PBL 과제수행계획서 (학생용)

<table>
<tr><td>팀명</td><td colspan="2">H</td></tr>
<tr><td rowspan="2">문제</td><td>문제 번호</td><td>주요내용</td></tr>
<tr><td>1</td><td>주어진 산업데이터를 활용하여 KNN 머신러닝 모델을 적용하고, 최적의 하이퍼파라메타(n_neighbors)를 구하고, 모델의 성능을 평가하시오.</td></tr>
<tr><td>학습목표</td><td colspan="2">① 연속형 데이터를 분석하고 시각화한 후 스케일링을 적용한다.
② KNN 모델에 대한 최적의 하이퍼파라메타를 구한다.
③ 하이퍼파라메타를 적용하여 모델의 성능을 평가한다.</td></tr>
<tr><td>가정/해결안</td><td colspan="2">처음에는 train set과 test set의 분리의 퍼센트에 따라서 결과가 달라질 것이라 예상했다. 그래서 random_state=7, stratify=y로 설정하였고, test_size를 변경했을 때 유의미한 결과를 가져올 것으로 가정하였다. 하지만 test_size만 변경하며 비교하였을 때, 유사한 결과만 나올 뿐 유의미한 결과는 없었다.
다음으로는 그리드 서치에서 best param을 뽑을 때, cv를 높이는 것이 더 좋은 결과를 가져올 것으로 가정하고 모델을 비교해 봤지만, cv를 변경해도 정확도를 더 높일 수 없었다.
n_neighbors의 경향성을 파악하기 위해 random search의 실행 결과를 n_neighbors를 기준으로 정렬하였고, 그 결과 전부 약 15이하일 때 가장 높은 정확도를 가지고 우하향 하는 그래프를 그린다는 공통점을 확인했다. 그래서 이 데이터 값에서는 n_neighbors가 대략적으로 15 이내에서 best param을 가진다는 경향성을 확인할 수 있었다.
이후 파악한 경향성에 따라 grid search의 범위 값을 0~15값으로 제한하여 Train 데이터에 gri search를 적용한 결과 best param가 3이 나왔고 이 값을 knn 모델에 적용하였다.
그 결과 test_size와 cv의 변화에 따라 정확도가 0.80~0.832 정도로 좁혀졌다. 이후 큰 변화 값이 보이지 않았기 때문에 random_state를 바꿔보기 시작했다. 그 결과 유의미한 결과를 보이기 시작했고 random_state가 42일 때 가장 좋은 정확도가 나왔다.
이미 앞에서 random search를 통해 15 이후 우하향 하는 경향성을 파악했기 때문에 grid search에서는 0~15의 범위 값으로 고정하여 best param을 추출하고 해당 값을 knn모델에 적용했다. doit 함수를 만들어서 파라미터로 random_state, test_size, cv만 변화를 주어 결과 값을 확인했다. doit 함수에 위의 3개의 값에 변화를 주며 확인한 결과 random_state가 42, test_size가 0.25, cv가 5일 때 가장 좋은 정확도인 0.85가 나왔다. 추가로 KNN모델 생성시 유클리디안 거리와 멘허튼 거리일 때도 비교해 보았으나 기본값인 유클리디안 거리일 때 가장 좋은 정확도가 나왔다.</td></tr>
<tr><td>이미 알고 있는 사실</td><td colspan="2">test_size의 값을 설정해서 훈련비율과 테스트 비율을 조절할 수 있는데 test 데이터 비율을 너무 줄이면 정확도가 오히려 낮아진다. 이를 염두에 두고 적절하게 비율을 설정하는 것이 중요하다.
데이터의 양이 많을수록 정규분포 형태로 수렴하게 되는데 일부 데이터를 가져오게</td></tr>
</table>

	되면 한쪽에 몰릴 수 있어서 골고루 퍼지게 설정하는 것이 중요하다. 데이터를 나눌 때 한쪽에 비중이 치우치지 않게 골고루 train과 test set에 분배하도록 하기 위해서 y를 이용해 stratify=y를 설정한다. 추후에 동일한 코드를 반복해도 랜덤으로 가져오는 동작을 동일하게 하기 위해서 random_state=k(k는 정수)를 설정한다.
더 알아야 할 사실	정확도에 영향을 줄 수 있는 다른 파라미터가 있을까? knn모델 이외에 본 데이터에 더 알맞은 모델은 무엇이 있을까? 전처리 과정에서 다른 스케일링을 사용했을 때 결과값이 어떻게 달라질까? 위의 3가지 물음에 대해 더 알아야 할 것 같다.
학습자원 (참고자료)	**지금까지 수업시간 및 팀 활동 때 작성했던 코드들, 수업 강의안**