Journal of Korean Economic Studies
2023, Vol.41, No.2, 67~106
https://doi.org/10.46665/jkes.2023.6.41.2.67

머신러닝 분석을 통한 학생들의 비만도 예측*

임민숙** · 정현주*** · 이진형****

본 연구는 머신러닝을 통해 비만에 영향을 미치는 변수를 분석하는 것이 목적이다. 초등·중등·고등학생들의 개인, 학교, 지역 특성 그리고 건강 관련 설문 자료를 바탕으로 최소자승법(OLS)과 머신러닝 모형인 릿지(Ridge Regression), 라쏘(Lasso Regression)와 엘라스틱넷(ElasticNet Regression) 분석을 통하여 비만을 추정하고 예측하였다. 또한 MSE(Mean Squared Error)를 이용하여 위 모형을 비교 평가하였다.

초등학생과 중·고등학생의 비만을 추정하는 모형을 각각 나누어 분석하였다. OLS로 분석한 결과, 식습관과 수면시간이 유의하게 나타났다. 머신러닝으로 분석한 결과, 식습관과 수면시간 외에도 생활 습관, 학교와 지역 특성이 비만 예측에 필요한 변수로 선택되었다.

본 연구는 머신러닝을 활용하여 각급 학교 학생별 맞춤형 비만 예방 정책 마련에 기초가 되는 연구라는 점에서 의의가 있다.

핵심주제어: 머신러닝, 릿지, 라쏘, 엘라스틱넷, 예측, MSE, 비만, 학생
경제학문헌목록 주제분류: I, I0, I1, I2

Ⅰ. 서론

학생의 건강은 미래의 건강한 사회를 만들기 위한 중요한 요소이고 성장기에 건강관리가 미흡한 경우 성인이 된 후에도 만성 질환 등으로 이어질 가능성이 높아 이는 사회·경제적 비용을 가중하는 결과를 초래한다. 우리나라의 비만 관련 사회·경제적 비용을 살펴보면, 2006년에는 4.8조 원에서 2016년에는 11.4조

* 본 연구는 한국교육환경보호원 건강증진센터 연구용역을 바탕으로 보완한 논문임.
** 제1저자, 성균관대학교 일반대학원 경제학과, E-mail: minlim1984@gmail.com
*** 공동저자, 한국교육환경보호원 연구위원, 전화: (043) 710-4019, E-mail: jeong0209@gmail.com
**** 교신저자, 성균관대학교 경제학과 교수, 전화: (02) 760-0263, E-mail: leejinh@gmail.com
논문투고일: 2023. 3. 10 수정일: 2023. 4. 16 게재확정일: 2023. 5. 15

원으로 GDP의 0.7% 규모이고 최근 10년간 2.4배가량 증가하였다.[1)]

학생 시기는 전 생애에 걸쳐 건강한 시기이지만 우리나라 학생의 특성상 신체활동 부족, 영양 불균형, 다양한 행동 습관 등에 기인한 비만 학생 증가라는 문제가 대두되고 있다.[2)] 이에 학생들의 비만을 예방하는 정책이 교육부 정책의 주요 의제가 될 것으로 예상되며 이를 위한 기초 통계 자료뿐만 아니라 비만 학생에게 미치는 요인을 분석할 필요성이 있다.

본 연구는 학생들의 건강조사 데이터를 토대로 학생의 특징과 학교 특징을 분석하는 연구를 함으로써 학생들의 비만을 유발하는 요인 및 정보를 얻을 수 있다. 학생들은 학교급에 따라 다른 행동 유형이나 성장 형태를 보이기 때문에 이러한 특징을 반영한 분석이 필요한 상황이다. 기존 연구에서는 각급 학교별, 지역별 학생을 대상으로 분석한 연구(허영희 외, 2006; 성선화 외, 2007; 하영미 외, 2014)만 있을 뿐 초·중·고등학생을 모두 분석한 연구는 드물다.

학생에게 적절한 건강 증진 교육과 건강 서비스를 제공하고 건강한 교육환경을 조성하는 것은 우리나라의 미래를 결정하는 중요한 과제이다. 이를 위해 각급 학교의 특성과 학생 개인의 특성 중 비만을 유발하는 요인이 무엇인지를 분석하는 연구를 통해 학생의 건강과 관련된 연구의 기반을 마련할 필요가 있다.

기존의 연구에서는 비만에 미치는 요인을 분석하기 위해 주로 전통적인 회귀모형을 사용하였다. 물론 이들의 연구는 비만 학생들의 정책에 공헌한 점이 있으나 단순히 전통적인 선형 회귀모형은 모형 설정의 오류가 있을 수 있으며 전통적 선형 회귀모형에 의한 분석을 통해 과잉적합(overfitting)을 할 수 있는 한계가 있다.

특히 전통적인 선형 회귀모형은 잔차제곱합(Residual Sum of Squares: RSS)을 최소화하는 가중치 백터를 구하는 방법을 통해 직선형 회귀선을 최적화하는 방식이다. 그렇기 때문에 유사한 특성이 비만에 영향을 미치는 경우에는 예측력이 높게 나올 수 있다. 하지만 현대 사회를 살고 있는 학생들은 다양한 식습관, 생활습관을 가지고 있기 때문에 이러한 경우에는 예측력이 많이 떨어져, 예측에 있어서는 정확도가 많이 낮아지는 한계점을 가진다. 또한 연구자의 자의에 의해 변수를 선정하고 이를 바탕으로 회귀분석을 하기 때문에 추정력과 예측력에 한계를 갖게 된다.

---

1) 국민건강보험공단, 비만의 사회경제적 영향 조사 개요 외, 2018년.
2) 교육부, 2015년 학교건강검사 표본조사 결과, 2014년에는 21.8%에서 2015년에는 25.0%로 약 3.0%p 증가함.

그러므로 현대의 다양한 생활 양식을 갖고 있는 학생들의 비만을 분석해야 하는 연구에서는 이들 전통적인 선형 회귀모형에 전통적 모형의 설정을 바탕으로 한 추정은 모형 설정의 오류와 추정방법의 한계로 인하여 예측 오차가 많이 생길 수 있다. 또한 과잉적합의 문제가 발생하여 과장되거나 혹은 잘못된 추정과 예측을 하게 되는 연구 결과를 낳기도 한다(김지환, 2021; 이재득, 2021; 송상윤, 2015). 그래서 이러한 전통적 선형 회귀모형에 의한 추정과 예측의 한계점을 보완하기 위해 최근 머신러닝(Machine Learning) 기법을 이용한 해외 연구가 많아지고 있다(Aaron Kreiner *et al.*, 2019; Alexandre Bonnet R. Costa *et al.*, 2021; Isaac K. Ofori *et al.*, 2022; Jose Manuel Pereira *et al.*, 2016).

우리나라 연구에서도 머신러닝에 의한 경제학적 분석이 존재하지만 주로 학생 진로, 물류, 항만, 지역 경제분석에 있어 연구만 있을 뿐 학생 비만에 관련한 연구는 거의 없는 실정이다.

본 연구의 목적은 학생의 비만을 유발하는 다양한 변수들을 머신러닝 방법을 이용하여 예측하고 비교 평가하는 데 있다. 머신러닝 분석방법은 기존의 선형 회귀분석 방법을 바탕으로 하고 있고 연구자가 변수 선정을 고려할 필요 없이 보다 객관적이고 종합적으로 학생들의 비만에 영향을 주는 변수를 알아볼 수 있다. 또한 MSE와 R^2값을 산출하여 머신러닝 모형 중에서도 예측력이 높은 모형을 보는 데 본 연구는 의의가 있다. 본 연구의 구성은 다음과 같다. 제Ⅱ절에서는 학생 비만에 관한 기존의 연구 결과와 머신러닝을 활용한 기존 연구를 소개하고, 제Ⅲ절에서는 학생 비만에 미치는 요인 분석을 위한 자료 및 계량모형을 소개한다. 제Ⅳ절에서는 실증분석 결과를 제시하고, 마지막 제Ⅴ절은 연구의 요약과 결론을 제시한다.

Ⅱ. 선행연구

1. 학생의 비만

학생의 비만은 국내뿐 아니라 세계적인 문제이기 때문에, 학생의 비만에 영향을 미치는 요인에 관한 연구들은 많이 이루어졌다. 특히 비만과 밀접한 관련성이 있는 식습관과 비만과의 연관성을 연구는 많이 찾아볼 수 있는데, 성선화 외(2007)는 전라북도 전주시 소재 4개교의 중학생 450명을 대상으로 한 설문 자료

를 바탕으로 학생들의 식습관과 비만과의 관계를 연구하였다. 비만이 아닌 학생은 비만인 학생에 비해 식사습관이 규칙적이고, 비만인 학생은 비만이 아닌 학생에 비해 비간식 횟수가 더 많고, 라면, 햄버거, 피자, 초콜릿, 사탕을 섭취하는 빈도가 높은 것으로 보고 되었다. 이규영 외(2008)는 우리나라 중·고등학교 학생들의 패스트푸드 및 탄산음료 섭취와 비만의 관계에 대해 지역별로 비교 연구하였다. 전국 중·고등학교 중 총 22개교 총 2,261명의 설문 자료를 연구 대상으로 분석하였는데, 과체중 학생은 대도시, 중소도시, 읍면 지역 중 중소도시에 가장 많았고, 단음식을 먹는 학생들이 거주하는 지역을 보면 대도시, 중소도시 학생들이 유의하게 많이 먹는 것으로 보고 하였다. 또한 아침식사를 먹는 학생은 농촌 지역 학생이 대도시나 중소도시 학생에 비해 유의하게 적었고, 대도시 학생의 경우 패스트푸드를 식사대용으로 먹는 경우가 55.7%로 가장 많고 농촌 지역 학생의 경우는 간식으로 먹는 경우가 가장 많다고 보고하였다. 해외에서도 문헌연구를 통해 식습관과 비만과의 관계를 통계 분석한 연구가 있는데, K Kuźbicka *et al.*(2013)은 유럽 학생들의 비만과 식습관을 연구하였는데, 음료수를 포함한 간식, 아침식사를 거르는 습관 그리고 TV 화면 앞에서 식사를 하는 방식은 비만을 촉진하는 중요한 요소라고 보고하였다.

학생의 식습관과 비만의 연관성을 분석한 연구들은 청소년들의 올바른 식생활을 유도하며 건강 증진과 영양교육을 위한 자료를 제안하는 데 큰 의미가 있지만, 식습관 외에도 학생들의 다양한 생활 양식이 건강에 미치는 영향도 적지 않다. 특히 우리나라 학생들의 수면시간이 부족한 상황이기 때문에 수면습관과 비만과의 연관성을 연구하는 것도 큰 의의가 있다. 하영미·박현주(2014)는 고등학생의 수면과 비만, 그리고 스크린 타임 사이에 어떠한 관계가 있는지 파악하기 위하여 서울시 11개 교육구청별로 각각 1개 학교를 무작위 추출하여 일반계 고등학교 학생과 전문계 고등학교 학생 총 1,763명을 대상으로 한 설문 자료를 분석하였다. 수면시간의 경우 일반계 고등학생이 전문계 고등학생에 비해 짧고, 일반계 고등학생의 수면시간과 비만의 관계는 유의하지는 않았으나 전문계 고등학생의 경우 수면시간이 짧을수록 비만일 확률이 유의미하게 높은 것으로 보고하였다. 홍민희(2019)도 청소년의 건강 행태와 비만과의 관련성을 분석하였는데 수면시간 6시간 이하에서 비만군, 6시간 초과에서 정상군이 모두 통계적으로 유의하게 높은 비율을 나타냈다고 보고하였다. 한편, 학령기 아동의 수면시간과 비만과의 관련성을 연구한 김유라 외(2011)는 수면시간과 비만은 유의미한 차이는 나타나지는 않는다고 밝혔다.

