

비지도 학습 모형

1. 주성분 분석과 차원 축소

1. 주성분 분석과 차원 축소

차원의 저주(Curse of Dimensionality)는 고차원 데이터 공간에서 발생하는 여러 문제를 말한다. 이는 주로 데이터 분석, 기계 학습, 최적화 등에서 나타난다. 주요 문제는 다음과 같다:

1. 데이터 희소성: 차원이 증가할수록 데이터 포인트들이 공간에서 희소해진다. 즉, 동일한 밀도의 데이터를 유지하려면 엄청나게 많은 데이터 포인트가 필요하다.

2. 거리 측정의 문제: 고차원 공간에서는 모든 점들이 서로 멀리 떨어져 있어, 거리 기반 알고리즘(예: k-최근접 이웃 알고리즘)의 성능이 저하된다. 두 점 사이의 거리가 비슷해져서 차별화가 어려워진다.

1. 주성분 분석과 차원 축소

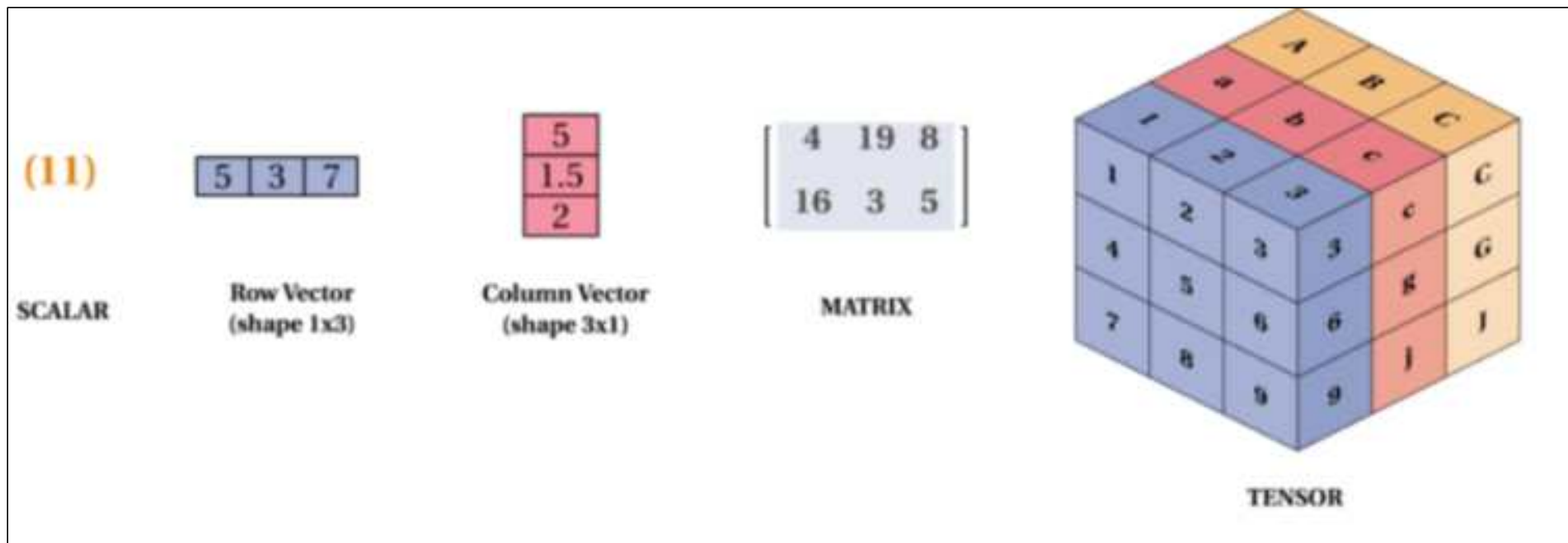
3. 계산 비용 증가: 차원이 늘어나면 계산 복잡도가 기하급수적으로 증가한다. 이는 메모리와 계산 시간의 큰 증가를 초래한다.

4. 과적합: 고차원 공간에서는 모델이 데이터를 너무 잘 맞춰서 훈련 데이터에는 잘 맞지만 새로운 데이터에는 잘 맞지 않는 과적합 (overfitting) 문제가 발생할 가능성이 높다.

따라서, 고차원 데이터를 다룰 때는 차원 축소 기법(예: 주성분 분석 (PCA), t-SNE)을 사용해 차원을 줄이거나, 차원 수에 맞는 충분한 데이터를 확보하는 것이 중요하다.

1. 주성분 분석과차원 축소

- 1차원에서는 데이터 포인트들이 선형적으로 배치되어 밀집되어 보인다.
- 2차원에서는 데이터 포인트들이 평면 상에 퍼지게 되어 밀도가 낮아진다.
- 3차원에서는 데이터 포인트들이 공간에 널리 퍼지게 되어 더욱 희소해진다.



차원의 저주 해결 방법

- 1. 차원 축소:** 주성분 분석(PCA), t-SNE, UMAP 등의 차원 축소 기법을 사용하여 고차원 데이터를 저차원으로 변환한다.
- 2. 피처 선택:** 데이터의 중요한 피처만 선택하여 차원을 줄이는 방법이다. 이는 모델의 성능을 유지하면서 불필요한 피처를 제거할 수 있다.
- 3. 정규화 기법:** L1, L2 정규화 등을 사용하여 모델의 복잡성을 줄이고 과적합을 방지한다.
- 4. 더 많은 데이터 수집:** 차원이 증가할수록 더 많은 데이터가 필요하므로, 가능한 한 많은 데이터를 수집하여 희소성을 줄인다.

