# <오전>

## [통계 이론]

### [통계의 기본 개념]

#### 개요

통계학은 데이터를 수집, 분석, 해석 및 표현하는 과학으로서, 이를 통해 불확실성을 다루고 데이터 기반 의사 결정을 지원하는 학문이다.

목적

* 데이터로부터 유의미한 정보 도출
* 현상 및 패턴 이해
* 연관성 파악
* 예측 및 추론
* 불확실성의 해소
* 의사결정 지원

중요성

* 데이터를 수치화하여 신뢰성을 부여함
* 의사 결정을 위한 근거 자료를 제시함
* 현상을 분석하여 실증 자료를 제시함

#### 표본의 추출

**• 확률적 표본 추출 (Probability sampling method)**

• 동일한 확률 하에서 표본을 추출하는 방법

• 무작위 표본 추출

• 난수표 등을 따라 모집단에서 표본을 기계적으로 추출하는 방법

• 체계적 표본 추출

• 모집단에서 특정한 규칙으로 표본을 추출하는 방법

• 층화 표본 추출

• 모집단을 특정 특성에 따라 여러 하위 집단으로 구분한 후, 집단의 규모에 비례하도록 추출하는 방법

**• 비확률적 표본 추출 (Non-probability sampling method)**

• 조사자가 자의로 표본을 추출하거나 조사 대상이 자발적으로 표본에 참여하는 방법

### [기술 통계]

#### 중심경향성 측정

• **중심경향성(Central Tendency)** :

• 표본 내의 원소들의 중심을 나타내는 지표.

• 평균, 중앙값, 최빈값 등 다양한 방식으로 표현됨

**• 평균(Mean) : 모든 값을 더한 후 데이터의 개수로 나눈 값**

• 장점: 데이터를 통합하여 하나의 대표값으로 표현.

• 단점: 이상치(outliers)에 민감하여 극단적인 값에 영향을 받음.

• 연속적이고 정규 분포를 따르는 데이터에서 대표값을 구할 때 유용

• 예시: {1, 2, 3, 4, 5}의 평균은 (1+2+3+4+5) / 5 = 3.

• 중앙값 (Median) : 표본 내의 원소들을 크기 순서대로 나열했을 때 중앙에 위치한 값

• 장점: 이상치의 영향을 덜 받음.

• 단점: 표본의 크기가 클 경우 정렬에 많은 시간 소요.

• 예시: {1, 2, 3, 4, 5}의 중앙값은 3. (짝수 개의 데이터인 경우, 두 중앙값의 평균을 사용)

• 비대칭적이거나 이상치가 있을 때 사용

• 최빈값 (Mode) : 표본에서 가장 자주 나타나는 값.

• 장점: 표본의 원소들의 빈도 분포를 잘 나타냄.

• 단점: 표본 내에 최빈값이 존재하지 않을 수 있으며, 복수 개의 최빈값이 있을 수도 있음.

• 예시: {1, 2, 2, 3, 4}의 최빈값은 2.

• 범주형 데이터에서 빈도가 가장 높은 값을 찾을 때 유용

• 가중평균 (Weighted Mean) : 표본의 원소들의 값에 가중치를 부여하여 계산한 평균.

• 모든 원소들의 값이 동일한 중요도를 가지지 않을 때 사용.

• 예시: {1, 2, 3}와 각각의 가중치 {0.1, 0.3, 0.6}이 있을 때

가중평균은 (1\*0.1 + 2\*0.3 + 3\*0.6) / (0.1 + 0.3 + 0.6)

• 예시: 5천원짜리 상품 8개와 9천원짜리 상품 2개를 구입했을 때

평균 구매값은 (5000 \* 8 + 9000 \* 2) / (8 + 2) = 5800

• 조화평균 (Harmonic Mean) : 원소들의 값들의 역수의 평균을 다시 역수로 변환하여 계산.

• 주로 비율이나 속도 데이터에서 사용.

• 예시: {1, 2, 3}의 조화평균은 3 / (1/1 + 1/2 + 1/3) = 약 1.636.

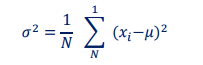
#### 산포도 측정

**• 산포도 (Measure of Dispersion) :**

• 통계에서 산포도를 측정하는 방법들은 데이터의 변동성을 파악하는 데 중요한 역할을 함• 분산 (Variance)

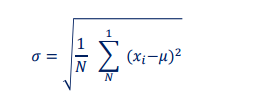
• 데이터 값들이 평균에서 얼마나 떨어져 있는지를 나타냄

• 계산 방법: 각 데이터 값에서 평균을 뺀 값을 제곱한 후, 이 값들의 평균을 구함.



• 표준편차(Standard Deviation)

• 분산의 제곱근으로, 데이터가 평균에서 얼마나 떨어져 있는지에 대한 표준적인 거리를 나타냄.



