ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 52, No.4: 593-602, December 2024
https://dx.doi.org/10.7469/JKSQM.2024.52.4.593

국민 건강 검진 데이터 기반 혈색소(헤모글로빈) 예측 모델링

정대원* · 황욱연*†

* 동아대학교글로벌비즈니스학과

The Hemoglobin Prediction Modeling Based on the National Health Data

Dae Won Jung* · Wook-Yeon Hwang*†

* Department of Global Business, Dong-A University

ABSTRACT

Purpose: Leveraging on the contemporary machine learning algorithms, we would like to improve the prediction performance of the existing MLR(Multiple Linear Regression) model to predict the blood hemoglobin levels.

Methods: The GBDT (Gradient Boosting Decision Trees) such as the XGBoost (Extreme Gradient Boosting), the LightGBM (Light Gradient Boosting Machine), and the CatBoost (Categorical Boost), the RF(Random Forests), and the MLP (Multi-Layer Perceptron) are adopted to build the new prediction models.

Results: The machine learning algorithms provide prediction performance better than the existing prediction model.

Conclusion: The proposed prediction models can be considered as an alternative better than the existing prediction model.

Key Words: The National Health Data, Blood Hemoglobin Levels, Machine Learning Algorithms, Prediction Model

● Received 23 August 2024, 1st revised 19 September 2024, accepted 25 September 2024

† Corresponding Author(wyhwang@donga.ac.kr)

* 이 논문은 2023학년도 동아대학교 연구년 지원에 의하여 연구되었음

1. 서 론

혈색소(헤모글로빈)는 혈액 내 산소와 이산화탄소를 운반하는 주요 단백질로, 건강한 신체 기능을 유지하는 데 중요한 역할을 한다. 세계보건기구(World Health Organization, WHO)에서 발표한 정상 혈색소 기준 수치보다 혈색소가 낮거나 높게 측정되면 영양 결핍과 질병을 의심할 수 있다. 대표적인 혈색소 수치와 관련된 질환이 빈혈이며, 혈색소의 구성 성분 중 하나인 철이 부족할 경우 혈색소가 낮게 측정되고 충분히 기능하지 못하여 빈혈이 발생한다(Llanos, 2016). 또한, 기준 수치보다 높게 측정되면 당뇨병, 지방간 등 대사성 질환과 심혈관 질환의 발생률과 사망률이 높아진다(Koivune et al., 2020). 이처럼 혈색소 수치의 이상은 다양한 건강 문제를 일으킬 수 있으므로 철저한 관리가 요구된다. 2022년 한국질병관리본부가 발표한 2021년 국민건강영양조사에 따르면(KDCA, 2021), 빈혈 유병률은 전체 인구의 9.3%로 보고되었고 여성은 남성에 비해 빈혈에 더 취약하며, 특히 70세 이상의 여성에서 빈혈 발생률이 매우 높다. 노인의 낮은 혈색소 수치는 건강상의 문제에 치명적이며, 65세 이상 노인에게 빈혈이 발생하면, 인지 기능의 장애, 치매 위험 증가, 적은 움직임으로 말미암은 골밀도 및 근육량 감소, 우울증 등의 문제가 발생할 수 있다(Jeong, 2011). 노인뿐만 아니라 심부전, 인간면역결핍바이러스(HIV) 감염, 암 등의 질환을 앓는 환자가 빈혈이 발생하면 사망률이 증가한다는 연구 결과도 있으며(Ezzekowitz et al., 2003), 이처럼 빈혈은 여러 방면에서 치명적인 건강 문제를 일으킬 수 있다.

이러한 연구 결과는 빈혈의 잠재적 심각성을 강조하지만, 많은 사람이 이에 대한 문제를 간과하고 있다. 특히, 최근 코로나19와 같은 팬데믹 상황에서 건강에 관한 관심이 증가하고 있으며, 비대면 진료의 중요성이 높아지고 있지만 기존의 혈색소 측정 방법은 대면 검사만 가능하며 침습적이고 시간이 많이 소요되어 빈혈을 확인하는 것에 어려운 점이 있다.

따라서 기존 방법의 한계점을 보완하기 위해 Hong and Hong(2021a)은 국민 건강 데이터를 사용하여 다중선형회귀분석(Multiple Linear Regression, MLR) 기반의 혈색소 예측 모델링을 제안하였다. Kavsaoglu et al.(2015)은 광혈류조영술(Photoplethysmography, PPG)센서 데이터와 다양한 건강 정보 데이터를 통한 혈색소 예측 모델링을 제안하였다. Yun(2020)은 시계열 데이터를 GRU(Gated Recurrent Unit) 모델에 적용하여 혈색소 수치 예측 모델을 구축하였다. 우리는 Hong and Hong(2021a)이 제안한 국민 건강 데이터를 사용한 MLR의 예측 성능을 향상시키기 위해 머신 러닝 알고리즘 기반의 새로운 혈색소 예측 모델을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 기존 혈색소 예측 모델들을 소개하고 제안되는 연구를 기술하였다. 3장에는 예측에 이용되는 이론적 배경을 소개하며 4장에서 실험과 결과를 소개한다. 마지막으로 5장에서 결론을 기술하였다.

