

연속성 데이터의 범주형 분류 최적화를 위한
머신러닝 모델 비교연구*†
-디지털 기반 보행 변인을 활용한
학생건강체력(PAPS) 중심으로-

이 성 문(제1저자)

대진대학교 일반대학원 기계설계·산업시스템학과(박사)

임 성 옥(교신저자)

대진대학교 산업경영공학과 (교수)

A Comparison study of machine learning models for
optimizing categorical classification of continuity data
-Focused on Physical Activity Promotion System (PAPS)
using Digital Gait Variables-

Lee, Sung Moon(First Author)

Department of Mechanical Design Industrial Systems Daejin University

Lim, Sung Uk(Corresponding Author)

Department of Industrial and Management Engineering Daejin University

Abstract

This study aims to optimize categorical classification of continuous data by

* 본고는 한국서비스경영학회 2024년 추계학술발표대회에서 발표한 내용을 기반으로 제작성한 것임

† 본고는 이성문의 2025년도 박사학위 논문 「논문제목」의 일부를 발췌하여 제작성한 것임

comparing various machine learning models, focusing on digital gait variables for assessing student health-related physical fitness (PAPS). The models evaluated include CatBoost, LightGBM, Random Forest, XGBoost, Gradient Boosting, and Support Vector Regressor. CatBoost consistently achieved the lowest Mean Squared Error (MSE) and highest R-squared (R²) scores across most fitness factors, indicating its suitability for categorical classification tasks. The findings suggest that leveraging digital gait variables can contribute to personalized health management for students.

Keywords : Continuity data, categorical classification, Digital Gait Variables, Machine Learning, Physical Activity Promotion System (PAPS)

접수일(2024년 10월 26일), 수정일(2024년 11월 13일), 게재확정일(2024년 11월 14일)

I. 서론

현대 사회에서 신체활동의 부족은 건강에 중대한 영향을 미치는 주요 문제로 대두되고 있다. 특히 청소년기는 신체적, 정신적 발달이 활발히 이루어지는 시기로, 이 시기에 신체활동이 부족할 경우 비만, 심혈관 질환, 정신 건강 문제 등 다양한 건강 문제가 발생할 수 있다. 연구에 따르면, 신체활동은 청소년의 건강과 발달에 필수적인 요소로서 체력 증진을 통해 이러한 건강 문제를 예방할 수 있는 중요한 수단이다 (Strong et al., 2005; Janssen & Leblanc, 2010).

학생들의 체력 평가는 주로 심폐지구력, 근력, 유연성, 순발력 등의 체력 요소를 측정하여 청소년의 전반적인 건강 상태를 평가하는 데 사용되고, 국내에서는 학생건강 체력평가(Physical Activity Promotion System; PAPS)를 수행하고 있는 실정이다(교육부, 2009). 그러나 이러한 평가 방법들은 일부 한계점을 가지고 있으며, 특히 실제 일상생활에서의 신체활동을 반영하는 데에는 제한적일 수 있다.

보행은 인간의 기본적인 움직임으로, 개인의 신체적 건강 상태를 반영하는 중요한 지표이다. 최근의 연구들은 디지털 기술을 활용한 보행 분석이 체력 평가 및 건강 관리에 효과적으로 사용될 수 있음을 강조하고 있다(Burnfield, 2010; Muro-de-la-Herran, Garcia-Zapirain, & Mendez-Zorrilla, 2014). 보행 분석은 특히 신체활동의 강도, 균형, 협응 등을 반영할 수 있어, 개인의 전반적인 건강 상태를 평가하는 데 매우 유용한 도구로 인식되고 있다(Perry, 2006). 이러한 디지털 보행 변수는 머신러닝 기술과 결합할 때 더욱 정교한 체력 평가와 예측을 가능하게 한다.

머신러닝 기술은 최근 의료 및 건강 평가 분야에서 크게 주목받고 있으며, 데이터 분석 및 예측의 정확성을 높이는 데 중요한 역할을 하고 있다(Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015). 특히, 머신러닝 알고리즘을 통해 대규모 데이터를 신속하고 정확하게 처리할 수 있으며, 이를 바탕으로 개인 맞춤형 건강 관리 전략을 제시할 수 있다. 이러한 기술적 발전은 PAPS와 같은 체력 평가 시스템에 디지털 보행 데이터를 도입하여, 청소년의 체력 상태를 보다 정확하게 평가할 수 있는 가능성을 열어주고 있다. 그러나 기존 연구에서는 PAPS와 같은 체력 평가 시스템에서 머신러닝을 활용한 연구가 제한적이었으며, 특히 디지털 보행 변수를 활용한 연구는 초기 단계에 머물러 있다.

본 연구의 목적은 디지털 보행 변수를 활용한 PAPS 중심으로한 연속성 데이터의 범주형 분류 최적화를 위한 머신러닝 모델 비교연구를 통해 학생 건강체력평가 시스템(PAPS)의 예측 모델을 개발하는 것이다. 이를 위해 다양한 머신러닝 알고리즘을 비교하고, 각 알고리즘이 PAPS의 예측에 얼마나 효과적인지를 평가하고자 한다. 특히, 본 연구는 기존 연구에서 나타난 한계를 극복하고, 보다 대규모의 데이터를 활용하여 모델의 일반화 가능성을 높이는 데 중점을 둔다. 이를 통해, 학생들의 체력 상태를 보다 정밀하게 평가하고, 맞춤형 운동 처방을 제공할 수 있는 체력 평가 시스템을 구축하는 데 기여하고자 한다.

이 연구는 디지털 보행 변수를 활용하여 학생건강체력평가(PAPS)를 예측하기 위한 다양한 머신러닝 모델의 성능을 비교 분석하는 데 목적을 두고 있다. 연구는 크게 데이터 수집, 전처리, 모델링, 모델 평가 및 분석의 단계로 구성된다.

II. 이론적 배경

2.1 머신러닝 모델

본 연구에서는 CatBoost Regressor, LightGBM Regressor, Random Forest Regressor, XGBoost Regressor, Gradient Boosting Regressor, Support Vector Regressor 등의 모델이 사용되었다.

CatBoost는 범주형 데이터를 효과적으로 처리하기 위해 설계된 고급 부스팅 알고리즘이다. CatBoost는 순차적 부스팅 방식을 통해 과적합을 방지하고, 비선형적이고 복잡한 데이터 구조를 처리하는 데 탁월한 성능을 발휘한다. 특히, 이 알고리즘은 범주형 변수 인코딩에 강점이 있어 범주형 데이터가 많은 실제 환경에서도 안정적인 예측

을 제공한다. CatBoost는 데이터의 불균형 문제를 해결하고, 학습 데이터와 검증 데이터 간의 간섭을 최소화하는 방식으로 설계되어 높은 예측 정확도를 자랑한다.

LightGBM은 Microsoft에서 개발한 경량화된 부스팅 알고리즘으로, 대규모 데이터셋에서 매우 빠른 학습 속도를 제공하는 것이 특징이다. LightGBM은 리프 중심 트리 성장 방식을 사용하여 각 리프 노드에서 최대 손실을 줄이는 방향으로 트리가 성장하도록 한다. 이를 통해 트리의 깊이가 깊어지더라도 학습 효율성을 높이며, 특히 대규모 데이터와 고차원의 문제를 다루는 데 강점을 보인다. 메모리 사용량이 적고 계산 속도가 빠르기 때문에 매우 복잡한 데이터셋에서도 효율적으로 모델을 훈련할 수 있다.