• 사분위수 범위(Interquartile Range, IQR)

• 데이터의 중앙 50%가 포함된 범위로, Q3(3사분위수)와 Q1(1사분위수)의 차이

• 극단값의 영향을 덜 받음



#### 왜도와 첨도

**• 왜도와 첨도 : 통계에서 데이터 분포의 특성을 설명하는 지표**

• 왜도 (Skewness) : 분포의 비대칭성을 측정하는 지표

• 첨도 (Kurtosis) : 분포의 뾰족함과 꼬리의 두꺼움을 측정하는 지표

### [확률과 분포]

#### 확률

• 표본 공간 (Sample Space) : 실험을 통해 나타날 수 있는 모든 결과들의 집합

• 사건 (Event) : 표본공간에 있는 일부 원소들로 이루어진 부분 집합

• 확률 : 특정 사건이 일어날 가능성을 나타내는 척도(사건이 발생할 가능성을 수치로 표현)

• 0에서 1 사이의 값을 가지며, 0은 사건이 절대로 일어나지 않음을 의미하며, 1은 사건이 반드시 일어남을 의미함

• 확률 함수 (Probability Function) : 사건이 발생할 확률을 나타내는 함수

• 규칙

• 덧셈 법칙 : 두 사건 𝐴, 𝐵 에 대하여, 𝑃 𝐴 ∪ 𝐵 = 𝑃 𝐴 + 𝑃 𝐵 − 𝑃(𝐴 ∩ 𝐵)

• 곱셈 법칙 : 두 사건 𝐴, 𝐵 가 독립 사건일 때, 𝑃 𝐴 ∩ 𝐵 = 𝑃 𝐴 × 𝑃

#### 확률 변수와 확률 분포

**• 확률 변수(Random Variable) :**

• 정의 : 특정 값이 나타날 가능성이 확률적으로 주어지는 변수

• 표본 공간의 각 표본점에 실수 값을 대응시키는 함수 역할을 함.

• 표본점 (Sample point) :

• 모집단에서 무작위로 뽑은 하나의 표본

• 확률 표본 (Random sample) :

• 모든 표본점(sample point)들이 동일한 확률로 추출된다는 조건 하에서 추출된 표본

• 이산 확률변수와 연속 확률변수로 구분

**• 이산형 확률변수 (Discrete random variable)**

• 0이 아닌 확률 값을 갖는 값이 셀 수 있는 경우의 확률 변수

• 예: 주사위를 굴렸을 때 나오는 눈의 수, 동전 던지기의 결과.

• 확률질량함수(Probability Mass Function) : 이산형 확률변수의 확률함수

• 이산형 확률변수에 의한 확률 분포 : 이항분포, 기하분포, 포아송분포 등

**• 연속형 확률변수 (Continuous random variable)**

• 가능한 값이 실수의 어느 특정 구간 전체에 해당하는 확률 변수

• 연속적인 값을 가짐.

• 예: 특정 시간 동안의 온도, 사람의 키

• 확률밀도함수(Probability Density Function) : 연속형 확률변수의 확률함수

• 연속형 확률변수에 의한 확률 분포 : 정규분포, 균일분포, t-분포, 카이제곱분포 등

#### 확률 분포

• 정의 : 확률 변수가 가질 수 있는 값들과 그 값들이 발생할 확률을 나타내는 함수

• 이산 확률 분포 (Discret Probability Distribution)

• 확률 변수가 취할 수 있는 값이 유한하거나 셀 수 있는 경우

• 확률질량함수(Probability Mass Function : PMF)로 표현

• 예 : 베르누이분포, 이항분포, 포아송분포

• 연속 확률 분포 (Continuous Probability Distribution)

• 확률 변수가 취할 수 있는 값이 연속적인 경우.

• 확률밀도함수(Probability Density Function : PDF)로 표현

• 예: 정규 분포, 지수 분포, 카이제곱 분포.

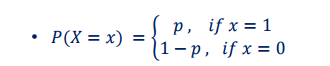
#### 확률 분포 – 이산 확률 분포

**• 베르누이 분포 (Bernoulli Distribution) :**

• 두 가지 결과(성공 또는 실패)가 있는 단일 시행의 확률분포

• 확률변수 𝑋 : 성공일때 1, 실패일때 0의 값

• 확률질량함수 (PMF) : (0 ≤ 𝑝 ≤ 1 는 성공 확률)

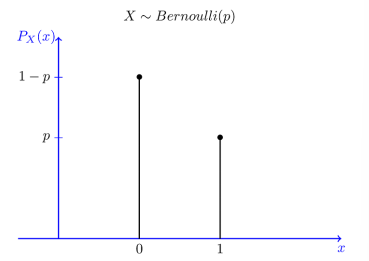


• 기대값 : 𝐸(𝑋) = 𝑝

• 예시 :

• 동전 던지기(앞면/뒷면)

