**Assignment 5**

202035349 심승민

**Assignment: Clean It, Fix It, Scale It**

1. Missing Value Handling

- `income[5]` → 평균으로 대체

- `purchases[10]` → 중앙값으로 대체

- 결측 처리 전후 값 출력

2. Outlier Detection & Handling

- IQR 기반으로 `income`, `purchases`의 이상치 탐지

- 이상치 값 및 인덱스 출력

- clip 적용

3. Feature Scaling

|  |  |  |
| --- | --- | --- |
| Feature | Scaling Method |  |
| age | Min-Max Scaling | 값의 범위가 제한적일 때 적합 |
| income | Z-score, Robust | 이상치 민감도 비교 |
| purchases | Log Transformation | 양의 왜곡 데이터에 적합 |
| [age, income, clicks] | Vector Norm | 유사도 기반 모델에 적합 |

4. Visualization

- Histogram (purchases): 로그 변환 전후 비교

- Boxplot (income): Robust scaling 전후 비교

**Assignment: Categorical Encoding**

1. Label Encoding (brands)

- Nike → 0, Adidas → 1, Puma → 2

2. Ordinal Encoding (sizes)

- Small → 1, Medium → 2, Large → 3

3. One-Hot Encoding (colors)

- 6x3 array 생성

4. Final Feature Matrix

- One-hot(colors) + ordinal(sizes) + label(brands) = (6 x 5)

Reflection

- Why is one-hot encoding better for colors than label encoding?

- 명목형 데이터이기 때문에 서열 관계 없음→ one-hot 적합

- Why is ordinal encoding okay for sizes?

- 서열(크기)의 의미가 있음 → ordinal 적합

**Assignment: Feature Selection & Preprocessing**

Part 1. Data Preparation

- 데이터 크기, feature 목록, target 분포 출력

- MinMaxScaler로 정규화

Part 2. Feature Selection

A. Chi-Square

- top 5 features: `worst concave points`, `mean concave points`, ...

B. Lasso

- Non-zero coef: `worst concave points`, `mean concavity`, ...

C. Tree-Based

- 중요 변수 top 5 시각화: `worst concave points`, `worst radius`, ...

Part 3. Comparison & Reflection

- Which features were selected consistently across methods?

- 모든 방법에서: `worst concave points`

- 두 방법에서: ‘worst radius’, ‘mean concavity’, ‘worst perimeter’, ‘mean concave points

- Did any method eliminate features that another considered important?

-Chi-Square나 Tree-based에서 중요하다고 판단된 feature(worst perimeter, mean concavity 등)제거

- Which method do you think is most trustworthy for this task, and why?

- Tree-based

- 비선형 반영 가능

- 직관적인 중요도 해석