Random Forest는 여러 개의 의사결정 나무(Decision Tree)를 결합하여 예측을 수행하는 대표적인 앙상블 기법이다. 랜덤 포레스트는 각 결정 트리의 예측 결과를 평균하거나 다수결 투표 방식으로 결합하여 최종 예측을 도출한다. 이 알고리즘의 강점은 트리의 개별적 약점을 보완할 수 있어, 단일 결정 트리에서 발생할 수 있는 과적합 문제를 효과적으로 방지한다는 점이다. 또한, 다양한 변수 간의 상호작용을 처리할 수 있어 데이터의 복잡한 패턴을 잘 포착할 수 있으며, 노이즈가 많은 데이터셋에서도 안정적인 성능을 유지한다.

XGBoost는 부스팅 알고리즘의 일종으로, 경사 하강법을 사용하여 반복적으로 모델의 성능을 개선하는 방법을 채택하고 있다. XGBoost는 일반적인 부스팅 기법보다 성능 최적화가 잘 되어 있으며, 과적합을 방지하기 위한 정규화(term)와 빠른 계산을 가능하게 하는 다양한 기법이 포함되어 있다. 예를 들어, XGBoost는 잔차에 기반한 학습 방식을 사용하여 오류를 줄여가며 모델을 점진적으로 개선하고, 이를 통해 데이터셋의 다양한 패턴을 학습할 수 있다. 또한, 빠른 계산 속도와 메모리 효율성을 동시에 달성하여 대규모 데이터셋에서 높은 성능을 보장한다.

Gradient Boosting은 부스팅 기법의 기본 형태로, 각 단계에서 예측 오차(residual)를 줄이기 위해 새로운 모델을 추가하는 방식으로 동작한다. 이전 모델이 만든 오차를 보완하는 방식으로 새로운 모델을 학습시키며, 이 과정이 반복되면서 성능이 점진적으로 향상된다. Gradient Boosting은 고차원의 복잡한 데이터에 대해 효과적인 예측을 수행할 수 있으며, 과적합을 방지하기 위해 학습 속도와 모델의 복잡도를 적절하게 제어할 수 있는 다양한 정규화 기법을 포함하고 있다. 특히, 순차적으로 예측 오차를 줄여 나가는 방식으로 데이터의 복잡한 패턴을 포착하고, 높은 예측 성능을 도출할 수 있는 능력이 뛰어나다.

Support Vector Regressor(SVR)는 서포트 벡터 머신(SVM)을 회귀 문제에 적용한 모델로, 데이터 포인트 사이의 마진을 최대화하면서 회귀를 수행하는 기법이다. SVR은 데이터가 선형 분리가 어려운 경우에도 커널 함수(kernel function)를 사용하여 고

차원 공간에서의 패턴을 탐색하며, 비선형 데이터에 대한 예측 성능이 뛰어나다. 특히, SVR은 소수의 서포트 벡터에 의존하여 학습을 진행하므로, 전체 데이터셋 중 중요한 데이터 포인트에 더 집중할 수 있다. 다만, 계산 비용이 크고 대규모 데이터셋에서는 학습 시간이 길어질 수 있으며, 매우 복잡한 데이터의 경우 효율성이 떨어질 수 있다는 단점이 있다.

2.2 학생건강체력평가시스템(PAPS)

학생건강체력평가제는 학생들의 체력과 건강 상태를 종합적으로 평가하고 이를 통해 건강 증진을 도모하는 체계적인 시스템이다(교육부, 2009). PAPS는 기존의 체력장 제도를 대체하며, 심폐지구력, 근력·근지구력, 유연성, 순발력, 비만 등의 다양한 요소를 종합적으로 평가한다. 기존의 체력장은 주로 운동 기능을 평가하는 데 중점을 두었던 반면, PAPS는 학생 개개인의 건강 상태를 종합적으로 평가하고, 이를 바탕으로 맞춤형 운동 처방을 제공하는 시스템으로 발전했다. 이로 인해 PAPS는 학생들의 체력뿐만 아니라 전반적인 건강 상태를 향상시키는 데 중요한 역할을 하게 되었다. 학생건강체력평가시스템 개요는 <표 1>와 같다.

<표 1> 학생건강체력평가시스템(PAPS) 개요

학생건강체력평가	
학년	초5 ~ 고3(초 4학년 선택)
특징	건강관련체력 위주: 심폐지구력, 근력·근지구력, 유연성, 체지방조절력 등 ※ 학생개인의 건강체력 측정에 초점
종목	5개 체력요인 12개 종목 선택형 심폐지구력(왕복오래달리기, 오래달리기-걸기, 스텝검사) 유연성(앉아윗몸앞으로굽히기, 종합유연성검사) 근력·근지구력[(무릎대고)팔굽혀펴기, 윗몸말아올리기, 약력검사] 순발력(50m 달리기, 제자리멀리뛰기) 체지방[체질량지수(BMI), 체지방률(%Fat)]
등급(점수)	1~5등급 - (종목별 20점, 전체 100점 만점) ※ 신체활동 성취동기 유발

측정	연 1회(학년 초) 다양한 IT 기술을 적용한 측정도구 사용
활용	자신의 건강체력 정보의 능가적 관리 및 운동처방을 받고 언제 어디서나 자신의 건강체력을 관리할 수 있는 웹 기반으로 구축 비만 학생의 선별 및 관리 선택평가가 있음

Ⅲ. 연구방법

3.1 머신러닝 비교 모형

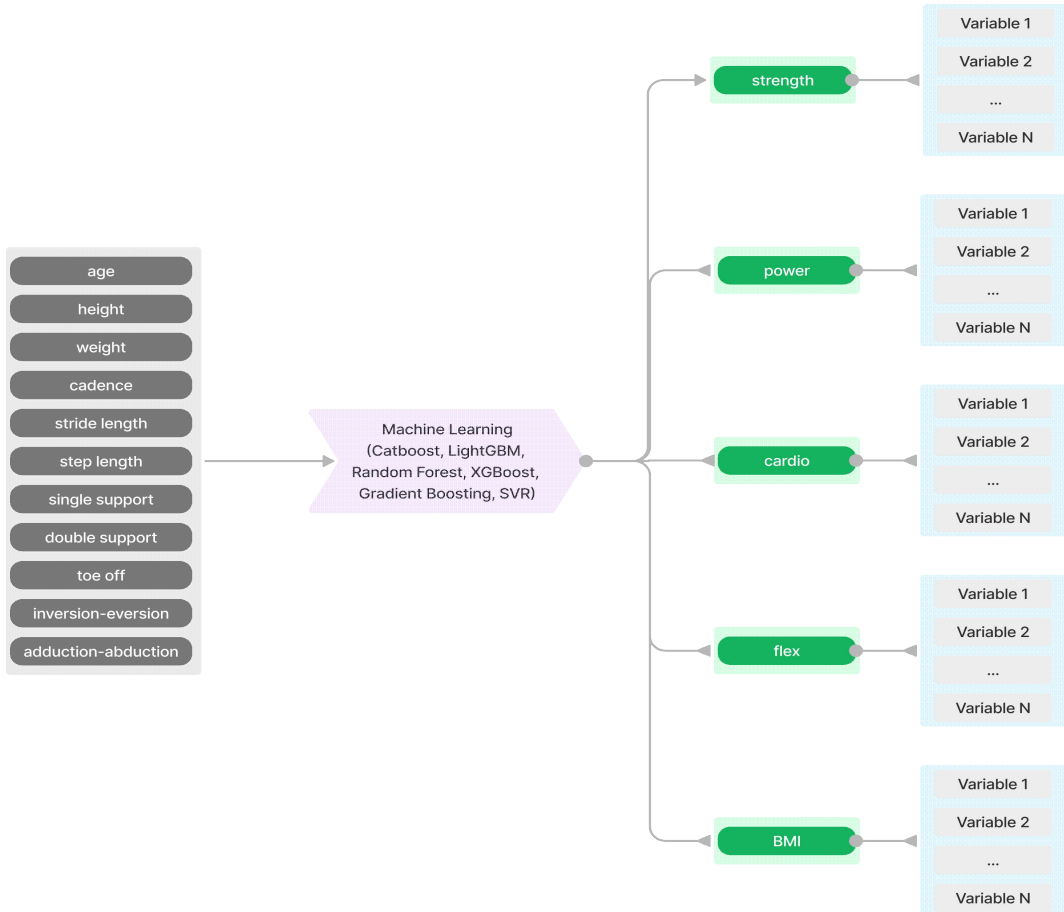
본 연구에서는 연속형 변수인 신체 정보(age, height, weight)와 보행 변수(cadence, Left Stride Length, Left Stride Length CV, Right Stride Length, Right Stride Length CV, Left Step Length, Left Step Length CV, Right Step Length, Right Step Length CV, Left Single Support, Left Single Support CV, Right Single Support, Right Single Support CV, Left Double Support, Left Double Support CV, Right Double Support, Right Double Support CV, Left Time of Toe Off, Left Time of Toe Off CV, Right Time of Toe Off, Right Time of Toe Off CV, Left ROM Inver-Ever, Left ROM Inver-Ever CV, Right ROM Inver-Ever, Right ROM Inver-Ever CV, Left ROM Dorsi-Plantar, Left ROM Dorsi-Plantar CV, Right ROM Dorsi-Plantar, Right ROM Dorsi-Plantar CV, Left ROM Adduct-Abduct, Left ROM Adduct-Abduct CV, Right ROM Adduct-Abduct, Right ROM Adduct-Abduct CV)를 사용하여 strength(근력), power(순발력), cardio(심폐지구력), flex(유연성), BMI 범주형 요인을 예측하는 데 목적을 둔다<그림 1>.