2. 선행연구 및 제안되는 연구

2.1 혈색소 예측의 선행연구

혈색소를 예측하기 위한 기존 연구에는 PPG 신호의 특성을 이용한 비침습적 헤모글로빈 수치 예측이 있다. Kavsaoglu et al.(2015)은 PPG 센서 신호와 환자의 간단한 신체 정보를 이용하여 혈색소 예측을 진행하였으며, 헤모글로빈 수치를 예측했을 때 성능은 RFS(Relief-based Feature Selection)를 통해 선정된 10가지 PPG 신호와 신체 정보를 사용한 SVR(Support Vector Regression)모델이 매우 좋은 결과를 보여주었다. 하지만 PPG 광신호를 사

용한 데이터를 수집하고 이용하기 어려운 문제점이 존재한다.

Yun(2020)의 말기 신부전 환자에서 순환 신경망을 이용한 혈색소 수치 예측과 조혈제 용량 권고 알고리즘 개발 연구가 있다. 이 연구에서는 수도권의 7개 대학병원에서 외래 혈액투석 중인 환자 466명을 대상으로 최대 5년간의 임상 데이터를 사용하였다. 그리고 혈색소 예측을 위해 MLR, MLP, GRU 그리고 XGBoost을 이용하여 예측 모형을 제작하고 성능을 비교 분석하였다. 연구 결과, GRU와 Gaussian noise layer를 결합한 모델이 가장 높은 성능을 보였지만, MLR을 제외한 다른 알고리즘들은 비슷한 성능을 보였다. 하지만 데이터가 특정 환자 층에 한정되어 있고 추적 연구의 특성상 시계열 데이터를 사용했기에 보편적으로 수집하기 어려운 변수들이 사용되었다. Ghosh et al.(2023)의 연구에서는 다양한 머신 러닝 알고리즘들을 사용하여 혈색소 수치를 추정하고 빈혈의 중증도를 예측하였다. 데이터는 BSMMU(Bangabandhu Sheikh Mujib Medical University)에서 수집한 전혈구계산치(Complete Blood Count, CBC)를 사용하였으며, 데이터 전처리 후 회귀 및 분류 기반 모델을 개발하여 비교했다. 신경망 모델이 혈색소 수치 추정에서 가장 높은 정확도와 낮은 오차로 우수한 성능을 보였고, RF를 이용한 모델이 빈혈 중증도 예측에서 가장 좋은 성능을 보여주었다. Dhakal et al.(2023)의 연구는 5세 이하 아동의 CBC 보고서를 사용하여 빈혈을 예측하기 위한 모델을 설계하였다. 데이터는 Kanti Children Hospital에서 수집한 700개의 데이터 기록으로 구성되었으며, 데이터 전처리 후 다양한 머신 러닝 알고리즘들을 적용하여 모델을 개발하였다. RF가 98.4%의 정확도로 최고 성능을 보였다. 그러나 특정 연령대의 데이터에 국한되었고 일반화의 어려움이 있었다. Aghajanian et al.(2024)의 연구는 산전 임상 데이터와 실험실 측정을 사용하여 출산 후 혈색소 수치를 예측하기 위해 머신 러닝 알고리즘들을 적용하였다. 데이터는 두 개의 학술 의료 센터에서 수집되었으며, 데이터 전처리 후 RF에 기반한 변수 선택을 거쳐 다양한 머신 러닝 알고리즘들을 사용하여 출산 후 혈색소 수치를 예측하였다. 신경망 모델이 가장 예측 좋은 예측 정확도를 보였다. 그러나 다양한 인구 기반 샘플을 통한 추가 연구가 필요하였고 특정 산모 집단에 국한되어 일반화가 어려운 한계점이 있다.