1. 주성분 분석과 차원 축소

주성분 분석과 차원 축소 소개

주성분 분석(PCA, Principal Component Analysis)은 고차원의 데이터를 저차원으로 변환하여 데이터를 단순화하는 차원 축소 기법이다. 이는 데이터의 중요한 구조를 유지하면서 차원을 줄여서 시각화나 추가 분석을 쉽게 할 수 있게 해준다. 주성분 분석은 다음과 같은 단계로 이루어져 있다:

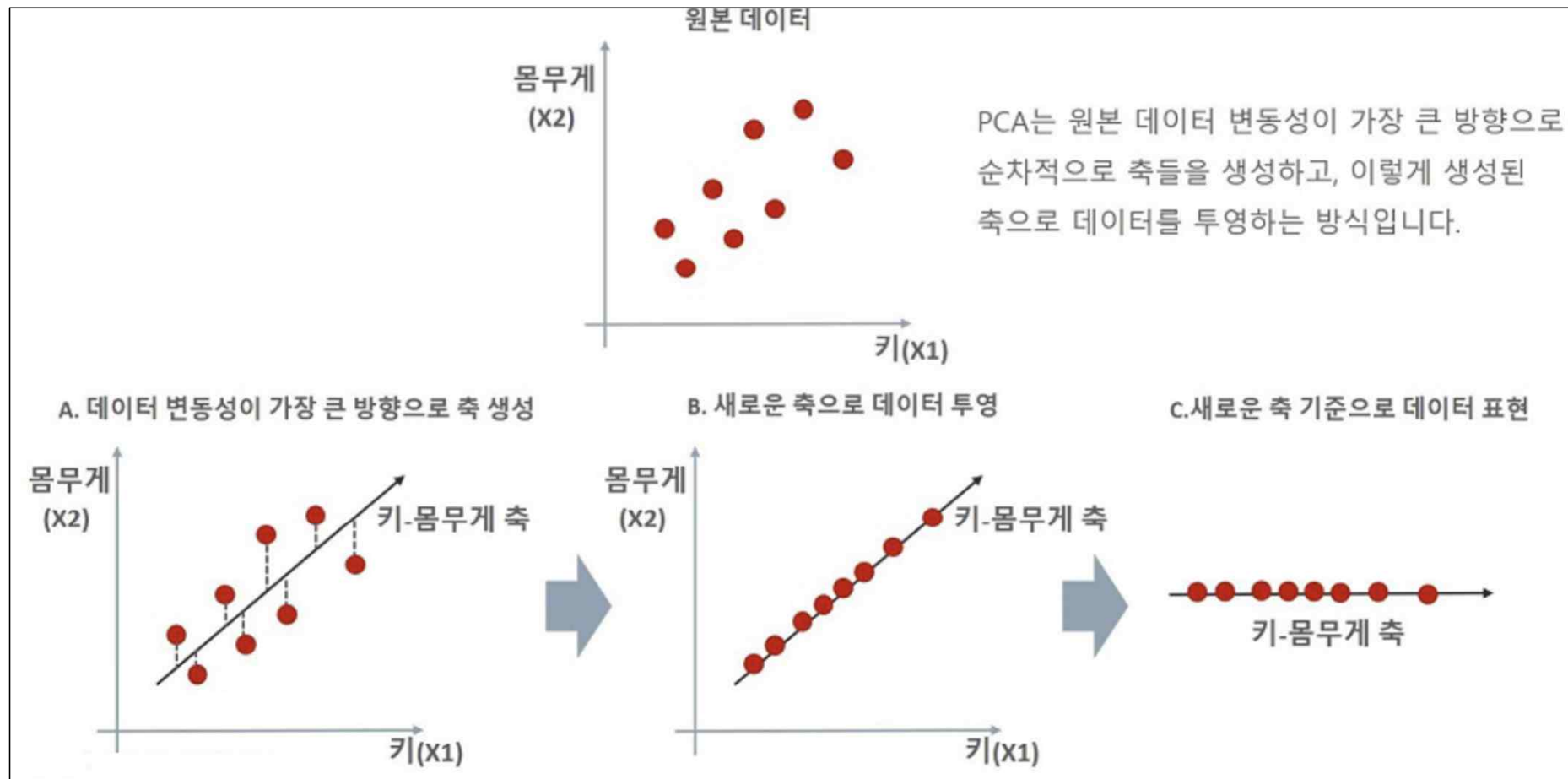
- 1. 데이터 표준화:** 주성분 분석을 수행하기 전에 데이터의 평균을 0, 분산을 1로 맞추는 표준화를 진행해야 한다.
- 2. 공분산 행렬 계산:** 표준화된 데이터의 공분산 행렬을 계산하여 변수들 간의 상관 관계를 파악한다.

1. 주성분 분석과 차원 축소

3. **고유 벡터와 고유 값 계산:** 공분산 행렬의 고유 벡터와 고유 값을 계산하여 데이터의 분산을 설명하는 주요 축을 찾는다.
4. **주성분 선택:** 고유 값을 기준으로 가장 큰 고유 값을 가지는 고유 벡터 (주성분)를 선택한다.
5. **데이터 변환:** 선택된 주성분으로 데이터를 변환하여 차원을 축소한다. 차원 축소는 데이터 시각화, 노이즈 제거, 계산 효율성 향상 등의 이점을 제공하며, 특히 고차원의 데이터를 다룰 때 유용하다.

1. 주성분 분석과 차원 축소

PCA(Principal Component Analysis) 동작 원리



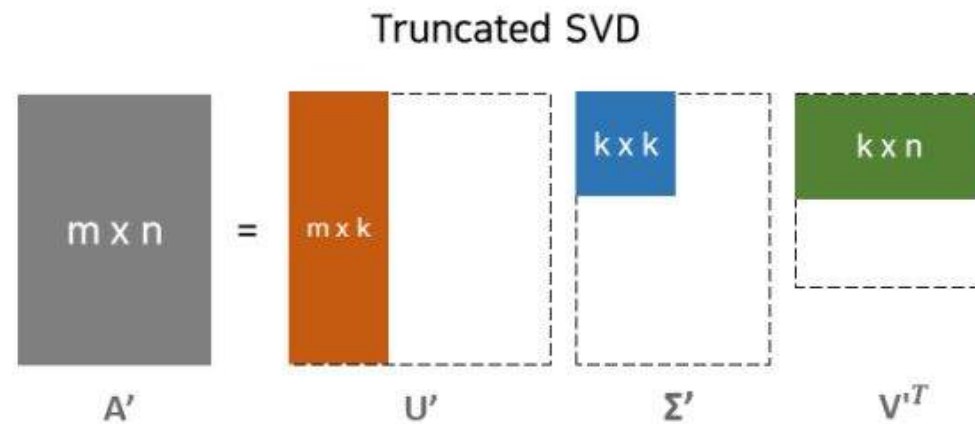
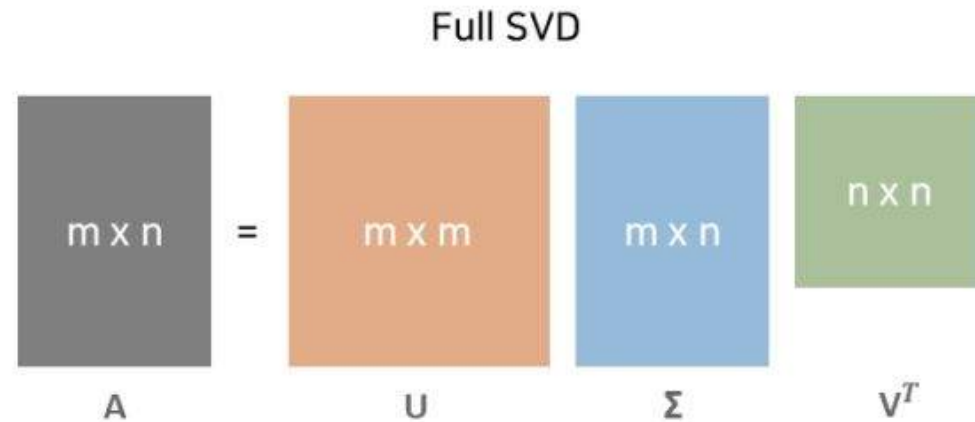
1. 주성분 분석과 차원 축소

