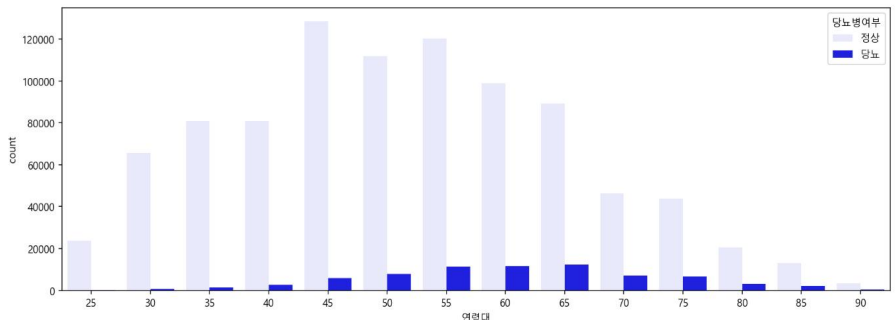
데이터 셋은 총 1000만개의 환자의 의료 검진 데이터이며 총 31개의 변수인 환자의 나이, 키, 몸무게 등의 일반정보와 혈당, 콜레스테롤, 혈당, 요단백, 당뇨병 판정 등 건강 검진 데이터에서 탐색적 데이터 분석방법과 전처리 과정을 통하여 9개의 독립변수로 선택하였다.

목적	분석방법	주요 내용
전체 데이터의 분포 특성	dist plot 분석	연속형 변수를 대상으로 데이터의 전체 분포를 확인
	Box Plot 분석	
변수간의 관련성 확인	산점도 분석	변수간 영향도와 연관성을 시각적으로 확인
	상관 분석	모든변수간의 상관성을 파악
질병(목표변수)별 영향인자 파악	로지스틱 회귀분석	vital few를 조사
	의사결정나무	
	그래디언트 부스팅	
	랜덤포레스트	
파생변수 만들기	도메인 지식 공부	목표변수(질병) 관련 논문보기
목표변수(당뇨, 고혈압) 예측	랜덤포레스트	분류 모델을 사용하여 환자의 질병 예측
	LGBM	
	Decision Tree	
	XGboost	
주요 변수 예측	선형 회귀 분석	회귀 모델을 사용해 vital few를 예측

[표 2] 독립변수 주요 내용 및 분석방법

2. 1차 탐색적 자료 분석(EDA: Exploratory Data Analysis)

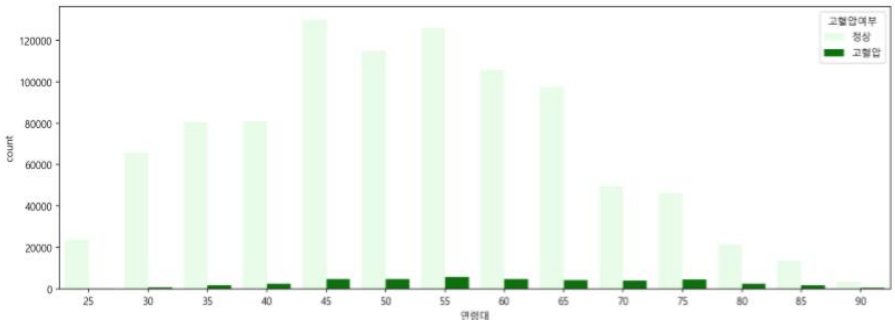
연구의 목표를 나타낸 그래프를 통해 질병의 주요 인자들과 신체적 조건에 대한 상관관계를 그래프로 파악을 하였다.



[그림 9] 연령대별 당뇨병 발생 비율

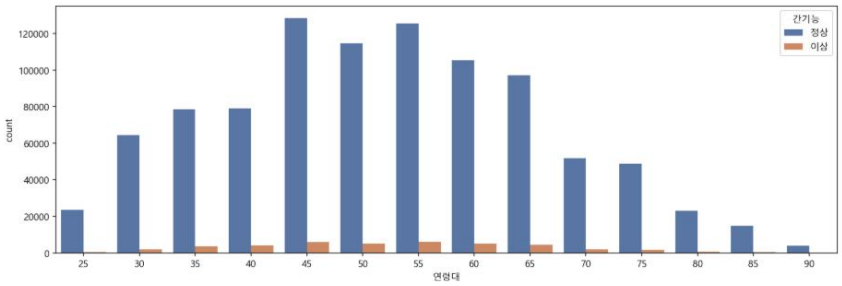
[그림 9]의 그래프 결과를 보면 당뇨병의 경우 35세 이후부터 일정

한 비율을 유지하고 있으며, 50~65세 인구에서 가장 많은 비율을 차지하고 있다.

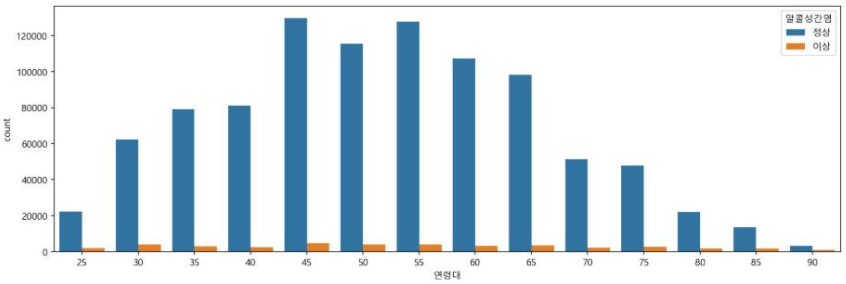


[그림 10] 연령대별 고혈압 발생 비율

[그림 10]의 그래프 결과를 보면 고혈압의 경우 35세 이후부터 일정한 비율을 유지하고 있다.

