

아시아교육연구 21권 1호

Asian Journal of Education

2020, Vol. 21, No. 1, pp. 191-215.

<https://doi.org/10.15753/aje.2020.03.21.1.191>

랜덤 포레스트를 활용한 읽기소양 수준에 따른 집단 결정요인 분석: PISA 2018 자료를 중심으로

손윤희(孫倫嬉)*

박현정(朴炫貞)**

박민호(朴民鎬)***

논문 요약

이 연구는 학생들의 읽기소양 수준에 따른 집단 분류에 주요한 영향을 미치는 변인을 탐색하는 데 목적이 있다. 이를 위하여 읽기소양 수준에 따른 집단을 일반집단, 기초수준 이하 집단으로 구분하였고, 랜덤 포레스트를 활용하여 집단 분류에 영향을 미치는 주요한 변수를 탐색하였다. 분석을 위하여 PISA 2018 자료를 이용하여 168개교에 재학 중인 6,630명의 학생을 연구대상으로 선정하였고, 학생, 학부모, 교사, 학교와 관련된 286개의 변수를 설명 변수로 모형에 포함하였다. 분석 결과, 기초수준 이하 집단 대비 일반집단으로 분류되는 데 있어 학생과 관련된 변수가 주요한 역할을 하는 것으로 나타났다. 구체적으로, 다양한 읽기전략과 관련된 변수, 읽기 활동에 대한 경험, 빈도 및 인식, PISA 검사 및 조사에 대한 어려움 및 노력 등이 주요한 역할을 하였다. 그 외에도 개인특성 및 가정배경이 집단 분류에 영향을 미쳤다. 또한, 이 연구에서는 랜덤 포레스트를 활용함으로써 변수들 간의 복잡한 관계를 고려하여 집단 분류와 결정 요인간의 비선형적 관계를 파악하였다. 마지막으로, 연구 결과를 토대로 학생의 읽기소양을 향상시키기 위해 필요한 교육적 지원을 제안하였다.

주요어 : 읽기소양, 기초수준 이하 집단, 집단 분류, 랜덤 포레스트, PISA 2018

* 1저자, 한국교육과정평가원 부연구위원(위촉)

** 공동저자, 서울대학교 교육학과 교수

*** 교신저자, 서울대학교 교육학과 박사수료, pmino85@snu.ac.kr

1. 서론

기술의 급격한 발전에 따라 현대 사회를 살아가고 있는 우리는 인터넷 서비스와 디지털 기기를 어려움 없이 이용하고 있다. 이는 가정에서 인터넷을 사용할 수 있는 학생의 비율이 지난 10년 동안 10%p 이상 증가하여, 2018년 기준으로 청소년의 95%가 가정에서 인터넷을 어려움 없이 사용한다는 결과를 통해서도 확인할 수 있다(OECD, 2019a). 따라서 청소년들은 인쇄된 매체뿐만 아니라 전자 매체를 통해 수많은 정보를 쉽게 접하고, 여러 경로를 통해 다양한 사람들과 의사소통을 하며 살아가고 있다. 이러한 과정에서 청소년들은 사실과 거짓이 혼재된 정보를 쉽게 접하고, 이러한 정보를 무비판적으로 수용할 수 있는 위험에 노출되어 있다. 따라서 학생들이 수많은 정보 속에서 자신의 목적에 맞는 정보를 선별하고, 선별된 정보 중에서 신뢰할 수 있는 정보를 판단할 수 있는 능력, 즉 읽기소양(reading literacy)을 갖추는 것이 중요하다.

읽기소양의 중요성이 강조됨에 따라 그 동안 읽기소양에 관한 연구가 여러 맥락에서 수행되어 왔다. 첫 번째 맥락에서는 다수의 연구에 의해 읽기소양과 개인특성요인, 가정, 교사 및 학교의 맥락 변인과의 관계가 탐색되었고(강대중, 염시창, 2013; 김혜숙, 2012; 송미영, 임현정, 김성숙, 2015; 이인화, 구남욱, 2019), 두 번째 맥락에서는 학생의 읽기소양 수준 또는 학생이 처한 환경에 따라 집단을 구분한 후, 보다 성취수준이 높은 집단 또는 역경을 극복하고 높은 성취를 보이는 집단에 속할 수 있는 요인이 탐색되었다(박현정, 하여진, 2011; 송미영, 김성숙, 이현숙, 김준엽, 2011; 정혜경, 조지민, 2013). 그 외에도 읽기소양과 교육과정을 비교하거나 평가방법에 따라 성취수준을 비교한 연구가 수행되었다(박혜영, 송미영, 2014; 이은하, 최은정, 2015).

이 연구에서는 기존 연구의 두 번째 맥락을 따라 읽기소양이 기초수준 이하인 집단과 그렇지 않은 집단을 분류하는 주요 변수를 탐색하고자 한다. 이는 읽기소양이 기초수준 이하인 학생들에게 적절한 교육적 지원이 제공되지 않을 경우, 학습 결손이 누적되어 학생들이 미래 사회를 살아가는 데 필요한 소양을 기르는 데 더 큰 어려움을 겪을 수 있기 때문이다. 같은 맥락에서 선행 연구는 기초학력 미달인 학생 또는 이 학생들의 비율이 높은 학교에 초점을 맞추어, 기초학력 미달 학생들의 성취를 향상시키기 위해 필요한 개인 및 가정 차원의 보호요인과 학교 및 정책 차원의 교육적 지원을 탐색하여 왔다(권재기, 2013; 김태은, 이화진, 오상철, 노원경, 2012; 송미영 외, 2011; 이효정, 김광주, 김성훈, 2018; 전명남, 양명희, 2015). 더욱이, 최근에는 ‘기초학력 지원 내실화 방안’이 발표됨에 따라 초1~고1 학생들의 기초학력을 진단하고, 그 결과를 토대로 학생지원 강화 방안을 모색하듯이, 성취수준이 낮은 학생에 대한 교육적 지원의 중요성이 강조되고 있다(교육부, 2019.3.28.).

이러한 관점에서 읽기소양 수준에 따른 성취집단을 결정하는 요인을 파악하여, 이를 토대로 학생들에게 적절한 교육적 개입을 제공함으로써 학생이 사회에서 살아가는 데 필요한 읽기소양을 증진시키는 것을 기대할 수 있다. 이때, 결정요인을 파악하는 데 있어 사용할 수 있는 정보를 최대한

많이 활용할 수 있다면, 교육맥락변인 간의 복잡한 관계를 고려하여 실제 상황을 보다 잘 반영할 수 있다. 이러한 맥락에서 의사결정나무(decision tree)는 하나의 모형에 많은 변수를 포함하고, 이 변수들 간의 비선형적 관계를 고려할 수 있다는 점에서 이점을 갖는다(김미림, 박민호, 2019). 하지만, 의사결정나무 모형은 주요 변수를 선택할 때, 각 노드(node)에서 한 변수만 사용하기 때문에 분석에 사용된 자료가 바뀌면 다른 결과를 나타내어 안정성이 떨어지는 제한점을 지닌다(유진은, 2015). 반면, 랜덤 포레스트(random forest)는 자료를 선택할 때뿐만 아니라 각 노드에서 변수를 선택할 때 무작위성을 반영하여 의사결정나무 모형의 한계를 보완하는 이점을 지닌다.

이러한 점에 근거하여 이 연구에서는 랜덤 포레스트를 활용하여 학생들의 읽기소양 수준에 따라 구분된 집단에 대한 결정요인을 파악하는 데 목적을 둔다. 구체적으로, PISA 2018 자료를 활용하여 읽기소양 수준에 따라 일반집단과 기초수준 이하 집단으로 구분한 후, 두 집단을 분류하는 데 주요한 영향을 미치는 요인을 추출하고자 한다. 특히, 랜덤 포레스트의 활용은 학생, 학부모, 교사, 학교와 관련된 많은 변수를 포함할 수 있기 때문에 보다 복잡한 구조에서 읽기소양 집단 결정에 주요한 영향을 미치는 변수가 무엇인지 파악할 수 있다. 이로부터 기존 연구 결과와 종합하여 학생들이 보다 더 높은 읽기소양을 갖추는 데 필요한 교육적 지원이 무엇인지 시사점을 도출하고자 한다.

II. 이론적 배경

1. 읽기소양에 대한 결정요인 분석 연구

PISA 2018에서는 읽기소양(reading literacy)을 “자신의 목적을 성취하고, 지식과 잠재적 능력을 계발하며, 사회에 참여하기 위해 텍스트를 이해·활용·평가·성찰하고, 다양한 텍스트 읽기 활동에 관여하는 능력”으로 정의하였다(조성민, 구남욱, 김현정, 이소연, 이인화, 2019; OECD, 2019b).

먼저, 읽기소양에 관한 결정요인에 관한 선행연구의 주요한 결과를 개인특성과 맥락변인에 초점을 맞추어 정리하면 다음과 같다. 개인특성 요인의 결과를 살펴보면, 많은 선행연구에서 여학생이 남학생보다 더 높은 읽기소양을 나타내고 있음을 확인하였다(강대중, 2015; 강대중, 염시창, 2013; 김혜숙, 2012; 박주현, 장우권, 2014; 송미영, 임현정, 김성숙, 2015; 이인화, 구남욱, 2019; 정혜경, 조지민, 2013). 한편, 읽기 활동에 대한 즐거움 및 흥미와 같은 긍정적인 태도를 가진 학생(강대중, 2015; 강대중, 염시창, 2013; 김혜숙, 2012; 박주현, 장우권, 2014; 정혜경, 조지민, 2013)이, 요약 및 이해와 같은 읽기 전략이 높은 학생(강대중, 2015; 강대중, 염시창, 2013)이, 읽기 활동에 더 많이 참여하는 학생(강대중, 2015; 강대중, 염시창, 2013)이, 온라인 자료와 같이 다양한 자료를 읽는 학생(김혜숙, 2012; 박현정, 하여진, 2011)이 더 높은 읽기소양을 보여, 학생수준에서는 읽기 활동과

관련된 정의적 영역과 적극적인 활동이 주요한 영향을 미치는 것으로 나타났다. 또한, 학교와 관련된 요인으로는 방과후 학교에 참여한 학생(정혜경, 조지민, 2013)이, 지각을 하지 않은 학생(이인화, 구남옥, 2019)이, 그리고 국어교과 수업 시수가 더 많은 학교에 재학하는 학생(이인화, 구남옥, 2019)이 더 높은 읽기소양을 보이는 것으로 나타났다.