• 제품 검수(불량/정상)

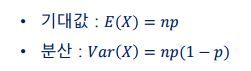


• 이항 분포 (Binomial Distribution):

• 일정한 횟수의 독립적인 베르누이 시행에서 성공의 횟수를 나타내는 분포

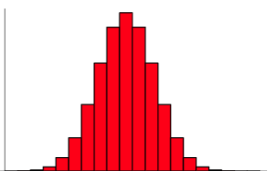
• 확률변수 𝑋 : 𝑛 번의 독립적인 시행 중 성공 횟수

• 확률질량함수 (PMF) : 𝑘 번 발생할 확률

* 
* 

• 예시 :

• 동전 여러 번 던지기 : 동전을 10번 던져서 앞면이 나오는 횟수



• 푸아송 분포 (Poisson Distribution):

• 단위 시간(또는 단위 공간)에서 특정 사건의 발생 빈도를 나타내는 분포

• 드물게 발생하는 사건의 빈도를 모델링할 경우 사용됨

• 확률변수 𝑋 : 일정한 시간에서 발생하는 사건의 횟수

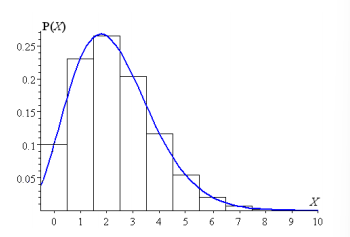
• 확률질량함수 (PMF) : (𝜆 는 단위 시간에서 사건이 발생하는 평균 횟수)



• 예시 :

• 콜센터 : 1시간 동안 걸려오는 전화의 수

• 교통량 : 특정 도로에서 1시간 동안 발생하는 교통사고 수



#### 확률 분포 – 연속 확률 분포

• 정규 분포 (Normal Distribution) :

• 종 모양의 대칭 분포로, 자연 현상에서 많이 나타나는 일반적인 분포

• 확률밀도함수 (PDF) : ( 𝜇 는 평균, 𝜎 는 표준편차를 의미)



• 특성 :

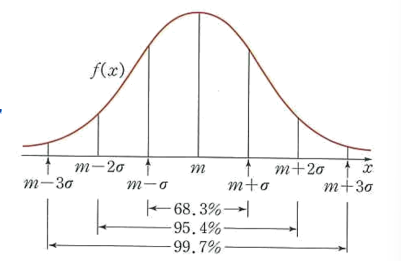
• 평균 𝜇 를 중심으로 대칭

• 확률변수의 68%가 𝜇 ± 𝜎, 95%가 𝜇 ± 2𝜎, 99.7%가 𝜇 ± 3𝜎 범위 내에 위치함

• 예시 :

• 키 : 성인의 키 분포

• 시험정수 : 대규모 시험의 성적 분포



• 지수 분포 (Exponential Distribution) :

• 사건 간의 시간 간격에 대한 분포. 주로 포아송 과정에서 시간 간격을 모델링하는 데 사용

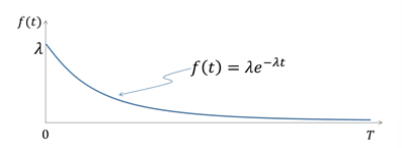
• 확률밀도함수 (PDF) : ( 𝜆 는 사건 발생률 )



• 예시 :

• 콜센터 대기 시간 : 다음 전화가 걸려올 때까지의 시간.

• 전자 부품 오동작 간격 : 전자 부품이 오동작 할 때까지의 시간.



• 카이제곱 분포 (Chi-Square Distribution)

• 정규 분포를 따르는 독립적인 변수들의 제곱의 합의 분포

• 모집단의 분산을 추정할 때 주로 사용됨

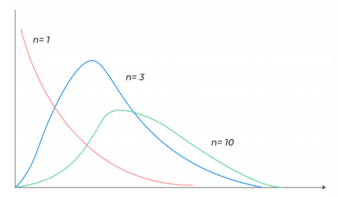
• 확률밀도함수 (PDF) : ( 𝑘 는 자유도, Γ 는 감마함수)



• 예시 :

• 적합도 검정 : 관찰된 데이터가 기대 분포와 일치하는지 검정

• 분산 분석 : 여러 집단의 분산을 비교



### [추정과 가설검정]

#### 추론

• 추론

• 표본을 활용하여 모집단의 특성을 추측함

• 모집단 전체를 조사할 수 없을 경우 주로 사용됨

• 추론의 방법

• 모수의 추정 :

• 미지수인, 모집단의 모수에 대한 추측 또는 추측값을 정확도와 함께 제시함

• 모수에 대한 가설검정 :

• 모집단의 모수에 대한 여러 가설들이 적합한지 여부를 표본으로부터 판단함

• 예시 : (직장인들의 연말 상여금 조사 예시)

