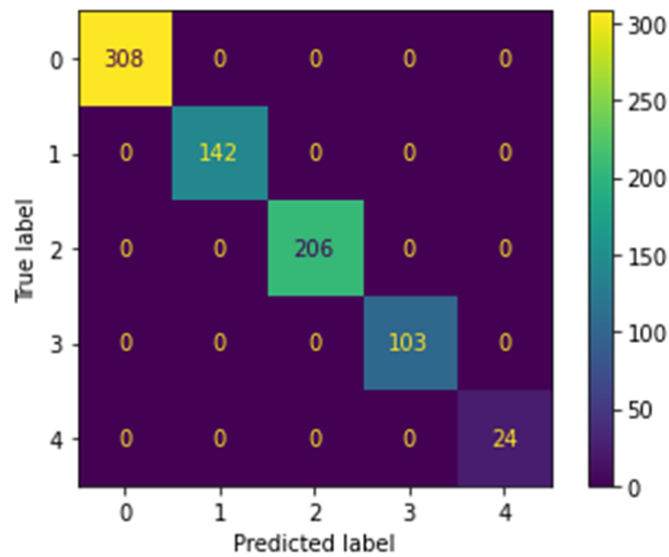








보고서

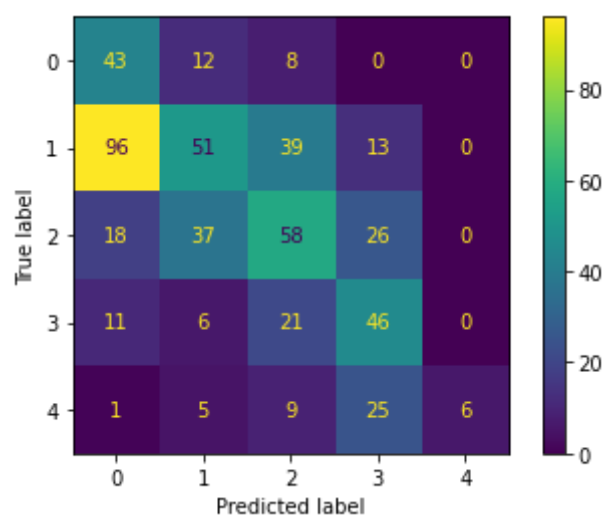


다음과 같은 결과를 보았을 때에는 100%의 분류를 해내었고,

Practice		Challenge		Search	
	ideal Score: 99.7957		akashkewar Score: 96.8335		ijeffking Score: 73.1359
Rank	Participant/team	Score	Time		
1	 ideal	99.7957	28/10/21 11:32 pm		
2	 akashkewar	96.8335	27/05/21 11:20 am		
3	 ijeffking	73.1359	24/05/21 06:41 pm		

test data의 정확도도 99.7957%로 상당히 높은 결과를 나타냈었으나,

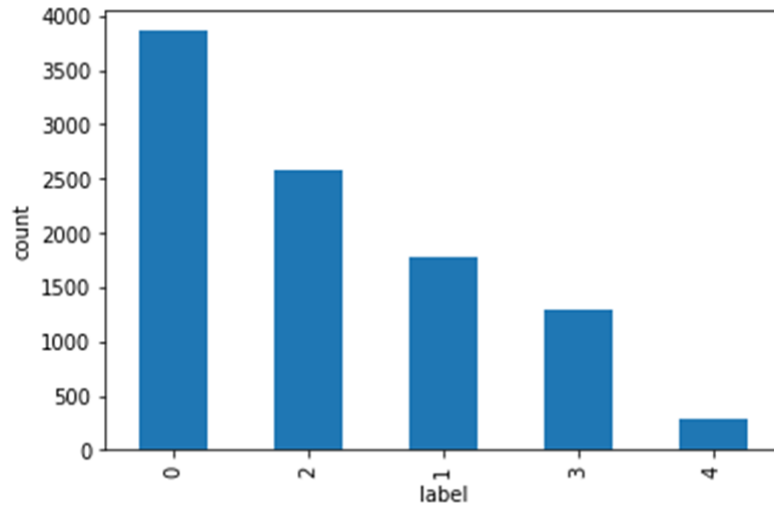
교수님께서 주신 데이터로 돌린 결과



다음과 같이 약 38%의 결과를 보이면서 아주 낮은 성능을 보였습니다.
때문에 해당 원인을 찾아보니 Data Leakage 현상이 발생한 것 같습니다.