또한 Hong and Hong(2021a)의 건강검진 빅데이터를 이용한 MLR 기반 혈색소 추정 방법에 관한 연구가 있다. 이 연구에서 그들은 건강보험공단에서 제공하는 국민건강검진 데이터(2014, 2015, 2017)를 이용하였고, MLR을 기반으로 하는 혈색소 추정 방법을 제안하였다. 피어슨 상관계수를 통해 혈색소와 높은 상관성을 보이는 7개의 변수를 선정하였고 그 7가지 변수 모두를 사용한 MLR이 가장 낮은 오차율을 보여주었다. 하지만 선형 상관관계를 통한 변수 선택을 진행하고 있어 비선형 상관관계를 고려하지 않은 문제점과 예측 모델로 MLR만 고려하고 있다는 한계가 존재하였다. Hong and Hong(2021b)의 다른 기존 연구는 신체 및 건강 정보를 이용한 회귀분석 기반 혈색소 추정이 있다. 이 연구에서는 국민건강검진 데이터(2014-2018)를 이용하였고, 기존 연구와 동일한 변수를 사용했지만 성별, 나이, 흡연의 유무로 집단을 구분함으로써 기존 연구와 차별화된 MLR을 기반으로 한 혈색소 추정 방법을 제안하였다. 남성 청년 흡연자 집단에서 가장 낮은 오차율이 계산되었고, 그 3가지 특징을 가진 집단들이 다른 집단들보다 더 낮은 오차율을 보였다. 하지만 이전 연구와 같이 변수 선택의 불편함과 MLR에 대한 한계점이 존재하였다.

2.2 제안되는 연구

우리는 기존 Hong and Hong(2021a) 연구와 같은 국민건강검진 데이터를 이용하되, MLR 대신 머신 러닝 및 딥러닝 알고리즘들을 적용하여 기존 연구 결과의 성능을 향상시킨 모델을 제안한다. 더 상세하게 설명하면 그들의 기존 연구에서 혈색소와 독립변수 간 상관계수를 활용하여 변수 선택을 고려하는 MLR과는 달리, 가능한 모든 변수를 고려하면서 머신 러닝 및 딥러닝 알고리즘들을 적용하여 최적의 예측모델을 구축하고자 한다. 이는 변수 간 복잡한 상호작용을 반영할 수 있게 하며 다양한 연령층과 건강 상태를 포함하는 데이터를 모두 사용하여 모델의 범용성을

높일 수 있다. Yun(2020)의 기존 연구에서 가장 예측 성능이 좋은 모델인 GRU 모델은 시계열 데이터에 효과적인 성능을 보여주는 신경망 모델이다. 그러나 본 연구에서는 시계열 데이터를 고려하지 않으므로 GRU는 제외하고 GBDT(LightGBM, XGBoost, CatBoost), RF, MLP 기법들을 이용하여 예측 모델을 구축한다.

본 연구는 Figure 1과 같이 진행되며 Hong and Hong(2021a)이 고려한 국민건강검진데이터(2014, 2015, 2017)를 훈련, 검증, 실험 데이터로 나누어 실험을 진행하고 다양한 예측 정확도를 평가하여 가장 예측 성능이 좋은 모델을 관찰한다.

