

머신러닝을 이용한 미숙아의 재원일수 예측 융복합 연구

김축환¹, 강성홍^{2*}

¹순천향대학교 천안병원 의료정보팀, ²인제대학교 보건행정학과

Convergence study to predict length of stay in premature infants using machine learning

Cheok-Hwan Kim¹, Sung-Hong Kang^{2*}

¹Medical Information Team, Soonchunhyang University Cheonan Hospital

²Dept. of Health Policy & Management, Inje University

요약 본 연구는 미숙아의 재원일수 예측 모형을 머신러닝 기법을 통해 개발하기 위해 수행되었다. 모형 개발을 위해 질병관리본부에서 수집한 퇴원손상심층조사 자료의 2011년부터 2016년까지 퇴원한 미숙아 6,149건을 이용하였다. 입원 초기 신경망 모형은 설명력(R^2)이 0.75로 다른 모형에 비해 우수하였다. 입원 초기 변수에 임상진단을 CCS(Clinical classification software)로 변환하여 추가 투입한 모형은 큐비스트(Cubist) 모형의 설명력(R^2)이 0.81로 랜덤 포레스트(Random Forests), 그라디언트 부스트(Gradient boost), 신경망(neural network), 벌집화 회귀(Penalty regression) 모형에 비해 성능이 우수하였다. 본 연구는 전국단위 데이터를 이용한 미숙아의 재원일수 예측 모형을 머신러닝을 통해 제시하고 그 활용 가능성을 확인하였다. 하지만 임상정보, 부모정보 등 데이터의 한계로 향후 성능 향상을 위한 추가 연구가 필요하다.

주제어 : 재원일수, 머신러닝, 큐비스트, 인공신경망, 융복합 연구

Abstract This study was conducted to develop a model for predicting the length of stay for premature infants through machine learning. For the development of this model, 6,149 cases of premature infants discharged from the hospital from 2011 to 2016 of the discharge injury in-depth survey data collected by the Korea Centers for Disease Control and Prevention were used. The neural network model of the initial hospitalization was superior to other models with an explanatory power (R^2) of 0.75. In the model added by converting the clinical diagnosis to CCS(Clinical classification software), the explanatory power (R^2) of the cubist model was 0.81, which was superior to the random forest, gradient boost, neural network, and penalty regression models. In this study, using national data, a model for predicting the length of stay for premature infants was presented through machine learning and its applicability was confirmed. However, due to the lack of clinical information and parental information, additional research is needed to improve future performance.

Key Words : Length of stay, machine learning, cubist, artificial neural network, convergence study

*Corresponding Author : Sung-Hong Kang(humkang@inje.ac.kr)

Received May 14, 2021

Revised June 7, 2021

Accepted July 20, 2021

Published July 28, 2021

1. 서론

한국사회는 저 출산과 노령사회화가 지속적으로 심화하고 있다. 통계청 지표에 따르면 2010년 1.22명이던 합계출산율이 2020년에는 0.84명으로 감소했다[1]. 출산율의 감소와 함께 임신연령의 증가, 난임 치료 등 사회 환경적 요인에 따라 미숙아 출생은 증가하고 있다[2]. 미숙아는 면역이 약하거나 신체 장기발달이 미숙한 상태에서 태어나지만 출생 후 의료서비스를 적기에 제공함으로써 건강한 사회의 일원으로 성장 할 수 있다[2]. 이를 위한 적절한 의료자원의 투입이 필요하며, 미숙아는 출생 직후 신생아집중치료실에서의 관리가 필수적이다. 2008년부터 권역별 신생아집중치료 병상의 불균형을 해소하고 미숙아 등의 고 위험 신생아에 대한 집중치료 접근성을 제고하기 위해 정부 차원의 지원 정책이 이루어지고 있다[4].

의료기관의 재원 기간 예측은 부족한 병상 자원의 효율적 관리를 가능하게 한다. 또한 합병증을 방지하기 위하여 조기 관리를 가능하게하고, 병원에 있는 인력 및 시설을 보다 능률적으로 관리하도록 한다[5]. 의료 질 개선 활동을 자극하고 향후 연구 및 의료 행위에 대한 근거를 제공 한다. 특히 미숙아 치료 과정에서 초기 불안한 산모에 대한 설명의 기초 자료가 될 수 있다[6]. 2019년 건강보험심사평가원의 신생아 중환자실 적정성 평가 결과에 의하면 평균재원일수는 13.6일, 4일에서 7일 이내 재원환자는 34.2%인 것으로 보고되고 있다. 이중 미숙아 상병 코드(P07)를 가진 비율은 25.0%로 가장 많았다[7].

Seaton은 영국의 국립 신생아 연구 데이터베이스에 등록된 24주 이상 31주 이하의 조산아에 대해 분석한 결과 사망환자의 절반이 10일 이하의 재원기간 안에 발생하는 것으로 확인 되었다. 출생 후 약 10일에 생존아의 특성을 파악하여 예상 재원기간을 부모에게 설명한다고 가정할 때 평균 체류기간은 92일 이었다[6]. 캘리포니아의 주산기품질관리협력(CPQCC)에 등록된 데이터의 출생 시 체중 401g이상 1000g이하의 극 초저체중아를 대상으로 한 재원일수 예측에서는 로그변환 선형 혼합모델, 음 이항 일반화선형모델이 유사한 결과로 고려되었다. 또한 출생체중과 재태 연령 요인에 따라 변동성이 큰 것으로 확인 되었다[8]. 중환자실의 재원일수 예측을 위한 여러 선행연구들은 크게 장기입원여부를 이항변수로 예측하거나, 입원일수를 연속변수로 예측하였다[9-11]. 최근 다수의 재원일수 예측 연구들에서 머신

러닝(Machine learning) 기반 예측 모델이 좋은 예측 성능을 보이는 것으로 보고되고 있다[11-16]. Brandon Thompson은 신생아의 행정자료를 이용하여 분석 자료의 사사분위(Q4 : 85%)에 속하는 장기입원 예측을 위한 기계학습 모델을 구축하였다. Zero R, Navie Bayes, 다층퍼셉트론(Multi-layer Perceptron), 단순로지스틱(Simple Logistic), 서포터벡터머신(Support Vector Machine), J48(C4.5), 랜덤포레스트(Random Forest), 랜덤트리(Random Tree) 모델을 평가한 결과 랜덤 포레스트 모델(ROC 0.877)을 우수한 모델로 제안하였다[16]. Whellan은 심부전환자의 재원일수 예측연구에서 입원 당시 이용 가능한 병원 특성과 임상 데이터를 사용하여 재원일수 예측인자를 연구하였다. 재원일수가 긴 환자가 재원일수가 짧은 환자보다 동반 질환이 더 많고 질병 중증도가 더 높았다[17]. 신경계질환을 대상으로 머신러닝을 이용한 또 다른 국내 연구에서는 환자의 동반상병을 이용한 중증도 보정 방법이 제시되었다. CCI(Charlson comorbidity index), ECI(Elixhauser comorbidity index), CCS(Clinical classification software)의 분류기준을 각각 적용하고, 회귀분석, 의사 결정나무, 랜덤 포레스트, 서포터 벡터 회귀, 신경망 모형을 적용하여 비교한 결과 CCS 진단군 분류를 통해 동반상병을 보정한 신경망 모형이 가장 모형 설명력(R^2 0.31)이 높았다[18]. 이와 같이 재원일수 예측에 있어 동반질환은 중요한 변수로 평가 받고 있다. 다른 한편으로는 퇴원손상심층조사 자료를 이용한 중증도 보정 재원일수를 타 의료기관에서 벤치마킹 활용하는 방안도 연구되었다[20]. 이에 본 연구에서도 선행연구와 같이 동반상병을 CCS 임상분류법에 따라 분류하여 적용하고, 선행연구의 기계학습에 사용된 여러 알고리즘의 적용을 고려하였다.

