AI-900 자격 시험
Microsoft Azure AI Fundamentals

Azure Machine Learning

1. Azure Machine Learning

https://learn.microsoft.com/ko-kr/training/modules/fundamentals-machine-learning/

Azure Machine Learning

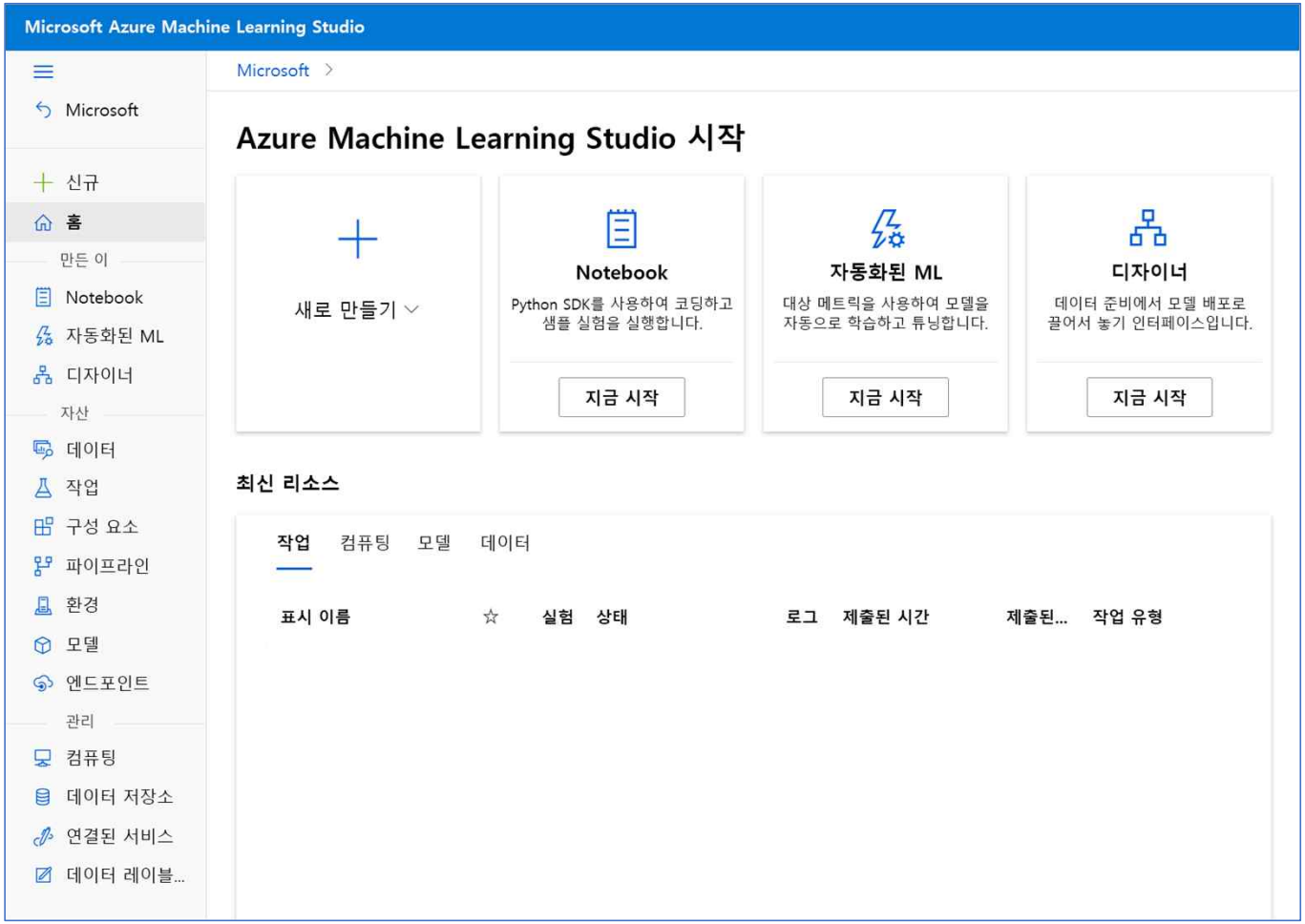
Machine Learning은 대부분의 인공 지능 솔루션의 토대가 되며, 지능형 솔루션을 만드는 과정은 기계 학습을 사용하여 수집한 이전 데이터를 사용하여 예측 모델을 학습시키는 것부터 시작합니다.

'Azure Machine Learning'은 기계 학습 모델을 학습하고 관리하는 데 사용할 수 있는 클라우드 서비스입니다.

Azure Machine Learning의 자동화된 Machine Learning 기능은 지도형 기계 학습 모델, 즉 학습 데이터가 알려진 레이블 값을 포함하는 모델을 지원합니다. 자동화된 Machine Learning을 사용하여 다음에 대한 모델을 학습시킬 수 있습니다.

- **분류**(범주 또는 클래스 예측)
- **회귀**(숫자 값 예측)
- **시계열 예측**(미래 시점의 숫자 값 예측)

Azure Machine Learning Studio



Azure Machine Learning Studio

Azure Machine Learning 컴퓨팅

핵심은 Azure Machine Learning이 기계 학습 모델을 학습시키고 관리하기 위한 서비스이므로 학습 프로세스를 실행하기 위해 컴퓨팅 리소스가 필요하다는 사실입니다. 컴퓨팅 대상은 모델 학습 및 데이터 탐색 프로세스를 실행할 수 있는 클라우드 기반 리소스입니다.

Azure Machine Learning 스튜디오 에서 데이터 과학 활동에 대한 컴퓨팅 대상을 관리할 수 있습니다. 다음 네 가지 종류의 컴퓨팅 리소스를 만들 수 있습니다.

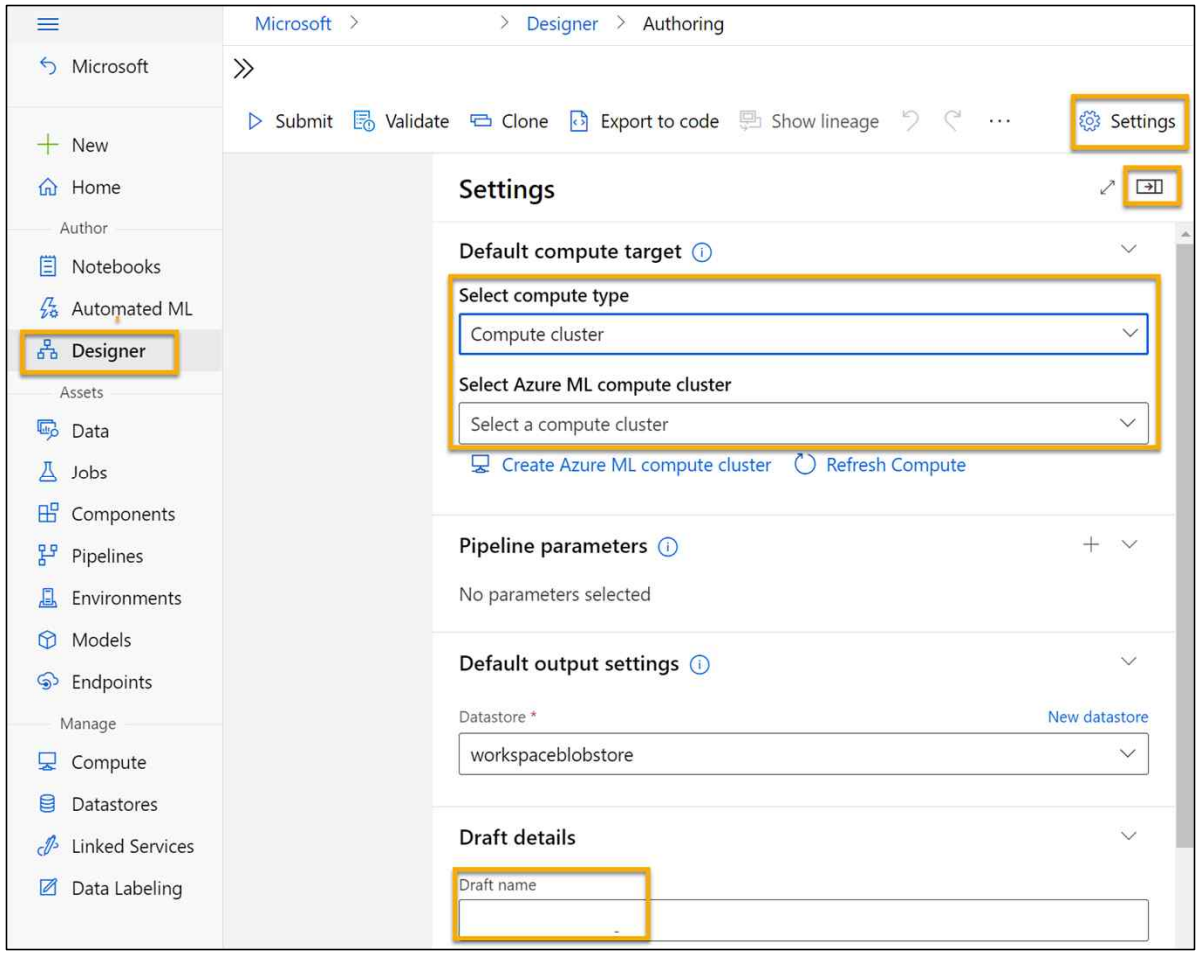
- **컴퓨팅 인스턴스**: 데이터 과학자가 데이터 및 모델을 작업하는 데 사용할 수 있는 개발 워크스테이션입니다.
- **컴퓨팅 클러스터**: 실험 코드의 주문형 처리를 지원하는 확장 가능한 가상 머신 클러스터입니다.
- **유추 클러스터**: 학습된 모델을 사용하는 예측 서비스의 배포 대상입니다.
- **연결된 컴퓨팅**: Virtual Machines 또는 Azure Databricks 클러스터와 같은 기존 Azure 컴퓨팅 리소스에 연결합니다.