• 표본조사를 통해 전국 직장인 상여금의 평균을 하나의 값으로 추정

• 표본조사를 통해 전국 직장인 상여금의 평균이 포함될만한 구간을 정함

• 표본조사를 통해 전국 직장인 상여금 평균이 5년전 평균값과 얼마다 다른지 판단함

#### 추정

• 점추정 (Point Estimation)

• 정의 : 모집단의 모수를 단일 값(점)으로 추정하는 방법

• 대푯값 : 일반적으로 표본평균, 표본분산 등이 사용됨

• 예시 : 표본평균이 50이라면 모집단 평균도 50으로 추정함

• 구간추정 (Interval Estimation)

• 정의 : 모집단의 모수를 포함할 것으로 예상되는 구간을 제시하는 방법

• 구성 : 신뢰구간(Confidence Interval)과 신뢰수준(Confidence Level)으로 구성

• 신뢰구간 : 모수가 포함될 것으로 예상되는 범위

• 신뢰수준 : 모수가 해당 구간에 포함될 확률

• 예시 : 95% 신뢰수준에서 신뢰구간이 [25, 55]라면 관심있는 모수가 25~55 사이에 있을 확률이 95%라는 의미

#### 신뢰구간, 신뢰수준

• 신뢰구간 (Confidence Interval) : 모수가 신뢰구간 안에 포함될 것으로 예상되는 범위

• 표본을 사용하여 계산하며 구간이 넓을 수록 모수를 포함할 확률이 높아짐

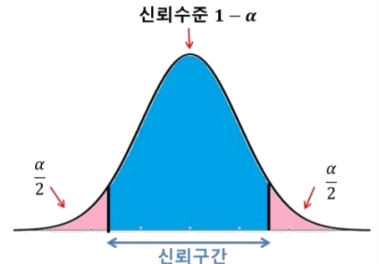
• 신뢰수준 (Confidence Level) : 모수가 신뢰구간에 포함될 확률

• 𝑃 𝑎 ≤ 𝜇 ≤ 𝑏 = 1 − 𝛼

• 90%, 95%, 99%가 자주 사용됨

• 유의수준 𝛼 :

• 모수가 신뢰구간에 포함되지 않을 확률



#### 가설 검정

• 가설 검정 : 어떤 추측이나 주장, 가설에 대해 타당성을 조사하는 작업

• 표본 통계량으로 모수 추정 시, 추정한 모수값이나 확률분포 등이 타당한지 평가하는 통계적 추론 방법

• 귀무가설 (Null Hypothesis, H0) : 버릴 것으로 예상하는 가설

• 대립가설 (Alternative Hypothesis, H1) : 실제 주장 또는 증명하려는 가설• 가설 검정 단계: 가설 수립 -> 유의수준 결정 -> 검정통계량 계산 -> 기각/채택 결정

#### 가설 검정 예시 (진통제 개발의 예시)

• 귀무가설/대립가설 수립 :

• 귀무가설 H0 : 새로운 진통제의 효과가 기존 진통제와 차이가 없다.

• 대립가설 H1 : 새로운 진통제가 기존 진통제보다 더 효과적이다.

• 유의수준 결정 :

• 유의수준을 0.05로 설정

(5%의 확률로 귀무가설이 참인데도 불구하고 기각할 가능성을 허용함)• 검정 통계량 계산 :

• 임상 실험을 통해 두 그룹의 환자들을 각각 새로운 진통제와 기존 진통제 투약

• 각 그룹에서 진통 정도를 수치화한 데이터 수집

• 두 그룹의 평균 진통 수치를 비교하여 검정 통계량(예: t-검정) 계산

• 기각/채택 결정 :

• P-value 계산 : 검정 통계량을 바탕으로 P-value를 계산

• P-value : 귀무가설 하에서 관측된 데이터가 발생할 확률 (H0을 지지하는 값)

• 결정 :

• P-value가 유의수준 0.05보다 작으면 귀무가설 기각 (효과 있다)

• 그렇지 않으면 귀무가설 채택 (효과 없다)

#### 가설 검정 – 오류

• 1종 오류 (Type I Error) :

• 정의 : 귀무가설이 참인데도 불구하고 이를 기각하는 오류

• 예시 : 새로운 진통제와 기존 진통제의 효과에 차이가 없는데, 새로운 진통제가 더 효과적이라고 결론 내리는 경우

• 결과 : 잘못된 긍정(False Positive) 결과 초래

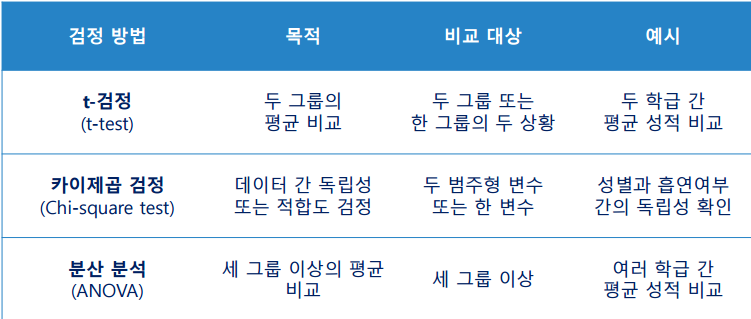
• 2종 오류 (Type II Error)