CCS(Clinical classification software) 임상 분류법은 미국 AHRQ(The Agency for Healthcare Research and Quality)에서 개발하여 모든 질병을 임상학적으로 군집화하는 방법이다. 개발 된 CCS 질병군을 이용하면 모든 동반상병을 259개의 질병군으로 분류할 수 있다[18,19].

퇴원손상심층조사의 동반질환을 이용한 재원일수 예측은 뇌졸중[13], 신경계질환[18], 심장병[19], 목 손상환자[21] 등 다양한 국내연구 사례가 있었으나, 중증도 보정과 재원일수에 영향을 미치는 변이 요인을 확인하는 목적이었다. 신생아 및 미숙아에 대한 재원일수 예측 국내연구는 확인 되지 않았다. 해외 사례는 주로 신생아

등록데이터를 관리하는 영국[6]과 미국[8,12], 인도[15] 등에서 미숙아의 신생아집중치료실 입원기간에 대한 일부 연구가 있었다. 이에 본 연구에서는 질병관리본부의 퇴원손상심층조사 자료와 다양한 머신러닝 기술을 이용하여 재원일수 예측 모델을 개발하여 제안하고자 한다.

2. 연구방법

2.1 연구방법

2.1.1 자료수집

본 연구는 질병관리본부로부터 퇴원손상심층조사 자료를 제공받아 미숙아를 대상으로 하였다. 퇴원손상심층조사는 질병관리본부에서 2004년 미국의 퇴원환자조사 방법론을 근거로 구축하여 2005년부터 매년 전국의 100병상 이상의 종합병원 및 병원에 입원하였다가 퇴원한 퇴원환자를 대상으로 170개의 병원의 표본 환자를 대상으로 수집된 자료이다[13,19].

2.1.2 분석대상자 추출

퇴원손상심층조사자료로부터 수집된 2011년부터 2016년 자료 1,362,455건으로부터 입원경로가 신생아실이면서 미숙아(premature infant) 또는 조산아(preterm infant) 질병코드(P07.0-P07.3)를 가진 환자 6,805명을 대상으로 하였다. 신생아집중치료실과 중환자실 입원환자를 대상으로 한 선행연구에서 사망으로 인한 조기 퇴원과 생존퇴원의 재원일수 예측의 차이와 재원일수 분포를 고려하여[6,9] 퇴원형태가 호전되어 귀가한 경우가 아닌 사망 또는 전원, 기타 421건, 재원일수가 174일을 넘는 6건을 제외하였다. 자료의 한계로 재태 연령과 신생아체중을 확인할 수 없는 229건을 제외한 6,149건을 최종 대상으로 하였다<Fig. 1>.

2.2 데이터의 전 처리

현실세계 데이터는 일반적으로 불완전하고 일관성이 없어 예측에 직접 사용 할 수 없거나 결과가 만족스럽지 않다. 이러한 데이터의 예측 결과를 향상시키기 위해 데이터 전 처리를 하게 된다. 데이터 전 처리에는 데이터 정리, 통합, 변환, 축소 등의 여러 가지가 있다[9].

본 연구에서는 재원일수를 목표 변수로, 입원기초변수 7개와 임상변수 38개를 설명 변수로 선정하였다. 통계적 특성 분석에 따르면 재원일수의 분포는 오른쪽으

로 심하게 치우치고 꼬리가 긴 왜곡된 분포 특성을 가지며 이는 선행연구와도 일치한다[8,9]. 치우친 분포는 Box-Cox 변환을 통해 정규분포 형태로 수정하였다. 범주형 변수는 One Hot Encoding 처리하여 0과 1로 구성된 숫자 형식으로 변환 하였다. 임상변수는 질병분류 코드를 CCS 진단군 분류에 따라 변경 후 역시 0과 1로 변환 하였다. CCS 진단 그룹의 발생 빈도수가 30건을 넘지 않는 변수에 대해 선택적으로 일부를 제거 하였다. 통계적 특성 분석과 함께 전 처리를 거친 데이터는 모델 개발용 훈련데이터(Train data) 70%와 모형 정확도 평가를 위한 검증데이터(Test data) 30%로 분할하였다.

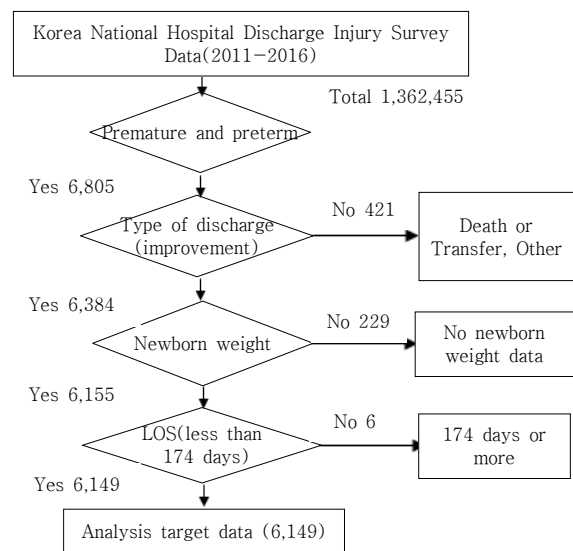


Fig. 1. Analysis subject extraction

2.3 분석방법

미숙아의 입원초기 확인 가능한 변수를 통하여 예측된 재원일수는 입원 후 시간이 지남에 따라 추가 확인되는 임상진단으로 계속 수정될 수 있다[6]. 본 연구에서는 입원초기변수와 임상변수를 구분하여 통계적 특성을 파악하고 머신러닝을 통한 예측모형을 구축 하였다. 예측 모형은 입원초기 모형과 임상변수가 추가 투입된 모형으로 구분하여 각각 모형 평가를 통해 최종 모델을 제안하고자 하였다. 통계분석 및 예측 모형 구축은 R 3.6.1과 R Studio를 사용 하였다. 머신러닝 모형 구축에는 caret 패키지를 사용 하였다. caret 패키지는 약 200개의 머신러닝 알고리즘에 대한 참고를 제공하며, 서로 다른 알고리즘을 비교할 수 있다[22].

2.3.1 통계적 특성분석

분석대상 미숙아의 일반적 특성, 제 특성에 따른 평균 재원일수를 살펴보았다. 재원일수는 등 분산의 가정을 만족하지 않아 Welch검정을 실시하고 Games-Howell를 통해 사후검정을 하였다.

2.3.2 예측 모형개발

모형 개발은 학습을 위해 분할된 70%의 데이터(Train data)를 이용하였다. 머신러닝 알고리즘의 선택은 선행연구에 따라 비교 가능한 여러 알고리즘이 고려되었다. 최종 모형은 벌점화 회귀(penalized regression), 그라디언트 부스트 머신(Gradient Boost), 랜덤 포레스트(Random Forest), 인공신경망(Neural Network), 큐비스트(Cubist)을 통해 모형을 구축하였다.