다음으로, 가정요인으로는 가정의 경제사회문화적 배경이 학생의 읽기소양에 주요한 영향을 미치고 있음을 다수의 연구를 통해 확인하였다(강대중, 2015; 강대중, 염시창, 2013; 김혜숙, 2012; 박현정, 하여진, 2011; 송미영, 임현정, 김성숙, 2015; 이인화, 구남옥, 2019). 그 외에도 가정에서 보유한 도서와 독서자원이 많을 경우(박주현, 장우권, 2014), 자녀가 재학하는 학교의 질에 대한 부모의 인식이 긍정적일수록(송미영, 임현정, 김성숙, 2015) 학생의 읽기소양이 더 높게 나타났다.

마지막으로, 교사와 학교와 관련된 요인으로는, 학교 또는 수업 분위기를 저해하는 학생의 방해요인이 적을수록(이인화, 구남옥, 2019; 정혜경, 조지민, 2013), 교사와의 관계가 높을수록(강대중, 2015; 송미영, 임현정, 김성숙, 2015), 교사 당 학생 수가 적을수록(강대중, 염시창, 2013) 학생의 읽기소양이 더 높은 것으로 나타났다. 또한, 읽기활동에 직접적인 영향을 미치는 환경으로 학교 도서관에서 보유한 도서의 수가 적거나 직원이 부족하여 업무에 지장이 많을 경우는 학생들의 읽기소양 성취에 방해요인으로 작용하는 것으로 나타났다(박주현, 장우권, 2014). 마지막으로, 학교에 평균적으로 가정의 사회경제적수준이 높은 학생이 재학하거나 읽기에 대한 즐거움이 높은 학생이 재학할 경우에 그 학교의 평균적인 읽기소양 향상에 정적인 영향을 미치는 것으로 나타났다(강대중, 2015; 강대중, 염시창, 2013; 김혜숙, 2012; 송미영, 임현정, 김성숙, 2015; 이인화, 구남옥, 2019).

다음으로, 읽기소양 수준에 따라 학생들을 집단으로 구분하고, 집단에 대한 결정요인을 분석한 주요 연구를 살펴보면 다음과 같다. 박현정과 하여진(2011), 송미영 외(2011)는 PISA 2009 자료를 활용하여 읽기소양의 수준에 따라 세 집단으로 구분한 후, 보다 높은 성취수준에 도달할 수 있는 요인을 파악하였다. 분석 결과를 살펴보면, 학생과 관련된 요인의 경우, 남학생보다 여학생이, 읽기에 대한 자기조절학습전략과 다양한 자료의 읽기 경험, 학교 및 학습활동에 대한 태도가 높을수록, 결석 및 결과 횡수가 낮을수록 읽기소양의 수준이 1수준 이하에서 2수준 이상으로 도달될 확률이 높은 것으로 나타났다. 그 외에도 부모의 교육기대수준, 교사 1인당 학생 수, 교과 외 창의적 활동 등과 관련된 맥락변인이 관련성을 보였다. 읽기소양의 수준이 4수준에서 5수준으로 도달되기 위해서 역시 학생의 성별, 읽기에 대한 자기조절학습전략, 다양한 자료의 읽기 경험, 학교 및 학습활동에 대한 태도가 영향을 미치는 것으로 나타났으며, 그 외에도 가정문화자본, 문제해결에 대한 개방성, 교사와 학생 관계 등이 추가적으로 관련성을 갖는 것으로 나타났다.

또한, 정혜경과 조지민(2013)은 사회경제적·문화적 지수와 읽기 성취의 잔차(residuals)를 활용하여 교육환경과 학업성취가 모두 취약한 집단과 비교하여 역경을 극복하고 높은 학업 성취를 보인 집단을 비교하였다. 분석한 결과, 학생들의 메타인지와 읽기에 대한 즐거움이 높을수록, 학업분위기

에 대하여 학생 간 방해가 적을수록, 여가목적으로 컴퓨터를 더 적게 사용할수록, 방과후 학교에 참여하는 학생이, 사회경제적·문화적으로 낮은 교육 환경을 극복하고 높은 성취를 보이는 것으로 나타났다.

2. 예측모형으로서의 랜덤 포레스트

Breiman(2001)에 의해 개발된 랜덤 포레스트는 분류·예측 기법으로, 기본적으로 의사결정나무 기법을 바탕으로 한다. 랜덤 포레스트는 다수의 붓스트랩(bootstrap) 표본으로 다수의 의사결정나무를 만들고 이를 종합하여 결과를 도출하는 기법이다. 따라서 랜덤 포레스트에 대한 설명에 앞서, 의사결정나무에 대한 언급이 필요하다. 의사결정나무는 의사결정규칙(decision rule)을 나무의 형태로 도표화하여 관심 대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 집단을 예측(prediction)하는 방법이다(최종후, 서두성, 1999). 의사결정나무 분석은 의사결정규칙(나무) 생성, 가지치기(pruning), 타당성 평가, 해석의 절차로 이루어진다(강현철, 서두성, 최종후, 1998; James, Witten, Hastie, & Tibshirani, 2013). 먼저, 의사결정규칙 생성 단계에서는 목적과 자료의 특성에 따라 분리 기준(split criterion)과 정지 규칙(stopping rule)으로 의사결정 규칙을 지정하여 나무를 생성한다. 이후에 분류 오류를 높이는 가지를 제거(pruning)한 후, 최종적으로 오류를 최소화하는 교차타당화(cross validation) 절차를 통해 반응변수와 예측변수를 가장 잘 연결하는 나무를 구축한다(김미림, 박민호, 2019).

이와 같은 의사결정나무 방법은 판별 분석, 회귀 분석, 신경망 분석 등과 같은 다른 예측 방법들에 비하여 비교적 분석이 간단하고, 나무구조를 통해 시각적으로 표현할 수 있기 때문에 연구자가 분석 과정을 쉽게 이해하고 설명할 수 있는 이점을 지닌다(김태진, 홍정식, 전운수, 박종률, 안태욱, 2018; 최종후, 서두성, 1999). 하지만 의사결정나무는 높은 설명력에 비하여 예측력이 떨어지고, 전체 자료를 훈련 자료(training data)와 시험 자료(testing data)로 나누어 교차타당화를 거치는 데도 불구하고 모형이 불안정하다는 등의 문제점이 알려져 있다(유진은, 2015; Friedman, Hastie & Tibshirani, 2000). 또한 나무의 크기가 커질수록 즉, 예측에 사용되는 변수가 늘어날수록 해당 표본을 잘 설명하나 다른 표본을 충분히 설명하지 못하는 과적합(overfitting) 문제를 지닌다. 따라서 나무의 크기나 가지치기 방식을 결정하는 데 있어 주의가 필요하다(Dietterich, 1995; Friedman et al., 2000; Rokach & Maimon, 2008).

이러한 의사결정나무 방법의 단점을 극복하기 위하여 Breiman(1996)은 하나가 아닌 다수의 의사결정나무를 생성하여 이를 종합하여 결과를 도출하는 앙상블 방법으로 배깅(bagging; bootstrap aggregating) 방법을 제시하였다. 배깅은 전체 자료에서 무작위 복원추출로 다수의 붓스트랩 표본을 생성하고 각각의 붓스트랩 표본으로부터 얻어진 다수의 의사결정나무의 결과를 종합하여 최종적인

결론을 도출한다. 이와 같이 무작위로 추출된 붓스트랩 표본으로부터 생성된 의사결정나무의 종합은 의사결정나무의 분산을 감소시켜 예측오차를 줄이며, 예측오차는 각각의 의사결정나무들 간 상관관이 낮을수록 더 작아진다(Breiman, 1996).

랜덤 포레스트는 표본을 무작위로 추출하는 배깅 방식에서 한발 더 나아간다. 배깅에서는 각 노드마다 모든 예측변수 안에서 최적의 분할(split)을 선택하는 방법을 사용하는 반면, 랜덤 포레스트는 각 노드에서 예측변수들을 임의로 추출하고, 추출된 변수 내에서 최적의 분할을 만들어 나가는 임의 노드 최적화(RNO; randomized node optimization) 절차를 추가한다. 따라서 랜덤 포레스트는 배깅 방법보다 의사결정나무들 간 상관을 낮춤으로써 예측오차를 더 줄이는 방식을 취한다(Breiman, 2001). 이러한 방식으로 인하여 랜덤 포레스트는 잡음(noise)이나 이상치(outlier)로부터 크게 영향을 받지 않는 강건함(robustness)을 가지며(Hamza & Larocque, 2005), 나무의 수를 늘릴수록 의사결정나무가 가지는 과적합 문제를 극복할 수 있다는 장점이 있다. 또한, 임의 노드 최적화 과정에서 여러 예측변수를 무작위로 조합하기 때문에 본 연구와 같이 예측변수가 많은 경우에도 안정적인 결과를 도출하며(Caruaana, Karampatziakis, & Yessenalina, 2008; Dudoit, Fridlyand & Speed, 2002), 특정 예측변수가 강력한 영향력을 가지더라도 각각의 의사결정나무에는 예측변수가 무작위로 포함되기에 이 변수의 영향력에 좌우되지 않고 다양한 예측변수의 영향을 분석할 수 있다는 장점을 가진다(James et al., 2013). 또한, 랜덤 포레스트는 의사결정나무 방식처럼 전체 자료를 훈련 자료와 시험 자료로 나누어 교차타당화하는 절차를 거치지 않는다. 따라서 붓스트랩 표본에 속하지 않은 Out-of-bag(이하 OOB) 자료를 시험자료 대신 활용할 수 있어 편의성을 지닌다(유진은, 2015).

한편, 랜덤 포레스트는 하나가 아닌 다수의 의사결정나무를 통해 최종적인 결과를 도출하기 때문에 직관적인 도표로 결과를 파악할 수 있었던 의사결정나무의 장점이 사라진다. 도표를 통한 직관적인 이해가 어려워진 대신 랜덤 포레스트에서는 중요도 지수(importance index)를 통해 예측변수의 영향력을 수치로 제시하고, 부분 의존성 도표(partial dependence plot)를 통해 각각의 예측변수가 반응변수에 어떻게 영향을 미치는지 제시한다(유진은, 2015). 예측변수가 예측 정확도에 기여하는 정도를 나타내는 중요도 지수는 붓스트랩 표본을 추출하였을 때, 표본에 속하지 않은 자료인 OOB 자료를 활용하여 산출되며, 각각의 의사결정 나무에 대해 원래 OOB 자료의 오분류율(error rate for classification) 평균과 해당 변수를 제거한 OOB 자료의 오분류율 평균의 차이를 표준오차로 나누어 표준화한 수치이다(Hastie et al., 2009). 또한, 부분 의존성은 모형에서 해당 예측변수 외의 나머지 예측변수들의 효과를 제외한 해당 변수만의 순수한 반응변수에 미치는 효과를 보여준다(Liaw & Wiener, 2002). 랜덤 포레스트 맥락에서는 각 변수의 수준에 따른 분류 확률 변화를 나타내어 분류 확률에 해당 변수가 상대적으로 기여하는 정도를 의미한다(Welling et al., 2016).

