100 XP

학습 파이프라인 만들기 및 실행

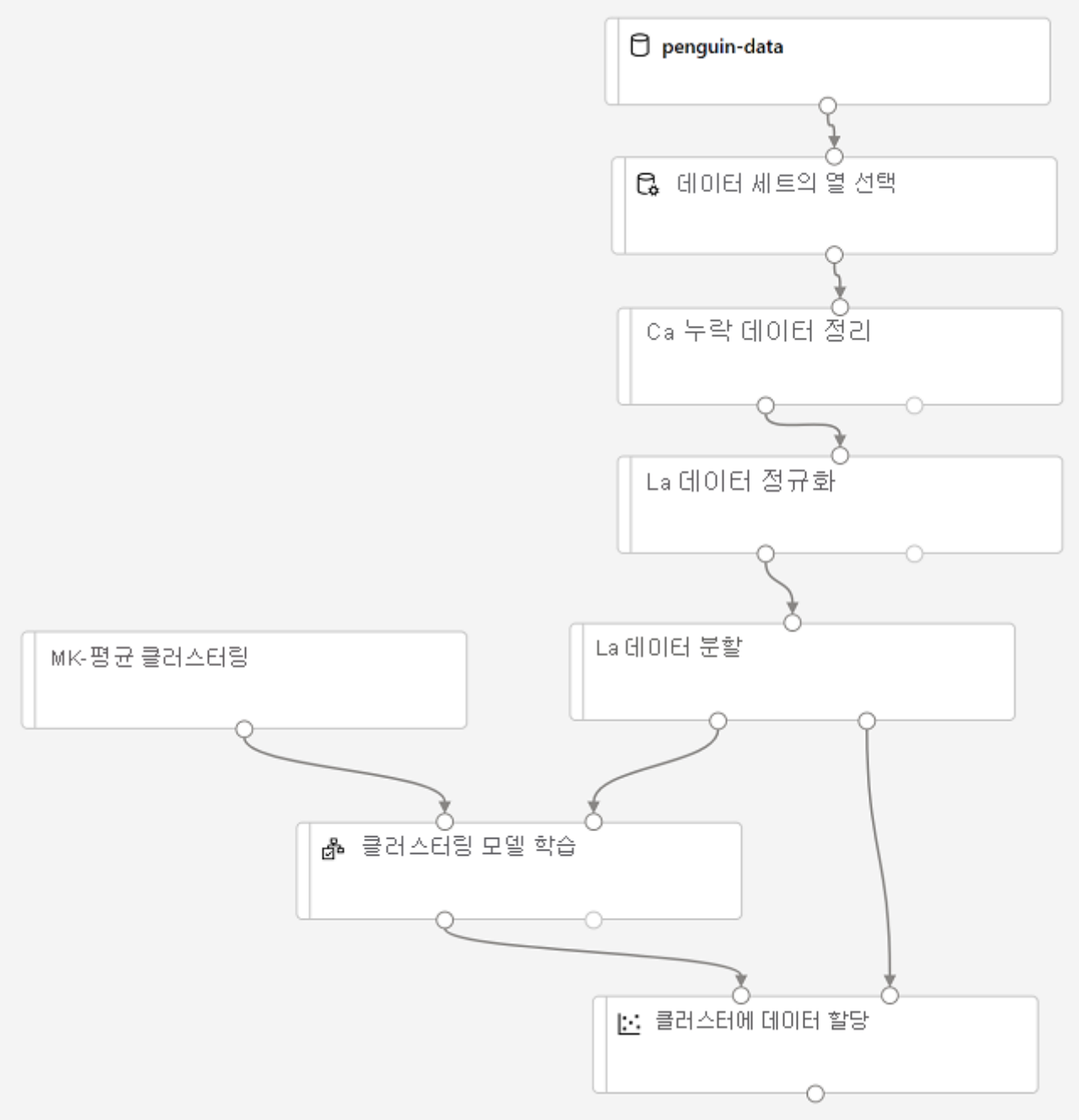
8분

데이터 변환을 사용하여 데이터를 준비한 후에는 이를 사용하여 기계 학습 모델을 학습할 수 있습니다.

학습 모듈 추가

클러스터링 모델을 학습하려면 데이터에 클러스터링 알고리즘을 적용하여 클러스터링을 위해 선택한 특징만을 사용해야 합니다. 데이터의 하위 집합을 사용하여 모델을 학습하고 나머지를 사용하여 학습된 모델을 테스트합니다.

