1/21 ML 과제

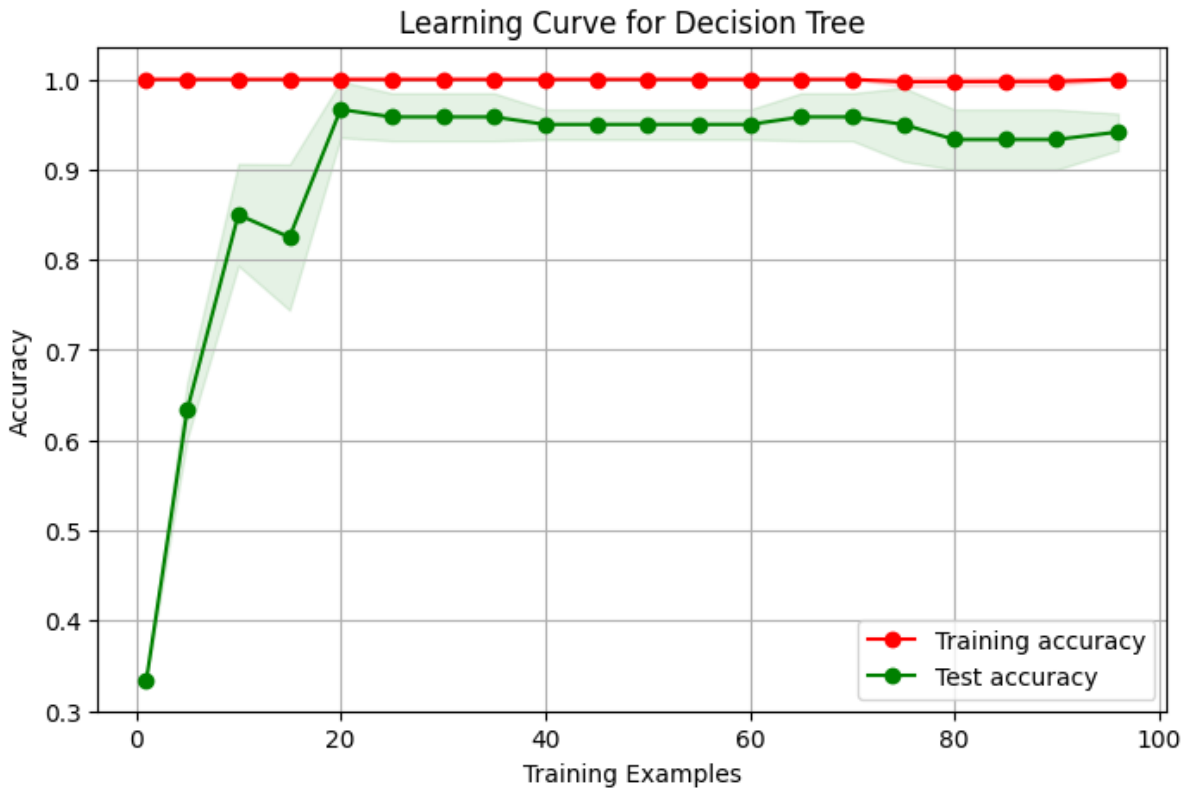
생성일	@2025년 1월 28일 오후 9:46
태그	

1. DT, RF
    1.1 Report the graphs as well as the train and test accuracy for both models in the report.
    1.2 What is the better model and please provide evidence and supporting arguments that back your decision?
2. PCA
    2.1 Report the plots in the report.
    2.2 What are the benefits and the disadvantages of PCA?