• 정의 : 대립가설이 참인데도 불구하고 귀무가설을 기각하지 않는 오류

• 예시 : 새로운 진통제가 기존 진통제보다 효과가 있음에도 불구하고, 두 진통제의 효과 차이가 없다고 결론 내리는 경우

• 결과 : 잘못된 부정(False Negative) 결과 초래

#### 가설 검정 방법



#### 가설 검정 방법 – t 검정

• t-검정(t-test) :

• 두 그룹 간의 평균의 차이가 유의미한지 확인하고자 할 때 주로 사용

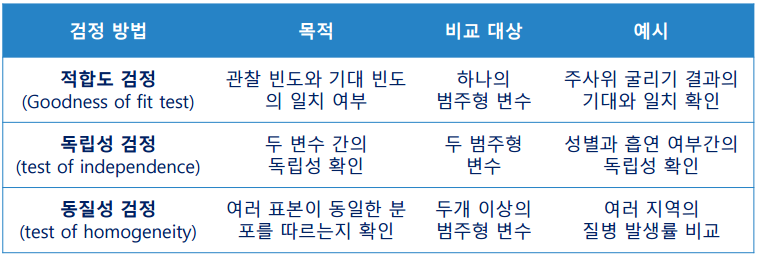


#### 가설 검정 방법 – 카이제곱 검정

• 카이제곱 검정(Chi-square test) :

• 범주형 데이터에서 기대 빈도와 관찰된 빈도 간의 차이를 확인할 때 사용

• 두 범주형 데이터에서 변수 간 독립성 여부를 검정할 때 사용



#### 가설 검정 방법 – 분산 분석

• 분산 분석(ANOVA : Analysis Of VAriance) :

• 3개 이상의 집단에 대한 평균 차이를 검증하는 분석 방법



### [상관분석 ]

#### 상관분석 – 개요

• 상관분석

• 두 변수 간의 관계를 정략적으로 평가하는 통계 기법

• 하나의 변수가 변할 때 다른 변수가 어떻게 변하는지 파악할 수 있음.

• 상관계수 (Correlation Coefficient)

• 두 변수 간의 관계의 강도와 방향을 나타내는 수치

• 상관관계의 방향

• 양의 상관관계 : 한 변수가 증가할 때 다른 변수도 증가하는 경향이 있는 경우

• 음의 상관관계 : 한 변수가 증가할 때 다른 변수는 감소하는 경향이 있는 경우

• 상관관계의 강도

• 강한 상관관계 : 상관계수의 절댓값이 0.7 이상

• 중간 상관관계 : 상관계수의 절댓값이 0.3~0.7 사이

• 약한 상관관계 : 상관계수의 절댓값이 0.3 미만



#### 상관분석

• 스피어만 상관계수 (Spearman's Rank Correlation Coefficient)

• 측정 대상 : 순위형 변수 간의 관계 측정 (비선형 관계일 경우에도 사용 가능)

• 계산 방법 : 각 변수의 순위를 매긴 후, 그 순위들 간의 상관계수를 계산함

(순위 차이에 따라 값을 계산하므로, 변수 간의 비선형 관계 반영 가능)

• 해석 :

• 𝑟𝑠 = 1 ∶ 완전한 양의 상관관계

• 𝑟𝑠 = −1 : 완전한 음의 상관관계

• 𝑟𝑠 = 0 : 상관관계 없음

• 장점 : 비선형 관계와 이상치에 민감하지 않음

• 단점 : 데이터의 순위 정보만 사용하여 정보 손실이 발생할 수 있음 (원래 데이터의 크기 정보 등)

• 예시: 학생 성적 순위와 스포츠 성적 순위 간의 관계.

(학생들의 성적과 스포트 성적 간의 관계 분석 시 점수를 순위로 변경하여 상관관계 분석)

#### 상관분석 – 개요

• 켄달의 타우 (Kendall's Tau, Kendall's Rank Correlation Coefficient)

• 측정 대상 : 변수의 순서 간의 상관성을 측정 (작은 데이터셋에 적합)

• 계산 방법 : 순위 쌍 간의 일치와 불일치를 비교하여 상관관계를 계산

• 해석 : 피어슨 상관계수와 동일하게 해석함 (𝜏)

• 장점 : 순위 정보에 기반하여 비선형 관계도 잘 반영

• 단점 : 계산이 복잡하며, 해석이 어려울 수 있음 (특히 데이터셋이 클 경우 계산 부담 증가)