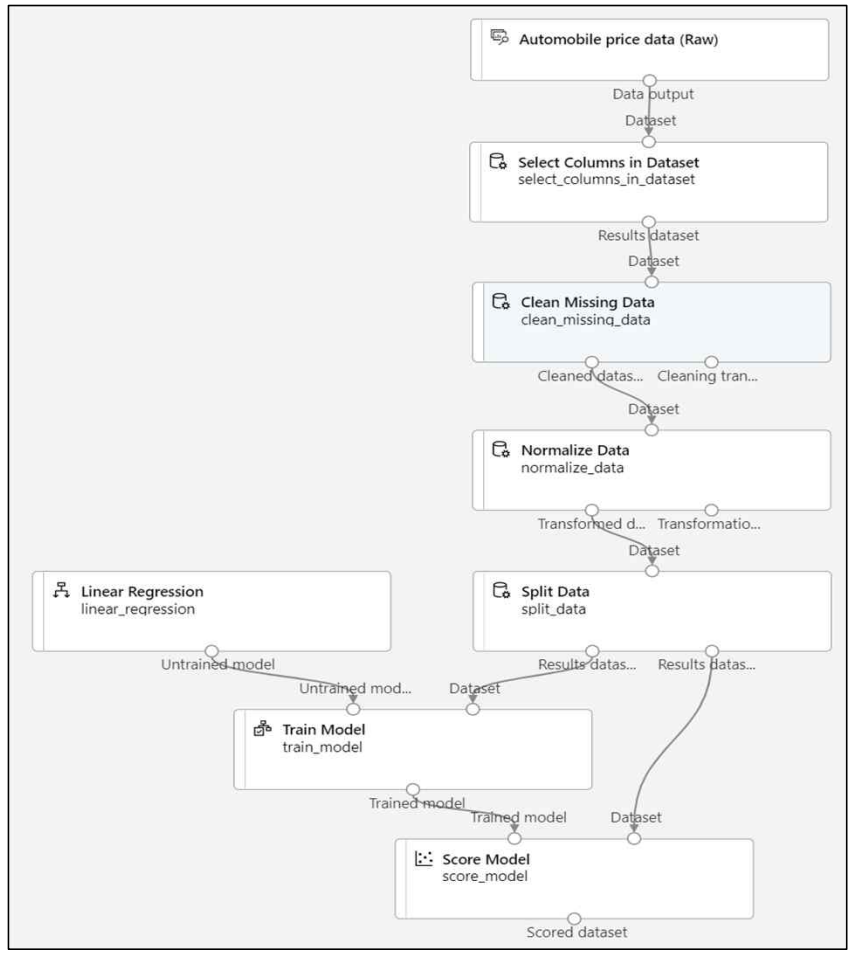
Azure Machine Learning

Azure Machine Learning 디자이너란 무엇인가요?

Azure Machine Learning 스튜디오에서 분류 기계 학습 모델을 작성하는 방법에는 여러 가지가 있습니다. 한 가지 방법은 기계 학습 모델을 학습, 테스트 및 배포하는 데 사용할 수 있는 디자이너라는 시각적 인터페이스를 사용하는 것입니다. 끌어서 놓기 인터페이스에서는 공유하고, 재사용하고, 버전을 제어할 수 있는 명확하게 정의된 입력 및 출력을 사용합니다.
파이프라인으로 알려진 각 디자이너 프로젝트에는 탐색을 위한 왼쪽 패널과 오른쪽의 캔버스가 있습니다. 디자이너를 사용하려면 모델에 필요한 빌딩 블록 또는 구성 요소를 식별하고 캔버스에 배치 및 연결하고 기계 학습 작업을 실행합니다.

Azure Machine Learning 디자이너





Azure Machine Learning : 데이터 전처리

데이터 전처리의 주요 작업

- **데이터 정리**: 누락된 값을 채우고, 시끄러운 데이터 및 이상치 값을 감지하고 제거합니다.
- **데이터 변환**: 데이터를 정규화하여 차원과 노이즈를 줄입니다.
- **데이터 감소**: 더 쉬운 데이터 처리를 위해 데이터 레코드 또는 속성을 샘플링합니다.
- **데이터 이산화** : 특정 기계 학습 방법에서 쉽게 사용할 수 있도록 연속 속성을 범주형 속성으로 변환합니다.
- **텍스트 정리**: 데이터 정렬 오류를 일으킬 수 있는 포함된 문자 (예: 탭으로 구분된 데이터 파일에 포함된 탭, 레코드를 손상시킬 수 있는 새 줄이 포함된 경우)를 제거합니다.

Azure Machine Learning : 피처 엔지니어링

Feature Engineering 이란?

훈련에 사용할 좋은 데이터(피처)들을 찾는 것이다.
에러, 이상치, 잡음으로 가득하면 결과가 좋지 않게 나오기 때문에 피처 엔지니어링이 필요하다. 기존의 피처를 사용해서 새로운 피처를 뽑아내는 작업을 피처 엔지니어링이라 부른다.

- Feature Selection(피처 선택) : 가지고 있는 피처 중에서 훈련에
  가장 유용한 피처들만 선택하는 방법이다
- Feature Extraction(피처 추출) : 피처를 결합하여 더 유용한
  피처를 만든다.

1. 회귀 모델

Azure Machine Learning : 회귀 모델

회귀는 원하는 결과를 예측하기 위해 변수 간의 관계를 이해하는 데 사용되는 기계 학습의 한 형태입니다. 회귀는 숫자 레이블, 변수에 따른 결과 또는 기능을 예측합니다. 예를 들어 자동차 판매 회사는 자동차의 특성(예: 엔진 크기, 좌석 수, 주행 거리)을 사용하여 가능한 판매 가격을 예측할 수 있습니다. 이 경우 자동차의 특성이 특징이며 판매 가격은 레이블입니다.

회귀는 모델이 학습을 통해 특징 조합을 레이블에 '맞추도록' 레이블에 대한 특징과 알려진 값을 모두 포함하는 데이터를 사용하여 모델을 학습하는 '지도(supervised)' 기계 학습 기법의 예입니다. 그러고 나서, 학습이 완료된 후 학습된 모델을 사용하여 레이블이 알려지지 않은 새 항목에 대한 레이블을 예측할 수 있습니다.

회귀 기계 학습 모델에 대한 시나리오

회귀 기계 학습 모델은 많은 산업에서 사용됩니다. 몇 가지 시나리오는 다음과 같습니다.

- 평방 피트와 방의 수와 같은 주택의 특성을 사용하여 주택 가격을 예측합니다.
- 날씨 및 토양 품질과 같은 농장 조건의 특성을 사용하여 작물 수확량을 예측합니다.
- 광고 로그와 같은 과거 캠페인의 특성을 사용하여 향후 광고 클릭 수를 예측합니다.