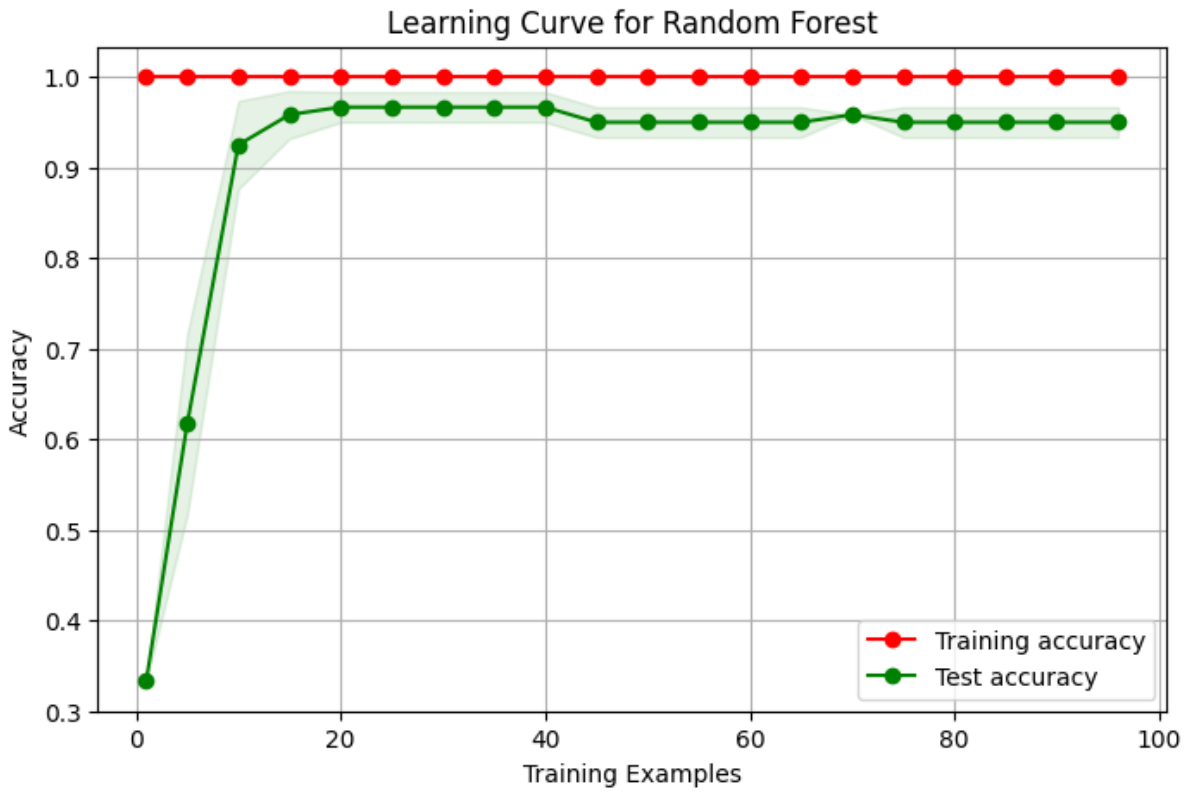
    Could you provide another dimensionality reduction method that can be used apart from PCA?
        장점
        단점
        대안 : Auto Encoder
3. SVM
    3.1 What is the difference between a soft margin SVM and hard margin SVM? Furthermore, can you provide advantages and disadvantages of both methods?
        Hard margin의 optimization problem.
        Soft margin의 optimization problem.
    3.2 Provide the plot in your report.

