발표 1

머신러닝을 활용한 후속출산의도를 예측하기 위한 모델 구축 및 네트워크 분석

엄연용(건양대학교 심리상담치료학과 교수)

Ⅰ. 서론

후속출산이란 첫째 자녀를 출산한 후 둘째 이상의 자녀를 출산하는 것으로 개인의 선택으로 이어지는 출산 의도를 말한다(이정원, 2009). 생애주기 모델에 따르면 첫째 자녀를 출산한 후 둘째 또는 그 이상의 후속 출산의 터울이 저출산 현상을 예측하는 데 중요하다고 보았다(Morgan, & Taylor, 2006). Kohler(2001)도 저출산의 원인 중 하나로 첫째 자녀 출산 후 둘째 이상의 후속출산을 하지 않는 사회 현상을 주목하였다. 우리나라 경우도 저출산 특징 중 하나로 후속 출산이 이어지지 않는다는 점이다. 2021년 전국 출산력 및 가족보건·복지실태조사에서 기혼여성의 출생아 수 분포의 종단적 특성을 살펴보면 1명의 범주는 지속적으로 증가하였으나(2015년 24.4%, 2018년 26.4%, 2021년 30.0%) 2명~3명 이상은(2015년 55.9%, 2018년 53.3%, 2021년 46.0%) 감소하는 것으로 나타났다(박종서, 외 2021). 또한 인구동향조사에서도 첫째 출산(-4.6%)보다 둘째 출산(-11.4%)의 감소가 더 큰 것으로 나타났다(임영일, 강현영, 서주희, 김경미, 2024). 이러한 현상은 한국만의 문제는 아니며 저출산 국가에서 반복적으로 나타나 기혼 여성의 후속출산에 초점을 맞추어 다각적인 방면에서 살펴볼 필요가 있다(Basten et al, 2014; Sleebos, 2003). 따라서 이 연구의 목적은 후속출산의도를 예측요인을 탐색하고 주요 요인을 확인하며 요인 간 상호관련성을 밝히고자 한다. 이를 위해 개인요인, 양육요인, 환경요인, 경제요인, 정책요인을 등 다양한 요인들을 복합적으로 조사한 한국 영유아 교육·보육 패널 자료를 활용하고자 한다.

Ⅱ. 연구방법

1. 연구대상

이 연구는 2022년 한국 영유아 교육·보육 패널 자료에서 배우자가 있는 어머니 20대 358명, 30대 1,925명, 40대 204명, 총 2,487명의 자료를 사용하였다.

2. 측정도구

후속출산의지를 결과변수로 설정하고 개인요인, 양육요인, 경제요인, 환경요인, 정책요인의 다섯 가지 범주에서 25개의 예측요인으로 설정하였다.

3. 분석방법

Google의 Colaboratory(Colab)를 사용하여 Random forest, Gradientboosting, Logistic Regression, Stacking 모델을 pipeline으로 구축하였고 K-fold 교차검증, GridSearch로 하이퍼파라미터를 적용하였다. metrics라이브러리로 Confusion Matrix를 산출하였다. feature_importances로 예측 요인의 중요도를 시각화하였다. networkx, matplotlib를 활용하여 네트워크를 분석하였다.

Ⅲ. 연구결과

1. 집단별 후속출산 여부

후속출산의도 연령별로 살펴보면 〈표 1〉과 같다.

〈표 1〉 집단별 출산 의도

(N=2,487)

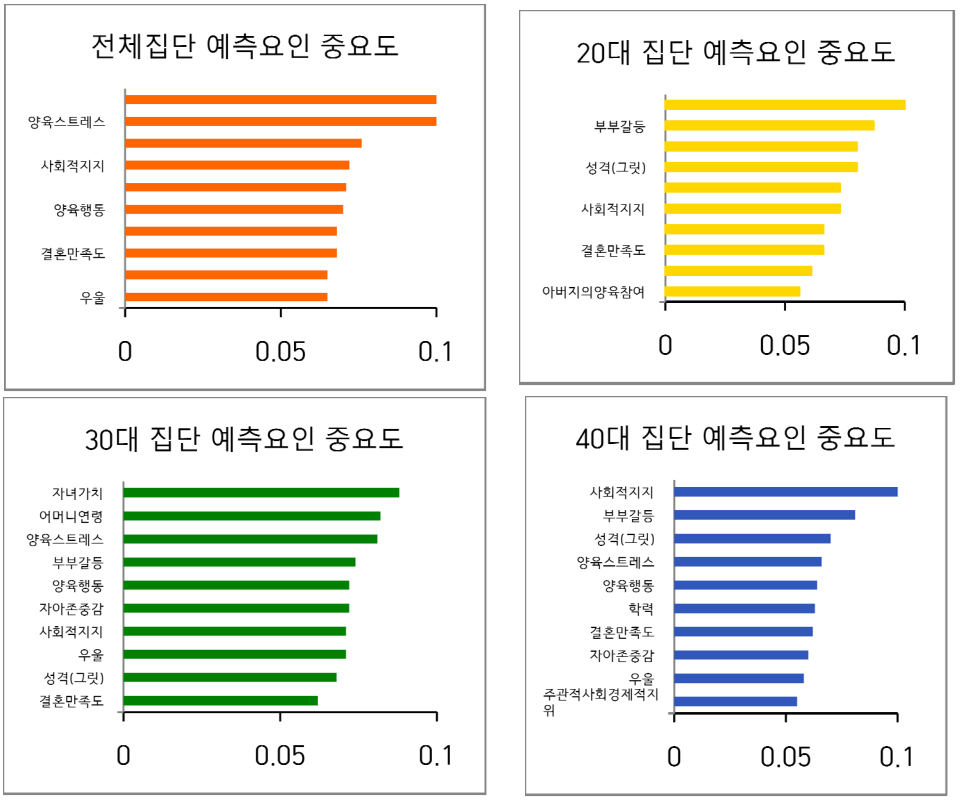
후속출산의도	20대(%)	30대(%)	40대(%)	Total(%)
낳지않음	164(45.8)	1,304(67.7)	181(88.7)	1,649(66.3)
낳겠음(또는 임신 중)	194(54.2)	621(32.3)	23(11.3)	838(33.7)