연구의 핵심은 주어진 신체 정보와 34개의 보행 변수를 활용하여 각 체력 요인에 대한 예측을 수행하고, 예측 성능이 가장 뛰어난 머신러닝 모델을 확인할 것이다. 이를 위해 CatBoost, LightGBM, Random Forest, XGBoost, Gradient Boosting, Support Vector Regressor와 같은 다양한 머신러닝 모델을 적용하고, 각 모델의 성능을 평가하였다.

3.2 평가모형

본 연구에서는 머신러닝 모델의 예측 성능을 평가하기 위해 세 가지 주요 평가 지표인 평균 제곱 오차(Mean Squared Error, MSE), 평균 절대 오차(Mean Absolute Error, MAE), 결정 계수(R-squared, R^2)를 사용하였다(최인묵, 2017) <표 2>. MSE(Mean Squared Error)는 모델이 예측한 값과 실제 값 사이의 오차를 제곱하여 그 평균을 구하는 방식으로 계산된다. MSE는 오차를 제곱하기 때문에 큰 오차에 더욱 민감하게 반응하며, 값이 작을수록 모델의 예측이 실제 값에 가까움을 의미한다. MSE는 모델이 큰 오차를 내는 경우 성능이 크게 떨어질 수 있다는 특징을 가지며, 예측 성능을 정밀하게 평가하는 데 적합하다. MAE(Mean Absolute Error)는 실제 값과 예측 값 사이의 차이를 절대값으로 변환하여 그 평균을 구하는 방식이다. MAE는 오차를 절대값으로 계산하기 때문에 모든 오차를 동일하게 취급하며, 값이 작을수록 모델이 예측을 더 정확하게 수행했음을 의미한다. MAE는 큰 오차에 대한 민감도가 MSE보다 낮아, 실제 값과 예측 값 간의 평균적 차이를 직관적으로 이해할 수 있다. R^2 (R-squared, 결정 계수)는 모델이 실제 값의 분산을 얼마나 잘 설명하는지를 나타내는 지표로, 1에 가까울수록 모델이 데이터를 잘 설명하고 있다는 것을 의미한다. R^2 는 0에서 1 사이의 값을 가지며, 1은 모델이 실제 데이터를 완벽히 설명하는 경우, 0은 예측 모델이 실제 데이터와 무관한 경우를 나타낸다. R^2 값이 음수일 경우, 모델의 예측 성능이 실제 평균 값보다도 낮다는 의미이며, 이는 모델의 예측력이 현저히 부족함을 나타낸다. MSE와 MAE는 모델의 예측값과 실제값 사이의 오차를 수치적으로 평가하는 지표이며, R^2 는 모델이 전체 데이터의 변동성을 얼마나 설명할 수 있는지를 비율로 나타낸다. 본 연구에서는 이러한 지표들을 사용하여 각 머신러닝 모델의 성능을 종합적으로 분석하고 비교하였다. 특히, MSE와 MAE는 예측값과 실제값 간의 차이를 수치적으로 평가하고, R^2 는 모델이 얼마나 데이터를 잘 설명하는지를 파악하는 데 중점을 두었다.

<그림 1> 머신러닝 모델 수행 흐름도



<표 2> 평가 모형 종류 및 설명

평가 지표	설명	해석
MSE (Mean Squared Error, 평균 제곱 오차)	MSE는 실제값과 예측값 간의 차이(잔차)를 제곱한 후 그 평균을 구하는 방식으로 계산	값이 작을수록 모델의 예측이 실제값에 가까움을 나타냄 MSE는 오류를 제공하기 때문에 큰 오류에 더 민감하게 반응하며, 이로 인해 매우 큰 오차가 있을 경우 모델의 성능이 크게 저하될 수 있음
MAE	MAE는 실제값과 예측값 간의 차이를	값이 작을수록 모델의 예측이

(Mean Absolute Error, 평균 절대 오차)	절대값으로 변환한 후 그 평균을 계산 예측값과 실제값 사이의 평균적 차이를 나타냄	실제값에 가까움을 의미 MAE는 모든 오류를 동일하게 취급하므로, 큰 오차에 대한 민감도가 MSE보다 낮음 실제값과 예측값 간의 절대적인 차이를 이해하는 데 유용함
R^2 (R-squared, 결정 계수)	R^2 는 예측 모델이 실제값의 변동을 얼마나 잘 설명하는지를 나타내는 지표 1에 가까울수록 모델이 데이터의 분산을 잘 설명하는 것을 의미 R^2 는 모델의 성능을 비율로 나타내며, 0은 모델이 예측에 실패한 경우, 1은 완벽한 예측을 한 경우임	값이 1에 가까울수록 모델이 데이터를 잘 설명하고 있음을 의미 0에 가까울수록 예측 능력이 떨어짐 음수 값이 나오는 경우 모델의 예측이 실제값보다도 안 좋은 성능을 보인다는 의미

IV. 실증분석

4.1 보행변인과 학생건강체력 데이터 수집 및 특성

본 연구에서는 디지털 기반 보행 변인 데이터를 수집하기 위해 발목형 관성 센서 (Inertial Measurement Unit) 기반의 시스템을 사용하여 분석하였다<그림 2>.