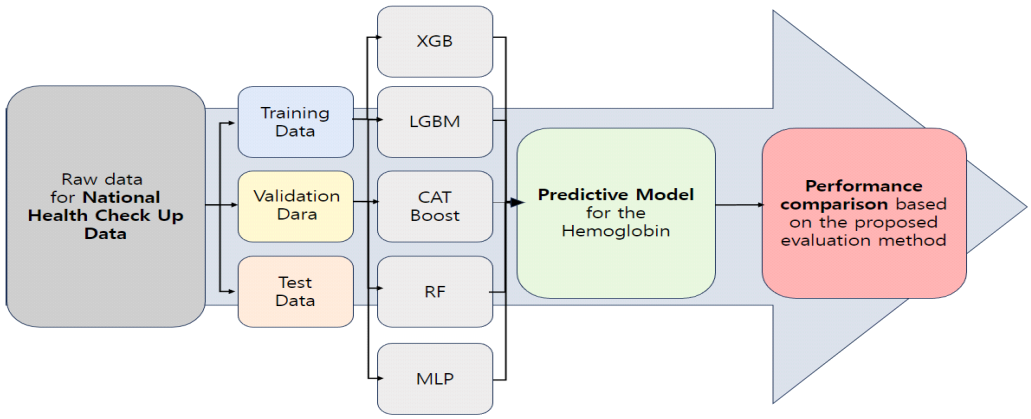

Figure 1. 연구 요약도

3. 이론적배경

3.1 CatBoost

CatBoost는 범주형 데이터를 처리하는 새로운 방법을 제시하고 있으며, 특히 CatBoost는 알고리즘에 내장된 기능을 통해 범주형 변수를 처리하는 데 있어 좋은 성능을 보인다. Prokhorenkova et al.(2017)에 따르면 CatBoost는 범주형 특성과 수치형 특성을 모두 처리할 수 있는 새로운 기능인 TS(Target Statistics)를 제공하고 범주형 자료의 중요도를 자동으로 추정할 수 있는 새로운 알고리즘을 이용하였다. 또한 기존의 Gradient boosting 방법들은 손실함수를 예측치에 대해 편미분한 기울기(Gradient) 값을 활용하므로, 발생하는 타겟누수(Target leakage)로 인한 과적합을 갖는다. 그래서 CatBoost가 트리를 학습하는 매 단계마다 다른 독립적인 데이터 세트를 이용하기 위해 임의의 순열에 따라 데이터 세트를 추출하여 독립적인 데이터 세트를 활용하는 Ordered boosting을 이용하고 있다. 또한 CatBoost는 트리를 구성할 때 의사결정나무(Decision tree)의 층마다 동일한 조건을 부여하여 좌우대칭의 형태로 만들어 과적합을 방지하는 ODT(Oblivious Decision Tree)를 구성한다.

3.2 XGBoost

XGBoost는 GBDT 알고리즘 중 빠른 속도를 보여주며 내부 알고리즘을 통해 높은 확장성을 보여준다. 이와 같은 이유로 데이터 분석 경진대회에서 좋은 성능을 보여 우승 알고리즘으로 자주 선정되었다. Chen and Guestrin(2016)

에 따르면 XGBoost는 기본적으로 약한 학습기인 의사결정나무를 이용하여 강력한 학습기로 예측을 진행하는 앙상블 기법의 알고리즘이다. XGBoost는 변수 선택, 변수 중요도 추정, 결측치 처리 등의 다양한 기능을 제공하고 규제(regularization)를 고려하고 과적합을 방지하여 모델의 성능을 향상시킬 수 있다. XGBoost는 대용량 데이터 세트에서도 뛰어난 성능을 보여주며 파이썬에서 병렬 처리 및 그래픽 처리 장치(Graphics Processing Unit, GPU) 가속화를 지원하여 빠른 학습과 예측이 가능하다.

3.3 LightGBM

LightGBM은 복잡하고 데이터 크기가 클 때 효율성과 확장성을 다른 기법에 비해 상당히 발전시킨 알고리즘이다. 다른 GBDT 알고리즘의 효율성과 확장성이 낮은 이유는 각 특성에 대해 모든 데이터를 읽어 가능한 모든 분할 지점의 정보를 추정해야 하기 때문이며, 이 문제를 해결하기 위해 LightGBM은 GOSS(Gradient-based One-Side Sampling)과 EFB(Exclusive Feature Bundling)라는 새로운 기술을 통해 기존 GBDT의 단점을 보완하였다.

LightGBM은 XGBoost와 유사한 알고리즘을 사용한다. 그러므로 LightGBM의 차별점인 GOSS와 EFB를 주로 설명하고자 한다. Ke et al.(2017)의 연구에서 LightGBM의 GOSS는 효율적인 데이터 샘플링 기법으로, Gradient 정보를 기반으로 한 샘플링을 수행한다. 큰 기울기를 가진 샘플을 유지하면서 작은 기울기를 가진 샘플에 대해 무작위 샘플링을 수행하는 것이 GOSS의 주요 목표이다. 또한 LightGBM의 EFB는 특정 조건에서 함께 분할되는 변수들을 묶어서 처리하여 학습 속도를 향상시키는 방법이며, 특정 변수가 다른 변수와 함께 분할될 때 이들을 하나로 묶어서 처리하고 업데이트한다. 이를 통해 효율적인 분할을 수행하고 모델의 학습 속도를 높인다.

3.4 Random Forest

Breiman(2001)에 따르면 RF는 배깅(Bagging)을 기반으로 하는 앙상블 트리 알고리즘이다. 분류 모델의 경우 Decision tree들이 전체 데이터에서 중복이 허락되고, 무작위로 선택된 데이터로 반복적으로 학습되어지는 과정을 배깅이라고 한다. 이 때 각각의 트리의 노드에서는 분할에 사용할 변수를 랜덤으로 선택한다. 마지막으로 최종 예측은 예측한 결과 중 가장 많이 선택된 클래스로 결정이 된다. 분류 모델과 유사하게 회귀 모델의 경우 의사결정나무들이 배깅에 의해 학습이 되고 모든 의사결정나무들의 예측값들의 평균을 최종 예측치로 고려한다.

