100 XP

데이터 살펴보기

10분

분류 모델을 학습하려면, 기존 특징(예측의 대상이 되는 엔터티의 특성)과 알려진 레이블 값(모델을 학습해 예측할 클래스 지표)이 포함된 데이터 세트가 필요합니다.

데이터 세트 만들기

Azure Machine Learning에서 모델 학습 및 기타 작업의 데이터는 주로 데이터 세트라는 개체에 캡슐화됩니다.

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.

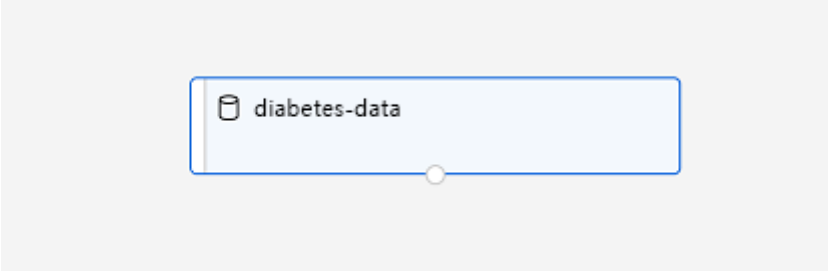
    - **기본 정보**:
        - **웹 URL**: https://aka.ms/diabetes-data
        - **이름**: diabetes-data
        - **데이터 세트 형식**: 테이블 형식
        - **설명**: 당뇨병 데이터
    - **설정 및 미리 보기**:
        - **파일 형식**: 구분 기호로 분리됨
        - **구분 기호**: 쉼표
        - **인코딩**: UTF-8
        - **열 머리글**: 첫 번째 파일의 머리글 사용
        - **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 당뇨병에 대한 테스트를 받은 환자의 세부 정보를 나타냅니다.

파이프라인 만들기

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

1. 작업 영역의 Azure Machine Learning Studio 에서 **디자이너** 페이지를 보고 + 를 선택하여 새 파이프라인을 만듭니다.
2. **설정** 창에서 기본 파이프라인 이름(*Pipeline-Created-on- date***)을 **Diabetes Training** 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 **aml-cluster** 컴퓨팅 클러스터를 선택합니다.
4. 디자이너의 왼쪽에서 **데이터 세트** 섹션을 확장한 다음 이전 연습에서 만든 **diabetes-data** 데이터 세트를 캔버스에 끌어다 놓습니다.
5. 캔버스에 있는 **diabetes-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.
7. 오른쪽으로 스크롤하여 **Diabetic** 열의 열 머리글을 선택하고 **0** 과 **1**, 두 값을 포함하고 있는지 확인합니다. 이러한 값은 모델에서 예측하는 레이블에 대해 가능한 두 가지 클래스를 나타내며, **0** 값은 환자가 당뇨병이 아니라는 의미, **1** 값은 당뇨병이라는 의미입니다.
8. 왼쪽으로 다시 스크롤하고 다른 열을 검토합니다. 이 열은 레이블을 예측하는 데 사용되는 특징을 나타냅니다. 이러한 열의 대부분은 숫자이지만 기능마다 크기가 다릅니다. 예를 들어 **Age** 값의 범위는 21~77인 반면, **DiabetesPedigree** 값의 범위는 0.078~2.3016입니다. 기계 학습 모델을 학습할 때 큰 값에 결과 예측 함수가 좌지우지되어서 작은 규모에 대한 특징의 영향을 줄이는 경우가 간혹 있을 수 있습니다. 일반적으로 데이터 과학자는 숫자 열을 비슷한 기준에 기초하도록 정규화하여 편중을 최소화합니다.
9. 다음과 같이 캔버스에서 데이터 세트를 볼 수 있도록 **diabetes-data의 결과 시각화** 창을 닫습니다.

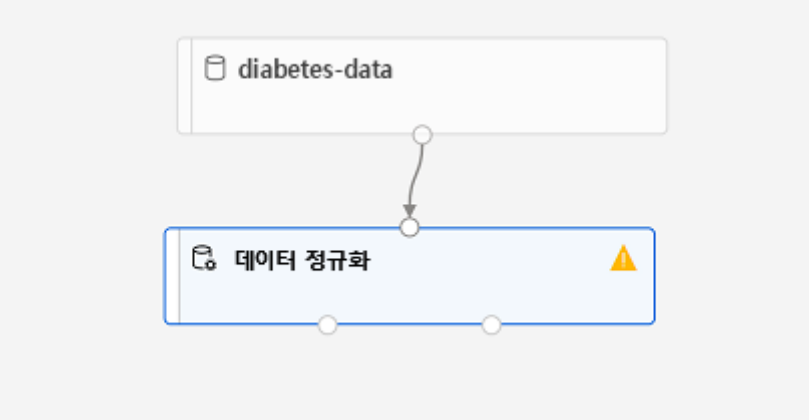


변환 추가

모델을 학습하려면 보통 데이터에 일부 전처리 변환을 적용해야 합니다.