벌점화 회귀(penalized regression)모형 중 L1 정규화를 적용한 라쏘 회귀(lasso regression)는 회귀계수에 벌점을 적용하는 회귀분석 방법으로 영향력이 거의 없는 변수의 회귀계수를 제거하여 모형을 단순화하는 장점이 있다. 최소제곱회귀모형이 관측치보다 변수가 많아지는 경우 행렬식이 0에 근사하여 분산이 무한에 수렴하는 문제를 Biase를 이용하여 분산을 낮추기 위해 고안된 방법이다[9,10].

랜덤 포레스트(Random Forest)는 다수의 의사결정 나무의 예측 평균을 이용하여 단일 회귀 트리보다 향상된 성능의 비선형 관계를 분석할 수 있는 모형이다[22]. 다수의 각 트리는 서로 다른 부트스트랩(Bootstrap) 데이터 세트에서 훈련되어 생성되며, 그 결과를 종합하여 최종 모형을 선택하는 의사결정나무 기반 앙상블 방법(ensemble methods)이다[18].

그라디언트 부스트(Gradient Boost)는 여러 개의 간단한 모델을 사용하거나 기본 학습기 혹은 성능이 약한 여러 학습기를 연결하여 성능을 높이는 앙상블 학습방식이다[24]. 데이터 셋에 대한 모델을 만든 후 오차의 수정(손실 함수)을 통해 오차를 최소화되는 방향으로 순차적으로 반복(경사 하강법)하여 최적화하는 머신러닝 기술이다. 매개 변수 튜닝이 필요하며 트레이닝 시간이 길고, 트리 기반 모델의 특성으로 고차원 데이터 셋에는 잘 동작하지 않는 단점이 있다[25]. 하지만 예측하기 어려운 데이터에 약한 기본 모델을 개발하여 순차적으로 모델을 추가함으로써, 이전 모델의 실수가 수정된 정밀한 최적의 모델을 만들 수 있는 장점이 있다.[26].

인공신경망(Neural Network)은 복잡한 인간의 뇌를 모방하여 알고리즘화한 것으로, 네트워크를 구성하기 위해 레이어를 연결하는 뉴런과 노드로 구성 된다. 입력층

(input layer)과 출력층(output layer)으로 구성된 레이어 사이에는 여러 은닉층(hidden layer)을 두어 활성화 함수와 가중치에 의해 오차를 최소화하는 모델을 생성하면서 각층을 통과한다. 고도로 구조화된 문제를 해결하는데 유용하지만 '설명이 가능하지 않은' 블랙박스 모델이라는 단점이 있다[18,24].

큐비스트(Cubist)는 RuleQuest(www.rulequest.com)에 의해 개발된 상용 프로그램으로 규칙 기반 모델이다. 회귀 트리 알고리즘과 선형 최소 제곱 회귀의 변형을 사용하여 각 모델링과 결합하는 하이브리드 트리 기반 접근 방식이다. 모델의 분석 결과는 규칙 모음으로 표현되며, 각 규칙에는 연관된 다 변량 선형 모델이 있다. 상황이 규칙의 조건과 일치 할 때마다 관련 모델이 예측 값을 계산하는 데 사용된다. 이러한 선형 모델이 상호 배타적일 필요가 없으므로 최종 예측에 도달하기 위해 출력 값이 평균화된다. 일반적으로 다변량 선형 회귀와 같은 간단한 기술로 생성 된 것보다 더 나은 결과를 제공하는 동시에 신경망보다 이해하기 쉬운 장점이 있다[5,28].

각 모형은 과 적합을 예방하고, 최적의 하이퍼파라미터(hyperparameter) 값을 얻기 위해 반복 10겹 교차 검증(repeated crossvalidation)을 시행하였다.

최종 모델 구축 전 선형회귀모형의 잔차를 통해 극단적인 이상치 관측(DFFITS, Cook's distance)값을 확인하였다[8]. 확인된 7건의 극단적인 데이터는 모델 적합성을 최적화하기 위해 모델개발용 데이터에서 제외 하였다.

2.3.3 모형의 평가

모형평가를 위해 분할된 30%의 데이터(Test data)를 이용해 각 모형의 예측 값을 산출하였다. 예측 값은 Box-Cox 변환의 역 변환한 후 리샘플링(postResample)을 통해 결정계수(R^2), 평균 제곱근 오차(Root mean square error), 절대 평균 오차(Mean absolute error)를 산출하여 각 모형의 성능을 비교 평가 하였다.

3. 연구결과

3.1 분석대상자의 특성

분석대상 미숙아는 6,149명으로 일반적인 특성은 Table 1과 같다. 다 태아 여부는 단 태아 출생이

70.7%, 쌍 태아 출생이 27.8% 였다. 재태 연령은 32주 이상 36주 이하가 68.4%, 37주 이상이 14.1%, 28주 이상 31주 이하가 13.0%, 28주 미만이 4.5%였다. 미숙아 체중은 2,000g이상 2,499g이하가 37.8%, 1,500g이상 1999g이하가 23.9%, 2500g이상이 22.7%, 1000g이상 1499g이하가 11.2%로 나타났다. 분석 대상자의 등록 거주지 행정구역이 읍, 면을 시골로 동을 도시로 구분했을 때 도시지역이 77.4%로 나타났다. 병상 규모는 500병상 이상 999이하가 60.5%, 1000병상 이상 23.0%로 대부분의 분석대상 미숙아가 대형병원 환자로 나타났다.

Table 1. General characteristics of Admission based Variable

Variable	n	%
Single vs. multiple fetuses		
Single	4,346	70.7
Twin	1,712	27.8
Other multiple	91	1.5
Gender		
Male	3,287	53.5
Female	2,862	46.5
Insurance		
Health insurance	6,036	98.2
Medical benefits	82	1.3
Other	31	0.5
Gestational age		
Less than 28 weeks	274	4.5
28-31	798	13.0
32-36	4,207	68.4
37 weeks or more	870	14.1
Weight		
Less than 1000g	273	4.4
1000g-1499g	687	11.2
1500g-1999g	1,467	23.9
2000g-2499g	2,327	37.8
2500g or more	1,395	22.7
Urban		
Rural	1,108	18.0
Urban	4,760	77.4
Unknown	281	4.6
Bed size		
100-299	519	8.4
300-499	491	8.0
500-999	3,723	60.5
1000 or more	1,416	23.0

임상변수 39개 중 상위 15개의 일반적인 특성은

Table 2와 같다. 가장 많은 임상 진단은 기타 주 산기 상태로 분석대상자 6,149명 중 62.8% 환자에게 나타났다. 다음으로는 용혈성 황달 및 주 산기 황달 55.5%, 호흡 곤란 증후군 23.6%, 심장 및 순환기 선천성 기형 20.3%, 패혈증 10.8%, 수술 받은 경우가 4.2%로 나타났다.

