논문 10

머신러닝 기법을 활용한 사교육비 지출 예측 모형 탐색

김 영 식[1] · 이 호 준[2] · 김 훈 호[3]

요 약

고등학교 유형이나 대학 입시의 큰 틀이 변화하면서 고등학교 학생들의 사교육 수요 또한 빠르게 변화하고 있을 것으로 예상되는 바, 고등학교 학생들에 초점을 맞추어 사교육 수요 현상을 새롭게 분석해 볼 필요가 있을 것으로 보인다. 이에 본 연구에서는 랜덤 포레스트 방법을 사용하여 사교육비 지출 예측 변수를 추출하고, 이를 활용하여 고정효과 및 확률효과 모형 분석을 실시함으로써 보다 적절한 사교육비 지출 예측 모형을 탐색하고자 하였다.

분석 결과에 따르면, '남성보호자 월평균 소득', '고등학교 계열', '고등학교 계열 선택 이유', '창의적 체험활동 만족도', '진학 관련 월평균 방과후학교 비용', '진로교육 활동 만족도', '여성보호자 최종학력' 등이 사교육비에 대한 예측력이 높은 것으로 나타났다. 랜덤 포레스트 분석 결과 도출된 사교육비 지출 관련 주요 변수들을 활용하여 적합한 고교생의 사교육비 지출 모형을 확인한 결과 OLS 및 고정효과 모형, 확률효과 모형 중 학교 확률효과 모형의 적합성이 가장 높았다.

증거 기반(evidence based) 의사 결정의 중요성이 정책 결정 과정에서 높아져 감에 따라 사교육비 지출 예측 모형에 대한 보다 합리적인 추정 및 이에 근거한 정책 대안 수립이 가능해지고 있다. 본 연구의 분석 결과 혹은 접근 방식을 보다 정교화함으로써 사교육과 관련된 원인을 보다 실증적으로 파악하는 한편, 이의 해소를 위한 정책 대안 수립을 위한 기초자료로 활용될 수 있다. 이런 맥락에서 정책 성과에 대한 예측 및 분류는 해당 정책의 성패를 선험적으로 가늠해 봄으로써 실패 확률을 줄이고 재정 효율성을 제고할 수 있다는 점에서 상당한 의의가 있다. 따라서 교육학 분야에서도 머신러닝을 활용한 실증분석이 향후 활발히 이루어질 필요가 있음을 제언한다.

주제어: 사교육비, 머신러닝, 랜덤포레스트, 고정효과, 확률효과

Ⅰ. 서론

2019년 3월에 발표된 교육부와 통계청의 '2018년도 초·중·고 사교육비 조사' 결과에 따르면, 전체 초·중·고등학교 학생들의 72.8%가 사교육에 참여하였으며, 전년도에 비해 1.7%p 증가한 것으로 나타났다(통계청, 2019.3.11.). 이를 학교급별로 살펴보면, 초등학생이 82.5%, 중학생이 69.6%, 고등학생이 58.5% 순이었다. 반면, 사교육비 규모는 고등학생들이 가장 많은 것으로 나

1) 경남대학교 교육학과 조교수
2) 한국교육개발원 민주시민교육연구실 부연구위원
3) 공주대학교 교육학과 조교수

타났다. 사교육에 참여하지 않은 학생들까지 포함한 초·중·고등학생 전체의 월평균 1인당 사교육비는 29만 1천 원으로 전년 대비 1만 9천 원 정도가 증가(7.0%)하였으며, 학교급별로는 고등학생이 32만 1천 원, 중학생이 31만 2천 원, 초등학생이 26만 3천 원 정도였다. 더욱이, 이러한 수치는 사교육을 참여하지 않은 학생들까지를 포함한 것으로, 실제 사교육 경험이 있는 학생들만을 대상으로 할 경우 고등학교 학생들의 1인당 월평균 사교육비는 54만 9천 원으로 증가한다. 중학교와 초등학교도 각각 44만 8천 원과 31만 9천 원으로 전년도에 비해 모두 소폭 증가(고 7.6%, 중 3.7%, 초3.9%)하였으며, 고등학교 학생들의 사교육비가 증가가 두드러진다. 더욱이, 각 가구에서 매월 고정적으로 지출해야 하는 항목들을 고려할 때, 가구 소득에서 사교육비가 차지하는 비중은 상당히 높은 수준이라고 할 수 있다.

사실, 정부는 과외금지 조치가 취해진 1980년 직후부터 사교육 수요를 억제하고 학부모들의 사교육비 부담을 줄이고자 지속적으로 노력해 왔다(문지영, 모은비, 서은정, 조정우, 2018). 그러나 2000년에 과외금지 조치가 위헌이라는 헌법재판소 결과가 발표된 이후 사교육 시장은 다시 빠르게 팽창하였다(송경오, 2013). 이에 노무현 정부는 2004년에 '공교육 정상화를 통한 사교육비 경감 대책'을 발표하였으며(교육인적자원부, 2004.2.17.), 이명박 정부에서는 2009년에 '공교육 경쟁력 향상을 통한 사교육비 경감 대책'을(교육과학기술부, 2009.6.3.), 그리고 박근혜 정부에서는 2014년에 '선행학습금지법'으로 불리고 있는 「공교육 정상화 촉진 및 선행교육 규제에 관한 특별법」을 제정하였다. 하지만, 교육부와 통계청의 사교육비 조사 결과에서도 나타나듯이(통계청, 2019.3.11.), 학생들의 사교육 참여 비율은 여전히 줄어들지 않고 있으며, 학부모들의 사교육비 부담이 가중되고 있을 뿐만 아니라 가계 소득에 따른 사교육비 지출 규모의 격차도 같이 증가하고 있어 교육 불평등 심화를 우려하는 목소리가 높다. 정부의 정책적 노력과 이에 따른 사교육비 시장의 변화를 분석한 유재봉, 조정우, 서은경, 김현철(2016)의 연구 또한 정부의 지속적이고 다양한 사교육비 경감 대책에도 불구하고 사교육 시장은 오히려 더욱 커져가는 경향을 보이고 있음을 지적하고 있다.

그러나 사교육 현상의 원인과 이를 해결하기 위한 방안에 대한 논의는 일치된 결론에 이르지 못하고 있다(김희삼, 2010; 문지영, 모은비 외, 2018). 가장 큰 이유는 각각의 연구들마다 분석대상이나 분석자료, 분석방법, 분석모형 등이 서로 다르기 때문이라고 할 수 있다(정동욱 외, 2012). 그러나 대부분의 연구들이 기존의 사교육 관련 이론이나 선행연구에 근거하여 비슷한 형태의 연구 모형을 설정하고 있을 뿐만 아니라, 유사한 변수들을 사용하여 분석을 실시하고 있음에도 불구하고 사교육 수요에 대한 각 변수들의 영향력이 서로 상반되게 나타나는 것은 상당히 중요한 문제라고 할 수 있다(송경오, 이광현, 2010; 송경오, 2013; 문지영, 김현철, 박혜연, 2018). 예를 들어, 한국교육종단연구(Korean Education Longitudinal Study, 이하 KELS) 자료를 사용하여 '방과후학교 프로그램' 및 '수준별 수업 운영의 충실성'이 사교육 수요(사교육 시간 및 비용)에 미치는 영향력을 분석한 이수정, 민병철(2009)의 연구에서는 사교육에 대한 유의미한 영향력이 존재한다고 주장한 반면, 같은 자료를 사용한 김희삼(2009)의 연구나 한국교육개발원에서 별도로 수집한 자료를 사용한 박소영(2008)의 연구에서는 방과후학교 프로그

램의 효과가 유의미하지 않다고 결론 내렸다. 또한, 이수정(2011)의 연구에서는 입학사정관 제도 도입 등의 대입제도 간소화 노력을 통해 사교육 참여 및 사교육비 지출이 감소하였다고 지적하였으나, 이필남(2011)과 김위정,김양분(2013)의 연구에서는 학생들이 입학사정관 전형을 선택한다 하더라도 사교육 참여나 사교육비 지출에 거의 차이가 없다고 결론 내렸다.

이처럼 그 동안 사교육 수요의 원인 분석과 해결 방안 모색을 위한 다양한 연구들이 수행되었으나, 학생들의 사교육 참여 및 사교육비 지출에 대한 논의들은 여전히 하나의 결론에 수렴되지 못하고 있다. 이에 대해 일부 연구자들은 그 동안 사교육 수요를 설명하기 위해 사용된 분석 모형이나 변수들이 상당히 제한적이었을 뿐만 아니라, 중요한 변수를 누락하거나 보다 직접적인 변수들을 통합 또는 분리해 내는데 실패했을 가능성을 제기하기도 한다(문지영, 모은비 외, 2018). 사교육에 대한 참여 여부 및 사교육비 지출 규모에 영향을 미치는 요인은 매우 다양하기 때문에 개인적·사회적 특성뿐만 아니라, 학교 특성이나 사교육을 둘러싼 정책적 요소 등을 함께 반영해야 한다는 것이다(김화경, 2017; 장지윤, 박인우, 장재홍, 2017; 박선영, 마강래, 2015; 이광현, 2013; 이수정, 2011).

더욱이, 그 동안 외국어고를 비롯한 특목고 입시 문제나 자사고 입시 과열 문제 등이 사회적 주목을 받으면서 사교육 관련 선행연구들이 주로 중학교 학생들의 사교육 수요에 보다 관심을 보여 왔기 때문에, 고등학생들의 사교육비 관련 연구들은 상대적으로 부족한 것이 사실이다. 2018년에 추진된 국가교육회의의 대입제도 공론화 이후 발표된 '2022 대입 개편안'이나 일반고, 자사고, 외고, 국제고 고입 동시 실시, 자사고 및 외고의 일반고 전환 등 고등학교를 중심으로 한 다양한 정책적 논의들이 거의 동시에 추진되면서 대입에 대한 학생 및 학부모들의 불안감이 증가했으며, 그러한 불확실성의 증가는 향후 고등학교 단계의 사교육 수요를 더욱 증가시킬 수 있다는 우려가 제기되고 있으나(한국일보, 2019.3.13.; 권민지, 2019.3.12.; 손수람, 2019.1.21.), 고등학교 학생들의 사교육 실태 및 수요를 설명하고 이를 예측하기 위한 분석 모형에 대한 연구들은 아직 부족한 실정이다.