2. 머신러닝

이러한 선행연구를 종합한 결과 일부 변수들의 통계적 유의성이 일관되지 않아 종합적인 시사점 도출 및 일반화에 어려움이 있었다. 이에 대한 이유로 연구마다 사용된 변수의 조합이 다르기에 사용된 변수의 영향력과 기여도가 변화될 수 있다(정애경 외, 2008).

Ferdowsy *et al.*(2021)은 비만과 비만이 아닌 두 집단으로부터 데이터를 수집하여 의사결정(Decision: DT), 랜덤 포레스트(Random Forest: RF), 로지스틱 회귀분석(Logostic Regression: LR) 등 9개의 방법을 적용하여 비만을 예측하였다. 이 중 LR이 97.09%의 확률로 가장 정확하게 비만을 예측한다고 밝혔다. 김은주 외(2020)는 한방 비만 프로그램을 시행한 과체중, 비만 성인 환자들을 대상으로 머신러닝 방법 중 DT, RF, LR 그리고 인공신경망(Artificial Neural Network: ANN)을 통해 체중 감량 예측에 효과적인지 분석하였다. 중위수 체중을 이용하여 랜덤 포레스트 모형을 사용하면 체중 감량 예측을 정확하게 할 수 있다고 밝혔다.

머신러닝 중 라쏘 회귀분석(Lasso Regression)을 이용하여 개인 단위의 비만을 예측한 국내 연구는 없다. 국내에서 라쏘 회귀분석은 다음과 같은 연구에 인용된 바 있다. 길은규 외(2018)는 지역단위 데이터를 사용하여 라쏘 회귀분석 방법을 통해 강원도의 강릉시와 원주시의 경제 성장률과 비만율이 특정 관계를 가진다는 가정하에 연구하였다. 분석 결과 경제 성장률의 흐름과 달리 비만율을 예측하기 힘들다고 밝혔다. 노민정(2019)은 머신러닝 기법 중 벌점회귀모형으로 분류되는 adaptive LASSO를 사용하여 청소년의 진로 결정과 관련된 변수를 살펴보았는데, 학생의 정서 문제, 진로정체감, 양육 방식, 학교생활 적응과 같이 선행연구에서 유의미하게 영향을 미친다고 밝힌 변수들이 진로 결정과 관련되어 있다고 보고하였다. 또한 직업관, 지역사회 인식, 휴대전화 보유 여부, 체험 활동 및 동아리 활동 참여 유무 변수들이 라쏘 분석을 통해 새롭게 진로 결정에 연관되어 있다고 보고하였다. 또한 이재득(2021)은 머신러닝 분석방법을 통해 부산의 경제 활성화를 위해 부산시의 7대 전략산업들의 22개 분야 중 어떠한 전략산업이 고용 효과가 있고 소득 증가를 가져올지 분석하였다. 정확한 추정과 예측을 위해 기존 계량경제 연구방법인 최소자승법(OLS)과 머신러닝 기법 중 릿지 회귀분석(Ridge Regression)과 라쏘 회귀분석을 사용하였다. 릿지 분석 결과 서비스플랫폼, 콘텐츠, 스마트금융산업으로 이루어진 지능정보서비스 산업이 고용

효과를 가장 높게 증가시키는 것으로 나타났고 소득도 가장 크게 증가시키는 것으로 나타났다. 라쏘 모형에서는 지능정보서비스 산업이 고용 효과를 가장 높게 증가시키는 것으로 나타났으나 소득을 증가하게 해주는 산업에는 지능정보서비스산업, MICE 그리고 글로벌 관광산업이 있다고 보고하였다.

Ⅲ. 분석 자료 및 분석방법

1. 분석 자료

본 연구의 분석 대상은 교육부 학생 건강검사 표본으로 2010년부터 2019년까지 초·중·고등학교 학생의 신체발달(키, 몸무게) 계측 자료와 자기기입식으로 건강 행태를 응답한 건강조사 자료이다. 분석단위는 학생이고 결측(missing) 데이터는 분석 대상에서 제외하였다.

2. 기초 통계 분석

분석 대상인 초등학생은 452,130명이고, 중학생은 304,190명 그리고 고등학생은 324,060명이다. 연도별로 살펴본다면 2010년은 총 186,022명, 2011년 180,368명, 2012년 86,851명, 2013년 84,424명, 2014년 82,516명, 2015년 84,782명, 2016년 82,828명, 2017년 80,446명, 2018년 107,594명, 2019년 104,189명이다.

학생들의 비만 인지 여부는 BMI[3] 수치를 바탕으로 판단하기 때문에 학생들의 신장과 체중의 특징이 중요하다. 각급 학교별 학생들의 초등학생의 평균 키는 135.68cm, 2011년 135.83cm, 2013년 135.98cm로 2010년 이후로 2019년까지 꾸준히 증가하고 있고, 중학생·고등학생의 경우에도 2010년부터 2019년까지 평균 신장은 지속적으로 증가하였다. 또한 평균 몸무게 역시 초등학생의 경우는 2010년에서 2013년까지 증가하였으나 2014년, 2015년에서는 감소하였고 다시 2016년부터 2019년까지 증가하였다. 중학생·고등학생의 경우는 2010년부터 2019년까지 증가하였다.

---

3) BMI=(몸무게)/(키)^2.

〈표 1〉 기초통계량[4)]

		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
각급 학교 (명)	초등	87,040	83,092	35,914	34,415	33,467	33,605	33,338	33,071	39,107	39,081
	중등	50,557	49,569	26,271	25,391	24,639	23,964	22,859	21,741	30,154	29,045
	고등	47,425	47,707	24,666	24,618	24,410	27,213	26,631	25,634	38,693	36,063
	합계	186,022	180,368	86,851	84,424	82,516	84,782	82,828	80,446	107,954	104,189
성별 (명)	남	99,864	97,329	44,962	43,660	42,557	43,137	42,219	40,950	55,027	52,994
	여	86,159	83,039	41,889	40,764	39,959	41,645	40,609	39,496	52,927	51,195
평균 키 (cm)	초등	135.68	135.83	135.98	136.02	136.02	136.10	136.21	136.21	136.44	136.62
	std.dev	11.63	11.69	11.78	11.89	12.01	12.00	12.15	12.15	12.06	12.20
	중등	160.90	161.08	160.72	161.03	161.32	161.70	161.95	162.04	162.22	162.38
	std.dev	8.01	7.98	7.85	7.87	7.84	7.79	7.81	7.87	7.87	7.90
	고등	167.46	167.47	166.77	166.64	166.75	166.56	166.65	166.69	167.04	167.19
	std.dev	8.23	8.27	8.24	8.25	8.25	8.26	8.30	8.26	8.32	8.39
평균 몸무게 (kg)	초등	34.01	34.08	34.21	34.56	34.35	34.46	34.69	34.92	35.16	35.40
	std.dev	10.49	10.52	10.67	10.84	10.94	10.98	11.30	11.48	11.43	11.57
	중등	54.13	54.52	54.26	54.88	55.04	55.37	55.99	56.17	56.79	56.98
	std.dev	11.88	11.85	11.81	11.94	12.09	12.15	12.56	12.63	12.07	13.40
	고등	61.53	61.66	61.20	61.56	61.78	61.94	62.43	63.01	63.32	63.68
	std.dev	12.36	12.27	12.25	12.34	12.83	12.80	13.37	13.73	13.85	14.28

3. 분석 모형

1) 전통적인 최소자승(OLS) 선형 회귀모형 추정

전통적인 계량경제학 분야에서는 일반적으로 선형 회귀모형인 최소자승(OLS) 모형을 통해 분석한다. 최소자승법 모형은 오차항에 대해 확률변수는 독립적이고, 확률분포는 모수까지 완벽하게 동일하다는 가정(Independent and Identically Distributed: IID가정)하에서 종속변수 y값은 $y = f(x) = ax + b$ 형태로 설정되며 독립변수들은 x들에 의해 설정된다.

이때 선형모형의 독립변수들의 종속변수에 대한 회귀분석에 의해 추정되는 y

4) 기초통계량은 이 연구의 분석을 위해 구축한 자료를 바탕으로 한 결과로 교육부가 발표한 학생 건강검사 표본 통계와 차이가 있을 수 있음.

값과 실제값들에 대한 차이를 나타내는 잔차가 최소가 되는 추정회귀 계수를 구한다. 또한 실제값과 예측값의 차이의 제곱의 합을 최소화시키는 추정계수들인 x값들의 추정치인 β값들을 구한다.

2) 머신러닝 분석 기법

하지만 최소자승자모형(OLS)과 같은 선형회귀모형은 x들을 연구자가 선별하는 과정에서 모형의 오류와 오차의 성질 등에 의해 과잉적합의 심각한 오류가 발생할 가능성이 있다. 이 점을 극복하기 위하여 중요한 변수를 선정하고 중요하지 않은 변수를 버리는 작업인 선택(selection)을 하게 된다. 또한 연구에서 사용하는 변수들이 고차원일 때 회귀계수에 벌점축소(shrinkage) 방법을 통해 데이터를 다루는데, 본 연구에서 사용할 릿지(Ridge), 라쏘(Lasso) 그리고 엘라스틱넷(ElasticNet)은 축소(shrinkage) 방법에서 대표적으로 사용하는 벌점회귀(regularized regression)모형이다(김지환, 2022; 송상윤, 2015).

벌점회귀모형은 회귀계수에 벌점을 부과하여 추정하는 계량기법이고, 특성상 연구자의 직관에 의존하지 않고 모든 잠재적인 설명변수들을 동시에 모형에 고려할 수 있는 장점이 있다. 더불어 벌점화 축소 추정 기법은 회귀계수의 편의(bias)는 반영하지만 분산(variance)을 줄여 예측오차를 낮출 수 있다(송상윤, 2015). 본 연구에서는 이를 확인하기 위하여 각 모형의 MSE(Mean Squared Error)를 이용하여 방법론들의 예측력을 확인하였다. 본 연구는 STATA 17(StataCorp. 2021)을 활용하였다.

3) 릿지 회귀, 라쏘 회귀 그리고 엘라스틱넷 회귀모형 추정

릿지 회귀분석의 규제항은 식 (1)과 같이 파라미터의 제곱의 합으로 이루어져 있다. 이것은 미분이 가능해 Gradient Descent 최적화가 가능하고, 파라미터의 크기가 작은 것보다 큰 것을 더 빠른 속도로 줄여준다.

$$L(\beta) = \min \sum_{i=1} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{p} \beta_j^2 \quad (1)$$

규제항의 λ 가 크면 클수록 릿지 회귀의 계수 추정치는 0에 가까워진다. 예를

들면, λ =0일 때는 규제항은 효과가 없고, 따라서 릿지 회귀분석은 최소제곱 추정치를 생성한다. 즉, λ 가 규제를 얼마나 부과하는가를 조절하는 역할을 한다.

라쏘 회귀계수는 식 (2)에서 나타나 있듯이 릿지 회귀와 비슷하게 생겼지만 규제항에 절댓값들의 합을 사용한다.

$$L(\beta) = \min \sum_{i=1} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{p} |\beta_j| \tag{2}$$

릿지 회귀분석에 비해 유의미하지 않은 변수들에 대해 계수를 0에 가깝게 추정해 주어 변수 선택 효과를 가져온다. 라쏘 모형은 파라미터의 크기에 관계없이 같은 수준의 제약(regularization)을 적용하기 때문에 작은 값의 파라미터를 0으로 만들어 변수를 모형에서 삭제하고 모형을 단순하게 만들어 해석을 용이하게 한다. λ 는 로그 스케일상에 균등 분포한 100개의 값을 이용하여 격자 탐색법(grid search)을 적용하고 교차 타당성(Cross Validation)을 통하여 예측오차의 평균 가장 작은 값을 선택한다. 대표적으로 K-fold Cross Validation 방법을 사용하는데, 전체 데이터 세트를 K개의 fold로 나누어 K번 다른 fold 1개를 테스팅 데이터(Testing Data)로, (K-1)개의 fold를 트레이팅 데이터(Traning Data)로 분할하는 과정을 반복함으로써 트레이닝 데이터와 테스팅 데이터를 교차 변경하는 방법론이다. 모든 데이터를 트레이닝 데이터와 테스팅 데이터로 활용하여 과적합 및 과소적합을 방지하고 더욱 일반화된 모형 생성이 가능하다.