1. DT, RF

1.1 Report the graphs as well as the train and test accuracy for both models in the report.



Decision Tree - Training Accuracy: 1.0000, Test Accuracy: 0.9333



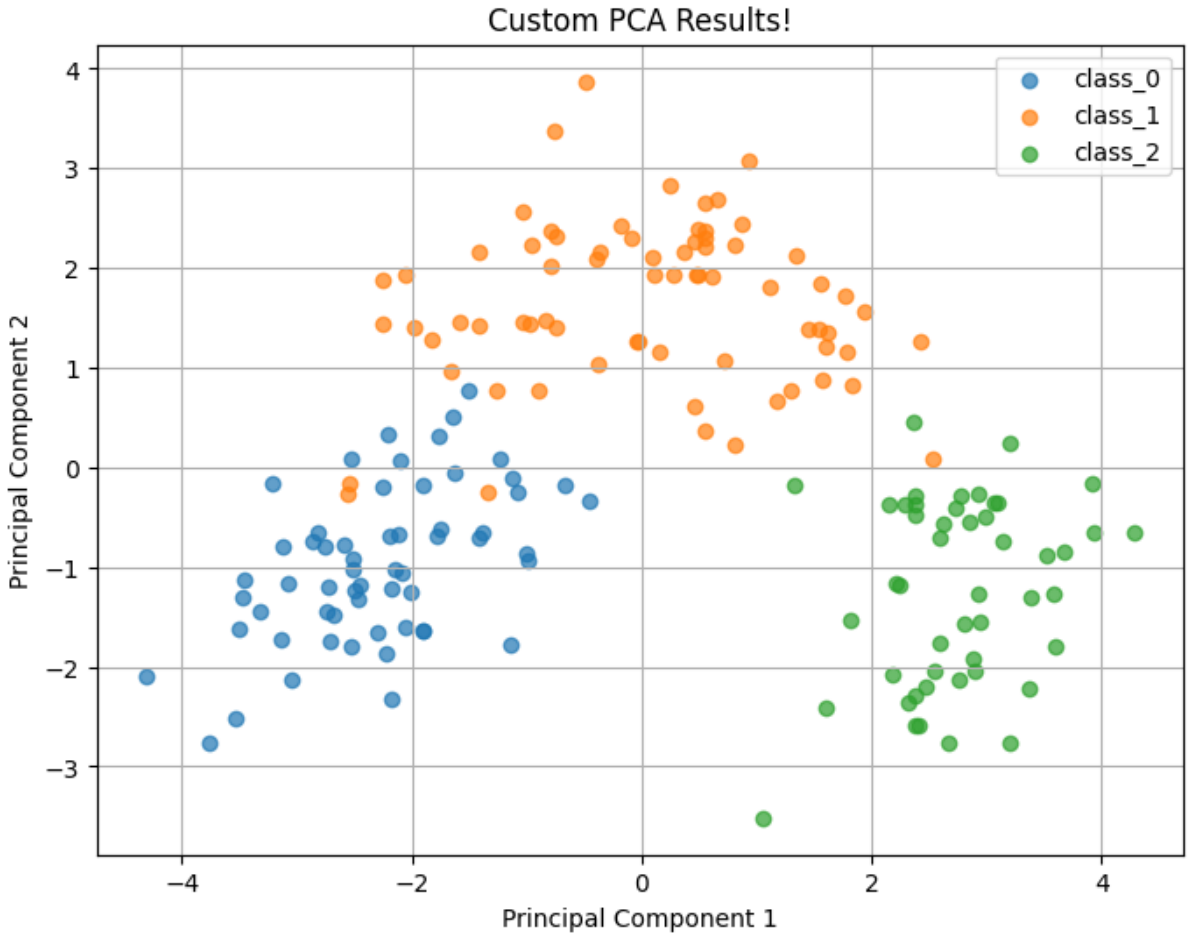
Random Forest - Training Accuracy: 1.0000, Test Accuracy: 0.9667

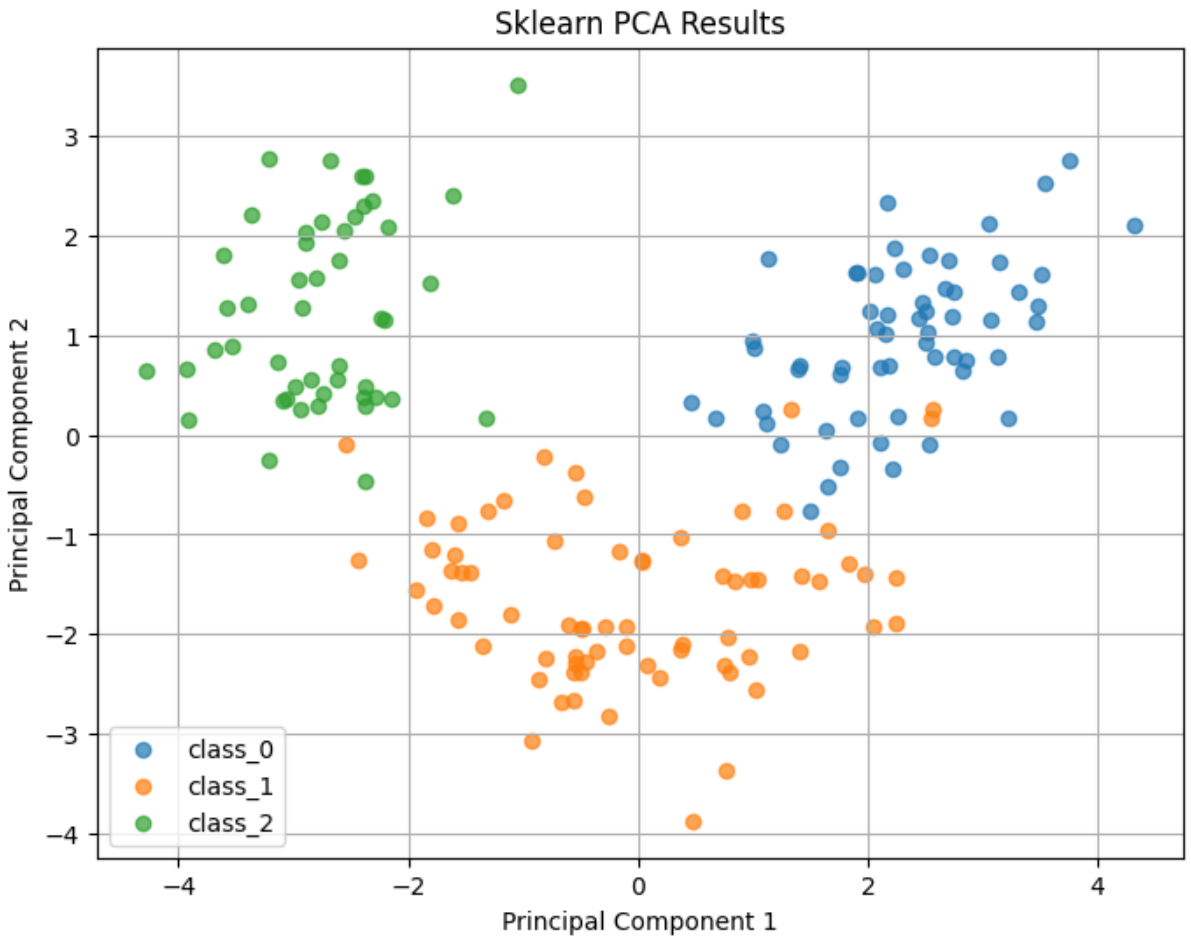
1.2 What is the better model and please provide evidence and supporting arguments that back your decision?

