

학습 파이프라인 만들기 및 실행

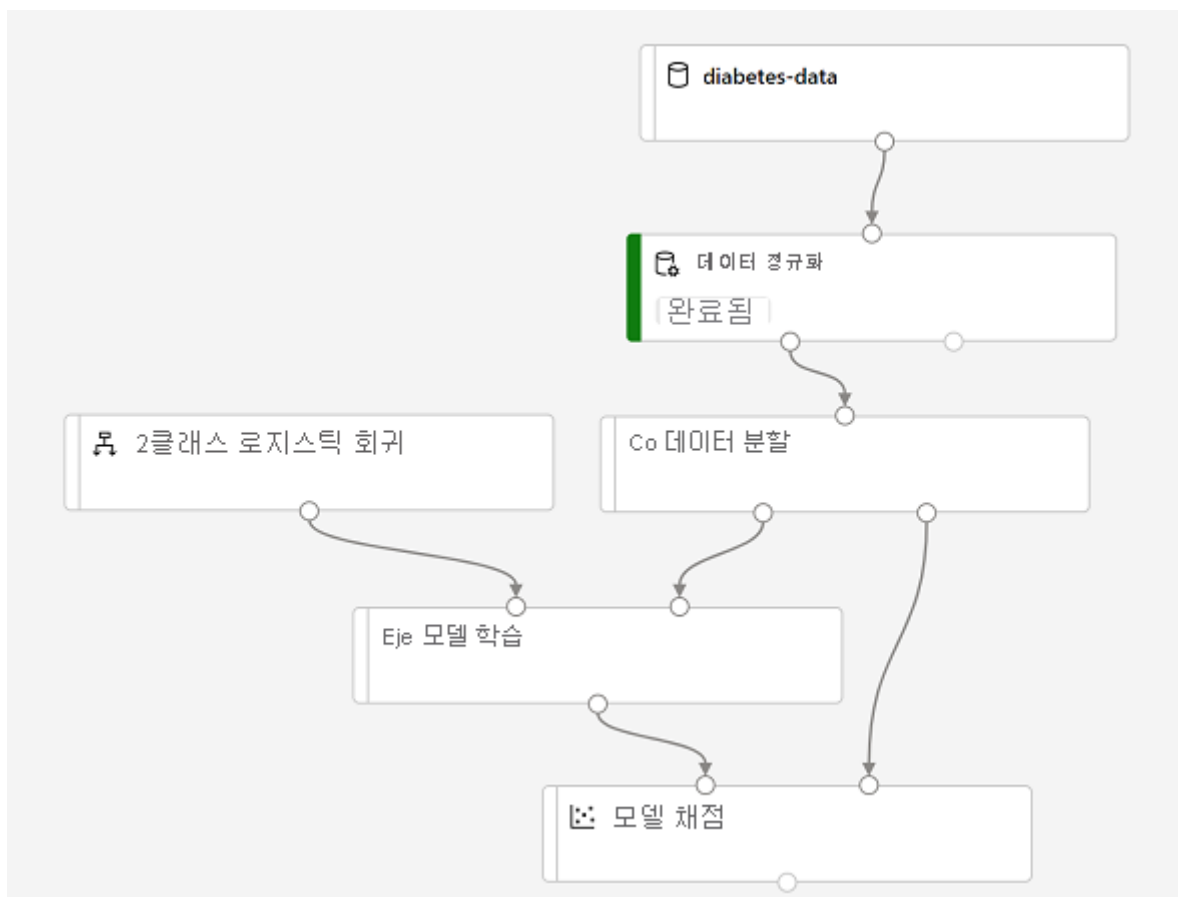
12분

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

학습 모듈 추가

모델을 학습하는 데 사용하는 데이터와, 학습된 모델을 테스트하는 데 사용할 데이터를 서로 구분하는 것이 일반적입니다. 이렇게 하면 모델에서 예측하는 레이블을 원래 데이터 세트의 실제 알려진 레이블과 비교할 수 있습니다.

이 연습에서는 다음과 같이 **당뇨병 학습** 파이프라인을 확장합니다.



아래 단계를 따르고, 필요한 모듈을 추가하고 구성할 때 위 이미지를 참조용으로 사용합니다.