<그림 2> 발목형 센서 보행 측정 장비(Smart Balance)



이 시스템은 컴퓨터와 소프트웨어(MotionCore, JEIOS, Korea), 관성 센서가 장착된 발목 센서 Smart Balance® SB-1(JEIOS, Korea)로 구성(Kim, Joo, Jeong, Jeon, & Jung, 2016; Lee, Youm, Jeon, Cheon, & Park, 2018; Yoo, Baek, & Kim, 2019) 되어 있으며, 관성 센서의 샘플링 주파수는 100Hz이고, ±6g의 3축 방향 가속도 측정

과 $\pm 500^\circ/s$ 의 3축 방향 각속도 측정이 가능한 것으로 좌우 발목 센서 내부에 탑재되어 있다. 발목 센서에 탑재된 블루투스 장비가 측정값을 보내고 컴퓨터와 연결된 신호 수집 장치에서 측정값을 수집하여 컴퓨터로 보낸다. 관성 센서의 지역 좌표계는 전후 방향을 X축, 좌우 방향을 Y축, 수직 방향을 Z축으로 설정하였다. 분석 변인(보빈도, 보장, 보폭, 외발지지기, 양발지지기, 발 이지 시점, 내/외반, 배/저측, 내/외전 가동범위)은 관성 센서 장치에서 취득된 가속도와 각속도를 가공하여 산출하였다. 발목 센서를 좌우 발목에 착용한 후 트레드 밀에서 3분 보행을 수행하였다.

학생건강체력 측정은 학생 건강 체력 평가 시스템(Physical Activity Promotion System; PAPS)를 기반으로 하였다(교육부, 2009). 학생들의 신장, 체중, BMI 외에 체력측정항목은 근력, 근파워, 심폐지구력, 유연성을 측정하였다. 근력은 악력으로 측정하였고, 근파워는 50m 달리기, 제자리 멀리뛰기, 심폐지구력은 20m 왕복오래달리기를 통해 측정되었으며, 유연성은 앉아윗몸앞으로굽히기로 측정하였다. 체력 측정 항목의 다양성을 평준화 시키기 위해서 현재 PAPS에서 규정하고 있는 등급으로 수치를 변환한 후 머신러닝 모델 분석에 적용하였다. 체력 측정 도구는 아래 <표 3>와 같다.

<표 3> 체력 요인 측정 도구

Variable	Measurement Tool	Manufacturer	Country
신장, 체중	신장체중측정기 (DS-103M)	제닉스	한국
근력	디지털악력계 (KS-301)	라비센	한국
순발력	제자리멀리뛰기매트 (OSC-A426)	오레인	한국
심폐지구력	팝스 왕복 오래달리기 음원	팝스 음원	한국
유연성	유연성 측정기 (HBT-1111)	한별	대만

보행 변인과 학생건강체력 변인 분석 데이터는 2021년부터 2024년 상반기 까지 8개 학교에서 누적 숫자 4890명의 학생(초등학생, 중학생, 고등학생)을 누적 학생 수 데이터를 기반으로 하였다. 측정된 보행 변인과 학생건강체력 변인의 평균 및 표준편차는 아래 <표 4>와 같다.

<표 4> 연구대상자의 인구통계학적 특성, 데이터 변인 평균 및 표준편차

Data Variable	Mean		Standard Deviation	
	Left	Right	Left	Right
나이(세) age	12.8		2.3	
신장(cm) height	156.0		11.7	
체중(kg) weight	50.9		14.8	
보빈도(steps/min) cadence	102.1		12.0	
보장(m) stride length	0.9	0.9	0.2	0.2
보장 분산(m ²) stride length CV	8.9	4.3	4.8	2.3
보폭(m) step length	0.4	0.5	0.1	0.1
보폭 분산(m ²) step length CV	4.5	0.2	2.4	0.1
외발지지기(%cycle) single support time	38.0	38.1	1.1	1.1
외발지지기 분산(%cycle) ² single support time CV	0.3	0.4	0.1	0.2
양발지지기(%cycle) double support time	24.1	24.1	1.8	1.8
양발지지기 분산(%cycle) ² double support time CV	0.4	0.2	0.2	0.1
발 이지 시점(%cycle) time of toe off	62.1	62.1	1.5	1.5
발 이지 시점 분산(%cycle) ² time of toe off CV	0.2	2.1	0.1	2.0
내/외반 가동범위(degree) ROM Inver-Ever	2.1	0.8	2.1	2.1
내/외반 가동범위 분산(degree) ² ROM Inver-Ever CV	0.7	1.4	0.2	0.4
배/저측 굴곡 가동범위(degree) ROM Dorsi-Plantar	-8.3	-8.5	2.7	2.6
배/저측 굴곡 가동범위 분산(degree) ² ROM Dorsi-Plantar CV	1.4	1.6	0.4	0.6
내/외전 가동범위(degree) ROM Adduct-Abduct	1.1	1.0	1.6	1.6

내/외전 가동범위(degree) ² ROM Adduct-Abduct CV	1.6	1.6	0.6	0.6
근력(등급)	3.0		1.0	
순발력(등급)	2.7		1.3	
심폐지구력(등급)	2.7		0.9	
유연성(등급)	2.0		1.0	
BMI(cm/m ²)	20.8		4.2	

4.2 머신러닝 모델 수행 방법

본 연구에서는 머신러닝 모델을 활용하여 특정 데이터셋의 타겟 변수 예측 성능을 평가하고 최적의 회귀 모델을 도출하기 위한 일련의 절차를 수행하였다. 이를 위해 다양한 회귀 알고리즘(Support Vector Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, CatBoost)을 비교 분석하고, 각 모델의 하이퍼파라미터 최적화 과정을 통해 성능을 극대화하였다. 이 연구는 머신러닝 모델의 성능 평가를 위해 데이터 전처리, 모델 훈련, 하이퍼파라미터 튜닝, 성능 평가 및 시각화 등의 방법론을 채택하였다. 이 절차를 통해 데이터 특성과 모델의 적합성을 검토하고, 과학적이고 객관적인 기준에 따라 최적의 회귀 모델을 선정하였다. 연구에 사용된 데이터는 csv 파일 형식으로 저장되어 있고, 다양한 변수들로 구성된 실험 데이터를 활용하였다. 파일 경로에 따라 .csv 형식의 파일을 자동으로 읽어 들여 pandas 라이브러리를 사용해 DataFrame 형식으로 로드하였다. 수집된 데이터는 결측값을 포함하고 있을 가능성이 있으며, 범주형 변수를 포함하는 경우가 있다. 이러한 데이터를 분석에 적합하게 만들기 위해 결측값을 포함한 샘플은 제거하였다. 이는 결측값이 있는 샘플이 전체 분석에 미치는 영향을 제거하고 모델의 학습 성능을 보장하기 위함이었다. 또한, 범주형 변수는 원-핫 인코딩(one-hot encoding)을 통해 이진 변수로 변환하였다. 이 과정에서 범주형 변수에 대해 더미 변수를 생성하고, 첫 번째 범주를 제거(drop_first=True)하여 다중공선성을 방지하였다. 타겟 변수(target_column)와 분석에서 제외할 변수(exclude_columns)를 제외한 모든 변수를 독립 변수로 설정하고, 타겟 변수는 종속 변수로 설정하였다. 이러한 변수를 x와 y로 나누어 분석의 기반으로 삼았다.