Random forest. 1) test accuracy 성능이 더 높았으며, 2) 적은 수의 sample에서 더 안정적이고 높은 test accuracy를 보여 과적합에 덜 민감한 강건성을 보이기 때문이다.

2. PCA

2.1 Report the plots in the report.





2.2 What are the benefits and the disadvantages of PCA?

Could you provide another dimensionality reduction method that can be used apart from PCA?

장점

고차원의 데이터를 그대로 분석에 활용하면, 데이터 차원이 밀집되지 않고, 유클리드 거리를 기반으로 하는 모델의 성능이 좋게 나타나지 않으며, 모델 학습 결과를 plot으로 확인하기 어려움 등등 여러 문제가 발생한다.

PCA는 고차원 데이터를, 가장 잘 설명하는 여러 orthogonal basis로 표현함으로써, 사용자는 전체 데이터의 정보가 너무 축소되지 않으면서도 충분히 적은 Principal component를 선택할 수 있다.

단점

PCA는 데이터를 잘 설명하는 여러 linear basis로 차원을 재구성하기 때문에 데이터의 구조가 선형적이지 않을 경우 PCA로 차원축소가 잘 이루어지지 않을 수 있다.

FA(Factor analysis) 처럼 축소한 basis에 이름을 붙일 수 없다. 축소된 PC1이 정확히 어떤 내용을 의미하는지 해석하기 어려울 수 있다.

Euclidian distance를 기반으로 하기 때문에, 각 변수의 스케일링 작업에 민감하다.

대안 : Auto Encoder

선형적 차원축소의 대안으로 딥러닝 기반 비선형적 차원 축소를 사용해 볼 수 있다. PCA와 동일하게 비지도 학습이며, 입력 데이터를 encoder와 decoder를 사용해 압축 후 복원하며 latent space로 비선형적 고차원 정보를 매핑할 수 있다.

3. SVM

3.1 What is the difference between a soft margin SVM and hard margin SVM? Furthermore, can you provide advantages and disadvantages of both methods?

Hard margin의 optimization problem.

$$\begin{aligned} \text{maximize:} \quad & \text{margin} = \frac{2}{\|\mathbf{w}\|} \\ \text{Subject to:} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i \end{aligned}$$

하드 마진은 데이터를 hyperplane으로 분리할 때, 모든 데이터가 margin 이상 떨어져 있어야 한다.