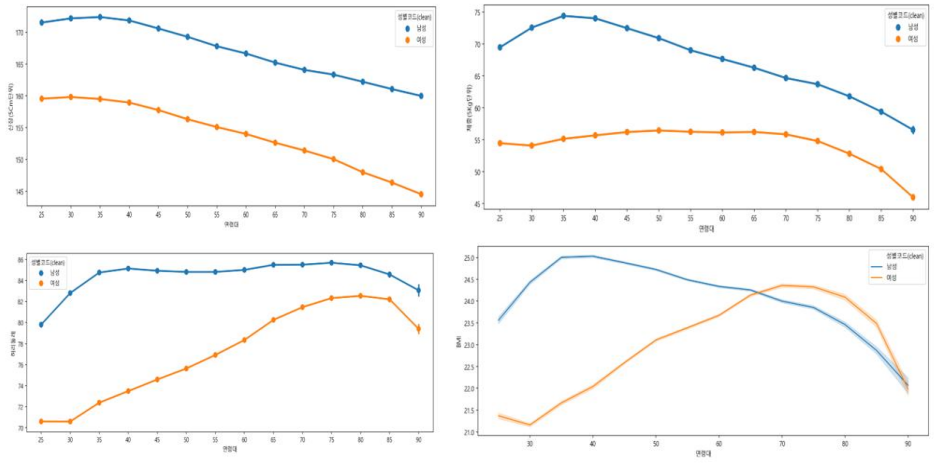


[그림 11] 연령대별 간기능 분포



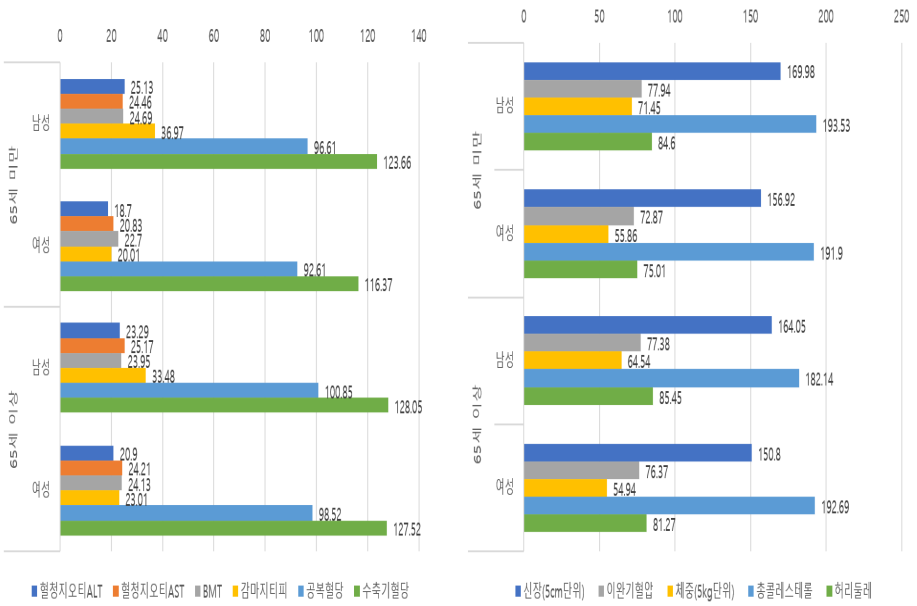
[그림 12] 연령대별 알콜성간염 분포

[그림 11]와 [그림 12] 그래프의 결과를 보면 간기능의 경우 연령이 높은 곳에서 분포가 있으나, 알콜성간염에는 25세부터 분포가 존재하고 있음을 확인할 수 있다.



[그림 13] 연령대별 신장,체중,허리둘레, BMI 분포

[그림 13]의 그래프 결과를 보면 연령이 높아질수록 신장과 체중은 줄어드나, 허리둘레의 경우는 증가하는 추세를 보이며, BMI수치의 경우 여성이 증가하고 남성이 감소하는 추세를 보이고 있다.

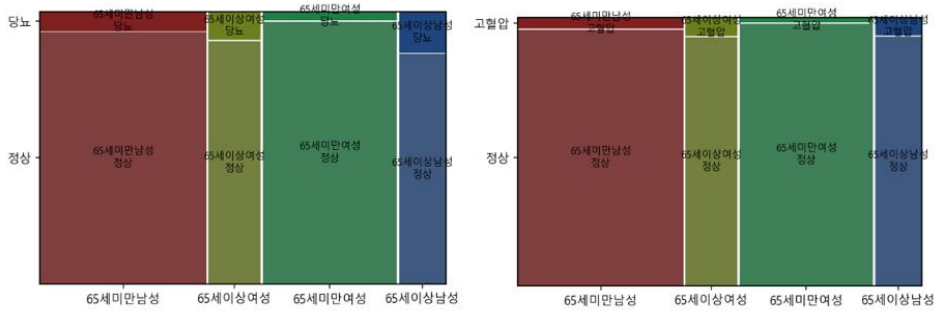


[그림 14] 연령대별 주요인자별 영향 분포도

[그림 14]의 그래프를 결과를 보면 고령이 되면서 남성의 경우 ALT값이 줄어드는 반면, 여성은 증가하고 있으며 BMI 수치의 경우도 남성은 감소하나 여성은 증가하고 있다.

또한 감마지티피의 경우 남성이 고령으로 갈수록 줄어드나 여성은 증가하는 추세이고 공복혈당은 두 성별 모두 증가하고 있으며 공복혈당은 당뇨의 주요인자이기도 하다. 수축기 혈압과 이완기 혈압도 두 성별 모두 증가하고 이 역시 고혈압의 주요인자이다.

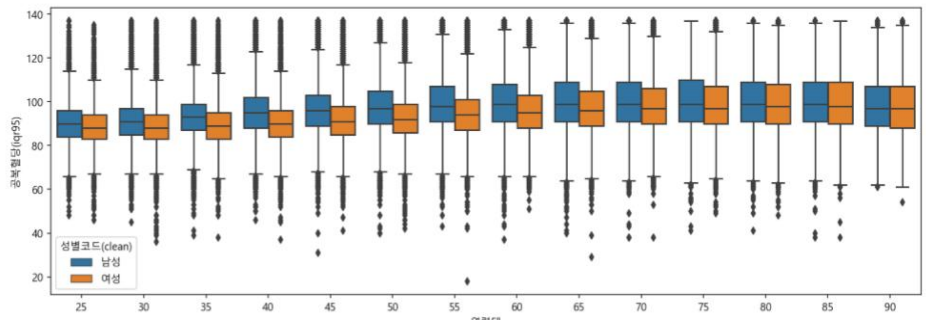
콜레스테롤 수치의 경우는 남성은 줄어드나 여성은 유지되고 있음을 확인할 수 있다.



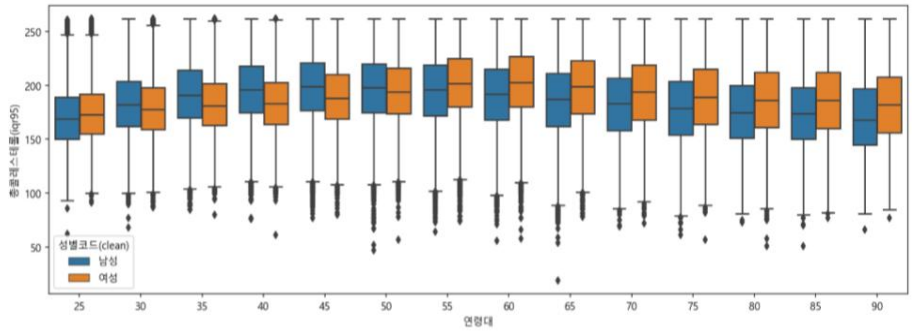
[그림 15] 연령대별 당뇨 및 고혈압여부 분포

[그림 15]의 Countplot을 보면 고혈압의 경우 65세 이상의 인구에서 당뇨별유병율 및 고혈압유병율이 높아짐을 알 수 있다.

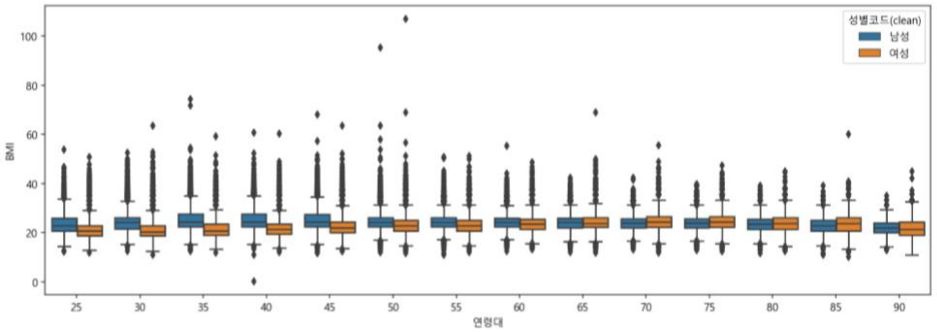
피어슨 상관계수 (Pearson Correlation Coefficient)는 데이터 안의 두 변수간의 선형관계 크기를 나타내는 값으로 공분산의 크기를 사용하여 -1에서 1까지의 값으로 나타내졌다. 절댓값이 1에 가까워질수록 강한 상관관계, 0에 가까울수록 약한 상관관계를 갖는 것이 확인되었다. 선택된 데이터들의 관계를 산점도(Scatter Plot)로 시각화하니 전체 데이터들을 확인할 때와 다르게 변수들 간 선형관계가 비교적 잘 나타났다.