본 연구에서는 다양한 머신러닝 회귀 모델의 성능을 평가하기 위해 Support Vector Regressor(SVR), Random Forest Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, CatBoost의 여섯 가지 모델을 사용하였다. SVR은 고차원 특성 공간에서 결정 경계를 설정하여 회귀 문제를 해결하는 모델로, 특정 수의 샘플만이

결정 경계에 영향을 미친다. 이는 고차원 데이터에 효과적이나, 하이퍼파라미터 선택에 민감하다(Vapnik, 1995). Random Forest Regressor는 여러 개의 결정 트리를 앙상블하여 예측 성능을 향상시키는 모델로, 개별 트리의 과적합을 방지하고 일반화 성능을 높인다(Breiman, 2001). Gradient Boosting Regressor는 순차적으로 트리를 생성하며, 이전 트리의 오차를 줄이는 방향으로 학습을 진행하는 모델이다(Friedman, 2001). XGBoost는 Gradient Boosting 알고리즘을 개선한 모델로, 계산 속도와 성능에서 우수한 결과를 보이며, 규제(regularization)를 통해 과적합을 방지한다(Chen & Guestrin, 2016). LightGBM은 XGBoost와 유사하지만, 대규모 데이터셋에서 더 빠르게 학습할 수 있도록 설계된 모델이다(Ke et al., 2017). CatBoost는 범주형 변수를 효율적으로 처리할 수 있는 Gradient Boosting 알고리즘으로, 자동으로 범주형 변수를 다루며, 매개변수 튜닝 없이도 우수한 성능을 발휘한다(Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018).

각 모델의 성능을 극대화하기 위해 하이퍼파라미터 튜닝을 수행하였다. 이를 위해 교차 검증(k-fold cross-validation)을 통한 그리드 탐색(Grid Search)을 사용하였다. GridSearchCV를 통해 각 모델의 다양한 하이퍼파라미터 조합을 탐색하였으며, 5-fold 교차 검증을 통해 각 모델의 성능을 평가하였다. 각 모델의 하이퍼파라미터를 최적화한 후, 최적 모델(best estimator)을 선택하고, 해당 모델의 성능을 검증 데이터셋에서 평가하였다. 모델의 성능은 평균 제곱 오차(Mean Squared Error, MSE), 평균 절대 오차(Mean Absolute Error, MAE), 결정 계수(R^2 Score)로 평가되었다. MSE는 실제 값과 예측 값의 차이를 제곱하여 평균낸 값으로, 오류의 크기를 제곱 단위로 측정한다. MAE는 예측 오류의 절대값을 평균낸 값으로, MSE와 달리 오류의 크기를 그대로 측정한다. R^2 는 모델이 데이터의 분산을 얼마나 설명하는지를 나타내는 지표로, 1에 가까울수록 모델이 데이터를 잘 설명한다.

각 모델의 성능 평가 결과는 시각적으로 표현하여 모델 간 비교를 용이하게 하였다. 성능 지표는 MSE, MAE, R^2 로 구분하여 플롯을 생성하였으며, 이를 통해 각 모델의 훈련 및 테스트 성능(20% 샘플)을 직관적으로 비교할 수 있도록 하였다. 결과적으로 가장 높은 R^2 값을 보인 모델을 최적 모델로 선정하였으며, 해당 모델의 최적 하이퍼파라미터 조합과 성능 지표를 제시하였다. 결과는 바(bar) 그래프로 시각화하였으며, 각 모델의 성능을 한눈에 비교할 수 있도록 구성하였다. 최종적으로 각 모델의 성능 평가 결과는 엑셀 파일로 저장하여 연구 자료로 활용하였다. 저장된 결과는 추후 연구에서 재현성(reproducibility)을 보장하며, 추가 분석을 위한 기초 데이터를 제공한다.

4.3 보행 변인 기반 체력요인 머신러닝 모델 분석 결과 및 해석

본 연구에서는 근력, 순발력, 심폐지구력, 유연성을 주요 타겟 변수로 설정하여 다양한 머신러닝 모델을 활용한 예측 성능을 평가하였다<표 5>. 사용된 모델은 CatBoost Regressor, LightGBM Regressor, Random Forest Regressor, XGBoost Regressor, Gradient Boosting Regressor, Support Vector Regressor(SVR), 총 6가지로, 각 모델의 하이퍼파라미터 최적화를 통해 최적의 예측 성능을 도출하였다.

<표 5> 체력 요인 머신러닝 모델별 성능 평가 결과

Model	체력 요인	Best Parameters	Trai	Tes	Trai	Tes	Trai	Tes	
			n	t	n	t	n	t	
			MS	MS	MA	MA	R2	R2	
			E	E	E	E			
CatBoost Regressor	근력		0.01	0.03	0.06	0.12	0.99	0.96	
	순발력	{ 'model__depth': 9, 'model__iterations': 200, 'model__learning_rate': 0.1 }	0.01	0.05	0.08	0.16	0.99	0.95	
	심폐지구력		0.01	0.05	0.06	0.14	0.99	0.95	
	유연성		0.01	0.07	0.08	0.17	0.99	0.96	
	BMI		0.18	0.63	0.33	0.58	0.99	0.97	
LightGBM Regressor	근력		{ 'model__learning_rate': 0.1, 'model__max_depth': -1, 'model__n_estimators': 200, 'model__num_leaves': 50 }	0.00	0.04	0.03	0.13	1.00	0.95
LightGBM Regressor	순발력	{ 'model__learning_rate': 0.1, 'model__max_depth': 20, 'model__n_estimators': 200, 'model__num_leaves': 50 }	0.00	0.06	0.04	0.17	1.00	0.94	
	심폐지구력		0.00	0.06	0.03	0.14	1.00	0.93	
	유연성		0.00	0.09	0.04	0.20	1.00	0.94	
	BMI		{ 'model__learning_rate': 0.1, 'model__max_depth': 20, 'model__n_estimators': 200, 'model__num_leaves': 100 }	0.02	0.67	0.07	0.58	1.00	0.97

Random Forest Regressor	근력	{'model__max_depth': None, 'model__max_features':	0.01	0.06	0.06	0.17	0.99	0.92
	순발력	'sqrt',	0.01	0.09	0.07	0.20	0.99	0.91
	심폐	{'model__n_estimators': 500} {'model__max_depth': 30, 'model__max_features':	0.01	0.08	0.06	0.18	0.99	0.92
	지구력	'sqrt',	0.02	0.15	0.09	0.26	0.99	0.91
	유연성	{'model__max_depth': None, 'model__max_features':	0.08	0.70	0.20	0.58	1.00	0.96
	BMI	'sqrt', 'model__n_estimators': 500}						
XGBoost Regressor	근력		0.00	0.04	0.02	0.13	1.00	0.95
	순발력	{'model__learning_rate': 0.1,	0.00	0.07	0.03	0.18	1.00	0.93
	심폐	'model__max_depth': 7,	0.00	0.06	0.03	0.14	1.00	0.93
	지구력	'model__n_estimators': 200}	0.00	0.10	0.03	0.20	1.00	0.94
	유연성		0.02	0.74	0.11	0.59	1.00	0.96
Gradient Boosting Regressor	근력		0.00	0.0	0.0	0.1	1.0	0.9
	순발력		0.00	4	2	2	0	5
	심폐	{'model__learning_rate': 0.1,	0.00	7	3	7	0	4
	지구력	'model__max_depth': 7,	0.00	0.0	0.0	0.1	1.0	0.9
	유연성	'model__n_estimators': 200}	0.00	6	3	4	0	4
	BMI		0.03	0.0	0.0	0.1	1.0	0.9
Support Vector Regressor	근력		0.04	0.0	0.1	0.1	0.9	0.9
	순발력	{'model__C': 10, 'model__epsilon': 0.1}	0.04	6	2	5	5	2
	심폐		0.06	0.1	0.1	0.1	0.9	0.8
	지구력		0.08	2	2	7	3	7
	유연성	{'model__C': 10, 'model__epsilon': 0.2}	1.77	0.1	0.1	0.2	0.9	0.9
	BMI	{'model__C': 10, 'model__epsilon': 0.5}		7	7	2	5	0
				2.0	0.9	0.9	0.9	0.9