Table 2. General characteristics of Clinical based Variable(Top15)

Variable	n	%
Other perinatal conditions	3,859	62.8
Hemolytic jaundice and perinatal jaundice	3,413	55.5
Respiratory distress syndrome	1,450	23.6
Cardiac and circulatory congenital anomalies	1,249	20.3
Septicemia	664	10.8
Presence or absence of Surgery	261	4.2
Retinal disease	239	3.9
Intestinal infection	221	3.6
Other congenital anomalies	179	2.9
Genitourinary congenital anomalies	178	2.9
Nutritional deficiencies	90	1.5
Birth trauma	90	1.5
Abdominal hernia	83	1.3
Digestive congenital anomalies	75	1.2
Other nutritional, endocrine and metabolic disorders	74	1.2

3.1.1 분석대상자의 평균 재원일수

분석대상 미숙아의 평균 재원일수는 23.6일, 중앙값은 14일, 사분위수의 1분위는 8일, 3분위는 30일, 최대값은 174일 이었다. 다 태아 출생여부는 단태 출생자의 평균 재원일수가 24.3일, 쌍 태아 출생이 21.3일, 기타 다 태아 출생이 30.7일로 평균 재원일수에 차이가 있는 것으로 나타났다($p < 0.01$). 태아의 재태 연령은 28주 미만이 23.6일, 28주 이상 31주 이하가 49.1일, 32주 이상 36주 이하가 15.0일, 37주 이상이 18.8일 이었다 ($p < 0.01$). 출생 시 체중은 1000g 미만이 94.4일, 1000g이상 1499g이하 54.5일, 1500g이상 1999g이하 25.9일, 2000g이상 2499g이하 1.3일, 2500g이상 10.9일로 나타났다($p < 0.01$). 병상 규모는 100병상 이상 299병상 이하가 18.1일, 300병상 이상 499일 이하 21.8일, 500병상 이상 999이하 23.8일, 1000병상 이상이 25.7일로 나타났다($p < 0.01$). 분석 대상자의 등록 거주지가 읍, 면 지역인 시골의 경우 23.2일, 도시인 경우 23.7일, 미상 23.4일로 통계적으로 유의미하지 않았다 ($p = 0.82$). Table 3

Table 3. Average length of stay according to characteristics of admission based variable

Variable	n	Mean±SD	F(p)
Total	6,149	23.6±24.8	
Single vs. multiple fetuses			
Single ^a	4,346	24.3±25.9	15.6(<0.01)
Twin ^b	1,712	21.3±21.7	c>a>b
Other multiple ^c	91	30.7±23.3	
Gender			
Male	3,287	23.7±24.5	0.2(0.65)
Female	2,862	23.4±25.1	
Insurance			
Health insurance	6,036	23.6±24.9	1.5(0.22)
Medical benefits	82	22.7±21.8	
Other	31	18.0±18.0	
Gestational age			
Less than 28 weeks ^a	274	95.6±26.7	1522(<0.01)
28-31 ^b	798	49.1±19.6	a>d>b>c
32-36 ^c	4,207	15.0±12.2	
37 weeks or more ^d	870	18.8±21.0	
Weight			
Less than 1000g ^a	273	94.4±28.0	1522(<0.01)
1000-1499 ^b	687	54.5±22.8	a>b>c>d>e
1500-1999 ^c	1,467	25.9±14.0	
2000-2499 ^d	2,327	12.3±9.4	
2500g or more ^e	1,395	10.9±10.8	
Urban			
Rural	1,108	23.2±23.6	0.20(0.82)
Urban	4,760	23.7±25.1	
Unknown	281	23.4±24.0	
Bed size			
100-299 ^a	519	18.1±16.9	1522(<0.01)
300-499 ^b	491	21.8±22.3	d,c,b>a
500-999 ^c	3,723	23.8±24.3	d>b
1000 or more ^d	1,416	25.7±28.8	

3.1.2 임상변수에 따른 평균 재원일수

임상 진단 변수에 따른 평균 재원일수는 기타 주 산기 상태로 진단받은 경우 27.8일로 진단받지 않은 16.5일과 차이가 있었다(p<0.01). 용혈성 황달 및 주 산기 황달을 진단받은 경우 26.1일, 진단받지 않은 경우 20.4일 있었다(p<0.01). 호흡 곤란 증후군을 진단받은 경우 26.1일, 진단받지 않은 경우 20.4일 있었다(p<0.01). 심장 및 순환기 선천성 기형을 진단받은 경우 44.8일, 진단받지 않은 경우 18.2일 있었다(p<0.01). 폐혈증을 진단받은 경우 41.0일, 진단받지 않은 경우 21.5일 있었다(p<0.01). 수술 받은 경우가 84.1일 수술 받지 않은 경

우가 20.9일로 재원일수가 짧았다(p<0.01). 출생 중 외상이 있는 경우 22.1일, 없는 경우가 23.6일로 나타났으나 통계적으로 유의미하지 않았다(p=0.56) Table 4.

Table 4. Average length of stay according to characteristics of Clinical based Variable (Top15)

Variable	n	Mean±SD	t(p)	
Other perinatal conditions	No	2,290	16.5±15.9	-20.1 (<0.01)
	Yes	3,859	27.8±28.0	
Hemolytic jaundice and perinatal jaundice	No	2,736	20.4±24.2	-9.0 (<0.01)
	Yes	3,413	26.1±25.0	
Respiratory distress syndrome	No	4,699	16.6±16.3	-33.4 (<0.01)
	Yes	1,450	46.3±32.7	
Cardiac and circulatory anomalies	No	4,900	18.2±17.9	-26.2 (<0.01)
	Yes	1,249	44.8±34.7	
Septicemia	No	5,485	21.5±22.1	-13.6 (<0.01)
	Yes	664	41.0±36.2	
Presence or absence of Surgery	No	5,888	20.9±20.1	-25.8 (<0.01)
	Yes	261	84.1±39.4	
Retinal disease	No	5,910	21.1±20.6	-26.5 (<0.01)
	Yes	239	85.3±37.3	
Intestinal infection	No	4,699	16.6±16.3	-33.4 (<0.01)
	Yes	1,450	46.3±32.7	
Other congenital anomalies	No	5,970	23.3±24.4	-3.1 (<0.01)
	Yes	179	31.4±34.4	
Genitourinary congenital anomalies	No	5,971	23.4±24.8	-2.7 (0.01)
	Yes	178	28.5±25.6	
Nutritional deficiencies	No	6,059	23.2±24.3	-6.4 (<0.01)
	Yes	90	49.9±39.7	
Birth trauma	No	6,059	23.6±24.7	0.59 (0.56)
	Yes	90	22.1±29.7	
Abdominal hernia	No	6,066	22.8±23.8	-16.0 (<0.01)
	Yes	83	77.8±32.2	
Digestive congenital anomalies	No	6,074	23.5±24.6	-2.4 (0.02)
	Yes	75	33.5±35.8	
Other nutritional, endocrine and metabolic disorders	No	6,075	23.4±24.6	-4.2 (<0.01)
	Yes	74	39.4±32.6	

3.2 예측 모델

3.2.1 변수 중요도

머신러닝의 데이터 셋에서 더 많은 변수를 가지는 것은 모델을 복잡하게 하고 해석을 더 어렵게 하고 과 적합이 발생하기도 한다. 변수 간에 중복이 있는지, 어떤 변수가 중요한 변수인지, 어떤 변수가 목표변수에 영향을 크게 주는 변수인지를 분석할 필요가 있다[23]. 입원 기초변수 모델에서 인공신경망 모형을 통한 변수 중요

도(Variable Importance)를 확인한 결과 출생 시 체중 2000g이상에서 2499g이하, 2500이상, 재태 연령 32주 이상에서 36주이하, 출생 시 체중 1500g이상에서 1999g 이하, 재태 연령 37주 이상 순으로 나타났다<Fig. 2>.