3.5 Multi-layer Perceptron

MLP는 신경망의 한 유형으로, 입력층, 은닉층, 출력층으로 이루어진 구조를 가지고 있다. Lee(2021)의 연구에서 MLP는 현실에 존재하는 많은 문제들은 비선형 경계를 가지므로 선형 경계만으로 모든 관계를 표현하는 것은 불가능하다고 지적하였다. 이런 한계점을 극복하고자 MLP는 선형 결합을 활성화 함수를 이용해 비선형으로 변환하여 계산한다. 각 층의 노드는 데이터의 특성을 나타내며, 활성화 함수를 통해 비선형성을 도입하여 복잡한 패턴을 학습할 수 있게 하였다. 학습은 출력값에 대한 입력값의 기울기를 출력층에서부터 계산하여 거꾸로 전파시키는 역전파(Backpropagation) 알고리즘을 통해 이루어지며 최적화를 위해 경사 하강법과 같은 알고리즘이 사용된다. MLP는 이미지 분류, 음성 인식 등 다양한 분야에서 활용되며, 적절한 Hyper parameter tuning과 규제를 통해 모델의 성능을 향상시킬 수 있다. 하지만 데이터와 모델 특성에 따라 조절이 필요하며, 주의 깊은 모델 설계와 학습이 필요하다.

4. 실험 및 결과

4.1 데이터 수집 및 전처리

대한민국의 질병관리청에서 관리하고 배포하는 2021년 국민건강영양조사 데이터(KDCA, 2021)를 혈색소를 예측하는 모델을 구축하기 위해 채택하였다. 이는 한국 국민 전연령의 건강검진 데이터와 다양한 데이터를 포함하고 있으며, 기존 Hong and Hong(2021a) 연구에서 사용된 동일한 데이터이다. 데이터 총 3,000,000개의 데이터와 36개의 칼럼이 존재한다. 실험에 제안된 기법들을 사용하기 위해 Hong and Hong(2021a) 연구에서 고려된 전처리 과정을 동일하게 진행하였다. 이들 중 ID와 같이 관측치를 식별하기 위해 이용되는 변수들은 중요도가 낮고 데이터 노이즈가 발생할 수 있어 제외하였으며, '구강 검진 수검 여부', '치아 우식증 유무', '치석 유무' 등 구강검진 관련 변수와 데이터 수집 일자 같은 결측치가 66%가 넘는 변수들은 제외하였다. 또한 피어슨 상관계수 절대값 0.003보다 낮아 종속 변수에 큰 영향을 보이지 않는 변수들은 연구에서 제외하였다. 그 후 사분범위(Inter Quartile Range, IQR)를 기반하여 극단적 이상치를 제외하고 음주 여부, 흡연 상태와 같은 범주형 자료는 One-Hot encoding을 통해 숫자형 자료로 변환하였다. 그 결과 총 2,972,311개의 데이터를 추출하였고 데이터에 사용되는 변수들과 제외된 변수들은 Table 1과 같다.

Table 1. 사용된 변수들 및 제외된 변수들

변수	변수명
사용 변수	연령대 코드(5세 단위), 성별 코드, 신장(5㎝ 단위), 체중(5㎏ 단위), 허리둘레, 이완기 혈압, 트라이글리세라이드, 감마지티피, 시력(좌), 시력(우), 청력(좌), 청력(우), 수축기 혈압, 식전혈당(공복혈당), 콜레스테롤, HDL콜레스테롤, 요단백, 혈청, 크레아티닌, 흡연상태, LDL콜레스테롤, AST, ALT, 음주 여부
제외된 변수	기준 연도, 가입자 일련번호, 시도 코드, 데이터 공개 일자, 데이터 기준일자, 구강검진 수검 여부, 치아우식증 유무, 치석 유무, 구강검진 수검 여부, 결손치 유무, 치아마모증 유무, 제3대 구치(사랑니) 이상

전체 데이터를 학습, 검증, 테스트 데이터로 나누었으며 학습 데이터를 전체 데이터의 60%, 검증 데이터를 20%, 테스트 데이터를 20%로 분리하여 학습을 진행하였다. 이러한 데이터 전처리를 거치고 XGBoost, LightGBM, CatBoost, RF, MLP를 사용하여 예측 모델링 작업을 수행하였으며, 다양한 알고리즘의 특성을 고려하여 예측 정확도를 극대화하고자 하였다.

