**논문번호:** J1\_202500029

**원논문 제목:** “설명가능한 인공지능 기반 비형식학습 효과성 분석 및 참여자 추천 애널리틱스: 지속 가능한 평생학습 유도를 위하여 (Sustainable Lifelong Learning: Explainable AI-Driven Analytics for Informal Learning Effectiveness and Personalized Recommendations)”

**안녕하십니까 한국빅데이터학회 학회지 에디터님,**

원고의 재제출과 심사자의 의견에 대해 답변할 수 있는 기회를 주셔서 감사합니다.

요청해주신 바와 같이 아래에 심사자별 코멘트에 대한 답변을 포함하였습니다.

심사자분들께서 지적해 주신 내용 외에도 일부 업데이트 사항을 반영 및 표시하여 논문을 개선하였습니다.

감사합니다,

유효정, 조은지, 이현서, 김경원

## 심사자 #1

### 심사자 #1, 심사내용 #1 (유효정, 조은지, 이현서)

|  |
| --- |
| 기존 연구 인용은 풍부하나, 2010년대 초반 논문 비중이 높습니다. 최근 국제 저널에서 발표된 AI·XAI 기반 평생학습 효과성 분석, 추천 시스템 연구를 추가 검토하고 본 논문과의 차별성을 강조해야 합니다.  (최신 문헌 보강: 선행연구 검토가 다소 구시대적입니다. 최근 국제 학술지에서 발표된 AI/XAI 기반 평생학습 분석 및 추천 시스템 관련 연구들을 추가로 반영하여, 본 연구의 차별성과 기여점을 보다 선명하게 제시할 필요가 있습니다.) |

**저자답변:** 이건 나도 걸리는 부분이긴 했는데.. 리뷰어가 바로 지적했네요;; (0) 엑셀에 선행연구 정리하던 방식과 유사하게.. (1) 구글 스칼라에서 “AI, XAI, Life Long Education, Continuing Education, Recommend System” 관련 선행연구들을 찾아서 정리 (2) 초록과 결론 부분만 보고 우리 연구와 비교할만한 인싸이트나 결과들만 빠르게 정리해서 작성해두세요. 최대한 많이 찾아보세요. 그걸 기반으로 글은 내가 쓰겠습니다. 간단하게 찾아본건 아래이니 포함해서 추가로 더 찾아 정리해주세요~

**추가사항:**

|  |
| --- |
|  |

### 심사자 #1, 심사내용 #2 (김경원)

|  |
| --- |
| 데이터 전처리 과정에서 사용된 언더샘플링·오버샘플링 기법의 타당성 설명이 부족합니다. 과적합(Overfitting) 위험에 대한 검증(예: 교차검증, 하이퍼파라미터 튜닝 과정)도 구체적으로 보강할 필요가 있습니다. 다양한 알고리즘 비교는 장점이지만, 왜 최종적으로 Random Forest가 가장 적합한지에 대한 이론적 근거를 보강하면 좋습니다.  (방법론적 보완: 데이터 전처리 과정(언더샘플링·오버샘플링)의 타당성 설명이 부족합니다. 또한 과적합 방지 및 교차검증 절차에 대한 구체적 기술이 필요합니다. 다양한 알고리즘을 비교한 점은 강점이나, 최종적으로 Random Forest를 주요 결과로 제시한 이유에 대해 이론적·실무적 근거를 보강하는 것이 바람직합니다.) |

**저자답변:** 중요한 지적을 해 주셔서 감사합니다. 총 3개의 질문으로 이해됩니다. (1) 데이터 전처리 과정의 타당성 보강 (2) 과적합 위험에 대한 검증 (3) Random Forest 선택의 이론적 근거. 우선 (1)에 대해서는 전처리 과정의 타당성을 높이기 위해 수치적인 근거를 반영하여 타당성을 높이고 아울러 기존보다 가독성을 높이기 위해 순서를 바꾸는 등 전면 재작성 하였습니다. 그리고 (2)에 대해서는, 본 연구에서 머신러닝과 딥러닝 알고리즘 모두에서 Validate 데이터로 하이퍼파라미터 튜닝 최적화를 통해 과적합을 대응하였습니다. 아울러, 하이퍼파라미터의 범위를 더욱 확장하여 실험을 다시 진행함으로써 최종 알고리즘은 XGBoost로 변경되었으며 예측 성능도 더욱 향상시킬 수 있었습니다. 관련 설명이 누락된 것으로 보여 본문과 실험결과를 모두 수정하였습니다. 마지막으로 (3)에서 XGBoost 알고리즘의 이론적 강점과 다른 알고리즘 대비 차별성, 그리고 데이터 특성과의 적합성과 실무적 의의를 “4.1 비형식학습 효과성 예측 평가(Train & Validate)” 부분에 아래와 같이 추가하였습니다.