III. 연구 방법

1. 분석 자료

PISA는 만 15세 학생을 대상으로 사회 경제적 삶에서 살아가는 데 필요한 핵심 지식과 기술을 평가하는 조사로, 학생들이 배운 지식을 학교 내 또는 학교 밖의 상황에 얼마나 적용할 수 있는지를 조사하는 연구이다(OECD, 2019b). PISA는 2000년부터 3년 주기로 주 영역을 변경하며 미래 사회에서 필요한 핵심 역량을 조사한다. PISA 2018은 읽기소양을 주 영역(major domain)으로, 수학, 과학, 글로벌 역량을 보조 영역(minor domain)으로 평가하고, 추가적으로 경제적 소양을 조사하였다. 또한, PISA 2018은 2단계 층화 표집방법에 근거하여 목표 모집단인 만 15세 학생들을 표집하였고, 한국의 경우에는 188개교에 재학 중인 6,650명의 학생을 대상으로 조사를 실시하였다. 이 연구에서는 6,650명의 학생 중에서 분석에 사용된 주요 변수에 무응답하거나 체계적으로 결측값을 가진 20명을 제외하여 최종적으로 6,630명의 학생을 연구 대상으로 선정하였다.

2. 주요 변수

1) 결과 변수

PISA 2018은 학생들의 읽기소양 수준을 8개의 수준(1c~1a수준, 2~6수준)으로 구분하는데(OECD, 2019b), 이 연구에서는 읽기소양 수준에 따라 학생들을 일반집단과 기초수준 이하 집단으로 구분하였다.

〈표 1〉 학생들의 읽기소양 수준에 따른 집단 비율

| 집단 | 읽기소양 수준 | 분석 대상 기준 | OECD 기준 |
|------------|---------|----------|---------|
| 일반집단 | 3수준 이상 | 65.7% | 65.3% |
| 기초수준 이하 집단 | 2수준 이하 | 34.3% | 34.7% |

주1. 랜덤 포레스트를 활용하는 데 있어 가중치를 적용할 수 없어 분석 대상 기준의 빈도분석은 가중치를 적용하지 않은 결과임.

구체적으로, 〈표 1〉에 제시된 바와 같이 읽기소양 수준이 3수준 이상에 해당하는 학생을 일반집단으로, 2수준 이하에 소속된 학생을 기초수준 이하 집단으로 정의하였다(김수진 외, 2011; 남민우, 김경희, 김희경, 2012). 학생들의 읽기소양 집단 분류를 위해 읽기소양과 관련된 10개의 유의측정값

(Plausible values, PVs)을 활용하여 각각의 유의측정값을 기준으로 소양 수준에 따라 집단을 구분한 후, 다수결의 법칙(majority rule)에 따라 소속 비율이 가장 높은 집단을 최종 집단으로 결정하였다 (유진은, 2017; Breiman, 1996). 예를 들어, 어떤 학생의 10개의 PVs의 개별 점수가 각각 1c, 1a, 2, 1b, 2, 3, 4, 5, 2, 3수준에 위치한다면, 다수결의 법칙에 따라 기초수준 이하 집단(2수준 이하)으로 구분하였다. 분석에 사용된 학생들의 집단 비율을 살펴보면, 일반집단은 65.7%, 기초수준 이하 집단은 34.3%로 나타나 OECD(2019b)의 읽기소양 수준에 따른 학생 비율과 유사하게 나타났다.

2) 설명 변수

학생들의 읽기소양 수준에 따른 집단 분류에 주요한 역할을 하는 요인을 탐색하기 위해 사용한 변수는 총 286개이다. 이 변수들은 크게 학생, 학부모, 학교 및 교사 관련 변수로 나뉘며, 각 요인의 하위 문항을 그대로 사용하거나 파생변수(derived variable)를 사용하였다. 분석에 사용된 변수에 대한 자세한 정보는 <표 2>와 같다.

<표 2> 읽기소양 수준에 따른 집단을 결정하기 위한 설명 변수

| 영역 (문항 수) | 설명 변수 | |
|------------|---|---|
| 학생 (177) | 개인 특성 (46) | 나이, 빠른 생일 여부, 성별, 출생국, 가정에서 사용하는 언어, 초등학교 입학 시점, 가정에서 보유한 문화적 자본, 가정에서 보유한 도서 수, 경쟁심, 작업 완성도, 실패에 대한 두려움, 성취 목표, 회복탄력성, 부모와의 관계 |
| | 일반 활동 (143) | 읽기에의 참여, 다양한 자료 읽기 활동, 학생의 책 읽는 방식, 읽기 시간, 온라인 읽기 활동, 읽기의 어려움에 대한 인지, PISA검사의 어려움에 대한 인지, PISA 검사 및 설문 응답에 대한 노력, 학생의 읽기 전략-이해하고 기억하기/요약하기, 학생의 디지털 읽기 전략-질과 신용성 파악 |
| | 국어 수업 (30) | 수업 분위기, 교사의 학습 지원, 교사의 명확한 지도, 교사 지원, 교사의 학생에 맞는 수업 방식의 변화, 교사의 피드백, 교사의 교수-학습에 대한 흥미, 학생의 학습 자극을 위한 교수 전략 |
| | 학교 (22) | 학교급, 학교 유형, 학습기회-자료/텍스트 길이/국어 수업 과제 및 숙제, 디지털 국어 학습 기회, 1교시 당 수업 시간(분) |
| ICT(10) | ICT 기기/인터넷 처음 사용/접속한 시기, 가정/학교에서 이용 가능한 ICT, 학교에서 인터넷 사용 정도, 평일/주말에 학교 밖에서 인터넷 사용 시간, 국어 수업에서/수업 이외에 학습에서 디지털 기기 사용 정도, 국어 수업에서 학생이나 교사의 디지털 기기 사용 여부 | |
| 교육 이력 (24) | 초등학교 입학 전 교육 및 보육 기간, 학년 유예, 전학 횟수, 국어 교과에 대한 주 학습시간, 등교 전이나 방과 후에 공부하는 시간 및 이유, 국어 정규 수업 외 추가적인 교육, 학교 공부에 대한 주변의 지원, 해외 교환 학생의 기회가 있을 때 참여 희망 여부, 자신이 생각하는 5년 후 자신의 모습, 본인 직업 기대 수준 | |

| | | |
|--------------------|--------------|---|
| 학부모 (55) | 개인특성 (26) | 부/모 출생국, 부/모 교육수준, 부/모 직업지위, 경제·사회·문화적 지위 지수, 학부모의 읽기에 대한 견해, 학부모가 취미로 독서하는 시간, 학부모가 다양한 읽기 자료/디지털 읽기 자료를 읽는 빈도, 학부모가 책/뉴스를 읽는 방식 |
| | 자녀 (29) | 가정 내 학습을 위한 학부모의 지원, 자녀가 초1때 가족의 활동 빈도, 자녀가 초1때 한 활동에서 사용한 언어, 자녀가 10살 무렵 자료를 읽은 빈도, 정서적 지지, 자녀에게 기대하는 학력 수준, 자녀가 초1때 무렵 추가적인 국어 수업 참여 여부, 연간 지출한 교육비 |
| 교사 및 학교 (54) | 교사(13) | 교육적 자원 및 교원 부족에 대한 교사의 인식, 국어 교사의 협력, 학생과의 관계에 대한 교사효능감, 교수 활동에 대한 교사효능감, 읽기 이해에 대한 학습 기회 제공, 읽기 활동에 대한 교사의 자극, 읽기 전략에 대한 교사의 지도, ICT 어플리케이션 이용, 국어수업 분위기, 교사의 직접지도, 교사의 피드백, 적응 수업 |
| | 학교(41) | 학교가 위치한 지역 규모, 설립유형, 학생-교사 비율, 학교 크기, 교수 방해요인, 학생평가 결과의 활용 목적, 국어수업의 평균 학급 크기, 정규 수업 외 추가 국어 수업, 국어수업에 대한 정책, 정규 교육과정 외의 활동, 디지털 기기 사용에 대한 환경, 학생 1인당 사용 가능한 컴퓨터 수 |

주1. 파생변수(derived variable)가 존재하는 경우에는 파생변수를 우선적으로 사용하였으며, 이는 이탤릭체로 표시하였음.

3. 연구 모형 및 연구 방법

학생들의 읽기소양 수준에 따른 집단을 분류하는 데 영향을 미치는 변수를 탐색하기 위해 PISA 2018 한국 자료를 활용하여 랜덤 포레스트를 수행하였다. 분석을 위해 R 3.6.2 프로그램과 Liaw & Wiener(2002)가 개발한 randomForest R 패키지의 4.6-14 버전을 활용하였다.

분석에 앞서, 분석 자료의 결측 처리를 위해 randomForest 패키지에 포함된 na.roughfix 함수를 활용하여 결측치를 대체하였다. na.roughfix 함수는 각각의 변수 단위로 결측 여부를 확인하고 결측치를 대체한다. 이 연구에서는 연속 변수에 대해서는 중앙값(median)으로 대체하였고, 명목 변수에 대해서는 최빈치(mode)로 대체하였다. Breiman(2003)에 따르면 RF는 결측에 대해 강건하고, 결측이 20%를 넘지 않으면 중앙값/최빈치 대체가 가장 적절한 접근법이다. 이 연구에서 사용한 286개 변수의 대부분이 95% 이상의 유효 응답 비율을 보였고, 19개 변수만이 95%에 미치지 못하였으나 80%를 넘는 유효 응답 비율을 보였다.