높은 정확도를 보인 원인을 분석 과정

DPhi의 test dataset의 분포는 아래와 같고,



수치적으로는

num0 = 39.42798774259448%

num1 = 18.07967313585291%

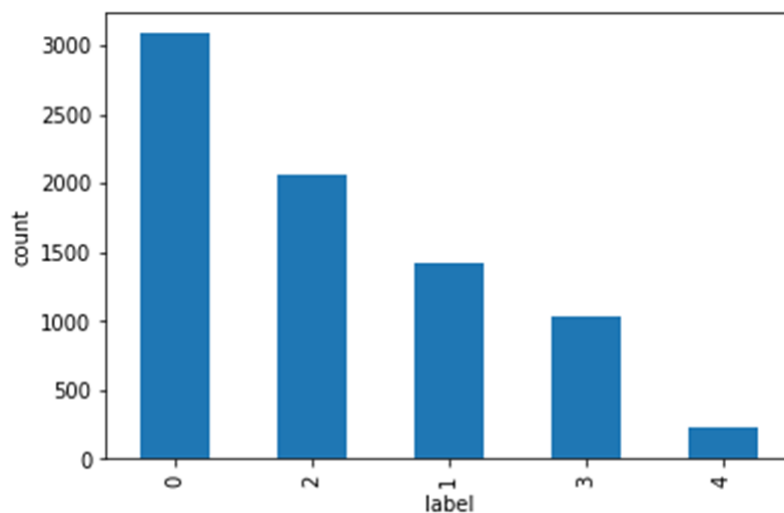
num2 = 26.353421859039837%

num3 = 13.125638406537282%

num4 = 3.0132788559754853%

데이터 양이 많은 순서 : 0 > 2 > 1 > 3 > 4

train dataset의 분포는 아래와 같습니다.



nn0 = 39.413447782546496%

nn1 = 18.087063151440833%

nn2 = 26.343756386674844%

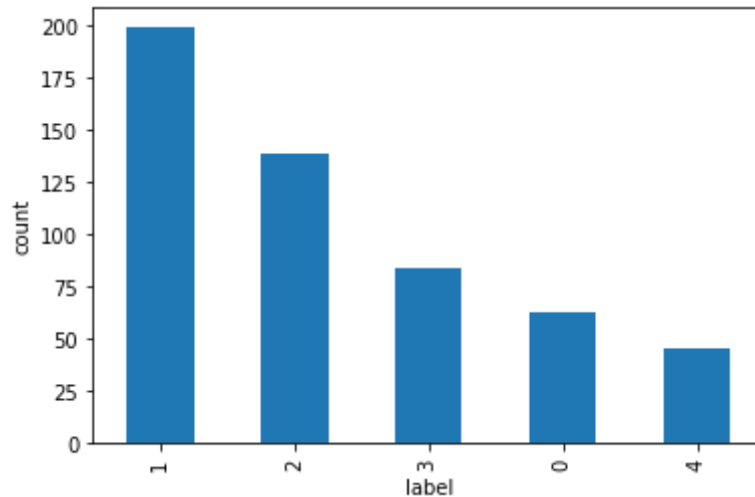
nn3 = 13.14122215409769%

nn4 = 3.014510525240139%

데이터 양이 많은 순서 : 0 > 2 > 1 > 3 > 4

두가지의 데이터셋이 동일한 분포를 가지고 있었고, 학습 데이터와 테스트 데이터의 분포가 엄청나게 유사하여
정확도가 높게 나온것 같습니다.

이렇게 생각하는 이유는



저희가 사용한 테스트 데이터셋의 분포는 1 > 2 > 3 > 0 > 4 순이지만, 학습 데이터셋의 분포는 그렇지 않았습니다.

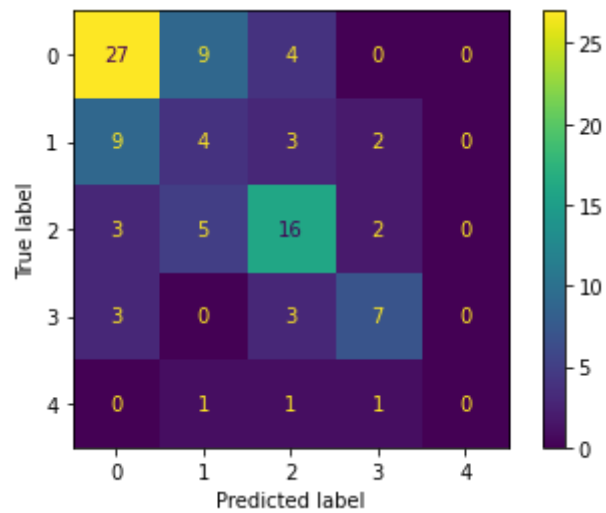
테스트 데이터셋은

label	사용된 데이터의 개수	모델이 예측한 라벨별 개수
0	63	169
1	199	111
2	139	135
3	84	110
4	46	6

예측된 개수의 분포를 보시면 0 > 2 > 1 > 3 > 4로 학습 데이터의 분포와 동일하게 예측의 분포가 맞춰지는 것을 알 수 있습니다.

하지만, 저희가 사용한 test 데이터셋의 분포를 학습데이터와 동일한 맞춘 경우의 분류 성능은 아래와 같습니다.

label	사용된 데이터의 개수	모델이 예측한 라벨별 개수
0	40	42
1	18	19
2	26	27
3	13	12
4	3	0



0 > 2 > 1 > 3 > 4의 순으로 예측 분포를 맞추는 것이 동일하였으나,
정확도 측면에서 약 54%로 좋지 못한 정확도를 보였습니다.

결과적으로, 99.80%로 엄청나게 높은 정확도가 나온 이유는 Data Leakage 현상과 학습 데이터와 테스트 데이터의 분포가 동일하였기 때문이라고 생각합니다.