**수정사항:**

|  |  |
| --- | --- |
| **[기존]**  (1) + (2)  (3)  없음 | **[변경]**  (1) + (2)  연구의 목표를 달성하기 위해 2018년부터 2022년까지의 “평생학습 개인실태조사” 데이터를 사용하였다. 데이터를 수집하는 설문조사의 내용이 큰 틀에서는 변화가 없다. 하지면 수집에 사용한 질문명이 조금씩 변경되기도 하기 때문에 2022년도를 기준으로 과거 데이터의 변수명을 동일하게 맞추었다. 또한 연도별로 수집되는 변수의 갯수는 차이가 있지만 5개년도 중 절반 이상인 3개년도 이상 공통적으로 존재하는 변수만을 선별하였다 <표 1>. 또한 샘플의 개수는 연도별로 차이가 있지만 값이 존재하지 않는 경우 결측치(NaN)으로 처리하고 연도별 구분을 위해 추가적인 연도 변수를 생성하였다. 데이터 병합이 완료된 후 모델링과 예측에 활용할 수 있는 전체 데이터는 51,369개의 샘플과 368개의 변수다. 이 중에서 변수들의 이상치 처리, 결측값 50% 이상 존재하는 변수들을 삭제하고 모델의 설명과 예측 성능을 위해 일부 변수에 대해 파생변수를 만들었다 <표 2>. 평균값을 통해 값을 일원화 하기도 하고, 유사 항목들의 여부를 확인후 이진화 하기도 하였다. 또한 학력, 소득, 직업 등 개인적 현황과 관련된 명목형 변수는 각 항목 값들에 따라 변수를 생성하여 설명하기 위해 더미변화화(Dummy Variable) 하여 최종적으로 103개의 변수로 변환되었다.  종속변수인 평생학습 효과성은 평생학습 참여가 삶의 질 향상에 얼마나 도움이 되었는지 “정신적 건강, 육체적 건강, 사회참여 만족도, 경제적 안정감” 4가지의 변수로 구분되어 있기에 4개 항목의 응답값 평균치를 기반으로 “평생학습 효과성” 변수를 생성하였다. 그리고 5점 척도 기준 4점 이상인 경우를 “만족” 그리고 중간값인 3점을 제외한 3점 미만의 경우를 “불만족”으로 변환하였다. 나머지 99개의 변수를 독립변수로 활용하였다.  그리고 전체 51,369명 중, 비형식교육 참여자는 18,117명(35.27%)이고 미참여자는 33,252명(64.73%)이다. 참여자는 18,117명은 8:2의 비율로 무작위 할당하여 각각 학습(Train)과 검증(Validate)로 활용하였다. 그리고 실제 비형식교육 미참여자 33,252명을 추천할 경우 어떤 만족도를 나타낼 것인지 예측에 활용하기 위한 미래 추천 테스트(Test) 데이터로 활용하였다.  그런데 학습 데이터를 사용해서 모델링을 할 때, 종속변수 라벨의 분포가 한쪽에 치우쳐진 경우 학습이 공정하게 이루어지지 못하는 불균형(Unbalanced) 상태로 인해 특정 라벨의 예측 성능이 떨어지는 이슈가 생긴다. 이러한 이슈를 해결하기 위해 일반적으로 샘플링을 통해 종속변수 라벨의 비율을 유사하게 반영하여 각 라벨별 학습이 공정하게 이루어질 수 있도록 반영할 수 있다. 