4.4 머신러닝 분석 결과 해석

본 연구에서는 CatBoost, LightGBM, Gradient Boosting, XGBoost, Random Forest, Support Vector Regressor(SVR) 등 다양한 머신러닝 모델을 활용하여 학생 건강체력평가시스템(PAPS)에서 측정된 체력 요인들을 예측하였다. 각 체력 요인(근력, 순발력, 심폐지구력, 유연성, BMI)에 대해 모델의 성능을 평가한 결과, 전반적으로 부스팅 계열의 모델(CatBoost, LightGBM, XGBoost, Gradient Boosting)이 가장 우수한 성능을 보였으며, 특히 CatBoost 모델이 대부분의 체력 요인에서 최상의 성능을 기록하였다. CatBoost Regressor는 모든 체력 요인과 BMI 예측에서 높은 결정 계수(R^2)와 낮은 평균 제곱 오차(MSE), 평균 절대 오차(MAE)를 기록하여 다른 모델들에 비해 일관되게 뛰어난 성능을 보였다. 이러한 결과는 CatBoost 모델이 다양한 데이터 특성에 대해 잘 일반화할 수 있으며, 데이터의 복잡한 패턴을 효과적으로 학습할 수 있음을 시사한다. 특히, 심폐지구력과 유연성 예측에서 CatBoost의 우수한 성능은 이 모델이 고차원적인 데이터에 대해서도 강력한 예측력을 보유하고 있음을 보여준다. 이와 유사한 결과를 보인 선행 연구에서는 CatBoost가 범주형 변수 처리가 뛰어나고, 데이터 불균형 문제에 강하다는 점이 강조되었다(Prokhorenkova et al., 2018). LightGBM Regressor는 CatBoost에 비해 전반적으로 성능이 다소 떨어졌으나, 여전히 높은 예측력을 보였다. 특히 근력과 순발력 예측에서 상대적으로 낮은 MSE와 높은 R^2 값을 기록하였다. LightGBM은 학습 속도가 빠르고, 대규모 데이터셋에 적합하다는 장점을 가지고 있다(Ke et al., 2017). 그러나 본 연구에서는 CatBoost에 비해 다소 과적합(overfitting)의 영향을 받을 가능성이 있는 것으로 나타났다. XGBoost Regressor는 전반적으로 Gradient Boosting Regressor과 유사한 성능을 보였으며, 특히 BMI 예측에서 다소 낮은 성능을 기록하였다. XGBoost는 계산 효율성과 성능에서 강점을 보이지만, 본 연구에서는 데이터의 복잡성에 따른 일반화 성능이 CatBoost에 비해 낮은 것으로 분석되었다(Chen & Guestrin, 2016). Gradient Boosting Regressor는 CatBoost와 유사한 부스팅 알고리즘을 사용하나, 일반화 성능이 다소 낮은 것으로 나타났다. 이는 하이퍼파라미터 설정과 모델의 복잡도에 따라 성능이 크게 달라질 수 있음을 시사한다(Friedman, 2001). Support Vector Regressor(SVR)는 모든 체력 요인과 BMI 예측에서 다른 모델들에 비해 가장 낮은 성능을 보였다. SVR은 고차원 공간에서의 선형 회귀를 수행하는 데 강점을 지니지만, 본 연구의 데이터 특성상 비선형성이 충분히 복잡하지 않았으며, 이에 따라 SVR이 다른 모델들보다 예측력이 떨어지는 것으로 나타났다(Vapnik, 1995). 다시 정리하면, 근력 예측에서는 CatBoost와 LightGBM 모델이 가장 높은 예측 정확도를 보였다. 특히 CatBoost 모델은 Test R^2 값이 0.96으로 가장 우수한 성능을 기록하였으며,

LightGBM 모델도 유사한 성능을 나타냈다. XGBoost와 Gradient Boosting 모델은 근력 예측에서 Test R^2 값이 각각 0.95로 비슷한 성능을 보였으나, Random Forest와 SVR 모델은 상대적으로 낮은 성능을 기록했다. SHAP 분석을 통해 확인된 주요 변수는 weight, height, age로, 특히 체중이 큰 영향을 미쳤으며, 이는 체중이 근력과 밀접한 관련이 있음을 시사한다. 보행 관련 변수들은 일부 모델에서 예측 성능을 높이는 데 기여했는데, 특히 Right ROM Adduct-Abduct, Right ROM Inver-Ever CV와 같은 보행 변수들이 모델의 예측 정확도를 향상시켰다. 이는 보행에서 발생하는 미세한 보행 변화가 근력에 영향을 미칠 수 있음을 의미한다.

순발력 예측에서도 CatBoost와 XGBoost 모델이 가장 우수한 성능을 보였으며, Gradient Boosting과 LightGBM 모델이 그 뒤를 이었다. Test R^2 값이 CatBoost에서 0.95, XGBoost에서 0.93으로 높은 성능을 기록했으며, 이들 모델이 순발력과 같은 복잡한 체력 요인을 잘 예측하는 것으로 나타났다. SHAP 분석에서는 weight와 height가 주요한 변수로 나타났으며, 체중이 클수록 순발력 예측값이 증가하는 경향을 보였다. 특히 보행 관련 변수인 Right Single Support CV, Left ROM Dorsi-Plantar CV와 같은 특성이 일부 모델에서 순발력 예측에 중요한 영향을 미치는 것으로 확인되었다. 이는 순발력과 보행 패턴 간의 상관관계를 나타내며, 보행의 미세한 변동이 순발력과 관련될 수 있음을 시사한다.

심폐지구력 예측에서는 CatBoost와 Gradient Boosting 모델이 가장 우수한 성능을 보였으며, Test R^2 값이 각각 0.96과 0.94로 기록되었다. LightGBM과 XGBoost 모델도 유사한 성능을 보였으나, Random Forest와 SVR 모델은 상대적으로 낮은 성능을 나타냈다. SHAP 분석에 따르면, 심폐지구력 예측에서 weight가 가장 중요한 변수로 나타났으며, 체중이 증가할수록 심폐지구력 예측값이 증가하는 경향을 보였다. 특히 Right ROM Inver-Ever CV, Right Single Support CV와 같은 보행 관련 변수가 XGBoost와 Random Forest 모델에서 중요한 예측 변수로 나타났는데, 이는 보행 패턴이 심폐지구력에도 밀접한 관련이 있을 수 있음을 시사한다. 보행 중에 발의 미세한 회전과 지지 패턴이 심폐지구력에 영향을 미친다는 점은 운동 수행 중 보행 안정성이 심폐 능력에 중요한 역할을 할 수 있음을 보여준다.

유연성 예측에서는 CatBoost와 LightGBM 모델이 Test R^2 값이 각각 0.97로 가장 높은 예측 성능을 보였다. Gradient Boosting과 XGBoost 모델도 유사한 성능을 보였으나, Random Forest와 SVR 모델의 성능은 다소 낮았다. SHAP 분석 결과, weight와 height가 유연성 예측에서 주요 변수로 나타났으며, 체중이 클수록 유연성 예측값이 감소하는 경향을 보였다. 이는 체중이 유연성에 부정적인 영향을 미칠 수 있음을 시사하며, 나이가 많을수록 유연성이 낮아지는 패턴도 확인되었다. 이와 함께, Left ROM Adduct-Abduct CV, Right ROM Inver-Ever CV와 같은 보행 관련 변수들이

유연성 예측에 중요한 영향을 미치는 것으로 나타났다. 이는 보행 중 발생하는 가동 범위와 보행 변화가 유연성과 밀접한 관련이 있음을 나타내며, 특히 보행 중 양측 발의 움직임 균형이 유연성에 중요한 요소로 작용할 수 있음을 시사한다.

