고려대학교 바이오의공학부 학생연구회 BEAT 18기 (2025-1)

활 동 보 고 서

|  |  |  |  |
| --- | --- | --- | --- |
| **성 명** | **유혜인**  **장현서**  **정다빈**  **황상엽** | **학 번** | 2024250017 |
| **2025년 5월 3일 4주차** | | **소속팀명** | 인공지능팀 |
| 1. 역할 분배  |  |  | | --- | --- | | 유혜인 | RandomForest  보고서 작성 | | 장현서 | SVM, MLP  ROC/PR curve/AUCROC/AUCPR 구하기 | | 정다빈 | Decisiontree  Confusion Matrix 그리기 | | 황상엽 | KNN  Hyperparameter 구하기 |  1. 데이터 전처리     뇌졸중 예측 데이터를 가져와 정보를 확인한다. bmi와 smoking\_status에 결측치가 존재하므로 이에 대해 imputation을 진행한다.    bmi의 결측치는 평균값으로 대체하고, smoking\_status의 결측치는 ‘unknown’으로 대체한다.    클래스 불균형을 완화하기 위해 오버샘플링을 진행한다.    모델 학습을 위해 특성과 타킷을 분리한다. stroke을 제외한 나머지 열을 특성으로, stroke 열을 타깃으로 설정한다.    모델 학습을 위해 문자열 데이터를 수치형 데이터로 변환한다.    데이터를 학습용 데이터와 평가용 데이터로 분할한다. 전체 데이터의 20%를 학습용 데이터로 사용한다.    데이터의 평균이 0, 표준편차가 1이 되도록 스케일링을 진행한다.   1. 모델 선정 및 성능 평가     5개의 모델 KNN, DecisionTree, RandomForest, SVM, MLP를 선정하여 교차검증과 평가 지표를 설정한다. 평가 지표는 f1\_score, ROC AUC, PR AUC로, 세 개의 평가 지표를 복합적으로 평가한다.    각각의 모델에 5-fold 교차검증을 차례대로 진행하고, 평가 지표에 대한 값을 출력한다.   1. 최종 모델 선택 및 성능 평가  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | KNN | DecisionTree | RandomForest | SVM | MLP | | f1\_score | 0.965 | 0.800 | 0.999 | 0.852 | 0.980 | | ROC AUC | 0.987 | 0.861 | 1.000 | 0.909 | 0.988 | | PR AUC | 0.975 | 0.811 | 1.000 | 0.849 | 0.974 |   각각의 모델에 대한 평가 지표는 위의 표와 같다.  RandomForest는 overfitting으로 판단하여 최종 모델은 MLP로 선택한다.        hyperparameter를 구하고, Grid Search 교차검증을 진행한다.    위에서 찾은 최적의 parameter를 적용하여 최종 MLP 모델을 완성하고, 학습용 데이터를 재학습한다.      최종 학습된 MLP 모델을 평가용 데이터에 적용해 예측 결과를 얻고, 이를 confusion matrix로 시각화한다.    ROC 곡선과 PR 곡선을 그리고, ROC AUC와 PR AUC를 각각 계산한다.    계산 결과 ROC AUC는 0.988, PR AUC는 0.978로, 모두 1에 거의 근접하다. 따라서 예측 결과가 거의 정확함을 확인할 수 있다. | | | |