100 XP

데이터 탐색

8분

클러스터링 모델을 학습하려면 클러스터링할 항목의 다양한 관찰 결과를 포함하는 데이터 세트가 필요합니다. 여기에는 개별 사례를 클러스터로 분리하는 데 유용한 개별 사례 간의 유사성을 확인할 수 있는 숫자 기능이 포함됩니다.

데이터 세트 만들기

Azure Machine Learning에서 모델 학습 및 기타 작업의 데이터는 주로 데이터 세트라는 개체에 캡슐화됩니다. 이 모듈에서는 세 가지 종의 펭귄에 대한 관찰을 포함하는 데이터 세트를 사용합니다.

1. Azure Machine Learning Studio 에서 **데이터 세트** 페이지를 확인합니다. 데이터 세트는 Azure ML에서 사용할 특정 데이터 파일이나 테이블을 나타냅니다.
2. 다음 설정을 사용하여 웹 파일에서 데이터 세트를 만듭니다.

    - **기본 정보**:
        - **웹 URL**: https://aka.ms/penguin-data
        - **이름**: penguin-data
        - **데이터 세트 형식**: 테이블 형식
        - **설명**: 펭귄 데이터
    - **설정 및 미리 보기**:
        - **파일 형식**: 구분 기호로 분리됨
        - **구분 기호**: 쉼표
        - **인코딩**: UTF-8
        - **열 헤더**: 첫 번째 파일의 헤더 사용
        - **행 건너뛰기**: 없음
    - **스키마**:
        - **경로** 이외의 모든 열 포함
        - 자동으로 검색된 형식 검토
    - **세부 정보 확인**:
        - 만든 후 데이터 세트를 프로파일링하지 않음

3. 데이터 세트를 만든 후에는 이를 열고 **탐색** 페이지를 보면서 데이터 샘플을 확인합니다. 이 데이터는 여러 번 펭귄을 관찰하면서 얻은 부리 길이와 두께, 날개 길이, 무게 측정값을 나타냅니다. 데이터 세트에는 세 가지 종의 펭귄이 표시되어 있습니다. 아멜리 펭귄, 젠투 펭귄, 턱끈 펭귄입니다.

> **참고**
>
> 이 연습에서 사용되는 펭귄 데이터 세트는 Kristen Gorman 박사와 Long Term Ecological Research Network 의 일원인 Palmer Station, Antarctica LTER 에서 수집 및 제공한 것입니다.

파이프라인 만들기

Azure Machine Learning 디자이너를 시작하려면 먼저 파이프라인을 만들고 작업하려는 데이터 세트를 추가해야 합니다.

1. 작업 영역의 Azure Machine Learning Studio 에서 **디자이너** 페이지를 보고 새 파이프라인을 만듭니다.
2. **설정** 창에서 기본 파이프라인 이름(*Pipeline-Created-on- date***)을 **Train Penguin Clustering** 으로 변경합니다(**설정** 창이 표시되지 않은 경우 상단 파이프라인 이름 옆에 있는 ⚙ 아이콘을 클릭).
3. 파이프라인을 실행할 컴퓨팅 대상을 지정해야 합니다. **설정** 창에서 **컴퓨팅 대상 선택** 을 클릭하고 이전에 만든 컴퓨팅 클러스터를 선택합니다.
4. 디자이너의 왼쪽 창에서 **데이터 세트** 섹션을 확장한 다음 이전 연습에서 만든 데이터 세트를 캔버스에 끌어다 놓습니다.
5. 캔버스에 있는 **penguin-data** 데이터 세트를 마우스 오른쪽 단추로 클릭(Mac의 경우 Ctrl+클릭)하고 **시각화** 메뉴에서 **데이터 세트 출력** 을 선택합니다.
6. 데이터의 스키마를 검토하여 다양한 열의 분포를 히스토그램으로 확인할 수 있습니다. 그런 다음 **CulmenLength** 열을 선택합니다. 데이터 세트는 다음과 비슷합니다.

7. 데이터 세트의 특성은 다음과 같습니다.