엘라스틱넷 회귀(ElasticNet Regressiion)모형은 식 (3)에서 볼 수 있듯이 릿지와 라쏘의 제약 조건을 모두 갖는 모형이다. 엘라스틱넷 회귀모형은 식에서 λ 값을 0과 1 사이에서 조절하여 비용함수의 패널티를 조절한다.

$$L(\beta) = \min \sum_{i=1} (y_i - \hat{y}_i)^2 + \lambda (\sum_{i=1}^{p} |\beta_j| + \sum_{i=1}^{p} \beta_j^2) \tag{3}$$

식 (1), (2), (3)에서 첫 번째 항에 해당하는 오차항의 표준편차는 곱셈 형태의 상수가 되어 못적함수를 최소화하고 그 과정에서 제거되면서 추론 기반의 라쏘 회귀분석의 알고리즘을 간단하게 하는 장점이 있지만 예측을 위한 라쏘 회귀분석 모형으로도 사용할 수 있다. 다만 예측의 목적에 사용하는 경우는 일관성을 유지하기 위해서 λ 의 선택 과정에서 교차 타당성을 이용하는 것이 반드시 필요

하다.

4. 예측력 평가지표

머신러닝의 예측력 평가지표는 다양하다. 이 중 김은주 외(2020)는 예측력을 평가하기 위해 DT, RF, LR, ANN으로 분석한 후 ROC(Receiver Operating Characteristics) 커브를 통해 AUC(Area Uder the Curve)를 활용하여 모델의 성능을 평가하였다. 다만 이 지표는 분류 척도(Classification Metrics)로서 어떻게 분류하여 예측하는 것이 더 적합한지를 평가할 경우 사용하는 방법이다. 회계척도(Regression Metrics)의 경우에는 MSE, RMSE, R-squared가 있는데, 이 지표들은 계량경제학에서는 주어진 모형의 예측력을 평가하기 위해 주로 개발되었다. 이 중 예측값과 실제값의 차이로 정의되는 예측 오차의 크기를 측정하기 위한 MSE와 일반적인 회귀모형에서 이용되는 R-squared를 사용하였다. 다만 머신러닝에서는 설명력을 나타내는 R-squared 값보다는 오차의 크기를 나타내는 MSE가 더 의미 있는 것으로 판단되어 MSE 값으로 최소자승(OLS) 모형과 머신러닝 모형인 릿지 회귀모형, 라쏘 회귀모형 그리고 엘라스틱넷 회귀모형들 간의 예측력을 평가하였다(이재득, 2021). 예측력을 평가하기 위한 테스팅 데이터는 전체 분석 대상 중 75%의 트레이팅 데이터 이외에 남은 25%를 이용하였다. 실제값과 예측값 차이의 제곱의 합을 최소화시키는 추정계수들인 x값들의 추정치인 β 값들을 구하고, 그 추정의 손실은 MSE에 의해 식 (4)와 같이 정의된다.

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \tag{4}$$

$\hat{y}_t$는 모형을 통해 도출한 예측값, y_t는 실제값, 그리고 T는 테스팅 데이터의 크기를 나타낸다.

Ⅳ. 분석 결과

1. 분석 대상 변수 선정

분석 대상에 사용된 변수는 〈표 2〉와 같다. 구체적인 변수의 설명은 부록에 작성하였다. 분석에 활용하기 위한 분석 변수 선정을 위해 다중공선성(VIF)을 검토한 결과 초등학생, 중·고등학생들은 다른 신체적 성장 과정을 겪고 있고, 다른 환경에서 성장하고 있으므로 보다 정확하게 비만에 영향을 미치는 요인을 실증분석을 하기 위해 초등학생을 대상으로 분석한 모형과 중·고등학생을 대상으로 분석한 모형을 세웠다.

종속변수는 비만 여부이고 더미변수를 사용하였다. 주요 설명변수로는 인적 특성인 연령, 월령, 성별을 사용하였다. 중·고등학생의 경우에는 월령이 크게 의미가 없을 수 있으나 초등학생의 경우에는 성장속도가 개인마다 다를 수 있어 월령이 의미가 있을 수 있다고 생각하여 변수에 포함하였다. 또한 라면, 음료수, 패스트푸드 등을 일주일 동안 먹는 횟수, 아침식사 습관을 식습관 관련 변수로 포함하였다.

특정 식품과 관련된 식습관은 일주일 동안 먹는 횟수에 대한 설문조사에서 1번부터 4번까지 응답하도록 되어 있다. 숫자가 더 클수록 해당 식품을 더 자주 먹는 것으로 나타난다. 조사 결과 학생들은 평균적으로 1주일에 우유와 같은 유제품을 음료수보다 더 자주 먹는 것으로 나타났으며, 육류를 패스트푸드보다 더 자주 먹는 것으로 나타났다. 아침식사 습관도 1번부터 4번까지 응답하도록 되어 있으며, 숫자가 클수록 더 자주 먹는 것으로 나타난다. 전반적으로, 학생들은 아침식사를 '대체로 먹는다'는 경향을 보인다.

또한 전체 학생 중 43.22%가 다이어트를 한 경험이 있다고 밝혔다. 이들은 다이어트를 위해 약을 복용한 것보다는 식단조절과 운동을 하였다는 것이 조사 결과로 나타났다. 자아신체상(body image)도 변수로 포함되어 있으며, 학생들은 1번부터 5번까지 응답하도록 되어 있다. 숫자가 클수록 학생들이 친구들과 비교해서 자신의 체형이 매우 뚱뚱한 편이라고 생각하는 것을 나타낸다. 분석 결과, 학생들은 평균적으로 자신의 체형을 '보통'으로 생각하고 있다는 것으로 나타났다.

또한 손 씻기, 양치질, 안전벨트, 안전장비 착용 등의 생활습관도 분석 대상에 포함하였다. 전체 학생 중 50% 이상이 손 씻기, 양치질, 안전벨트 착용을 잘

〈표 2〉 기초통계량 및 변수 설명

	변수		평균	표준편차	빈도	비율(%)	분석 모형
종속 변수	비만 여부	O	0.128	0.334	942,225	87.21	초·중·고등학생
		×			138,155	12.79	
인적 특성	연령		12.334	3.438			
	월령		153.464	41.054			
	성별	남	1.479	0.499	562,699	52.08	
		여			517,681	47.92	
식습관	라면		1.889	0.537			
	음료수		2.098	0.720			
	패스트푸드		1.727	0.567			
	육류		2.546	0.698			
	우유/유제품		3.022	0.936			
	과일		2.900	0.852			
	채소		2.852	0.878			
	아침식사		3.325	0.996			
다이어트 경험	아무것도 안 함	No	0.568	0.495	466,966	43.22	
		Yes			613,414	56.78	
	식단조절	No	0.241	0.428	820,005	75.90	
		Yes			260,375	24.10	
	약을 먹음	No	0.013	0.113	1,066,445	98.71	
		Yes			13,935	1.29	
	운동으로 감량	No	0.317	0.465	737,814	68.29	
		Yes			342,566	31.71	
자아 신체상	자아신체상		3.049	0.982			
생활 습관	하루 수면량		2.636	1.052			
	손 씻기	No	0.816	0.388	195,877	18.43	
		Yes			866,992	81.57	
	양치질	No	0.926	0.261	78,176	7.35	
		Yes			985,008	92.65	
	안전벨트 착용	No	0.593	0.491	433,162	40.74	
		Yes			630,021	59.26	
	안전장비 착용	No	0.335	0.472	706,289	66.45	
		Yes			356,577	33.55	
	주 3회 이상 운동	No	0.526	0.499	211,957	47.40	초등학생
		Yes			235,183	52.60	
	TV시청	No	0.358	0.479	286,645	64.13	
		Yes			160,349	35.87	
	게임	No	0.167	0.373	372,120	83.22	
		Yes			75,009	16.78	
정신적 요소	괴롭힘	No	0.051	0.219	424,495	94.93	
		Yes			22,689	5.07	

	현금갈취	No	0.014	0.117	440,962	98.61	
		Yes			6,222	1.39	
	신체접촉	No	0.013	0.113	441,372	98.70	
		Yes			5,822	1.30	
	가출 생각	No	0.052	0.223	423,781	94.77	
		Yes			23,389	5.23	
	가족 지지	No	0.853	0.354	65,678	14.69	
		Yes			381,491	85.31	
	체벌 경험	No	0.039	0.194	429,707	96.10	
		Yes			17,454	3.90	
	상담 요청	No	0.024	0.153	436,104	97.59	
		Yes			10,759	2.41	
	가족 흡연	No	0.424	0.494	257,413	57.56	
		Yes			189,762	42.44	
	가족 음주	No	0.173	0.379	369,676	82.67	
		Yes			77,520	17.33	
	무기력감	No	0.036	0.187	430,950	96.38	
		Yes			16,204	3.62	
	수업 태도 교정	No	0.054	0.226	422,176	94.60	
		Yes			24,085	5.40	
	고민 상담 희망	No	0.027	0.163	434,407	97.25	
		Yes			12,267	2.75	
생활 습관	운동		2.039	0.989			중·고등 학생
	1년 동안 치료 경험		1.581	0.929			
	게임	No	0.295	0.456	433,249	70.46	
		Yes			181,614	29.54	
	음란물/채팅	No	0.041	0.199	589,647	95.89	
		Yes			25,300	4.11	
정신적 요소	괴롭힘	No	0.034	0.181	594,312	96.63	
		Yes			20,738	3.37	
	고민 상담	No	0.787	0.41	131,294	21.35	
		Yes			483,785	78.65	
	가정 문제	No	0.153	0.36	520,589	84.65	
		Yes			94,378	15.35	
	가출 생각	No	0.074	0.261	569,669	92.63	
		Yes			45,351	7.37	
	폭력위협	No	0.01	0.101	608,690	98.97	
		Yes			6,356	1.03	
	학교 문제 상담요청	No	0.055	0.227	581,252	94.55	
		Yes			33,529	5.45	

	흡연/음주 상담희망	No	0.014	0.119	606,229	98.57	
		Yes			8,779	1.43	
	성 문제 상담희망	No	0.012	0.108	607,503	98.83	
		Yes			7,219	1.17	
	고민 상담희망	No	0.113	0.317	542,259	88.71	
		Yes			68,997	11.29	
지역	서울				130,739	12.10	초·중·고등학생
	부산				31,947	5.73	
	대구				56,258	5.21	
	인천				58,460	5.41	
	광주				39,845	3.69	
	대전				43,039	3.98	
	울산				37,887	3.51	
	세종				10,284	0.95	
	경기				202,376	18.73	
	강원				47,550	4.40	
	충북				47,202	4.37	
	충남				53,943	4.99	
	전북				52,480	4.86	
	전남				57,368	5.31	
	경북				67,588	6.26	
	경남				83,006	7.68	
	제주				30,408	2.81	
학교	설립 형태	공립			871,136	80.63	
		국립			5,904	0.55	
		사립			203,340	18.82	
	남녀공학	O			197,093	18.24	
		×			883,287	81.76	
	학생 수 합계		828.954	382.336	382.336		
	학급당 학생 수		33.022	16.406	16.406		
	교원 수 합계		54.836	21732	21732		
	남자 교원 수		23.221	14.848	14.848		
	여자 교원 수		31.615	16.930	16.930		
	교원 1인당 학생 수		14.811	3.717	3.717		

하고 있으나, 헬멧이나 안전장비 착용은 33.55%만 실천하고 있다는 것이 나타났다. 손 씻기, 양치질은 위생과 관련된 행동으로 각종 질병 예방에 대한 가장 기본적인 방법이며, 진혜정 외(2013)는 청소년의 칫솔질 1회 증가 시 손 씻기 횟수가 1.5배 증가하는 것으로 밝혔다. 특히 치아우식증 유병률과 칫솔질 횟수는 유의한 관계가 있다고 보고되었다(박신영, 2017; 이종화 외, 2015). 그리고 양치질을 하지 않는다면 비만인 사람은 숙주의 면역력과 염증을 변화시킬 가능성이 더

크기 때문에 손 씻기와 양치질과 같은 생활습관은 비만과 밀접한 관련이 있다고 사료되었다(김선일 외, 2018).