이 연습에서는 다음과 같이 **Train Penguin Clustering** 파이프라인을 확장합니다.



필요한 모듈을 추가하고 구성하는 동안 위 정보를 참조로 사용하여 아래의 단계를 따릅니다.

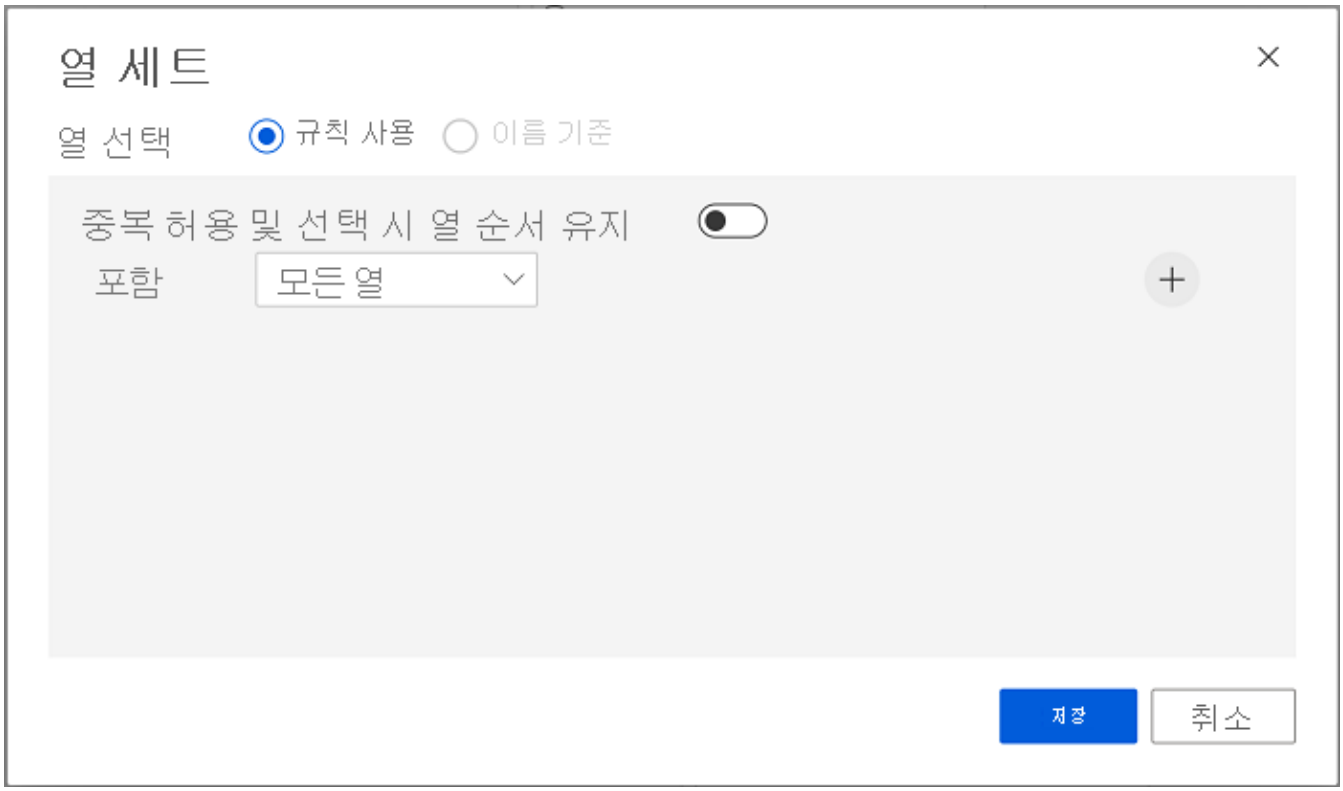
1. 아직 열려 있지 않은 경우 **Train Penguin Clustering** 파이프라인을 엽니다.
2. 왼쪽 창의 **데이터 변환** 섹션에서 **데이터 분할** 모듈을 **데이터 정규화** 모듈의 아래에 있는 캔버스로 끌어 놓습니다. 그런 다음 **데이터 정규화** 모듈의 왼쪽 출력과 **데이터 분할** 모듈의 입력을 연결합니다.
3. **데이터 분할** 모듈을 선택하고, 다음과 같이 설정을 구성합니다.