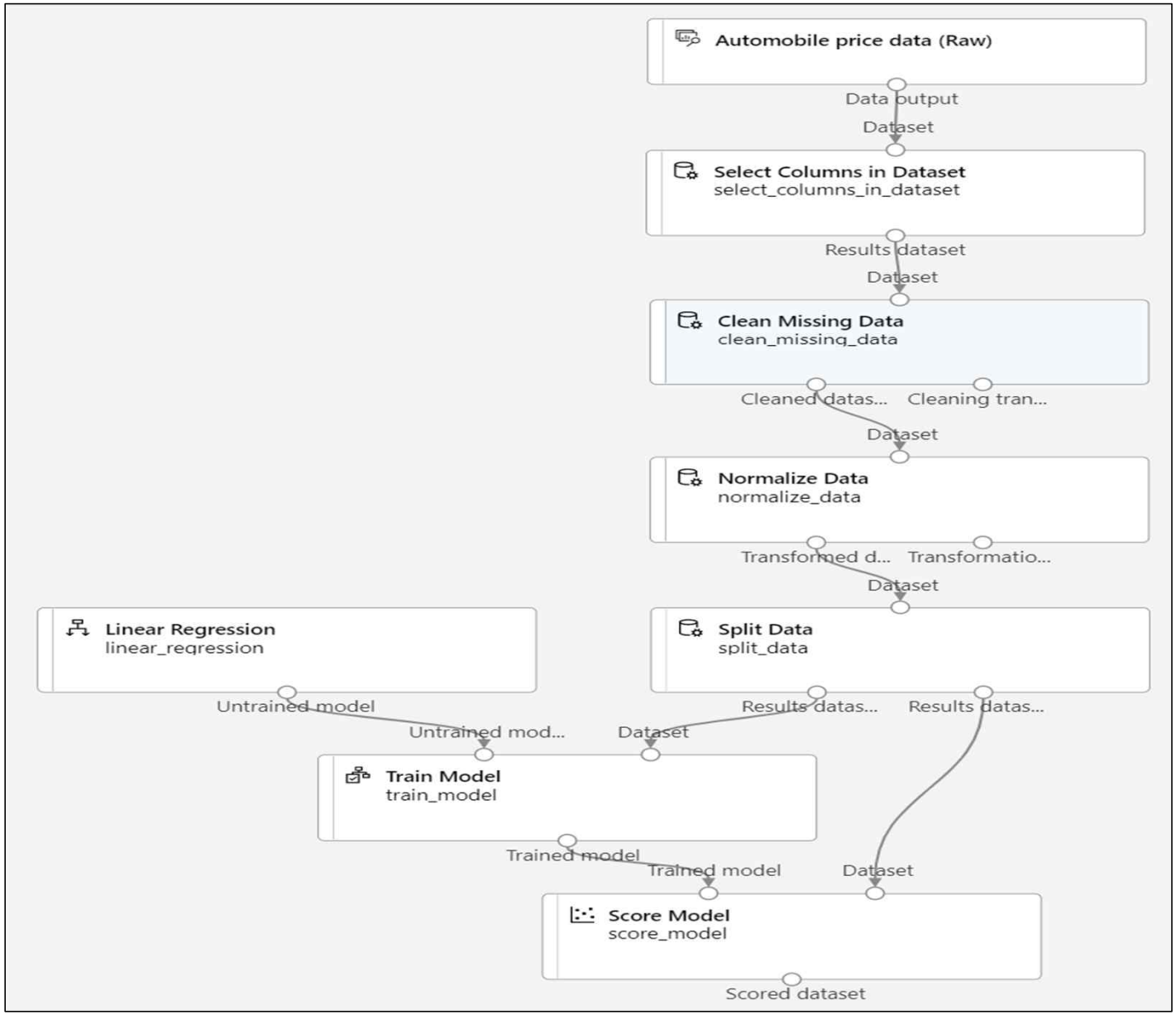
Azure Machine Learning : **회귀 모델 구현**

회귀 기계 학습 모델을 학습하고 평가하는 단계를 다음과 같이 생각할 수 있다.

1.데이터 준비 : 데이터 집합의 기능 및 레이블을 식별합니다. 필요에 따라 데이터를 사전 처리 또는 정리 및 변환한다.
2.모델 학습 : 데이터를 학습 및 유효성 검사 집합이라는 두 그룹으로 분할합니다. 학습 데이터 집합을 사용하여 기계 학습 모델을 학습시킨다. 유효성 검사 데이터 집합을 사용하여 기계 학습 모델에서 성능을 테스트한다.
3.성능 평가 : 모델의 예측이 알려진 레이블과 얼마나 가까운지 비교한다.
4.예측 서비스 배포 : 기계 학습 모델을 학습한 후 학습 파이프라인을 실시간 유추 파이프라인으로 변환해야 한다. 그런 다음, 다른 사용자가 사용할 수 있도록 모델을 서버 또는 디바이스에 애플리케이션으로 배포할 수 있다.

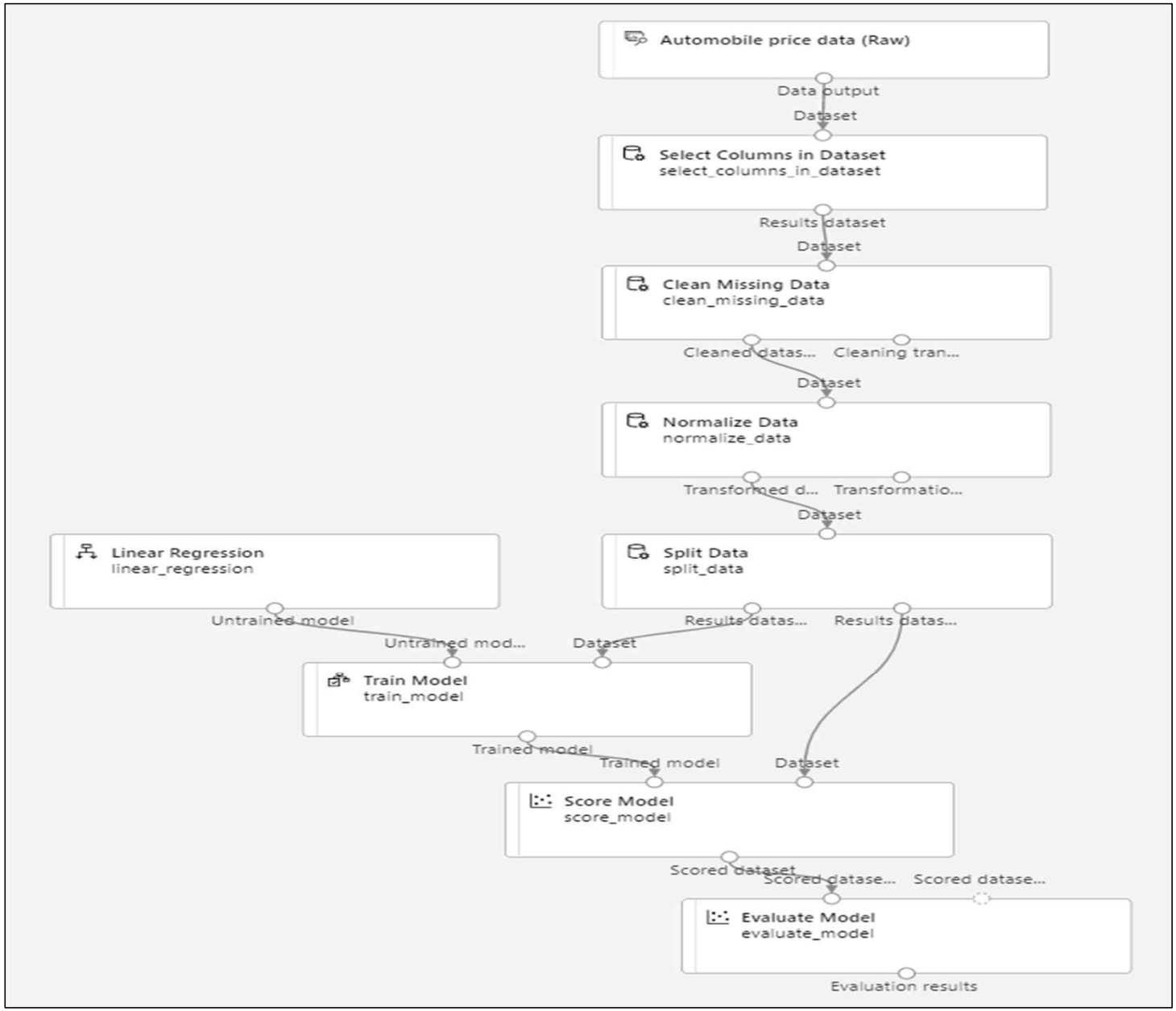
Azure Machine Learning 디자이너에서 파이프라인 만들기 **(회귀모델 : 학습)**

https://microsoftlearning.github.io/AI-900-AIFundamentals/instructions/02a-create-regression-model.html