이처럼 고등학교 유형이나 대학 입시의 큰 틀이 변화하면서 고등학교 학생들의 사교육 수요 또한 빠르게 변화하고 있을 것으로 예상되는 바, 고등학교 학생들에 초점을 맞추어 사교육 수요 현상을 새롭게 분석해 볼 필요가 있을 것으로 보인다. 이에 본 연구에서는 머신러닝(machine learning) 기법의 하나인 랜덤 포레스트(Random Forest) 방법을 사용하여 사교육비 지출에 대한 예측력이 높은 변수들을 추출하고, 이를 활용하여 사교육비를 종속변수로 하는 일반 OLS 및 고정효과 모형, 확률효과 모형 분석을 실시함으로써 보다 적절한 사교육비 지출 예측 모형을 탐색하고자 하였다. 분석을 위해 한국교육고용패널Ⅱ(Korean Education and Employment PanelⅡ, 이하 KEEPⅡ)의 고등학생 2차 년도(2017년) 조사 자료를 사용하였다.

Ⅱ. 선행연구 및 이론적 고찰

1. 사교육의 개념 및 범위

한국교육개발원이나 한국직업능력개발원, 한국청소년정책연구원 등의 정부출연 연구기관뿐만 서울시교육청이나 경기도교육청 등의 시·도교육청 등에서 학생 관련 패널 자료를 구축·운영하기 시작하면서, 초·중·고등학교 학생들의 사교육 현황 및 영향 요인에 대한 분석 연구가 활발해졌다. 사교육 연구에 활용 가능한 자료로 우선, 통계청에서 매년 조사하고 있는 '초·중·고 사교육비 조사'가 있으며, 한국교육개발원에서 매년 실시하고 있는 '한국교육종단연구(KELS)'의 2005 코호트와 2013 코호트, 같은 기관에서 초·중·고등학교 학교급을 3년 주기로 조사하고 있는 '학교 교육 실태 및 수준 분석 연구', 2017년부터 일제 고사 방식에서 표집 평가 방식으로 변경된 한국교육과정평가원의 '국가수준 학업성취도 평가', 한국청소년정책연구원의 '한국아동·청소년 패널조사(NYPI)', 한국직업능력개발원의 '한국교육고용패널(KEEP)' 2004 코호트 및 2016 코호트, 서울시교육청의 '서울교육종단연구(SELS)' 2010 코호트, 경기도교육청의 '경기교육종단연구(GEPS)' 2012 코호트 등이 있다.

그런데 이들 각 조사마다 정의하고 있는 사교육의 개념이나 그것에 포함되는 구체적인 활동의 사례들이 조금씩 다른 것을 확인할 수 있다.

〈표 1〉 사교육 조사 및 패널 자료별 사교육비 범위

조사 및 자료	사교육비 범위
통계청 사교육비 조사	• 초·중·고 학생들이 학교의 정규 교육과정 이외에 사적인 수요에 의해서 학교 밖에서 받는 보충교육을 위해 개인이 부담하는 비용 • 학원수강, 개인 및 그룹 과외, 방문학습지, 인터넷 및 통신 강의 등 • 방과후학교, EBS 교재, 어학연수, 진로진학상담 등은 별도 조사
한국교육개발원 한국교육종단연구	• 학원(단과반, 종합반), 과외(개인과외, 그룹과외), 학습지 및 통신·인터넷 과외 등
한국직업능력개발원 한국교육고용패널	• 학원 수강과 개인 과외, 그룹 과외, 학습지, 유료 인터넷 및 통신 과외, 학교 내 방과후학교, EBS 특강 등
서울시교육청 서울교육종단연구	• 학원이나 과외, 학습지 등을 사용하는 활동을 의미하되, EBS 교육방송 시청 제외
경기도교육청 경기교육종단연구	• EBS 교육방송이나 방과후학교, 자율학습 등을 제외한 학교 외 교육활동으로, 과외, 학원수강, 학습지 및 인터넷 강의 등

예를 들어, 통계청에서 매년 발표하고 있는 초·중·고등학교 학생들의 사교육비 조사에서 정의하고 있는 사교육비의 범위는 '초·중·고 학생들이 학교의 정규 교육과정 이외에 사적인 수요에 의해서 학교 밖에서 받는 보충교육을 위해 개인이 부담하는 비용'에 해당한다(통계청, 2019.3.11.). 조사에 포함되는 사교육 유형은 학원수강이나 개인 및 그룹 과외, 방문학습지, 인터넷 및 통신 강의 등으로 이에 소요되는 교재비 및 수강료를 포함한다. 다만, 방과후학교 참

여나 EBS 교재 구입, 어학연수 참여, 진로진학 학습상담 등은 사교육비와 성격이 다른 것으로 판단하여 별도의 항목으로 분리하여 조사하고 있다. 반면, 한국교육개발원의 '한국교육종단연구'에서는 '학원(단과반, 종합반), 과외(개인과외, 그룹과외), 학습지 및 통신·인터넷 과외' 등을 모두 포함하고 있다. 서울시교육청의 '서울교육종단연구'와 경기도교육청의 '경기교육종단연구'에서는 '과외나 학원, 학습지, 인터넷 강의' 등을 사용하는 학교 외 활동을 사교육으로 규정하였으며, 앞선 조사들과 마찬가지로 학교 내 방과후학교나 EBS 교육방송 교재 구입 및 특강 시청은 사교육에서 제외하고 있다.

반면, 본 연구에서 사용하고 이는 '한국교육고용패널' 2016 코호트의 일반고 학생용 설문지에서는 개별 과목별 사교육 경험 여부와 참여 시간, 월 평균 사교육 비용 등을 조사하고 있는데, 사교육의 범위에 포함되는 활동으로 학원 수강과 개인 과외, 그룹 과외, 학습지, 유료 인터넷 및 통신 과외, 학교 내 방과후학교, EBS 특강 등을 제시하였다. 앞선 통계청 조사 자료나 서울교육종단연구의 조사 자료와 달리, 사교육 활동 범위에 EBS 특강이나 학교 내 방과후학교 과외 활동이 모두 포함되어 있다는 점이 특징적이라 할 수 있다.

2. 고등학생의 사교육비 지출에 영향을 미치는 변수 탐색

고등학교 학생들의 사교육비 지출 규모에 관한 연구는 두 가지 유형으로 구분해 볼 수 있다. 첫 번째 유형은 사교육비 영향 요인에 대한 탐색적 연구로, 새롭게 발표된 실태 조사 자료 또는 패널 조사 자료를 활용하거나 새로운 분석방법 또는 분석모형을 사용하여 사교육비 지출 규모에 영향을 미치는 요인을 탐색하는 연구들이 이에 해당한다. 일반계 고등학교 학부모의 사교육비 지출 결정 변인에 대한 '경로분석(path analysis)'을 실시한 김현진(2004)은 일반계 고등학생의 사교육비 지출이 주로 도시 거주 여부나 부모의 학력 및 사회경제적 수준과 같은 개인적 배경 요인들에 의해 영향을 받고 있으며, 공교육 내실화나 대입제도 개선, 학부모의 교육관 및 사회풍토 개선 등과 같은 정책적 차원의 교육비 경감 대책들은 유의미한 영향을 미치지 못한다고 밝혔다. 박균달과 김현진(2011)은 고교평준화 제도와 부모소득, 거주지의 도시 여부, 부학력, 모학력, 성적 등의 변수가 시설, 교사, 학교생활 등에 대한 학교불만족 변수를 매개로 학생들의 사교육비 지출에 미치는 영향을 살펴보았다. 분석 결과에 따르면, 고교평준화 정책은 고등학교 학생의 사교육비 지출에 유의미한 영향을 미치지 않았다. 반면, 월평균 소득이 높으며 도시 지역에 거주할수록, 외고 학생들이 일반고 학생들보다, 진학희망 대학 수준이 높을수록 사교육비 지출이 증가했으며, 외고 학생들의 경우 학교에 대한 불만족이 높을수록 사교육비 지출이 증가하는 것으로 나타났다. 초·중·고 학생의 사교육비 영향요인을 분석한 이혜정, 송종우(2014)는 회귀분석과 분류분석 방법을 사용하여 사교육비에 영향을 미치는 중요변수가 무엇인지를 탐색하였는데, 분석 결과에 따르면, 대도시가 중소도시 보다, 가구소득이 높을수록, 일반고, 중학교, 특성화고, 초등학교 순으로 사교육비 지출이 많은 것으로 나타났다.

최근에는 패널 조사 자료를 사용하여 관찰 불가능한 개인 및 환경의 이질적 특성을 적절히

통제하고 선택편의를 최소화하기 위한 새로운 분석 방법들이 적용되고 있다. 예를 들어, 송경오, 이광현(2010)은 '패널 확률효과 토빗모형(panel random effects Tobit model)'을 사용하여 고등학교 학생의 사교육 참여 시간 및 사교육비 지출 규모에 영향을 미치는 학교 및 교육 특성 변수를 분석하였으며, 거주지 규모(대도시, 중소도시, 읍면)와 방과후학교 참여 여부의 상호작용 항을 투입하여 방과후학교 변수의 영향력을 세분화하여 살펴보았다. 분석 결과에 따르면, 고등학교 시기의 사교육 참여 및 사교육비 지출은 1학년 시기에 집중 투자가 이루어졌다가 2학년 시기에 조금 감소하며, 3학년 시기에 다시 증가하는 경향을 보였다. 그리고 방과후학교나 수준별 수업, 자율학습 모두 고등학생들의 사교육 수요를 약화시키는 것으로 나타났으며 특히, 방과후학교 참여 비율이 높은 학교에서 학생들의 사교육 참여 및 사교육비 지출이 크게 감소하는 것으로 나타났다. 그 외에도 해당 연구는 기간제 교사 비율이 낮을수록, 교과협의회가 활성화된 학교일수록 사교육비 지출이 감소한 것을 보고하였다.