즉, yi 클래스 1,-1에 대해 두 데이터의 모델 계산 값이 hyperplane과 최대한 멀리 떨어져 있어야 함을 의미한다. (margin 최대화)데이터가 쉽게 분할되는 경우 모델 학습과 해석이 용이하지만, 완벽히 선형적으로 분리되지 않으면 적용이 어렵다. margin을 최대화 할 때, 모든 데이터를 만족시키기 위해 너무 작은 margin을 갖는 hyperplane을 고르는 등 과적합 될 가능성이 있다.

Soft margin의 optimization problem.

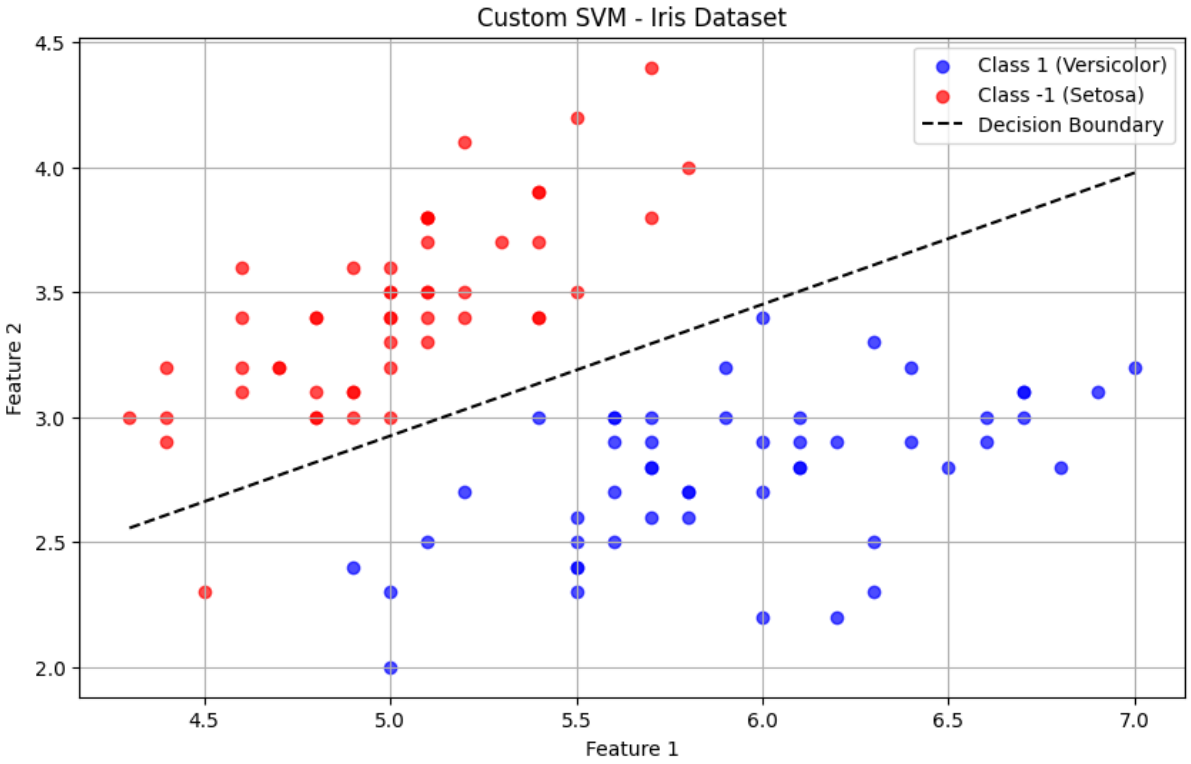
$$\min \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \right)$$

$$\begin{aligned} \text{Subject to:} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{aligned}$$

소프트 마진은 하드 마진에서 제약조건을 완화한다. 제약조건에 slack variable을 추가해 조건을 완화하고, 최적화 시 규제항을 포함해 slack variable의 크기 최소화를 목적함수에 반영한다.

이 slack variable을 통해 일부 데이터가 margin보다 가까이 분포하는 것을 허용해, 약간의 분류 오차가 있는 데이터에 대해서도 보다 일반화된 hyperplane을 선정할 수 있다.

3.2 Provide the plot in your report.


Custom SVM - Iris Dataset
Class 1 (Versicolor)
Class -1 (Setosa)
Decision Boundary
Feature 1
Feature 2
4.5
5.0
5.5
6.0
6.5
7.0
2.0
2.5
3.0
3.5
4.0
4.5