Azure Machine Learning 디자이너에서 파이프라인 만들기 **(회귀모델 : 평가)**

- 평균 절대 오차(MAE)
- 루트 평균 제곱 오차(RMSE)
- 상대 제곱 오차(RSE)
- 상대적 절대 오류(RAE)
- 결정 계수 (R^2)

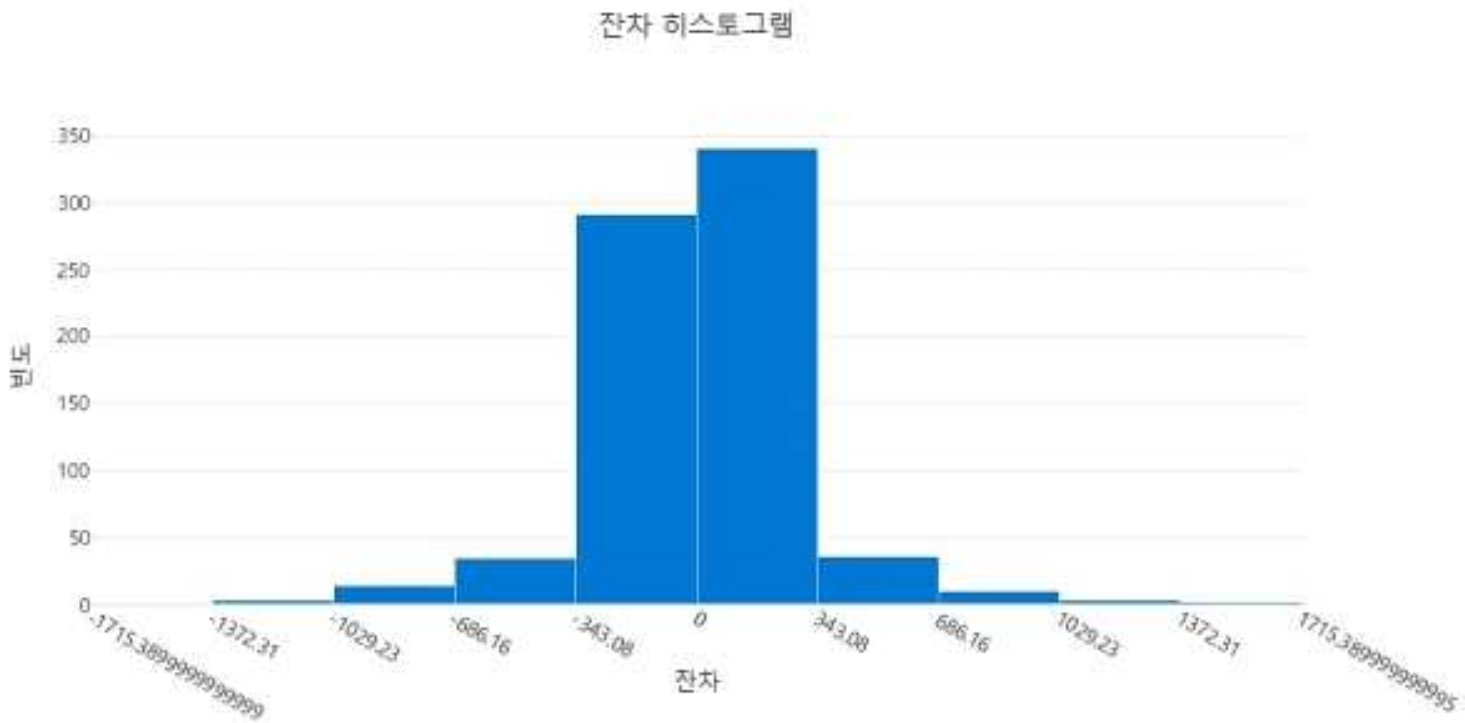


Azure Machine Learning : 회귀 모델 성능 평가 지표

- **MAE(평균 절대 오차)**: 예측 값과 실제 값 사이의 평균 차이입니다. 해당 값은 레이블과 동일한 단위(이 경우 달러)를 기준으로 합니다. 해당 값이 낮을수록 모델의 예측 정확도가 높아집니다.
- **RMSE(평균 제곱 오차)**: 예측 값과 실제 값 사이의 평균 제곱 차이의 제곱근입니다. 결과는 레이블과 동일한 단위(달러)를 기준으로 하는 메트릭입니다. MAE(위)보다 차이가 크면 개별 오차의 분산이 크다는 것을 나타냅니다(예: 일부 오차는 매우 작고 다른 오차는 큰 경우).
- **RSE(상대 제곱 오차)**: 예측 값과 실제 값 간의 차이에 대한 제곱을 기준으로 한, 0에서 1사이의 상대 메트릭입니다. 해당 메트릭이 0에 가까울수록 모델의 성능이 더 뛰어납니다. 해당 메트릭은 상대적이므로 레이블이 다른 단위인 모델을 비교하는 데 사용할 수 있습니다.
- **RAE(상대 절대 오차)**: 예측 값과 실제 값 간의 절대 차이에 대한 제곱을 기준으로 한, 0에서 1사이의 상대 메트릭입니다. 해당 메트릭이 0에 가까울수록 모델의 성능이 더 뛰어납니다. RSE와 마찬가지로 해당 메트릭은 레이블의 단위가 서로 다른 모델을 비교하는 데 사용할 수 있습니다.
- **결정 계수(R^2)**: 해당 메트릭은 일반적으로 '결정 계수'라고 하며 모델이 설명하는 예측 값과 실제 값 간의 분산이 어느 정도인지 개략적으로 알려 줍니다. 해당 값이 1에 가까울수록 모델의 성능이 더 뛰어납니다.
  https://www.codingprof.com/3-ways-to-calculate-the-root-relative-squared-error-rrse-in-r/

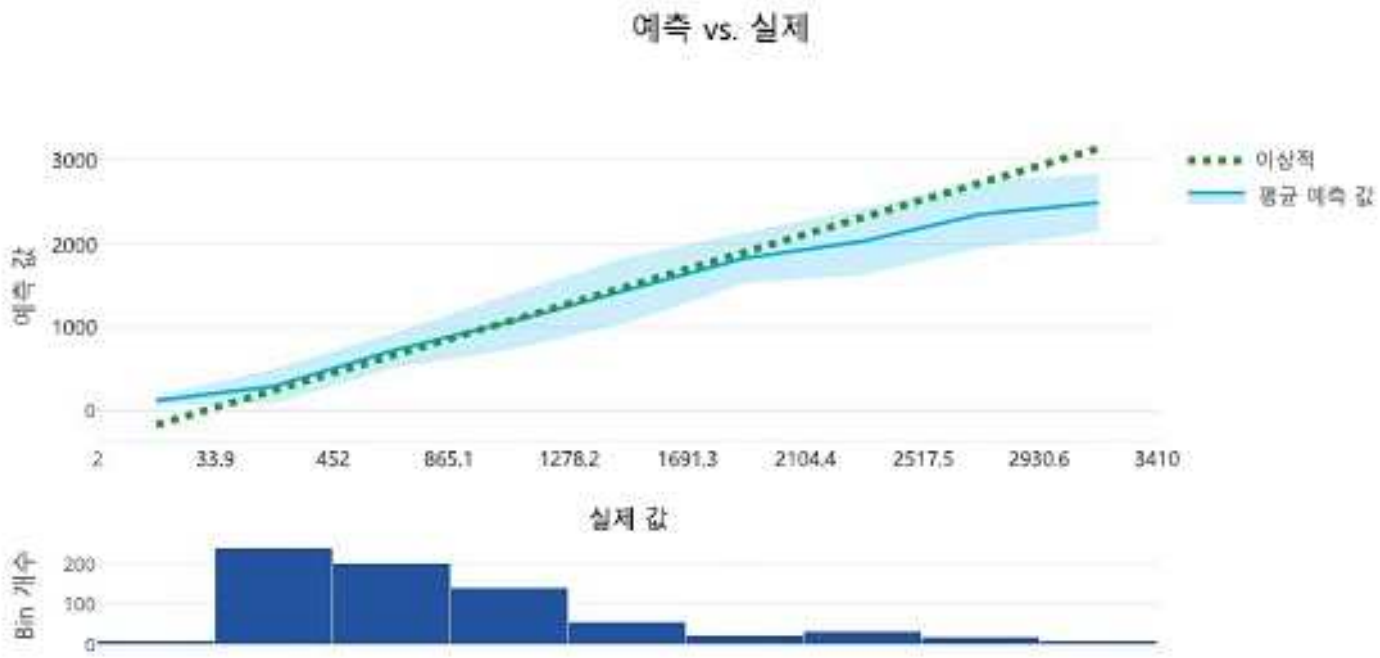
Azure Machine Learning : 회귀 모델 성능 평가

잔차 히스토그램은 잔차 값 범위의 빈도를 보여 줍니다. 오차는 모델에서 설명할 수 없는 예측 값과 실제 값의 차이, 즉 오류를 나타냅니다. 가장 자주 발생하는 오차 값이 0 주위에 클러스터링되어야 합니다. 오류가 적고, 특히 양 극단의 오류는 거의 없는 것이 좋습니다.



