Clustering project

어종 식별 및 분류 군집 모델

pronounce12@gmail.com
조현상
github notion

01 어종 식별 및 분류
소개 및 분석 목적

1. 목적 :
'fish.csv' 데이터셋을 사용하여 다양한 어종을 식별하고 분류하는 시스템을 개발합니다. 이 시스템은 어류 판매업체 또는 수산시장에서 어종을 자동으로 분류하는 데 사용될 수 있습니다.

2. 프로젝트 전제 조건:
- 데이터셋은 **1159마리**의 물고기 데이터로 구성됩니다.
- 데이터는 물고기의 길이, 크기, 무게등에 관한 정보를 각각의 데이터셋에 포함하고 있습니다.
- 프로젝트의 주요 관심사 및 중점 분석 대상은 "**Weight**" 컬럼으로, 이 컬럼은 물고기의 크기를 나타내는 중요한 변수입니다. 이를 통해 어종 식별 및 분류 모델을 개발하고자 합니다.

3. 컬럼 소개

Fish.csv

- **Species (어종)**: 물고기의 종을 나타냅니다. 이는 구분 가능한 개별 어종의 이름을 나타내며, 생물학적 분류에 따른 명칭입니다.
- **Weight (무게)**: 물고기의 체중을 나타냅니다. 일반적으로 그램(g)이나 킬로그램(kg) 단위로 측정되며, 어류의 크기와 건강 상태를 나타내는 지표입니다.
- **Length1 (표준 길이)**: 물고기의 머리부터 꼬리 지느러미 시작 부분까지의 길이를 의미합니다. 표준 길이(Standard Length)라고도 불리며, 어류의 성장률을 평가하는 데 사용됩니다.
- **Length2 (전장)**: 물고기의 머리부터 꼬리 지느러미 끝까지의 길이를 의미합니다. 전장(Total Length)이라고도 하며, 어류의 전체적인 크기를 측정하는 데 쓰입니다.
- **Length3 (체장)**: 물고기의 몸통 부분, 즉 머리 끝부터 꼬리 지느러미 시작 부분까지의 길이를 말합니다. 체장(Fork Length)이라고도 불리며, 어류의 체형을 나타내는 데 유용합니다.
- **Height (높이)**: 물고기의 몸통의 높이를 나타냅니다. 보통 지느러미를 포함하지 않고 몸통만을 기준으로 측정하며, 어류의 체형과 관련된 특성입니다.
- **Width (폭)**: 물고기의 가장 넓은 부분의 폭을 나타냅니다. 이는 보통 물고기의 몸통 가로 길이를 의미하며, 어류의 체형을 파악 할 수 있는 요소입니다.

02 Data 전처리

결측치 및 중복행 확인

isna().sum()

```
Out[41]: Species    0
         Weight     0
         Length1    0
         Length2    0
         Length3    0
         Height     0
         Width      0
         dtype: int64
```

duplicated().sum()

```
Out[42]: 0
```

결측치나 중복행은 데이터의 질을 떨어뜨릴 수 있습니다.
다행히도, 'fish.csv' 데이터셋을 검토한 결과, 결측치나 중복행이 존재하지 않다는 것을 확인할 수 있었습니다. 즉, 데이터셋에 포함된 모든 행은 완전하고 분석의 정확성을 높여줍니다.

Label 인코딩

sklearn의 LabelEncoder를 사용하여 scaled_fish_df 데이터프레임의 'Species' 열에 있는 범주형 데이터를 숫자로 변환

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
scaled_fish_df['Species'] = encoder.fit_transform(scaled_fish_df['Species'])
print(encoder.classes_)
```

