# Law 통합 분석 리포트

데이터 드리프트 분석 보고서

생성일시: 2025년 07월 03일 12시 48분

## Dataset Information & Statistics

### Train Dataset

**Total Rows:** 4078
**Columns:** 4 (class, text, cleaned_facts, doc_len)

**Preview**

| Index | class | text | cleaned_facts | doc_len |
|---|---|---|---|---|
| 1 | i | 2) 또한 자동차 열쇠에 대한 배타적인 점유를 취득하는 이상 인근에 주차된 해당 자동차를 사실상 지배하는 것은 매우 용이하고, | 자동차 열쇠 배타 점유 취득 이상 인근 주차 해당 자동차 지배 | 18 |
| 2 | w | 자기의 행위로 인하여 타인의 사망이라는 결과를 발생시킬 만한 가능성 또는 위험이 있음을 인식하거나 예견하면 족한 것이며 그 인식이나 예견은 확정적인 것은 물론 불확정적인 것이... | 자기 행위 타인 사망 결과 발생 가능 위험 인식 예견 인식 예견 확정 확정 미필 고의 인정 바 | 29 |
| 3 | u | 이 사건의 경우 앞서 든 증거들에 의하여 인정되는 다음과 같은 사정들을 고려하여보면 | 사건 경우 증거 인정 다음 사정 고려 | 12 |
| 4 | c | 나. 피고인은 본건 범행 당시 술에 취하여 심신상실 또는 심신미약의 상태에 있었다. | 피고인 범행 당시 술 심신 상실 심신 미약 상태 | 12 |
| 5 | d | ② 피해자가 피고인에게 먼저 시비를 걸며 부엌에 있던 식칼을 꺼내들고 피고인을 위협하자 피고인은 이에 대응하여 피해자를 밀어 넘어뜨리고 판시와 같이 피해자를 식칼로 찔렀는데, | 피해자 피고인 시비 부엌 식칼 피고인 위협 피고인 대응 피해자 판시 피해자 식칼 | 23 |

### Valid Dataset

**Total Rows:** 1935
**Columns:** 4 (class, text, cleaned_facts, doc_len)

**Preview**

| Index | class | text | cleaned_facts | doc_len |
|---|---|---|---|---|
| 1 | i | 피해자가 입은 상해의 부위 및 정도는 사망에 이를 수 있는 정도의 위험한 수준이었고, | 피해자 상해 부위 정도 사망 정도 위험 수준 | 13 |
| 2 | c | 이 사건 범행 당시 피고인에게 미필적으로나마 살 | 사건 범행 당시 피고인 | 13 |

| | | | | |
|---|---|---|---|---|
| | | 인의 범의가 있었음을 충분히 인정할 수 있다. | 미필 살인 범의 인정 | |
| 3 | e | 이러한 손상은 둔기에 의해 발생 가능한 손상으로 보인다는 소견을 밝히고 있는 점, | 손상 둔기 발생 가능 손상 소견 | 12 |
| 4 | c | 피고인 A에게 살인의 고의는 없었고, | 피고인 살인 고의 | 5 |
| 5 | c | 피고인이 이 사건 범행 당시 음주로 인하여 사물을 변별할 능력이나 의사를 결정할 능력을 상실하였다거나 그 능력이 미약한 상태에 이르렀다고는 보이지 아니하므로, | 피고인 사건 범행 당시 사물 변별 능력 사 결정 능력 상실 능력 상태 | 21 |