〈표 2〉 사교육비 영향 요인을 탐색한 주요 선행연구

연구자	분석자료	분석방법	독립변수*
김현진 (2004)	한국교육개발원 수집 자료	경로분석	월평균 사교육비 지출, 월평균 가계소득, 거주 지역, 부 및 모의 학력, 학교불만족, 점수위주 선발제도 강도, 자녀교육에 대한 부모의 관심 및 걱정, 학력 및 학벌주의 문화풍토
송경오, 이광현 (2010)	한국교육종단연구	패널자료분석 중 확률효과 토빗모형	개인특성: 성별, 가구소득, 부모학력, 교육기대수준, 학습흥미도 학교배경: 설립유형, 남녀공학여부, 도시소재 여부 교육활동: 자율학습참여, 수준별수업, 방과후학교 교육의질: 학교교육만족도, 교사관심정도, 교사이해, 수업분위기 교육환경: 학운위, 학부모회, 교과협의회, 교사간관계, 학급당학생수, 교원노조, 기간제교사
박균달, 김현진 (2011)	저자들의 선행연구 및 자료	통계적 메타분석	부학력, 모학력, 소득, 거주지역, 성적, 평준화 여부, 학교교육, 사교육여부, 대학교 진학 희망, 학교불만족, 내신성적, 정의적 성취 수준 등
송경오 (2013)	국내 사교육 관련 선행연구 문헌	통계적 메타분석	사회·심리적 특성: 학교만족도, 교사협력, 교사 열의, 학급문화 조직·행정적 특성: 수준별, 방과후, 리더십, 교사 잡무 시간, 학교규모, 학급규모, 시설수준, 기간제 비율 제도적 특성: 평준화여부, 학교소재지, 학교SES
이광현 (2013)	한국교육개발원 학교실대조사	토빗모형 헤크만모형 경향점수매칭	성별, SES, 부모성별기대, 평균성적, 학생만족도, 학부모만족도, 교사성취압력, 교사지원, 교사열의, 수업분위기, 설립유형, 지역, 남녀공학, 학교규모, 학급규모, 교사만족도, 교사협력문화, 학교평균 SES, 수업지도성, 학생1인당 교수학습 활동비, EBS 시청, 방과후학교 참여 여부
문지영, 모은비 외 (2018)	서울교육종단연구	벌점 회귀모형	수업태도, 수업분위기, 진로성숙도, 진학정보 인지정도, 주당공부시간, 학업성적, 방과후학교 여부, EBS 수업 여부, 고등학교 선택 시 고려사항, 남녀공학 여부, 남녀합반 여부, 교사특성, 가계소득, 거주지역, 부모의 경제활동 여부, 부모와의 진로 대화 여부

다른 한 가지 유형은 특정 변인이 사교육비 지출 규모에 미치는 영향력을 분석하고, 그 결과가 가진 정책적 의미와 사교육 문제 완화를 위한 시사점을 도출하는 연구들이라 할 수 있다. 사교육 수요와 관련된 초기 연구들은 주로 학교 특성에 따른 차이에 초점을 맞추었는데, 고교평준화 지역 소재 여부나 공·사립으로 구분되는 학교 설립유형에 따른 사교육 수요의 차이를 실증적으로 확인하고자 하였다. 예를 들어 김현진과 최상근(2004)은 한국교육개발원의 '사교육 실태 및 사교육비 규모 분석 연구'에서 수집된 자료를 활용하여 고교평준화 실시 지역 여부에 따라 학부모들의 사교육비 지출 규모에 차이가 있는지를 분석하였다. 분석 결과에 따르면, 평준화 지역 여부는 사교육비 지출에 영향을 미치지 못하는 것으로 나타났다. 한국교육고용패널조사(KEEP) 자료를 바탕으로 고교평준화 제도가 사교육비 지출에 미치는 영향을 분석한 채창균(2006)의 연구나 한국청소년패널조사(KYPS) 자료를 사용하여 고교평준화 또는 경쟁 선발 전형 학교(특목고, 자사고 등) 재학 여부에 따른 사교육비 차이를 분석한 강태중(2009)의 연구 역시 평준화 제도가 사교육비 지출에 영향을 미치지 못한다고 결론 내렸다. 가장 최근에는 강소랑(2016)이 한국교육종단연구(KELS) 자료를 사용하여 고교평준화 여부가 사교육 여부, 사교육 시간, 사교육 비용 등에 미치는 영향을 분석하였는데, 앞선 다른 연구들과 마찬가지로 평준화 지역과 비평준화 지역 사이에 유의미한 차이가 없음을 다시 확인하였다.

이후 특목고 진학 계획 또는 특목고 재학의 사교육비 유발 효과나 고등학교의 유형에 따른 사교육비 지출 차이를 검증하기 위한 연구들도 활발하게 실시되었다. 김성식과 송혜정(2009)은 중학교 1학년부터 고등학교 1학년까지의 4개년도 한국교육종단연구(KELS) 패널자료를 사용하여 특목고 진학 계획과 학교 불만족 요인이 사교육비에 미치는 영향을 분석하였다. 분석 결과를 살펴보면, 중학교 기간 동안은 학교 불만족 요인이 사교육비에 영향을 미치지 않았으나, 고등학교 1학년 시점에는 사교육을 증가시키는 요인으로 작용하였다. 그리고 중학교 시기에 특목고 진학 계획을 가진 학생들의 사교육 수요는 상당하지만, 진학 경쟁이 종료된 고등학교 1학년 시점에는 사교육비가 크게 감소하였다. 하준경(2010)은 특목고 정원의 변화가 가계의 사교육비 지출에 미치는 영향을 '경제학적 이론 모형'과 '시뮬레이션 기법'을 활용하여 분석하였는데, 그에 따르면 특목고 비중의 확대가 사교육비에 미치는 영향은 일반고의 질이 어떠한지에 가장 큰 영향을 받는 것으로 나타났다. 즉, 사교육비 절감이라는 정책적 목표를 달성하기 위해서는 무엇보다도 먼저 일반고의 질적 수준 제고 노력이 중요함을 지적하였다. 반면, 신혜진(2017)은 '3수준 다층성장모형'을 사용하여 서울교육종단연구(SELS) 자료에 나타난 고등학교 유형별 사교육비 지출 차이를 분석하였다. 분석 결과를 살펴보면, 특목고와 자사고에 진학한 학생들은 일반고 학생들에 비해 중학교 당시 많은 사교육비를 지출하였으며, 이러한 사교육비 지출 차이는 고등학교 진학 후에도 그대로 유지되고 있는 것으로 나타났다.

학교 특성에 따른 사교육 수요의 차이에 주목했던 이들 연구들과 달리 최근에는 학교교육 활동이나 입시 제도의 변화에 따른 사교육 수요의 변화를 분석하는 연구들이 보다 활발하게 실시되고 있다. 예를 들어, 채재은, 임천순, 우명숙(2009)과 정동욱 외(2012), 장지윤, 박인우, 장재홍(2017) 등은 EBS 교육 프로그램 시청이 고등학생의 사교육비 지출

에 미치는 영향을 분석하였으며, 심은석, 박균달, 김현진(2013)과 문지영, 김현철 외(2018) 등은 방과후학교 참여가 사교육비 경감에 미치는 영향을 분석하였다. 반면, 이광현(2013)은 방과후학교 참여와 EBS 인터넷 방송 시청 변수 모두에 주목하고 있는데, 토빗모형과 표본선택을 보정하기 위한 헤크만(Heckman) 모형뿐만 아니라, 방과후학교 참여와 EBS 시청을 학생들의 의도적인 자기선택행위로 가정하고 선택편의(self-selection bias) 문제를 해결하기 위해 경향점수 매칭 모형(propensity score matching model)을 함께 활용하였다. 분석 결과에 따르면, 방과후학교 참여는 일관되게 사교육비 지출을 감소시켰으며, EBS 시청은 토빗모형과 경향점수매칭 모형에서 사교육비 지출에 부적인 영향을 미치는 것으로 나타났다.

대학 입시제도의 변화에 따른 고등학생 사교육 참여 또는 사교육비 지출 변화를 분석한 연구도 상당히 많이 이루어졌다. 우선, 이수정(2011)은 수능 9등급제 및 수시선발 도입을 특징으로 하는 2002년 대입제도 개선안과 수능 점수 폐지 및 입학사정관제 도입, 기회균등할당제 도입 등으로 대표되는 2008년 대입제도 개선안이 사교육 시간 및 사교육비 지출에 미친 영향을 비교·분석하였다. 분석 결과에 따르면, 2008년 대입제도 개선안 도입 이후 사교육 참여 비율과 사교육비 지출 모두 크게 감소한 한 것으로 나타났다. 그리고 2008학년도 대학 입시에서 일반전형과 특별전형을 선택한 학생 간 혹은 수시모집과 정시모집을 선택한 학생 간에 사교육비 지출 규모에 차이가 있는지를 분석한 이수정, 조원기(2014)의 연구 결과를 살펴보면, 수시모집 입학생이 정시모집 입학생에 비해 고등학교 3학년 시기의 사교육비 지출 규모가 작았으나, 일반전형 학생과 특별전형 학생 사이에는 사교육비 규모의 차이가 없는 것으로 나타났다. 이필남(2011)은 한 걸음 더 나아가 입학사정관 전형을 준비하는 고등학교 3학년 학생과 그렇지 않은 학생의 사교육비 지출 차이를 분석하였는데, 경향점수매칭과 토빗 모형을 적용하여 분석한 결과에 따르면 입학사정관제 여부는 사교육 수요 증감에 유의미한 영향을 미치지 않는 것으로 나타났다. 2013년에는 김위정, 김양분(2013)이 다시 한 번 입학사정관제 전형 지원 계획에 따른 사교육비 지출 차이를 분석하였다. 다만, 앞선 이필남(2011)의 연구가 입학전형 방법에 대한 학생들만의 응답 결과를 분석 대상으로 하고 있던 것과 달리, 김위정, 김양분(2013)은 대입전형 방법 선택에 있어 부모의 영향력이 상당하다는 점을 고려하여 입학사정관제 전형 지원 계획에 대한 학생의 응답과 학부모의 응답을 별도로 분석하였으며, 양자의 응답 결과를 종합하여 둘 중 하나라도 입학사정관제 전형 지원 의사가 있으면 '있음'으로 간주하여 추가 분석을 실시하였다. 이들의 분석 결과에 따르면, 입학사정관제 전형 지원 의사는 사교육비 지출 규모에 영향을 미치지 않았으나, 사교육 참여 여부에는 유의미한 차이들이 나타났으며 특히, 읍·면 지역에 거주하는 입학사정관제 지원 희망 학생들의 사교육 참여 비율이 보다 높은 것으로 나타났다.