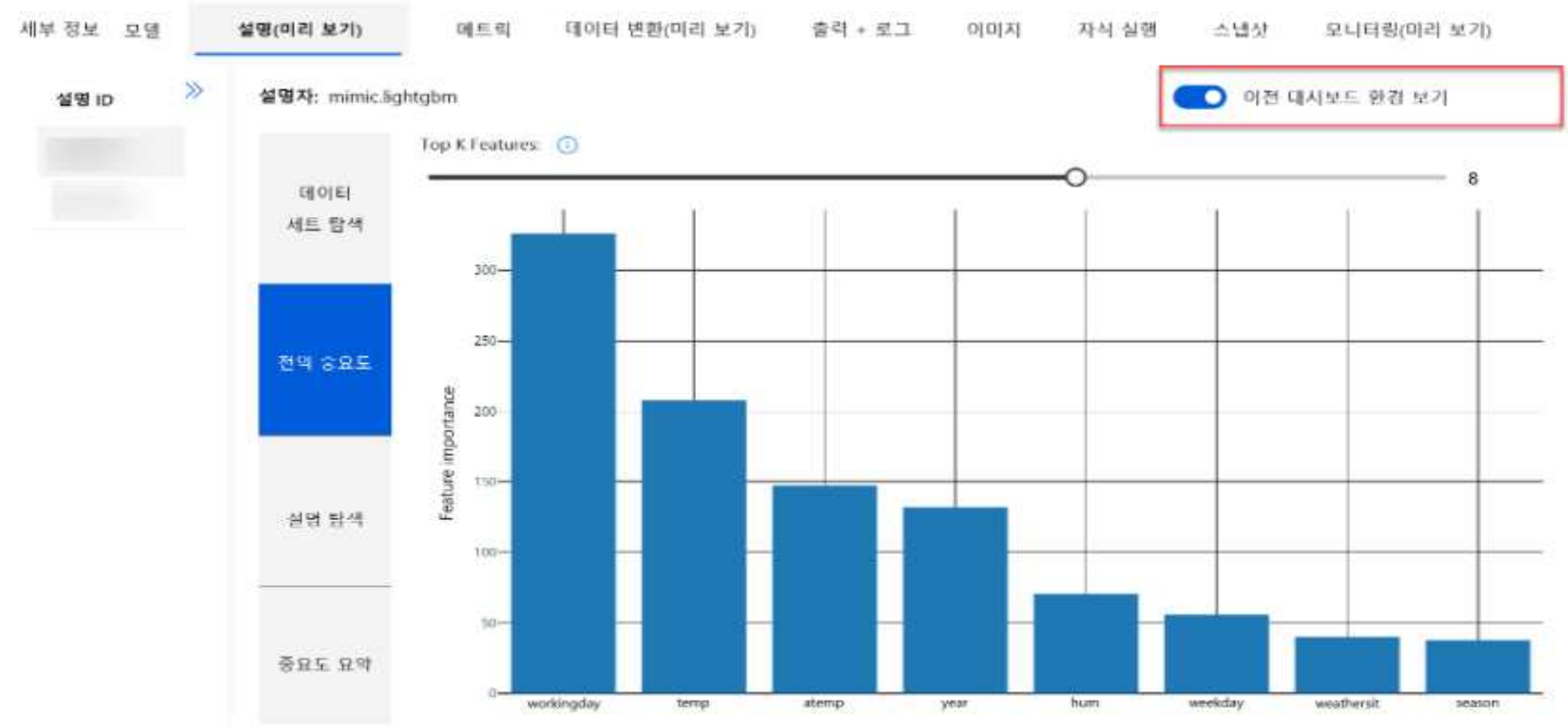
Azure Machine Learning : 회귀 모델 성능 평가 지표

예측 값 및 참 값 차트는 예측 값이 참 값과 밀접하게 상호 연관되는 대각선 추세를 표시합니다. 점선은 완벽한 모델의 성능을 보여 줍니다. 모델의 평균 예측 값 선이 점선에 가까울수록 성능이 더 높은 것입니다. 꺾은선형 차트 아래의 히스토그램은 참 값의 분포를 보여 줍니다.



Azure Machine Learning : 회귀 모델 성능 평가 지표

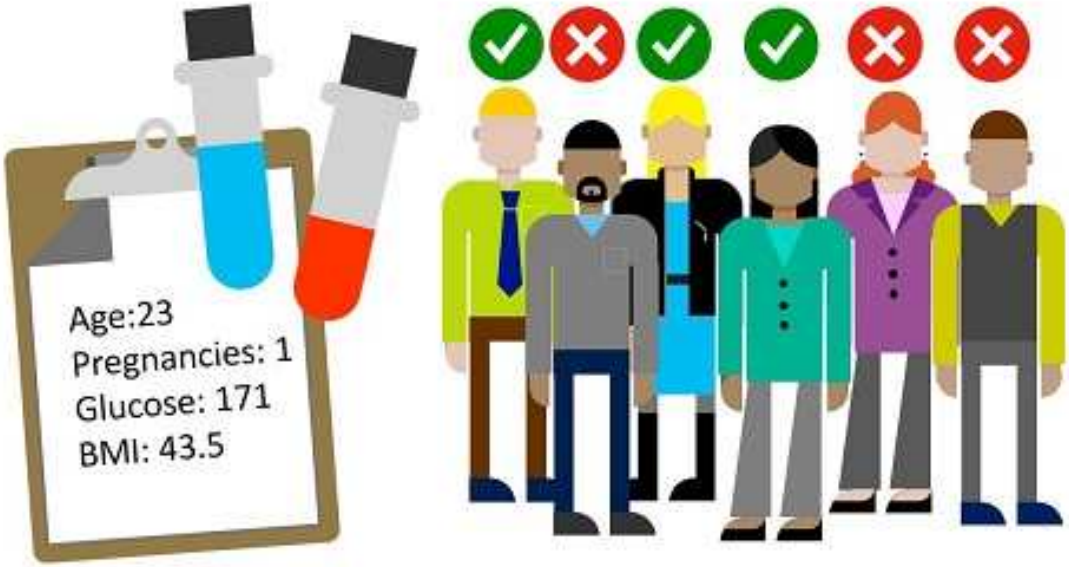
설명 탭을 선택합니다. 설명 ID를 선택한 다음, **집계 기능 중요도**를 선택합니다. 이 차트는 다음과 같이 데이터 세트의 각 기능이 레이블 예측에 영향을 미치는 정도를 보여 줍니다.



2. 분류 모델

Azure Machine Learning : 분류 모델

분류는 항목이 속한 범주 또는 클래스를 예측하는 데 사용되는 기계 학습의 한 형태입니다. 예를 들면, 진료소는 환자가 당뇨병이 걸릴 위험이 있는지를 예측하기 위하여 환자의 특성 (나이, 체중, 혈압 등)을 이용할 수 있습니다. 이 경우 환자의 특성이 특징이며, 레이블은 각각 당뇨병이 없거나 있는 환자를 나타내는 **0** 또는 **1**이라는 분류입니다.



회귀와 마찬가지로, 분류는 모델이 학습을 통해 특징 조합을 레이블에 맞추도록 레이블에 대한 특징과 알려진 값을 모두 포함하는 데이터를 사용하여 모델을 학습하는 감독형 기계 학습 기술의 예입니다. 그러고 나서, 학습이 완료된 후 학습된 모델을 사용하여 레이블이 알려지지 않은 새 항목에 대한 레이블을 예측할 수 있습니다.