```
[[-0.65394306 -0.35089978 -0.27568191 -0.11674094  0.60107609 -0.23493547]
 [-0.48982158 -0.26799285 -0.19015795 -0.01231045  0.83273284 -0.06040232]
 [-0.3188617  -0.27835621 -0.17115262 -0.02101299  0.80807105  0.17823631]
 ...
 [ 0.40528707  1.06730943 -2.08818615  0.58363546 -0.42170268  0.60481611]
 [ 1.16180698  1.47564744  1.10711406  0.9483317  -0.16748728 -0.60111857]
 [-0.12350935  1.69359805  0.15281522  1.0346741   0.3647323  -1.25527698]]
```

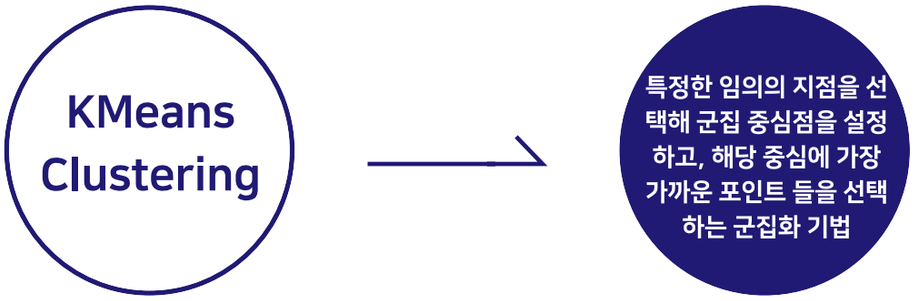
정규화 및 PCA

	pca1	pca2
0	-0.452688	-0.564158
1	-0.108067	-0.658039
2	0.057986	-0.708710
3	0.341283	-0.483139
4	0.616239	-0.589666
5	0.766291	-0.616550
6	0.978295	-0.737514
7	0.580465	-0.442993
8	0.839971	-0.598185
9	1.069580	-0.570499

PCA

PCA는 **차원 축소 기법**으로, 많은 특성을 가진 데이터셋에서 주요 정보를 유지하면서 데이터셋의 차원을 줄이는 방법입니다. PCA는 고차원 데이터에서 중요한 구조를 발견하거나, 노이즈를 제거하고, 효율적인 시각화를 가능하게 합니다.

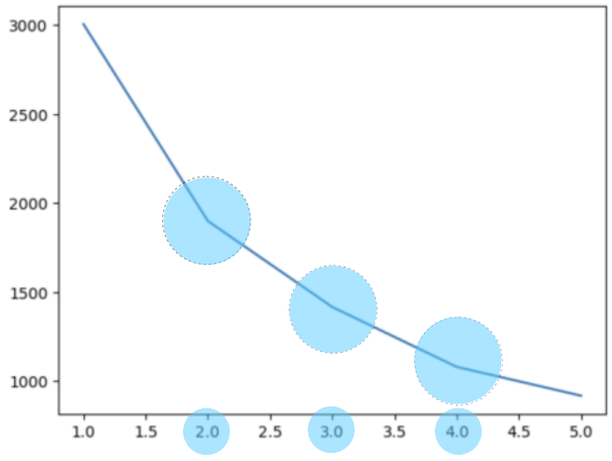
04 사용한 Clustering 기법



KMeans Clustering

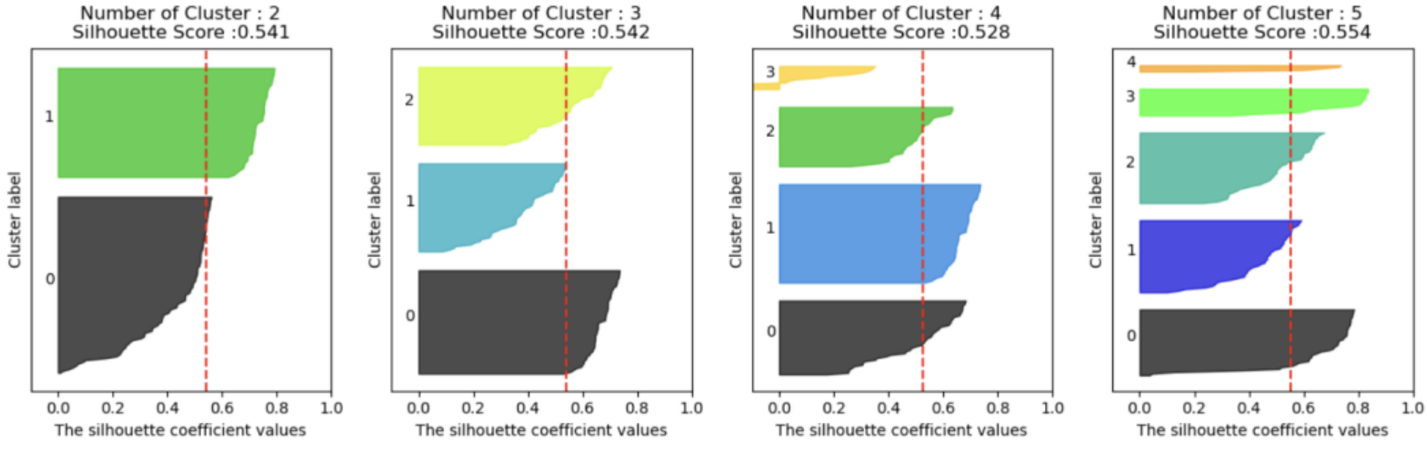