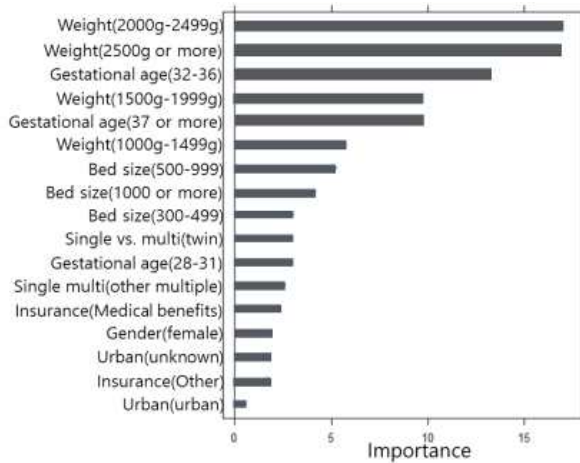


Fig. 2. Admission based Variable importance (Neural Network model)

입원기초변수에 임상변수가 추가 투입된 큐비스트 모형의 변수는 20개의 각 모형에서 생성한 규칙을 통해 알아보았다. 각 규칙에는 재태 연령 32주 이상에서 36주이하, 호흡 곤란 증후군, 출생 시 체중 2000g이상에서 2499g이하, 재태 연령 37주 이상, 출생 시 체중 1500g이상에서 1999g이하 순으로 나타났다. 조건(if) 별 출력(then)된 다 변량 회귀식에서 변수의 출현율을 확인한 결과 추가 투입된 임상변수는 수술여부, 기타 주산기 상태, 호흡 곤란 증후군, 패혈증, 심장 및 순환기 선천성 기형, 장의 감염 순이었다. Table 5

Table 5. Variables used in the 20 committee, 0 neighbors Cubist model

Percent appearance in a rule conditions (%)	Percent appearance in a multivariate regression models (%)	Variable
50%	65%	Gestational_age(32-36)
50%	87%	Respiratory distress syndrome
6%	99%	Weight(2000g-2499g)
2%	67%	Gestational_age(37 or more)
1%	99%	Weight(1500g-1999g)
	98%	Weight(2500g or more)
	94%	Presence or absence of Surgery

88%	Other perinatal conditions
79%	weight(1000g-1499g)
76%	Septicemia
76%	Bed size(1000 or more)
74%	Cardiac and circulatory anomalies
73%	Intestinal infection
70%	gestational_age(28-31)
56%	Hemolytic jaundice and perinatal jaundice
53%	Bed size(500-999)
43%	Bed size(300-499)
41%	Single vs. multi(twin)
38%	Other congenital anomalies
32%	Retinal disease
24%	Gender(female)
20%	Urban(urban)
10%	Birth trauma
1%	Urban(unknown)

3.2.2 모델별 매개변수

모델의 학습 효율을 결정하는 각각의 하이퍼파라미터(hyper-parameter)는 caret 패키지의 튜닝 기능을 이용하여 RMSE값이 최소화된 상태 값을 설정 하였다. 반복 5회 10겹 교차 검증(repeated cross validation)을 사용하였다. 튜닝 길이(tuneLength)는 튜닝 성능과 시스템의 효율성을 고려하여 20으로 설정하였다. 튜닝 길이는 튜닝 매개 변수 설정을 위한 그리드의 크기를 제어 하여 복잡한 매개변수 후보 값을 선택한다. 크게 설정하면 성능은 좋아지지만 시간이 많이 걸리는 등 효율성이 떨어질 수 있다[28]. 구체적인 모델별 하이퍼파라미터 값은 Table 6과 같다. 별점화 회귀 모형의 알파는 1로 라쏘(LASSO)회귀 모형을 사용 하였다.

Table 6. Hyperparameter optimal tuning value for each model

Models	Admission based Models of parameter	Admission and Clinical Models of parameter
GLMN	alpha=1 lambda=0.0003140304	alpha=1 lambda=0.001275978
RF	mtry=5	mtry=7
GBM	n.trees=100, interaction.depth=5, shrinkage=0.1, n.minobsinnode=10	n.trees=1000, interaction.depth=2, shrinkage=0.1, n.minobsinnode=10
NNT	size=3, decay=0.06812921	size=3, decay=0.1
CUB	committees=20, neighbors=0	committees=20, neighbors=0

3.2.3 모형의 예측

입원초기 신경망 모형은 17개의 투입변수로 1개의 은닉층에 3개의 노드를 가지는 모형이다. 활성함수로 선형출력이 사용되었다. Fig. 3. 생성된 신경망 모형에 검증 데이터(Test data)를 투입하여 예측된 재원일수를 그래프로 표현하면 Fig. 4와 같았으며 모형의 설명력(R²)는 0.75였다. 검증용 재원일수 최소값이 0일, 최대값이 162일인 반면 예측 값의 최소값 7일, 최대값 116일로 단기 입원 환자와 장기 입원 환자의 예측 성능이 현저히 떨어지는 것으로 확인 되었다. 반면 입원초기변수에 임상변수가 추가 투입된 큐비스트 모델에서는 장기 입원환자의 예측이 여전히 나쁘지만 비교적 고른 분포를 보였다. 모형의 설명력(R²)는 0.81로 향상 되었다. Fig. 5

입원초기변수에 임상변수가 추가 투입된 큐비스트 모형의 결과는 Table 7과 같은 규칙조건(if)에 따른(then) 다변량회귀식을 산출하였다. 재태 연령이 32주 이상에서 36주 이하에 속하지 않고, 출생 시 체중 2000g이상에서 2499g이하인 경우의 다변량 회귀식은 $2.020307+1.34 \times \text{수술여부}+0.63 \times \text{재태 연령}$ 28주 이상에서 31주 이하+ $0.4 \times \text{호흡 곤란 증후군}+0.24 \times \text{순환기 선천성 기형}-0.22 \times \text{병상규모}$ 1000명 이상+ $0.17 \times \text{용혈성 황달}$ 및 주 산기 황달+ $0.12 \times \text{기타 주 산기 상태}-0.11 \times \text{출생 시 체중}$ 2000g이상에서 2499g이하- $0.12 \times \text{2500g이상}+0.28 \times \text{장 감염(ccs135)}+0.14 \times \text{패혈증}+0.06 \times \text{병상 규모}$ 500병상 이상 999병상 이하- $0.06 \times \text{출생 시 체중}$ 1500g이상에서 1999g 이하- $0.05 \times \text{성별}-0.05 \times \text{재태 연령}$ 32주 이상에서 36주이하- $0.06 \times \text{재태 연령}$ 37주 이상의 회귀식으로 출력되었으며, 이 회귀식이 같은 규칙 중 추정된 절대평균오차가 가장 낮게 나왔다(MAE 0.35). 이런 형태의 10개의 규칙은 20개의 모델에서 반복 산출되어 47개의 규칙을 출력하였으며, 이 중 추정된 절대평균오차가 가장 작은 다변량 회귀식은 Table 9와 같다.

