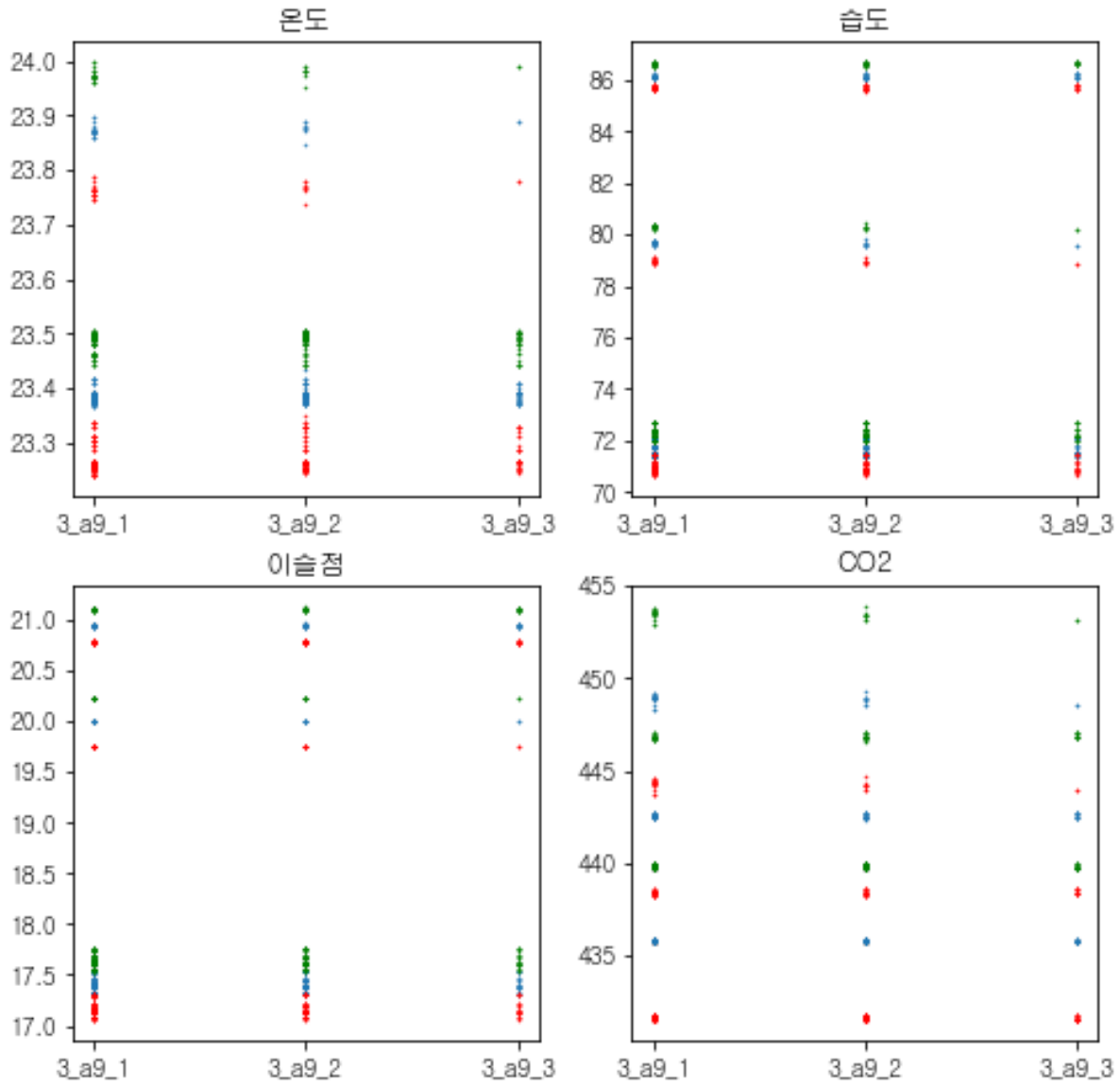
식물병해충 AI 경진대회 환경데이터 분석

1. 파프리카_다량원소결핍(P)_초기 (3_b7_1)는 내부 습도 평균이 낮은 쪽으로 이상치가 많이 발생하는 현상을 보인다. 또한, train데이터 셋에 파프리카 다량원소결핍(P)의 경우 초기만 존재하므로 이 병의 발생원인 중 하나로 낮은 습도를 꼽을 수 있다. 다만, 이미지만의 구분에서도 틀리지 않고 구분해내는 질병코드이다.