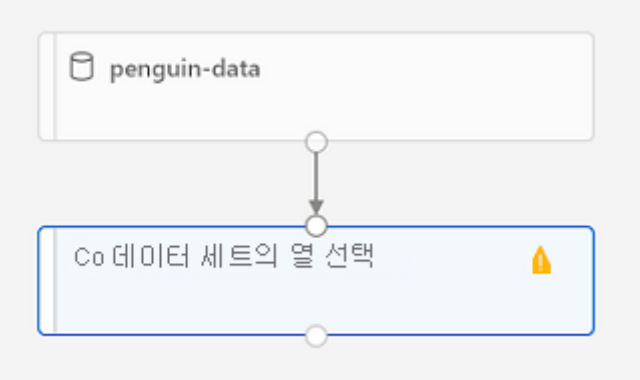
    - 데이터 세트에는 다음 열이 포함되어 있습니다.
        - **CulmenLength**: 펭귄의 부리 길이(밀리미터)입니다.
        - **CulmenDepth**: 펭귄의 부리 깊이(밀리미터)입니다.
        - **FlipperLength**: 펭귄의 날개 길이(밀리미터)입니다.
        - **BodyMass**: 펭귄의 무게(그램)입니다.
        - **Species**: 종을 나타냅니다(0:"Amelie", 1:"Gentoo", 2:"Chinstrap").
    - **CulmenLength** 열에 누락된 값이 2개 있습니다(**CulmenDepth**, **FlipperLength** 및 **BodyMass** 열에도 누락된 값이 2개 있음).
    - 측정값은 규모가 다릅니다(수십 밀리미터에서 수천 그램까지).

8. 파이프라인 캔버스의 데이터 세트를 볼 수 있도록 데이터 세트 시각화를 닫습니다.

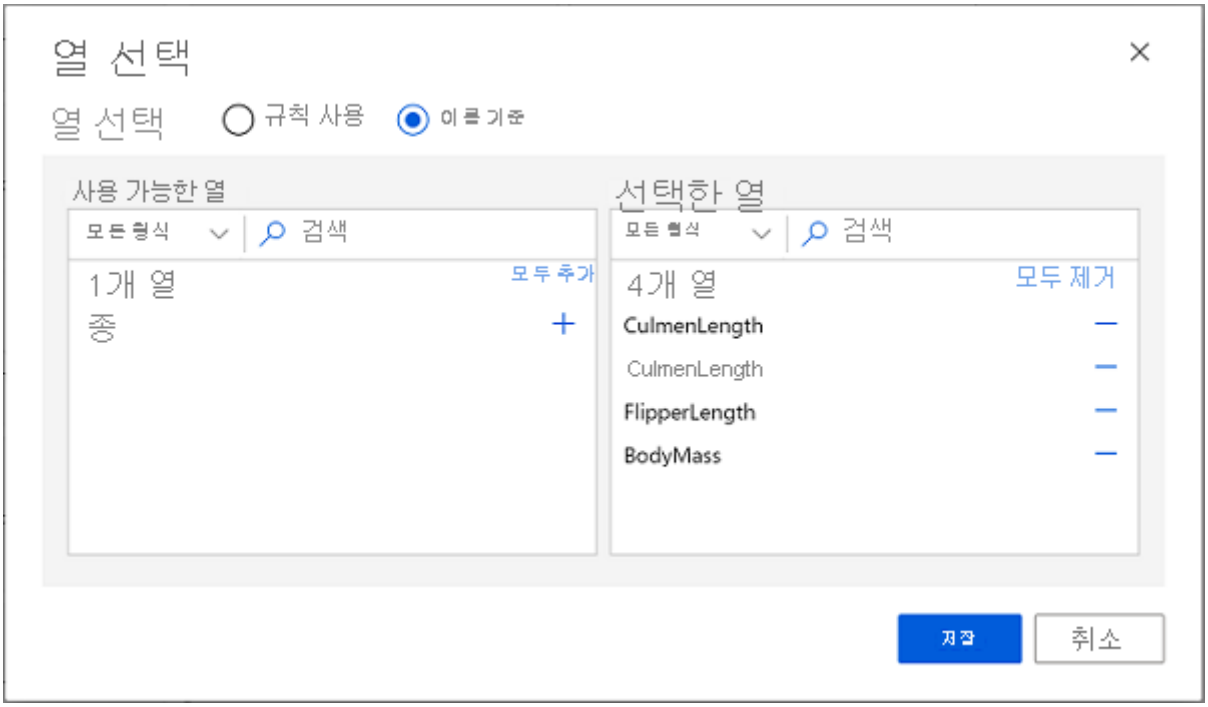
변환 적용

펭귄 관찰 결과를 클러스터링하기 위해 측정값만을 사용하고 종 열은 버리도록 하겠습니다. 또한 값이 누락된 행을 제거하고, 규모가 비슷해지도록 숫자 측정값을 정규화해야 합니다.