4.2 모델 학습

전처리된 데이터를 GBDT(XGBoost, LightGBM, CatBoost), RF, MLP를 이용하여 학습을 시킨 후 검증 데이터를 통해 모델의 예측 정확도를 평가하였다. 또한 머신 러닝 및 딥러닝 모델을 학습할 때 필요한 Hyper parameter는 파이썬 기본 설정값 및 베이지안 최적화(Snoek et al.,2012)를 검증 데이터에 적용한 후 테스트 데이터를 통해 최종 예측을 진행하였으나 예측 정확도는 차이가 없었다. 그래서 우리는 머신 러닝 및 딥러닝 모델을 학습할 때 필요한 Hyper parameter에 대해서 파이썬 기본 설정값을 모든 실험에 적용하였고, 최종적으로 Hong and Hong(2021a)이

제안한 MLR 모델과 본 연구에서 제안된 머신 러닝 및 딥러닝 모델들과 비교하였다. 모든 알고리즘과 실험은 파이썬으로 구현되었다. 우리가 제안하는 머신러닝 및 딥러닝 모델은 Hyper parameter에 대해서 파이썬 기본 설정값을 고려하므로 머신 러닝 및 딥러닝 모델에 해당되는 파이썬 함수에 데이터만 입력하여 구현된다.

4.3 예측 정확도 지표

Hong and Hong(2021a)은 평균 절대 비 오차(MAPE(오차율))만 예측 정확도 지표로 고려하였으나, 우리는 예측 모델을 최적화 후 예측 성능을 평가하기 위해 평균 제곱 오차(MSE), 평균 절대 비 오차(MAE), 평균 제곱근 오차(RMSE), 평균 절대 비 오차(MAPE(오차율))를 고려하였으며 이러한 예측 성능 지표들은 아래의 과정을 통해 산출된다.본 연구에서 제안된 알고리즘들로 학습된 모델을 활용하여 테스트 데이터에 대한 예측치를 산출하였고, 그 결과를 실제 데이터와 비교하여 정확도를 평가한다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \cdots\cdots (1))$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \cdots\cdots (2)$$

$$RMSE = \sqrt{MSE} \quad \cdots\cdots (3)$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \cdots\cdots (4)$$

y_i : 예측변수의 실제값

$\hat{y}_i$: 예측변수의 예측값

n : 총 데이터의 수

4.4 실험 결과

Table 2. 테스트 데이터에 대한모델별 예측 성능 평가 결과

Model / Accuracy	MLR	MLP	RF	LightGBM	XGBoost	Cat Boost
MSE	1.406	1.000	0.994	0.986	0.988	0.982
RMSE	1.186	1.000	0.997	0.993	0.994	0.991
MAE	0.929	0.783	0.781	0.774	0.775	0.773
MAPE	6.767	5.692	5.686	5.661	5.665	5.649

Table 3. 검증 데이터에 대한모델별 예측 성능 평가 결과

Model / Accuracy	MLR	MLP	RF	LightGBM	XGBoost	Cat Boost
MSE	1.026	0.986	0.991	0.984	0.984	0.982
RMSE	1.013	0.993	0.996	0,992	0.992	0,991
MAE	0.788	0.774	0.776	0.774	0.774	0.773
MAPE	5.773	5.664	5.679	5.661	5.660	5.654

Table 2에 따르면 CatBoost, LightGBM, XGBoost, RF, MLP 등 다양한 모델을 이용한 예측 성능이 Hong and Hong(2021a)의 MLR 보다 향상된 예측 성능을 보여준다. 이는 동일한 데이터를 사용하여 예측 정확도를 비교한 결과이며, 이전 연구와 비교했을 때, 본 연구의 MAPE(오차율) 성능이 약 1.1% 향상되었음을 알 수 있다. 뿐만 아니라, MAE에서도 이전 모델이 0.929인 반면, 본 연구에서 개발된 모델은 약 0.15 더 적은 MAE 값을 보여주며 더 정확한 예측성능을 보여준다. 이는 피어슨 선형 상관관계를 활용하여 변수를 선택한 MLR을 적용하는 것보다, 본 연구에서 제안하는 방법을 통해 모델을 개발하는 것이 예측 성능을 개선하는 데 도움이 될 수 있음을 입증한다.

또한 딥러닝 기법인 MLP와 GBDT(CatBoost, LightGBM, XGBoost)를 비교하였을 때 예측 정확도 부분에서 GBDT가 조금 더 좋은 성능을 보여주었다. 이러한 결과는 정형 데이터에서 여러 딥러닝 모델보다 GBDT의 성능이 우수한 사례들을 보여주는 연구결과와 일치한다(Shwartz-Ziv and Armon, 2022).

Table 3에 따르면 검증 데이터에 대한 예측 성능 분석 결과, 본 연구에서 사용된 다양한 모델들이 안정적인 성능을 보여주었다. 특히, MAPE 값이 5.65%에서 5.77% 사이로 모든 모델이 일관된 예측 성능을 기록하였으며, 이는 데이터를 효과적으로 학습하고 있음을 시사한다. 모형의 안정성은 테스트 데이터와 검증 데이터 간의 성능 차이가 거의 없다는 점에서 입증되었다. 특히 GBDT 계열 모델들은 검증 데이터에서도 높은 성능을 유지하며, 데이터의 복잡한 구조를 잘 처리하는 것으로 나타났다.