2. 파프리카_파프리카흰가루병(3_a9)은 6월 20, 27일, 7월 4일 데이터 측정 시작일이 세 개의 날이 존재하고 모두 비슷한 기간으로 계절의 영향은 크지 않은 것으로 판단됨.

3.



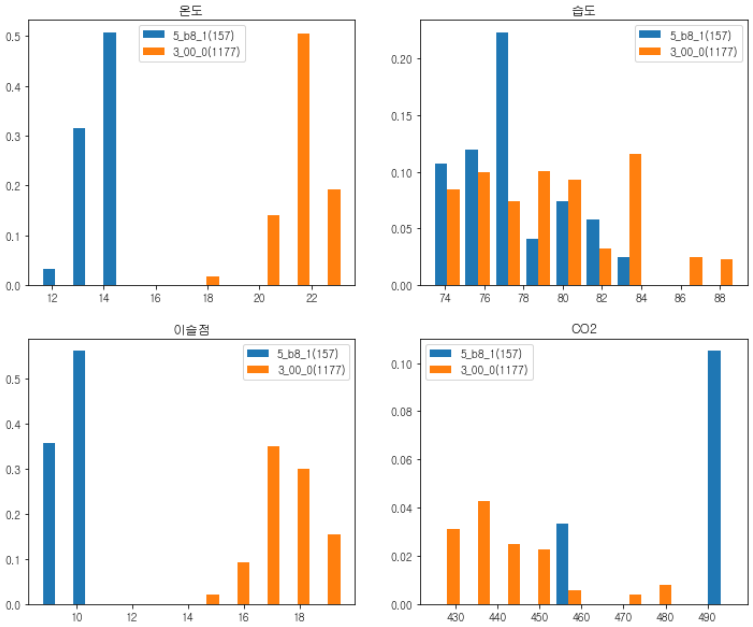
파프리카흰가루병(3_a9)의 진행정도를 기준으로 온도, 습도, 이슬점, CO2의 데이터 분포를 비교해보았을 때 진행정도를 구분할 정도의 특색있는 분포를 보이지 않음. 따라서, 파프리카흰가루병의 진행정도를 구분하기 위해서는 이미지 데이터로 또 다른 모델을 만드는 것이 더 좋은 방법이라고 생각됨. -> 학습 v3 결과 중 파프리카흰가루병을 제외하고 예측결과가 상대적으로 낮은 레이블에 대해서 환경변수의 분포 비교가 필요함.

4. 예측이 상대적으로 떨어지는 분류

번	정답	예측	오답률(%)
1	고추_다량원소결핍(K)_초기(5_b8_1)	파프리카_정상(3_00)	0.6369
2	고추_다량원소결핍(N)_초기(5_b6_1)	고추_다량원소결핍(K)_초기(5_b8_1)	0.6757
3	토마토_토마토흰가루병_중기(2_a5_2)	시설포도_정상(6_00)	0.5291
4	파프리카_다량원소결핍(K)_초기(3_b8_1)	파프리카_파프리카흰가루병_중기(3_a9_2)	0.6536
5	파프리카_다량원소결핍(N)_초기(3_b6_1)	파프리카_다량원소결핍(K)_초기(3_b8_1)	0.7042
6	파프리카_정상(3_00)	오이_정상(4_00)	0.0850

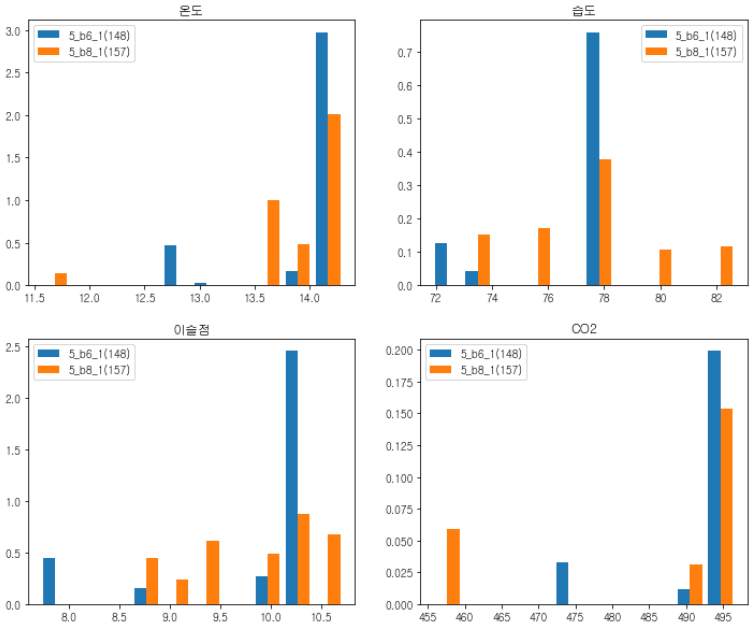
각 라벨의 모든 객체들의 온도, 습도, 이슬점, CO2 데이터의 분포를 시각화해서 의미있게 사용할 수 있는지 확인해보았다.