Azure Machine Learning : 분류 모델

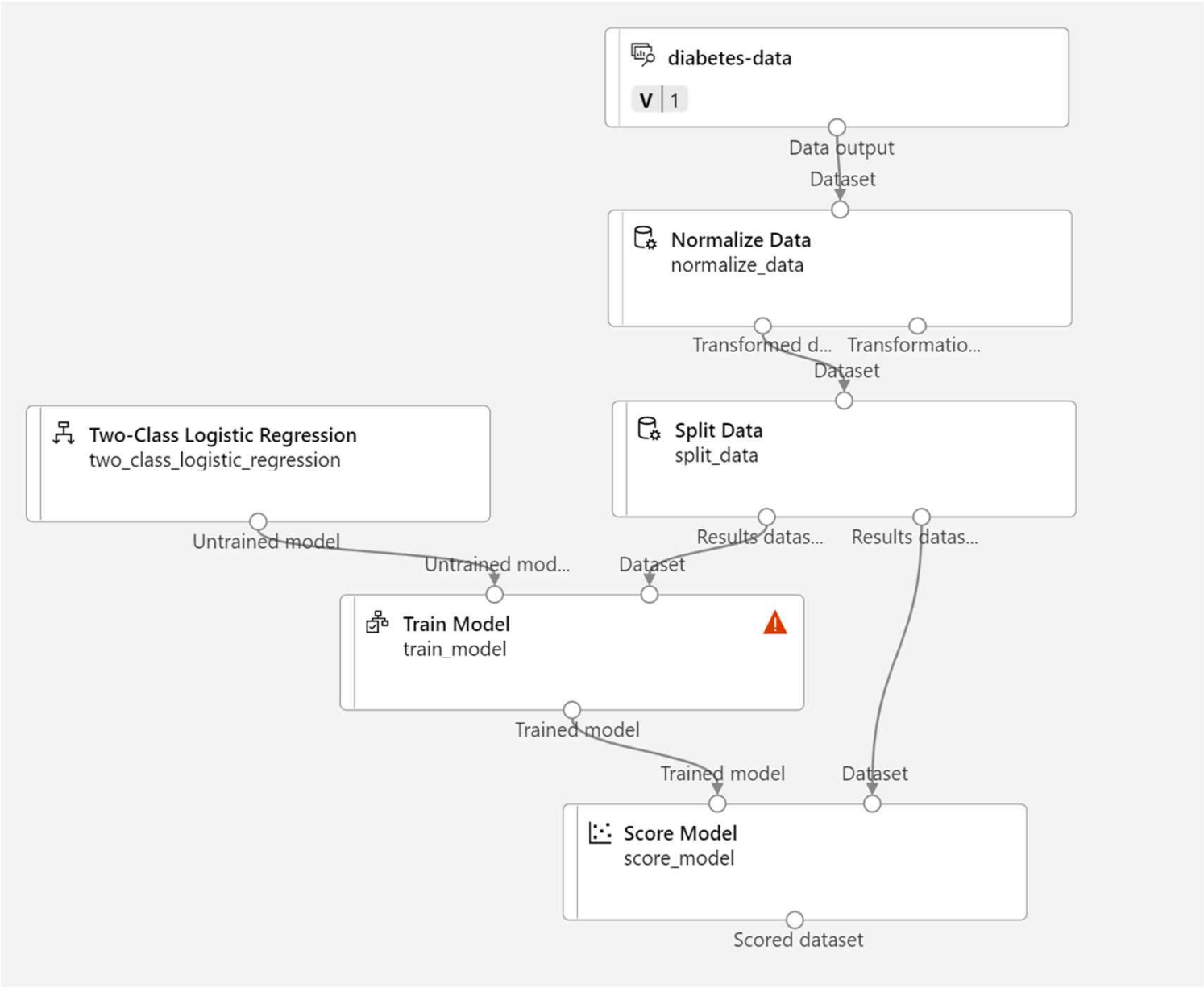
기계 학습 모델 분류 시나리오

분류 기계 학습 모델은 많은 산업에서 사용됩니다. 몇 가지 시나리오는 다음과 같습니다.

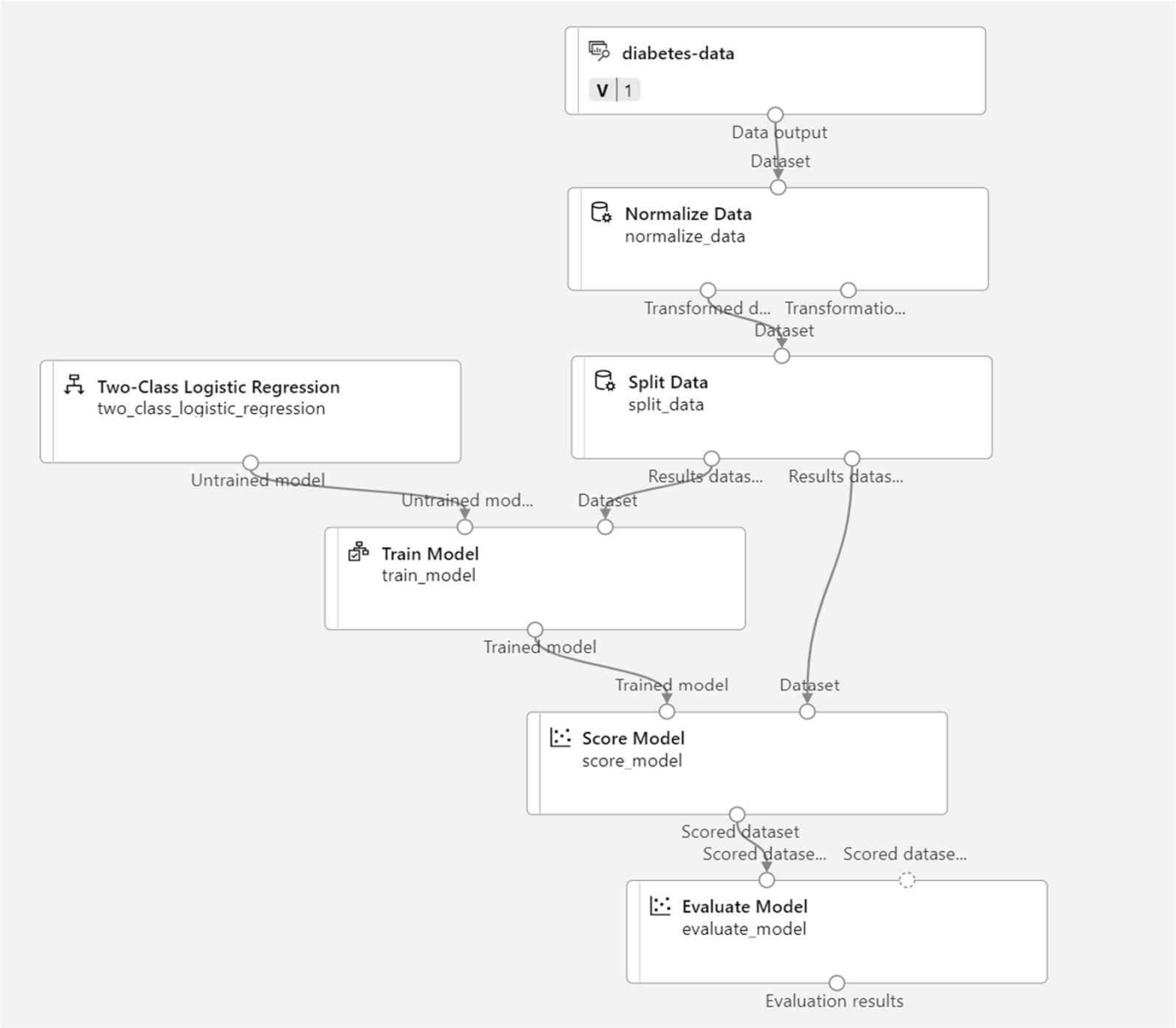
- 임상 데이터를 사용하여 환자가 아플지 여부를 예측합니다.
- 기록 데이터를 사용하여 은행이 대출을 제공해야 하는지 여부를 예측합니다.
- 중소기업의 특성을 사용하여 새로운 벤처가 성공할지 예측합니다.