[그림 16] 연령대 및 성별대 공복혈당 상관관계

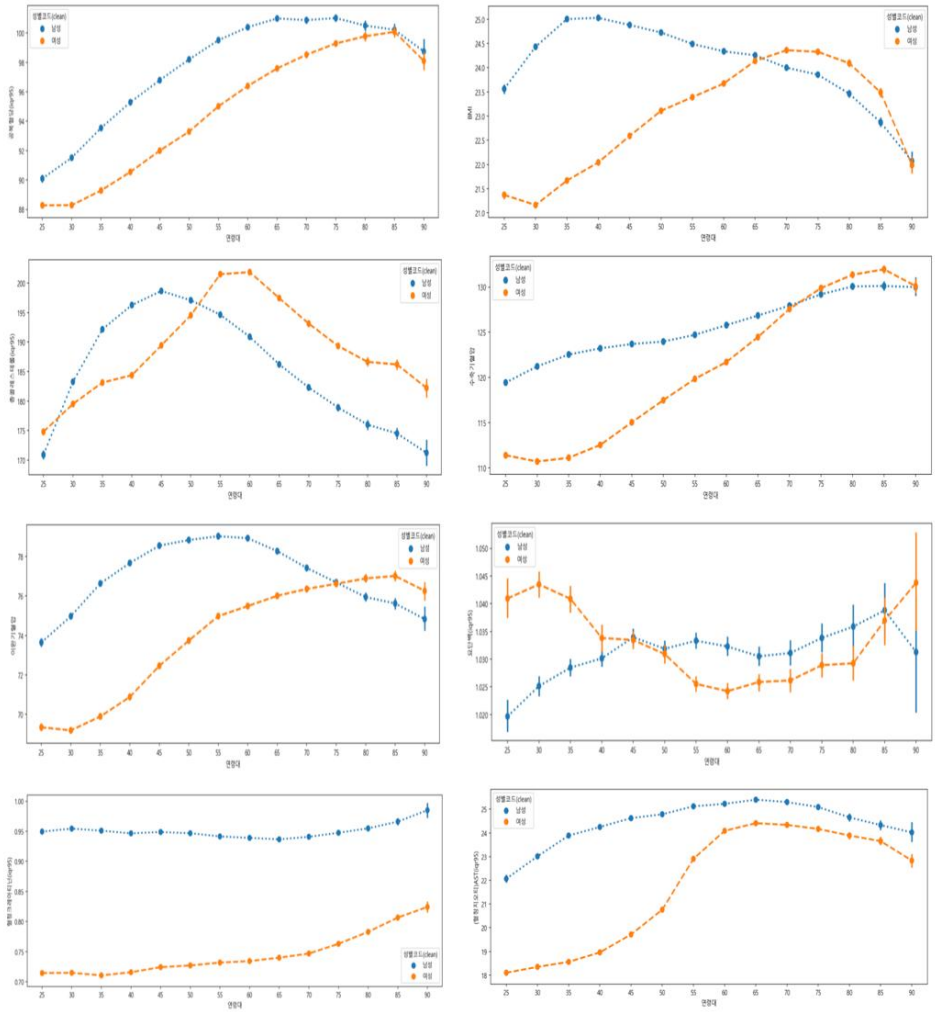


[그림 17] 연령대 및 성별대 총콜레스테롤 상관관계



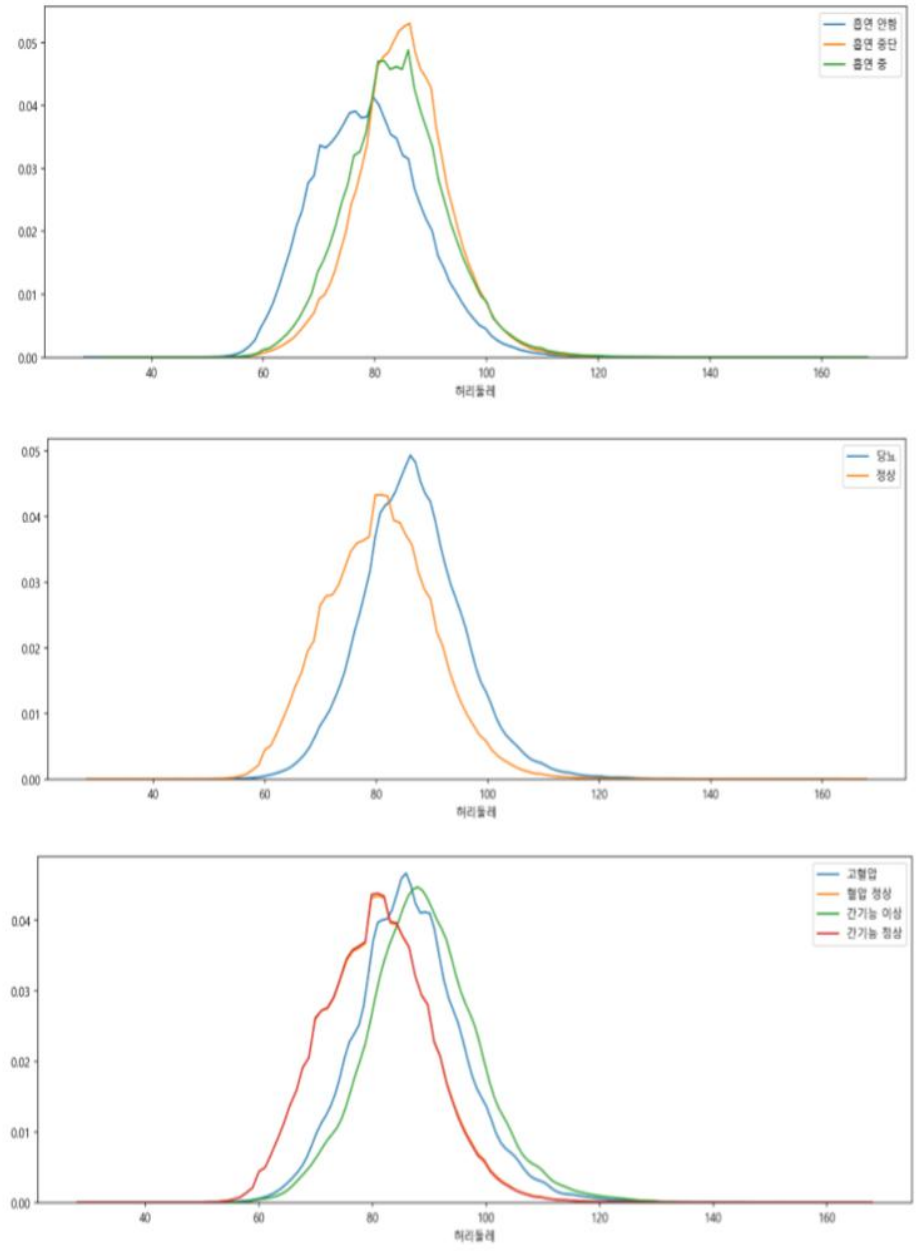
[그림 18] 연령대 및 성별대 BMI 상관관계

또한 상관계수를 통해 다중공선성(Multicollinearity) 문제의 가능성을 발견할 수 있었다. 이는 독립변수 사이에 가장 강한 상관관계가 존재하는 것으로, 통계모형의 계수 신뢰구간이 커져서 각각의 회귀계수에 대한 해석을 어렵게 하는 문제를 발생할 수 있다고 알려졌다(박성현, 2007). 따라서 진단을 통해 분석변수를 조정하거나 일반적인 회귀모형이 아닌 다른 모형의 사용이 필요했다. 다중공선성을 검증할 수 있는 방법으로는 추정된 회귀모형에서 변량의 팽창정도를 의미하는 VIF(Varinance Inflation Factor)가 10 이상이 되거나 공차한계(Tolerance)가 0.1 이하이면 다중공선성에 문제가 있다고 판단할 수 있다(권순호, 이진우, 정건희, 2017). [그림 19]의 그래프를 통해 주요 변수들간의 상관관계 여부를 파악할 수 있다.