또한 GBDT에 내장된 변수 중요도를 통해 주요 변수들을 확인할 수 있다. Figure 2에서 예측 성능이 가장 높게 나온 CatBoost의 경우 성별, 연령대, ALT, 감마지티피, 흡연상태가 주요 변수로 나타났으며 이는 Hong and Hong(2021b)에서 연령대, 성별, 흡연유무가 전체 집단을 구분하는 인자로 선택되었는지 설명할 수 있다.

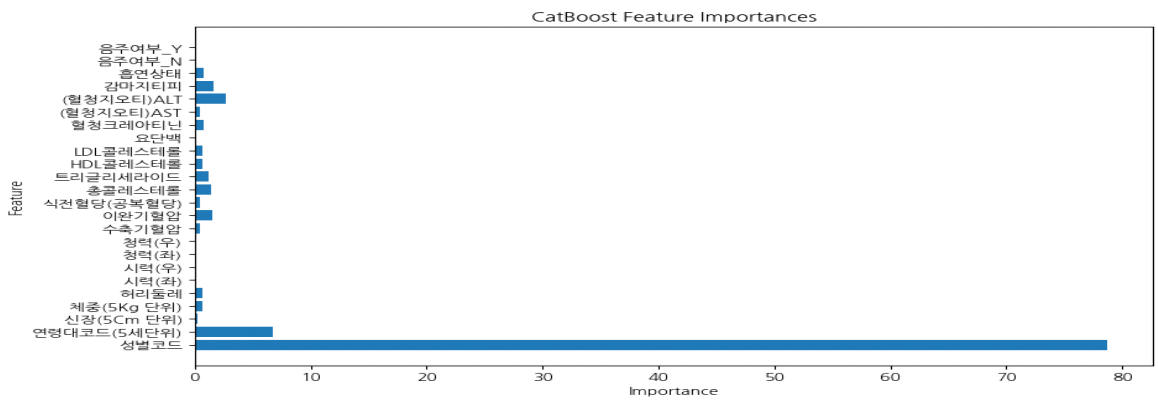

Figure 2. CatBoost의 변수 중요도

5. 결론

우리는 Hong and Hong(2021a)이 제안한 기존의 MLR보다 향상된 예측 정확도를 갖는 모델을 구축하고자 Hyper parameter에 대해서 파이썬 기본 설정값을 사용하는 머신 러닝 및 딥러닝 알고리즘들을 제안하였다. 우리는 실험 결과를 통해 기존 MLR보다 제안된 Hyper parameter에 대해서 파이썬 기본 설정값을 사용하는 머신 러닝 및 딥러닝 알고리즘들이 더 좋은 예측 성능을 갖는 것을 확인하였다. 헤모글로빈 측정기에 대한 미국 식품의약국(U.S. Food and Drug Administration, FDA) 허가와 미국 실험실 표준인증(Clinical Laboratory Improvement Amendments, CLIA) 기준은 MAPE(오차율) 7% 이내이며(Hong and Hong, 2021a) 본 연구에서 제안된 모델들은 모두 기준치보다 낮은 오차율을 보여주고 있다.

본 연구 결과를 활용하여 거동이 힘들고 빈혈에 대하여 위험성이 있는 노년층 환자와 간단히 자신의 혈색소 수치에 대하여 알고 싶은 환자들에게 기존의 데이터를 활용하여 간단히 혈색소의 예측이 가능할 것으로 사료된다. 이러한 연구는 헬스케어 분야에서의 실용적 가치와 함께, 다양한 집단을 위한 맞춤형 건강 관리에 기여할 것으로 기대된다. 결론적으로, 본 연구는 기존 연구의 성능보다 더 나은 예측 성능과 실용적 가치를 지닌 예측 모델을 제안하였다는 점에서 의의를 가진다. 우리의 연구 결과는 향후 다양한 헬스케어 분야에서 널리 활용될 수 있을 것으로 사료된다(Choi et al., 2023, Jung et al., 2018, Shang et al., 2016).

REFERENCES

Aghajanian, S., Jafarabady, K., Abbasi, M., Mohammadifard, F., Bakhtiari, M. B., Shokouhi, N., Gargari, S. S., and Bakhtiyari. M. 2024. Prediction of Post-Delivery Hemoglobin Levels with Machine Learning Algorithms. Scientific Reports 14(13953). https://doi.org/10.1038/s41598-024-64278-z.