• 예시 : 직무 수행 순위와 승진 순위 간의 관계

## [실습 시작 전 이론 설명 / 날씨 데이터를 활용한 로켓 발사 예측 모델 구현]

### [머신러닝 개요]

#### 인공지능(AI, Artificial Intelligence)

인공지능(AI, Artificial Intelligence) : 인간이 가지고 있는 지적 능력을 컴퓨터에서 구현하는 기술

• 고도의 문제 해결 능력을 가진 인공적인 지능

• 인간의 학습능력, 추론능력, 지각능력을 인공적으로 구현하려는 컴퓨터 과학의 세부 분야

• 머신러닝, 딥러닝 등의 기술을 기반으로 데이터분석, 추론, 예측, 분류, 자연어처리, 추천 등에 활용됨

#### 인공지능 발전 요인

빅데이터 : 최근 몇년간 데이터의 막대한 증대 (스마트폰, SNS)

클라우드 컴퓨팅 : 강력한 컴퓨팅 파워로의 손쉬운 접근

딥러닝 알고리즘 : 인공신경망 기술의 급격한 발전

#### 인공지능 vs 머신러닝 vs 딥러닝

• 인공지능 (Artificial Intelligence)

인간의 학습능력, 추론능력, 자연어 이해능력 등을 컴퓨터 프로그래밍으로 실현한 기술

• 머신러닝 (Machine Learning)

인공지능 연구 분야 중 하나로, 기계가 코드로 명시하지 않은 동작을 정제된 데이터로부터 학습하여 스스로 패턴을 찾아 작업을 실행할 수 있도록 하는 알고리즘

• 딥러닝 (Deep Learning)

머신러닝의 일부로, 인간의 뇌 신경망을 모방한 인공신경망의 일종. 일반적인 머신러닝 모델과 다르게 데이터를 사람이 추출해서 학습시키는 것이 아니라 데이터 자체를 전달하여 학습시키며 인공신경망 구조를 이용함

#### 머신러닝(Machine Learning)

머신러닝 : 주어진 데이터 속에서 규칙을 학습하는 알고리즘을 연구하는 데이터 과학의 한 분야• 데이터에 기반하여 규칙을 파악하고, 이 규칙에 따라 새로운 입력값의 결과를 예측