행렬의 고유값 분해

SVD(특이값 분해

Singular Value

Decomposition),



2. 주성분 분석(PCA) / 차원축소 구현 실습

2. 주성분 분석(PCA)/차원축소 구현 실습

iris.csv 엑셀로 열어 확인하기

	A	B	C	D	E
1	sepal leng	sepal widi	petal leng	petal widt	label
2	5.1	3.5	1.4	0.2	0
3	4.9	3	1.4	0.2	0
4	4.7	3.2	1.3	0.2	0
5	4.6	3.1	1.5	0.2	0
6	5	3.6	1.4	0.2	0
7	5.4	3.9	1.7	0.4	0
8	4.6	3.4	1.4	0.3	0
9	5	3.4	1.5	0.2	0
10	4.4	2.9	1.4	0.2	0
11	4.9	3.1	1.5	0.1	0
12	5.4	3.7	1.5	0.2	0
13	4.8	3.4	1.6	0.2	0
14	4.8	3	1.4	0.1	0
15	4.3	3	1.1	0.1	0
16	5.8	4	1.2	0.2	0
17	5.7	4.4	1.5	0.4	0
18	5.4	3.9	1.3	0.4	0
19	5.1	3.5	1.4	0.3	0
20	5.7	3.8	1.7	0.3	0

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2
150 rows x 5 columns					

2. 주성분 분석(PCA)/차원축소 구현 실습



데이터셋 설명해줘(한글로)



이 데이터셋은 유명한 "아이리스(Iris)" 데이터셋입니다. 각 행은 하나의 꽃 샘플을 나타내며, 각 열은 해당 샘플의 특성을 나타냅니다. 데이터셋의 열은 다음과 같습니다:

1. ``sepal length (cm)`` - 꽃받침의 길이 (센티미터)
2. ``sepal width (cm)`` - 꽃받침의 너비 (센티미터)
3. ``petal length (cm)`` - 꽃잎의 길이 (센티미터)
4. ``petal width (cm)`` - 꽃잎의 너비 (센티미터)
5. ``label`` - 꽃의 종류를 나타내는 라벨 (0, 1, 2로 구분됨)

각 라벨은 다음과 같은 꽃 종류를 의미합니다:

- 0: Iris-setosa
- 1: Iris-versicolor
- 2: Iris-virginica

이 데이터셋은 머신러닝 및 데이터 분석 실습에 자주 사용되며, 꽃의 특성들을 바탕으로 꽃의 종류를 예측하는 모델을 만드는 데 유용합니다. [3-]