[그림 19] 주요변수들간의 상관관계 그래프

[그림 20]의 그래프 결과에서 허리둘레에 따른 질병 여부 상관관계를 탐색적 분석 결과를 볼 수 있는데, 허리둘레와 흡연여부와의 관계, 허리둘레와 당뇨 여부의 상관관계, 허리둘레와 고혈압 및 간기능 이상 여부 상관관계를 파악할 수 있다.



[그림 20] 허리둘레와 주요변수들간의 상관관계 그래프

3. 종속변수와 독립변수 정의

모형 추정을 위해 데이터 탐색 분석을 통해 정리된 데이터들을 각각의 종속변수인 당뇨병 진단 여부 및 당뇨합병증 진단여부로 사용하였다. 독립변수는 각각의 종속변수에 사용하게 될 연령대, 성별, BMI 수치, 신장, 체중, 허리둘레, 수축기혈압, 이완기혈압, 요단백, 콜레스테롤, 감마지티피로 이뤄진 공통 독립변수로 사용하였다. EDA를 통하여 최종 선정된 독립변수는 총 11개이고, 종속변수는 1개로 정의하여 [표 3]에서 나타내고 있다.

[표 3] 종속변수와 독립변수 정의

변수종류	변수이름
종속변수	당뇨합병증 진단 여부
독립변수 (인구)	30대(%)
	50대(%)
	60대(%)
	70대(%)
	80대 이상(%)
독립변수 (의료)	성별
	BMI 수치
	신장
	체중
	허리둘레
	수축기혈압
	이완기혈압
	요단백
	콜레스테롤
	감마지티피

4. 분석 모형 형성

데이터 축소 후 정의되어진 독립변수들을 통해 각 종속변수들마다 가장 적절한 변수의 조합들로 의사결정나무모델, Random Foreset 모델, Gradient Boosting 모델, Ada Boosting 모델, 로지스틱 모델 모형을 형성하였다.

1) 의사결정나무(DTA: Decision Tree Analysis)

의사결정나무분석(decision tree analysis)은 자료 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화하는 데이터마이닝 기법 중의 하나로 의사결정규칙을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행한다. 분류 또는 예측의 과정이 나무구조에 의한 추론규칙(induction rule)에 의해서 표현되기 때문에 신경망분석, 판별분석, 회귀분석 등의 방법들에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(강현철 외, 2006). 또한, 의사결정나무분석은 판별분석이나 회귀분석 등과 같은 모수적(parametric) 모형을 분석하기 위해서 사전에 이상치(outlinear)를 검색하거나 분석에 필요한 변수를 찾아내고 모형에 포함되어야 하는 상호작용효과를 찾아내는 데 사용되는 분석 방법으로, 선형성(linearity)이나 정규성(normality) 또는 등분산성(equal variance) 등의 가정을 필요로 하지 않는 비모수적 방법이다.

의사결정나무분석은 하나의 나무모형구조로 표현되며 마디(node)라고 불리는 구성요소들로 이루어져 있다. 마디는 그 기능에 따라 뿌리마디(root node), 자식마디(child node), 부모마디(parent node), 끝마디(terminal node), 중간마디(internal node), 가지(branch)로 분류된다. 의사결정나무분석의 단계는 다음과 같다. 먼저, 분석의 목적과 자료구조에 따라 적절한 분리기준(split criterion)과 정지규칙(stopping rule)

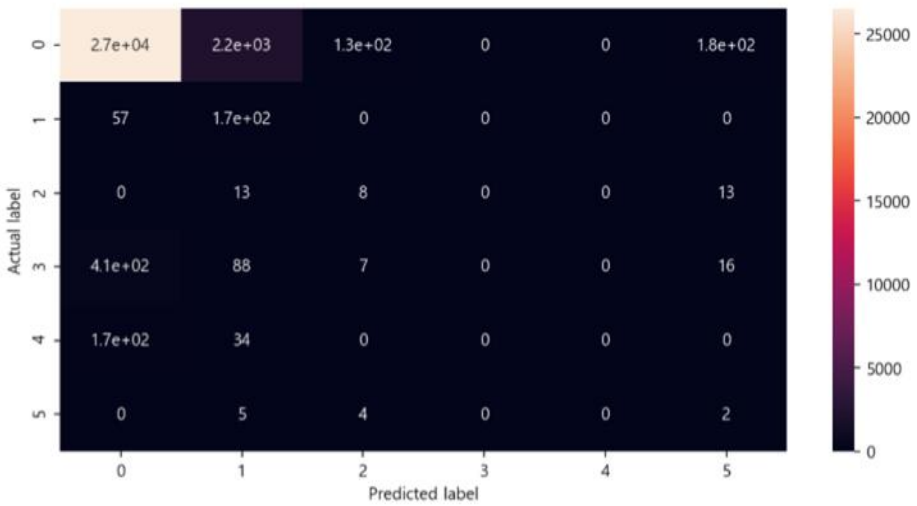
을 지정하여 나무구조가 형성된다. 여기서, 분리기준이란 하나의 부모 마디로부터 자식마디들이 형성될 때, 독립변수의 선택과 범주의 병합이 이루어질 기준을 의미하며, 정지규칙이란 더 이상의 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 여러 가지 규칙을 의미한다. 다음으로는 형성된 의사결정나무에서 분류오류(classification error)를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거한 후 의사결정나무의 타당성을 평가하게 된다.

의사결정나무의 타당성 평가는 이익도표(gain chart)나 위험도표(risk chart) 또는 검증용 자료에 의한 교차타당성(cross validation) 등을 통해 이루어지며 타당성 평가 이후에는 분석 결과를 해석하고 최종 예측모형을 구축하게 된다.

이러한 의사결정나무분석을 위해 CHIDE(Chi-squaredAutomatic Interaction Detection)(Kass, 1980), CART(Classification And Regression Trees)(Breiman et al., 1984), C4.5(1993) 등과 같은 다양한 알고리즘이 제안되어 있으나 가장 보편적으로 사용되고 있는 알고리즘은 CART 알고리즘이다. CART는 지니지수(Gini Index, 이산형 종속변수인 경우 적용) 또는 분산의 감소량(Variance reduction, 연속형 종속변수인 경우 적용)을 이용하여 이지분리(binary split)를 수행하는 알고리즘이다.

지니지수란 각 마디에서의 불순도(impurity)를 측정하는 지수로, 지니지수를 최대한 감소시켜 주는 설명변수와 그 변수의 최적분리를 자식마디로 선택하게 된다. 분산의 감소량은 각 마디의 다양도를 나타내는 측도로서 예측오차를 최소화하는 것과 동일한 기준으로 분산의 감소량을 최대화하는 기준의 최적분리에 의해서 자식마디를 형성하게 된다.