## Test Dataset

**Total Rows:** 780
**Columns:** 4 (class, text, cleaned_facts, doc_len)

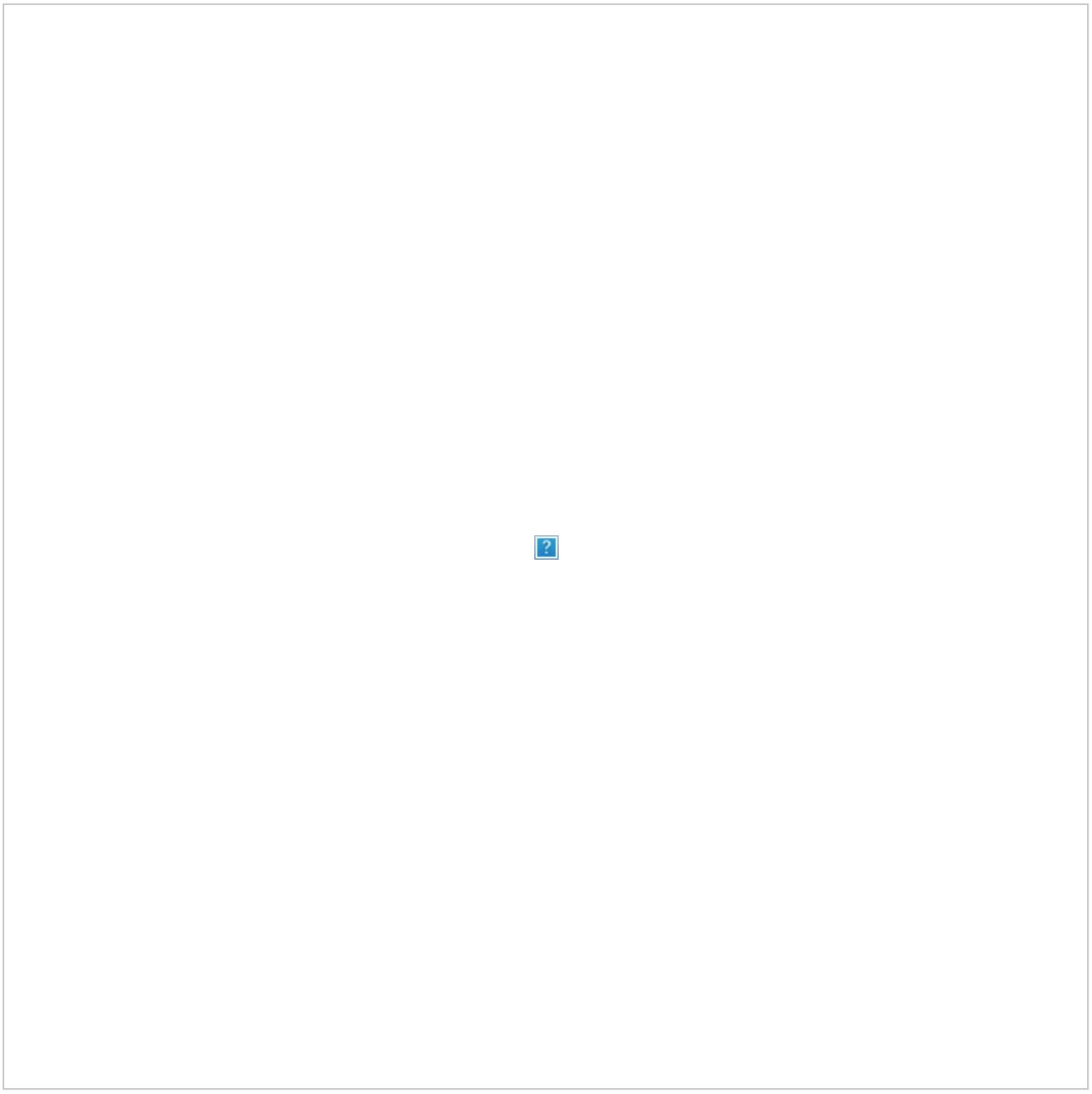
### Preview

| Index | class | text | cleaned_facts | doc_len |
|---|---|---|---|---|
| 1 | u | 이 법원이 적법하게 채택하여 조사한 증거에 의하여 인정되는 아래와 같은 사정 즉, | 법원 채택 조사 증거 인정 아래 사정 | 12 |
| 2 | u | 위와 같은 사정들을 모아보면, | 사정 | 4 |
| 3 | u | 앞서 살핀 증거에 의하여 인정되는 다음과 같은 사실 또는 사정을 종합하면, | 증거 인정 다음 사실 사정 종합 | 11 |
| 4 | u | 피고인과 변호인은, | 피고인 변호인 | 2 |
| 5 | d | (증거기록 제178쪽). | 증거 기록 쪽 | 2 |