2. 후속출산의도 모델의 성능 평가

전체 집단에서 최적의 성능을 보인 모델은 Logistic Regression으로 CA .70, Precision .68, Recall .70, F1 .68이었다. 20대 집단은 Random Forest으로 CA .60, Precision .63, Recall .62, F1 .63이었다. 30대 집단은 Gradientboosting으로 CA .69, Precision .66, Recall .69, F1 .66이었다. 40대 집단은 Random Forest으로 CA .88, Precision .84 Recall .88, F1 .85이었다.

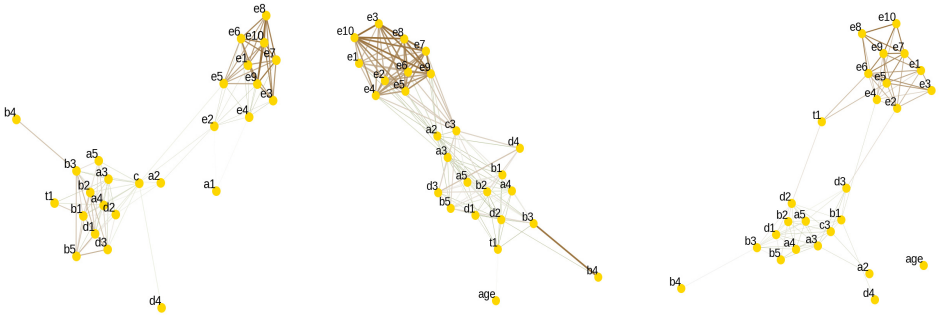
3. 예측 요인의 중요도

후속출산의도에 영향을 미치는 예측 요인의 중요도를 중요도를 산출하였다. 집단별로 살펴보면 전체 집단의 예측요인의 중요도는 [그림 1]과 같다.



[그림 1] 예측요인 중요도

4. 네트워크 분석



[그림 2] 네트워크 요인 분석

a1=어머니연령,a2=학력,a3=우울,a4=성격(그릿),a5=자아존중감,b1=양육스트레스,b2=양육행동,b3=자녀가치,b4=양육신념,b5=아버지의양육참여,c=주관적사회경제적지위,d1=결혼만족도,d2=부부갈등,d3=사회적지지,d4=지역사회양육환경만족도,e1=육아기근로시간단축,e2=육아휴직,e3=가족돌봄휴가,e4=임신기간근로시간단축,e5=시간외근로금지,e6=야간및휴일근로제한,e7=시차출퇴근제,e8=선택근무제,e9=재택근무제,e10=원격근무제,t1=후속출산의도

Ⅳ. 논의 및 결론

본 연구 결과 첫째, 후속출산의도를 예측하는 랜덤포레스트 모델, 그래디언부스팅 모델, 로지스틱 회귀 모델, 스태킹 모델은 F1 값을 기준으로 모두 .64~.85 예측력을 보였다. 둘째, 후속출산의도를 예측하는 중요도는 집단에 따라 상이하였다. 셋째, 네트워크 분석을 통해 후속 출산에 영향을 미치는 요인들의 상호 관련성은 집단에 따라 다르게 나타났다. 종합적인 지원 시스템을 구축하여 어머니들에게 다양한 측면에서 지원을 제공해야 할 필요가 있다. 본 연구 결과를 토대로 예측요인을 예측하고 지원하는 데 있어 중요한 자료와 방향성을 제시하였으며 각 연령대별로 맞춤형 정책과 프로그램을 개발하고, 어머니들의 정신건강과 양육지원을 강화함으로써 후속출산을 촉진하는 데 실증적인 자료로 활용될 수 있을 것이다.