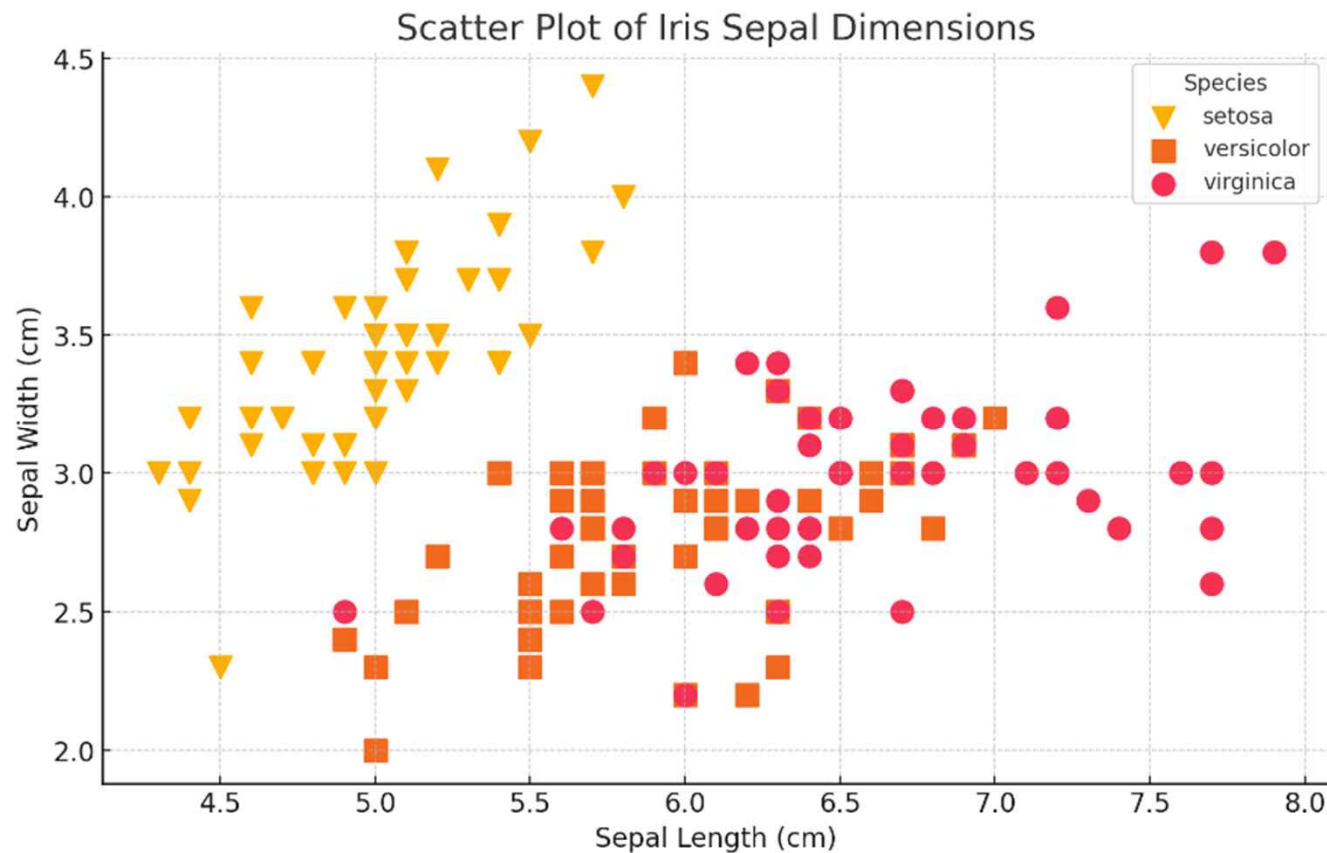
꽃잎

꽃받침



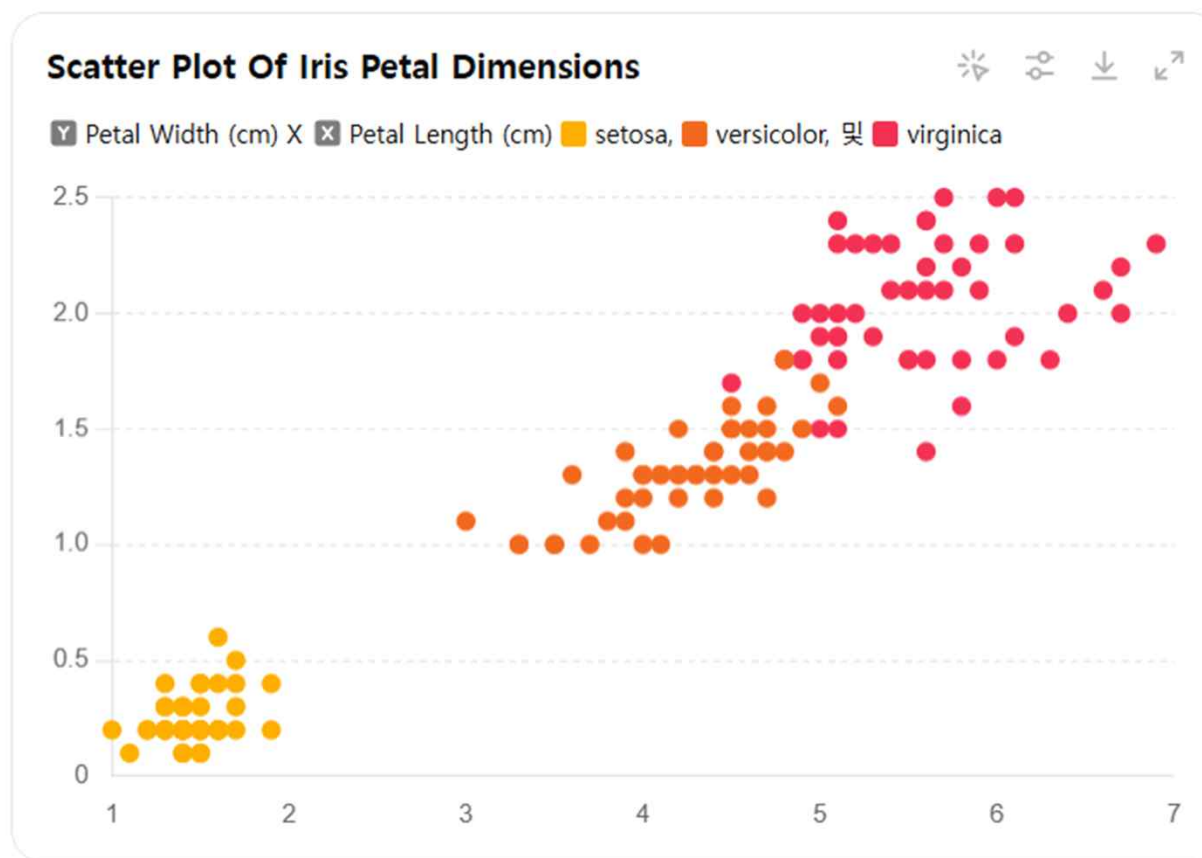
2. 주성분 분석(PCA)/차원축소 구현 실습

앞에 두 컬럼을 x,y축으로 해서 레이블 'setosa' : 세모, 'versicolor' : 네모, 'virginica': 동그라미로 산점도를 그려줘(레전드도 표시해줘)



2. 주성분 분석(PCA)/차원축소 구현 실습

세번째와 네번째 두 컬럼을 x,y축으로 해서 레이블 'setosa' : 세모, 'versicolor' : 네모, 'virginica': 동그라미로 산점도를 그려줘(레전드도 표시해줘)



2. 주성분 분석(PCA)/차원축소 구현 실습

앞에서부터 네 개의 피처를 PCA를 사용하여 축소하여 테이블로 만들어줘.