1. 아직 열려 있지 않다면 이전 단원에서 만든 **당뇨병 학습** 파이프라인을 엽니다.
2. 왼쪽 창의 **데이터 변환** 섹션에서 **데이터 분할** 모듈을 **데이터 정규화** 모듈의 아래에 있는 캔버스로 끌어 놓습니다. 그런 다음 **데이터 정규화** 모듈의 '변환된 데이터 세트'(왼쪽) 출력을 **데이터 분할** 모듈의 입력에 연결합니다.

3. 데이터 분할 모듈을 선택하고, 다음과 같이 설정을 구성합니다.

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

4. 왼쪽 창에서 **모델 학습** 섹션을 확장하고 **데이터 분할** 모듈 아래에 있는 캔버스로 **모델 학습** 모듈을 끌어다 놓습니다. 그런 다음 **분할 데이터** 모듈의 Result dataset1(왼쪽) 출력을 **모델 학습** 모듈의 데이터 세트(오른쪽) 입력에 연결합니다.

5. 학습하고 있는 모델은 **당뇨** 값을 예측하므로 **모델 학습** 모듈을 선택하고 해당 설정을 수정하여 **레이블 열** 을 **Diabetic**(당뇨)으로 설정합니다(대문자 표시 및 철자 정확히 일치).

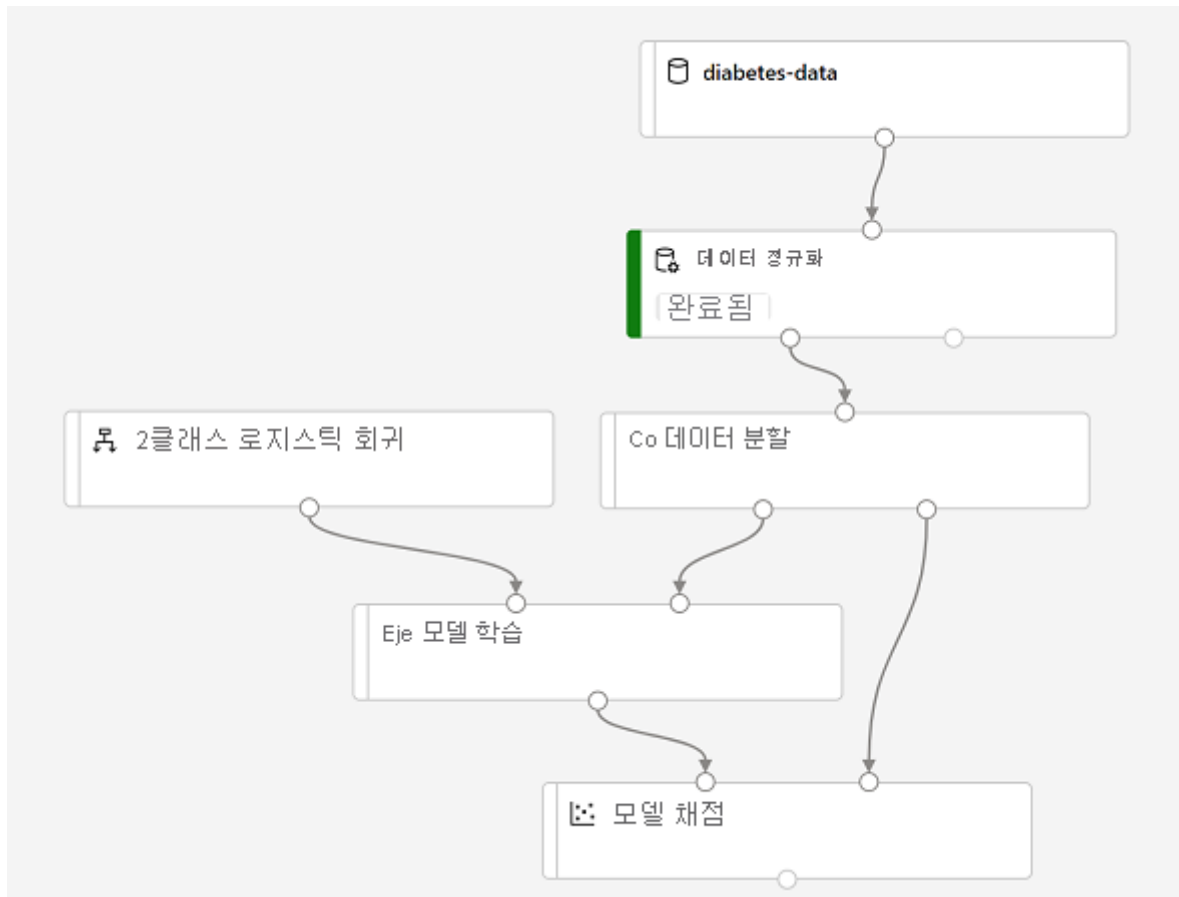
6. 모델에서 예측하는 **Diabetic** 레이블은 클래스(0 또는 1)이므로 분류 알고리즘을 사용하여 모델을 학습해야 합니다. 특히 가능한 클래스가 두 가지이므로 이진 분류 알고리즘이 필요합니다. **기계 학습 알고리즘** 섹션을 펼치고 **분류** 에서 **2 클래스 로지스틱 회귀 분석** 모듈을 **데이터 분할** 모듈의 왼쪽, **모델 학습** 모듈 위의 캔버스로 끌어 놓습니다. 그런 다음 출력을 **모델 학습** 모듈의 **학습되지 않은 모델**(왼쪽) 입력에 연결합니다.