Ⅲ. 데이터 및 분석방법

1. 분석 대상

본 연구는 고3 학생들의 사교육비 지출과 관련된 변수를 추정하기 위해 한국직업능력개발원에서 제공하는 한국교육고용패널Ⅱ(Korean Education and Employment PanelⅡ: KEEPⅡ)의 2차년도(2017년) 조사 자료를 활용하였다. 한국교육고용패널Ⅱ는 2004년부터 진행한 한국교육고용패널Ⅰ에 이어 2010년대 중·고교생에 대한 조사를 새롭게 진행하기 위해 시작된 조사로, 2016년 당시 고등학교 2학년에 재학 중이던 학생과 그 보호자(학부모) 및 담임교사, 참여 학교의 학교행정가에 대한 조사를 실시하였으며, 2차년도 조사인 2017년에는 학생 패널(당시 고3) 9,157명을 대상으로 조사를 실시하였다. 본 연구에서는 2차년도 조사에 참여한 고3 학생 9,157명 중 진학 관련 월평균 사교육 비용(431명) 및 취업/자격증 관련 월평균 사교육 비용 관련 문항(101명)에 응답하지 않은 534명을 제외한 8,625명을 대상으로 분석을 실시하였다.

2. 변수 설명

본 연구는 고3 학생들의 사교육비 지출과 관련된 변수들을 예측하고, 해당 변수들을 설명변수로 하는 OLS 및 고정효과 모형, 확률효과 모형 등을 활용하여 보다 적절한 사교육비 예측 모형을 도출하고자 한다. 이를 위해 우선 고3 학생들의 사교육비 지출액을 타내는 변수를 구성하기 위하여 KEEP Ⅱ 2차년도 조사 자료 중 학생들을 대상으로 '진학 관련 월평균 사교육비용'(Y17SA02149)과 '취업/자격증 관련 월평균 사교육 비용'(Y17SA02150)를 활용하여 두 변수 값을 합한 '월평균 사교육 비용' 변수를 생성하였다. 이에 따르면 분석 대상 8,625명들의 월평균 사교육 비용 평균은 29.1만원임, 표준편차는 5.9만원인 것으로 나타났다.

〈표 3〉 고교생의 월평균 사교육 비용에 대한 기초 통계량

월평균 사교육 비용	Obs	평균	표준편차	최소	최대
	8,625명	29.08(만원)	5.92(만원)	0	30,000,000

연속변수를 종속변수로 하는 전통적인 OLS 회귀분석과 마찬가지로 랜덤 포레스트의 경우 또한 종속변수를 예측하기 위한 설명변수를 필요로 한다. 다만, 랜덤 포레스트의 경우 다양한 설명변수들의 상호작용과 비선형성을 고려하여 추정 결과를 얻을 수 있음과 함께, 많은 설명변수를 모형에 포함시키더라도 자유도 감소의 문제를 일으키지 않는다는 장점을 지니고 있기에 가능한 많은 변수를 포함하여 고3 학생들의 사교육비 지출에 영향을 미치는 변수를 예측할 수 있다(최필선, 민인식, 2018). 이에 본 연구에서는 KEEP Ⅱ 2차년도 자료에서 제공하는 변수 중 추정 결과의 안정성을 확보하기 위하여 전체 응답자 중 50% 이상이 결측값을 보인 변수들

을 제외한 191개의 설명변수를 사용하여 고3 학생들의 사교육비 지출에 영향을 미치는 요인을 분석하였다. 분석에 활용된 설명변수를 범주별로 제시하면 다음의 <표 4>와 같다.

〈표 4〉 랜덤 포레스트에 활용된 설명변수 및 범주

범주	설명변수
학교생활	학교 유형, 고교 소재지, 고교 계열, 고교 계열 선택 이유, 재학 중인 학교 유형 선택 이유, 재학 중인 학교 유형 선택 시기, 학교 선택시 영향을 준 사람(1, 2순위), 고등학교 생활 만족도, 고등학교 시설 인식, 교사 인식, 고3 담임교사 인식, 교과 교사 만족도, 수업 태도, 수업 분위기, 내신등급(국어, 수학, 영어, 과학, 사회, 음악, 미술, 체육), 전반적 진로교육 만족도, 진로교육 및 활동 경험 여부 및 만족도, 진로에 관한 질문, 학교 내외 수상 여부
학습 및 사교육	방과후자율학습 운영 및 참여 여부, 방과후학교 경험 여부, (국어/영어/수학/사회/과학/제2외국어/논술/예체능별) 방과후학교 경험 여부, 취업/자격증 방과후학교 참여 여부, 진학 관련 방과후학교 비용, 취업/자격증 관련 월평균 방과후학교 비용, 혼자 공부하는 시간(주중, 주말), 현장 체험 경험 여부 및 횟수, 현장 체험 만족도, 자격증 유무
가정생활	거주 형태, 현재 거주지, 현재 거주 지역 규모, 가정생활 전반적 만족도, 부모와의 활동 빈도, 부모와의 대화 빈도, 남성(여성) 보호자의 학력/경제활동 상태/직장 고용형태/월평균 소득/직장 종사자수, 보호자의 부동산/금융소득, 해외여행 경험
여가생활	여가 시간(평일/휴일), 여가 시간 활동 빈도, 독서 경험 여부 및 독서량, 주요 독서 분야, 독서 관련 인식, 한달 용돈 액수, 용돈 소비 1/2순위
재학 중 근로경험	재학 중 근로경험 여부
진로계획 및 직업의식	희망 교육 수준, 미래 직업 결정 여부, 미래 직업 관련 인지 수준, 직업을 갖는 이유, 직업생활의 성공조건, 인생을 사는 데 중요한 것(가치관)
건강	건강상태, 일주일 평균 운동시간, 수면 시간, 아침 식시, 고민/걱정거리, 흡연 및 음주 여부
개인적 특성	성별, 행복도, 다문화수용성, 자아효능감

본 연구는 <표 4>에 제시된 변수들을 대상으로 고 3 학생들의 월평균 사교육비 지출과 관련하여 상대적인 예측력이 높은 변수들을 탐색한 후, 이들을 활용하여 고 3 학생들의 사교육비를 종속변수로 하는 OLS 및 고정효과 모형, 확률효과 모형 분석을 실시함으로써 보다 적절한 사교육비 지출 예측 모형을 탐색하고자 하였다. 랜덤 포레스트 결과 도출된 변수들에 대한 내용은 IV장 분석 결과의 2절과 3절에 제시되어 있다.

3. 분석방법

1) 랜덤 포레스트(Random Forest)

본 연구는 고교생의 사교육 지출에 영향을 미치는 변수를 예측하기 위해 랜덤 포레스트(Random Forest 기법)을 활용하였다. 본 연구는 191개의 설명변수를 이용하여 고교생의 사교

육비 지출액을 추정하는데, 이러한 자료에 대해 전통적인 OLS 회귀분석 모형을 적용할 경우 변수의 수가 많아짐에 따라 자유도 감소로 인한 오류 확률이 높아지게 된다.

이러한 전통적인 접근 방식의 한계를 극복하기 위한 랜덤 포레스트는 일종의 데이터 마이닝 기법 중 하나로서 의사결정나무 모형을 기저로 하여 무작위성을 최대로 부여함으로써 예측오차를 줄이는 것으로, 상대적으로 높은 예측력 및 모형 안정성을 지니며(유진은, 2015), 투입되는 설명변수가 늘어나더라도 자유도 감소의 문제를 야기하지 않으므로 본 연구의 목적에 적합한 분석방법이라고 할 수 있다.

본 연구에서는 8,625명의 표본을 일반적으로 많이 이용되는 7:3의 비율로 훈련자료(training data)와 시험자료(test data)로 나누어(유진은, 2015; 최필선 외, 2018), 각 자료별로 예측 모형으로부터 도출된 사교육비 추정값과 실제 사교육비 지출액 간의 상관계수값과 둘 간의 편차값을 확인함으로써 예측 모형의 성능을 점검하였다. 훈련 자료는 실제 랜덤 포레스트 분석이 이루어지는 자료로서 해당 자료를 통해 종속변수에 대한 예측 모형을 도출해내게 된다. 일반적으로 랜덤 포레스트 기법은 훈련 자료로부터 도출된 예측 모형을 시험 자료에도 적용하여 유사한 결과가 도출되는지를 확인함으로써 모형의 강건성을 확보하게 된다. 한편, 랜덤 포레스트에서는 모형에 투입된 설명 변수 중 일정 개수의 설명 변수를 무작위로 선택하여 의사결정나무를 생성하는데, 본 연구에서는 Breiman(2001)이 제안한 룰(설명변수의 개수/3)에 따라 각 의사결정나무마다 63개의 예측 변수를 사용하는 것으로 설정하였다.

다음으로 표본과 변수를 무작위로 선택해 의사결정나무를 생성하는 과정을 1,000회 반복하여 모형의 예측오차를 분석함과 동시에, 사교육비 지출에 대한 설명변수의 예측 중요도를 나타내기 위해 평균제곱오차비율증가(%IncMSE)와 노드순도증가(IncNodePurity)를 확인하였다. 해당 중요도 지수가 높은 순으로 사교육비 지출에 영향을 미치는 주요 상위 10개 변수들을 제시하였고, 해당 설명변수가 반응변수에 미치는 한계효과를 측정한 부분의존성도표(Partial Dependence Plot: PDP)를 제시하여 각 변수의 범주별 취업 효과를 추정하였다. 이러한 분석을 위한 도구로는 R(ver.3.6.1)의 randomForest package를 사용하였다.