BMI 예측에서도 CatBoost와 LightGBM 모델이 가장 우수한 성능을 보였으며, Test R² 값이 각각 0.97로 기록되었다. XGBoost와 Gradient Boosting 모델도 Test R² 0.96으로 유사한 성능을 보였으며, Random Forest와 SVR 모델은 상대적으로 낮은 성능을 보였다. SHAP 분석에서는 weight와 height가 주요 변수로 나타났으며, 체중이 증가할수록 BMI 예측값이 증가하는 경향을 보였다. 특히 체중이 BMI 예측에서 가장 중요한 변수로 나타났으며, 이는 체중이 BMI에 미치는 영향을 명확하게 설명해준다. 보행 관련 변수들, 예를 들어 Right Single Support CV, Right ROM Inver-Ever와 같은 변수들이 일부 모델에서 BMI 예측에 중요한 영향을 미치는 것으로 나타났으며, 이는 보행 패턴과 신체 지지 시간이 BMI와 관련이 있음을 시사한다. 전반적으로, CatBoost와 LightGBM 모델이 모든 체력 요인에서 가장 높은 예측 성능을 보였으며, 이들 모델은 복잡한 데이터 구조를 처리하는 데 있어 강점을 나타냈다. Gradient Boosting과 XGBoost 모델도 유사한 성능을 보였으나, CatBoost와 LightGBM에 비해 다소 낮은 성능을 기록하였다. Random Forest 모델은 전반적으로 안정적인 성능을 보였으나, 부스팅 계열 모델보다는 성능이 떨어졌으며, SVR 모델은 모든 체력 요인에서 상대적으로 낮은 성능을 보였다. SHAP 분석을 통해 모든 체력 요인 예측에서 weight, height, age와 같은 신체적 특성이 중요한 변수로 나타났으며, 특히 체중은 대부분의 모델에서 가장 큰 영향을 미치는 변수로 확인되었다. 또한, 보행 관련 변수들인 Right Single Support CV, LeftROM Inver-Ever CV와 같은 보행 관련 변수들이 일부 모델에서 예측 성능을 향상시키는 데 중요한 역할을 했으며, 이는 보행 패턴과 보행의 변동이 체력 요인과 밀접한 관련이 있음을 시사한다. 이는 PAPS에서 측정되는 체력 요인들이 신체적 특성과 더불어 보행 패턴에도 영향을 받는다는 것을 보여주며, 모델 성능 향상을 위해 주요 신체 변수뿐만 아니라 보행과 보행 관련 변수를 신중하게 고려해야 함을 의미한다.

V. 결 론

5.1 연구 결과의 의의와 기대효과

본 연구는 머신러닝 모델을 사용하여 체력 요인(근력, 순발력, 심폐지구력, 유연성)

과 체질량지수(BMI)를 예측하고, 이를 통해 다양한 모델의 성능을 비교 평가했다는 점에서 중요한 의의를 가진다. 기존의 체력 평가 방법은 주로 물리적 검사와 측정에 의존해 왔으나, 본 연구는 데이터 기반의 예측 모델을 활용하여 보다 정교하고 효율적인 체력 평가 방법을 제시했다. 이러한 접근은 다양한 분야에서 체력 및 건강 지표를 평가하고 관리하는 새로운 패러다임을 제시하는 데 기여할 수 있다.

첫째, 본 연구의 가장 큰 의의는 CatBoost Regressor 모델이 체력 요인과 BMI 예측에서 일관되게 우수한 성능을 보였다는 점이다. 이는 CatBoost가 다양한 데이터 특성에 대해 높은 일반화 능력을 보유하고 있으며, 복잡한 패턴을 효과적으로 학습할 수 있음을 시사한다. CatBoost는 다른 머신러닝 모델들에 비해 범주형 변수를 다루는 데 탁월하며, 데이터 불균형 문제에 강한 면모를 보인다(Prokhorenkova et al., 2018). 이러한 결과는 CatBoost가 체력 및 건강 관련 데이터 분석에서 유망한 도구임을 보여준다. 더 나아가, CatBoost의 강력한 예측 성능은 건강 관리와 체력 평가 분야에서 새로운 표준을 세우는 데 기여할 수 있다. 이는 특히 스포츠 과학, 재활 치료, 공공 보건 정책 수립 등에서 실질적인 응용 가능성을 제시한다.

둘째, 본 연구는 다양한 머신러닝 모델을 체계적으로 비교함으로써, 각 모델의 강점과 약점을 명확히 규명했다. 예를 들어, LightGBM Regressor는 CatBoost에 비해 다소 낮은 성능을 보였으나, 빠른 학습 속도와 대규모 데이터 처리에 강점을 보였다(Ke et al., 2017). 이는 대량의 데이터를 빠르게 처리하고 분석해야 하는 상황에서 LightGBM이 유리할 수 있음을 시사한다. 반면, Random Forest Regressor는 결정 트리를 통한 예측 성능이 강력하나, 일부 체력 요인 예측에서 일반화 성능이 낮은 것으로 나타났다(Breiman, 2001). 이러한 분석은 각 모델의 특성을 이해하고, 상황에 맞는 최적의 모델을 선택하는 데 중요한 기준을 제공한다.

셋째, 본 연구는 머신러닝 모델을 활용한 체력 평가의 잠재적 가능성을 탐색했다는 점에서 의미가 있다. 기존의 체력 평가 방법은 주로 정적이고 일회적인 측정에 의존했으나, 머신러닝 모델은 데이터를 지속적으로 수집하고 분석함으로써 동적이고 적응적인 체력 평가를 가능하게 한다. 이는 예를 들어, 선수의 체력 상태를 실시간으로 모니터링하고, 이에 따라 훈련 프로그램을 조정하는 데 유용하게 활용될 수 있다. 또한, 일반인의 건강 관리에서도 개인의 체력 데이터와 BMI를 기반으로 맞춤형 운동 및 영양 프로그램을 제공하는 데 기여할 수 있다. 이러한 개인 맞춤형 접근은 공공 보건의 중요한 목표인 질병 예방과 건강 증진에 있어 큰 효과를 기대할 수 있다.

넷째, 연구 결과는 머신러닝 모델을 활용한 예측 결과의 해석 가능성에 대한 새로운 접근을 제시했다. 복잡한 머신러닝 모델, 특히 CatBoost와 같은 모델은 높은 예측력을 보유하지만, 그 예측 결과를 해석하는 데 어려움이 있을 수 있다. 본 연구는 모델의 예측 결과를 다양한 성능 지표(MSE, MAE, R^2)를 통해 분석함으로써, 결과의

해석 가능성을 높이고, 실질적인 응용에 필요한 정보를 제공했다. 이는 체력 평가와 건강 관리에서 머신러닝 모델이 실질적으로 활용될 수 있는 가능성을 높여준다.

마지막으로, 본 연구는 머신러닝 모델이 건강 관리와 체력 평가 분야에서 어떻게 적용될 수 있는지에 대한 구체적인 사례를 제공했다. 이는 향후 유사한 연구를 수행하는 연구자들에게 중요한 참고 자료가 될 수 있으며, 머신러닝 모델이 다양한 건강 데이터를 분석하고 예측하는 데 있어 강력한 도구로 사용될 수 있음을 입증했다. 이는 머신러닝이 기존의 정적이고 일회적인 건강 평가 방법을 보완하거나 대체할 수 있는 잠재력을 지니고 있음을 보여준다.