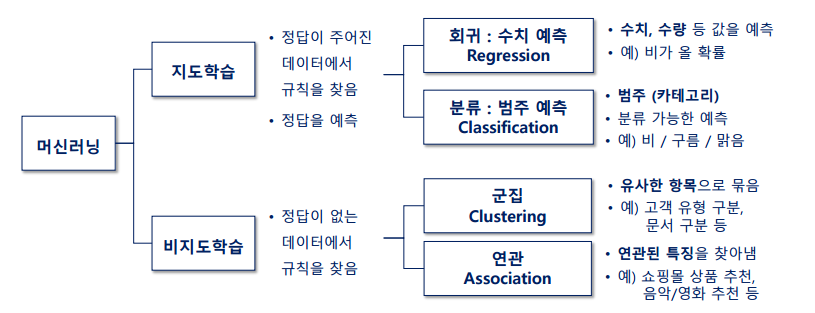
#### 머신러닝 유형

정답 데이터 유무에 따라 지도학습과 비지도학습으로 구분

- 지도학습 : 명시적인 정답 데이터를 가지고 규칙을 찾음

- 비지도학습 : 정답 데이터 없이 규칙을 찾음

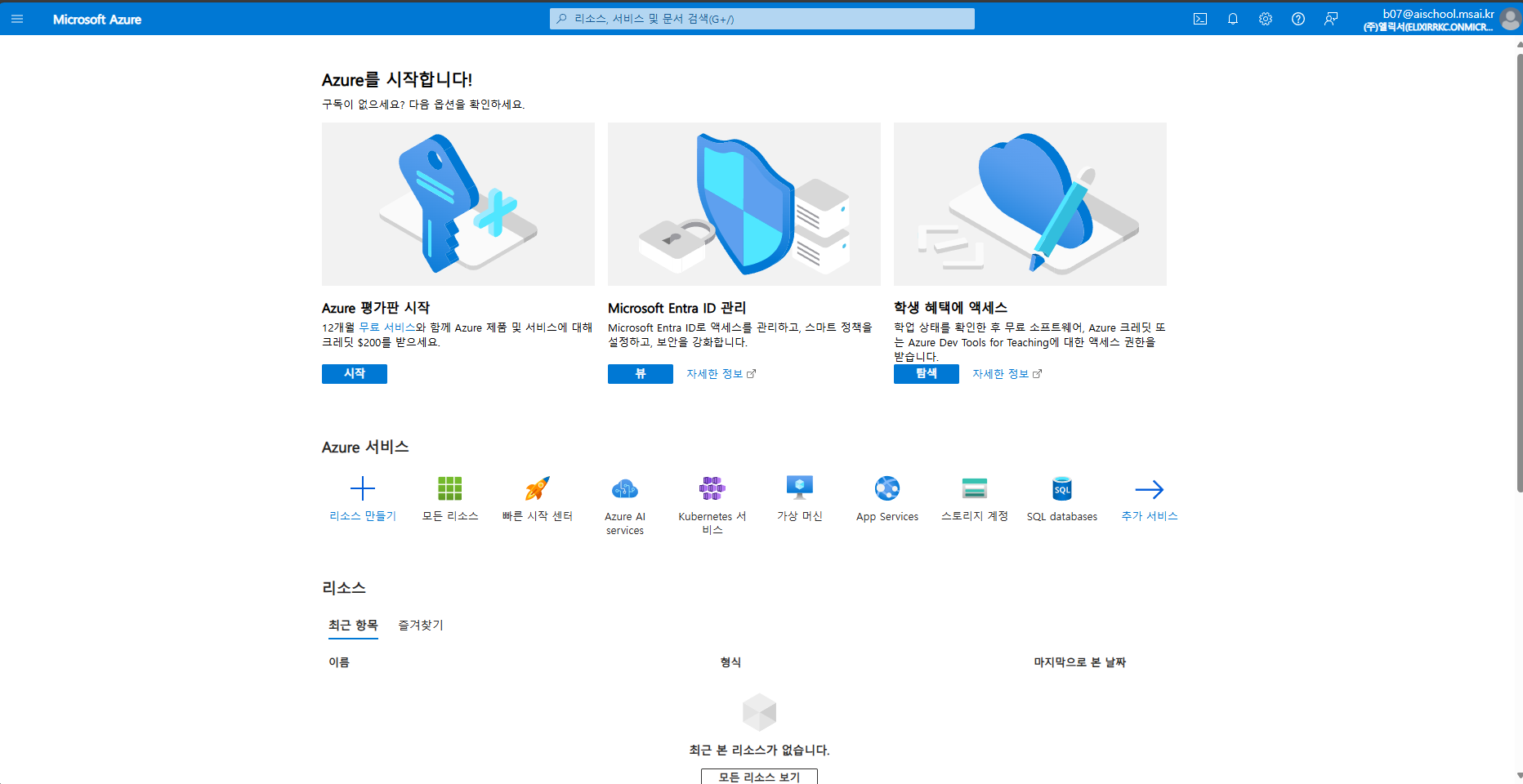
#### 머신러닝 유형 정리

지도학습은 수치 등의 값을 예측하는 회귀, 예/아니오 등의 범주를 예측하는 