100 XP

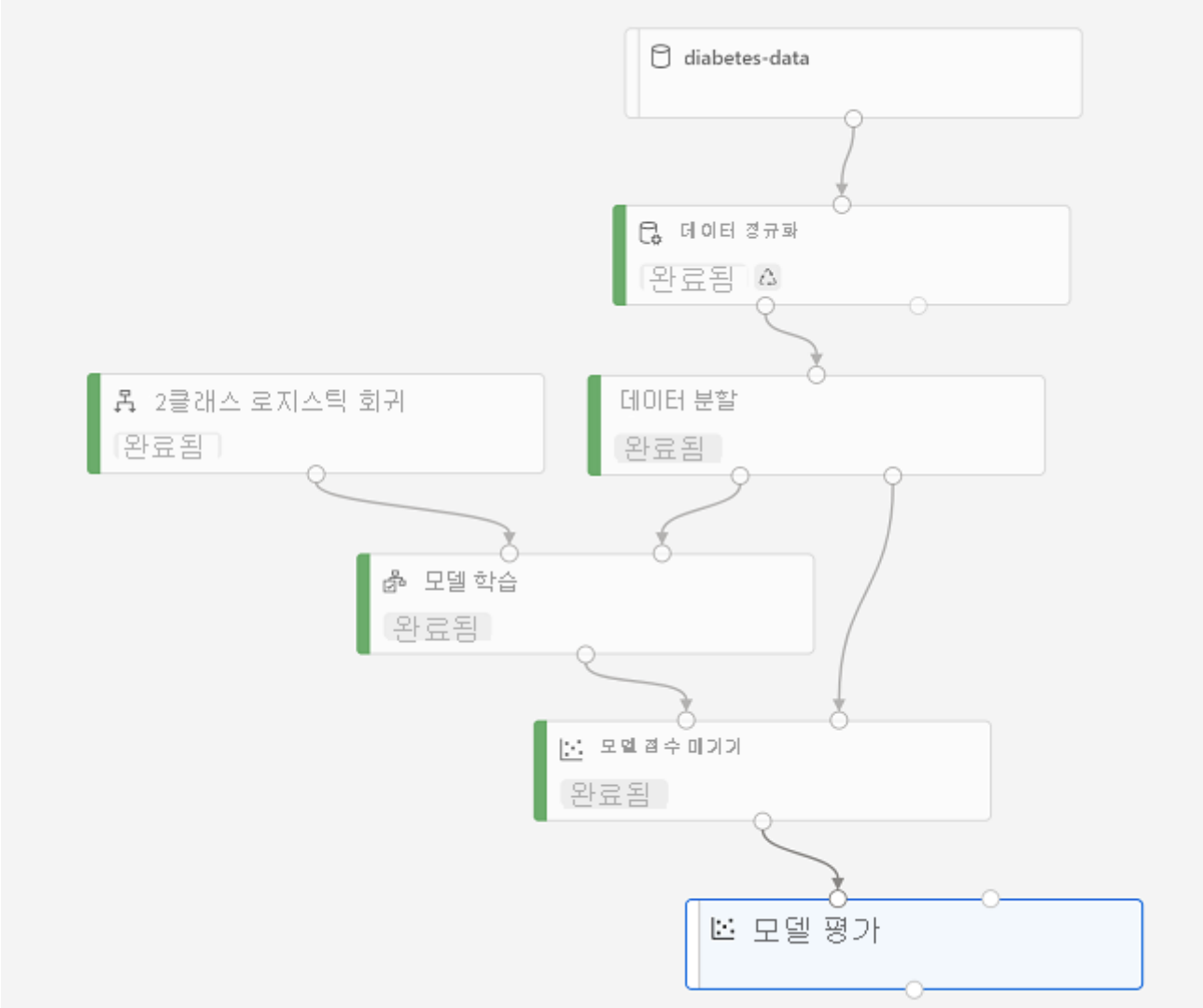
# 분류 모델 평가

10분

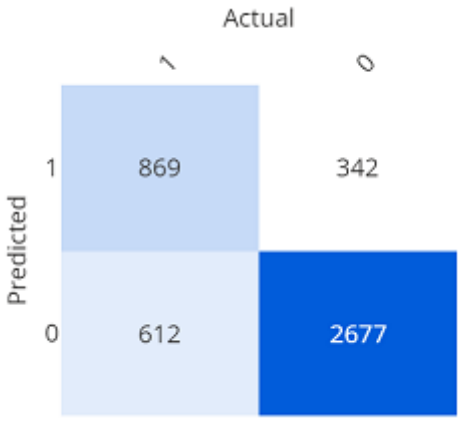
저장해서 모델 채점에 사용한 유효성 검사 데이터에는 레이블에 대해 알려진 값이 포함됩니다. 따라서 모델의 유효성을 검사하기 위해 유효성 검사 데이터 세트를 채점할 때 예측된 레이블 값과 레이블의 true 값을 비교할 수 있습니다. 이 비교에 따라 모델의 성과를 설명하는 다양한 메트릭을 계산할 수 있습니다.

## 모델 평가 모듈 추가

1. 아직 열려 있지 않다면 이전 단원에서 만든 **당뇨병 학습** 파이프라인을 엽니다.
2. 왼쪽 창의 **모델 채점 및 평가** 섹션에서 **모델 평가** 모듈을 **모델 채점** 모듈 아래에 있는 캔버스로 끌어다 놓고, **모델 채점** 모듈의 출력을 **모델 평가** 모듈의 **점수가 매겨진 데이터 세트**(왼쪽) 입력에 연결합니다.
3. 파이프라인이 다음과 같아야 합니다.



4. **제출** 을 선택하고 **mslearn-diabetes-training** 이라는 기존 실험을 사용하여 파이프라인을 실행합니다.
5. 실험이 완료될 때까지 기다립니다.
6. 실험이 완료되면 **모델 평가** 모듈을 선택하고 설정 창에서 **출력 + 로그** 탭의 **평가 결과** 섹션에 있는 **데이터 출력** 에서 **시각화** 아이콘을 사용하여 성능 메트릭을 봅니다. 이러한 메트릭은 데이터 과학자가 유효성 검사 데이터를 기준으로 모델이 얼마나 잘 예측하는지를 평가하는 데 도움이 될 수 있습니다.
7. 가능한 각 클래스에 대해 예측 값과 실제 값 수의 집계인 모델의 혼동 행렬을 확인합니다. 이와 같이 두 가지 가능한 값 중 하나를 예측하는 이진 분류 모델의 경우 혼동 행렬은 클래스 **0** 과 **1** 에 대한 예측 값과 실제 값을 보여 주는 2x2 표이며, 다음과 유사합니다.