2) OLS

본 연구에서는 랜덤 포레스트 분석 결과에 기초하여 우선, 고3 학생들의 월평균 사교육비용을 종속변수로 하는 전통적인 교육생산함수를 활용하여, 종속변수와 관련성이 있는 학생 특성 변수들을 통제한 후 랜덤 포레스트에서 도출된 주요 설명변수들이 고교생의 사교육비 지출에 미치는 영향력을 살펴보았다.

3) 학교 고정효과 모형 & 학교 확률효과 모형

위와 같은 중다회귀모형에 기반한 교육생산함수 추정은 학생들이 속한 개별 학교의 고유한 특성을 모형에 고려하지 못한다는 단점을 지닌다. 특히, KEEP II와 같이 학교를 우선 추출한

후 해당 학교에 재학 중인 학생들을 조사한 자료의 경우 개별 학생들과 밀접히 연관된 단위 학교의 특성을 고려하지 못할 경우 추정된 계수의 일치성 및 효율성과 관련된 제약이 따른다.

고정효과모형은 이와 같이 관측 불가능한 개체 특성 오차항(본 연구에서는 고유한 학교 특성)을 고정된 상수로 간주하여 이를 추정하는 분석방법으로, 이를 통해 처치변수와 종속변수 간의 관계를 보다 엄밀하게 도출할 수 있다는 장점이 있다. 개체특성 오차항를 제거하기 위해 주로 활용되는 방법은 차분모형(differencing model)으로서, 이는 패널데이터의 각 시점 데이터에서 각 패널의 전체평균데이터를 차분한 후에 이에 대한 OLS 추정을 실시한다. 이와 같은 모형은 패널 개체의 개체 특성 오차항과 처치변수 간에 상관관계가 있더라도 회귀계수에 대한 일치 추정량을 얻을 수 있으며, 차분을 통해 오차항의 시간 개념이 사라져 자기상관 문제 또한 해결할 수 있다는 장점을 지닌다(민인식, 최필선, 2010).

그러나 해당 모형은 동일한 조사 대상이 반복적으로 측정된 패널 데이터에 적용 가능한 것으로, 본 연구와 같이 패널 조사의 특정 연도 자료를 활용하는 횡단면 조사 자료인 경우에는 각 조사 대상이 속한 학교의 고유한 개체 특성이 학교별로 서로 다르면서 고정되어 있다고 가정한 최소제곱더미변인 고정효과모형(Least Squares Dummy Variables Fixed Effect Model)을 활용할 수 있다. 이에 본 연구에서는 각 학교별로 상수항을 서로 다르게 추정하는 최소제곱더미변인(LSDV: least squares dummy variables) 모형을 활용하였으며, 이 수식은 아래와 같다.

$$(2)\ y_{ij} = \sum_{j=1}^{k} \alpha_j + \sum_{q=1}^{l} X'_{qij}\beta_q + \epsilon_{ij}$$

(i : 학생, j : 학교, l : 설명변수의 수, k : 전체학교 -1, α_j : j 학교의 고유 특성,

β_q : 회귀계수 벡터, X'_{qij} : 개인의 독립변수 벡터, ϵ_{ij} : 오차항)

한편, 이러한 고정효과 모형와 같이 학교의 고유한 특성이 고정된 것이 아니라, 일종의 확률변수로서 학교 간에 서로 독립이며 동분산을 지닌다는 가정을 할 경우에는 확률효과 모형을 활용할 수 있다. 특히, 이러한 확률효과 모형은 위계적 선형모형(HLM)의 Random Intercept Model과 동일한 모형으로서 본 연구와 같은 다층 구조의 자료 분석에 적합한 모형으로 알려져 있다. 이상을 고려할 때 학교의 고유한 특성을 고정된 것으로 볼 것인지, 확률적으로 변화하는 확률 변수로 볼 것인지에 따라 고정효과 모형과 확률효과 모형으로 구분할 수 있다.

이러한 고정효과 모형과 확률효과모형 중 분석에 적합한 모형의 선택은 두 모형의 추정치 간에 차이가 없다는 Hausman Test을 통해 결정한다. 위의 연구모형에서 처치변수와 학교 개체 특성 간의 상관관계가 없다는 가정이 참이라면, 즉 처치변수의 외생성(exogeneity)이 확보된다면 고정효과 추정량에 비해 확률효과 모형의 추정량이 더 효율적인 추정량인 것으로 알려져 있다. 그러나 설명변수의 외생성이 확보되지 못한다면 확률효과 모형 추정량은 일치추정량이 되지 못하는 단점이 있다. 이에 본 연구에서는 고정효과모형과 확률효과 모형을 모두 활용하여 분석을 실시한 이후, 설명변수의 외생성을 검증하는 Hausman Test를 실행함으로써 고정

효과모형과 확률효과모형의 분석 결과 중 보다 나은 추정량을 선택하여 보고하였으며, 그 결과를 분석 결과표에도 제시하였다.

Ⅳ. 분석 결과

1. 랜덤 포레스트를 활용한 고교생의 사교육비 지출 관련 요인 탐색

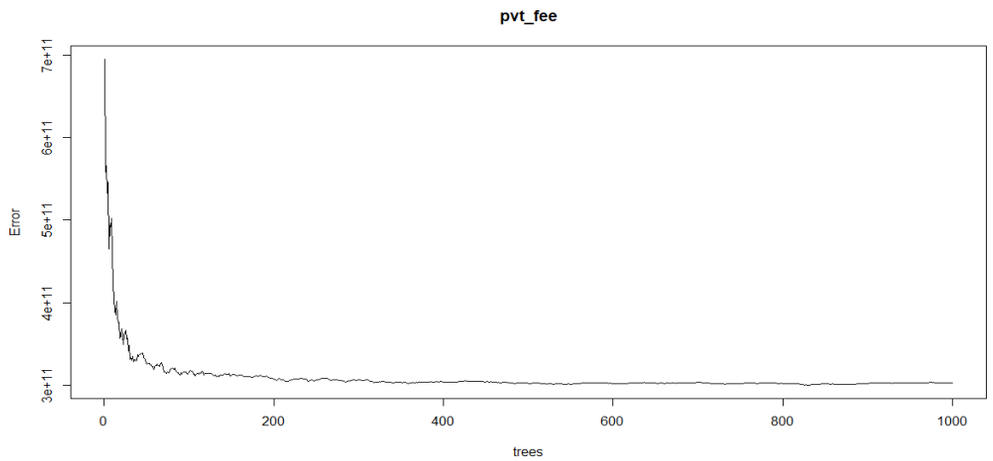
본 연구는 우선 랜덤 포레스트를 활용하여 고교생의 사교육비 지출에 영향을 미치는 요인을 분석하였다. 우선 반응변수인 사교육 비용은 연속형 변수로, 사교육 참여 여부와 같은 이분형 변수를 분류할 때 확인하는 정분류율, 특이도, 민감도를 통해 예측성과를 확인할 수 없다. 본 연구에서는 랜덤 포레스트 추정 값과 실제 사교육 비용 간의 상관 분석을 통해 추정 값과 실제 값이 서로 상관이 있는지 확인하는 한편, 실제 사교육비와 추정값 간의 차이를 살펴보았다. 다음 표는 훈련 자료와 시험 자료의 추정 값과 실제 사교육 비용 간의 상관 분석 결과를 나타낸 것으로, 시험 자료의 추정값과 실제 사교육비 간의 상관 계수가 0.446로 나타나, Rea & Parker(2005)의 기준으로 상대적으로 강한 양의 상관관계에 있는 것을 확인하였다. 이와 함께 실제 사교육 비용과 추정값 간의 차이는 5% 수준 내외임을 확인하였다.

〈표 5〉 랜덤 포레스트 추정값과 실제 사교육비 간의 상관 분석 결과

모형	훈련 자료	시험 자료
상관 계수	.973***	.446***
(실제값-추정값) 평균	-13,554 (4.66%)	12,883 (4.44%)

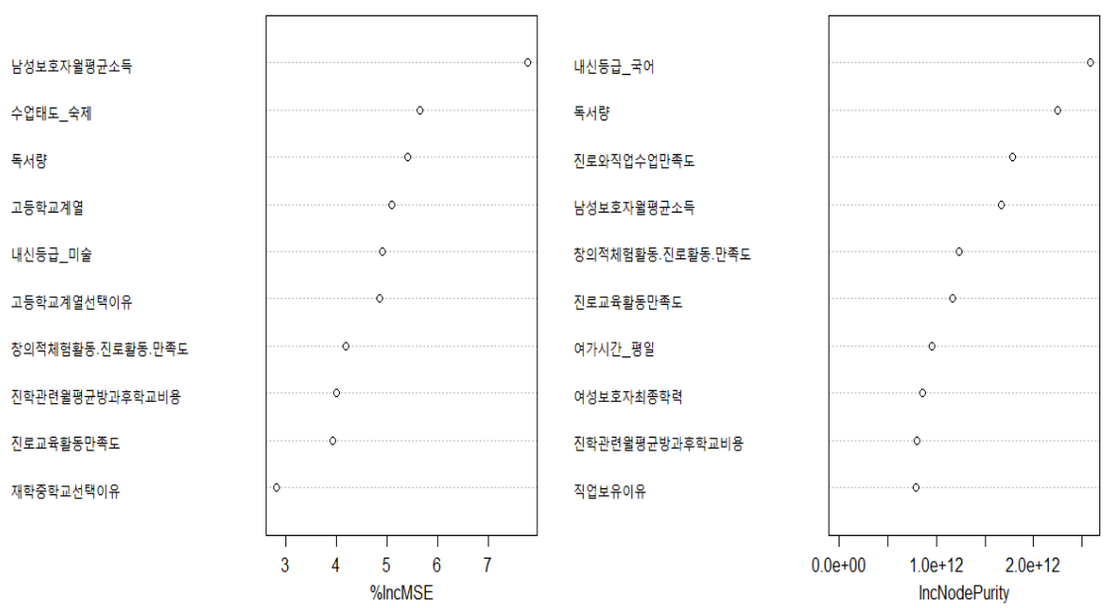
* 범례: *** 〈 .01 ** 〈.05 * 〈.1.