(1)



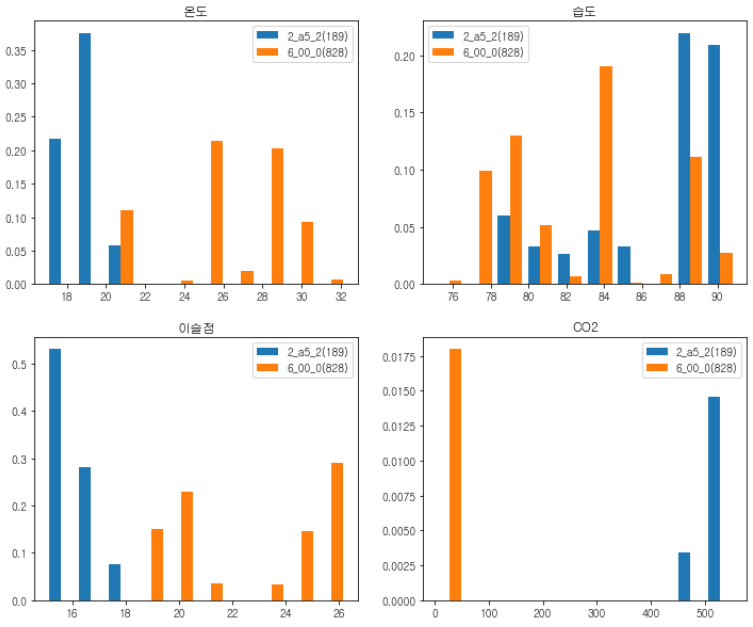
온도와 이슬점면에서 확실한 차이가 보여진다. 습도에서는 분포로써 라벨을 분류하기엔 적합하지 않고 CO2는 나뉘는 모습이긴 하지만 조금 더 살펴볼 필요가 있다..

(2)



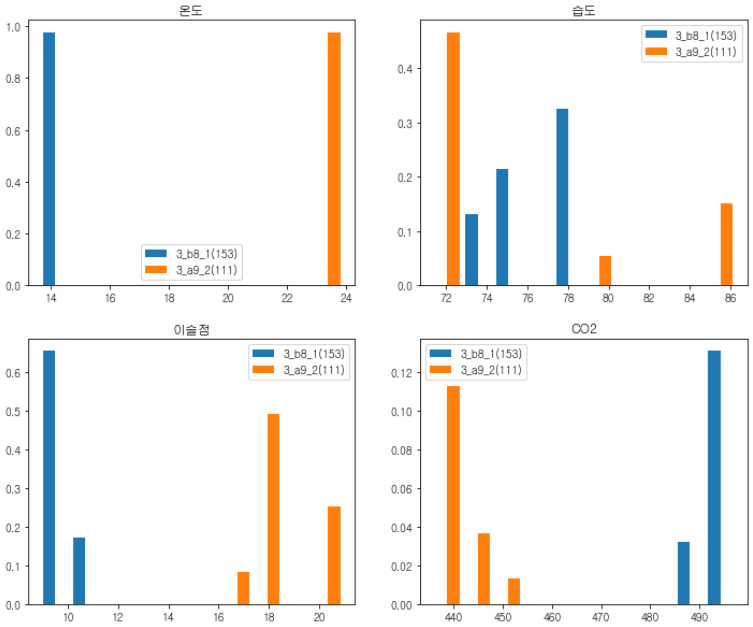
모든 면에서 분포에 큰 차이가 없는 것으로 나타난다. 같은 작물에 비슷한 종류의 질병이라 환경 데이터가 미치는 영향이 크지 않다고 볼 수 있다. 따라서, 파프리카흰가루병 진행정도 구분과 마찬가지로 이미지를 활용한 학습방식의 변화로 정답률을 올리는 것이 더 좋은 방법으로 생각된다..