랜덤 포레스트 분석 과정에서 생성한 붓스트랩 표본 수는 총 50,000개이고, 초기에 표본이 적을 때는 오차율이 다소 변화하다가 2,000번째 표본 이후에서는 오차율이 안정화되었다. 각각의 의사결정나무 모형의 각 마디에서 사용하는 변수 개수의 최적화 값은 randomForest 패키지에 포함된 tuneRF 함수를 활용하여 16으로 나타났다. 이는 Breiman(2001)이 제시한 $\text{floor}(\sqrt{p})$ (p = 전체 예측변수 개수)로 구한 값($\text{floor}(\sqrt{286}) = 16$)과 일치하였다. 또한, 최대 마디의 수는 의사결정나무의 가치가 충분히 생성될 수 있도록 20개로 설정하였다.

붓스트랩 표본의 표집 사례 수에 있어서는 각 집단의 비율에 차이가 있는 불균형 자료(imbalanced data)의 경우 분류 결과가 대체로 사례 수가 많은 집단 쪽으로 편향된다는 점(김미림, 박민호, 2019)을 감안하여 표집 비율을 조정하였다. 편향 보정을 위한 표집 비율 조정 방법은 사례 수가 많은 쪽을 적게 표집하거나(under-sampling) 사례 수가 적은 쪽을 많이 표집하는(over-sampling) 방식을 취할 수 있다(Van Hulse, Khoshgoftaar, & Napolitano, 2007). 이 연구에서는 사례수가 적은 ‘기초이하 수준 집단’이 실질적인 의미가 있다는 점과 선행 연구(김미림, 박민호, 2019; Drummond & Holte, 2003)를 참조하여 일반집단을 적게 표집하는 방식으로 붓스트랩 표본을 표집하였다.

랜덤 포레스트 분석 결과를 평가하기 위하여 훈련자료 및 시험자료의 정분류율(accuracy), 민감도(sensitivity), 특이도(specificity)를 산출하여 모형의 분류 정확도와 과적합 여부를 판단하였다. 정분류율은 일반집단과 기초수준 이하 집단의 모든 사례들 중 원래의 집단에 속하는 것으로 예측된 비율을 나타내고, 민감도는 연구의 초점 집단인 기초수준 이하 집단에 대한 정분류율을, 특이도는 일반집단에 대한 정분류율을 의미한다.

분류에 영향을 미치는 변수 선별은 중요도 지수를 사용하였고, 중요도 지수 순으로 상위 30개 변수(분석에 포함된 전체 286개 설명변수의 10% 정도)를 중요도 지수와 함께 결과에 제시하였다. 또한, 이 변수들에 대해 부분 의존성 도표를 제시하여 각 변수의 수준에 따라 ‘일반집단’ 또는 ‘기초수준 이하 집단’에 속할 확률의 변화를 시각화하였다. 부분 의존성 도표는 Welling, Refsgaard, Brockhoff, & Clemmensen(2016)이 개발한 R 패키지인 forestFloor 1.11.1 버전을 사용하여 도출하였다.

IV. 연구 결과

1. 집단 분류 결과

훈련 자료와 시험 자료에 대한 랜덤 포레스트의 분류 결과는 <표 3>과 같다. 분석 결과를 순차적으로 살펴보면, 훈련 및 시험 자료의 정분류율이 각각 .7840과 .7753으로 나타나 실제 집단과 예측된 집단이 일치하는 비율이 높은 것으로 나타났다. 또한, 민감도는 기초수준 이하 집단의 정분류율을 의미하며, 훈련 및 시험 자료에서 각각 .7384와 .7257로 나타났다. 마지막으로, 일반집단의 정분류율을 의미하는 특이도는 훈련 및 시험 자료에서 각각 .8078과 .8011로 나타났다. 이와 같은 결과는 집단 분류에 있어 과적합 문제가 발생하지 않고, 분류의 정확성이 양호하다는 것을 보여준다.

〈표 3〉 랜덤 포레스트의 분류 정확도

| 구분 | 정분류율 | 민감도 | 특이도 |
|-------|-------|-------|-------|
| 훈련 자료 | .7840 | .7384 | .8078 |
| 시험 자료 | .7753 | .7257 | .8011 |

2. 기초수준 이하 집단에 대한 주요한 설명 변수 분석

읽기소양의 기초수준 이하 집단을 분류하는 데 있어 중요도 지수가 높은 30개의 설명 변수를 정리한 결과는 〈표 4〉와 같다. 분석 결과를 살펴보면, 먼저, 읽기전략(요약, 이해 및 기억, 디지털 읽기 전략-질과 신용성 파악)과 관련된 변수가 중요도 지수가 높은 30개의 변수 중 30%(9개)를 차지하는 것으로 나타났다. 특히, 요약과 관련된 읽기전략, 디지털 읽기전략과 관련된 변수가 중요도 지수가 높은 상위 5개의 변수 중 4개의 변수로 나타났다. 따라서 읽기소양 수준에 따른 집단을 분류하는 데 있어 학생의 읽기전략이 중요한 역할을 하고 있음을 확인하였다.

또한, 읽기 활동과 관련된 다양한 변수가 중요도 지수가 높은 30개의 변수 중 약 26.7%(8개)로 나타났다. 이는 과거에 책을 읽은 빈도, 온라인 읽기 활동에 참여하는 정도, 읽기 어려움에 대한 학생들의 인식수준, 읽기 활동에 참여하는 것에 대한 태도 등이 읽기소양 수준에 따른 집단을 분류하는 데 주요한 역할을 하는 것을 나타낸다.

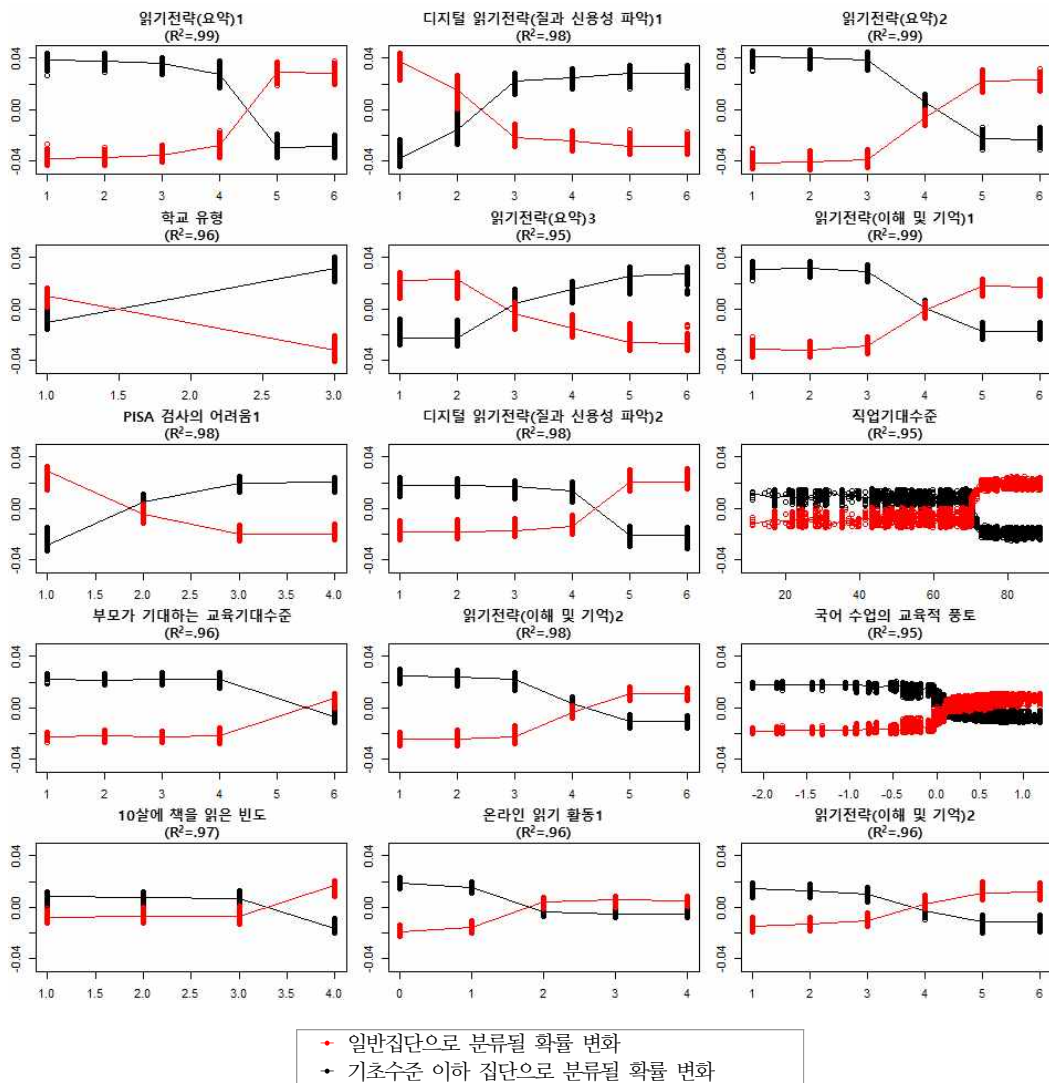
또한, PISA 검사 및 설문 문항 응답과 관련된 변수도 집단 분류에 대하여 상당 부분을 기여하는 것으로 나타났다. 구체적으로, PISA 검사의 글을 읽고 답하는 과정에서 어려움을 느끼거나 검사 및 조사에 응답하는 과정에서 노력한 정도를 나타내는 변수가 중요도 지수 상위 30개의 변수 중 약 16.7%(5개)를 차지하는 것으로 나타났다.

그 외에도, 학생이 재학하는 학교의 계열, 본인 또는 부모의 기대 직업 및 교육수준, 집에서 보유한 도서 수, 경제·사회·문화적 지위 지수(ESCS)와 같은 개인특성요인과 가정배경요인도 중요한 변수로 나타났다. 교사 및 학교와 관련된 요인으로는 국어 수업의 교육적 풍토, 학교에서 사용 가능한 ICT 수도 중요한 변수임을 확인하였다.

다음으로는, 중요도 지수가 높은 30개의 변수의 각 수준에 따라 기초수준 이하 집단 대비 일반집단에 분류될 확률 변화를 살펴보았다. 이를 위해 〈표 4〉에 제시된 30개 설명변수의 영향력을 나타내는 부분 의존성 도표를 [그림 1]과 [그림 2]와 같이 제시하였다. 각 도표에서 X축은 각 변수의 척도를 나타낸다. 또한, Y축은 해당 변수의 각 수준에 따른 분류 확률 변화를 나타내므로 각 집단에 대한 분류 확률 변화의 합은 항상 0으로 나타난다(Welling et al., 2016).