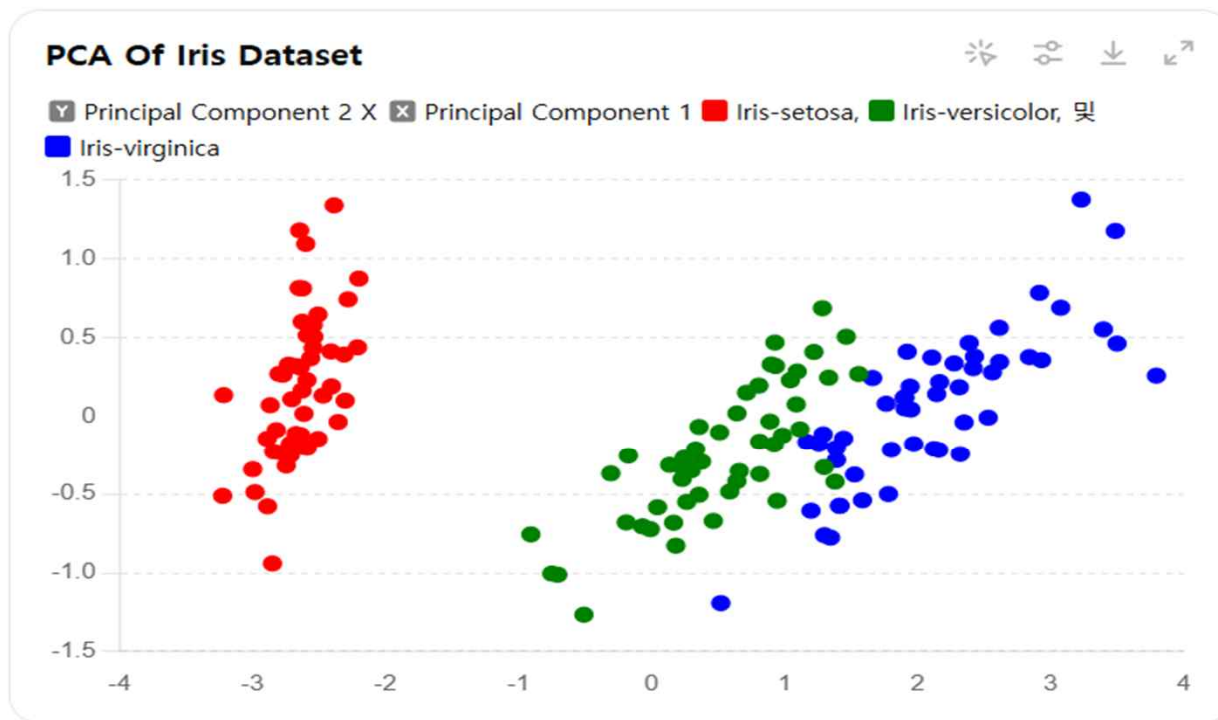


PCA Data				
	Principal Component 1	Principal Component 2	label	
1	-2.684125625969541	0.319397246585103	0	
2	-2.7141416872943274	-0.1770012250647791	0	
3	-2.888990569059296	-0.14494942608555708	0	
4	-2.7453428556414092	-0.31829897925191564	0	
-	-2.72871653655452	0.32675451293492	-	

PCA를 사용하여 네 개의 피처를 두 개의 주성분으로 축소한 결과를 보여주는 테이블을 생성했습니다. 새로운 테이블에는 "Principal Component 1", "Principal Component 2" 두 개의 주성분과 원래의 레이블("label")이 포함되어 있습니다. 이 테이블을 통해 데이터의 차원을 축소한 후의 값을 확인할 수 있습니다. [↔]

2. 주성분 분석(PCA)/차원축소 구현 실습

앞 두개의 피처를 각각 x,y축으로 하고 레이블을 각각 다른 색으로 구별하여 산점도를 그려줘(그래프의 legend에 레이블의 종류를 표시해줘).



위의 산점도는 PCA로 축소된 두 개의 주성분을 각각 x축과 y축으로 하여 그린 것입니다. 각 점은 아이리스 데이터의 개별 샘플을 나타내며, 색상은 각기 다른 품종을 구별합니다. 레이블은 그래프의 범례에 표시되어 있어, 각 색상이 어떤 품종을 나타내는지 쉽게 확인할 수 있습니다. [↗]

3. 군집화(Clustering)

3. 군집화(Clustering)

군집화(Clustering)

데이터를 비슷한 특성을 가진 그룹으로 묶는 비지도 학습 기법을 말한다. 군집화는 데이터 내의 자연스러운 그룹이나 패턴을 식별하는데 사용되며, 이는 여러 분야에서 다양한 목적으로 활용될 수 있다.

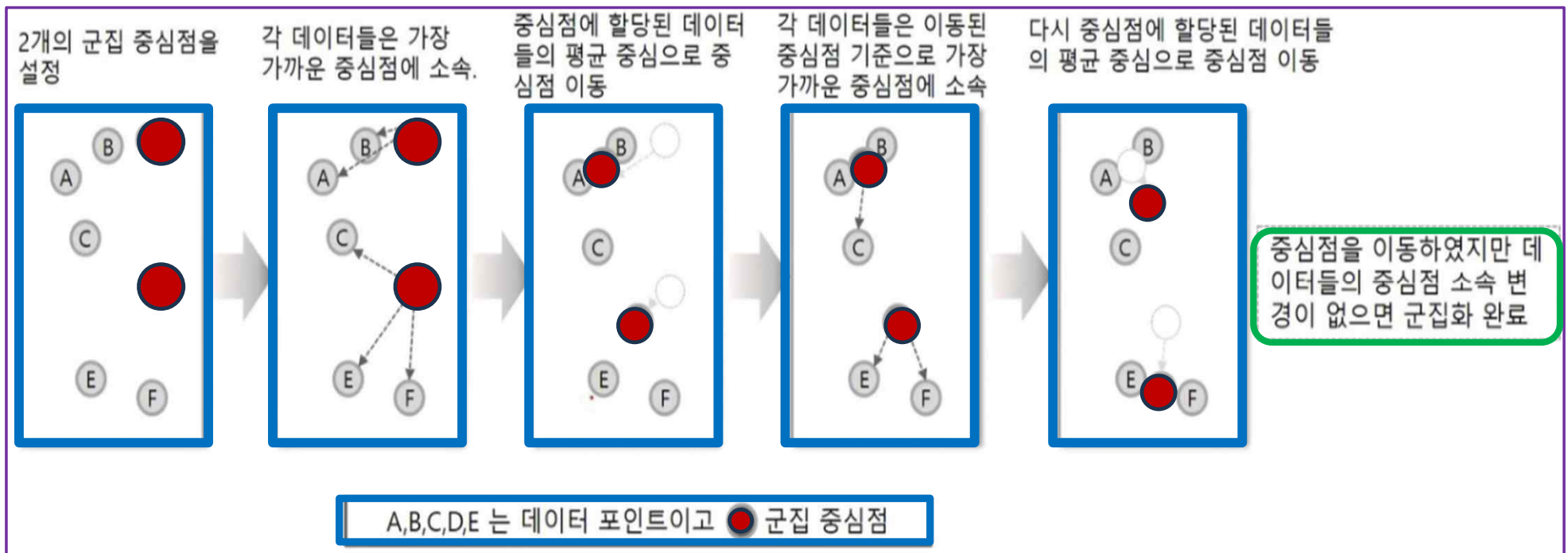
군집화 알고리즘 종류

군집화 알고리즘에는 여러 가지가 있지만, 대표적인 알고리즘으로는 **K-means, 계층적 군집화(Hierarchical Clustering), DBSCAN** 등이 있다.