1. 왼쪽 창에서 **데이터 변환** 섹션을 펼칩니다. 여기에는 모델 학습 전에 데이터 변환에 사용할 수 있는 다양한 모듈이 포함되어 있습니다.
2. 펭귄 관찰을 클러스터링하기 위해 측정값만을 사용하고 종 열은 무시하도록 하겠습니다. 따라서 **데이터 세트에서 열 선택** 모듈을 **penguin-data** 모듈 아래에 있는 캔버스로 끌어다 놓고 **penguin-data** 모듈 하단의 출력을 **데이터 세트에서 열 선택** 모듈 상단의 입력에 연결합니다.



3. **데이터 세트에서 열 선택** 모듈을 선택하고 오른쪽에 있는 **설정** 창에서 **열 편집** 을 선택합니다. 그런 다음 **열 선택** 창에서 **이름 기준** 을 선택하고 + 링크를 사용하여 열 이름 **CulmenLength**, **CulmenDepth**, **FlipperLength** 및 **BodyMass** 를 선택합니다.



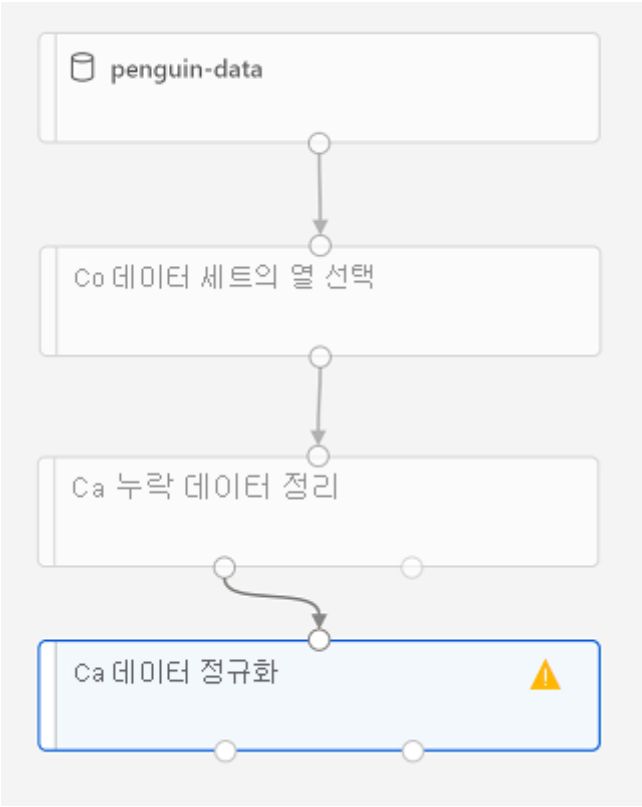
4. **데이터 세트에서 열 선택** 모듈 설정을 저장하고 디자이너 캔버스로 돌아옵니다.
5. **누락된 데이터 정리** 모듈을 **데이터 세트에서 열 선택** 모듈 아래에 있는 캔버스로 끌어다 놓은 다음 이렇게 연결합니다.

6. **누락된 데이터 정리** 모듈을 선택하고 오른쪽의 설정 창에서 **열 편집** 을 클릭합니다. 그런 다음 **열 선택** 창에서 **규칙 사용** 을 선택하고 **모든 열** 을 포함합니다.