(3)



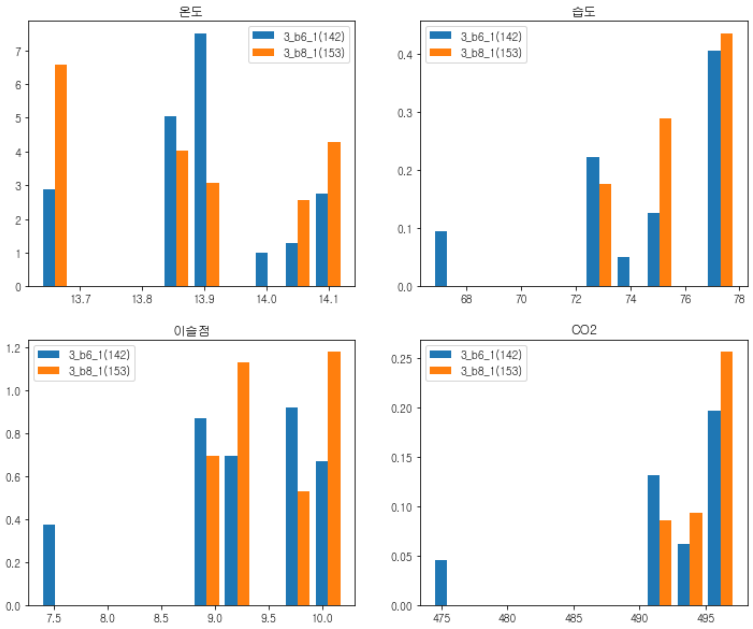
1번 비교와 마찬가지로 온도와 이슬점에서 확실한 차이가 보여진다. 습도는 분포가 비슷하고 CO2에서 확연한 차이가 보이지만 사실상 6_00_0의 값들은 거의 0에 수렴한다(데이터 전처리 과정에서 nan값이 대체되면서 이러한 모습이 나옴). 따라서, 온도와 이슬점을 활용하는 것이 두 객체 분류에 도움을 줄 것으로 예상됨.

(4)



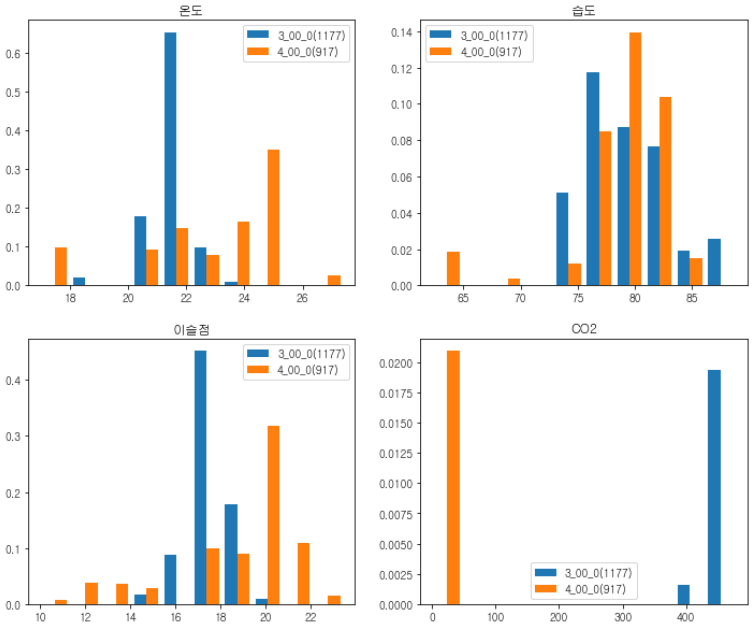
온도와 이슬점에서 분포에 대한 차이를 보이고, 습도는 분포가 비슷하다. 다만, CO2에서 확연한 차이가 드러남. 같은 파프리카 작물에서 서로 다른 질병이지만 환경데이터의 분포가 극명하게 나타난다. 0.6535%의 오답률 해소를 위해 환경데이터 중 CO2를 포함시키는 것도 하나의 방법이 될 것이라고 생각된다.

(5)



2번 비교와 마찬가지로 같은 작물 비슷한 질병이라 환경 데이터의 분포가 거의 동일하게 나타나는 것으로 보인다.

(6)



서로 다른 작물의 정상 객체들을 구분하는 비교이다. CO2를 제외한 세 변수의 분포는 거의 동일하게 나타난다. 따라서, 환경데이터에 CO2를 포함하여 학습시키는 것이 좋은 방법이라고 생각된다.

5. 4번과 동일한 방식으로 각 환경데이터들을 평균 값에서 최소, 최대값들로 바꿔서 비교를 진행. 평균과 최소, 최대값들의 분포가 거의 똑같은 모습을 보인다.

6. 결론 : 환경데이터 종류 중 결측치 값이 높지 않은 컬럼들 12가지를 골라서 온도, 습도, 이슬점, CO2의 평균값, 이 중 온도 평균, 이슬점 평균, CO2 평균 3가지 값만 가지고 학습을 진행하는 것이 좋아보인다.