❗ 참고

분류 모델을 학습하는 데 사용할 수 있는 여러 알고리즘이 있습니다. 이를 선택하는 데 도움이 필요하면 [Azure Machine Learning 디자이너용 기계 학습 치트 시트](#) 를 살펴보세요.

7. 학습된 모델을 테스트하기 위해, 원래 데이터를 분할할 때 따로 분리해 두었던 유효성 검사 데이터 세트를 '채점'합니다. 즉, 유효성 검사 데이터 세트의 특징에 대해 레이블을 예측합니다. **모델 채점 및 평가** 섹션을 펼치고 **모델 채점** 모듈을 **모델 학습** 모듈 아래에 있는 캔버스로 끌어다 놓습니다. 그런 다음 **모델 학습** 모듈의 출력을 **모델 채점** 모듈의 **학습된 모델**(왼쪽) 입력에 연결하고 **데이터 분할** 모듈의 **결과 데이터 세트 2**(오른쪽) 출력을 **모델 채점** 모듈의 **데이터 세트**(오른쪽) 입력에 연결합니다.

8. 파이프라인이 다음과 같아야 합니다.



학습 파이프라인 실행

이제 학습 파이프라인을 실행하고 모델을 학습할 준비가 되었습니다.

1. **제출** 을 선택하고 **mslearn-diabetes-training** 이라는 기존 실험을 사용하여 파이프라인을 실행합니다.
2. 실험이 완료될 때까지 기다립니다. 5분 이상 걸릴 수 있습니다.
3. 실험이 완료되면 **모델 채점** 모듈을 선택하고 설정 창에서 **출력 + 로그** 탭의 **점수가 매겨진 데이터 세트** 섹션에 있는 **데이터 출력** 에서 **시각화** 아이콘을 사용하여 결과를 봅니다.
4. 오른쪽으로 스크롤하면 **Diabetic** 열(레이블의 알려진 실제 값이 포함됨) 옆에 예측 레이블 값을 포함하는 **Scored Labels**(점수가 매겨진 레이블)이라는 새 열이 있으며, 여기에는 예측 레이블 값과 0과 1 사이의 확률을 포함하는 **Scored Probabilities**(점수가 매겨진 확률) 열이 포함됩니다. 이는 '양성' 예측의 확률을 나타내므로 0.5보다 큰 확률은 예측 레이블이 1(당뇨)이 되고 0과 0.5 사이의 확률은 예측 레이블이 '0'(당뇨 아님)이 됩니다.*_
5. _ 모델 채점 결과 시각화* 창을 닫습니다.

모델에서 **Diabetic** 레이블의 값을 예측하지만 그 예측은 어느 정도 안정적일까요? 이를 평가하려면 모델을 평가해야 합니다.

다음 단원: 분류 모델 평가