기계 학습을 이용한 체외 수정 전 출생 가능성 예측

개요

체외 수정(IVF)은 인간 불임 문제를 해결하는 인기 있는 방법 중 하나로, 자궁내막증, 난자의 질 저하, 부모의 유전 질환, 배란 문제, 정자 또는 난자에 해를 끼치는 항체 문제, 자궁경부 점액에서 정자가 살아남지 못하는 문제, 낮은 정자 수 등의 다양한 원인으로 인해 선택됩니다. 그러나 IVF는 성공을 보장하지 않으며, 높은 비용과 불확실한 결과로 인해 부담이 큰 선택이 될 수 있습니다. IVF 과정에서 고려해야 할 요소가 많기 때문에, 의료진이 정확하게 출생 가능성을 예측하는 것은 어려운 일입니다.
본 연구에서는 인공 지능(AI)을 이용하여 출생 가능성을 예측하는 모델을 개발하였습니다. 특히, 난자 및 정자가 기증자가 아닌 부부에게서 유래된 경우를 중심으로 예측을 수행하였습니다. 공공 데이터셋인 인간 수정 및 배아학 관리국(HFEA)에서 제공한 데이터를 활용하여 다양한 AI 알고리즘(고전적인 기계 학습, 심층 학습, 앙상블 학습)을 비교 분석하였습니다. 실험을 통해 혼동 행렬, F1-점수, 정밀도, 재현율, ROC 곡선 등의 지표를 활용하여 모델 성능을 평가하였습니다. 특징 선택(feature selection) 기법을 적용한 경우와 적용하지 않은 경우를 비교하여 모델의 성능을 분석하였으며, 랜덤 포레스트(Random Forest) 모델이 특징 선택 없이 F1-점수 76.49%로 가장 높은 성능을 보였습니다.

연구 방법론

데이터셋 설명

이 연구에서 사용된 데이터는 2010년부터 2016년까지 HFEA에서 수집한 익명화된 등록 데이터입니다. 총 495,630개의 환자 기록과 94개의 임상 특성이 포함되어 있으며, 연구 목적에 따라 141,160개의 환자 기록을 필터링하여 사용하였습니다. 데이터는 수치형, 범주형, 텍스트형으로 구성되었으며, 연구에서는 임상적으로 유의미한 30개의 특성만을 선택하였습니다.

데이터 전처리

데이터 전처리 과정에서는 다음과 같은 단계를 수행하였습니다.

- 난자와 정자가 동일한 부부에게서 유래된 경우만을 포함
- 자극 사용 여부(stimulation used)가 기록된 데이터만 선택
- 비정상적인 연령(예: 999) 값 제거
- 연령 및 기타 범주형 데이터 인코딩
- 목표 변수(출생 여부)를 이진 분류 문제로 변환
- 데이터 불균형 문제를 해결하기 위해 음성 샘플을 줄여 균형을 맞춤
- 상관 행렬을 이용하여 상관관계가 높은 특성을 제거하여 데이터 차원 축소

모델 학습

본 연구에서는 기계 학습, 심층 학습, 앙상블 학습을 적용하여 출생 가능성을 예측하였습니다. 학습에 사용된 모델은 다음과 같습니다.

- 기계 학습: 로지스틱 회귀(Logistic Regression), K-최근접 이웃(KNN), 다층 퍼셉트론(MLP), 결정 트리(Decision Tree)
- 심층 학습: 1차원 신경망(1-D Neural Network)
- 앙상블 학습: 랜덤 포레스트(Random Forest), AdaBoost, 투표 분류기(Voting Classifier)

모델 학습은 두 가지 설정에서 진행되었습니다.

1. **특징 선택 없이 학습**: 전체 25개의 특성을 사용하여 모델을 학습
2. **특징 선택 후 학습**: 선형 SVC 및 트리 기반 특징 선택 기법을 적용하여 중요한 특성만을 선택하여 모델을 학습

결과 및 논의

특징 선택 없이 학습한 경우

모델 성능을 비교한 결과, 랜덤 포레스트 모델이 F1-점수 76.49%, 재현율 76%, ROC AUC 84.60%로 가장 우수한 성능을 보였습니다.

특징 선택 후 학습한 경우

선형 SVC 및 트리 기반 특징 선택을 적용한 결과, 전반적으로 성능이 하락하는 경향을 보였습니다. 가장 높은 성능을 보인 모델은 AdaBoost로, ROC AUC 77.60%를 기록하였습니다. 그러나 전체적으로 특징 선택을 적용한 경우보다 적용하지 않은 경우가 더 나은 결과를 보였습니다.

결론

본 연구에서는 기계 학습 및 심층 학습을 이용하여 IVF 시술 전 출생 가능성을 예측하는 모델을 개발하였습니다. 특히, 랜덤 포레스트 모델이 가장 우수한 성능을 보였으며, F1-점수 76.49%와 ROC AUC 84.60%를 기록하였습니다. 하지만 본 연구의 모델은 단일 데이터 출처(HFEA)에서 수집된 데이터에 기반하였기 때문에 일반화 가능성이 제한될 수 있습니다. 또한, 흡연, 음주, 카페인 섭취 등 생활 습관 요인이 포함되지 않아 예측 성능을 더욱 향상시킬 여지가 있습니다.

향후 연구에서는 보다 다양한 IVF 클리닉에서 데이터를 수집하여 모델의 일반화 성능을 개선하고, 생활 습관 요인을 포함하여 보다 정밀한 예측 모델을 개발하는 것이 필요합니다.