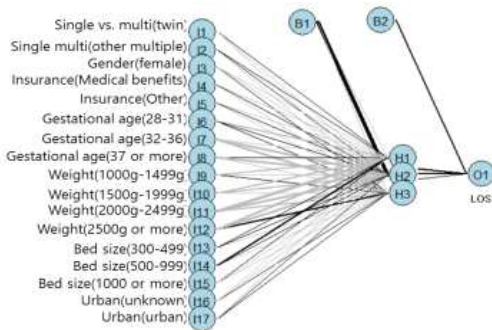


Fig. 3. Admission-based neural network model

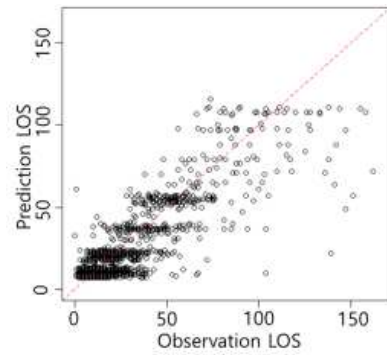


Fig. 4. Admission based prediction of neural network model (R² 0.75)

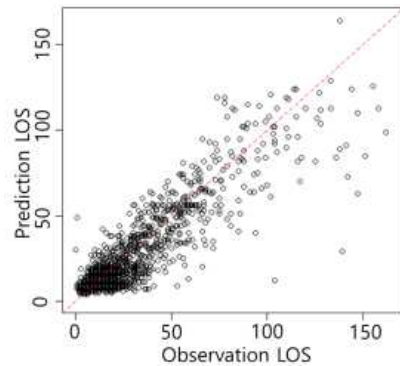


Fig. 5. Admission and Clinical based prediction of Cubist model (R² 0.81)

Table 7. Ten Cubist rules with smallest Mean Absolute Error (MAE).

Rule	Consequent	MAE
gestational_age3 <= 0, weight4 > 0	$2.020307+1.34\text{opyn1}+0.63\text{gestational_age2}+0.4\text{ccs221}+0.24\text{ccs213}-0.22\text{bedsize4}+0.17\text{ccs222}+0.12\text{ccs224}-0.11\text{weight4}-0.12\text{weight5}+0.28\text{ccs135}+0.14\text{ccs2}+0.06\text{bedsize3}-0.06\text{weight3}-0.05\text{gender1}-0.05\text{gestational_age3}-0.06\text{gestational_age4}$	0.35
gestational_age3 > 0	$3.718505-1.45\text{weight4}-1.64\text{weight5}-0.92\text{weight3}+0.86\text{opyn1}+0.31\text{ccs221}-0.4\text{weight2}-0.16\text{gestational_age3}+0.16\text{ccs213}-0.18\text{gestational_age4}-0.14\text{bedsize4}+0.16\text{ccs2}+0.27\text{ccs135}+0.1\text{ccs224}+0.05\text{ccs222}-0.04\text{single_multi2}-0.05\text{gestational_age2}$	0.33
gestational_age3 <= 0, weight3 > 0	$3.859788-0.82\text{weight4}-0.58\text{weight3}-0.57\text{weight5}-0.64\text{gestational_age4}+0.25\text{ccs221}-0.23\text{weight2}+0.14\text{bedsize3}-0.14\text{gestational_age3}+0.33\text{opyn1}-0.19\text{gestational_age2}+0.2\text{bedsize2}+0.07\text{ccs224}+0.1\text{ccs2}-0.05\text{gender1}+0.06\text{ccs213}+0.11\text{ccs135}+0.06\text{ccs87}+0.02\text{ccs222}$	0.34
gestational_age3 <= 0, weight3 <= 0, weight4 <= 0	$4.295847-0.52\text{weight5}-0.42\text{gestational_age4}-0.37\text{gestational_age2}-0.28\text{weight3}-0.32\text{weight2}-0.2\text{weight4}+0.18\text{ccs221}+0.37\text{opyn1}+0.13\text{ccs224}+0.19\text{ccs2}+0.1\text{bedsize3}-0.05\text{ccs222}+0.09\text{ccs87}+0.06\text{bedsize2}-0.03\text{gestational_age3}$	0.31

ccs2211 <= 0	4.253301-1.34weight4-1.52weight5-0.86gestational_age3-1gestational_age4-0.8weight3+0.79opyn1-0.34gestational_age2+0.23ccs213-0.28weight2+0.38ccs87+0.33ccs217+0.1ccs222+0.1bedsize3+0.26ccs135-0.11bedsize4+0.08ccs224+0.14bedsize2+0.12ccs2-0.05single_multi2-0.05urban1+0.03ccs2211-0.02gender1	0.35
ccs2211 > 0	4.446906-1.21weight4-1.3weight5-0.76weight3-0.69gestational_age3-0.31weight2-0.26gestational_age4-0.27gestational_age2+0.17ccs224+0.32opyn1+0.18ccs2-0.1ccs222+0.15ccs87	0.31
gestational_age3 <= 0, weight4 <= 0	4.308512-0.97weight3-0.62weight4-0.58gestational_age4-0.4weight2+0.25bedsize3-0.36gestational_age2+0.25ccs221+0.21weight5+0.26bedsize2+0.32opyn1+0.17ccs2+0.17urban9+0.08bedsize4	0.35
gestational_age4 > 0	4.092114-1.95weight4-1.16weight3+1.16opyn1-0.25gestational_age4-0.19weight5+0.13ccs222-0.2weight2-0.08gestational_age3+0.06ccs221-0.06gestational_age2+0.03bedsize3+0.03ccs224+0.04ccs2	0.40
gestational_age3 <= 0, gestational_age4 <= 0	4.674826-1.29weight4-0.9weight3-0.53weight2-0.39gestational_age2+0.19ccs221-0.2gestational_age4+0.14weight5-0.11ccs222+0.18bedsize2+0.25opyn1+0.1bedsize3+0.1ccs224-0.04gestational_age3+0.16ccs143+0.04ccs2	0.34
gestational_age3 <= 0	4.464732-1.91weight4-1.07weight3-0.64gestational_age4+0.72opyn1-0.39weight2-0.26gestational_age2+0.15bedsize3+0.12ccs221-0.1weight5-0.04gestational_age3	0.37

3.3 모형평가

입원초기모델과 임상변수가 추가 투입된 모델을 통해 예측모형의 성능을 비교 하였다. 개발된 예측 모형은 인공신경망, 큐비스트, 그라디언트 부스트, 랜덤 포레스트, 벌점화 회귀 모형 이었다. 각 모형 설명력(R²), 절대 평균 오차(MAE), 평균 제곱근 오차(RMSE)를 비교한 결과 입원 초기 모델에서는 인공신경망 모형이 성능이 가장

Table 8. Evaluation of admission based models

Models	RMSE	R ²	MAE
Neural Network	12.71	0.75	7.93
Cubist	12.97	0.74	7.94
Gradient Boosting	12.86	0.74	7.97
Random Forest	12.89	0.74	7.99
Penalized Regression	13.46	0.72	8.17

Table 9. Evaluation of the admission and clinical based models

Models	RMSE	R ²	MAE
Cubist	11.03	0.81	7.00
Gradient Boosting	11.15	0.80	7.04
Neural Network	11.12	0.80	7.08
Random Forest	11.37	0.80	7.15
Penalized Regression	11.44	0.79	7.29

좋았고(R² 0.75), 임상변수가 추가 투입된 모델에서는 큐비스트 모형(R² 0.81)이 성능이 가장 뛰어난 것으로 확인 되었다. 또한 입원초기 확인 가능한 변수만으로도 재원일수 예측이 가능 했으며, 임상 진단 변수를 추가 투입함으로써 모형의 설명력(R²)이 향상되고 오차가 줄어드는 것을 확인 하였다. Table 8, Table 9.