본 연구의 학습 데이터가 대표적인 불균형 데이터로 비형식학습을 만족한 참여자는 5,721명이지만 만족하지 못한 참여자는 678명로 매우 적다. 이러한 경우, 만족한 사람들을 언더샘플링(Undersampling)하거나 불만족한 사람들을 오버샘플링(Oversampling) 할 수 있다. 우리는 비형식학습 만족도가 높은 사람들의 특징이 주된 관심사이고 동시에 가장 샘플의 수를 많이 보존할 수 있는 방식인 오버샘플링을 적용하였다. 마지막으로, 변수의 응답값을 0과 1사이로 값으로 동일하게 변환하여 학습과정에서 변수들의 비교를 공정하게 학습할 수 있도록 스케일링(Scaling) 하였다.  마지막으로, 모델링의 과정에서 학습데이터로 훈련을 하고 검증데이터로 하이퍼파라미터(Hyperparameter)를 최적화하여 학습데이터에 과도하게 적합하지 않고, 새로운 데이터에서도 안정적인 일반화 성능을 확보하는 방식으로 과적합을 대응하였다. 마지막으로 최종 추정된 하이퍼파라미터와 알고리즘을 사용해서 학습 및 검증데이터를 재학습 한 후, 테스트 데이터로 실제 미래 만족도를 예측하여 미래에 만족할 사람들은 어떤 요인을 나타낼지 분석하며 실험을 마무리하였다.  (3)  본 연구에서 다양한 머신러닝 알고리즘을 비교한 결과, 최종적으로 XGBoost가 가장 우수한 예측 성능을 보였다. 이는 단순히 실험적 결과를 넘어, XGBoost의 이론적 특성과 본 연구 데이터의 구조적 특성이 맞아떨어진 결과로 해석할 수 있다. 우선, XGBoost는 Gradient Boosting 기반 알고리즘으로, 변수 간 복잡한 비선형적 상호작용을 포착할 수 있으며, 정규화를 통해 과적합을 효과적으로 억제한다. Random Forest 대비 편향(Bias) 감소 효과를 확보하면서도 분산(Variance)를 안정적으로 제어한다. 반면 LightGBM은 Leaf-wise 성장 전략으로 속도는 빠르지만 소규모 데이터에서는 과적합 위험이 상대적으로 크고, CatBoost는 범주형 변수 처리에 강점이 있으나 본 연구와 같이 연속형 및 혼합형 변수가 중심인 데이터셋에서는 추가적인 이점을 크게 제공하지 못하였다.  따라서 XGBoost는 본 연구의 데이터 특성(변수 간 복잡한 관계, 잠재적 불균형 구조, 결측치 존재 가능성 등)에 최적으로 부합하며, 동시에 다양한 응용 분야에서 축적된 실무적 활용 경험을 통해 모델의 안정성, 일반화 가능성, 해석 가능성을 균형 있게 제공한다는 점에서 최종 결과를 대표하는 알고리즘으로 선택하는 것이 타당하다. 이러한 맥락에서 XGBoost는 본 연구 주제에 대한 가장 적합한 방법론적 선택일 뿐 아니라, 실제 정책 및 현장 적용 측면에서도 높은 활용 가치를 지닌다고 볼 수 있다. |