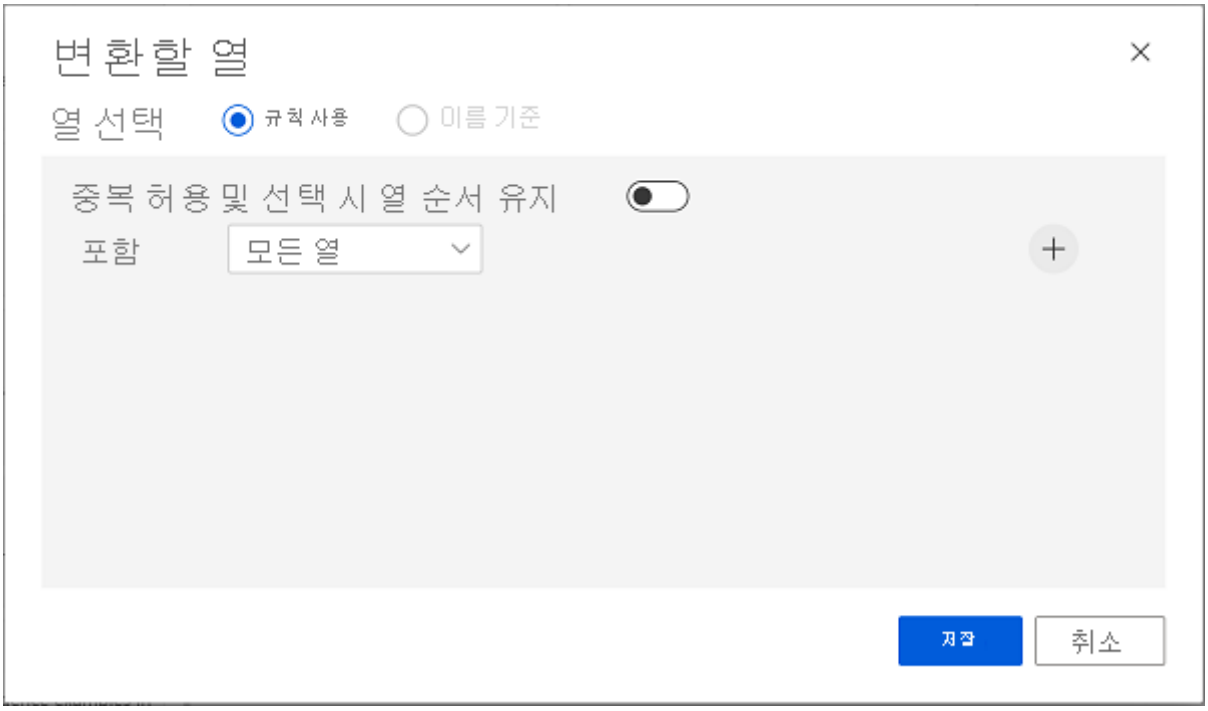
7. **누락된 데이터 정리** 모듈을 선택한 상태로 설정 창에서 다음 구성을 설정합니다.

   - **최소 누락 값 비율**: 0.0
   - **최대 누락 값 비율**: 1.0
   - **정리 모드**: 전체 행 제거

8. **데이터 정규화** 모듈을 **누락된 데이터 정리** 모듈 아래에 있는 캔버스로 끌어다 놓습니다. 그런 다음 **누락된 데이터 정리** 모듈의 가장 왼쪽에 있는 출력과 **데이터 정규화** 모듈의 입력을 연결합니다.



9. **데이터 정규화** 모듈을 선택하고 오른쪽에 있는 **설정** 창에서 **변환 메서드** 를 MinMax 로 설정하고 **열 편집** 을 선택합니다. 그런 다음 **열 선택** 창에서 **규칙 사용** 을 선택하고 **모든 열** 을 포함합니다.



10. **데이터 정규화** 모듈 설정을 저장하고 디자이너 캔버스로 돌아옵니다.

파이프라인 실행

데이터 변환을 적용하려면 파이프라인을 실험으로 실행해야 합니다.

1. 파이프라인이 다음과 유사해야 합니다.

2. **제출** 을 선택하고, 컴퓨팅 클러스터에서 **mslearn-penguin-training** 이라는 새 실험을 사용하여 파이프라인을 실행합니다.
3. 실행이 끝날 때까지 기다립니다. 5분 이상 걸릴 수 있습니다. 실행이 완료되면 모듈은 다음과 같습니다.

변환된 데이터 보기

이제 모델 학습을 위한 데이터 세트가 준비되었습니다.

1. 완료된 **데이터 정규화** 모듈을 선택하고 오른쪽 **설정** 창의 **출력 + 로그** 탭에서 **변환된 데이터 세트** 에 대한 **시각화** 아이콘을 선택합니다.
2. 데이터를 보면 **Species** 열이 제거되어 누락된 값이 없고 4가지 모든 특징에 대한 값이 일반적인 규모로 정규화되었음을 알 수 있습니다.
3. 정규화된 데이터의 결과 시각화를 닫습니다.

데이터 세트에서 사용하려는 특징을 선택 및 준비했으므로 이제 이를 사용하여 클러스터링 모델을 학습할 준비가 되었습니다.

다음 단원: 학습 파이프라인 만들기 및 실행

계속 >