다음으로 의사결정트리 개수별 예측오차 비율의 변화를 [그림 1]과 같이 제시하였다. 랜덤 포레스트는 표본과 변수를 무작위로 선택해 의사결정트리를 생성하는 과정을 반복하는데, 본 연구에서는 최대 1,000개의 의사결정트리를 생성하도록 설정한 후 오차율 변화를 확인하였다. 이에 따르면 의사결정트리가 약 200개 이상이 되면 이후부터는 일정 값으로 수렴하는 것으로 나타났다. 따라서 본 연구에서 최대 1,000개의 의사결정트리를 생성하도록 설정한 것은 오차율을 수렴하게 하는 데에 충분한 숫자임을 확인하였다.



[그림 1] 의사결정트리 개수별 사교육 비용에 대한 예측 오차 변화

이어서 고교생의 사교육 비용에 영향을 주는 설명변수의 중요도를 분석하기 위해 평균제곱오차(Mean Squared Error)와 노드순도를 기준으로 중요도 지수를 산출하였으며, 아래 그림은 중요도가 높은 순으로 상위 10개의 변수를 지수별로 제시한 것이다.

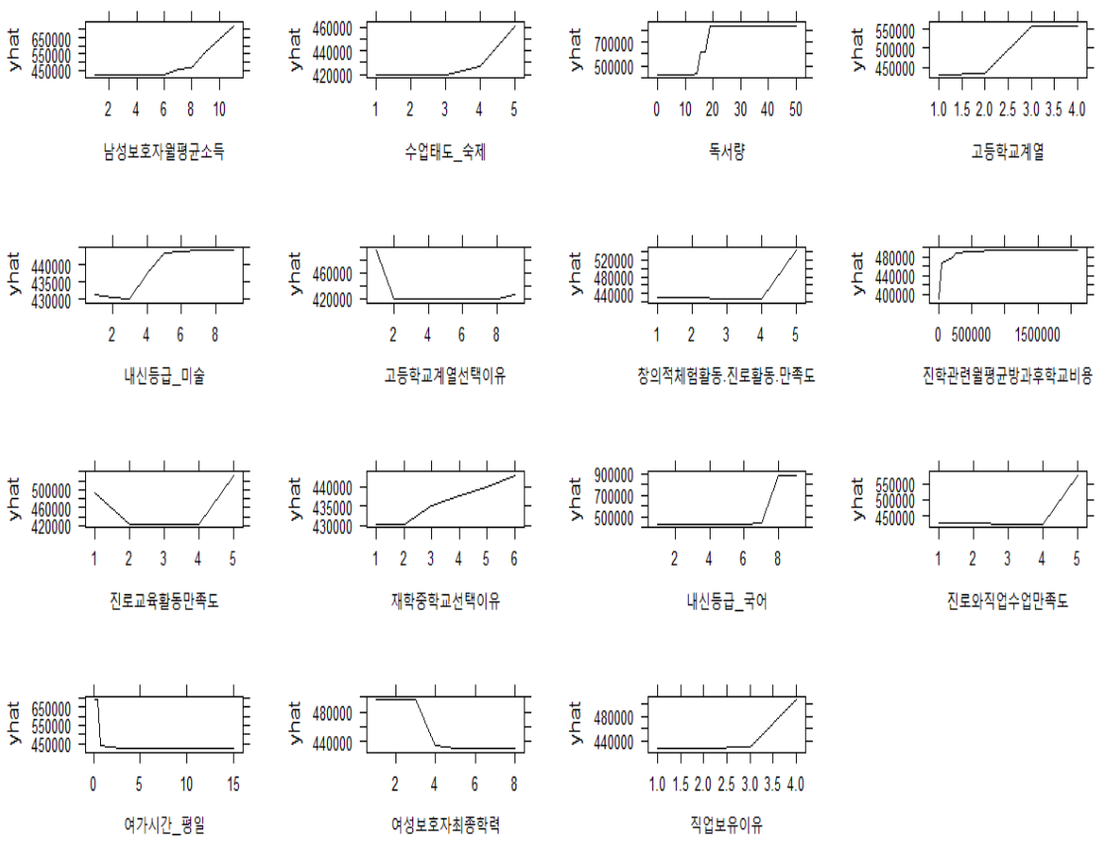


[그림 2] 고교생의 사교육비 지출 관련 중요도 지수 상위 10개 변수

평균제곱오차를 기준으로 보면 '남성보호자 월평균 소득' 문항의 중요도 지수가 가장 높게 나타났다. 다음으로 '수업태도-숙제', '독서량', '고등학교 계열', '내신등급-미술' 문항이 뒤를 이었다. 이어서 '고등학교 계열 선택 이유', '창의적 체험활동 만족도', '진학 관련 월평균 방과후 학교 비용', '진로교육 활동 만족도', '재학중 학교 선택 이유' 문항의 순서로 중요도 지수가 높

은 것으로 확인되었다.

노드순도를 기준으로 보면, ‘내신등급_국어’ 문항의 중요도 지수가 가장 높게 나타났다. 다음으로 ‘독서량’, ‘진로와 직업 수업 만족도’, ‘남성보호자 월평균 소득’, ‘창의적 체험활동 만족도’ 문항이 뒤를 이었다. 이어서 ‘진로교육활동 만족도’, ‘여가시간-평일’, ‘여성보호자 최종학력’, ‘진학 관련 월평균 방과후학교 비용’, ‘직업 보유 이유’ 문항의 순서로 중요도 지수가 높은 것으로 확인되었다.



[그림 3] RF 분석 결과 주요 변수별 부분의존성 도표(Partial Dependence Plot)

다음으로, 확인된 고교생의 사교육 비용 관련 주요 변수에 대해 부분의존성 도표(Partial Dependence Plot)를 작성하였다. 가령 남성보호자의 소득이 높을수록, 그리고 숙제를 꼬박꼬박 잘하는 학생일수록, 독서량이 많을수록 사교육 비용을 높게 예측하였고, 질적 변수인 고등학교 계열에 따라 사교육비 지출을 다르게 예측하는 것으로 나타났다.

2. 고교생의 사교육비 지출 예측 모형 탐색

앞서 살펴본 바와 같이 고교생의 월평균 사교육 비용에 대한 랜덤 포레스트 분석 결과에 기초한 주요 예측 변수들의 기초 통계량은 아래와 같다.

〈표 6〉 주요 변수 기술통계표

구분		사례 수	평균	표준편차	최솟값	최댓값
월평균 사교육 비용		8,622	290838.20	592528.40	1	30000000
남성보호자 월평균 소득		6,692	1.48	0.43	0	2.40
수업태도_숙제		8,557	3.63	0.94	1	5
독서량		5,522	2.11	3.01	0.1	50
고교 계열	문과	4,990	0.51	0.50	0	1
	이과	4,990	0.46	0.50	0	1
	예체능	4,990	0.03	0.16	0	1
(3학년 1학기) 미술 내신 등급		5,546	3.37	1.96	1	9
계열 선택 이유	원하는 직업과 관련	4,985	0.21	0.41	0	1
	원하는 대학 전공 고려	4,985	0.19	0.39	0	1
	나의 적성 고려	4,985	0.20	0.40	0	1
	내신에 유리	4,985	0.14	0.35	0	1
	수능 성적이 잘 나와서	4,985	0.03	0.17	0	1
	선택 계열 과목이 좋아서	4,985	0.09	0.28	0	1
	선택 계열 과목을 잘해서	4,985	0.02	0.14	0	1
	미선택 계열 과목이 싫어서	4,985	0.06	0.23	0	1
	미선택 계열 과목을 못해서	4,985	0.07	0.25	0	1
창의적 체험활동(진로활동) 만족도		6,068	3.50	0.79	1	5
진학 관련 월평균 방과후학교 비용		8,527	64,768.34	195,904.30	1	3000000
진로교육/활동 만족도		8,558	3.43	0.85	1	5
재학중 학교 선택 이유	대학 진학	8,621	0.40	0.49	0	1
	가정 형편	8,621	0.01	0.12	0	1
	성적에 맞추어서	8,621	0.14	0.35	0	1
	적성 고려	8,621	0.18	0.38	0	1
	취업	8,621	0.17	0.38	0	1
	기타	8,621	0.09	0.29	0	1
(3학년 1학기) 국어 내신 등급		8,099	3.88	1.65	1	9
진로와 직업 수업 만족도		6,887	3.45	0.79	1	5
여가시간_평일		8,331	2.92	1.98	0.17	15
여성 보호자 최종 학력	무학	8,203	0.01	0.09	0	1
	초등학교	8,203	0.01	0.11	0	1
	중학교	8,203	0.03	0.16	0	1
	고등학교	8,203	0.57	0.50	0	1
	2-3년제 대학	8,203	0.10	0.30	0	1
	4년제 대학	8,203	0.26	0.44	0	1
	대학원(석사)	8,203	0.02	0.15	0	1
	대학원(박사)	8,203	0.00	0.07	0	1

구분		사례 수	평균	표준편차	최솟값	최댓값
직업 보유 이유	자신과 가족의 생계 유지	8,605	0.47	0.50	0	1
	사회 봉사 및 참여	8,605	0.06	0.23	0	1
	자아실현	8,605	0.38	0.49	0	1
	사회적 인정	8,605	0.09	0.28	0	1

본 연구는 고교생의 사교육 비용에 대한 랜덤 포레스트 분석 결과에 근거하여 <표 6>에 제시된 변수들을 활용하여 OLS, 학교 고정효과 모형 및 학교 확률효과 모형을 통한 사교육비 예측 모형들을 도출하였고, 모형 간의 비교를 통하여 보다 적합한 모형을 확인하고자 하였다.