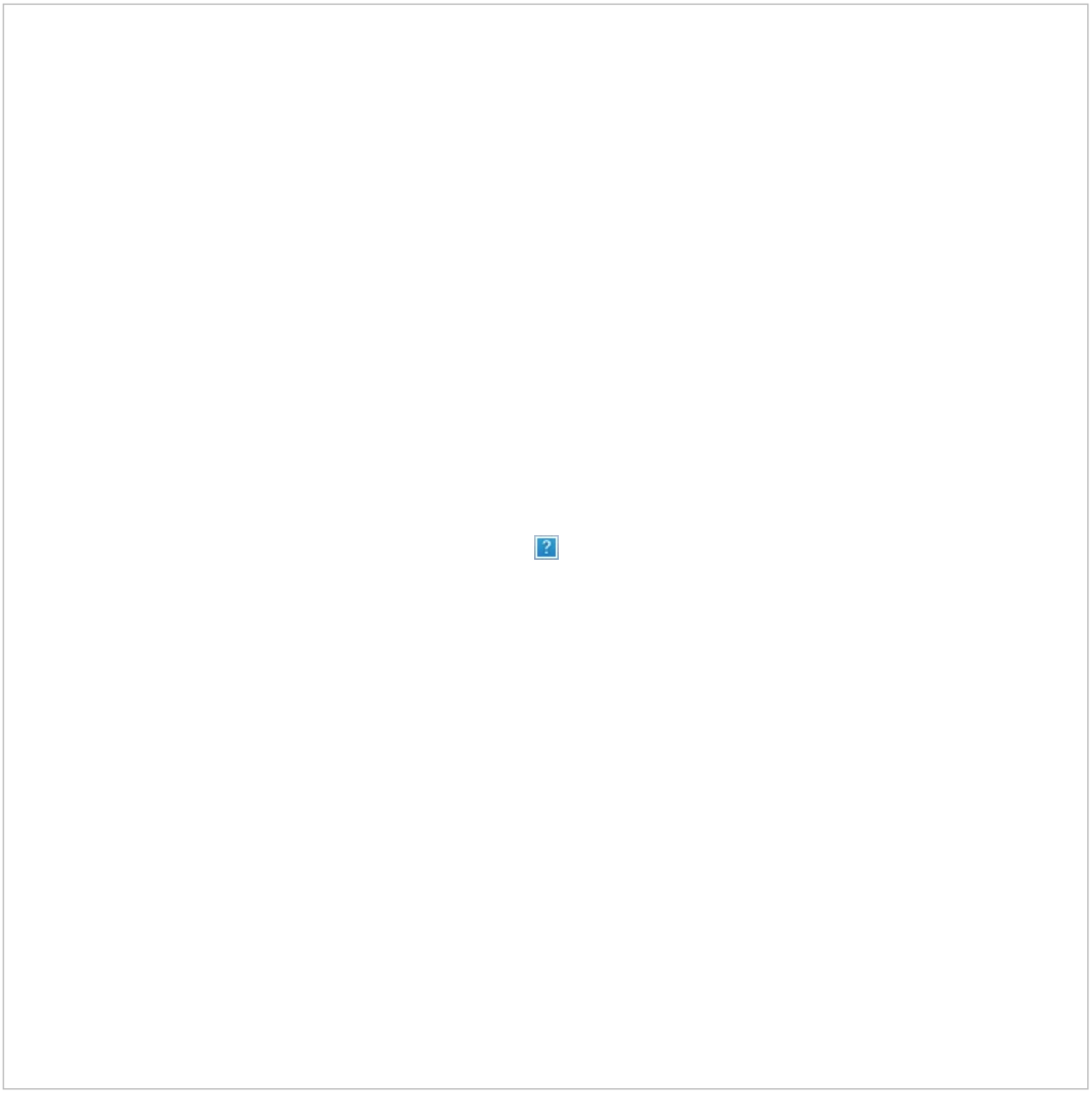
---

## Visualizations

### Class Distribution

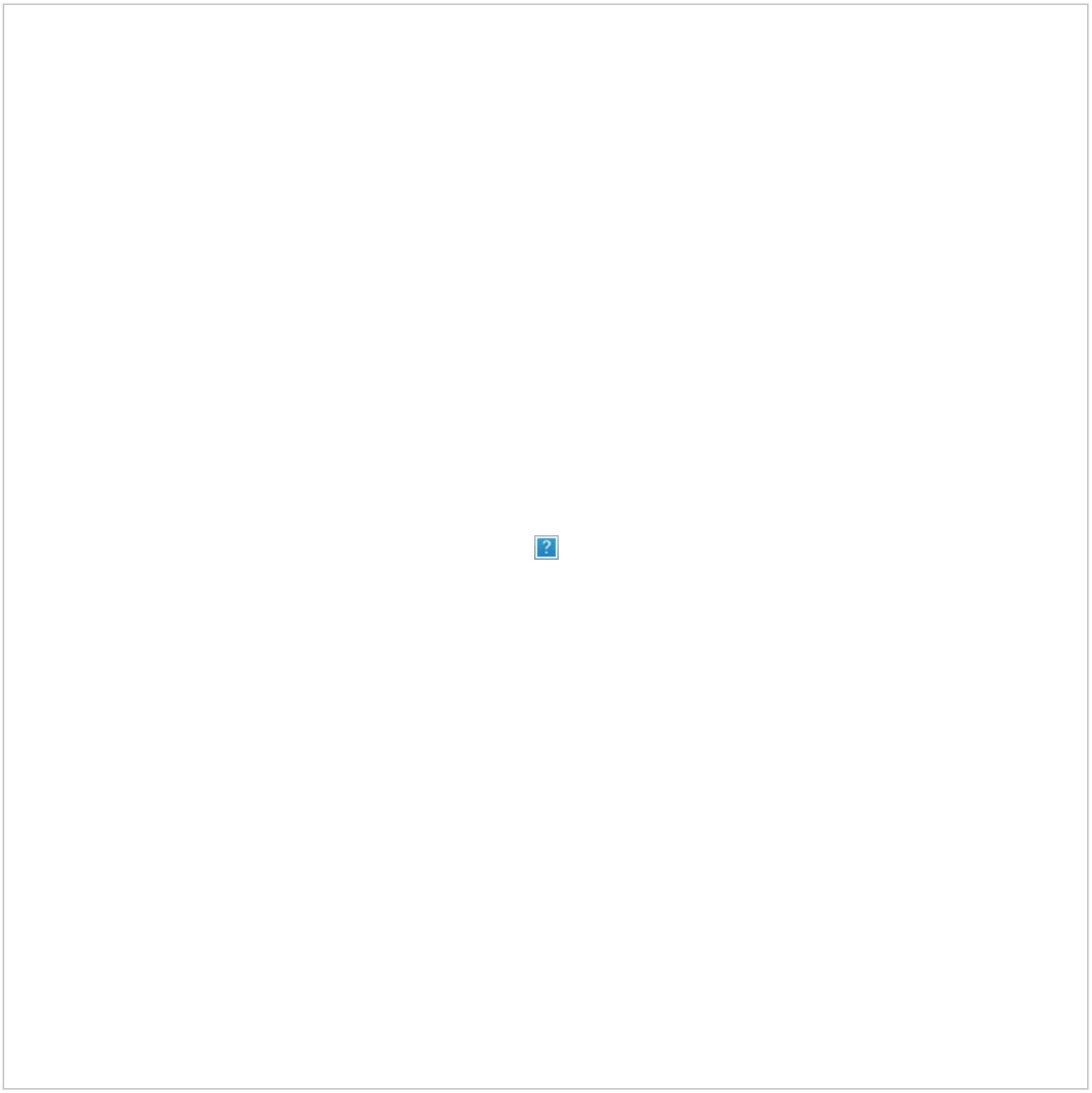


## Document Length Distribution



| Dataset | Longest Sentence | Shortest Sentence | Mean Sentence Length | Sum of Sentences |
|---|---|---|---|---|
| Train | 70 | 1 | 13 | 4078 |
| Validation | 64 | 1 | 14 | 1935 |
| Test | 95 | 1 | 13 | 780 |