Azure Machine Learning 디자이너에서 파이프라인 만들기 **(분류 모델 : 학습)**

https://microsoftlearning.github.io/AI-900-AIFundamentals/instructions/02b-create-classification-model.html



Azure Machine Learning 디자이너에서 파이프라인 만들기 **(분류 모델 : 평가)**



Azure Machine Learning : 혼동 행렬(Confusion Matrix)

		실제	
		Positive (1)	Negative (0)
예측	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Azure Machine Learning : 혼동 행렬(Confusion Matrix)

- 참 양성 (True Positive): 모델이 올바르게 심장 질환이 있음을 예측했을 경우
- 거짓 양성 (False Positive): 모델은 심장 질환이 있음을 예측했는데 실제 관측은 심장 질환이 없을 경우
- 거짓 음성 (False Negative): 모델은 심장 질환이 없음을 예측했는데 실제 관측은 심장 질환이 있을 경우
- 참 음성 (True Negative): 모델이 올바르게 심장 질환이 없음을 예측했을 경우

Azure Machine Learning : 분류 모델 평가 지표

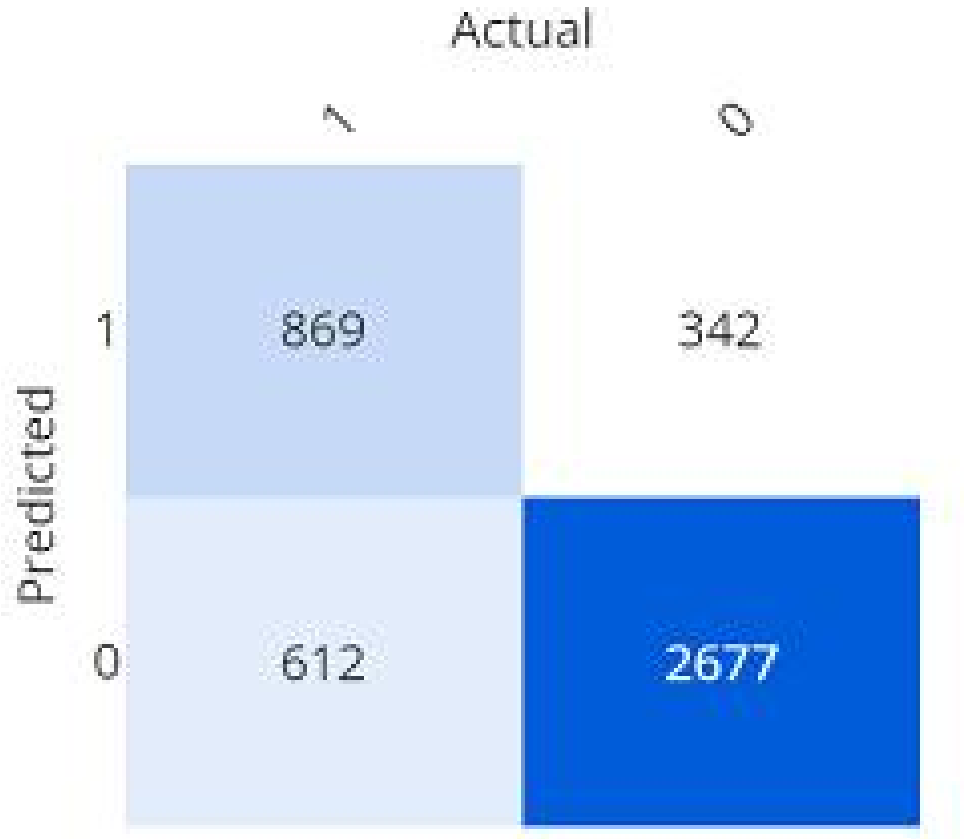
	설명	산출식
Accuracy (정확도)	맞게 검출한 비율	(TP+TN)/ (TP+TN+FP+FN)
Precision (정밀도)	P로 예측한 것 중 실제 P의 비율	TP/ (TP+FP)
Recall (진양성율)	실제 P를 P로 예측	TP/ (TP+FN)
False Alarm (위양성율)	실제 N을 P로 예측	FP/ (FP+TN)

https://truman.tistory.com/179

Azure Machine Learning : 혼동 행렬(Confusion Matrix)

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.

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 그리드이며, 우측 그림과 유사합니다.



Azure Machine Learning : 분류 모델 평가 지표

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.

메트릭은 다음과 같은 혼동 행렬에서 파생될 수 있습니다.

- **정확도**: 올바른 예측(진양성 + 진음성)과 총 예측 수의 비율입니다.
- **정밀도**: 올바르게 식별된 양성 사례의 비율입니다(진양성 수를 진양성 수와 위양성 수의 합으로 나눈 비율).
- **재현율**: 양성으로 분류된 사례 중 실제로 양성인 비율입니다(진양성 수를 진양성 수와 위음성 수의 합으로 나눈 비율).
- **F1 점수**: 전체 메트릭은 기본적으로 정밀도와 재현율을 결합합니다.

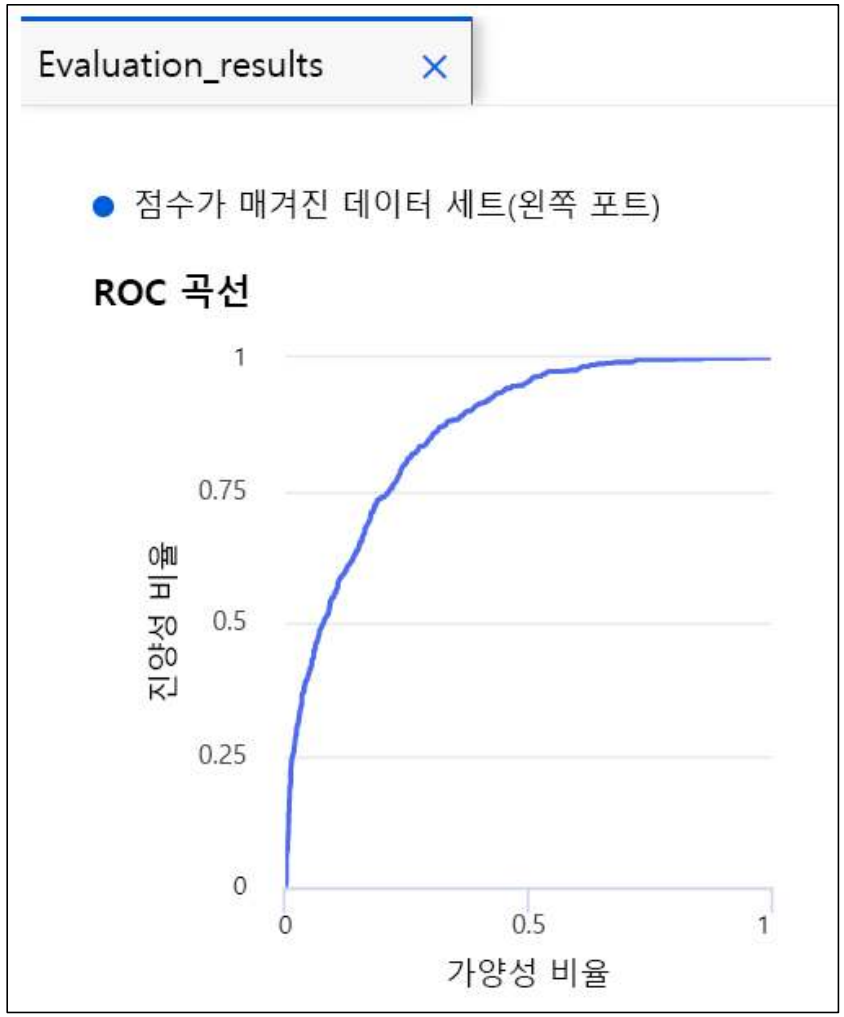
이러한 메트릭의 정확도는 가장 직관적입니다. 하지만 모델의 작동 성능을 측정하는 것만큼 정확도를 사용하는 데도 주의를 기울여야 합니다. 모집단의 3%만이 당뇨병이 있다고 가정합니다. 항상 **0**을 예측하여 97% 정확도로 모델을 만들 수 있지만 당뇨병 사례를 올바르게 예측하는 데는 도움이 되지 않습니다. 이러한 이유로 대부분의 데이터 과학자는 정밀도와 재현율 같은 다른 메트릭을 사용하여 분류 모델의 성능을 평가합니다.