〈표 4〉 읽기소양 수준 집단 분류의 주요 설명 변수

| 중요도 순서 | 중요도 지수 | 설명 변수 |
|-----------|-----------|---|
| 1 | 94.84 | 읽기전략(요약): 가장 중요한 문장에 밑줄을 그으면서 텍스트를 읽은 후 자신의 글로 요약 |
| 2 | 93.19 | 디지털 읽기전략(질과 신용성 파악): 가급적 빨리 신청서를 작성하기 위해 링크 클릭 |
| 3 | 91.88 | 읽기전략(요약): 가장 중요한 사실들의 요약한 글에 포함되었는지 신중하게 확인 |
| 4 | 82.17 | 학교 유형: 1-일반계, 2-전문계(특성화) |
| 5 | 81.30 | 읽기전략(요약): 가능한 한 많은 문장을 정확하게 옮겨 적음 |
| 6 | 80.16 | 읽기전략(이해 및 기억): 텍스트를 자신의 말로 요약 |
| 7 | 73.54 | PISA 검사의 어려움: 이해할 수 없는 단어 많음 |
| 8 | 72.23 | 디지털 읽기전략(질과 신용성 파악): 스마트폰 제공에 대한 안내가 있었는지 알아보기 위해 휴대전화 회사의 웹 사이트 확인 |
| 9 | 70.11 | 직업기대수준 |
| 10 | 68.58 | 부모가 기대하는 교육기대수준 |
| 11 | 65.23 | 읽기전략(이해 및 기억): 텍스트의 중요한 부분에 밑줄 그음 |
| 12 | 62.02 | 국어 수업의 교육적 풍토 |
| 13 | 59.77 | 10살에 책을 읽은 빈도 |
| 14 | 58.15 | 온라인 읽기 활동: 온라인으로 실용적인 정보 검색 |
| 15 | 57.62 | 읽기전략(이해 및 기억): 텍스트를 읽은 후에 다른 사람과 그 내용에 대해 토론 |
| 16 | 57.27 | 학교에서 한 학생이 사용 가능한 컴퓨터 수 |
| 17 | 56.63 | 온라인 읽기 활동: 특정한 주제에 관해 알기 위하여 온라인 정보 검색 |
| 18 | 56.51 | PISA 검사의 어려움: 여러 쪽에 걸쳐 있는 글들을 읽으면서 내용 제대로 파악하지 못함 |
| 19 | 53.85 | 디지털 읽기전략(질과 신용성 파악): 보낸 사람의 이메일 주소 확인 |
| 20 | 53.09 | PISA 검사의 어려움: 너무 어려운 글들이 많음 |
| 21 | 51.98 | 읽기 어려움에 대한 인지: 나는 좋은 독자이다 |
| 22 | 51.66 | 읽기에의 참여: 나에게 독서는 시간 낭비이다 |
| 23 | 46.94 | 읽기 어려움에 대한 인지: 나는 어려운 글도 이해할 수 있다 |
| 24 | 46.64 | 다양한 자료 읽기 활동: 소셜류 |
| 25 | 43.22 | 조사 응답에 대한 노력 |
| 26 | 43.16 | 집에서 보유한 도서 수 |
| 27 | 42.73 | 경제·사회·문화적 지위 지수 |
| 28 | 40.11 | 검사 응답에 대한 노력 |
| 29 | 39.30 | 학교에서 사용 가능한 ICT |
| 30 | 38.50 | 읽기에의 참여: 나는 필요한 정보를 찾기 위해서만 독서를 한다 |



[그림 1] 주요 설명 변수의 부분의존성 도표(중요도 순서 1~15에 해당하는 변수)

영역별로 나누어 주요 결과를 살펴보면, 첫째, 중요도 지수가 1, 3, 5번째로 높았던 요약과 관련된 읽기전략 변수는 2페이지의 텍스트를 글로 요약하는 과제를 제시하고, 각 전략이 얼마나 유용한지를 묻는 문항이다. 부분 의존성 도표를 살펴보면, 가장 중요한 문장을 중심으로 자신의 글로 요약(중요도 순서: 1)하는 것이 중요하다고 생각하는 학생일수록 기초수준 이하 집단보다 일반집단으로 분류될 확률이 높아지는 것으로 나타났다. 더욱이, 일정 수준(4점)을 넘어서는 순간에 일반집단으로 분류될 확률이 급격하게 증가하는 것으로 나타났다. 또한, 중요한 내용이 요약된 글에 포함되었는지 확인(중요도 순서: 3)하는 전략이 유용하다고 생각하는 학생일수록 일반집단으로 분류될 확률이 높

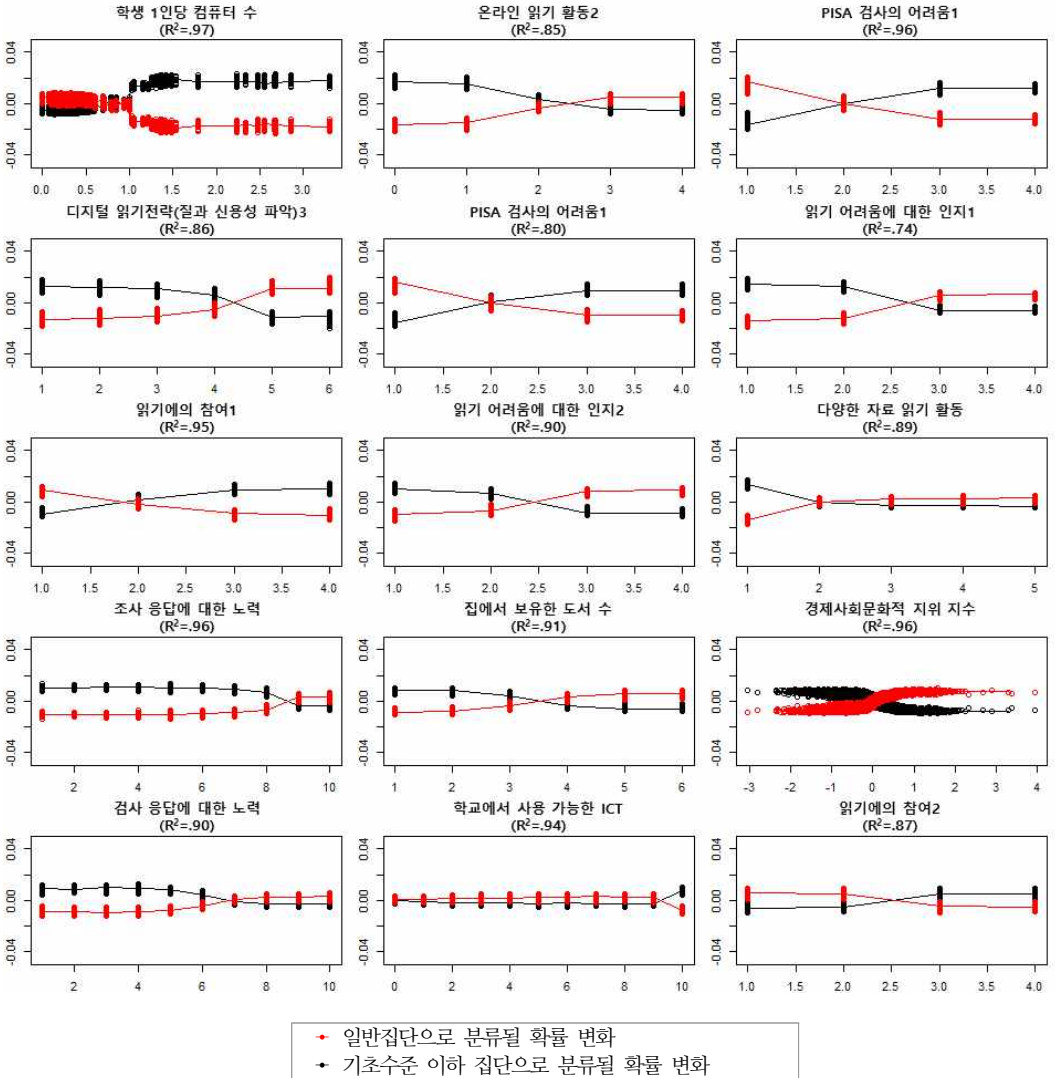
아졌다. 특히, 이 전략이 유용하다고 생각하는 학생(5점 이상)보다 이 전략이 유용하지 않다고 생각하는 학생(3점 이하)들을 읽기소양 수준의 집단을 분류하는 데 더 큰 영향력을 나타냈다. 반면, 가능한 한 많은 문장을 그대로 옮겨 적는 전략(중요도 순서: 5)이 요약이 유용하다고 생각하는 학생일수록 읽기소양의 일반집단보다 기초수준 이하 집단으로 분류될 확률이 높게 나타났다.

또한, 디지털 읽기전략(질과 신용성 파악)과 관련된 변수는 당첨된 스마트폰을 받기 위해 링크를 클릭하여 필요한 정보와 함께 신청서를 작성하라는 이메일을 받은 상황에서, 텍스트의 질과 신용성을 파악하는 데 각 전략이 얼마나 적절한지 묻는 문항이다. 분석 결과를 살펴보면, 가급적 빨리 신청서를 작성하기 위하여 링크를 클릭하는 전략(중요도 순서: 2)이 적절하지 않다고 생각하는 학생일수록, 또는 스마트폰 제공에 대한 안내가 있었는지 확인하기 위해 휴대전화 회사의 사이트를 확인하거나 이메일 발신자의 주소를 확인하는 전략(중요도 순서: 8)이 텍스트의 질과 신용성을 파악하는 데 적절하다고 생각하는 학생일수록 읽기소양 일반집단에 분류될 확률이 증가하는 것으로 나타났다.

또한, 텍스트에 있는 정보를 이해하고 기억하는 전략으로 중요한 내용에 밑줄을 긋거나(중요도 순서: 11) 텍스트를 읽은 후에 다른 사람과 그 내용에 대해 토론하는 전략(중요도 순서: 15)이 유용하다고 생각하는 학생일수록 일반집단으로 분류될 확률이 증가하는 나타났다. 특히, 중요한 내용에 밑줄을 긋는 전략에 대한 부분 의존성을 살펴보면, 변수의 낮은 수준에서의 분류 확률 변화 폭이 높은 수준에서의 변화 폭보다 상대적으로 크게 나타났다. 따라서 이 전략이 유용하다고 생각하는 학생들보다 유용하지 않다고 생각하는 학생을 분류하는 데 더 큰 기여를 하는 것으로 나타났다.