### Word Cloud



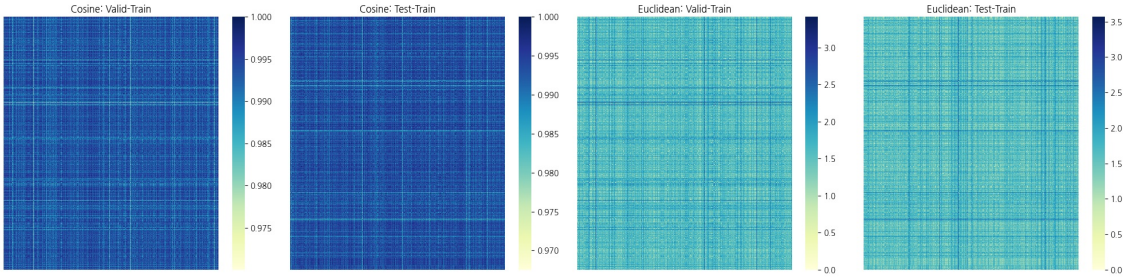
**통계 요약 코멘트:**

- 데이터의 전반적인 특성 요약:
- 본 데이터셋은 세 개의 서로 다른 평가 간소화(Train, Validation, Test)로 구성되어 있습니다.
- Train: 4078개의 문서, 평균 문장 길이 13단어
- Validation: 1935개, 평균 문장 길이 13단어
- Test: 780개, 평균 문장 길이 12단어
- 각 평가 간소화마다 문서 수와 평균 문장 길이가 약간씩 다른 것을 볼 때, 데이터가 다양한 형태로 수집되었음을 알 수 있습니다.
- 키워드와 길이 등으로부터 데이터 성격이나 주제의 변화 가능성을 추론:
- 모든 세 개의 평가 간소화(Train, Validation, Test)는 비슷하게 "피고인", "피해자", "범행"과 같은 공통적인 키워드를 포함하고 있어, 이들이 사건 처리 과정에서 중요한 역할을 할 수 있음을 나타냅니다.
- train, validation, test 데이터셋 간의 차이 분석:
- Train에서 validation으로 넘어가는 동안 문서 수는 11개 증가했지만 평균 문장 길이는 동일합니다. 이는 평가 간소화 프로세스가 효율적으로 작동하고 있음을 시사하며, 텍스트 데이터를 처리하는 데 있어서 일관성이 유지되고 있다는 것을 나타냅니다.
- Test는 다른 두 평가 간소화와 동일한 문서 수와 평균 문장 길이를 가지고 있습니다. 이는 유사한 결과가 도출되었음을 나타내며, 이 세 평가가 실제로 동일한 테스트 조건을 가지고 진행되었을 가능성이 높습니다.

## Data Drift Analysis Results

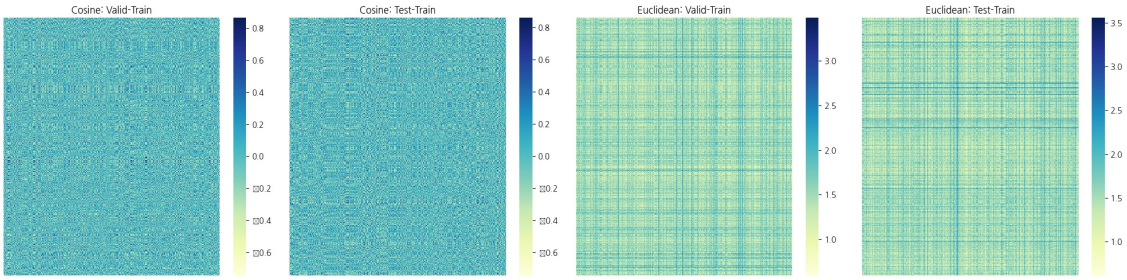
**Train Embeddings:** (4078, 768)
**Valid Embeddings:** (1935, 768)
**Test Embeddings:** (780, 768)