(1) 분석 결과



Best GridSearchCV Accuracy : 0.89
Test Set Accuracy : 0.89

[그림 21] 의사결정나무 분석결과

의사결정나무 분석의 결과로 훈련데스트 정확도는 0.89 이고 테스트 세트 정확도는 0.89 으로 높은 예측률을 보이고 있지만 오차행렬 결과값을 보면 질병을 질병으로 잘 맞추는지 일치확률 지표값은 1번만 높게 나타나고 있다.

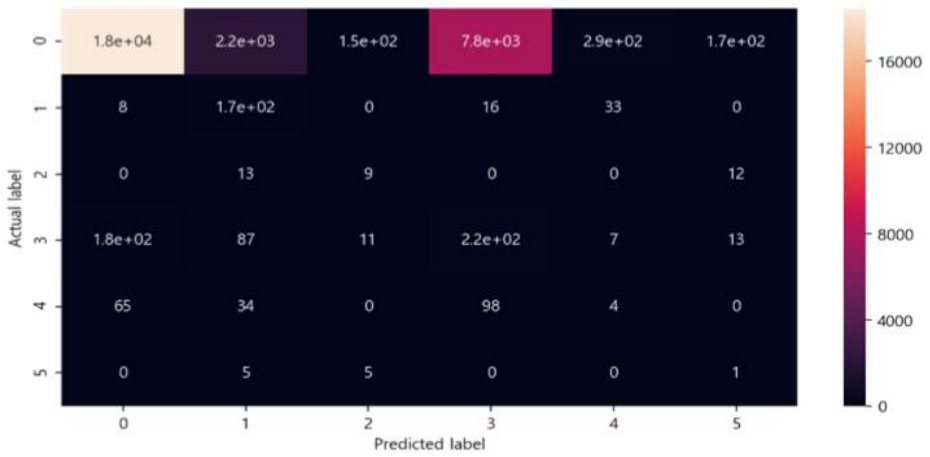
2) 랜덤포레스트(Random Forest)

랜덤포레스트(Random Forest)는 의사결정나무를 여러 개 구축하여 예측 값을 내는 기계학습 기법이다. 랜덤포레스트를 이루고 있는 각 의사결정나무는 전체 학습 데이터 셋으로부터 일부 무작위 복원 추출된 학습 데이터와 설명변수에 의해 구축된다. 각 의사결정나무마다

학습한 데이터 셋이 다르기 때문에 구축된 나무의 모형과 그로부터 예측된 값이 모두 다르게 된다. 랜덤포레스트에서 하나의 의사결정나무의 학습 데이터가 매번 다르게 형성되듯 학습되지 않은 데이터 역시 매번 다르게 추출된다. 의사결정나무 구축에 사용되지 않은 데이터는 모델의 검증용으로 사용되는데, 이를 OBB (Out-Of-Bag)라 한다. 랜덤포레스트의 전체 의사결정나무에서 OOB로 선택된 횟수는 개별 개체마다 다르고 선택되었을 때 분류되는 값도 나무마다 다르게 예측된다. 이러한 특성은 개별 개체에 대한 예측 확률을 산출할 수있게 해주는데, 예를 들어 이진분류 문제 중 Y = 1로 예측할 확률은 (OOB로 사용됐을 때 1로 예측한 횟수)/(OOB로 사용된 횟수)로 계산된다. 이렇게 모든 개체는 확률로 최종 예측 값을 가지게 되고 임계값(Cut-off-Value)에 따라 Y=1 또는 Y = 0으로 분류된다. 본 연구에서는 이러한 확률값을 모기업의 특정 거래처 기업이, 모기업의 가치사슬 기업일 가능성으로 사용하였다. 이 모델의 장점은 첫째, 나무 하나의 정확도는 학습된 데이터 셋이 적고 온전하지 않기 때문에 떨어질 수 있으나, 이들을 종합하여 예측한 최종 정확도는 단순 의사결정나무알고리즘 보다 우수하다. 둘째, 대수의 법칙에 의해 숲의 크기(나무 수)가 커질수록 일반화 오류, 흔히 알려진 이름으로는 오분류율이 특정한계 값으로 수렴하게 되고 과적합(over-fit)현상을 보이지 않는 안정적인 모델로 구축된다. 셋째, 개별 의사결정나무들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 잡음이나 이상치로부터 크게 영향을 받지 않는다. 랜덤포레스트는 빈도가 불균형한 이항분류의 예측에 있어 가장 우수한 예측력을 보인 것으로 보고되고 있다. 실제 전자세금계산서 데이터를 보면 한 기업의 매입/매출에 출현한 거래처 중 가치사슬 내 속하는 거래처들은 일부이고, 일시적인 거래를 하는 기업이거나 소모품 등 쉽게 대체 가능한 업종의 기업들이 대부분 차지하고 있다. 이러한 불균형 현상을 고려했을 때 이와 같은 환경에서 우수한 예측력을 보이는 랜덤포레스트

는 가치사슬 기업의 분류에 효과적으로 적용될 수 있을 것으로 보인다. 랜덤포레스트는 분류와 회귀 문제 모두에 적용될 수 있으나, 범주형 예측 값을 다루는 분류문제에 주로 활용되고 있다. 랜덤포레스트를 분류 문제에 활용한 국내 연구로는 39개의 재무제표 변수를 사용하여 기업의 채권 등급 분류를 진행한 연구가 존재하는데, 이 연구는 인공신경망, 서포트벡터머신, 다변량판별분석 그리고 랜덤포레스트에 모두 적용하여 모델별 성능을 비교하였고 그중 랜덤포레스트에서 가장 우수한 성능을 보였다. 해외 연구사례에도 인공신경망, 서포트벡터머신 알고리즘에 비해 랜덤포레스트가 우수한 분류 성능을 보인 사례가 있는데, 이 연구의 경우 전자 혀(Electronic Tongue) 데이터를 통해 오렌지 음료 및 중국 식초를 구별해내는 연구였다. 랜덤포레스트가 항상 다른 알고리즘들에 비해 더 우수한 분류 성능을 보이는 것은 아니지만 인공신경망 또는 서포트벡터머신과 비교할 때 명확한 이점은 존재한다. 상대적으로 적은 파라미터만을 사용자가 선택하면 된다는 점에서 서포트벡터머신, 인공신경망보다 학습의 편의성이 높다고 할 수 있다.

(1) 분석 결과



```
Best GridSearchCV Accuracy :  0.662
Test Set Accuracy :  0.629
```

[그림 22] 훈련데이터 세트 정확도

랜덤포레스트 분석의 결과로 훈련데스트 정확도는 0.662 이고 테스트 세트 정확도는 0.629 으로 다소 낮은 예측률을 보이고 있지만 오차행렬 결과값을 보면 질병을 질병으로 잘 맞추는지 일치확률 지표값은 1번과 3번은 아주 높으며 그 외 지표값도 어느정도 수치를 보여주고 있다.