1. K-means 군집화

- **초기화**: 데이터 포인트 중에서 k 개의 초기 중심을 무작위로 선택한다.
- **할당**: 각 데이터 포인트를 가장 가까운 중심에 할당한다.
- **중심 업데이트**: 각 군집의 중심을 해당 군집에 속한 데이터 포인트의 평균으로 업데이트한다.
- **반복**: 중심이 더 이상 변하지 않을 때까지 할당과 중심 업데이트를 반복한다.
- 이 알고리즘은 단순하고 빠르지만, 초기 중심의 선택에 따라 결과가 달라질 수 있다.

K-means 군집화 동작 과정



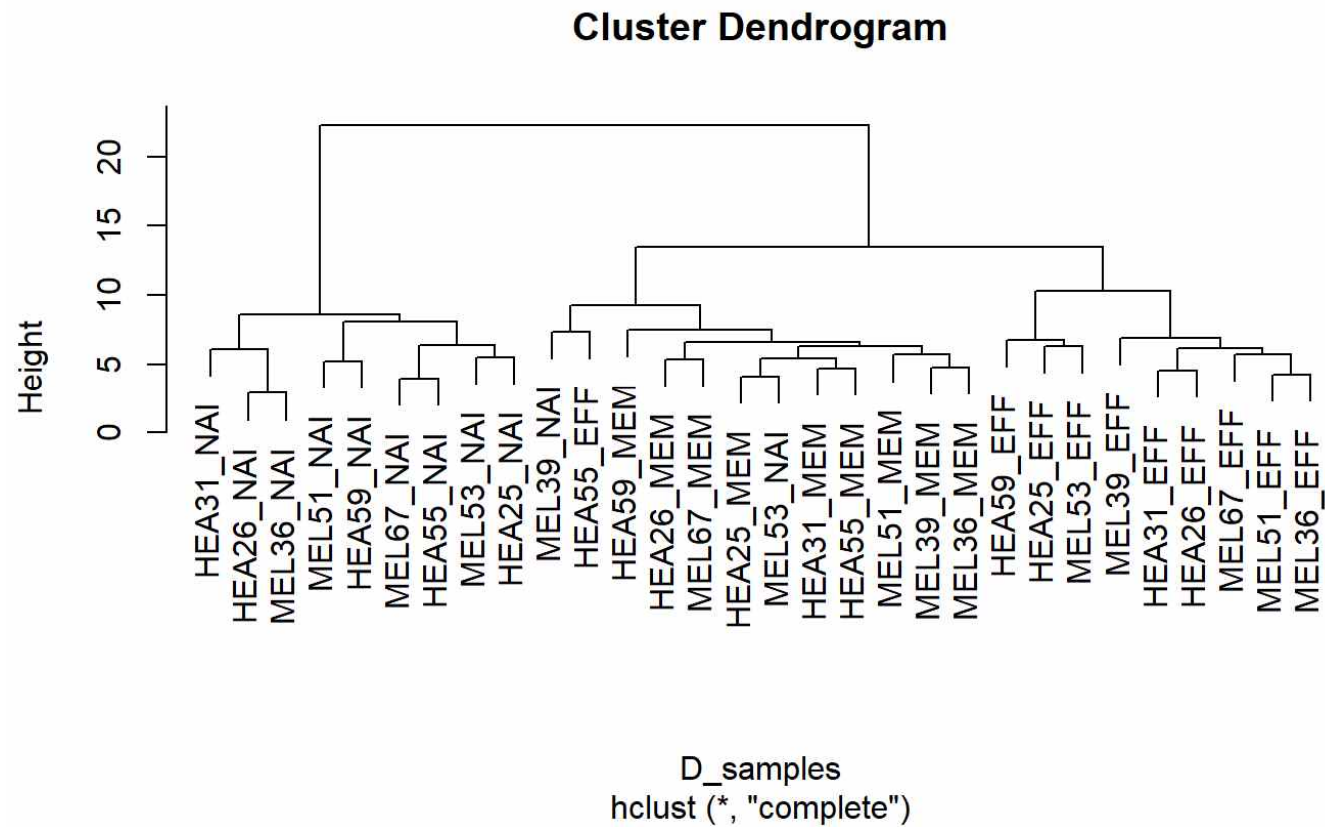
K-means 군집화의 특징

- N개의 표본을 k개의 군집으로 나눈다
- 군집화 중 가장 많이 사용, 쉽고 간결하다, 대용량 데이터 처리 가능
- 속성(변수)의 개수가 너무 많으면 정확도가 떨어져서 PCA로 차원 감소가 필요
- 몇 개의 군집을 선택해야 할지 어렵다
- 이상치 데이터에 취약: 이상치로 인해 중심점이 잘못 설정 될 수 있다

2. 계층적 군집화(Heirarchical clustering)

- **초기화**: 각 데이터 포인트를 하나의 군집으로 간주한다.
- **병합**: 가장 가까운 두 군집을 병합하여 새로운 군집을 형성한다.
- **반복**: 모든 데이터가 하나의 군집이 될 때까지 병합 과정을 반복한다.
- **덴드로그램**: 군집화 과정을 덴드로그램이라는 트리 구조로 시각화하여 적절한 군집 수를 선택할 수 있다.
- 이 방법은 군집의 수를 사전에 지정할 필요가 없지만, 데이터가 많을 경우 계산 비용이 많이 들 수 있다.

계층적 군집화



3. DBSCAN

- **밀도 기반 군집화**: 데이터 포인트의 밀도를 기반으로 군집을 형성한다.
- **핵심 포인트**: 일정 반경 내에 최소 데이터 포인트 수를 만족하는 포인트를 핵심 포인트로 정의한다.
- **군집 확장**: 핵심 포인트를 중심으로 이웃 포인트를 포함하여 군집을 확장한다.
- **노이즈 처리**: 어느 군집에도 속하지 않는 포인트를 노이즈로 처리한다.
- 이 알고리즘은 불규칙한 모양의 군집을 잘 찾아내며, 군집의 수를 미리 지정할 필요가 없다.