둘째, 읽기 활동과 관련된 설명 변수의 결과를 살펴보면, 10살에 책을 읽은 빈도(중요도 순서: 13)가 높은 학생일수록 기초수준 이하 집단보다 일반집단으로 분류될 확률이 높은 것으로 나타났다. 특히, 수준에 따라 분류 확률의 변화를 비교하면, 책을 거의 읽지 않거나(1점) 일주일에 한두 번 책을 읽는 학생(3점)에 대한 분류 확률의 변화의 폭은 유사하게 나타났다. 반면, 책을 매일 또는 거의 매일(4점 척도) 읽는 학생들을 일반집단으로 분류할 확률이 증가하는 것으로 나타나, 책을 읽는 빈도의 수준에 따라 집단을 분류하는 데 기여하는 정도의 차이가 다르게 나타남을 확인하였다. 또한, 실용적인 정보(중요도 순서: 14)를 검색하거나 특정한 주제(중요도 순서: 17)에 관해 알기 위하여 온라인 정보를 검색하는 활동을 많이 한 학생일수록 일반집단으로 분류될 확률이 증가하는 것으로 나타났다. 더욱이, 온라인 읽기 활동을 모르거나 거의 하지 않는 학생(1점 이하)의 경우에는 읽기소양 기초수준 이하 집단으로 분류될 확률 변화가 높은 반면, 최소 한 달에 몇 번이라도 온라인 읽기 활동을 한 학생(2점 이상)의 경우에는 각 집단으로 분류되는 데 뚜렷한 차이를 보이지 않았다. 또한, 자신이 좋은 독자(중요도 순서: 21)이거나 어려운 글도 이해(중요도 순서: 23)할 수 있다고 생각하는 학생들, 즉, 읽기에 대한 어려움이 없다고 인식하는 학생들일수록 일반집단으로 분류될 확률이 높아지는 것으로 나타났다. 반면, 독서가 시간 낭비(중요도 순서: 22)이거나 필요한 정보를 찾기 위해서만 독서(중요도 순서: 30)를 하는 경향이 높은 학생일수록 일반집단보다 기초수준 이하 집단으

로 분류될 확률이 상승하는 것으로 나타났다. 마지막으로, 스스로 원해서 소설류의 자료를 자주 읽는 학생(중요도 순서: 24)일수록 일반집단으로 분류될 확률이 상승하는 것으로 나타났다. 더욱이, 부분의존성 도표를 살펴보면, 소설류의 자료를 전혀 또는 거의 읽지 않은 학생(1점)들은 기초수준 이하 집단으로 분류될 확률이 높은 반면, 일 년에 몇 번이라도 자료를 읽는 학생(2점 이상)들의 경우, 각 집단으로 분류될 확률이 유사한 것으로 나타났다.



[그림 2] 주요 설명 변수의 부분의존성 도표(중요도 순서 16~30에 해당하는 변수)

셋째, PISA 검사 및 조사에 대한 응답과 관련된 결과는 다음과 같다. 먼저, PISA 검사의 글을 읽고 답하는 문제에 대하여 이해할 수 없다는 단어가 많거나, 너무 어려운 글들이 많다고 느낀 학생(중요도 순서: 7, 20)일수록 일반집단보다 기초수준 이하 집단으로 분류될 확률이 증가하는 것으로 나타났다. 또한, 구조상의 이유로 여러 페이지에 걸쳐 있는 글들을 읽으면서 내용을 제대로 파악하는 데 어려움을 느낀 학생(중요도 순서: 18)일수록 일반집단보다 기초수준 이하 집단으로 분류될 확률이 높아지는 것으로 나타났다. 반면, PISA 검사 문항을 풀거나 설문 문항에 응답하는 데 있어 열심히 노력한 학생(중요도 순서: 28, 25)일수록 일반집단으로 분류될 확률이 증가하는 것으로 나타났다. 한편, 설문 응답에 대한 노력의 부분 의존성 도표를 살펴본 결과, 8점과 9점 사이에서 집단 분류의 양상이 다르게 나타났으며, 9점 이상일 경우에는 집단 분류 확률의 변화가 큰 차이를 보이지 않았다.

넷째, 학생의 개인특성, 가정 및 학교 맥락변인에 관한 결과를 요약하면 다음과 같다. 먼저, 학생이 재학하는 학교의 계열(중요도 순서: 4)이 집단 분류 확률에 영향을 미치는 것으로 나타났다. 일반계 학교에 재학하는 학생의 경우, 각 집단으로 분류되는 데 기여하는 정도는 큰 차이를 보이지 않았지만, 전문계열 학교에 재학하는 학생의 경우, 기초수준 이하 집단으로 분류될 확률이 일반집단으로 분류될 확률보다 큰 차이를 보이며 높게 나타났다. 또한, 학생 또는 부모가 기대하는 직업 또는 교육 수준이 높은 학생(중요도 순서: 9, 10)일수록 일반집단으로 분류될 확률이 높은 것으로 나타났다. 이때, 직업기대수준의 경우에는 일정 수준을 넘어서는 순간 일반집단과 기초수준 이하 집단을 분류하는 정도가 크게 나타나는 특징을 보였다. 반면, 부모의 교육기대수준의 경우에는 자녀가 석·박사 학위를 취득하길 원하는 경우(6점)에는 두 집단으로 분류되는 정도가 유사하게 나타난 반면, 자녀에 대한 교육기대수준이 상대적으로 낮은 경우(4점 이하)에는 일반집단보다 기초수준 이하 집단으로 분류되는 경향이 큰 차이를 보이며 더 높게 나타나는 특징을 보였다. 또한, 가정에서 보유한 도서 수(중요도 순서: 26)의 경우, 가정에서 100권 이하의 책을 보유한 경우(3점 이하)에는 기초수준 이하 집단으로 분류될 확률이 증가하는 반면, 100권 초과 책을 보유한 경우(4점 이상)에는 일반집단으로 분류되는 모습을 보였다. 또한, 가정의 ESCS(중요도 순서: 27)의 경우에는 평균을 기점으로 두 집단으로 분류되는 정도가 다르게 나타났다. 즉, 평균보다 높은 ESCS 가정의 학생은 일반집단으로 분류되는 반면, ESCS가 평균 이하인 가정의 학생은 기초수준 이하 집단으로 분류되는 경향을 보였다. 이때, 읽기소양 집단 분류에 미치는 영향이 ESCS가 높아지거나 낮아짐에 따라 급격한 차이를 보이지 않고, 일반적으로 유사한 영향력을 갖는 특징을 보였다. 마지막으로, 학교맥락변인의 경우, 국어 수업의 교육적 풍토(중요도 순서: 12)가 좋을수록, 학교에 학생이 사용할 수 있는 컴퓨터의 수(중요도 순서: 16)가 적을수록, 또는 학교에 학생이 사용할 수 있는 다양한 ICT 기기(중요도 순서: 29)가 구비되어 있는 정도가 높을수록 학생이 일반집단으로 분류되는 경향이 높게 나타났다.

V. 논의 및 결론

기술의 발전으로 인해 학생들은 다양한 매체 경로를 통해 수많은 정보를 경험하고 있다. 따라서 학생들이 정보를 이해하고, 자신의 목적에 맞춰 정보를 선택하고, 해석 및 종합할 수 있는 능력이 중요하게 되었다. 이러한 관점에서 이 연구는 읽기소양의 수준이 기초수준 이하인 학생과 그렇지 않은 학생을 분류하는 주요한 변수를 탐색함으로써, 학생의 읽기소양을 향상시킬 수 있는 방안을 탐색하는 데 목적을 두었다. 이를 위해 PISA 2018 자료를 활용하여 6,630명 학생을 연구 대상으로 선정하였고, 학생, 학부모, 교사, 학교장의 설문 자료 중 총 286개의 설명 변수를 사용하였다. 이후, 랜덤 포레스트를 활용하여 읽기소양 수준에 따라 정의된 일반집단과 기초이하 수준 집단을 분류하는 데 주요한 변수가 무엇인지 탐색하였다. 주요 결과에 대한 요약과 관련된 시사점은 다음과 같다.

첫째, 학생의 읽기소양 수준에 따라 기초수준 이하 집단과 일반집단을 분류하는 데 있어 읽기전략이 중요한 변수임을 확인하였다. 구체적으로, 텍스트를 이해하고 기억하기, 요약하는 읽기전략뿐만 아니라 온라인 상황에서 텍스트의 질과 신용성을 파악하는 디지털 읽기전략과 관련된 다수의 변수가 중요도 지수가 높은 변수(30%)로 선별되었다. PISA 2009 자료를 활용한 선행연구의 경우, 다양한 읽기전략(예: 이해 및 기억, 요약, 통제)이 읽기소양과 정적인 관계를 보이거나 보통집단과 하위 집단을 분류하는 데 유의한 영향을 미치고 있음을 확인하였다(강대중, 2015; 김재철, 2005; 박현정, 하여진, 2011). 따라서 이 연구 결과는 읽기전략이 읽기소양과 관련이 있다고 본 선행연구와 같은 맥락을 보이고 있다. 이러한 결과를 토대로 읽기소양의 수준 향상을 위하여 다음과 같은 방안을 제안할 수 있다. 다양한 교과 수업에서 읽기활동과 관련된 교육활동이 이루어질 때, 학생들이 적절한 전략을 사용하는지를 교사, 동료, 또는 학생이 직접 점검할 수 있는 기회를 제공해야 한다. 예를 들어, 학생이 문장을 그대로 옮겨 적는 전략을 사용하고 있다면, 글에서 가장 중요한 문장이 무엇인지 확인하고 자신의 글로 요약한 후, 요약된 글에 중요한 부분이 포함되었는지 점검할 수 있도록 도움을 주어야 한다. 또한, 글을 이해하고 기억하는 데 있어 텍스트를 자신의 말로 요약하고, 다른 사람과 그 내용에 대하여 토론을 하는 전략이 유용하다고 생각하는 학생이 일반집단으로 분류될 확률이 높게 나타났다. 따라서 학생들이 텍스트를 이해하고 활용하는 데 있어 동료와 대화 및 토의 기회를 충분히 제공하여 자신의 전략을 반성하고, 전략을 주도적으로 사용할 수 있는 기회를 부여해야 한다(박영목, 2006). 이때, 읽기소양 수준을 고려하여 학생들에게 친숙하거나 비교적 길이가 짧은 텍스트를 순차적으로 제공하여 학습 부담을 완화할 필요가 있다.