5.2 추후 연구과제

본 연구에서 도출된 결과는 다양한 가능성을 제시하며, 향후 연구에서 보다 깊이 있는 탐구와 확장을 필요로 한다. 첫째, 본 연구에서는 체력 요인 네 가지와 BMI를 중심으로 분석을 수행했으나, 더 다양한 건강 지표와 체력 요소를 포함하는 연구가 필요하다. 예를 들어, 혈압, 혈당 수치, 심박수 등 중요한 생리적 변수를 포함한 종합적인 건강 지표 예측 연구가 필요하다. 이러한 연구는 개인의 전반적인 건강 상태를 평가하는 데 있어 보다 종합적이고 정밀한 접근을 가능하게 할 것이다. 또한, 이러한 변수를 포함한 연구는 머신러닝 모델이 다양한 건강 지표를 동시에 예측하는 데 있어 어떻게 적용될 수 있는지에 대한 중요한 통찰을 제공할 수 있다.

둘째, 본 연구에서는 단일 데이터셋을 사용하여 모델을 평가하였으나, 다양한 인구 집단과 환경에서 수집된 데이터를 활용한 연구가 필요하다. 예를 들어, 연령, 성별, 인종, 생활 습관 등이 다양한 데이터셋을 사용하여 모델의 일반화 성능을 평가하고 개선할 필요가 있다. 이는 모델이 특정 데이터셋에 과적합되는 문제를 피하고, 다양한 조건에서의 예측 성능을 평가할 수 있는 기회를 제공할 것이다. 또한, 이러한 연구는 모델이 실제 환경에서 어떻게 성능을 발휘할 수 있는지를 평가하는 데 중요한 역할을 할 것이다.

셋째, 본 연구에서 제안된 모델을 실제 체력 평가 및 건강 관리 시스템에 적용하는 연구가 필요하다. 실제 환경에서 모델의 성능을 평가하고, 이를 통해 얻은 피드백을 바탕으로 모델을 개선하는 과정을 통해, 머신러닝 모델이 실질적인 건강 관리 도구로 자리 잡을 수 있을 것이다. 예를 들어, 스포츠 팀이나 피트니스 센터에서 본 연구의 모델을 활용하여 선수나 회원들의 체력 상태를 모니터링하고, 이에 따라 맞춤형 훈련 프로그램을 제공하는 시스템을 구축할 수 있다. 이러한 시스템은 체력 향상뿐만 아니라, 부상 예방과 재활 치료에도 큰 도움이 될 것이다.

넷째, 본 연구는 특정 시점의 데이터를 사용하여 예측을 수행하였으나, 시간에 따른

체력 변화와 건강 지표의 변동성을 반영하는 동적 예측 모델을 개발하는 연구가 필요하다. 이는 시계열 데이터나 연속적인 데이터 수집을 통해 개인의 체력 상태와 건강 지표가 시간에 따라 어떻게 변화하는지를 분석하고, 이에 따라 예측 모델을 구축하는 것을 의미한다. 이러한 연구는 특히 장기적인 건강 관리와 질병 예방에 있어 중요한 정보를 제공할 수 있다. 예를 들어, 개인의 체력 상태가 시간이 지남에 따라 어떻게 변화하는지를 분석함으로써, 질병 발생 가능성을 조기에 예측하고, 이를 예방하기 위한 조치를 취할 수 있다.

마지막으로, 본 연구는 주로 성능 지표를 중심으로 모델을 평가했으나, 사용자의 만족도와 실용성을 고려한 연구가 필요하다. 이는 모델의 예측 성능뿐만 아니라, 사용자가 모델을 얼마나 쉽게 이해하고 활용할 수 있는지를 평가하는 것을 포함한다. 예를 들어, 사용자 인터페이스(UI) 및 사용자 경험(UX) 디자인을 통해 모델의 사용성을 개선하고, 이를 통해 사용자의 만족도를 높이는 연구가 필요하다. 이러한 연구는 머신러닝 모델이 실제 응용에서 얼마나 효과적으로 사용될 수 있는지를 평가하는 데 중요한 역할을 할 것이다.

종합적으로, 본 연구는 머신러닝 모델을 활용한 체력 요인 및 BMI 예측에 대한 새로운 접근을 제시하였으며, CatBoost Regressor가 가장 우수한 성능을 보임으로써, 향후 유사한 연구에 중요한 참고 자료를 제공할 수 있을 것이다. 향후 연구에서는 본 연구의 결과를 바탕으로 다양한 건강 지표와 체력 요소를 포함한 종합적인 예측 모델을 개발하고, 이를 실제 환경에서 적용하여 모델의 성능을 평가하고 개선하는 것이 필요하다. 이러한 연구는 체력 평가와 건강 관리 분야에서 머신러닝 모델이 더욱 널리 활용될 수 있는 가능성을 열어줄 것이다.

참고문헌

- 교육부 (2009), *학생건강체력평가제 (PAPS) 측정 매뉴얼*, 교육부.
- 최인묵. (2017). 메타회귀분석을 이용한 여행사광고 커뮤니케이션 조절효과에 관한 연구: 국내 광고매체 연구의 4개 유형분류를 중심으로: 국내 광고매체 연구의 4개 유형분류를 중심으로, *서비스경영학회지*, 18(4), 205-229.
- Breiman, L. (2001). Random forests, *Machine learning*, 45(1), 5-32.
- Burnfield, M. (2010). Gait analysis: normal and pathological function, *Journal of Sports Science and Medicine*, 9(2), 353.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785-794.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine, *Annals of statistics*, 1189-1232.
- Janssen, I., & LeBlanc, A. G. (2010). Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *International journal of behavioral nutrition and physical activity*, 7, 1-16.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Kim, Y. K., Joo, J. Y., Jeong, S. H., Jeon, J. H., & Jung, D. Y. (2016). Effects of walking speed and age on the directional stride regularity and gait variability in treadmill walking. *Journal of Mechanical Science and Technology*, 30 (6), 2899-2906.
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.
- Lee, M., Youm, C., Jeon, J., Cheon, S. M., & Park, H. (2018). Validity of shoe-type inertial measurement units for Parkinson's disease patients during treadmill walking. *Journal of Neuroengineering and Rehabilitation*, 15.
- Mannini, A., & Sabatini, A. M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*, 10(2),

1154-1175.

- Muro-De-La-Herran, A., Garcia-Zapirain, B., & Mendez-Zorrilla, A. (2014). Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors*, 14(2), 3362-3394.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
- Strong, W. B., Malina, R. M., Blimkie, C. J., Daniels, S. R., Dishman, R. K., Gutin, B., ... & Trudeau, F. (2005). Evidence based physical activity for school-age youth. *The Journal of pediatrics*, 146(6), 732-737.
- Vapnik, V. (1995). The nature of statistical learning theory. *Springer science & business media*.
- Yoo, J. I., Baek, K. W., & Kim, J. S. (2019). Gait profile of K-pop art player measured with a shoe-type inertia measurement device during treadmill walking. *JEP online*, 22 (7), 110-119.

* 저자소개 *

· 이 성 문(sfknight@hanmail.net)

광운대학교 경영대학원에서 MIS 전공으로 석사학위를 취득하였다. 주요관심 분야는 디지털 측정, 운영관리, 품질보증이다.

· 임 성 욱(sulim@daejin.ac.kr)

성균관대학교 일반대학원에서 산업공학 박사학위를 취득하였으며, 현재 대진대학교 산업경영공학과 교수로 재직 중이다. 주요 관심분야는 품질지표, 식스시그마 경영, 디지털 품질혁신 등이다.