군집화의 활용 분야

- 1. 마케팅:** 고객 세분화를 통해 맞춤형 마케팅 전략을 수립할 수 있다. 비슷한 구매 패턴을 가진 고객들을 군집화하여 특정 그룹에 맞는 마케팅 캠페인을 진행할 수 있다.
- 2. 의료:** 환자의 증상 데이터를 군집화하여 비슷한 질병 패턴을 가진 환자 그룹을 식별할 수 있다. 이를 통해 맞춤형 치료법을 제안할 수 있다.
- 3. 이미지 처리:** 비슷한 특성을 가진 이미지나 패턴을 군집화하여 이미지 검색이나 분류 작업을 효율적으로 수행할 수 있다.
- 4. 문서 분류:** 텍스트 데이터를 군집화하여 주제별로 문서를 분류하거나 검색 시스템을 개선할 수 있다.

4. 군집화(Clustering) 구현 실습

4. 군집화(Clustering) 구현 실습

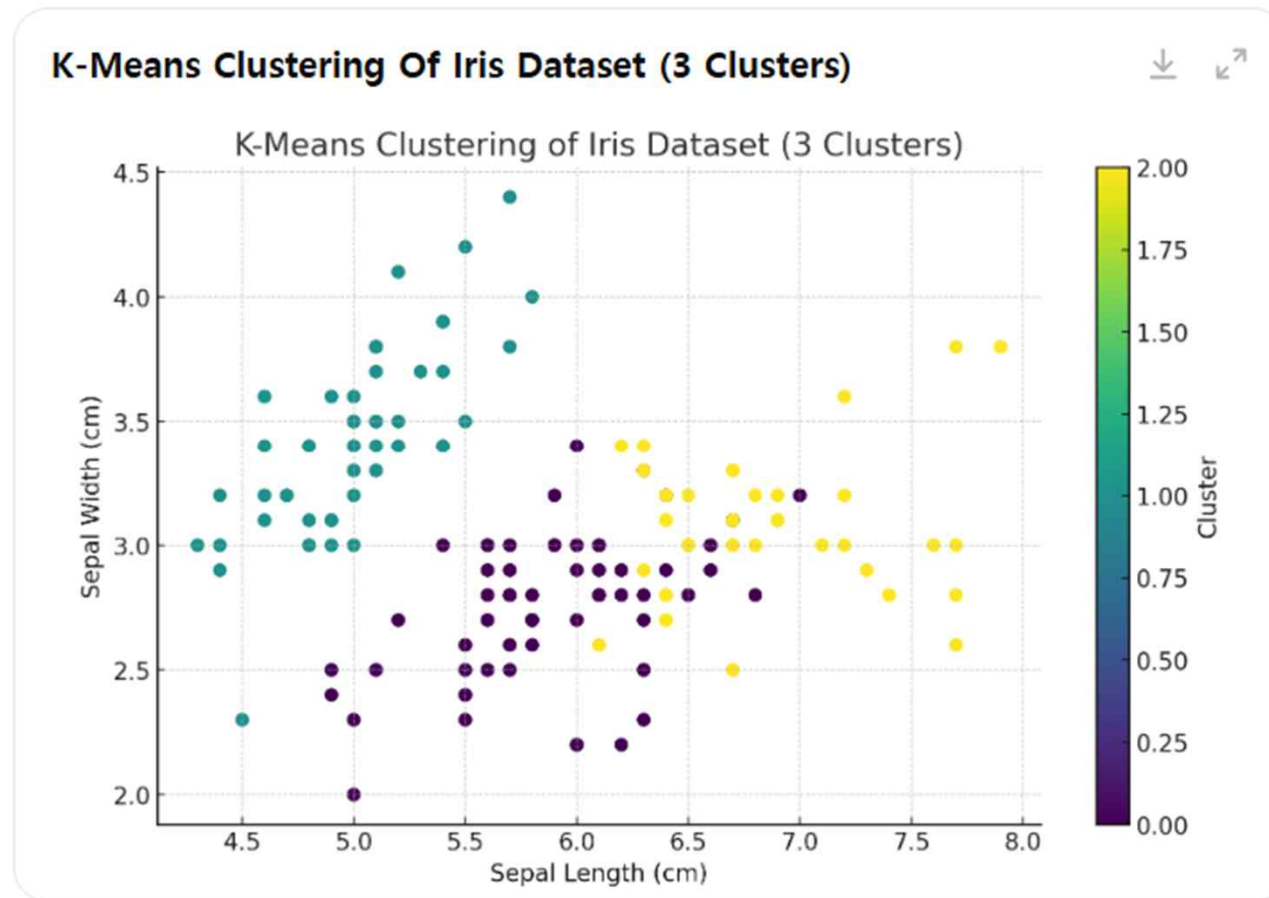
iris.csv 엑셀로 열어 확인하기

	A	B	C	D	E
1	sepal leng	sepal widi	petal leng	petal widt	label
2	5.1	3.5	1.4	0.2	0
3	4.9	3	1.4	0.2	0
4	4.7	3.2	1.3	0.2	0
5	4.6	3.1	1.5	0.2	0
6	5	3.6	1.4	0.2	0
7	5.4	3.9	1.7	0.4	0
8	4.6	3.4	1.4	0.3	0
9	5	3.4	1.5	0.2	0
10	4.4	2.9	1.4	0.2	0
11	4.9	3.1	1.5	0.1	0
12	5.4	3.7	1.5	0.2	0
13	4.8	3.4	1.6	0.2	0
14	4.8	3	1.4	0.1	0
15	4.3	3	1.1	0.1	0
16	5.8	4	1.2	0.2	0
17	5.7	4.4	1.5	0.4	0
18	5.4	3.9	1.3	0.4	0
19	5.1	3.5	1.4	0.3	0
20	5.7	3.8	1.7	0.3	0

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2
150 rows x 5 columns					

4. 군집화(Clustering) 구현 실습

K-means 를 사용하여 3개의 군집으로 만들어줘.



4. 군집화(Clustering) 구현 실습

Iris Data With Clusters						
	(cm)	petal width (cm)	label	cluster		
1		0.2	0	1		
2		0.2	0	1		
3		0.2	0	1		
4		0.2	0	1		
5		0.2	0	1		
6		0.4	0	1		

I have applied K-means clustering to the Iris dataset, forming three clusters. The clusters are visualized in the scatter plot, with colors indicating different clusters. You can also see the cluster assignments in the provided data.