혼동 행렬에는 예측 값과 실제 값이 모두 1(진양성이라고 함)인 경우가 왼쪽 상단에 표시되고 예측 값과 실제 값이 모두 0(진음성)인 사례가 오른쪽 하단에 표시됩니다. 다른 셀에는 예측 값과 실제 값이 서로 다른 경우(위양성 및 위음성)가 표시됩니다. 행렬의 셀에는 색이 지정되어 있으며 셀에 표시되는 사례가 많을 수록 색이 더 진해집니다. 결과에서는 왼쪽 위에서 오른쪽 아래로 색이 진한 셀을 대각선으로 찾아 보면서 모든 클래스에 대해 정확하게 예측하는 모델을 식별할 수 있습니다(즉, 예측 값이 실제 값과 일치하는 셀). 다중 클래스 분류 모델의 경우(가능한 클래스가 두 개 이상인 경우), 동일한 접근 방식이 가능한 각 실제 값 및 예측 값 수의 조합을 테이블화하는 데 사용됩니다. 따라서, 세 가지 클래스가 가능한 모델은 예측 및 실제 레이블 일치 셀이 대각선을 이루는 3x3 행렬이 만들어집니다.

8. 다음이 포함된 혼동 행렬의 왼쪽에서 메트릭을 검토합니다.

    - **정확도**: 올바른 예측(진양성 + 진음성)과 총 예측 수의 비율입니다. 즉, 모델이 당뇨병을 올바르게 예측하는 비율은 어느 정도입니까?
    - **정밀도**: 올바르게 식별된 양성 사례의 비율입니다(진양성 수를 진양성 수와 위양성 수의 합으로 나눈 비율). 즉, 모델에서 당뇨병이 있는 것으로 예측한 모든 환자 중에 실제로 당뇨가 있는 환자의 수는 얼마입니까?
    - **재현율**: 양성으로 분류된 사례 중 실제로 양성인 비율입니다(진양성 수를 진양성 수와 위음성 수의 합으로 나눈 비율). 즉, 실제로는 당뇨가 있는 모든 환자 중 모델이 식별할 수 있는 환자의 수는 얼마입니까?
    - **F1 점수**: 전체 메트릭은 기본적으로 정밀도와 재현율을 결합합니다.