## Embedding Distance (Original Dimension)

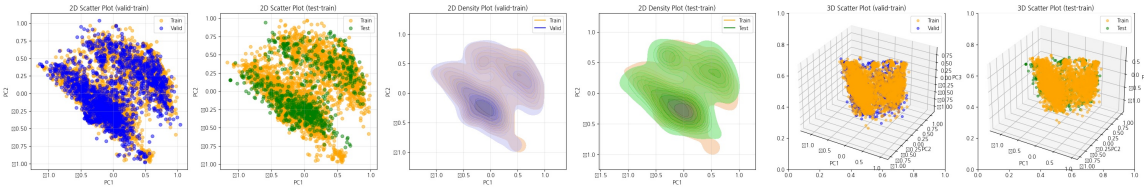



**PCA Reduced Dimension:** 100

## Embedding Distance after PCA




## Embedding Visualization after PCA




## Quantitative Drift Scores

```
- MMD: score = 0.0005, drift = False
- Wasserstein Distance: score = 0.0600, drift = True
- KL Divergence: score = 0.0100, drift = False
- JensenShannon Divergence: score = 0.0600, drift = True
- Energy Distance: score = 0.0000, drift = False
```

## Drift Analysis Summary

데이터 드리프트 분석 보고서

1. MMD (Mean Absolute Deviation)

MMD는 각 데이터 포인트가 평균으로부터 얼마나 떨어져 있는지를 측정하는 지표입니다. 여기서는 수치로 변환된 값이 0.0005로 나타났습니다. 이는 모델의 예측 성능이나 정확도를 나타낼 수 있지만, 구체적인 해석은 데이터셋에 따라 달라질 수 있습니다.

2. Wasserstein Distance

riesien 거리는 각 데이터 포인트 쌍 사이의 거리를 측정하는 지표입니다. 여기서는 0.0600으로 나타났으며, 모델이 시간에 따라 어떻게 변하는지를 나타내는 중요한 지표로 해석될 수 있습니다. 실제로 이 값은 drift가 발생했음을 나타냅니다.

3. KL Divergence (Kullback-Leibler Divergence)

KL divergence는 두 분포 간의 차이를 측정하는 비음수적 함수입니다. 여기서는 0.0100으로 나타났으며, 데이터셋이 시간에 따라 어떻게 변하고 있는지를 평가하는 데 사용할 수 있습니다.

4. Jensen-Shannon Divergence
Jensen-Shannon Divergence은 KL divergence를 2로 나눈 값이며, 양수 및 음수 값을 모두 가질 수 있는 지표입니다. 여기서는 0.0600으로 나타났으며, 데이터셋 간의 차이를 측정하는 데 사용할 수 있습니다.
5. Energy Distance
에너지 거리는 두 데이터 포인트 사이의 거리를 측정하는 지표입니다. 여기서는 0.0000으로 나타났으며, 이는 모델이 동일한 값을 예측하고 있음을 나타냅니다.
drift 발생 판단 근거
- Wasserstein Distance는 0.0600의 값으로 실제로 drift가 발생했음을 확인하였습니다.
drift가 발생할 경우 모델 성능에 미치는 영향
- 데이터셋 간의 차이가 큰 경우, 모델의 예측 성능이 저하될 수 있습니다. 특히, 특정 시간이 지나면서 모델이 시간에 따라 어떻게 변하는지를 정확하게 반영하지 못한다면 예측 정확도가 떨어질 가능성이 높습니다.
전반적 해석 및 요약
데이터셋 간 차이를 종합적으로 해석하기 위해서는 여러 지표들을 함께 고려해야 합니다. MMD와 KL divergence는 데이터의 변동성을 직접적으로 나타내는 반면, stewardship 거리와 에너지 거리는 모델이 시간에 따라 어떻게 변하는지를 평가하는 데 유용합니다. 이 경우, 우선적으로 stewardship 거리가 0.0600인 것을 통해 drift가 발생했음을 확인할 수 있으며, 이는 모델 예측 성능에 영향을 미칠 가능성이 높습니다.
데이터셋 간 차이를 해석하기 위해서는 여러 지표들을 종합적으로 고려하여 데이터의 변동성, 시간에 따른 변화 등을 평가하는 것이 중요합니다. 이를 바탕으로 적절한 모델을 선택하고, 지속적인 모니터링을 통해 모델 성능을 유지하는 것이 필요합니다.