비만과 안전벨트 착용은 비만과 밀접한 관련이 있다고 보고된 바 있고, 비만인 학생들은 그렇지 않은 학생들에 비해 안전벨트를 사용하기에 불편함을 느낄 수 있다고 밝혔다(Price *et al.*, 2011). 안전장비 착용 역시 이와 같은 맥락으로 비만과 관련이 있다고 판단하여 분석 대상으로 삼았다.

따돌림 경험, 고민상담 희망 등 정신적 요소, 스크린 타임, 운동습관 등은 초등학생과 중·고등학생의 설문이 각각 달라 따로 모형을 분리하여 분석하였다. 특히, 스크린 타임[5]을 분석 대상 데이터에서는 초등학생의 경우는 하루에 TV 시청과 게임을 2시간 이상 하는지를 조사하였다. TV는 초등학생의 35%가 2시간 이상 시청을 하고 게임은 16.78%가 2시간 이상 하는 것으로 나타났다. 또한 중·고등학생의 경우는 음란물 채팅을 자주하는지 여부와 게임을 2시간 이상 하는지 여부를 조사하였다. 중·고등학생의 29.54%가 하루 2시간 이상 게임하는 것으로 나타났다. 학생들의 스크린 타임이 증가할수록 수동적 신체활동 시간이 늘고 높은 에너지를 소비할 수 있는 신체활동의 시간이 줄어들게 됨에 따라 비만과 밀접한 관련이 있다고 보고 된 바 있다(김재등 외, 2009).

그리고 스크린 타임은 수면량과도 관련이 있는데 하루 수면량에 관한 변수는 1부터 4번으로 되어 있고 숫자가 클수록 하루 수면시간을 더 많이 가지는 것을 의미한다.

정신적인 측면을 나타내는 변수들을 살펴보면, 초등학생 중 14.69%가 가족의 지지를 받지 못한다고 생각하고 있으며, 가족 구성원 중에서 술을 많이 마시는 것으로 인해 걱정하는 학생은 17.33%이다. 또한 초등학생 중 2.75%가 고민 상담을 희망한다는 결과가 나왔다. 중·고등학생의 경우, 11.29%가 고민 상담을 희망하는 것으로 나타났는데, 이는 초등학생들에 비해 중·고등학생들이 고민이 늘어나는 것으로 해석할 수 있다.

이규영(2008)은 과체중 학생의 비율은 대도시, 중소도시, 읍면 지역 중 중소도시에서 높다는 것을 밝혀 학생이 거주하고 있는 지역은 분석에 있어서 유의미하게 영향을 미칠 것이라고 생각하여 분석 대상 변수로 선택하였다. 또한 학생이 속해 있는 학교 특성도 의미가 있을 것이라고 판단하여 학교 특징인 학교의 설립 형태, 교원 수 합계, 1인당 교원 수를 학교 관련 변수로 포함하였다.

---

5) 스크린 타임은 TV, 컴퓨터 등의 전자기기 앞에서 보내는 시간으로 낮은 에너지를 소모하는 활동을 말한다(하영미 외, 2014).

2. 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 분석

본 절에서는 학생들의 특징이 비만에 영향을 미치는 요인을 분석하기 위해 학생들의 식습관, 생활습관, 주변 환경 그리고 학교, 지역 특징을 중심으로 머신러닝 기법의 학생들의 비만에 대한 예측모형을 도입하였다. 이를 위해서 〈표 2〉에서 obes_1 변수, 즉 비만인지 여부를 종속변수로 선정하였고 나머지 변수들을 각급 학교 모형에 맞게 독립변수로 선택하여 추정하였다.

초등학생, 중·고등학생들의 개인적 특성, 학교와 지역적 특성 중 어떠한 특성이 비만에 영향을 미치는지 분석하기 위하여 전통적인 선형 회귀모형인 최소자승법(OLS)을 먼저 사용한 후, 머신러닝 추정 모형인 릿지 회귀모형 분석, 라쏘 회귀모형과 엘라스틱넷 회귀모형 분석방법으로 추정하고 예측력을 평가하였다.

본 연구에서는 머신러닝 기법을 활용한 분석을 위하여 각각의 예측모형을 평가하기 위해 트레이닝 데이터 세트(Training Dataset)와 테스팅 데이터 세트(Testing Dataset)를 75%와 25%로 각각 나누어 사용하였다.

1) 초등학생의 비만에 대한 예측 분석

본 절에서는 먼저 비만 여부를 종속변수로 선택한 후, 설명변수들인 식습관, 수면습관 그리고 통제변수 등으로 이루어진 총 11개 변수로 최소자승법(OLS)을 분석하고 분석 대상인 모든 변수로 머신러닝 기법인 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 분석으로 추정하여 비교한다.

⑴ 최소자승법(OLS) 추정

본 연구는 초등학생의 식습관이 비만에 미치는 영향을 알아보기 위해 최소자승법(OLS)을 활용하여 추정하였다. OLS 모형 분석을 위하여 선행연구(허영희 외, 2006; 성선화 외, 2007; 하영미 외, 2014)를 바탕으로 변수들을 선정하여 분석하였다.

식 (5)에서 $Obes_i$ 변수는 종속변수로서 비만인지 여부이고, 주요 독립변수인 식 (5)에서의 $Eating_i$ 변수들은 식습관에 관련된 변수들로서 라면, 음료수, 패스트푸드, 육류, 우유/유제품을 1주일에 몇 회 먹는지와 아침식사 습관에 관한 변수이다. 또한 $Sleep_i$ 는 하루에 수면시간이 얼마인지에 관한 변수이다. 식 (5)에서 Z_i 는 주요 통제변수로서 선행연구에 따라 학생들의 개인적인 특성인 거주

도시, 성별과 나이 그리고 조사연도를 통제변수로 추정하였다.

$$Obes_i = \beta_0 + \beta_1 Eating_i + \beta_2 Sleep_i + \beta_3 Z_i + \epsilon_i \quad (5)$$

〈표 3〉은 설명변수들과 종속변수를 OLS 회귀모형으로 추정한 결과를 나타낸다. 총 관측치 중 75%를 무작위로 트레이닝 데이터로 선정하여 335,545명을 대상으로 회귀분석하였다.

〈표 3〉을 보면 초등학생들의 라면, 음료수, 패스트푸드, 우유/유제품의 섭취가 비만에 유의적인 영향을 미치고 있는 것으로 나타난다. 반면 육류의 섭취는 비만에 음(-)의 영향을 미치고 있는 것으로 나타났다. 또한 수면시간은 비만에 음의 영향을 미치고 있는 것으로 나타났다. 종합해 보면 초등학생은 라면, 음료수, 패스트푸드는 적게 먹고, 아침식사를 자주 하고 육류를 많이 먹으면 비만일 가능성이 낮아진다. 또한 수면시간도 많을수록 비만일 가능성이 낮아짐을 알 수 있다.

〈표 3〉 초등학생 비만 OLS 분석 결과

		종속변수
		비만
식습관(Eating)	라면	0.00343*** (0.00120)
	음료수	0.00583*** (0.000932)
	패스트푸드	0.0102*** (0.00111)
	육류	−0.0135*** (0.000933)
	우유/유제품	0.00253*** (0.000655)
	아침식사 습관	−0.00585*** (0.000734)
수면습관(Sleep)	수면시간	−0.0184*** (0.000750)
통제변수(Z)	거주 도시	0.000369*** (4.42e−05)
	조사연도	0.00434*** (0.000187)
	나이	−0.00133*** (0.000331)
	성별(남=0, 여=1)	−0.0369*** (0.00113)
	Constant	−8.496*** (0.377)
	Observations	335,545
	R-squared	0.009

주: *** p<0.01, ** p<0.05, * p<0.1.

(2) 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 추정

〈표 4〉는 가공이 완료된 전체 자료 관측치로 머신러닝 방법인 릿지 회귀모형, 라쏘 회귀모형 그리고 엘라스틱넷 회귀모형으로 비만인지 여부를 예측한 결과이다. 릿지 회귀모형의 경우 443,920개 가운데 무작위로 선정된 75%에 해당하는 332,952개의 관측치로 구성된 트레이닝 데이터를 이용하였고, 라쏘 회귀모형의 경우 333,060개의 관측치로 구성된 트레이닝 데이터, 엘라스틱넷 회귀모형의 경우 333,690개의 관측치로 구성된 트레이닝 데이터를 이용하여 분석하였다.

분석 결과, 머신러닝 기법 3개의 모형의 조율 변수 λ 를 포함한 모형의 전반적인 적합도 지표들을 알 수 있다. 트레이닝 데이터를 분석하여 얻은 CV-예측오차의 관점에서 라쏘 회귀모형과 엘라스틱넷 회귀모형이 최선의 모형으로 선택될 수 있는 것을 확인할 수 있다.

〈표 5〉는 머신러닝을 통해 비만 예측모형의 추정 결과이다. 추정치의 값은 비만에 영향을 미치는 정도를 말한다(송상윤, 2015). 머신러닝 기법을 통해 예측한 결과 선행연구를 바탕으로 구축한 모형 OLS에서 분석 대상으로 삼았던 변수들인 식습관, 수면습관, 거주 도시, 나이 그리고 성별은 머신러닝 모형에서도 채택하였다. 또한 Brixval *et al.*(2012)의 선행연구에서 언급한 따돌림이나 괴롭힘을 경험하였는지 여부와 자아신체상도 비만 예측모형에서 분석 대상으로 사용하였다.

머신러닝 모형에서는 추가적으로 다이어트 경험, 손 씻기, 양치질, 안전장비 착용 등이 분석 대상으로 사용되었고 남녀공학 여부, 설립 형태, 교원 1인당 학생 수 등과 같은 학교관련 변수도 분석 대상으로 사용되었다는 것을 알 수 있다. 연령보다는 월령을 릿지, 라쏘 그리고 엘라스틱넷 모든 모형에서 분석 대상으로 사용하였다. 이는 초등학생의 경우 성장이 완전하게 이루어지지 않고 월령별로 성장속도가 다르기 때문에 연령보다 월령이 정확한 분석이 가능하다는 것으로 해석할 수 있다.