https://truman.tistory.com/240

Azure Machine Learning : ROC / AUC

ROC 곡선 및 AUC 메트릭

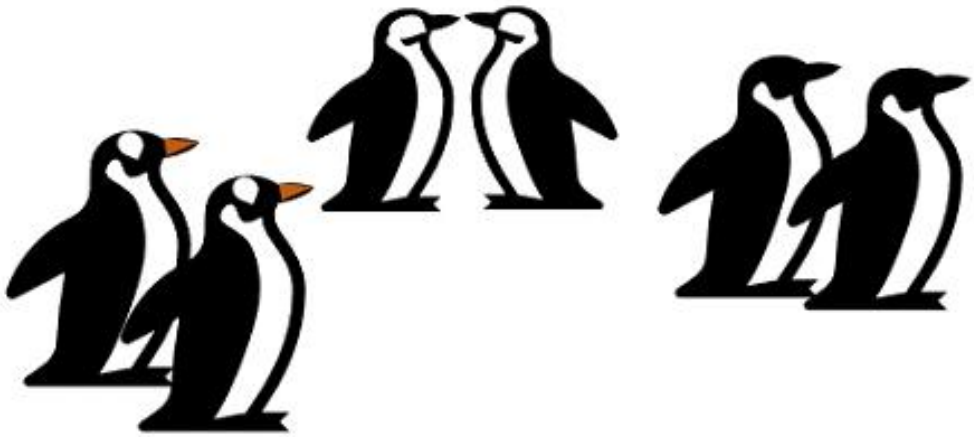
재현율에 대한 또 다른 용어는 **진양성 비율**이며, 여기에는 실제 음성 사례 수 대비 양성으로 잘못 식별된 음성 사례의 수를 측정하는 **위양성 비율**이라는 메트릭이 따라온다. 0과 1 사이의 가능한 모든 임계값에 대해 이러한 메트릭을 그리면 **ROC 곡선**이라고 하는 곡선이 생성된다 (ROC는 *Receiver Operating Characteristic*을 의미하지만 대부분의 데이터 과학자는 이를 ROC 곡선이라고 부름). 이상적인 모델에서는 곡선은 좌상향하는 방향으로 나가 차트의 전체 영역을 포함한다. **AUC** 메트릭에 해당하는 *곡선 아래 영역* (0~1 범위의 값)이 클수록(**1에 가까울수록**) 모델의 성능이 뛰어나다. **평가 결과**에서 ROC 곡선을 검토할 수 있다.



3. 클러스터링 모델

Azure Machine Learning : 클러스터링 모델

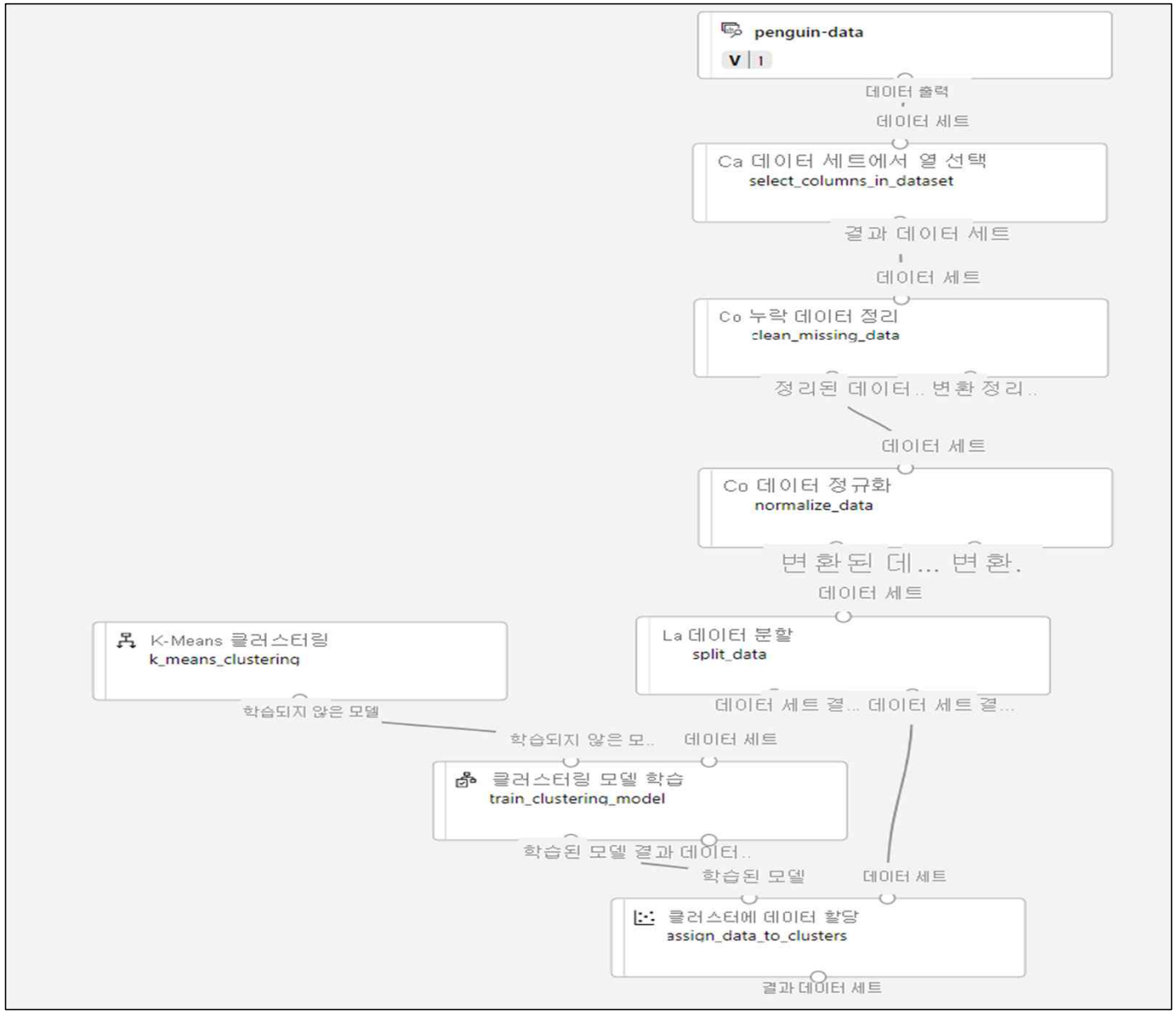
클러스터링(clustering)은 유사한 항목을 특성에 따라 클러스터로 그룹화하는 데 사용되는 기계 학습의 한 형태입니다. 예를 들어 연구원은 펭귄을 측정하고 비율의 유사성에 따라 그룹화할 수 있습니다.



클러스터링은 감독되지 않은 기계 학습의 한 예로, 해당 특성 또는 기능에 따라 항목을 클러스터로 분리하는 모델을 학습합니다. 모델을 학습하는 이전에 알려진 클러스터 값(또는 레이블)은 없습니다.

Microsoft Azure Machine Learning 디자이너를 사용하여 코드를 작성할 필요 없이 끌어다 놓기 시각적 인터페이스를 사용하여 클러스터링 모델을 만들 수 있습니다.

Azure Machine Learning : 클러스터링 모델 (학습)



Azure Machine Learning : 클러스터링 모델

K-평균 알고리즘은 항목을 지정한 클러스터의 수, 즉 *K* 값으로 그룹화합니다. **K-평균 클러스터링** 모듈을 선택해 오른쪽 창에서 **중심 수** 매개 변수를 **3**으로 설정합니다.

① **참고**

펭귄 측정값과 같은 데이터 관찰을 다차원 벡터라고 생각할 수 있습니다. K-평균은 다음과 같은 방식으로 작동합니다.

a. *K* 좌표를 *n*차원 공간의 중심이라는 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.

Azure Machine Learning : 클러스터링 모델

1. Azure Machine Learning 디자이너 파이프라인을 사용하여 K-평균 클러스터링 모델을 학습 및 테스트합니다. 모델에서 3개의 클러스터 중 하나에 항목을 할당하고자 합니다. 이를 위해 K-평균 클러스터링 모듈의 어떤 구성 속성을 설정해야 하나요?

- ◉ 중심의 수를 3으로 설정
  - ✔ **정답입니다. K 클러스터를 만들려면 중심의 수를 K로 설정해야 합니다.**
- ○ 난수 초기값을 3으로 설정
- ○ 반복을 3으로 설정

2. Azure Machine Learning 디자이너를 사용하여 클러스터링 모델에 대한 학습 파이프라인을 만듭니다. 이제 추론 파이프라인에서 모델을 사용하고자 합니다. 모델에서 클러스터 예측을 추론하기 위해 사용해야 하는 모듈은 무엇인가요?

- ○ 모델 채점
- ◉ 클러스터에 데이터 할당
  - ✔ **정답입니다. 클러스터에 데이터 할당 모듈을 사용하여 학습된 클러스터링 모델에서 클러스터 예측을 생성합니다.**
- ○ 클러스터링 모델 학습

감사합니다