4. 고찰

본 연구는 전국단위의 조사가 이루어지는 퇴원환자심층조사 자료를 이용하였다. 전국 단위의 표본 자료를 이용한 미숙아의 재원일수 예측을 위해 다양한 머신러닝 기법을 탐색하고 최적의 모형을 제안하는 것을 목표로 하였다. 미숙아는 출생과 동시에 입원 치료하게 되는 특성을 고려하여 신생아실이나 분만실을 통해 입원되는 환자만을 대상으로 하였으며, 타 병원 전원과 치료 중 사망을 제외하였다. 머신 러닝에는 무엇보다도 각 변수들의 이해가 중요하며 피쳐 엔지니어링(feature engineering)을 통해 변수의 변환, 표준화, 정규화, 그룹화, 중요도에 따른 선택 또는 제거 등이 이루어진다. 이러한 과정을 거쳐 생성되는 모델은 전통적인 통계 이론에 기반 하지만 그 목적은 예측에 더 중점을 둔다. 선행연구에서 제안하였던 다양한 머신러닝 기법을 고려하여 모형을 구현하고 비교하였다. 최종 제안된 큐비스트 모형은 예측의 해석이 용이하고, 예측 모형에 적용된 규칙 뒤에 있는 논리를 확인 할 수 있을 뿐만 아니라 재원일수에 영향을 미칠 수 있는 변수를 파악하고 이해하는데 장점이 있다. 큐비스트 모형을 적용한 울혈성 심부전 환자 재원일수를 예측한 Turgeman의 사례에서 학습 데이터에서 R² 0.84의 결과를 보였다[5]. 이 데이터는 입원병력, 외래병력, 활력징후, 혈액검사, 동반질환에 대한 보다 다양한 임상 데이터를 사용할 수 있었던 차이가 있음을 고려한다면 본 연구 결과와 유사한 예측 성능을 보였다. Singh는 재태 연령에 따라 그룹을 나누고 산산기 및 산전 요인, 영양편차, 약물, 임상진단에 대한 독립 변수를 이용하여 회귀모형을 구축하였다. 검증 데이터에서 32주 미만의 R²이 0.95, 32주 이상 34주 미만에서 0.74, 34주 이상 37주 미만에서 0.62를 보였다[15]. 본 연구에서도 입원초기모형은 장기입원에 대하여 예측 성능이 떨어지고, 예측값의 분포가 고르게 나타나지 않았다. 그럼에도 불구하고 입원초기변수가 기본적인 7개로 제한적이라는 점을 고려한다면 인공신경망 모형이 R²

0.75의 결과는 의미 있는 결과이다. 선행연구에서도 급성심근경색 및 심부전환자의 인공신경망 모형을 이용한 재원일수 예측에서 좋은 성능을 보였다[27]. 뉴 잉글랜드 지역 신생아중환자실 환자의 재원일수를 예측한 또 다른 사례에서는 출생 체중, 재태 연령, 신생아 급성생리학점수(SNAP-PE : score for neonatal acute physiology perinatal extension), 신생아이환율평가 지수(MAIN : morbidity assessment index for newborns)를 사용하여 예측하였다.(R^2 0.79). 이 모델은 출생 시 알려진 요인을 넘어 생후 첫 주 질병을 판단하여 모델링함으로써 성능이 개선된다고 제안 하였다[12]. 본 연구에서는 신생아 급성생리학점수와 신생아이환율평가 지수 대신 수술여부와 CCS 임상진단 분류만을 사용하였음에도 입원초기변수에 임상변수를 추가 투입한 모델에서는 모형 설명력 R^2 이 0.81의 큐비스트 모형을 구현할 수 있었다.

기존 전통적 통계기법이 통계적 가정을 기반으로 재원일수에 영향을 미치는 변수를 찾아내는데 효과적이었고, 재원일수 관리를 위한 방안을 제시하는데 맞춰진 사례가 많았다[13]. 본 연구는 재원일수를 입원 초기에 예측하고 예측 자료를 기반으로 보호자에게 설명의 기초 자료를 제공하는데 활용하는 것을 고려하였다. 이러한 활동은 보호자에게 치료에 대한 불확실성을 제거해 주는데 도움이 된다. 본 연구를 통해 제한된 데이터만으로도 머신러닝의 여러 알고리즘을 통해 재원일수의 예측 가능성을 확인 하였다. 미숙아의 특성상 태어나면서부터 알 수 있는 기초적인 정보 즉, 재태 연령, 출생 시 체중, 출생 직후 활력징후, 엄마의 기본 정보 및 병력 등 출생 병원에서는 누락되기 힘든 기본 정보가 있다. 추가 적인 임상 및 부모의 정보가 결합된다면 예측 성능의 향상도 기대된다.

본 연구에서는 미숙아 환자의 재원일수를 입원초기 기본정보 또는 입원 후 임상 진단 정보로 예측 가능하고, 여러 머신러닝 기법을 통해 그 예측 성능을 확인 하였다는데 본 연구의 의의를 갖는다.

5. 결론

본 연구는 미숙아의 입원초기 기본정보 또는 임상 진단 자료로 재원일수의 예측을 시도 하였다. 재원일수 예측에 있어 여러 머신 러닝 기법을 비교하여 입원초기에는 신경망 모형, 임상 진단 변수 추가 투입 모형에는 큐

비스트 모형이 성능이 뛰어난 것으로 확인 되었다. 미숙아의 특성상 입원초기 또는 입원 일주일 후 많은 임상정보가 확인 가능해지고 이를 통해 재원일수를 예측 하는 것은 의료기관의 병상 자원을 효율적으로 관리 할 수 있고 의료 질 개선의 노력을 촉진 할 수 있다. 또한 머신러닝 기법의 적절한 활용으로 보다 정확한 예측이 가능해짐으로써 이를 활용한 보호자 설명의 기초 자료로 활용할 수 있다.

본 연구에 사용된 데이터의 한계로 포함되지 못한 산모의 정보, 태아의 출생 시 상태 즉, 아프가 점수(Apgar score) 또는 신생아이환율평가 지수 등과 같은 자료를 포함한 좀 더 고도화된 예측 모델과 사망 또는 미숙아 외의 신생아 중환자실 환자에 대한 추가 연구를 통해 신생아 중환자실 자원의 효율적 관리를 위한 예측 모델도 가능할 것이다. 전국단위 데이터를 이용하여 머신러닝을 이용한 미숙아의 재원일수 예측 모형을 제시하고 그 활용 가능성을 확인 했었는데 본 연구에 의의가 있다.