- 나중에 **AUC** 로 돌아갑니다.

이러한 메트릭의 정확도는 가장 직관적입니다. 하지만, 모델의 작동 성과를 측정하는 것만큼 단순 정확도를 사용하는 데도 주의를 기울여야 합니다. 모집단의 3%만이 당뇨병이 있다고 가정합니다. 항상 **0** 을 예측하는 모델을 만들 수 있으며 이때의 정확도는 97%이지만 아주 유용하지는 않습니다. 이러한 이유로 대부분의 데이터 과학자는 정밀도와 재현율 같은 다른 메트릭을 사용하여 분류 모델의 성능을 평가합니다.

9. 메트릭 목록 위에 **임계값** 슬라이더가 있는지 확인합니다. 분류 모델에서 예측하는 것은 가능한 각 클래스의 확률입니다. 이 이진 분류 모델의 경우 양성(당뇨병 있음)으로 예측될 확률은 0과 1 사이에 있는 값입니다. 기본적으로 당뇨병 예측 확률이 0.5를 넘으면 클래스 예측은 1이 되고, 예측이 임계값보다 작으면 환자가 당뇨병이 **아닐** 확률이 더 크므로(클래스의 확률은 다 더하면 1이 됨) 예측 클래스는 0이 됩니다. 임계값 슬라이더를 이동하여 혼동 행렬에 미치는 영향을 확인하세요. 왼쪽(0)으로 이동하면 재현율 메트릭은 1이 되며, 오른쪽(1)으로 이동하면 재현율 메트릭이 0이 됩니다.

10. **ROC 곡선**(ROC는 수신된 연산자 특징(Received Operator Characteristic)을 의미하지만 대부분의 데이터 과학자는 ROC 곡선이라고 부름)의 임계값 슬라이드 위를 확인합니다. 재현율에 대한 또 다른 용어는 **진양성 비율** 이며, 여기에는 실제 음성 사례 수 대비 양성으로 잘못 식별된 음성 사례의 수를 측정하는 **위양성 비율** 이라는 메트릭이 따라옵니다. 이러한 메트릭을 0과 1 사이에서 가능한 모든 임계값에 대해 서로를 대조하면 곡선이 도출됩니다. 이상적인 모델에서는 곡선은 좌상향하는 방향으로 나가 차트의 전체 영역을 포함합니다. 곡선 아래 영역(0~1 범위의 값)이 클수록 모델의 성능이 뛰어나며, 이는 아래에 다른 메트릭과 함께 나열된 **AUC** 메트릭입니다. 이 영역이 모델의 성능을 어떻게 표시하는지 이해하려면 ROC 차트의 왼쪽 아래에서 오른쪽 위로 향하는 대각선을 생각해 보세요. 각 환자의 상태를 추측하거나 동전 던지기로 정하는 경우에 예상되는 결과를 나타냅니다. 반은 맞고 반은 틀릴 거라 예측할 수 있으므로 대각선 아래의 영역은 AUC 0.5를 나타냅니다. 모델에 대한 AUC가 이진 분류 모델의 AUC보다 높을 경우 모델은 임의의 추측보다 나은 결과를 도출합니다.

11. **모델 평가 결과 시각화** 창을 닫습니다.

기능 엔지니어링과 전처리를 최소한으로 했기 때문에 이 모델의 성능이 모두 좋은 것은 아닙니다. **2 클래스 의사 결정 포리스트** 와 같은 다른 분류 알고리즘을 사용하고 결과를 비교해 볼 수 있습니다. **데이터 분할** 모듈의 출력을 여러 **학습 모델** 과 **채점 모델** 모듈에 연결할 수 있으며 두 번째 **채점 모델** 모듈을 **평가 모델** 모듈에 연결하면 나란히 비교할 수도 있습니다. 이 연습에서는 완벽한 모델을 학습하는 것이 아니라 분류 및 Azure Machine Learning 디자이너 인터페이스를 소개하는 것으로 충분합니다.

---

## 다음 단원: 유추 파이프라인 만들기

계속 >