    - **분할 모드**: 행 분할
    - **첫 번째 출력 데이터 세트에서 행의 비율**: 0.7
    - **무작위 초기값**: 123
    - **계층화된 분할**: 아니요

4. 왼쪽 창에서 **모델 학습** 섹션을 확장하고 **데이터 분할** 모듈에 있는 캔버스로 **클러스터링 모델 학습** 모듈을 끌어다 놓습니다. 그런 다음 **분할 데이터** 모듈의 *Result dataset1*(왼쪽) 출

력을 **클러스터링 모델 학습** 모듈의 데이터 세트(오른쪽) 입력에 연결합니다.

5. 클러스터링 모델은 원래 데이터 세트에서 선택한 모든 특징을 사용하여 데이터 항목에 클러스터를 할당해야 합니다. **클러스터링 모델 학습** 모듈을 선택하고 설정 창에 있는 **매개 변수** 탭에서 **열 편집** 을 선택하고 **규칙 사용** 옵션을 사용하여 다음과 같이 모든 열을 포함합니다.



6. 학습 중인 모델은 특징을 사용하여 데이터를 클러스터로 그룹화하므로 클러스터링 알고리즘을 사용하여 모델을 학습해야 합니다. **기계 학습 알고리즘** 섹션을 확장한 뒤, **클러스터링** 에서 **K-평균 클러스터링** 모듈을 **penguin-data** 데이터 세트의 왼쪽, 그리고 **클러스터링 모델 학습** 모듈 위에 있는 캔버스로 끌어다 놓습니다. 그런 다음 출력을 **클러스터링 모델 학습** 모듈의 **학습되지 않은 모델**(왼쪽) 입력에 연결합니다.

7. 'K-평균' 알고리즘은 항목을 지정한 클러스터의 수, 즉 **K** 값으로 그룹화합니다. _ *K-평균 클러스터링**을 선택하고 설정 창에 있는 **매개 변수** 탭에서 **중심의 수** 매개 변수를 **3** 으로 설정합니다.

> **참고**
>
> 펭귄 측정값과 같은 데이터 관찰을 다차원 벡터라고 생각할 수 있습니다. K-평균은 다음과 같은 방식으로 작동합니다.
>
> a. *K* 좌표를 *n* 차원 공간의 중심 이라는 무작위로 선택된 지점으로 초기화합니다. 여기서 *n* 은 특징 벡터의 차원 수입니다.
> b. 특징 벡터를 동일한 공간의 지점으로, 각 지점을 가장 가까운 중심에 할당합니다.
> c. 중심을 그에 할당된 지점의 중앙으로 이동합니다(평균 거리를 기준으로 함).
> d. 이동 후에 가장 가까운 중심에 지점을 다시 할당합니다.