4. 군집화(Clustering) 구현 실습

각 데이터가 속한 군집 중심점을 알려줘.

1. Cluster 0:

- Sepal Length: 5.901613 cm
- Sepal Width: 2.748387 cm
- Petal Length: 4.393548 cm
- Petal Width: 1.433871 cm

2. Cluster 1:

- Sepal Length: 5.006000 cm
- Sepal Width: 3.428000 cm
- Petal Length: 1.462000 cm
- Petal Width: 0.246000 cm

3. Cluster 2:

- Sepal Length: 6.850000 cm
- Sepal Width: 3.073684 cm
- Petal Length: 5.742105 cm
- Petal Width: 2.071053 cm

4. 군집화(Clustering) 구현 실습

label을 기준으로 군집에 속한 개수를 테이블로 보여줘.



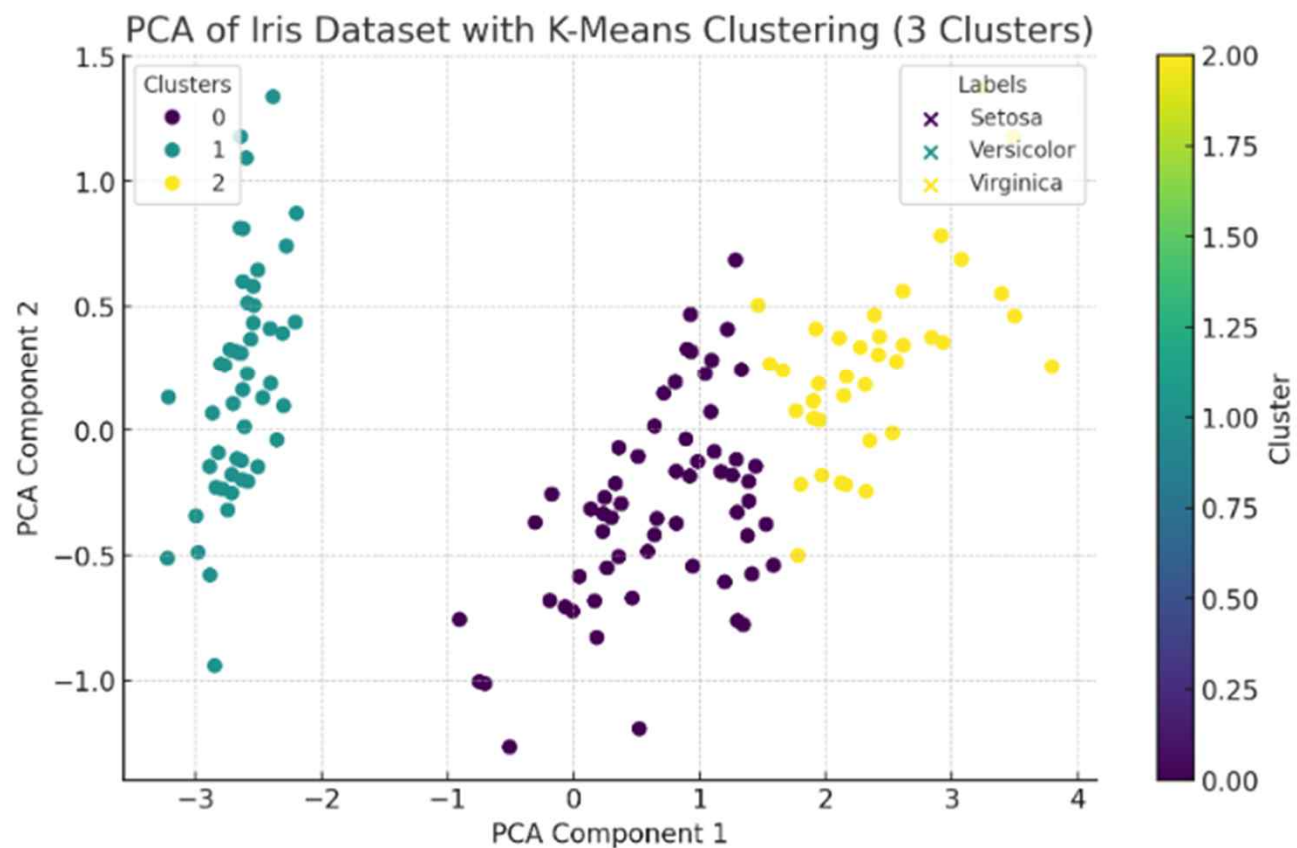
Cluster Label Counts					↓	↗
	Label	0	1	2		
1	0	0	50	0		
2	1	48	0	2		
3	2	14	0	36		

군집화 결과 시각화하기

군집화 된 데이터의 앞 네 개의 피처를 PCA를 사용하여 2개로 차원 축소한 다음 이 두개의 피처를 각각 x,y축으로 하고 군집값을 각각 다른 색으로 구별하여 산점도를 그려줘(그래프의 legend에 레이블의 종류를 표시해줘).

4. 군집화(Clustering) 구현 실습

PCA Of Iris Dataset With K-Means Clustering (3 Clusters)



- **Setosa** (Cluster 1)
- **Versicolor** (Cluster 0)
- **Virginica** (Cluster 2)

군집 평가(cluster Evaluation)

: 실루엣 계수(Silhouette Coefficient)를 사용한다

3개로 군집화한 평가지표인 실루엣 계수를 알려줘(한글로).



3개로 군집화한 K-means 클러스터링의 실루엣 계수는 약 **0.553**입니다. 이 점수는 군집들이 적절히 분리되고 내적으로 응집되어 있음을 나타냅니다. 실루엣 계수가 1에 가까울수록 군집화가 더 잘 되어 있음을 의미합니다.



4. 군집화(Clustering) 구현 실습

2개 3개 4개 5개 6개로 각각 군집화한 후 각각의 실루엣 계수를 비교해줘.

Silhouette Scores For Different Clusters			↓ ↗
	Number of Clusters	Silhouette Score	
1	2	0.6810461692117467	
2	3	0.5528190123564101	
3	4	0.49805050499728803	
4	5	0.48874888709310643	
5	6	0.3648340039670036	

2개로 군집화 했을 때 실루엣 계수가 가장 크다

감사합니다