3) 그래디언트 부스팅 (Gradient Boosting)

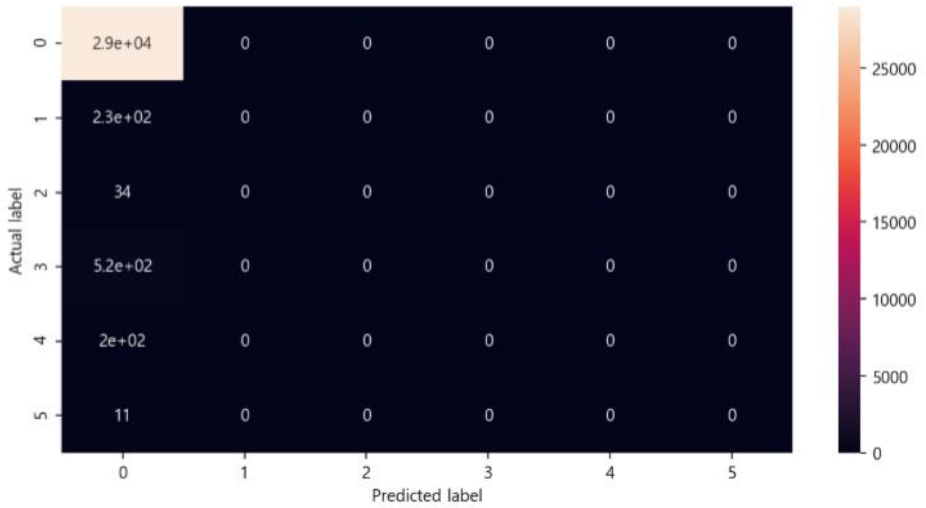
기계학습에서 부스팅(Boosting)이란 비교적 부정확한 약한 학습기(Weak Learner)를 묶어서 보다 정확하고 강한 학습기(Strong Learner)를 만드는 방식을 뜻한다. 일단 정확도가 낮더라도 첫 번째 트리 모델을 만들고, 드러난 약점(예측 오류)은 두 번째 트리 모델이 보완한다. 이와 같은 방법으로 다음트리 모델에서 약점을 계속하여 보완하여 결국에는 강한 학습기를 구축한다.

손실함수(Loss Function)는 예측 모델의 오류를 정량화하며, 이러한 손실함수 값을 최소화하는 모델 내 파라미터를 찾기 위하여 일반적인 기계학습 모델들은 경사 하강(Gradient Descent) 방식을 사용한다. 그래디언트 부스팅은 이러한 파라미터 손실함수 최소화 과정을 모델 함수(f_i) 공간에서 수행하며, 손실함수를 모델 파라미터가 아니라 다음과 같은 (1)수식에 의해 현재까지 학습된 트리 모델 함수로 미분한다.

$$f_{i+1} = f_i - \rho \frac{\delta J}{\delta f_i} \qquad (1)$$

즉, 그래디언트 부스팅 모델에서 트리 모델 함수 미분값은 현재까지 학습된 모델의 약점을 나타내는 역할을 하며, 다음 트리 모델의 피팅을 수행할 때 그 미분값을 사용하여 약점을 보완하여 성능을 Boosting한다.

(1) 분석 결과



Best GridSearchCV Accuracy : 0.966
Test Set Accuracy : 0.967

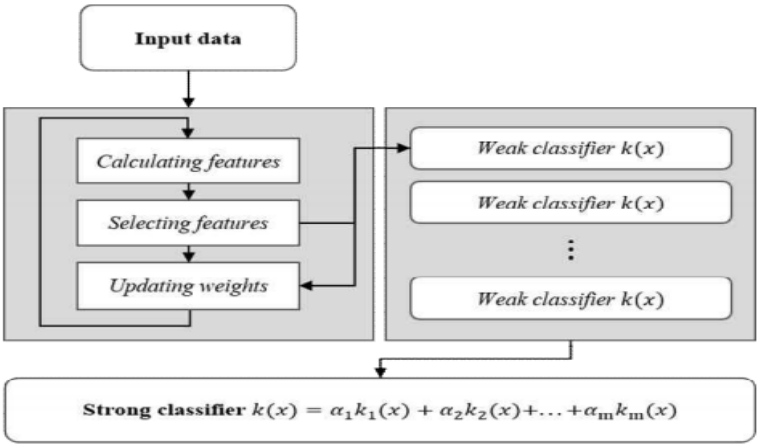
[그림 23] 훈련데이터 세트 정확도

그래디언트 부스팅 분석의 결과로 훈련데스트 정확도는 0.966 이고 테스트 세트 정확도는 0.967 으로 높은 예측률을 보이고 있지만 오차행렬 결과값을 보면 질병을 질병으로 잘 맞추는지 일치확률 지표값은 모든 변수들에 있어서 0으로 나왔다.

4) 에이다 부스팅 (Ada Boosting)

에이다 부스트(Adaboost)는 기계학습 메타 알고리즘(machine

learning meta algorithm)으로 성능을 향상하기 위하여 다른 학습 알고리즘과 결합하여 사용한다. 에이다부스트에서 부스트(boost)는 미리 정해진 수의 모형 집합을 학습하는 것이 아니라 하나의 모형에서 시작하여 모형 집합에 포함할 모형을 추가하는 것이다. 예측 성능이 낮은 개별 분류기를 약한 분류기(weak classifier)라 하며 약한 분류기를 조합하여 더 좋은 성능을 발휘하는 강한 분류기(strong classifier)로 만드는 방법이다. 적응형 부스트(adaptive boost)로 약한 분류기들이 상호보완하도록 단계적으로 학습하며 이들을 조합하여 최종 강한 분류기로 성능을 증폭시킨다.



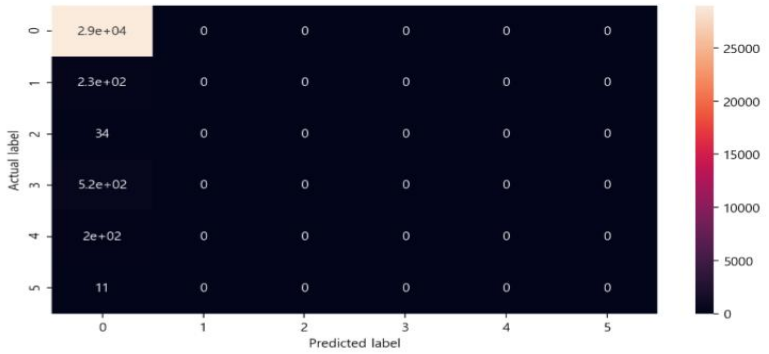
[그림 24] Abstract of Adaboost

[그림 24]와 같이 에이다 부스트를 추상화하여 살펴보면 약한 분류기들을 한 번에 하나씩 단계적으로 학습시킬 때 먼저 학습된 분류기가 잘못 분류한 결과 정보를 다음 분류기의 학습 시 사용하여 이전 분류기의 단점을 보완 하도록 한다. 이는 약한 분류기가 잘못 분류한 표본에 가중치를 두어 더하는(adaptive) 방법으로 잘못 분류되는 데이터에 더 집중하여 학습하고 분류할 수 있도록 한다.

에이다 부스트는 구현이 쉽고 오차가 적으며 또한, 성능을 높이거나

조절해야 할 매개변수가 적다. 개별 분류기들의 성능이 떨어지더라도 각각의 성능이 무작위 추정보다 조금이라도 더 낫다면 최종 모형은 강한 분류기로 수렴한다. 에이다 부스트는 훈련 과정은 모형의 예측 능력을 향상하는 것으로 생각되는 특징들만 선택하고 이는 차원 수를 줄이는 동시에 필요 없는 특징은 고려하지 않음으로써 잠재적으로 수행 시간을 개선한다. 에이다 부스트는 앙상블 기법인 부스팅 알고리즘의 하나로 단일 의사결정나무를 학습하는 것이 아닌 다수의 의사결정나무를 생성하여 단일 의사결정나무의 단점을 보완하고 예측 정확성과 일반화 성능을 높이고자 하였다. 단일 의사결정나무는 데이터를 학습과정 중 과적합되어 잘못된 결정을 하는 단점이 있으므로 여러 개의 의사결정나무를 생성하여 각 의사결정나무가 생성될 때마다 발생하는 오차를 줄여나가며 예측하기 때문에 하나의 트리를 사용하는 것보다 좋은 결과를 도출한다. 또한, 에이다 부스트는 다른 트리 알고리즘보다 적은 하이퍼 파라미터를 사용하기 때문에 건설현장에서 안전에 관련된 데이터분석을 하는 안전관리자 또는 분석가들이 적용하기 용이한 장점이 있다.