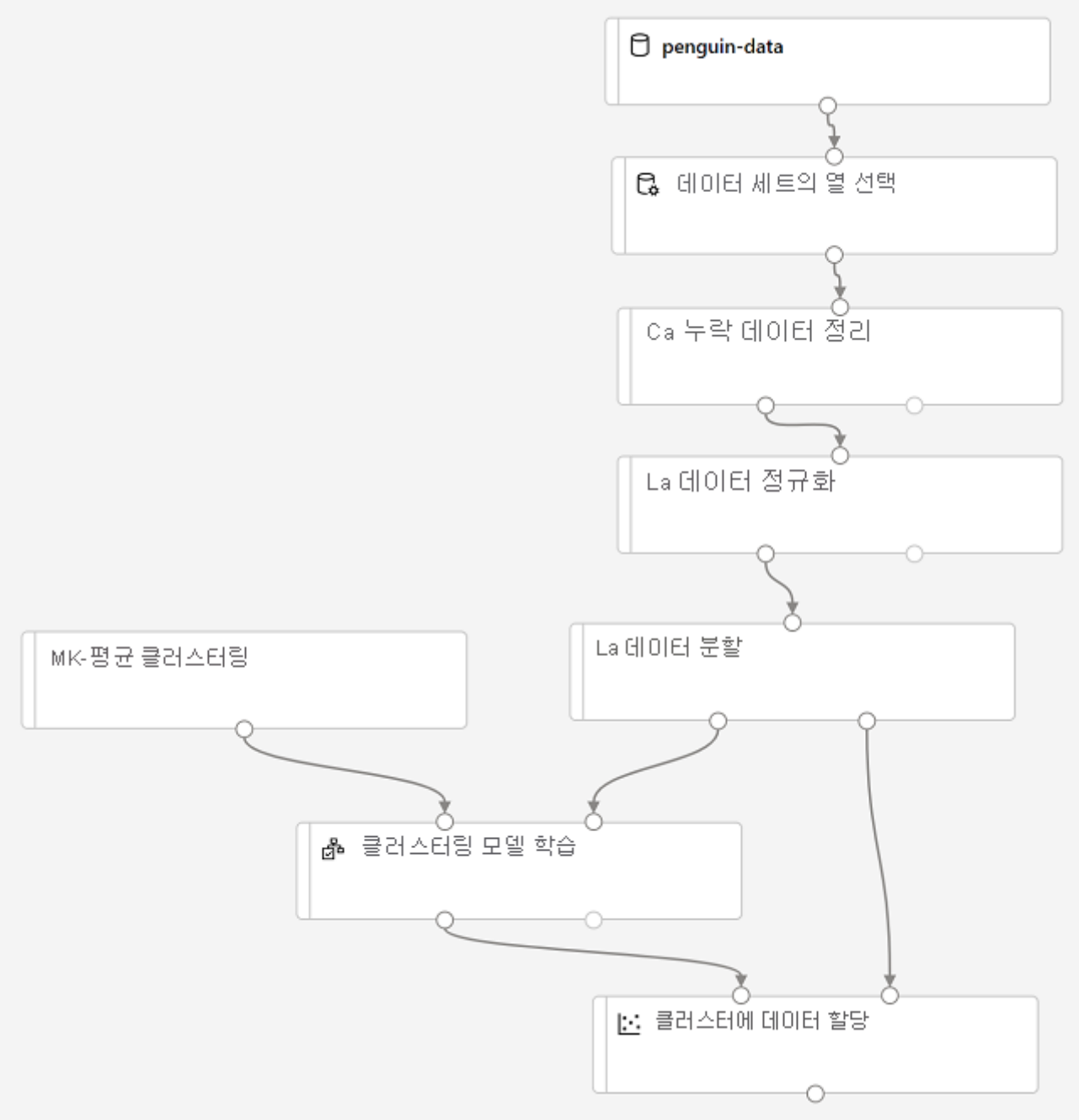
e. 클러스터 할당이 안정화되거나 지정된 반복 횟수가 완료될 때까지 3단계와 4단계를 반복합니다.

8. 데이터의 70%를 사용하여 클러스터링 모델을 학습한 후에는 나머지 30%를 사용하여 데이터를 클러스터에 할당하는 모델을 사용함으로써 이를 테스트할 수 있습니다. **모델 채점 및 평가** 섹션을 확장하고 **클러스터에 데이터 할당** 모듈을 **클러스터링 모델 학습** 모듈 아래에 있는 캔버스로 끌어다 놓습니다. 그런 다음 **클러스터링 모델 학습** 의 **학습된 모델**(왼쪽) 출력을 **클러스터에 데이터 할당** 모듈의 **학습된 모델**(왼쪽) 입력에 연결합니다. 그리고 **데이터 분할** 모듈의 **결과 데이터 세트2**(오른쪽) 출력을 **클러스터에 데이터 할당** 모듈의 **데이터 세트**(오른쪽) 입력에 연결합니다.

학습 파이프라인 실행

이제 학습 파이프라인을 실행하고 모델을 학습할 준비가 되었습니다.

1. 파이프라인이 다음과 같아야 합니다.



2. **제출** 을 선택하고, 컴퓨팅 클러스터에서 **mslearn-penguin-training** 이라는 기존 실험을 사용하여 파이프라인을 실행합니다.
3. 실험이 완료될 때까지 기다립니다. 5분 이상 걸릴 수 있습니다.
4. 실험이 완료되면 **클러스터에 데이터 할당** 모듈을 선택하고 설정 창에서 **출력 + 로그** 탭의 **결과 데이터 세트** 섹션에 있는 **데이터 출력** 에서 **시각화** 아이콘을 사용하여 결과를 봅니다.
5. 오른쪽으로 스크롤하여 각 행이 할당된 클러스터(0, 1 또는 2)를 포함하는 **할당** 열을 확인합니다. 또한 이 행을 나타내는 지점에서 각 클러스터의 중심까지의 거리를 나타내는 새 열이 있습니다. 지점과 가장 가까운 클러스터가 할당된 클러스터입니다.
6. **클러스터에 데이터 할당** 시각화를 닫습니다.

모델에서 펭귄 관찰에 대한 클러스터를 예측할 때 그 예측은 얼마나 안정적인가요? 이를 평가하려면 모델을 평가해야 합니다.

다음 단원: 클러스터링 모델 평가

계속 >