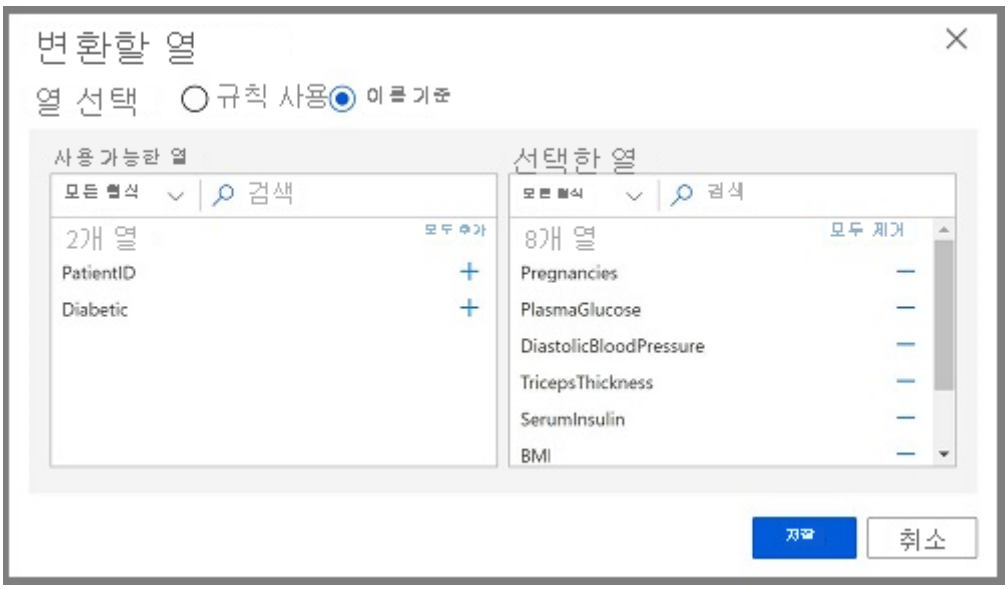
1. 왼쪽 창에서 **데이터 변환** 섹션을 펼칩니다. 여기에는 모델 학습 전에 데이터 변환에 사용할 수 있는 다양한 모듈이 포함되어 있습니다.
2. **데이터 정규화** 모듈을 **diabetes-data** 데이터 세트 아래에 있는 캔버스로 끌어다 놓습니다. 그런 다음 **diabetes-data** 데이터 세트 하단의 출력을 다음과 같이 **데이터 정규화** 모듈 상

단의 입력에 연결합니다.



3. **데이터 정규화** 모듈을 선택하고 설정을 확인합니다. 여기서 변환 방법 및 변환할 열을 지정해야 한다는 것을 알 수 있습니다.
4. 이미지에 표시된 것처럼 변환을 **MinMax** 로 설정하고 다음 열을 이름으로 포함하도록 열을 편집합니다.

    - **Pregnancies**
    - **PlasmaGlucose**
    - **DiastolicBloodPressure**
    - **TricepsThickness**
    - **SerumInsulin**
    - **BMI**
    - **DiabetesPedigree**
    - **Age**

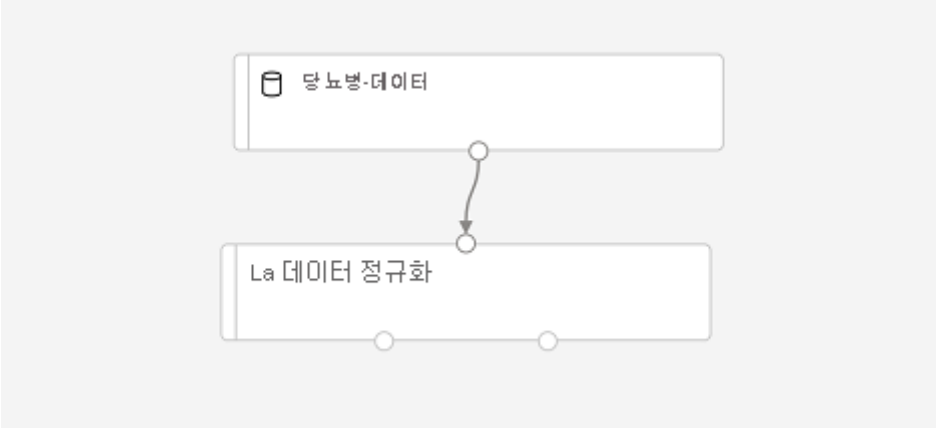


데이터 변환은 숫자 열을 정규화하여 동일한 규모로 배치하므로, 큰 값이 있는 열이 모델 학습을 좌지우지하지 않도록 합니다. 일반적으로 학습을 위한 데이터를 준비하려고 이와 같은 일련의 사전 처리 변환을 적용하지만 이 연습에서는 작업을 간단하게 유지합니다.

파이프라인 실행

데이터 변환을 적용하려면 파이프라인을 실험으로 실행해야 합니다.

1. 파이프라인이 다음과 유사해야 합니다.



2. **제출** 을 선택하고 컴퓨팅 클러스터에서 **mslearn-diabetes-training** 이라는 새 실험으로 파이프라인을 실행합니다.
3. 실행이 완료될 때까지 기다려 주세요. 몇 분 정도 걸릴 수 있습니다.

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 보고 선택한 숫자 열이 공통 배율로 정규화된 것을 확인합니다.
3. 정규화된 데이터의 결과 시각화를 닫습니다.

다음 단원: 학습 파이프라인 만들기 및 실행

계속 >