(1) 분석 결과



Best GridSearchCV Accuracy : 0.966
Test Set Accuracy : 0.967

[그림 25] 에이다부스트 분석 결과

에이다 부스트 분석의 결과로 훈련데스트 정확도는 0.966 이고 테스트 세트 정확도는 0.967 으로 높은 예측률을 보이고 있지만 오차행렬 결과값을 보면 질병을 질병으로 잘 맞추는지 일치확률 지표값은 앞서 그래디언트 부스팅 분석결과와 같이 모든 질병에서 일치확률값이 0으로 나왔다.

5) 로지스틱 (Logistic)

로지스틱 회귀분석은 범주화를 통해 나타낼 수 있는 종속변수와 독립 변수간의 관계식을 이용하여 두 개 이상의 집단을 분류하고자 할 경우에 사용되는 통계적 분석기법이다. 이는 선형회귀분석과 매우 유사한 기법으로, 선형회귀분석에서는 각 독립변수들의 계수 값을 해석하는 반면, 로지스틱 회귀분석은 회귀분석을 통해 얻어진 계수의 역로그(Anti-log) 값에 해당하는 OR값에 대하여 해석을 수행한다는 차이점이 존재한다.

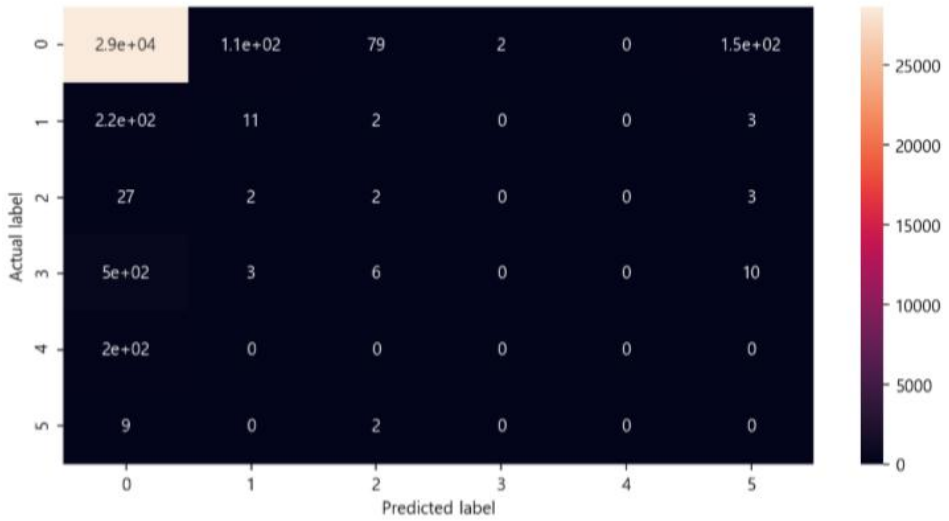
로지스틱 회귀분석에서는 연속형 독립변수 뿐 아니라 이산형 독립변수도 사용할 수 있으며, 이에 대한 어떠한 가정도 필요하지 않다는 장점을 가지고 있어, 독립변수 간 정규분포를 따르며 집단 간 분산과 공분산이 동일하다는 가정을 하고 통계적 분석을 수행하는 판별분석보다 더 선호되고 있다.

독립변수가 하나인 로지스틱 회귀모형은 여러 개의 독립변수를 나타내는 다중 로지스틱 회귀모형으로 쉽게 확장을 할 수 있으며, 다음과 같은 수식으로 나타낼 수 있으며, 해당 식에서 X는 각각의 독립변수를 의미하며, β는 각 독립변수의 회귀계수를 의미한다.

$$P_n = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n} \qquad (2)$$

로지스틱 회귀분석은 모형구조에 의한 연관성 및 교호작용의 유형을 설명할 수 있으며, 모수의 추론을 통해 반응 값에 대한 설명변수의 영향력을 평가할 수 있다. 뿐만 아니라 예측확률을 기반으로 판별 및 분류를 수행할 수 있어, 의학, 통신, 금융 등 다양한 업계에서 로지스틱 회귀분석을 이용하여 어떠한 사건이 발생할 확률을 예측하는 작업을 수행하고 있다.

(1) 분석 결과




```
Best GridSearchCV Accuracy :  0.955
Test Set Accuracy :  0.956
```

[그림 26] 로지스틱 분석 결과

랜덤포레스트 분석의 결과로 훈련데스트 정확도는 0.955 이고 테스트 세트 정확도는 0.956 으로 높은 예측률을 보이고 있지만 오차행렬 결과값을 보면 질병을 질병으로 잘 맞추는지 일치확률 지표값은 전반적으로 수치를 보여주고 있지만 각 질병대 질병 일치율은 거의 0에 가까운 수치로 나타났다.

Ⅳ. 연구결과

각 종속변수별 최종 채택된 예측 모형을 살펴보면 랜덤포레스트를 사용하는 독립변수들은 11개로 구성되었고, 이들 독립변수들로 이루어진 모형으로 각 독립변수의 계수들은 다음의 [표 4]로 나타났다.

[표 4] 최종선정모델 계수

독립변수	변수 계수
(Intercept)	-54.9372
60대	0.5228
여성	-0.1752
BMI	-0.1652
신장	-0.0549
체중	0.0502
허리둘레	0.0204
수축기혈압	0.2601
이완기혈압	0.2634
요단백	0.2272
콜레스테롤	0.2271
감마지티피	-0.0029

Ⅴ. 결론 및 제언

1. 연구 결과 요약 및 시사점

본 연구에서는 두 개 이상의 독립변수들 사이의 관계 파악을 통해 종속변수의 값을 설명하고 이를 예측할 수 있는 통계적 기법인 다양한 딥러닝 모형을 이용하여 당뇨병 위험도 및 당뇨합병증 위험도 비율에 대한 예측 모형을 형성하였다. 각 전처리된 데이터들을 적용하여 분석한 모형들은 훈련데이터와 검증데이터 기반으로 교차검증을 통해 모형의 적용성을 평가, 최종적인 모형 선정과 예측이 이루어졌다. 연구를 통해 도출된 결론은 다음과 같다.

다양한 딥러닝 관련 분석모형들의 비교 결과 당뇨합병증을 예측하기에 가장 적절한 모형으로 랜덤포레스트(Random Forest) 모형이 선택되었다. 당뇨합병증 예측 모형으로 약 63%의 예측률과 예측에 영향을 미치는 각 질병 변수에 대한 일치값들이 가장 높게 나타났다. 또한 P.value 5% 이내에 유의한 인자(당뇨합병증에 영향을 주는 인자)들은 상수 -54.9372, 60대(0.5228), 여성(-0.1752), BMI(-0.1652), 신장(-0.0549), 체중(0.0502), 허리둘레(0.0204), 수축기혈압(0.2601), 이완기혈압(0.2634), 요단백(0.2272), 콜레스테롤(0.2271), 감마지티피(-0.0029)으로 나타났으며 당뇨합병증에 걸릴 확률에 대한 회귀식은 Log(p/(1-p))=-54.9372+0.5228(60대)-0.1752(여성)-0.1652(BMI)-0.0549(신장)+0.0502(체중)+0.0204(허리둘레)+0.2601(수축기혈압)+0.2634(이완기혈압)+0.2272(요단백)+0.2271(콜레스테롤)-0.0029(감마지티피)이다. 이에 따른 결과 회귀식은 차후 당뇨합병증을 예측하는 프로그램 제작에 활용된다면 당뇨합병증 위험도를 자가진단 할 수 있는 서비스를 할 수 있을 것이라고 본다.