〈표 7〉 사교육비 예측 모형별 비교

		OLS	Fixed Effect Model	Random Effect Model
남성보호자 월평균 소득		1.449*** (0.364)	1.275*** (0.400)	1.403*** (0.364)
수업태도_숙제		0.264* (0.182)	0.076* (0.201)	0.222* (0.182)
독서량		-0.050* (0.070)	-0.038* (0.077)	-0.042* (0.070)
고교 계열	문과	-2.099** (0.966)	-2.506** (1.088)	-2.130** (0.965)
	이과	-1.760* (0.972)	-2.101* (1.100)	-1.789* (0.972)
(3학년 1학기) 미술 내신 등급		0.042* (0.096)	-0.053* (0.111)	0.025* (0.097)
계열 선택 이유	원하는 대학 전공 고려	-0.167* (0.477)	0.004* (0.514)	-0.130* (0.475)
	나의 적성 고려	0.597* (0.464)	0.965* (0.501)	0.651* (0.461)
	내신에 유리	0.297* (0.526)	0.947* (0.588)	0.441* (0.528)
	수능 성적이 잘 나와서	0.418* (0.960)	0.490* (1.066)	0.395* (0.960)
	선택 계열 과목이 좋아서	-0.977* (0.589)	-0.931* (0.657)	-0.974* (0.588)
	선택 계열 과목을 잘해서	0.386* (1.114)	0.705* (1.157)	0.442* (1.104)
	미선택 계열 과목이 싫어서	0.155* (0.728)	-0.392* (0.827)	0.068* (0.727)
	미선택 계열 과목을 못해서	0.840* (0.749)	0.960* (0.808)	0.849* (0.745)
창의적 체험활동(진로활동) 만족도		-0.009* (0.282)	0.181* (0.309)	0.040* (0.282)
진학 관련 월평균 방과후학교 비용		0.432*** (0.026)	0.451*** (0.029)	0.435*** (0.026)
진로교육/활동 만족도		0.046* (0.200)	0.067* (0.216)	0.061* (0.199)

		OLS	Fixed Effect Model	Random Effect Model
재학중 학교 선택 이유	대학 진학	-0.084* (0.481)	0.449* (0.528)	-0.002* (0.479)
	가정 형편	-0.614* (2.053)	0.402* (2.171)	-0.525* (2.040)
	성적에 맞추어서	-0.178* (0.599)	0.231* (0.659)	-0.131* (0.598)
	적성 고려	-1.381** (0.632)	-0.239* (0.688)	-1.171* (0.630)
	취업	0.166* (1.308)	0.495* (1.437)	0.256* (1.304)
(3학년 1학기) 국어 내신 등급		-0.335*** (0.113)	-0.270** (0.125)	-0.325*** (0.113)
진로와 직업 수업 만족도		-0.107* (0.277)	-0.203* (0.308)	-0.153* (0.277)
여가시간_평일		-0.260*** (0.082)	-0.324*** (0.093)	-0.269*** (0.082)
여성 보호자 최종 학력	초등학교	2.949* (2.975)	4.557* (3.374)	3.400* (2.964)
	중학교	2.529* (2.852)	3.492* (3.227)	2.834* (2.840)
	고등학교	2.303* (2.605)	3.578* (2.990)	2.709* (2.595)
	2-3년제 대학	2.966* (2.630)	4.367* (3.020)	3.399* (2.621)
	4년제 대학	2.912* (2.611)	3.926* (2.995)	3.257* (2.601)
	대학원(석사)	3.068* (2.776)	4.307* (3.176)	3.464* (2.767)
	대학원(박사)	2.487* (3.039)	3.826* (3.469)	2.879* (3.028)
직업 보유 이유	사회 봉사 및 참여	0.066* (0.689)	0.305* (0.755)	0.130* (0.687)
	자아실현	-0.072* (0.332)	-0.109* (0.365)	-0.082* (0.332)
	사회적 인정	0.656* (0.561)	0.615* (0.615)	0.647* (0.560)
상수		5.016* (3.067)	4.139* (3.493)	4.703* (3.058)
n		1,195	1,195	1,195
R2		0.269	0.280	0.275
Lagrangian multiplier test				6.78***
Hausman Test				40.48

* 주: 제시값(β)은 회귀계수이며 ( )안의 값은 표준오차임.
* 범례: *** 〈 .01 ** 〈.05 * 〈.1.

<표 7>의 분석 결과에 따르면 랜덤 포레스트를 통하여 도출된 사교육비 예측 요인 모두가 OLS 및 고정효과 모형, 확률효과 모형에서 사교육비 지출과 통계적으로 유의한 관련성을 맺고 있음을 확인하였다. 변수 간 다중공선성 및 상호작용의 효과를 모두 고려하는 랜덤 포레스

트의 특성을 고려할 때 <표 7>의 분석 결과는 상당히 신뢰할 만 것임을 알 수 있다.

본 연구는 고교생의 사교육비 지출 예측 모형 간의 비교를 위하여 Breusch & Pagan(1980)이 제안한 Lagrange Multiplier(LM) 검정과 Hausman 검정(Hausman test)를 실시하였다. 우선 전체 오차항의 분산에서 학교 특성 오차항의 분산이 차지하는 비중에 대한 LM 검정은 패널 오차항의 개체 특성을 고려할 필요가 있는지에 대한 중요한 정보를 제공해준다. 확률효과 모형을 활용한 학교 고유 특성 유무에 대한 LM 검정 결과, 학교 특성 오차항이 0이라는 귀무가설이 기각되어, 조사 패널들이 속한 학교의 고유한 개체특성을 고려할 필요가 있음을 알 수 있다.

그리고 고정효과 모형과 확률효과 모형에 대한 하우스만 검정 결과 패널의 개체특성과 설명변수 간의 상관관계가 0이라는 귀무가설을 기각할 수 없음에 따라 확률효과모형을 선택하는 것이 바람직함을 알 수 있다. 이에 따르면 고교생의 사교육 비용에 대한 예측 모형을 탐색하는데 있어서는 학교의 고유한 특성을 확률변수로 간주하는 확률 효과 모형(Random Effect Model)이 보다 타당함을 알 수 있다.

Ⅴ. 결론 및 제언

본 연구는 한국교육고용패널Ⅱ 2차년도(2017년) 자료에 대하여 랜덤 포레스트 기법을 활용하여 고교생의 사교육비 지출을 예측하는 요인을 탐색하는 데에 일차적 목적을 두었고, 191개 설명변수를 활용한 모형의 예측성과를 확인한 결과 랜덤 포레스트의 추정 모형이 적합한 모형임을 확인하였다. 다음으로 고교생의 사교육 비용에 영향을 주는 설명변수의 중요도를 확인하여 상위 15개의 설명변수를 도출한 후 해당 변수들을 활용하여 적합한 사교육비 지출 예측 모형을 확인하고자 하였다. 본 연구의 주요 결과를 정리하면 다음과 같다.

우선, 고교생의 사교육 비용에 영향을 미치는 설명변수 중 중요도가 높은 상위 15개 변수를 평균제곱오차와 노드불순도 지수를 활용하여 확인한 결과, '남성보호자 월평균 소득', '수업태도-숙제', '독서량', '고등학교 계열', '내신등급-미술', '고등학교 계열 선택 이유', '창의적 체험활동 만족도', '진학 관련 월평균 방과후학교 비용', '진로교육 활동 만족도', '재학중 학교 선택 이유', '내신등급_국어' '진로와 직업 수업 만족도', '여가시간-평일', '여성보호자 최종학력', '직업 보유 이유' 등의 문항이 사교육비에 대한 예측력이 높은 것으로 나타났다. 이를 통해 볼 때 고교생의 사교육비 지출에는 부모의 소득 및 학력과 같은 SES의 영향력이 여전히 높은 것으로 나타났다. 이러한 분석 결과는 많은 이들이 우려하는 바와 같이 부모의 SES에 따른 사교육의 양극화, 이로 인한 학력 격차의 유지 및 양극화가 심각한 문제로 대두될 수 있음을 알 수 있다. 이에 다양한 사교육 대체 활동(EBS, 방과후학교, 학교 교육 강화 등)을 통하여 부모의 SES에 따른 사교육 격차가 학력의 격차로 이어지는 가능성을 줄이는 데 적합한 정책적 노력을 강구할 필요가 있음을 알 수 있다.

한편, 랜덤 포레스트 분석 결과 도출된 사교육비 지출 관련 주요 변수들을 활용하여 OLS 및 고정효과 모형, 확률효과 모형을 통하여 보다 적합한 고교생의 사교육비 지출 모형을 확인한 결과 학교 확률효과 모형의 적합성이 가장 높은 것으로 나타났다. 최근 들어 증거 기반(evidence based) 의사 결정의 중요성이 정책 결정 과정 및 교육 거버넌스 과정에서 높아져 감에 따라 사교육비를 포함한 다양한 교육 현상에 대한 보다 합리적인 추정 및 이에 근거한 정책 대안 수립의 필요성 또한 확대되고 있는 추세이다. 이에 본 연구의 분석 결과 혹은 접근 방식을 보다 정교화함으로써 이를 효과적으로 활용할 경우 현재 우리나라 학부모와 학생들이 고통받고 있는 사교육과 관련된 원인을 보다 실증적으로 파악하는 한편, 이의 해소를 위한 정책 대안 수립을 위한 기초자료로 활용될 수 있을 것으로 보인다.

본 연구는 최근 들어 주목받고 있는 대표적인 머신러닝 기법 중 하나인 랜덤 포레스트를 활용하여 고등학생들의 사교육비 지출과 관련된 예측 요인을 도출하고 주요 변수들의 예측력을 비교하고, 예측 모형 간의 비교를 통해 향후 사교육 수요 관련 연구에 있어 머신러닝 기법의 활용 가능성을 확인하였다는 점에서 의의를 찾을 수 있다. 기존 사회과학의 실증분석은 회귀분석에 기초한 인과관계 추정이 핵심을 이루어 왔다. 이로 인해 예측 및 분류는 상대적으로 관심의 대상에서 벗어나 있었던 것이 사실이나, 정책 성과에 대한 예측 및 분류는 해당 정책의 성패를 선험적으로 가늠해 봄으로써 실패 확률을 줄이고 재정 효율성을 제고할 수 있다는 점에서 상당한 의의가 있다고 볼 수 있다. 따라서 교육학 분야에서도 머신러닝을 활용한 실증분석이 향후 활발히 이루어질 필요가 있다.