또한, 이 연구에서는 추가적으로 디지털 상황에서 이루어지는 읽기전략이 읽기소양 수준에 따른 집단을 분류하는 데 영향을 미치는 것으로 나타났다. 예를 들어, 이메일을 받은 상황에서 텍스트의 질과 신용성을 파악하는 전략 중 회사의 사이트를 확인하거나 이메일 발신자의 주소를 확인하는 전략을 사용하는 학생이 일반집단으로 분류될 확률이 높았다. 더욱이, PISA 2018에서 읽기소양을 정

의하는 데 있어 자신에게 필요한 정보를 선택하고, 정보의 신용을 판단할 수 있는 능력인 ‘평가’가 추가된 만큼, 읽기소양에서 정보의 질과 신용성을 파악하는 디지털 읽기전략의 중요성이 강조되고 있다(조성민 외, 2019). 또한, PISA와 PIRLS는 소양을 평가하기 위해 컴퓨터 기반 평가를 시행하고 있으며, 국가수준 학업성취도 평가는 핵심역량을 평가하기 위해 지필 검사에서 구현하기 어려운 신유형 문항을 컴퓨터 기반 평가에 도입하는 방안을 중장기적으로 계획하는 바와 같이, 국내외적으로 컴퓨터에 기반한 평가를 시행 및 준비하고 있다(김희경 외, 2019; Martin, Mullis, & Foy, 2015; OECD, 2019b). 이처럼 디지털 형태에서 읽기소양이 평가 및 요구되고 있는 만큼, 교육 활동 과정에서 디지털 형태의 자료를 읽는 전략을 활용할 수 있는 교수학습이 이루어져야 한다. 추가적으로, 이 연구에서는 온라인에서 특정한 주제의 정보나 실용적인 정보를 검색하는 활동 빈도가 읽기소양 수준에 따른 집단을 분류하는 데 영향을 미치는 것을 확인하였다. 따라서 다양한 교과와 수업에서 페이퍼 기반의 자료뿐만 아니라 온라인에서의 읽기자료, 애니메이션 및 동영상에서 제공되는 읽기 자료를 제공함으로써, 학생이 디지털 맥락의 다양한 자료에서 자신이 필요로 하는 정보를 선별하고, 그 정보의 질과 신용성을 판단하는 읽기전략을 학습할 수 있는 기회를 충분히 제공해야 한다.

둘째, 읽기소양 집단을 구분하는 데 읽기활동의 경험과 태도가 주요한 영향을 미치는 것으로 나타났다. 이러한 결과는 선행연구와 일치된 결과를 보여준다(강대중, 2015; 강대중, 염시창, 2013; 김재철, 2005; 김혜숙, 2012; 박주현, 장우권, 2014; 박현정, 하여진, 2011; 정혜경, 조지민, 2013). 먼저, 과거의 독서 빈도는 읽기소양 집단을 분류하는 데 중요한 변수로 선정되었다. 즉, 10살 무렵에 매일 또는 거의 매일 책을 읽었던 경험이 있는 학생은 기초수준 이하 집단보다 일반집단으로 분류될 확률이 높은 것으로 나타났다. 또한, 현재 시점에서 다양한 자료에 대한 읽기 활동을 살펴보면, 학생이 스스로 원하여 잡지, 만화, 비소설류, 신문의 자료를 읽는 활동은 읽기소양 집단을 분류하는 데 주요한 영향을 미치지 않은 반면, 학생이 스스로 원해 소설 자료를 읽는 활동은 집단 분류에 주요한 변수로 나타났다. 더욱이, 부분 의존성 결과에 근거하면, 소설을 전혀 읽지 않거나 거의 읽지 않는 학생은 기초수준 이하 집단으로, 학생이 스스로 원하여 1년에 몇 번이라도 소설을 읽은 학생은 일반 집단으로 분류될 확률이 높게 나타났다. 마지막으로, 읽기활동에 대한 태도의 경우, 학생들이 독서를 시간 낭비라고 생각하거나 필요에 의해서만 독서를 하는 경우에는 기초수준 이하 집단으로 분류될 확률이 높게 나타났다. 반면, 학생들이 자신이 좋은 독자라고 생각하거나 어려운 글도 이해할 수 있다고 인지하는 학생일수록 기초수준 이하 집단보다 일반집단으로 분류될 확률이 높게 나타났다.

이러한 결과는 현재의 읽기 활동뿐만 아니라 과거의 독서활동이 장기적으로 학생의 읽기소양에 영향을 미칠 수 있음을 보여준다. 또한, 읽기 활동에 대한 긍정적인 태도를 갖춘 학생일수록 더 높은 읽기소양을 갖추고 있음을 알 수 있다. 이러한 점에 근거하여 먼저, 학생들이 다양한 자료를 자주 읽고, 읽기 활동에 흥미와 긍정적 태도를 갖출 수 있도록 가정에서의 교육 및 환경이 필요함을 확인하였다. 이를 위하여 가정에서는 학생의 연령에 적절한 다양한 자료를 비치하거나 지역 시설을 통해

다양한 자료를 접할 수 있는 기회를 마련하고, 지속적인 읽기 활동을 장려할 수 있는 분위기를 형성할 수 있다. 이를 통해 읽기 활동이 의무적인 활동이거나 시간 낭비의 활동이라고 인식하기보다는, 학생이 흥미를 느껴 읽기 활동을 지속하기를 기대할 수 있다. 다음으로, 학급, 학교 및 지역사회에서도 학생의 읽기 활동을 장려하기 위한 교육활동 및 환경을 구축할 필요가 있다. 이를 위해 다양한 자료에 접근할 수 있는 경로를 간소화함으로써 자료를 손쉽게 접할 수 있는 환경을 구축하고, 읽기 활동 이후에 자신의 경험과 생각을 공유할 수 있는 장을 마련해야 한다. 이로부터 학생들의 읽기 경험의 폭을 확장하여 주체적으로 읽기 활동에 참여하기를 기대할 수 있다.

셋째, 이 연구에서는 PISA 검사에 대한 학생의 응답이 읽기소양 집단을 분류하는 데 있어 주요한 영향을 미치는 것으로 나타났다. 이는 개별문항에 초점을 맞추어 해석할 필요가 있다. 먼저, 어려운 단어나 이해하기 어려운 글들이 많아 검사에 어려움을 느낀 학생이 기초수준 이하 집단으로 분류되는 경향을 보였다. 이러한 결과는 검사에 대한 지각된 어려움이 검사에 대한 걱정을 야기하여 성취에 영향을 미친다고 해석함과 동시에(Hong, 1999; Li, Lee, & Solmon, 2007), 학생의 읽기소양의 수준이 낮기 때문에 문항을 푸는 데 어려움을 느꼈다고 해석할 수 있다. 한편, 글이 여러 쪽에 나누어 제시되어 내용을 파악하는 데 어려움을 겪은 학생이 기초수준 이하 집단으로 분류될 확률이 높게 나타났다. 이러한 결과를 통해 문항의 구조가 방해요인으로 작용하여 측정하고자 하였던 읽기소양을 제대로 측정하지 못한 결과라고 해석할 수 있다.

넷째, 이 연구에서는 랜덤 포레스트를 활용하여 읽기소양 집단 분류에 영향을 미치는 주요 변인을 파악함으로써 다음과 같은 의미를 갖는다. 먼저, 이 연구에서는 학생, 학부모, 교사 및 학교와 관련된 많은 변수를 모형에 포함하여 변수들 간의 복잡한 관계 속에서 읽기소양에 주요한 영향을 미치는 변인을 파악했다는 점에서 이점을 갖는다. 따라서 많은 변수들 중에서 선행연구에서 언급되었던 학생의 읽기전략, 읽기 활동이 주요한 영향을 미친다는 것을 확인할 수 있었다. 더욱이, 요인과 관련된 하위문항의 평균값 대신 요인과 관련된 모든 하위문항을 모형에 포함하여 보다 구체적으로 관계를 파악했다는 점에서 의미를 갖는다. 따라서 읽기전략에 대한 구체적인 내용에 따라 읽기소양 수준에 따른 집단과의 관계를 비교하였다. 예를 들어, 텍스트를 요약하는 읽기 전략의 경우, 중요한 문장에 밑줄을 그으며 자신의 글로 요약하고, 중요한 사실이 요약한 글에 포함되었는지 확인하는 전략이 유용하다고 생각하는 학생은 일반집단으로 분류될 확률이 높은 반면, 가능한 한 많은 문장을 정확하게 옮겨 적는 전략이 유용하다고 생각하는 학생은 기초수준 이하 집단으로 분류될 확률이 높게 나타났다. 이러한 결과를 통해 읽기소양을 향상시키기 위하여 학생들이 어떠한 읽기전략을 취하는 것이 보다 적절한지 확인할 수 있다. 다음으로, 이 연구에서는 랜덤 포레스트를 활용함으로써 집단 분류와 변수들 간의 비선형적인 관계를 파악할 수 있다는 점에서 이점을 갖는다. 예를 들어, 학생이 스스로 원하여 소설을 자주 읽을수록 기초수준 이하 집단보다 일반집단에 속할 확률이 높다고 판단하는 것 이상으로 정보를 파악할 수 있었다. 즉, 평소에 학생이 소설을 전혀 읽지 않거나 거의 읽지 않은

경우에는 기초수준 이하 집단으로 분류될 확률이 높지만, 일 년에 몇 번이라도 소설을 읽을 경우에는 두 집단으로 분류될 확률이 유사한 정도로 나타났다. 이와 같이 변수들과 집단 분류 확률과의 구체적인 관계를 제공해준다는 점에서 이 연구는 학생들에게 필요한 교육적 지원이 무엇인지 단서를 제공해준다.

이러한 이점에도 불구하고, 이 연구는 다음과 같은 한계점을 보이며, 이를 바탕으로 후속 연구에 대하여 다음과 같은 제언이 가능하다. 첫째, 이 연구에서는 PISA 자료를 활용하였음에도 불구하고 학생이 학교에 내재된 구조를 반영하지 못하고 학생 수준에서 분석을 실시하였다. 따라서 후속 연구에서는 다층 구조를 반영하여 데이터 마이닝 기법을 적용하여 연구 결과의 타당도를 검증해볼 필요가 있다(Sela & Simonoff, 2012). 둘째, 이 연구에서는 읽기소양 수준에 따라 기초수준 이하 집단, 일반집단과 같이 두 집단을 정의한 후, 두 집단 분류에 대한 결정 요인을 파악하였다. 후속 연구에서는 읽기소양 수준에 따라 기초수준 이하 집단, 보통집단, 우수집단과 같이 집단을 보다 세분화하여 집단 분류에 대한 결정 요인을 비교함으로써, 읽기소양 수준에 따라 필요한 교육적 지원이 무엇인지 파악할 필요가 있다.

