PBL 과제수행계획서 (학생용)

팀명	E조	
문제	문제 번호	주요내용
	1	**주어진 데이터를 활용하여 수업시간에 배운 머신러닝 모델을 적용하고, 최적의 하이퍼파라메타(n_neighbors)를 구하고, 모델의 성능을 평가하시오.**
학습목표	1) feature들을 해석하기 2) one-hot encoding과 scaling 등 data preprocessing 을 수행하기 3) SVM, Logistic Regression, KNN, RandomForestClassifier, Voting Classifier model 등 지금까지 학습한 모델 중에서 두 가지 이상을 사용하여 최적의 파라메타를 찾아서 예측하기 4) 성능을 평가	
가정/해결안	1) 구글링을 통해 feature의 뜻을 파악한다. 2) 범주형 데이터에 원핫인코딩을 적용하고, KNN일 때만 연속형 데이터에 Min-Max 스케일링을 적용한다. 3) KNN 모델과 의사결정나무, Random Forest의 최적의 파라메타를 Random Search로 대략 갈피를 잡은 후 Grid Search를 통해 정확한 값을 구한다. 4) Confusion Matrix를 이용해 각 모델의 정확도와 재현율을 확인한다.	
이미 알고 있는 사실	heart.csv에서 종속변수 y값은 HeartDisease로 정해져 있다. 최적의 하이퍼파라메터를 구하는 방법은 Grid Search를 통해 구할 수 있다. 다양한 머신러닝 모델 기법(Random Forest, KNN, 의사결정나무, SVC, 회귀분석 등)과 모델의 성능을 평가하는 방법.	
더 알아야 할 사실	앞으로 더 심화된 머신러닝의 기법을 배우고 데이터를 실질적으로 분석하고 싶다. 모델의 성능을 평가할 수 있는 다른 기법이 더 있는지 궁금하다.	
학습자원 (참고자료)	**수업시간에 배운 실습파일, 구글링(feature 설명, 이상치 보간 등)**	