〈표 4〉 초등학생 비만 예측모형의 적합도 지표

	모형	λ	Non-zero 변수의 수	표본 내 R^2	표본 외 R^2	CV 예측오차평균
초등학생 비만 모형	Ridge	0.01397	54	0.2191	0.2188	0.082512
	Lasso	0.00017	44	0.2191	0.2188	0.082507
	ElasticNet	0.00034	43	0.2191	0.2188	0.082507

〈표 5〉 초등학생 비만 예측모형의 추정 결과

	선택변수	Ridge	Lasso	ElasticNet
		추정치	추정치	추정치
1	월령	−0.0196	−0.0254	−0.0254
2	성별	−0.0210	−0.0212	−0.0212
3	태어난 해	0.0013	0.0015	0.0015
4	라면	−0.0012	−0.0009	−0.0009
5	음료수	0.0016	0.0014	0.0014
6	패스트푸드	0.0013	0.0011	0.0011
7	육류	0.0010	0.0009	0.0009
8	우유/유제품	0.0004	0.0001	0.0001
9	과일	−0.0082	−0.0081	−0.0081
10	채소	−0.0031	−0.0029	−0.0029
11	아침식사	0.0016	0.0014	0.0014
12	다이어트 경험; 아무것도 안 함	−0.0165	−0.01575	−0.0157
13	다이어트 경험; 식단조절	0.0293	0.0295	0.0295
14	다이어트 경험; 약	0.0056	0.0056	0.0056
15	다이어트 경험; 운동	0.0186	0.0187	0.0187
16	하루 수면량	−0.0075	−0.0075	−0.0075
17	자아신체상	0.1200	0.1217	0.1217
18	손 씻기	−0.0014	−0.0012	−0.0012
19	양치질	−0.0027	−0.0026	−0.0026
20	안전벨트 착용	0.0038	0.0037	0.0037
21	안전장비 착용	0.0028	0.0027	0.0027
22	주 3회 이상 운동	−0.0007	−0.0005	−0.0005
23	하루 TV 2시간 이상 시청	0.0059	0.0058	0.0058
24	하루 게임 2시간 이상	0.0064	0.0063	0.0063
25	괴롭힘/따돌림	0.0006	0.0004	0.0004
26	현금갈취	0.0029	0.0027	0.0027
27	신체접촉	−0.0002	−5.1100	−5.1700
28	가출 생각	−0.0012	−0.0011	−0.0011
29	가족지지	−0.0005	−0.0003	−0.0003
30	체벌 경험	−0.0017	−0.0016	−0.0016
31	가족 흡연	0.0051	0.0049	0.0049
32	무기력감	−0.0011	−0.0010	−0.0010
33	고민상담 희망	−0.0003	−0.0003	−0.0001

34	국립학교 여부	−0.0004	−0.0001	−0.0001
35	사립학교 여부	0.0003	0.0005	0.0005
36	도서벽지 여부	0.0009	0.0011	0.0011
37	시지역 여부	−0.0036	−0.0007	−0.0007
38	읍지역 여부	−0.0013	0.0007	0.0007
39	특별/광역시 여부	−0.0036	−0.0009	−0.0009
40	학급당 학생 수	−0.0044	−0.0046	−0.0046
41	남자 교원 수	0.0041	0.0040	0.0039
42	여자 교원 수	−0.0026	−0.0021	−0.0026
43	교원 1인당 학생 수	−0.0049	−0.0044	−0.0044
44	수업 태도 교정	−0.0001	−4.4900	
45	나이	−0.0057		
46	도시규모	0.0005		
47	상담요청	−0.0001		
48	가족 음주	0.0000		
49	공립학교 여부	−0.0002		
50	면지역 여부	−0.0021		
51	중소도시 여부	−0.0017		
52	학교설립 형태	0.0002		
53	학생 수 합계	0.0005		
54	교원 수 합계	−0.0008		

반면 라쏘 회귀분석 모형과 엘라스틱넷 분석 모형에서는 가족 음주 때문에 고민인지 여부, 상담요청, 소속 학교의 학생 수 합계 변수는 분석 대상으로 선택하지 않았다.

구체적인 추정치 값을 살펴보면 라쏘의 경우 자아신체상과 식단조절을 하였던 경험이 비만에 양(+)의 영향을 미친 것으로 나타났다. 월령, 라면을 먹는 식습관은 음(−)의 영향을 미치는 것으로 나타났다.

(3) 최소자승법모형, 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 예측력 평가

계량경제학에서는 주어진 모형의 예측력을 평가하기 위해 다양한 지표들이 개발되었다. 특히 예측값과 실제값의 차이로 정의되는 예측오차의 크기를 측정하기 위해서도 다양한 수단이 개발되었지만 본 연구에서는 일반적으로 사용되고 있는

〈표 6〉 초등학생 비만 예측모형의 예측력 평가 결과

모형	sample	MSE	R^2	관측치
OLS	Training Data	0.1036659	0.0192	339,098
	Testing Data	0.1032621	0.0204	113,032
Ridge	Training Data	0.0824807	0.2191	332,952
	Testing Data	0.0820078	0.2209	110,968
Lasso	Training Data	0.0824769	0.2191	333,060
	Testing Data	0.082006	0.2210	111,013
ElasticNet	Training Data	0.0825233	0.2194	333,690
	Testing Data	0.0820656	0.2213	111,253

MSE 지표를 이용하여 OLS, 릿지, 라쏘 그리고 엘라스틱넷의 모형 간의 예측력을 평가하였다(이재득, 2021).

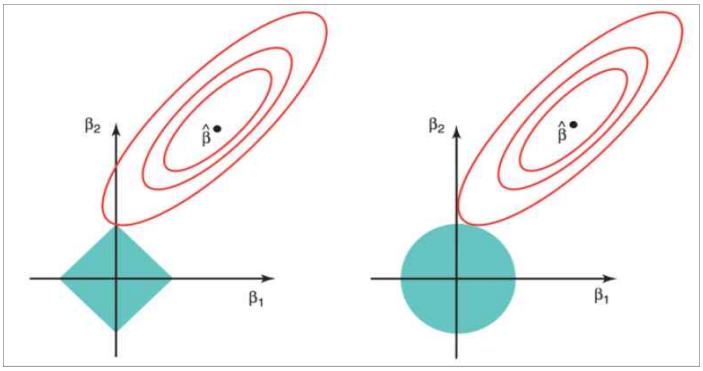
예측력을 평가하기 위해 테스팅 데이터는 전체 자료에서 모형 개발에 이용한 75%의 트레이닝 데이터 이외에 남은 25%를 이용하였다.

〈표 6〉은 트레이닝 데이터와 테스팅 데이터를 사용한 예측모형들의 예측력 평가 결과를 나타낸다. 〈표 6〉에서 볼 수 있듯이, 초등학생의 비만 예측의 경우에는 트레이닝 데이터와 테스팅 데이터의 MSE가 가장 적은 라쏘 회귀분석 모형이 적합하다는 것을 알 수 있다.

OLS, OLS, 릿지, 라쏘 그리고 엘라스틱넷 모형을 비교를 하자면, OLS 모델에서 식별하는 변수는 11개, 머신러닝 모형 중 릿지 모형에서 식별하는 변수는 총 54개, 라쏘는 44개, 엘라스틱넷은 43개이다. 머신러닝을 기반으로 한 모형은 OLS보다 더 많은 변수들을 가지고 MSE를 계산하기 때문에 가지고 있는 결측치가 많아 관측치가 OLS에 비해 적을 수밖에 없다.

특히 라쏘 모형에서는 관측치 감소가 더 두드러질 수밖에 없는데, 〈그림 1〉의 왼쪽 그래프에서 나타나 있듯, 라쏘는 선형제약(L1제약) 하에서 $L(\beta) = \min \sum_{i=1} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{p} |\beta_j|$ 식에서 $\lambda \sum_{i=1}^{p} |\beta_j|$ 에서 살아남을 β_j를 선정하기 때문에 코너해를 가질 경우 β_2만을 살아남고, β_1이 사라진다. 〈그림 1〉의 오른쪽 그래프에서 볼 수 있는 것처럼 릿지 모형에서는 원형제약(L2제약) 하에서 β_j를 선정하기 때문에 코너해를 갖는 경우가 없기 때문에 모든 β값들이 살아남게 된다.

〈그림 1〉 라쏘와 릿지 모형의 해[6)7)]



따라서 릿지 회귀분석 모형에서는 이론적으로 다중공선성에 의해 생략되는 일부 경우를 제외하고 모든 변수 54개가 분석 대상이 되어 본 연구의 결과처럼 관측치 수가 OLS에 비해 적은 경우가 생길 수 있다. 라쏘는 일부 β는 분석 대상이 되고 일부는 사라지기 때문에 상대적으로 릿지에 비해 많이 나올 수 있다.

2) 중·고등학생의 비만에 대한 예측 분석

본 절에서는 앞서 분석한 초등학생의 비만에 대한 분석 모형과 동일하게 먼저 비만인지 여부를 종속변수로 선택한 후, 설명변수들인 식습관, 수면습관 그리고 통제변수 등으로 이루어진 총 11개 변수로 최소자승법(OLS)을 분석하고 분석 대상인 모든 변수로 머신러닝 기법인 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 분석으로 추정하여 비교한다.

(1) 최소자승법(OLS) 추정

선행연구를 바탕으로 식 (5)와 같이 중·고등학생의 비만에 영향을 미치는 변수를 선정하고 트레이닝 데이터 총 471,188명을 대상으로 분석하였다. OLS 분석 결과 관측치는 461,292명이었다.

〈표 7〉은 OLS로 추정한 결과이다. 중·고등학생의 라면, 음료수, 패스트푸드,

6) Hui Zou and Trevor Hastie(2005), p. 304.
7) Lasso 그리고 고차원 문제와 오버피팅, 「이미지」, 2023. 1. 27., http://freesearch.pe.kr/archives/4473

〈표 7〉 중·고등학생 비만 OLS 분석 결과

		종속변수
		비만
식습관(Eating)	라면	−0.0192***(0.000953)
	음료수	−0.00621***(0.000733)
	패스트푸드	−0.0104***(0.000955)
	육류	−0.00976***(0.000735)
	우유/유제품	0.00669***(0.000554)
	아침식사 습관	−0.00908***(0.000474)
수면습관(Sleep)	수면시간	−0.00380***(0.000624)
통제변수(Z)	거주 도시	0.000358***(3.91e−05)
	조사연도	0.00733***(0.000165)
	나이	0.00358***(0.000325)
	성별(남=0, 여=1)	−0.0323***(0.00103)
	Constant	−14.54***(0.331)
	Observations	461,292
	R-squared	0.009

주: *** p<0.01, ** p<0.05, * p<0.1.

육류의 섭취는 음의 영향을 미치는 것으로 나타났다. 또한 아침식사는 '거의 먹을수록' 비만일 가능성은 낮아지고 수면시간이 많을수록 비만일 가능성이 낮아지는 것을 알 수 있다. 반면 우유/유제품 섭취는 비만에 양의 영향을 미치는 것으로 나타났다. 인적 특성을 보면 나이가 많을수록 그리고 여자보다는 남자가 비만일 가능성이 높은 것으로 나타났다. 종합해 보면, 우리나라 중·고등학생은 수면시간이 많을수록, 아침식사를 자주 할수록 비만일 가능성이 낮아짐을 알 수 있다.

(2) 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 추정

〈표 8〉은 가공이 완료된 전체 자료 관측치로 머신러닝 방법인 릿지 회귀모형, 라쏘 회귀모형 그리고 엘라스틱넷 회귀모형으로 중·고등학생의 비만을 예측한 결과이다. 릿지 회귀모형의 경우 607,449개 가운데 무작위로 선정된 75%에 해당하는 445,569개의 관측치로 구성된 트레이닝 데이터를 이용하였고, 라쏘와 엘라스틱넷 회귀모형의 경우 456,374개의 관측치로 구성된 트레이닝 데이터를 이용하여 분석하였다.

〈표 8〉에서 볼 수 있듯 머신러닝 기법 3개의 모형의 조율 변수 λ 를 포함한

〈표 8〉 중·고등학생 비만 예측모형의 적합도 지표

	모형	λ	Non-zero 변수의 수	표본 내 R^2	표본 외 R^2	CV 예측오차평균
중·고등학생 비만 모형	Ridge	0.01420	53	0.1965	0.1963	0.093058
	Lasso	0.00005	47	0.1965	0.1963	0.093053
	ElasticNet	0.00010	47	0.1965	0.1963	0.093053

모형의 전반적인 적합도 지표들을 알 수 있었다. 트레이닝 데이터를 분석하여 얻은 CV-예측오차의 관점에서 분석한 결과 라쏘 회귀모형과 엘라스틱넷 회귀모형이 최선의 모형으로 선택될 수 있는 것을 확인할 수 있다.