참고문헌

- 강대중(2015). PISA 2009와 NAEA 2009에 모두 참여한 학생 및 학교의 학교효과 분석. **학습자중심 교과교육연구**, 15(10), 1-19.
- 강대중, 염시창(2013). PISA 2009 읽기성취도 데이터를 활용한 학교효과 분석. **교과교육학연구**, 17(2), 323-345.
- 강현철, 서두성, 최종후(1998). Enterprise Miner의 의사결정나무분석 알고리즘. SAS 사용자 컨퍼런스 발표 자료집, 169-186.
- 교육부(2019). 한 아이도 놓치지 않고 기초학력 책임진다. 교육부 보도자료(2019.3.28.).
- 권재기(2013). 일반계 고등학교의 기초학력미달 변화에 따른 학교유형 분류 및 종단적 변화. **교육평가연구**, 26(5), 1037-1066.
- 김미림, 박민호(2019). 랜덤 포레스트를 활용한 대학생의 최초 취업 사교육 참여 시점별 특성 분석. **교육연구논총**, 40(1), 1-33.
- 김수진, 김경희, 김희경, 남민우, 시기자, 신진아, 이인호, 임해미(2011). 국가수준 학업성취도 평가와 국제 학업성취도 평가 연계를 통한 우리나라 학생들의 성취 특성 비교(RRE 2011-1). 한국교육과정평가원.
- 김재철(2005). 성별과 학습선호도 유형에 따른 독서흥미, 독서량, 인지전략, 학문적 자신감, 읽기소양의 구조적 관계. **교육심리연구**, 19(1), 93-114.
- 김태은, 이화진, 오상철, 노원경(2012). 초등학교의 기초학력 미달 여부에 영향을 미치는 변인들의 효과 분석. **교육심리연구**, 26(2), 505-521.
- 김혜숙(2012). ICT 활용이 학업성취도에 미치는 영향: PISA 2009 한국 자료를 중심으로. **아시아교육연구**, 13(1), 1-22.
- 김희경, 김완수, 김수진, 정혜경, 김미림, 김성훈(2019). 국가수준 학업성취도 평가 점수 체제 개선 및 결과 활용도 제고 방안(RRE 2019-3). 한국교육과정평가원.
- 남민우, 김경희, 김희경(2012). PISA와 NAEA에서 나타난 우리나라 학생들의 국어 능력 특성 연구. **새국어교육**, 89, 155-186.
- 박영목(2006). 전략적 과정 중심 읽기 지도 방안. **독서연구**, 16, 269-296.
- 박주현, 장우권(2014). PISA 2009 학업성취도에 대한 학교도서관 변인의 영향력 분석. **정보관리학회지**, 31(3), 331-351.
- 박현정, 하여진(2011). PISA 읽기소양 우수 및 취약집단 학생들에 대한 결정요인 변화. **교육평가연구**, 24(4), 921-942.

- 박혜영, 송미영(2014). PISA 2012 PRA와 DRA에 나타난 우리나라 남녀 학생들의 읽기 성차 특성 분석. **국어교육**, 145, 349-377.
- 송미영, 김성숙, 이현숙, 김준엽(2011). 학교교육 개선을 위한 학생의 학업성취 수준 결정요인 분석. **교육평가연구**, 24(2), 261-289.
- 송미영, 임현정, 김성숙(2015). PISA 성취수준 향상에 미치는 교육맥락변인의 차별적 효과 분석. **한 국교육**, 42(1), 249-273.
- 유진은(2015). 랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법. **교육평가연구**, 28(2), 427-448.
- 유진은(2017). 기계학습을 통한 TIMSS 2011 중학생의 수학 성취도 관련 변수 탐색. **교원교육**, 33(1), 43-56.
- 이은하, 최은정(2015). 핵심역량과 국어과 교육과정의 연계 가능성 탐색: PISA 2009 읽기 영역을 중심으로. **아시아교육연구**, 16(2), 201-230.
- 이인화, 구남욱(2019). PISA 2015 읽기 영역 결과에 나타난 한국 학생들의 성취 특성 및 교육맥락변 인의 영향력 분석 연구. **독서연구**, 50, 113-144.
- 이효정, 김광주, 김성훈(2018). 중학교 기초학력 미달률에 미치는 학교환경과 교장공모제의 효과. **교 육과정평가연구**, 21(1), 173-195.
- 전명남, 양명희(2015). 중학교 기초학력미달 학생의 학습실태와 요구 조사. **교육학논총**, 36(2), 21-39.
- 정혜경, 조지민(2013). PISA 2009 읽기와 사회경제적 배경 변인에 기반한 역경 극복 학생 특성 분석. **교육과정평가연구**, 16(1), 161-184.
- 조성민, 구남욱, 김현정, 이소연, 이인화(2019). OECD 국제 학업성취도 평가 연구: PISA 2018 결과 보고서(RRE 2019-11). 한국교육과정평가원.
- 조지민, 김수진, 이상하, 김미영, 옥현진, 임해미(2011). 2011년 국제 학업성취도 평가 연구 (PISA/TIMSS): PISA 2009 결과에 기반한 읽기 영역 성취 특성 비교(RRE 2011-4-3). 한국교육과정평가원.
- 최종후, 서두성(1999). 데이터마이닝 의사결정나무의 응용. **통계분석연구**, 4(1), 61-83.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Breiman, L. (2003). Manual for setting up, using, and understanding random forest V4.0. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf (2020.3.14. 검색).
- Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the International Conference on*

- Machine Learning* (ICML 2008), 96–103.
- Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)*, 27(3), 326–327.
- Drummond C., & Holte., R. C. (2003). Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. *Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II*, 1–8.
- Dudoit, S., Fridlyand, J., & Speed., T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97, 77–87.
- Hamza, M., & Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8), 629–643.
- Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*(2nd edition). New York, NY: Springer Science & Business Media.
- Hong, E. (1999). Test anxiety, perceived test difficulty, and test performance: Temporal patterns of their effects. *Learning and Individual Differences*, 11(4), 431–447.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.
- Li, W., Lee, A., & Solmon, M. (2007). The role of perceptions of task difficulty in relation to self-perceptions of ability, intrinsic value, attainment value, and performance. *European Physical Education Review*, 13(3), 301–318.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- Martine, M. O., Mullis, I. V. S., & Foy, P. (2015). Chapter 3. Assessment design for PIRLS, PIRLS literacy, and ePIRLS in 2016. In *PIRLS 2016 assessment framework*(2nd ed). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/pirls2016/framework.html> (2020.3.14. 검색).
- OECD. (2019a). *PISA 2018 assessment and analytical framework*. Paris: OECD Publishing.
- OECD. (2019b). *PISA 2018 results(Volume I): What students know and can do*. Paris: OECD Publishing.
- Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 86, 169–207.

- Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. *Proceedings of the International Conference on Machine Learning (ICML 2007)*, 935-942.
- Welling, S. H., Refsgaard, H. H. F., Brockhoff, P. B., & Clemmensen, L. K. (2016). Forest floor visualizations of random forests. ArXiv-preprints. Retrieved from <http://arxiv.org/abs/1605.09196>. (2020.1.30. 검색).

* 논문접수 2020년 2월 10일 / 1차 심사 2020년 3월 6일 / 게재승인 2020년 3월 18일

* 손윤희: 서울대학교 지구과학교육과를 졸업하고 서울대학교 교육학과에서 교육측정평가 전공으로 석사 및 박사학위를 취득하였다. 현재는 한국교육과정평가원 부연구위원(위촉)으로 재직 중이며, 국가수준 학업성취도 평가 관련 업무를 수행하고 있다. 주요 관심분야는 다층자료 분석, 매개효과 분석, 혼합모형, 구조방정식모형 등이다.

* E-mail: sonyoonhee0106@gmail.com

* 박현정: 서울대학교 교육학과를 졸업하고 동 대학원에서 석사학위를 취득하였으며 University of Minnesota에서 교육측정 및 통계 전공으로 박사학위를 취득하였다. 현재 서울대학교 교육학과 교수(교육측정평가 전공)로 재직 중이며, 주요 관심분야는 국제비교분석, 종단자료의 분석, 다층자료 분석 등이다.

* E-mail: hjp@snu.ac.kr

* 박민호: 서울대학교 사범대학 물리교육과를 졸업하고, 서울대학교 사범대학 교육학과에서 석사학위를 취득하였으며, 현재 동대학원 박사과정을 수료하였다. 주요 관심분야는 문항반응이론, 빅데이터분석, 머신러닝 등이다.

* E-mail: pmino85@snu.ac.kr

Abstract

Random Forest Analysis of Factors Influencing the Students' Reading Literacy Levels: Using PISA 2018 Korea Data

Son, Yoonhee^{*}
Park, Hyun-Jeong^{**}
Park, Minh^{***}

The purpose of this study is to explore the variables that have a major impact on the classifying of groups by reading literacy level. To research the study, the groups were divided to the general group and the below basic literacy level group by reading literacy level and the random forest was used to explore the main variables to classify the group by reading literacy level. And the subjects of this study were the 6,630 students in the 168 schools who participated in PISA 2018 and the 286 variables of student, parents, teacher, and school were used for analysis. As a result, the variables of student were found to play a major role in classifying students into the general group rather than the below basic literacy level group. Specifically, the group classification was influenced by the variables related to the various reading strategies, the experiences, frequency and cognition of reading activities, and the difficulty or effort of the PISA testing or survey. Also, other variables related to individual and family backgrounds affected the classification of groups. In addition, by using the random forest, this research confirmed the nonlinear relationship between group classification and affecting variables, taking into account complex relationships among variables. Based on these results, this study suggested the educational supports needed to improve the students' reading literacy.

Key words: reading literacy, below basic literacy level group, group classifying, random forest, PISA 2018

* First author, Associate research fellow(temporary), Korea Institute for Curriculum and Evaluation

** Professor, Seoul National University

*** Corresponding author, Ph. D. Candidate, Seoul National University