### 심사자 #1, 심사내용 #3 (유효정, 조은지, 이현서)

|  |
| --- |
| 본 연구의 주요 강점은 정책 현장 활용 가능성이지만, 구체적 제언이 부족합니다. 예를 들어, 개인정보 보호 문제, 디지털 격차, 지역별 학습 기회 차이 등 실제 정책 적용 시 발생할 수 있는 제약 요인을 논의해야 합니다. 이를 바탕으로 교육기관·정부가 실제 적용할 수 있는 구체적 실행 방안을 제시하면 논문의 실천적 가치가 크게 강화됩니다.  (정책적·실천적 시사점 구체화: 논문이 정책적 활용 가능성을 강조하고 있으나, 개인정보 보호, 디지털 격차, 지역별 학습 기회 불평등 등 실제 적용상의 제약 요인에 대한 논의가 부족합니다. 교육기관이나 정부가 현실적으로 적용할 수 있는 구체적 실행 전략을 제시하면 논문의 실천적 가치가 한층 강화될 것입니다. 종합적으로, 본 논문은 게재 잠재력이 충분하나 위의 사항들을 충실히 보완할 것을 권고합니다. 수정 후 게재(Revision Required)가 적합하다고 판단됩니다.) |

**저자답변:** 경영과학회에 발표하는 은지가 작성한 논문의 경우 “토의” 부분이 추가되었어요 최종버전에서. 이 리뷰어도 실험결과를 기반으로 토의를 작성하라는 뜻입니다. 아이데이션 해서 자유롭게 활용방법을 작성해보세요.

**수정사항:**

|  |  |
| --- | --- |
| **[기존]** | **[변경]** |

## 심사자 #2

### 심사자 #2, 심사내용 #1 (유효정, 조은지, 이현서)

|  |
| --- |
| 현재 1장에서는 서론과 문헌조사가 함께 제시되어 있습니다. 1장에는 본 연구의 연구배경과 이론적 배경을 간단하게 언급하고, 2장을 분리하여 문헌조사를 실시하십시오. 특히 한국 교육과 전세계 교육에 있어서 비형식 학습관련 비교를 시행해 주십시오. 왜 한국에서 이 평생학습 연구가 필요한지 그 필요성을 강조해 주십시오. 특히 비형식 학습 이외에 사이버대학 들도 평생교육의 측면에서 교육들을 시행하고 있는데 일반적인 평생학습과 사이버 대학과의 비교도 추가해 주십시오. 비형식 학습의 성공사례, 실패 사례에 대해서도 설명하고 본 연구의 필요성과도 연결해 주십시오. |

**저자답변:** 논문의 이해도를 높일 수 있는 구조적인 코멘트를 주셔서 감사합니다. 기존 1장을 2개의 장으로 분라하여 1장은 간결한 연구배경과 연구 목표 및 기여를, 2장에서는 개념, 통계, 선행연구, AI 관련 적용 흐름, 연구 필요성으로 묶어서 구조를 분리했습니다.

**추가사항:**

|  |
| --- |
|  |

### 심사자 #2, 심사내용 #2 (김경원)

|  |
| --- |
| 2장 문헌조사에서는 는 구조적 문제를 해결하기 위해 1장은 서론으로, 2장은 "이론적 배경 및 선행연구 검토"로 분리하여 논문의 논리적 체계성을 강화해야 합니다. 특히 2장에서 본 연구에서 사용하는 연구방법론들을 표로 비교 분석을 추가해 주십시오. 해당 방법론을 본 연구에서 사용해야 하는 이유에 대해서 보다 상세하게 제시해 주십시오. |

**저자답변:**

**수정사항:**

|  |  |
| --- | --- |
| [기존] | [변경] |
|  |  |

⬛

### 심사자 #2, 심사내용 #3 (김경원)

|  |
| --- |
| 2장에서는 추가적으로 본 연구에서 사용한 주요 변수들을 표로 제시해 주십시오. |

**저자답변:**

**수정사항:**

|  |  |
| --- | --- |
| [기존] | [변경] |

### 심사자 #2, 심사내용 #4 (유효정, 조은지, 이현서)

|  |
| --- |
| 표 4에서는 질문 문항을 전체적으로 제시하지 말고 요약하여 정리하고, 전체 문장은 appendix 형태로 추가해 주십시오 |

**저자답변:** 표 4를 복사해서 일단 맨뒤로 빼세요. 그리고 표 4의 질문들을 묶어서 어떠한 질문들로 구성되어 있는지 clustering 해서 작성해보세요. 예를 들어 개인정보 10문항, 직업정보 5문항 등등

**수정사항:**

|  |  |
| --- | --- |
| [기존] | [변경] |

### 심사자 #2, 심사내용 #5 (김경원)

|  |
| --- |
| 비형식 학습 이외의 기존의 연구들과 비교분석을 결론에서 추가해 주십시오. |

**저자답변:**

**수정사항:**

|  |  |
| --- | --- |
| [기존] | [변경] |

■