〈표 9〉에서는 머신러닝을 통해 비만 예측모형의 추정 결과를 살펴볼 수 있다. 중·고등학생 비만을 예측하기 위한 모형에서 선행연구를 바탕으로 구축한 모형 OLS에서 분석 대상으로 삼았던 변수들인 식습관, 수면습관, 거주 도시, 나이 그리고 성별 변수를 머신러닝 모형에서도 채택하였다. 추가적으로 머신러닝 모형에서는 다이어트 경험, 양치질, 안전장비 착용 등이 분석 대상으로 사용되었음을 알 수 있다. Brixval *et al.*(2012)의 선행연구에서 언급한 따돌림이나 괴롭힘을 경험하였는지 여부도 비만 예측모형에서 분석 대상으로 사용하였다. 이를 통해 정신적인 스트레스나 신체적 스트레스 역시 비만에 영향을 미칠 수 있는 요인으로 해석할 수 있다. 또한 남녀공학 여부, 설립 형태, 교원 1인당 학생 수 등과 같은 학교 관련 변수도 분석 대상으로 사용되었다는 것을 알 수 있다.

라쏘와 엘라스틱넷 회귀모형에서는 손 씻기, 하루 30분 이상 운동 여부, 성문제 상담희망 여부, 여자 교원 수는 분석 대상으로 사용하지 않은 것으로 나타났다.

〈표 9〉 중·고등학생의 비만 예측모형의 추정 결과

	선택변수	Ridge	Lasso	ElasticNet
		추정치	추정치	추정치
1	연령	-0.0029	-0.0039	-0.0039
2	월령	0.0060	0.0070	0.0069
3	도시규모	0.0026	0.0026	0.0026
4	성별	-0.0376	-0.0376	-0.0376
5	태어난 해	0.0028	0.0028	0.0026

6	라면	−0.0024	−0.0024	−0.0024
7	음료수	−0.0018	−0.0018	−0.0018
8	패스트푸드	−0.0045	−0.0021	−0.0045
9	육류	0.0009	−0.0045	−0.0045
10	우유/유제품	−0.0026	0.0009	0.0009
11	과일	−0.0027	−0.0026	−0.0021
12	채소	0.0057	0.0057	0.0057
13	아침식사	−0.0023	−0.0022	−0.0022
14	다이어트 경험; 아무것도 안 함	−0.0053	−0.0049	−0.0049
15	다이어트 경험; 식단조절	0.0112	0.0111	0.0111
16	다이어트 경험; 약	0.0132	0.0133	0.0133
17	다이어트 경험; 운동	0.0012	0.0009	0.0009
18	하루 수면량	0.0006	0.0005	0.0005
19	자아신체상	0.1384	0.1407	0.1407
20	양치질	−0.0116	−0.0117	−0.0117
21	안전벨트 착용	0.0047	0.0046	0.0046
22	안전장비 착용	0.0079	0.0080	0.0046
23	1년 동안 치료 경험	0.0014	0.0013	0.0013
24	하루 게임 2시간 이상	0.0080	0.0079	0.0079
25	음란물/채팅	0.0024	0.0023	0.0023
26	괴롭힘/따돌림	0.0041	0.0042	0.0042
27	고민상담 대상 여부	−0.0038	−0.0038	−0.0038
28	가정 문제 걱정	−0.0004	−0.0003	−0.0003
29	가출 생각	−0.0012	−0.0012	−0.0012
30	고민상담 희망	0.0006	0.0006	0.0006
31	공립학교 여부	−0.0003	−0.0009	−0.0009
32	국립학교 여부	0.0003	0.0001	0.0001
33	도서벽지 여부	0.0009	0.0010	0.0010
34	면지역 여부	−0.0015	−0.0010	−0.0010
35	시지역 여부	−0.0018	−0.0100	−0.0100
36	읍지역 여부	0.0004	0.0007	0.0007
37	중소도시 여부	0.0019	0.0022	0.0022
38	특별/광역시 여부	−0.0015	−0.0009	−0.0009
39	남녀공학 여부	−0.0059	−0.0059	−0.0059
40	학생 수 합계	−0.0085	−0.0100	−0.0099
41	학급당 학생 수	−0.0120	−0.0123	−0.0123

42	교원 수 합계	0.0048	0.0088	0.0088
43	남자 교원 수	0.0045	0.0025	0.0025
44	교원 1인당 학생 수	0.0012	0.0022	0.0022
45	학교설립 형태	0.0003		
46	손 씻기	0.0000		
47	하루 30분 이상 운동	0.0000		
48	성 문제 상담희망	0.0000		
49	사립학교 여부	0.0003		
50	여자 교원 수	0.0022		
51	폭력위협	0.0012		
52	상담희망	0.0009		
53	흡연/음주 상담희망	0.0008		
54	가족지지		0.0011	0.0011
55	체벌 경험		0.0009	0.0009
56	가족 흡연		0.0008	0.0008

(3) 최소자승법모형, 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 예측력 평가

중·고등학생의 비만 예측모형의 예측력을 평가하기 위해 〈표 10〉과 같이 테스팅 데이터의 MSE를 살펴본 결과, 라쏘 회귀모형의 MSE가 가장 작고 R^2의 값도 높기 때문에 중·고 전체 학생들의 비만에 영향을 미치는 변수를 분석하기 위해서는 라쏘 회귀모형이 예측에 적합하다는 것을 확인할 수 있다.

〈표 10〉 중·고등학생의 비만 예측모형의 예측력 평가 결과

모형	sample	MSE	R^2	관측치
OLS	Training Data	0.1122294	0.0302	471,188
	Testing Data	0.1121675	0.0297	157,062
Ridge	Training Data	0.0930359	0.1965	455,569
	Testing Data	0.0928247	0.1995	151,880
Lasso	Training Data	0.0930691	0.1964	456,374
	Testing Data	0.0927017	0.1994	152,115
ElasticNet	Training Data	0.0930691	0.1964	456,374
	Testing Data	0.0927018	0.1994	152,115

3) 초·중·고등학생의 비만에 대한 예측 분석

앞에서 초등학생과 중·고등학생으로 나누어 분석하였다면, 본 절은 모든 학생에게 공통적으로 해당하는 설문과 변수를 분석 대상으로 하여 분석하였다. OLS와 머신러닝 기법으로 분석하고 이를 비교한 결과를 살펴보고자 한다.

(1) 최소자승법(OLS) 추정

선행연구를 바탕으로 학생의 비만에 영향을 미치는 변수를 선정하였고 트레이닝 데이터에 속하는 810,285명을 대상으로 식 (5)에 따라 분석한 결과는 〈표 11〉과 같다. OLS 분석 결과 관측치는 796,780명이다. 초·중·고등학생의 라면, 음료수, 육류의 섭취, 아침식사 습관은 비만에 1%의 유의수준에서 음의 영향을 미치는 것으로 나타났다. 하지만 우유 및 유제품의 섭취는 비만에 유의한 수준에서 비만에 양의 영향을 미치는 것으로 나타났다. 수면시간과 비만에는 음의 관계가 있다는 것을 알 수 있다. 종합해 보면 식습관에서는 우유와 같은 유제품을 많이 먹으면 비만일 가능성이 높지만 아침식사를 자주하게 되면 비만일 가능성은 낮

〈표 11〉 초·중·고등학생 비만 OLS 분석 결과

		종속변수
		비만
식습관(Eating)	라면	−0.0109*** (0.000739)
	음료수	−0.00177*** (0.000574)
	패스트푸드	−0.000531 (0.000723)
	육류	−0.0128*** (0.000570)
	우유/유제품	0.00481*** (0.000421)
	아침식사 습관	−0.00764*** (0.000395)
수면습관(Sleep)	수면시간	−0.0105*** (0.000470)
통제변수(Z)	거주 도시	0.000391*** (2.92e−05)
	조사연도	0.00595*** (0.000123)
	나이	0.000803*** (0.000149)
	성별(남=0, 여=1)	−0.0345*** (0.000757)
	Constant	−11.73*** (0.248)
	Observations	796,780
	R-squared	0.008

주: *** p<0.01, ** p<0.05, * p<0.1.

아진다는 점이 일관되게 나타났다. 또한 충분한 수면시간이 비만의 가능성을 낮춘다고 볼 수 있다. 이 분석 결과는 각급 학교별로 분석한 결과와 일치한다.

(2) 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 추정

〈표 12〉는 STATA 17 프로그램을 통해 전처리가 작업이 완료된 전체 자료 관측치로 머신러닝 방법인 릿지, 라쏘 그리고 엘라스틱넷으로 비만인지 여부를 예측하였다. OLS의 경우 1,080,380개의 관측치 중 무작위로 선정된 75%에 해당하는 810,285개의 관측치로 구성된 트레이닝 데이터를 이용하였고 릿지, 라쏘 그리고 엘라스틱넷 회귀모형의 경우 793,863개의 관측치로 구성된 트레이닝 데이터를 사용하여 분석하였다.

분석 결과, 조율 변수 λ, 0이 아닌 변수의 개수 그리고 표본 내·외의 R^2를 포함한 모형의 전반적인 적합도 지표들을 보여 주고 있다. 트레이닝 데이터의 경우, CV-예측오차의 관점에서 라쏘와 엘라스틱넷 모형이 최선의 모형으로 선택될 수 있음을 알 수 있다.

〈표 13〉은 초·중·고등학생 비만 예측모형의 추정 결과를 나타낸다. 분석 결과, 앞서 선행연구를 바탕으로 구축한 모형 OLS에서 분석 대상으로 삼았던 변수들인 식습관, 수면습관, 거주 도시, 나이, 성별들은 머신러닝 모형에서도 채택한 것을 볼 수 있다.

또한 머신러닝 모형에서는 추가적으로 다이어트 경험, 손 씻기, 양치질, 안전벨트 등이 분석 대상으로 사용되었고, 소속 학교 관련 변수로는 남녀공학 여부, 설립 형태, 교원 1인당 학생 수 등이 분석 대상으로 사용되었다. 그러나 라쏘와 엘라스틱넷 모형의 경우 학교 특성과 관련된 변수 중 교원 수 합계 변수는 분석 대상으로 사용하지 않았다.

〈표 12〉 초·중·고등학생 비만 예측모형의 적합도 지표

	모형	λ	Non-zero 변수의 수	표본 내 R^2	표본 외 R^2	CV 예측오차평균
초·중·고등학생 비만 모형	Ridge	0.01410	42	0.2003	0.2002	0.089009
	Lasso	0.00009	38	0.2003	0.2003	0.089004
	ElasticNet	0.00016	38	0.2003	0.2003	0.089004

〈표 13〉 초·중·고등학생 비만 예측모형의 추정 결과

	선택변수	Ridge	Lasso	ElasticNet
		추정치	추정치	추정치
1	연령	−0.0122	−0.0076	−0.0077
2	월령	−0.0144	−0.0204	−0.0204
3	도시규모	0.0008	0.0005	0.0005
4	성별	−0.0320	−0.0325	−0.0325
5	태어난 해	0.0018	0.0018	0.0018
6	라면	−0.0010	−0.0008	−0.0008
7	음료수	0.0003	0.0002	0.0002
8	패스트푸드	−0.0006	−0.0005	−0.0005
9	육류	−0.0046	−0.0046	−0.0046
10	우유/유제품	0.0017	0.0015	0.0015
11	과일	−0.0057	−0.0056	−0.0056
12	아침식사	−0.0004	−0.0003	−0.0003
13	다이어트 경험; 아무것도 안 함	−0.0156	−0.0157	−0.0157
14	다이어트 경험; 식단조절	0.0138	0.0135	0.0135
15	다이어트 경험; 약	0.0103	0.0103	0.0103
16	다이어트 경험; 운동	0.0036	0.0031	0.0031
17	하루 수면량	−0.0033	−0.0033	−0.0034
18	자아신체상	0.1320	0.1341	0.1341
19	손 씻기	−0.0003	−0.0001	−0.0001
20	양치질	−0.0087	−0.0088	−0.0088
21	안전벨트 착용	0.0039	0.0038	0.0038
22	안전장비 착용	0.0047	0.0046	0.0046
23	하루 게임 2시간 이상	0.0055	0.0055	0.0055
24	사립학교 여부	−0.0009	−0.0016	−0.0016
25	도서벽지 여부	0.0008	0.0009	0.0009
26	면지역 여부	−0.0019	−0.0007	−0.0007
27	읍지역 여부	0.0003	0.0015	0.0015
28	중소도시 여부	−0.0020	−0.0008	−0.0008
29	특별/광역시 여부	−0.0027	−0.0011	−0.0011
30	학교설립 형태	−0.0009	−0.0009	−0.0010
31	남녀공학 여부	−0.0048	−0.0047	−0.0048
32	학생 수 합계	−0.0037	−0.0028	−0.0030
33	학급당 학생 수	−0.0096	−0.0100	−0.0101

34	남자 교원 수	0.0067	0.0071	0.0072
35	여자 교원 수	−0.0033	−0.0024	−0.0023
36	교원 1인당 학생 수	0.0034	0.0032	0.0034
37	조사연도	0.0113	0.0114	0.0114
38	학교급	0.0106	0.0123	0.0124
39	시지역 여부	0.0018		
40	교원 수 합계	0.0018		
41	공립학교 여부	0.0054		
42	국립학교 여부	0.0001		

(3) 최소자승법(OLS)모형, 릿지 회귀모형, 라쏘 회귀모형과 엘라스틱넷 회귀모형 예측력 평가

초·중·고등학생의 비만 예측모형의 예측력을 평가하기 위해 테스팅 데이터의 MSE를 살펴본 결과, 〈표 14〉에서 나타나 있듯이 엘라스틱넷의 MSE가 가장 작고 R^2의 값도 높기 때문에 초·중·고 전체 학생들의 비만에 영향을 미치는 변수를 분석하기 위해서는 엘라스틱넷 회귀분석 모형의 예측력이 좋은 모형으로 나타났다.