KMeans Clustering은 데이터에 대한 **사전 정보 없이도 신속하게 대용량 데이터**에 적용할 수 있는 직관적이고 범용적인 중심점 기반의 분할적 군집화 방법으로, 각 데이터가 다른 군집의 중심점보다 자신이 속한 군집의 중심점에 가깝도록 **원형**에 가까운 군집을 형성하며 이는 데이터가 실제로 원형일 때 가장 효과적입니다.

05 KElbow, Silhouette Coefficient



KElbow graph

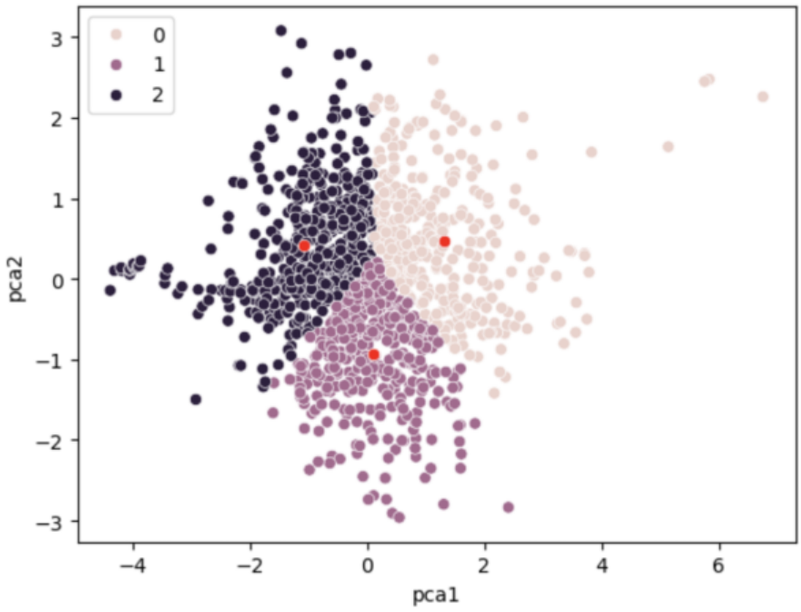
엘보우 방법은 적절한 클러스터 수 K를 결정하기 위해 사용되며, 클러스터 수에 따른 비용 함수의 값을 플로팅하여 그래프를 그립니다.
여기서 비용 함수는 일반적으로 각 클러스터 내 데이터 포인트와 클러스터 중심점 사이의 거리의 제곱합(Sum of Squared Distances)을 의미합니다.
어종 식별 및 분류 모델에서는 **클러스터 2, 3, 4가 가장 적절**한 클러스터 수로 확인되었습니다.



Silhouette Coefficient

실루엣 계수(Silhouette Coefficient)는 클러스터링의 품질을 측정하는 지표 중 하나로, 각 데이터 포인트가 얼마나 잘 클러스터링되었는지를 수치적으로 나타냅니다.
실루엣 계수는 -1에서 1 사이의 값을 가지며, 높은 값일수록 클러스터링 결과가 좋다고 판단할 수 있습니다.
어종 식별 및 분류 모델에서는 **클러스터 3이 가장 적절**한 클러스터 수로 확인되었습니다.

05 PCA 산점도



PCA Scatter Plot

PCA Scatter Plot의 각 점은 개별 데이터 포인트를 나타내며, 색상은 각 포인트가 속한 클러스터를 나타냅니다.
어종 식별 및 분류 모델에서는 세 개의 클러스터(0, 1, 2)가 있으며, 각 클러스터의 중심점은 빨간색 점으로 표시되어 있습니다.
클러스터의 중심점은 해당 클러스터에 속한 모든 데이터 포인트들의 평균 위치를 나타냅니다.

12 느낀점

선형회귀와 OLS

차원 축소 기법인 PCA를 적용하는 과정에서 몇 가지 어려움에 직면했습니다. 원본 데이터의 해석성을 유지하면서도 중요한 정보를 보존하는 것은 쉽지 않았지만, 이러한 도전을 통해 데이터의 본질적인 구조에 대한 깊은 이해를 얻을 수 있었습니다. 또한, PCA 산점도를 통해 군집화된 데이터를 시각화함으로써, 데이터 내 패턴을 더 명확히 볼 수 있었습니다.

감사합니다