REFERENCES

- [1] Korean Statistical Information Service(2021). Population dynamics and dynamics rate. *KOSIS(Online)*. https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B8000F&conn_path=12
- [2] H. S. Lee. & S. J. Kang. (2020). Factors Influencing Post-traumatic Growth in Mothers with Premature Infants Admitted to the Neonatal Intensive Care Unit. *Child Health Nursing Research*, 26(2), 267-276. DOI : 10.4094/chnr.2020.26.2.267
- [3] Yuri Lee. (2016). A Comparative Study on the Support Policy for Prematurity in Some OECD Countries. *Health and Social Welfare Review*, 36(4), 63-96. DOI : 10.15709/hswr.2016.36.4.63
- [4] B. H. Kang, K. A. Jung, W. H. Hahn, K. S. Shim, J. Y. Chang & C. W. Bae. (2011). Regional Analysis on the Incidence of Preterm and Low Birth Weight Infant and the Current Situation on the Neonatal Intensive Care Units in Korea, 2009. *Journal of the Korean Society of Neonatology*, 18(1), 70. DOI : 10.5385/jksn.2011.18.1.70
- [5] L. Turgeman, J. H. May & R. Sciuilli. (2017). Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. *Expert Systems with Applications*, 78, 376-385. DOI : 10.1016/j.eswa.2017.02.023
- [6] S. E. Seaton, L. Barker, E. S. Draper, K. R. Abrams, N. Modi & B. N. Manktelow. (2018). Estimating neonatal

- length of stay for babies born very preterm. *Archives of Disease in Childhood – Fetal and Neonatal Edition*, 104(2), F182–F186.
DOI : 10.1136/archdischild-2017-314405
- [7] Health insurance review and Assessment services. (2019). Results of the 1st neonatal intensive care unit adequacy evaluation. *HIRA(Online)*.
https://www.hira.or.kr/cms/open/04/04/12/2019_07.pdf
- [8] H. C. Lee, M. V. Bennett, J. Schulman & J. B. Gould. (2013). Accounting for variation in length of NICU stay for extremely low birth weight infants. *Journal of Perinatology*, 33(11), 872–876.
DOI : 10.1038/jp.2013.92
- [9] C. Li et al. (2019). Prediction of Length of Stay on the Intensive Care Unit Based on Least Absolute Shrinkage and Selection Operator. *IEEE Access*, 7, 110710–110721.
DOI : 10.1109/access.2019.2934166
- [10] S. Grampurohit & S. Sunkad. (2020). Hospital Length of Stay Prediction using Regression Models. In 2020 IEEE International Conference for Innovation in Technology (INOCON). *IEEE*.
DOI : 10.1109/inocon50539.2020.9298294
- [11] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, & S. Yoo. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLOS ONE*, 13(4), e0195901.
DOI : 10.1371/journal.pone.0195901
- [12] G. J. Bender, D. Koestler, H. Ombao, M. McCourt, B. Alskinis, L. P. Rubin & J. F. Padbury. (2013). Neonatal intensive care unit: predictive models for length of stay. *Journal of Perinatology*, 33(2), 147–153.
DOI : 10.1038/jp.2012.62
- [13] B. K. Choi, S. W. Ham, C. H. Kim, J. S. Seo, M. H. Park, & S. H. Kang. (2018). Development of Predictive Model for Length of Stay(LOS) in Acute Stroke Patients using Artificial Intelligence. *Journal of Digital Convergence*, 16(1), 231–242.
DOI : 10.14400/JDC.2018.16.1.231
- [14] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig & S. Koblar. (2020). Machine Learning in the Prediction of Medical Inpatient Length of Stay. *Internal Medicine Journal*.
DOI : 10.1111/imj.14962
- [15] H. Singh et al. (2021). Designing a bed-side system for predicting length of stay in a neonatal intensive care unit. *Scientific Reports*, 11(1).
DOI : 10.1038/s41598-021-82957-z
- [16] B. Thompson, K. O. Elish, and R. Steele. (2018). Machine Learning-Based Prediction of Prolonged Length of Stay in Newborns. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA).
DOI : 10.1109/ICMLA.2018.00236
- [17] D. J. Whellan et al. (2011). (2011). Predictors of Hospital Length of Stay in Heart Failure: Findings from Get With the Guidelines. *Journal of Cardiac Failure*, 17(8), 649–656.
DOI : 10.1016/j.cardfail.2011.04.005
- [18] J. H. Park & S. H. Kang. (2019). A Study on the Development of Predictive Model for Severity-Adjusted Length of Stay in Nervous System Patients Using Machine Learning. *Health and Social Welfare Review*, 39(1), 390–427.
DOI : 10.15709/hswr.2019.39.1.390
- [19] J. B. Lee & H. Woo. (2020). Determinants of Length of Stay in Ischemic Heart Disease Patients. *Journal of Health Informatics and Statistics*, 45(1), 52–59.
DOI : 10.21032/jhis.2020.45.1.52
- [20] Y. H. Choi, Y. J. Kim & S. H. Kang. (2015). The Development of Convergence Bench-making system on length of stay. *Journal of digital convergence*, 13(5), 89–99.
DOI : 10.14400/JDC.2015.13.5.89
- [21] S. O. Hong, Y. T. Kim, J. H. Park & S. H. Kang (2015). The Variation of Factors of Severity-Adjusted Length of Stay (LOS) in Injury of Neck. *Health and Social Welfare Review*, 35(2), 561–583.
DOI : 10.15709/hswr.2015.35.2.561
- [22] M. Kuhn, (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5).
DOI : 10.18637/jss.v028.i05
- [23] J. Elrod et al. (2021). Using Artificial Intelligence to Obtain More Evidence? Prediction of Length of Hospitalization in Pediatric Burn Patients. *Frontiers in Pediatrics*, 8.
DOI : 10.3389/fped.2020.613736
- [24] Y. C. Woo, S. Y. Lee, W. Choi, C. W. Ahn & O. K. Baek. (2019). Trend of Utilization of Machine Learning Technology for Digital Healthcare. *Data Analysis, Electronics and Telecommunications Trends*, 34(1), 98–110.
DOI : 10.22648/ETRI.2019.J.340109
- [25] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg & O. Lyashevskaya. (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of translational medicine*, 7(7), 152–152.
DOI : 10.21037/atm.2019.03.29
- [26] Y. Zhang & A. Haghani. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324.
DOI : 10.1016/j.trc.2015.02.019
- [27] P. F. (Jennifer) Tsai et al. (2016). Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. *Journal of Healthcare Engineering*, 2016, 1–11.
DOI : 10.1155/2016/7035463

- [28] Q. Shao, R. C. Rowe, and P. York. (2007). Investigation of an artificial intelligence technology—Model trees. *European Journal of Pharmaceutical Sciences*, 31(2), 137-144.
DOI : 10.1016/j.ejps.2007.03.004

김 축 환(Chok Hwan, Kim)

[경력]



- 2006년 8월 : 순천향대학교 산업정보대학원 의료공학과 의료정보학전공 (공학석사)
- 2003년 6월 ~ 현재 : 순천향대학교 천안병원 의료정보팀
- 관심분야 : 전자의무기록, 의료데이터 분석, 머신러닝

· E-Mail : kchkam@naver.com

강 성 홍(Sung-Hong Kang)

[경력]



- 1990년 2월 : 서울대학교 보건대학원 보건관리학과 (보건학석사)
- 1997년 2월 : 인제대학교 일반대학원 보건학과 (보건학박사)
- 1998년 3월 ~ 현재 : 인제대학교 보건행정학과 교수
- 관심분야 : 보건정보, 의무기록, 데이터마이닝, 건강증진

· E-Mail : hcmkang@hanmail.net