〈표 14〉 초·중·고등학생의 비만 예측모형의 예측력 평가 결과

모형	sample	MSE	R^2	관측치
OLS	Training Data	0.1085593	0.0247	810,285
	Testing Data	0.1093955	0.0247	270,095
Ridge	Training Data	0.089000	0.2003	793,863
	Testing Data	0.0897476	0.1999	264,666
Lasso	Training Data	0.0889961	0.2003	793,863
	Testing Data	0.0897435	0.2000	264,666
ElasticNet	Training Data	0.0889959	0.2003	793,863
	Testing Data	0.0897433	0.2000	264,666

Ⅴ. 결론

현재 우리나라뿐 아니라 전 세계적으로 비만은 주요한 건강 문제로 주목받고 있다. 더욱이 성장기에 있는 학생들의 비만은 주요한 현안이고 학생들을 건강하게 만들기 위해 많은 노력을 하고 있다.

이러한 노력의 일환인 학생들의 건강 증진 정책의 효율적인 지원과 육성을 위해서는 먼저 학생들의 비만에 영향을 미치는 요인에 대한 정확한 예측과 추정이 필요하다. 그러나 전통적 선형 회귀분석 모형은 학생들의 다양한 특성에 따라 미칠 수 있는 다양한 변수로 분석을 하다 보면 과잉적합 문제가 발생하여 비만에 관한 정책을 덜 정확하게 제안할 수 있다.

그리하여 본 연구는 초등학생, 중학생과 고등학생의 비만에 미치는 요인을 실증적으로 분석할 때, 최소자승법(OLS) 모형에 의한 추정뿐 아니라 과잉적합 문제를 극복하고 면밀한 예측을 위해 규제항을 도입한 머신러닝 기법인 릿지, 라쏘 그리고 엘라스틱넷을 통해 예측하여 비교하였다.

초·중·고등학생을 대상으로 한 신체 계측 자료와 건강 관련 설문 자료를 이용하여 비만에 영향을 주는 외생적인 변수로 정의하여 분석하였고 그 결과 종합해서 요약해 보면 다음과 같다.

첫째, 초등학생의 비만을 예측한 결과를 살펴보면 OLS 모형의 경우 식습관 중 라면, 음료수, 패스트푸드, 우유 및 유제품의 1주일 동안 섭취하는 횟수와 비만은 유의하게 양의 영향을 미치는 것으로 나타났다. 머신러닝 3가지 모형 중에서는 라쏘 회귀모형이 예측력이 좋은 모형으로 나타났으며 라쏘 모형에서는 OLS 모형에서 유의하게 나타난 변수 외에 식습관 관련된 변수 중에서 과일, 채소 섭취 횟수, 생활습관과 관련된 손 씻기 횟수, 양치질 횟수, 안전벨트 착용, 운동 횟수, 일주일에 TV 2시간 이상 시청하는 횟수 등이 영향을 미치는 것으로 선택되었다. 또한 학교와 관련된 변수로 학교의 소재지, 사립학교 여부, 교원 1인당 학생 수가 영향을 미치는 것으로 선택되었다. 그 밖에 다이어트 경험 여부와 따돌림 경험 여부 등이 선택되었다.

둘째, 중·고등학생의 비만을 예측한 결과를 살펴보면 OLS 모형의 경우 식습관 중 라면, 음료수, 패스트푸드를 1주일 동안 섭취하는 횟수와 아침식사를 하는 횟수는 비만과 유의하게 음의 영향을 미치는 것으로 나타났다. 머신러닝 3가지 모형 중에서는 라쏘 회귀모형이 예측력이 좋은 모형으로 나타났으며 라쏘 모형

에서는 OLS 모형에서 유의하게 나타난 변수 외에 식습관 관련된 변수 중에서 과일, 채소 섭취 횟수, 생활습관과 관련된 변수로는 손 씻기 횟수, 양치질 횟수, 안전벨트 착용, 운동 횟수, 일주일에 게임 2시간 이상 시청하는 횟수와 음란물 채팅 횟수 등이 영향을 미치는 것으로 선택되었다. 또한 학교와 관련된 변수로 학교의 소재지, 사립학교 여부, 교원 1인당 학생 수가 영향을 미치는 것으로 선택되었다. 그리고 다이어트 경험 여부, 따돌림 경험 여부와 가정 문제에 대한 걱정 여부가 비만에 영향을 미치는 것으로 선택되었다.

셋째, 초·중·고등학생의 비만을 예측한 결과를 살펴보면 OLS 모형의 경우 식습관 중 라면, 음료수를 1주일 동안 섭취하는 횟수와 아침식사를 하는 횟수는 비만에는 유의하게 음의 영향을 미치는 것으로 나타났다. 우유 및 유제품의 섭취 횟수는 비만과 양의 관계에 있는 것으로 나타났다. 머신러닝 3가지 모형 중에서는 엘라스틱넷 회귀모형이 예측력이 좋은 모형으로 나타났으며 라쏘 모형에서는 OLS 모형에서 유의하게 나타난 변수 외에 식습관 관련된 변수 중에서 과일 섭취 횟수, 생활습관과 관련된 변수로는 손 씻기 횟수, 양치질 횟수, 안전벨트 착용, 운동 횟수, 일주일에 게임 2시간 이상 시청하는 횟수가 영향을 미치는 것으로 선택되었다. 또한 학교와 관련된 변수로 학교의 소재지, 사립학교 여부, 남녀공학 여부, 교원 1인당 학생 수가 영향을 미치는 것으로 선택되었다. 그리고 다이어트 경험 여부, 자아신체상이 비만에 영향을 미치는 것으로 선택되었다.

결론을 종합해 보면 선행연구에서는 라면, 패스트푸드, 탄산음료와 같은 간식 횟수가 비만에 양의 영향을 미친다고 밝혔는데, 라쏘 모형 분석 결과 음료수나 패스트푸드의 경우에는 선행연구 결과와 일치하게 양(+)의 효과로 나타났으나 라면을 먹는 습관의 경우에는 음(−)의 영향으로 나타났다. 그리고 자신이 뚱뚱하다고 생각할수록 비만에 양(+)의 영향을 미치는 결과 역시 선행연구의 결과와 일치한다.

생활습관을 비교해 보자면 게임 횟수와 TV 시청 횟수는 비만에 양(+)의 영향을 미치는 것으로 나타나 선행연구의 결과와 일치하였다. 한편, 안전장비 착용의 경우에는 착용을 덜 하는 학생일수록 비만일 가능성이 높다는 선행연구 결과와 달리 양(+)의 영향을 미치는 것으로 나타났다.

지역변수의 경우 이규영(2008)은 대도시, 중소도시, 읍·면 지역 중 중소도시에 과체중 학생이 비만일 확률이 높다고 하였는데, 초등학생 모형, 초·중·고등학생 모형에서는 일치하는 결과를 볼 수는 없지만 중·고등학생의 경우에는 특별/광역시는 비만에 음(−)의 영향, 중소도시는 비만에 양(+)의 영향을 미쳐 선행연

구와 일치하였다.

본 연구 결과의 정책적 의미를 살펴보면 다음과 같다. 먼저 초등학생과 중·고등학생의 비만을 더욱 정확하게 예측하기 위해서는 기존의 선형회귀모형보다는 과잉적합과 모형설정의 오류를 줄이는 라쏘 회귀모형으로 추정할 필요가 있다. 만약 학교의 급을 구분 없이 초·중·고등학생 모두를 대상으로 비만을 예측하기 위해서는 엘라스틱넷 회귀모형으로 추정할 필요가 있다.

그러므로 학생들의 비만에 미치는 요인을 정확하게 예측하고 분석하여 학생들의 건강한 식습관뿐 아니라 생활습관, 정서적인 케어를 지원하는 정책을 통해 각급 학교 학생들에게 맞는 건강 증진 정책을 활성화할 수 있도록 해야 할 것이다.

그러나 연구의 완성도를 높이기 위해 향후 더욱 긴 시간 동안 학생들에 대한 자료들을 수집하고 현재는 관찰할 수 없지만 학생의 비만에 영향을 미치는 데이터의 수집이 가능하다면 더욱 엄밀한 분석이 될 수 있겠지만, 이는 추후 과제로 남긴다.

부 록

〈부표 1〉 변수 설명

변수명		변수 설명		분석 모형
종속변수	비만 여부	0. 비만 아님	1. 비만	초등, 중·고등
인적 특성	연령	만 나이(6세 미만 18세 초과 결측)		
	월령	개월 수		
	성별	1. 남	2. 여	
식습관	라면	일주일 동안 OO을 대체로 몇 번 먹습니까?		
	음료수			
	패스트푸드	1. 먹지 않음	2. 1~2번	
	육류	3. 3~5번	4. 매일	
	우유/유제품			
	과일			
	채소			
	아침식사	아침식사는 어떻게 합니까?		
		1. 거의 안 먹음	2. 대체로 안 먹음	
		3. 대체로 먹음	4. 거의 먹음	
다이어트 경험	아무것도 안 함	0. No	1. Yes	
	식단조절			
	약을 먹음			
	운동으로 감량			
자아신체상	자아신체상	친구들과 비교해서 자신의 체형이 어떻다고 생각하십니까?		
		1. 매우 마른 편	2. 약간 마른 편	
		3. 보통	4. 약간 살이 찐 편	
		5. 매우 살이 찐 편		
생활습관	하루 수면량	1. 6시간 이내	2. 6~7시간	
		3. 7~8시간	4. 8시간 이상	
	손 씻기	밥을 먹기 전이나 밖에서 놀다 돌아와서 비누로 손을 씻는다.		
		0. No	1. Yes	
	양치질	하루에 두 번 이상 이를 닦는다.		
		0. No	1. Yes	
	안전벨트 착용	자동차를 탈 때 안전벨트를 맨다.		
		0. No	1. Yes	
	안전장비 착용	인라인스케이트, 롤러블레이드, 스케이트보드, 자전거 등을 탈 때 헬멧을 쓰고 보호대를 착용한다.		
		0. No	1. Yes	
	주 3회 이상 운동	일주일에 세 번 이상 숨이 차거나 땀이 날 정도로 운동을 합니까?		초등
		0. No	1. Yes	
	TV 시청	TV를 하루에 2시간 이상 본다.		
		0. No	1. Yes	