Auvinen, J., Tapio, J., Karhunen, V., Kettunen, J., Serpi, R., Dimova E. Y., Gill, D., Soininen, P., Tammelin, T., Mykkänen, J., Puukka, K., Kähönen, M., Raitoharju, E., Lehtimäki, T., Korpela, M. A., Raitakari, O. T., Kiukaanniemi, S. K., Järvelin, M. R., and Koivunen, P. 2020. Systematic Evaluation of the Association between Hemoglobin Levels and Metabolic Profile Implicates Beneficial Effects of Hypoxia. American Association for the Advancement of Science 7(29):1-12.

Breiman, L. 2001. Random Forests. Machine Learning 45(1):5-32.

Chen, Y. and Guestrin, C. 2016. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data mining, Association for Computing Machinery, San Francisco, USA. p 785-794.

Choi, H. C., Seo, S. K., Kwon, J. Y., Park, S. C., Chang, H. J. 2023. Medical Staff's Awareness of Infected Patient Transfer Robots: Using SERVQUAL and AHP. Journal of Korean Society for Quality Management 51(3):381-401.

Dhakal, P., Khanal, S., and Bista, R. 2023. Prediction of Anemia Using Machine Learning Algorithms. International Journal of Computer Science & Information Technology 15(1):15-30

Ezekowitz, J. A., McAlister, F. A., and Armstrong, P. W. 2003. Anemia is Common in Heart Failure and is Associated with Poor Outcomes: Insights from a Cohort of 12,065 Patients with New-Onset Heart Failure. Circulation 107(2):223-225.

Hong, S. H. and Hong, K. S. 2021a. A Study on the Estimation of Hemoglobin based on Regression Using Physical and Health Information. Proceedings of the Korea Society for Industrial Systems Conference, Korea Information Processing Society, Seoul, Korea. p 553-555.

Hong, S. H. and Hong, K. S. 2021b. A Study on the Estimation Method of Hemoglobin Based on Linear and Multiple Regression Analysis Using Health Examination Big Data. Journal of the Institute of Electronics and Information Engineers 58(9):42-50.

Jeon, G. R., Lee, C. H., Jung, S. M., & Choi, J. G. 2024. The Effect of Characteristics of Social Intelligence Robots on Satisfaction and Intention to Use: Focused on User of Single Person Households. Journal of Korean Society for Quality Management 52(1):95-113.

Jeong, C. W. 2011. Anemia in the Elderly. The 62nd Fall Conference of the Korean Academy of Internal Medicine in 2011. The Korean Journal of Medicine, Jeonju, Korea. p 153-156.

Kavsaoğlu, A. R., Polat, K., and Hariharan, M. 2015. Non-Invasive Prediction of Hemoglobin Level Using Machine Learning Techniques with the PPG Signals Characteristics Features. Applied Soft Computing 37:983- 991.

KDCA. 2021. Korea National Health and Nutrition Examination Survey. https://knhanes.kdca.go.kr/knhanes/main.do.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu T. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Proceedings of the 31st International Conference on Neural Information Processing Systems,Curran Associates Inc., Long Beach, USA, p 3149-3157.

Lee, J. H. 2019. The Review of Deep Learning. Master's thesis, Ewha Womans University.

Llanos, M. J. 2016. Significance of Anemia in the Different Stages of Life. Enfermería Global 15(3):407- 418.

Mugdha, A. G., Pinki, F. T., and Talukdhar, S. K. 2023. Hemoglobin Estimation and Anemia Severity Prediction Using Machine Learning Algorithms. 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), IEEE, Dhaka, Bangladesh, p 1-6.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. 2018. CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems,Curran Associates Inc., Montréal, Canada, p 6639-6649.

Shang, M., Shin, Y. H., Lee, C., W. 2016. The Influence of the IoT Based Healthcare User's Experience Value on the Usage and Continuous Use Intention - Focused on Xiaomi Mi Band User in China -. Journal of Korean Society for Quality Management 44(3):689-706.

Shwartz-Ziv, R. and Armon, A. 2022. Tabular Data: Deep Learning Is Not All You Need. Information Fusion 81:84-90.

Snoek, J., Larochelle, H., and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Curran Associates Inc., Lake Tahoe, USA, p 2951-2959.

Yun, H. R. 2020. Development of Hemoglobin Prediction and Erythrocyte Stimulating Agent Recommendation Algorithm (HPERA) Using Recurrent Neural Network in End - Stage Kidney Disease Patients. PhD's thesis, Yonsei University.

저자소개

정대원 학부에서 글로벌 비즈니스를 전공했으며 주요관심분야는 데이터 분석, 마케팅 조사, 머신 러닝 등이다.

황욱연 North Carolina State University에서 통계학 석사 및 박사 학위를 취득하였으며, 현재 동아대학교 국제대학 글로벌비즈니스학과에서 부교수로 근무하고 있다.