한편, 본 연구는 다음과 같은 제한점들을 가지고 있어 아래와 같이 후속연구 제안을 통해 이를 개선해 가고자 한다. 첫째, 다층 자료, 종단자료, 인과분석 등과 연계하여 정교한 머신러닝 기법을 적용할 필요가 있다. 본 연구는 한국교육고용패널 Ⅱ의 2차년도 자료, 즉 횡단면 자료를 기반으로 분석을 실시하였다. 해외에서는 이미 다층 자료 및 종단 자료에 대해 랜덤 포레스트 기법을 적용한 연구들도 탐색적으로 실시되고 있음을 고려할 때(Hajjem et al., 2014; Huang et al., 2016), 향후 다년도 자료를 사용하여 더욱 정교하고 엄밀한 머신러닝 기법 활용 연구가 수행될 필요가 있을 것으로 보인다. 한편, 사교육비에 대한 보다 정확한 추정은 설명변수와 사교육비 간의 인과관계에 근거할 때 가능하다. 본 연구에서 활용한 랜덤 포레스트는 실질적으로 변수 간의 인과관계 탐색을 위한 분석방법이기보다는 예측(prediction)에 보다 주안점을 두고 있는 분석방법이다. 이에 다음에는 머신러닝 기법을 활용한 인과성 추정 기법에 대한 심도 있는 고민과 활용을 통해 예측력이 강화된 사교육비 추정 모형을 도출할 수 있기를 기대해본다.

둘째, 사교육 관련 데이터를 보강하여 실증적으로 규명해 왔던 사교육비 결정요인 모델을 정교하게 보완하려는 노력이 필요하다. 데이터의 한계로 인해 본 연구는 초중등단계에서 사교육 참여율이나 사교육비 규모와 관련한 영향 요인들을 실증적으로 규명했던 다른 연구들과 마찬가지로(김현진, 2004; 성낙일, 홍성우, 2008; 송경오, 이광현, 2010), 학원 특성(학원 지도 과목, 강사 질, 수업 질 등)이나 지역 여건(ex. 지역 사교육 시장) 등과 같이 사교육비 지출 규모

에 영향을 미치는 요인들을 종합적으로 고려하지 못하였다. 현재 통계청에서 조사하는 '초중고 사교육비 조사'는 사교육비 규모를 정확하게 파악하는 데 조사의 방점을 두고 있어, 사교육비 규모를 결정하는 요인을 규명하는 데 현실적인 어려움이 있다. 실제 정책 담당자들은 조사 문항 개선, 표집 대상 확대 등의 방식으로 '초중고 사교육비 조사'를 개선하기 위해 노력하고 있으나, 이런 접근 방식은 사교육비 지출의 실태 파악을 넘어서 사교육비 증가의 원인을 밝혀내는 데 적합한 정보를 충분히 수집하기에는 제한적이다. 따라서 향후 정확한 사교육비 실태 파악과 사교육비 증가 원인을 종합적으로 분석할 수 있는 사교육비 조사 체계를 마련할 필요가 있다.

참고문헌

강소랑(2016). 고교 평준화가 사교육에 미치는 영향. 2016 한국정책학회 추계학술대회 겸 국제학술대회 자료집(pp. 5~29.).

강태중(2009). 고등학교 '평준화' 배정과 경쟁 선발이 사교육비 지출에 미치는 영향 분석. 교육사회학연구, 19(2), 1~30.

교육과학기술부(2009.6.3.). 공교육 경쟁력 향상을 통한 사교육비 경감 대책. 보도자료.

교육인적자원부(2004.2.17.). 공교육 정상화를 통한 사교육비 경감대책. 보도자료.

권민지(2019.3.12.). 초·중·고 사교육비 매년 최고치 경신. 경기매일(주소: http://www.kgmaeil.net/news/articleView.html?idxno=208482, 접속일: 2019.09.20.)

김성식, 송혜정(2009). 학교 불만족과 특목고 진학 경쟁이 사교육 시간과 비용의 변화에 미치는 영향. 교육사회학연구, 19(4), 21~46.

김위정, 김양분(2013). 대입 입학사정관 전형 지원 계획이 사교육비 지출에 미친 영향 분석. 교육사회학연구, 23(4), 86~117.

김현진(2004). 사교육비 지출 결정 변인 구조 분석. 교육행정학연구, 22(1), 27~45.

김현진, 최상근(2004). 고교평준화제도와 사교육비 지출의 관계 분석. 한국교육, 31(1), 365~383.

김화경(2017). 경향점수를 이용한 학교급별 사교육 요인 분석: 수학 사교육을 중심으로. 학습자중심교과교육연구, 17(7), 49-66.

김희삼(2009). 사교육비 지출에 영향을 주는 학교특성의 분석. 노동경제논집, 32(3), 27~59.

김희삼(2010). 학업성취도, 진학 및 노동시장 성과에 대한 사교육의 효과 분석. 한국교육개발원 연구보고서 2010-05. 서울: 한국교육개발원.

문지영, 김현철, 박혜연(2018). 사교육비 및 사교육참여율에 대한 방과후학교의 효과. 교육행정학연구, 36(1), 329~354.

문지영, 모은비, 서은경, 조정우(2018). 벌점화 회귀모형을 사용한 사교육비 관련요인 탐색. 한국교육, 45(1), 111~137.

박균달, 김현진(2011). 사교육 효과와 원인에 관한 메타 분석. 교육논총, 31, 75~104.

박선영, 마강래(2015). 지역의 교육환경이 사교육비 지출에 미치는 영향에 관한 연구. 지역연구, 31(3). 3-17.

박소영(2008). 방과후 학교와 EBS 수능강의의 사교육비 경감 효과. 교육행정학연구, 26(1), 391~411.

손수람(2019.1.21.). 10년만에 사교육 참여 상승반전.. 모든 학교급 일제히 상승. 베리타스 알파(주소: http://www.veritas-a.com/news/articleView.html?idxno=141533, 접속일: 2019.09.20.)

송경오(2013). 사교육 수요에 영향을 주는 학교 특성에 대한 메타분석적 접근. 교육과학연구, 44(1), 1~29.

송경오, 이광현(2010). 일반계 고등학교 학생의 사교육 수요에 영향을 미치는 학교교육 특성에 대한 패널분석. 교육행정학연구, 28(4), 301~326.

신혜진(2017). 고교유형별 서울시 학부모의 사교육비 지출의 종단적 분석. 교육행정학연구, 35(4), 259~285.

심은석, 박균달, 김현진(2013). 서울시 초·중·고등학교 학생의 방과후학교 참여가 사교육비 경감에 미치는 효과. 중등교육연구, 61(2), 361~388.

유재봉, 조정우, 서은경, 김현철(2016). 정책변화에 따른 사교육 시장의 변화 연구. 서울: 성균관대학교 사교육혁신 교육연구소.

유진은(2015). 랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법. 교육평가연구 28(2), 427-448.

이광현(2013). 사교육 경감 정책 효과 분석. 교육사회학연구, 23, 111-138.

이수정(2011). 대입제도의 변화가 사교육비 지출에 미친 영향 분석. 교육재정경제연구, 20(1), 127~147.

이수정, 민병철(2009). 사교육 수요와 학업성과에 영향을 주는 학교 특성 분석. 교육재정경제연구, 18(4), 179~206.

이수정, 조원기(2014). 대입전형에서 학생부 내신 반영 강화 정책과 고고생의 사교육비 지출간의 관련성 분석. 고용직업능력개발연구, 17(3), 125~150.

이필남(2011). 대학 입학사정관전형 지원 계획과 사교육 수요 관계 분석. 교육재정경제연구, 20(4), 125~151.

이혜정, 송종우(2014). 초,중,고 사교육비 영향요인 분석. 응용통계연구, 27(7), 1125-1137.

장지윤, 박인우, 장재홍(2017). EBS 방송 시청이 인문계 고등학생의 사교육비 지출에 미치는 영향. 한국교육문제연구, 35(3), 129~150.

정동욱, 박현정, 하여진, 박민호, 이호준, 한유경(2012). EBS 교육 프로그램의 사교육 경감 효과 분석: 서울특별시 중·고등학교를 중심으로. 교육행정학연구, 30(3), 21~42.

채재은, 임천순, 우명숙(2009). 방과후학교와 수능강의가 사교육비 및 학업성취도에 미치는 효과 분석. 교육재정경제연구, 18(3), 37~62.

채창균(2006). 고교평준화가 사교육비 지출에 미친 영향에 대한 실증분석. 교육사회학연구, 16(2), 163-179.

최필선, 민인식(2018). 머신러닝 기법을 이용한 대졸자 취업예측 모형. 직업능력개발연구, 21(1), 31-54.

통계청(2019.3.11.). 2018년 초중고 사교육비 조사 결과. 보도자료.

하준경(2010). 특목고가 가계의 사교육비 지출에 미치는 영향. 경제분석, 16(3), 156~191.

한국일보(2019.3.13.). 사교육비 역대 최고치 기록, 문재인 정부 대책 있기는 한가. 한국일보 사설. (주소: https://www.hankookilbo.com/News/Read/201903121622349578, 접속일: 2019.09.20.)

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

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. The review of economic studies, 47(1), 239-253.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313-1328.

Huang, L., Jin, Y., Gao, Y., Thung, K. H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2016). Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. Neurobiology of aging, 46, 180-191.

Rea, L. M., & Parker, R.A. (2005). Designing & Conducting Survey Research A Comprehensive Guide (3rd Edition). San Francisco, CA: Jossey-Bass.