2. 연구 한계 및 발전방향

선택된 모형의 설명력이 다소 미흡하여 당뇨합병증이 잘 예측되었다고 볼 수 없었다. 1년 단위의 데이터를 분석하였기에 최소 3년이상의 데이터들을 분석하였다면 더욱 예측력이 좋은 모형 형성이 됐을 것이다. 또한 예측의 정확도보다는 질병을 질병으로 얼마나 잘 맞추는가에 대한 지표를 파악하는 것이 중요하기 때문에 Class Weight를 부여하는 방법으로 모델링을 진행 한다면 더욱 정확한 예측모형이 나왔을 것이라고 본다. 이러한 한계는 분석할 수 있는 본 연구의 H/W의 사양이 떨어지는 이유도 있긴 하다.

참 고 문 헌

강임옥·서수라·이애경. 2006. “건강위험평가를 활용한 건강증진사업 활성화 방안.” 국민건강보험공단 연구보고서.

계묘진. 2012. “Lasso를 기반으로 한 로지스틱회귀모형 연구.” 계명대학교 석사학위 논문.

권순호·이진우·정건희. 2017. “인공신경망 및 다중회귀 모형을 이용한 대설피해 추정 함수 개발.” 한국재방학회논문집. 제17권. 2호. pp. 315-325.

고민정·한준태·이정석·임은실, 2019. “검진 사후관리의 질 제고를 위한 건강위험 예측모형의 타당성 평가”. 국민건강보험공단 연구보고서.

김인철. 2018. “다중회귀분석, ARIMA모형과 신경회로망을 이용한 전력 수요 예측.” 인하대학교 석사학위 논문.

김홍표. 2018. “LASSO 회귀분석을 활용한 PGA 선수스코어에 영향을 미치는 요인분석.” 연세대학교 석사학위 논문.

박인근·홍지후·강남규·김성호·정구범. 2019. 『4차 산업혁명 현장 전문가가 알려주는 빅데이터 분석과 활용』. 파주: ㈜제이펍.

박성현. 2007. 『회귀분석』. 민영사.

신호철, 2004. 건강위험평가, 대한임상건강증진학회 춘계학술대회지, S26-S32.

이정찬, 2010. “건강관리서비스 도입방안 검토와 대안모색”. 대한의사협회의료정책연구소 연구보고서, 1-264.

이형탁. 2019. “머신러닝 예측 알고리즘을 이용한 선박 접안속도에 영향을 미치는 요인 분석.” 한국해양대학교 석사학위 논문.

조경희·박영민·지수혜·추정은·임현선. 2014. 『개인별 맞춤형 통합 건강관리 프로그램 연구 및 개발』 일산병원 연구보고서.

조비룡·조희경·박진호·김주영·강승완·권혁태. 2007. 『건강위험평가 개선을 위한 연구개발』서울대학교 의과대학 가정의학교실·국민건강보험공단.

조희숙. 2003.『의료서비스 관여도가 고객 만족도 및 애호도에 미치는 영향』이화여자대학교 대학원 박사학위논문.

Anderson, D, R., & Staufacker, M. J. 1996. “The impact of worksite-based health risk appraisal on health related outcomes”. A review of the literature. American Journal of Health Promotion, 10(6), 499-508.

Chawla, N. v., & Davis, D. A. 2013. “Bringing Big data to Personalized Healthcare: A Patient-centered Framework”.Journal of General internal Medicine, 28(3).

Caser, W, Rajkumar, B., and Kotagiri, R. 2016. “Big Data: Principles and Paradigms.” Elsevier, Inc.

Hui, Z., and Trevor, H. 2005. “Regularization and Variable Selection via the Elastic Net.” Journal of the Royal Statistical Society. Series B (Statistical Methodology).Vol. 6.(2). 301-320.

Park IB, Balk SH. 2009. “Epidemiologic characteristics of diabetes mellitus in Korea: current status of diabetic patients using Korean Health Insurance Database”. Korea Diabetes Journal. 133:357-362.

Kevin, P.M. 2016. 『Machine Learning: A Probabilistic Perspectiv e』. The MIT Press.

Leavell, H. R., & Clark, E. G. 1958. “Preventive Medicone for the Doctor in his Community”. An Epidemiologic Approach. 2nd edition, New york.

Lee DW, Park CY, Song SJ. 2011. “Study on survey of knowledge

and awareness level of diabtic retinopathy in type 2 diabetes patients: results from Seoul metro-city diabetes preventionn program survey". Journal of the Korean Ophthalmological Society :52:1296-1301.

Robert, T. 1996. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological). Vol. 58(1). 267-288.

Stanton, M. W. 2002. "Expanding patient-centered card to empower patients and assist providers" Research in Action. Agency for Healthcare Rerearch and Quality, Rockville.

Wallace, R. B. 2019. "Primary Prevention" Encyclopedia of public Health. Retrived April 23.

국 문 초 록

머신러닝 기법을 활용한 합병증 예측모형에 관한 연구 :국민건강데이터를 중심으로

김신영
빅데이터·산업보안학과 빅데이터전공
남서울대학교 대학원

기존의 다양한 질병예측에 관한 연구들은 설문지 및 코호트DB를 그대로 사용하여 다양한 합병증 위험도 예측을 그대로 반영하지 못했다는 한계를 가졌다. 합병증에 관한 다각적인 분석을 위해서는 상당한 노력과 전문성이 요구되기에 이러한 복잡한 분석과정을 자동으로 지원해줄 수 있는 다양한 방법들을 적용할 필요가 있다. 그래서 본 연구에서는 국민건강공공데이터들을 분석하여 당뇨합병증에 주요하게 영향을 미치는 요인이 무엇인지 그리고 당뇨합병증의 위험도를 예측하는 모델이 무엇인지를 확인하고자 하였다. 이에 대한 분석으로 머신러닝 예측기법 중 의사결정나무, 랜덤포레스트, Gredient Boosting, Ada Boosting, 로지스틱 회귀분석을 활용하였다. 그 결과 가장 적합한 모형인 랜덤포레스트 예측모형을 통해 예측률을 확인하였고 당뇨합병증 위험도를 예측하는 당뇨합병증 위험도 예측 회귀식이 도출하였다.

주제어 : 당뇨합병증, 기계학습, 예측분석, 랜덤포레스트, 회귀모형

ABSTRACT

A Analysis in Complication Prediction Using Machine Learning Prediction Algorithm : Focusing on National Health Data

Sin Young Kim
Dept. of Big Data·Industry Security, Big Data Major
The Graduate School of Namseoul University

Existing studies on disease prediction had limitations of using the questionnaire and cohort DB as they did not reflect various complication risk predictions. Multilateral analysis of complications requires considerable effort and expertise, and it is necessary to apply a variety of methods that can automatically support this complex analysis process. In this study, we analyzed public health data to find out what factors affect diabetic complications and what models predict the risk of diabetic complications. Decision trees, random forests, Gredient Boosting, Ada Boosting, and logistic regression analysis were used in the machine learning prediction techniques. As a result, the prediction rate was confirmed through the random forest prediction model, which is the most suitable model, and the regression equation for diabetic complication risk predicting the risk of diabetic complication was derived.

Key words: Diabetes complications, Machine learning, Predictive analysis, Random forest, Regression model