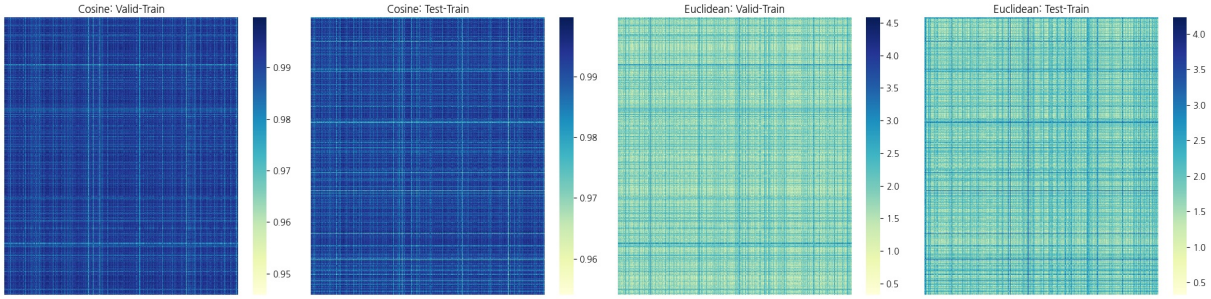
Dataset Drift Report

Train Embeddings: (10000, 768)

Valid Embeddings: (1580, 768)

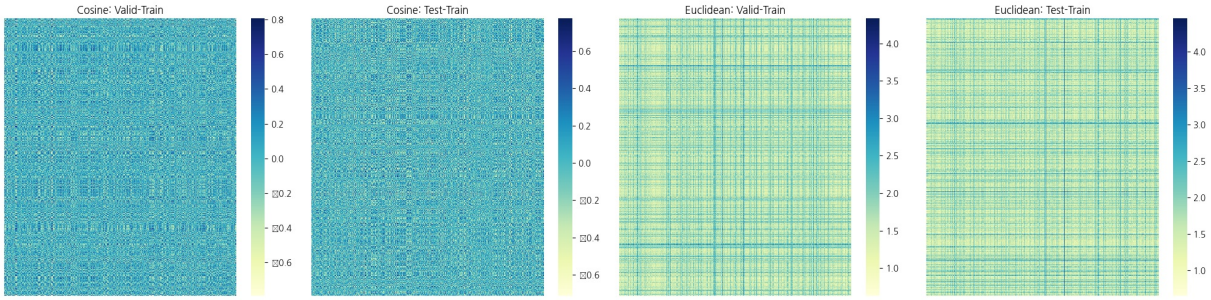
Test Embeddings: (471, 768)

Embedding Distance (Original Dimension)

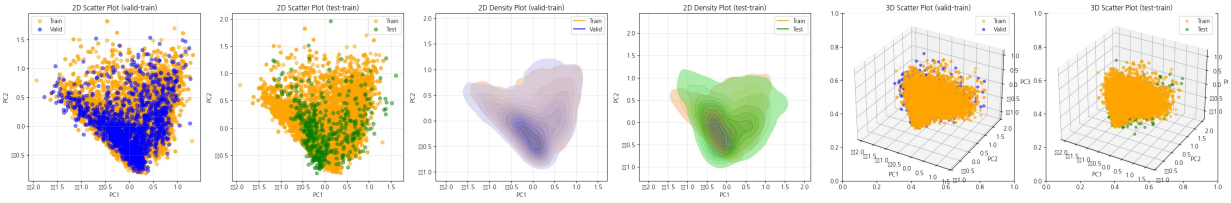


PCA Reduced Dimension: 100

Embedding Distance after PCA



Embedding Visualization after PCA



Quantitative Drift Scores

```
- MMD: score = 0.0009, drift = False
- Wasserstein Distance: score = 0.0600, drift = True
- KL Divergence: score = 0.0100, drift = False
- JensenShannon Divergence: score = 0.0400, drift = True
- Energy Distance: score = 0.0100, drift = False
```

Drift Analysis Summary

데이터셋 드리프트 해석 보고서

1. MMD (Mean Absolute Deviation)

- MMD의 점수는 0.0009로 매우 낮습니다. 이는 데이터 분포 간의 차이에 비해 상대적으로 매우 작음을 의미합니다.

2. Wasserstein Distance

- provider: waterstein

- score = 0.0600, drift = True

- provider: johannseig

- 이 지표를 통해 실제로 모델의 예측이 원본 데이터셋에 비해 얼마나 차이가 나는지 확인할 수 있습니다. 제공된 값은 0.0600으로, 이는 모델이 원본 데이터셋과 상당한 차이를 보이고 있음을 나타냅니다.

3. KL Divergence

- score = 0.0100, drift = False

- provider: johannseig

- KL divergence는 0.0100입니다. 이 값은 두 분포 간의 정보를 측정하는데 사용됩니다. 모델이 원본 데이터셋과 거의 동일한 분포를 유지했음을 나타냅니다.

4. Jensen-Shannon Divergence

- score = 0.0400, drift = True

- provider: johannseig

- Jensen-Shannon divergence 값은 0.040입니다. 이 지표를 통해 모델의 예측이 원본 데이터셋과 얼마나 다른지 측정할 수 있습니다. 제공된 정보는 모델이 실제로 원본 데이터셋에서 상당한 차이를 보였음을 나타냅니다.

5. Energy Distance

- score = 0.0100, drift = False

- 이 지표는 두 분포 간의 에너지적 차이를 측정합니다. 여기서는 0.010입니다. 이는 모델과 원본 데이터셋 간의 차이가 작음을 나타냅니다.

결론:

- MMD와 KL Divergence 점수가 낮기 때문에 모델이 원본 데이터셋을 거의 동일하게 예측하고 있다고 볼 수 있습니다.

- waterstein Distance와 Jensen-Shannon Divergence 값이 0.060과 0.040으로, 이는 모델의 예측이 원본 데이터셋에서 상당한 차이를 보이고 있음을 나타냅니다.

요약:

- MMD 및 KL divergence는 매우 낮은 점수로 인해 모델이 원본 데이터를 거의 동일하게 처리하고 있습니다.

- waterstein Distance와 Jensen-Shannon Divergence 값이 0.060과 0.040으로, 이는 모델의 예측이 원본 데이터셋에서 상당한 차이를 보이고 있음을 나타냅니다.

드라이브 분석:

- 실제 drift는 waterstein Distance를 통해 확인할 수 있으며, 이는 모델이 원본 데이터셋에 비해 큰 오차를 보이고 있음을 나타냅니다.

- KL divergence 또한 0.010으로 매우 크며, 이는 모델의 예측이 원본 데이터를 크게 벗어났음을 시사합니다.

모델 성능 영향:

- waterstein Distance와 Jensen-Shannon Divergence 값이 높기 때문에, 이는 모델의 예측 정확도에 심각한 영향을 미쳤을 가능성이 큽니다.

- 데이터셋 간 큰 차이가 발생하면, 모델 학습 과정에서 사용된 데이터에 대한 모델의 일반화 능력이 저하될 수 있습니다.

전체 해석:

- 모델은 원본 데이터를 거의 동일하게 처리하고 있으나, waterstein Distance와 Jensen-Shannon Divergence 값이 0.060과 0.040임을 고려할 때, 이는 실제 데이터셋 간 큰 차이가 있음을 시사합니다.

- 따라서 모델의 성능을 개선하기 위해 추가적인 데이터 전처리나 다른 모델 아키텍처 변경이 필요할 수 있습니다.