	게임	인터넷이나 게임을 하루에 2시간 이상 한다.		
		0. No	1. Yes	
정신적 요소	괴롭힘	지난 1년 동안 친구들에게 괴롭힘이나 따돌림을 당한 적이 있다.		
		0. No	1. Yes	
	현금갈취	돈을 빼앗는 친구가 있다.		
		0. No	1. Yes	
	신체접촉	내 몸을 자주 만지는 사람이 있다.		
		0. No	1. Yes	
	가출 생각	집을 나가고 싶을 때가 자주 있다.		
		0. No	1. Yes	
	가족 지지	우리 가족은 나의 이야기를 잘 들어주고 나의 감정을 존중해 준다.		
		0. No	1. Yes	
	체벌 경험	자주 매를 맞는 편이다.		
		0. No	1. Yes	
	상담요청	가정 및 학교생활 문제로 선생님의 상담이 필요하다.		
		0. No	1. Yes	
	가족 흡연	같이 사는 사람 중에 담배를 피우는 사람이 있다.		
		0. No	1. Yes	
	가족 음주	같이 사는 사람 중에 술을 너무 많이 마셔서 걱정되는 사람이 있다.		
		0. No	1. Yes	
	무기력감	모든 것이 귀찮고 희망이 없는 것처럼 느껴진다.		
		0. No	1. Yes	
	수업태도 교정	공부시간에 선생님께 자주 혼난다.		
		0. No	1. Yes	
	고민상담 희망	고민이나 괴로운 일로 상담을 받고 싶다.		
		0. No	1. Yes	
생활습관	운동	하루 30분~1시간 이상 숨이 차거나 땀이 날 정도의 운동을 일주일에 며칠이나 합니까?		중·고등
		1. 거의 안 함	2. 1~2일 정도	
		3. 3~4일 정도	4. 5일 이상	
	1년 동안 치료 경험	지난 1년 동안 사고나 외상 때문에 병원, 보건실에서 치료를 받은 적이 있습니까?		
		1. 없음	2. 1번	
		3. 2번	4. 3번 이상	
	게임	인터넷이나 게임을 하루에 2시간 이상 한다.		
		0. No	1. Yes	
	음란물/채팅	음란물을 보거나 성인사이트에서 채팅을 자주 한다.		
		0. No	1. Yes	
정신적 요소	괴롭힘	지난 1년 동안 친구들에게 괴롭힘이나 따돌림을 당한 적이 있다.		
		0. No	1. Yes	
	고민상담	고민이 있거나 괴로울 때 의논할 수 있는 사람이 있다.		
		0. No	1. Yes	
	가정 문제	가정(가족) 내의 문제에 대해 걱정이 된다.		
		0. No	1. Yes	

	가출 생각	지난 1년 동안 가출하는 것을 심각하게 생각해 본 적이 있다.		
		0. No	1. Yes	
	폭력위협	가정이나 학교에서 폭력으로 인해 자신의 안전이 위협을 받고 있다.		
		0. No	1. Yes	
	상담요청	학교 생활 문제로 전문가의 상담을 받고 싶다.		
		0. No	1. Yes	
	흡연/음주 상담희망	술이나 담배 문제로 전문가의 도움을 받고 싶다.		
		0. No	1. Yes	
	성 문제 상담희망	성 문제로 전문가의 상담을 받고 싶다.		
		0. No	1. Yes	
	상담희망	고민이나 괴로운 일로 상담을 받고 싶다.		
		0. No	1. Yes	
지역	도시	11. 서울	21. 부산	초등, 중·고등
		22. 대구	23. 인천	
		24. 광주	25. 대전	
		26. 울산	27. 세종	
		41. 경기	42. 강원	
		43. 충북	44. 충남	
		45. 전북	46. 전남	
		47. 경남	49. 제주	
	도서벽지 여부	0. No	1. Yes	
	면지역 여부	0. No	1. Yes	
	시 여부	0. No	1. Yes	
	읍지역 여부	0. No	1. Yes	
	중소도시 여부	0. No	1. Yes	
	특별/광역시 여부	0. No	1. Yes	
학교	설립 형태	1. 공립	2. 국립	
		3. 사립		
	남녀공학 여부	0. 남녀공학	1. 남녀공학 아님	
	학생 수 합계	학교 전체 학생 수, 단위: 명		
	학급당 학생 수	한 학급당 학생 수, 단위: 명		
	교원 수 합계	교원 수, 단위: 명		
	남자 교원 수	남 교원 수, 단위: 명		
	여자 교원 수	여 교원 수, 단위: 명		
	교원 1인당 학생 수	학생 수 / 교원 수, 단위: 명		

참 고 문 헌

건강보험공단, “비만의 사회경제적 영향 조사 연구 개요 외,” 2018(https://www.medifonews.com/news/article.html?no=142677, 2023. 1. 27. 인출).

교육부, “2015년 학교건강검사 표본조사 결과,” 2015(https://www.moe.go.kr/sn3hcv/doc.html?fn=88828c2819629b8293723c99137f1066&rs=/upload/synap/202301/, 2023. 1. 27. 인출).

김선일 외, “청소년의 비만도에 따른 칫솔질 실천과 손 씻기의 연관성 분석,”『한국학교·지역보건교육학회지』제19권 제2호, 2018, 65~76.

김은주 외, “한의 체중 조절 프로그램에 참여한 과체중, 비만 환자에서의 머신러닝 기법을 적용한 체중 감량 예측 연구,”『대한한의학회지』제41권 제2호, 2020, 58~79.

김재등 외, “초등학교 고학년 아동의 비만 정도에 따른 식습관, 운동습관 및 생활습관의 차이,”『한국사회체육학회지』제38권, 2009, 855~865.

김지환, “신용평가 분석 예제를 활용한 Lasso 벌점 통계적 학습 모형 비교,”『통계연구』제22권, 2022, 103~117.

노민정·유진은, “Adaptive LASSO를 통한 진로 결정 관련 변수 탐색,”『열린교육연구』제27권 제4호, 2019, 133~155.

성선화 외, “전주지역 중학생의 성별 및 비만판정에 따른 식행동 비교연구,”『한국식품영양과학회지』제36권 제8호, 2007, 995~1009.

송상윤, “예대금리차 결정요인 모형의 측력 비교 연구–Ridge, LASSO 및 ElasticNet 방법론을 중심으로,”『금융지식연구』제13권 제3호, 2015, 41~65.

이규영 외, “우리나라 중·고등학교 학생들의 패스트푸드 및 탄산음료 섭취에 관한 지역별 비교연구,”『한국학교보건학회지』제21권 제2호, 2008, 47~60.

이재득, “릿지 회귀와 라쏘 회귀모형에 의한 부산 전략산업의 지역경제 효과에 대한 머신러닝 예측,”『Journal of Korea Port Economic Association』제37권 제1호, 2021, 197~215.

이종화 외, “중·고생의 치아우식증과 구강건강행태와의 관련성 연구: 제9차(2013년) 청소년 건강행태 온라인 조사,”『한국치위생학회지』제15권 제1호, 2015, 119~127.

정애경·김계현·김동민, “진로상담: 진로미결정 및 관련변수에 관한 국내연구 메

타 분석,"『상담학연구』제9권 제2호, 2008, 551~564.
진혜정 외, "한국 청소년의 위생습관 중 잇솔질과 손 씻기의 연관성,"『한국치위생학회지』제13권 제1호, 2013, 82~88.
하영미 외, "고등학생의 수면부족, 비만, 스크린 타임 사이의 관련성 연구,"『Journal of Korean Biological Nursing Science』제16권 제2호, 2014, 80~89.
_____, "초등학교 비만 아동과 정상 체중 아동의 체중 조절 실태와 식습관에 관한 연구,"『The East Asian Society of Dietary Life』제16권 제3호, 2006, 272~280.
_____, *Stata Lasso Reference Manual Release 17*, A Stata Press Publication, 2021.
Aaron, Kreiner and John Duca, "Can Machine Learning on Economic Data better Forecast the Unemployment Rate?," *Applied Economics Letters*, 27, 2019, 1434~1437.
Brixval, Carina S., Signe L. B. Rayce, Mette Rasmussen, Bjørn E. Holstein, and Pernille Due, "Overweight, Body Image and Bullying—an Epidemiological Study of 11- to 15-years Olds," *European Journal of Public Health*, 22(1), 2012, 126~130.
Ferdowsy, Faria *et al.*, "A Machine Learning Approach for Obesity Risk Prediction," *Current Research in Behavioral Sciences*, 2, 100053, 2021.
Hui, Zou and Trevor Hastie, "Regularization and Variable Selection via the ElasticNet," *Journal of the Royal Statistical Society*, 67(2), 2005, 301~320.
Jose, Manuel Pereira, Mario Basto, and Amelia Ferreira da Silva, "The Logistic Lasso and Ridge Regression in Predicting Corporate Failure," *Procedia Economics and Finance*, 39, 2016, 634~641(https://doi.org/10.1016/S2212-5671(16)30310-0).
Kuźbicka, K. *et al.*, "Bad Eating Habits as the Main Cause of Obesity among Children," *Pediatric Endocrinology, Diabetes and Metabolism*, 19(3), 2013, 106~110.
Kil, E. *et al.*, "Analysis of the Relationship between Regional Economic Growth and Obesity by Using Lasso Regression," *In Proceedings of*

the Korea Information Processing Society Conference, 25(2), Korea Information Processing Society, 2018, 565~568.

Ofori, Isaac K., Camara K. Obeng, and Simplice A. Asongu(2022), "What Really Drives Economic Growth in Sub-Saharan Africa? Evidence from the Lasso Regularization and Inferential Techniques," *Journal of the Knowledge Economy*(https://doi.org/10.1007/s13132-022-01055-1).

Osmani, Teixeira C. Guillén, João Victor Issler, and Yihao Lin, "Machine Learning and Oil Price Point and Density Forecasting," *Energy Economics*, 2021(https://doi.org/10.1016/j.eneco.2021.105494).

Park, Sin-Young, "The Associated Factors with Subjective Oral Symptoms Experience in Obesity Adolescent," *Journal of Korean Society of Dental Hygiene*, 17(5), 2017, 757~767.

Price, J. H., J. A. Dake, J. E. Balls-Berry, and M. Wielinski, "Seat Belt Use among Overweight and Obese Adolescents," *Journal of community health*, 36, 2011, 612~615.

Tibshirani, R., "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 1996, 267~288.

[Abstract]

Prediction of Obesity in Students through Machine Learning Analysis

Minsook Lim* · Hyunjoo Jeong** · Jinhyung Lee***

The purpose of this study is to analyze variables that affect obesity through machine learning in elementary school. Obesity was estimated and predicted using OLS and machine learning models such as Ridge Regression, Lasso Regression, and ElasticNet Regression based on personal, school, regional characteristics, and health-related survey data of middle and high school students. In addition, the above models were compared and evaluated using MSE (Mean Squared Error).

The models for estimating obesity in elementary, middle, and high school students were analyzed by dividing them. As a result of the OLS analysis, eating habits and sleeping time were found to be significant. The machine learning analysis revealed that lifestyle, school, and regional characteristics were also important variables in predicting obesity, in addition to eating habits and sleeping hours.

This study is meaningful because it provides the basis for preparing customized obesity prevention policies for each school student using machine learning.

Keywords: Machine Learning, Ridge, Lasso, ElasticNet, MSE, obesity, students, prediction

JEL Classification: I, I0, I1, I2

* First Author, Sungkyunkwan University, Department of Economics, Doctoral Student, E-mail: minlim1984@gmail.com

** Coauthor, Korea Educational Environments Protection Agency, Research Fellow, Tel: +82-43-710-4019, E-mail: jeonghj@schoolkeepa.or.kr

*** Corresponding Author, Sungkyunkwan University, Department of Economics, Professor, Tel: +82-2-760-0263, E-mail: leejinh@gmail.com