분류로 나눠지며 비지도학습은 유사한 항목을 그룹화한 군집 등이 있음

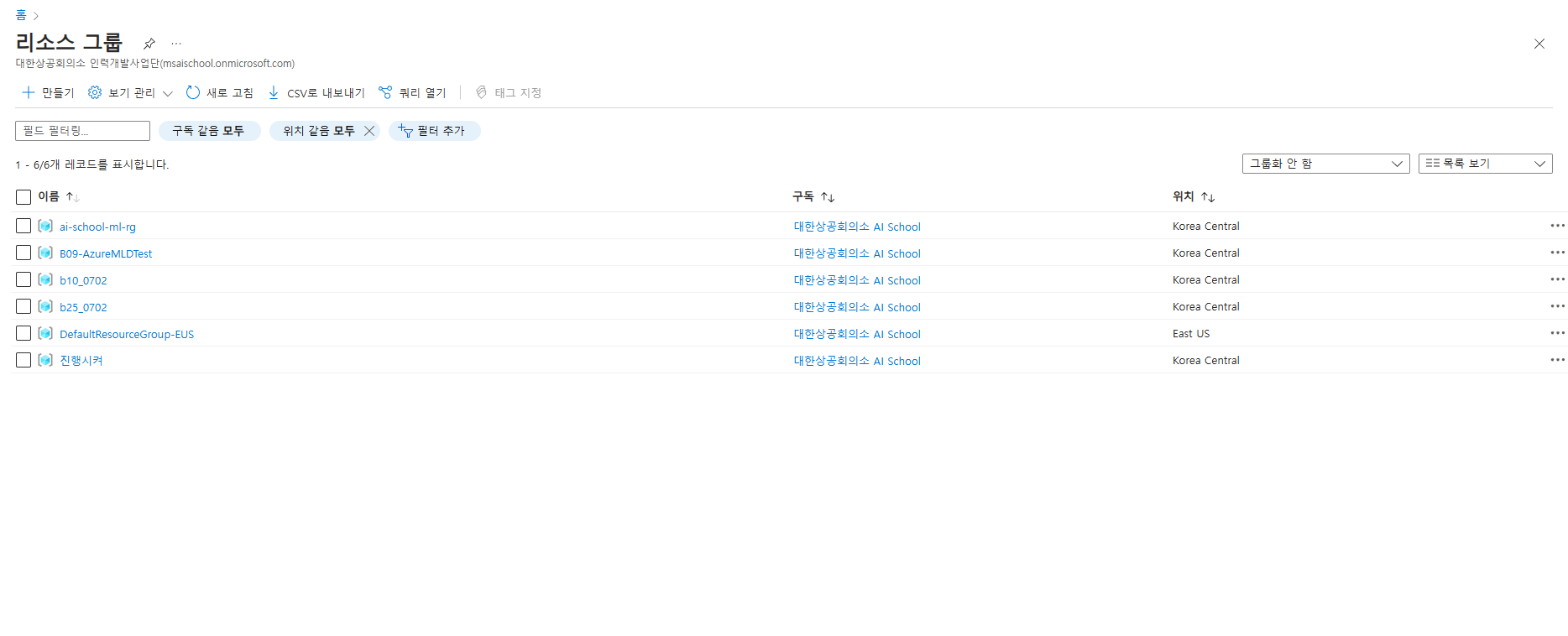
# <오후>

## [실습 시작/ 날씨 데이터를 활용한 로켓 발사 예측 모델 구현]

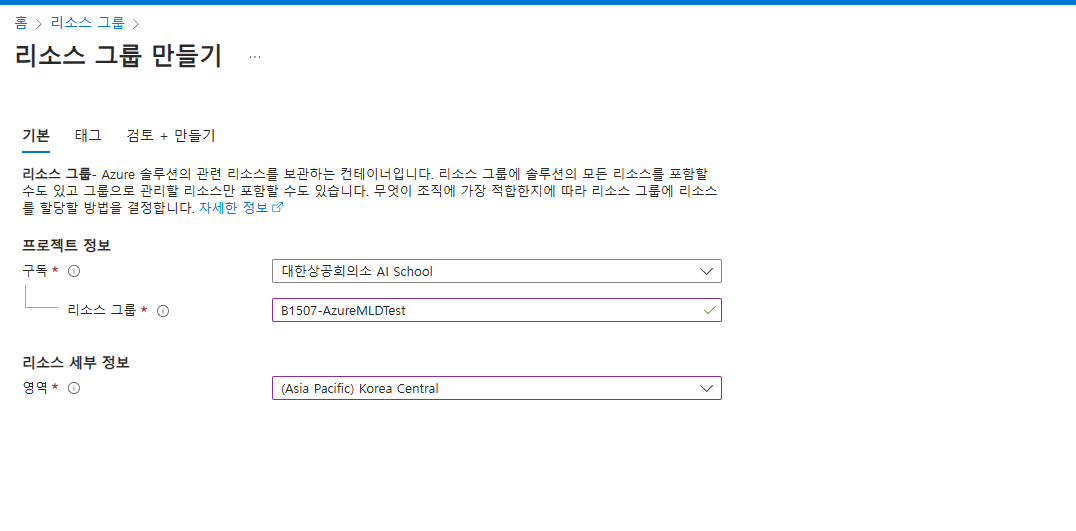
Azure 화면



리소스 그룹



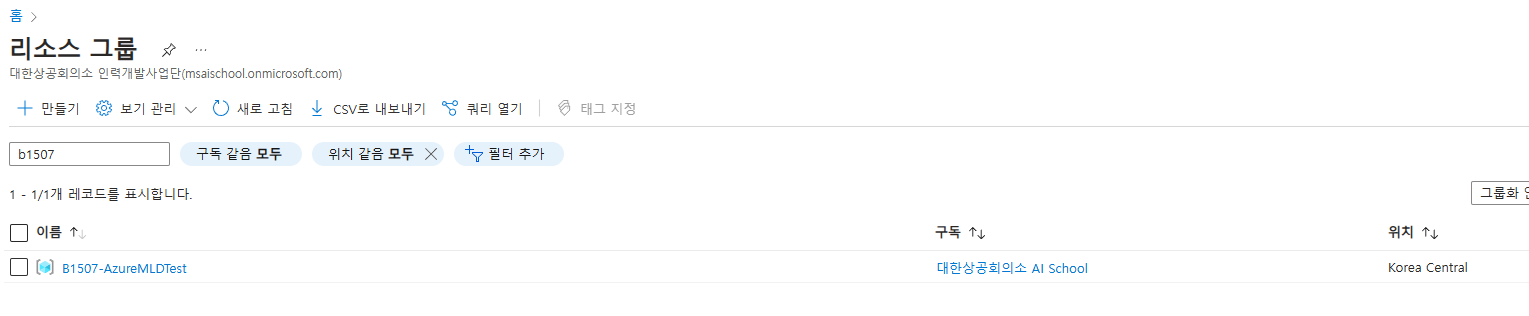
리소스 그룹 생성 모습

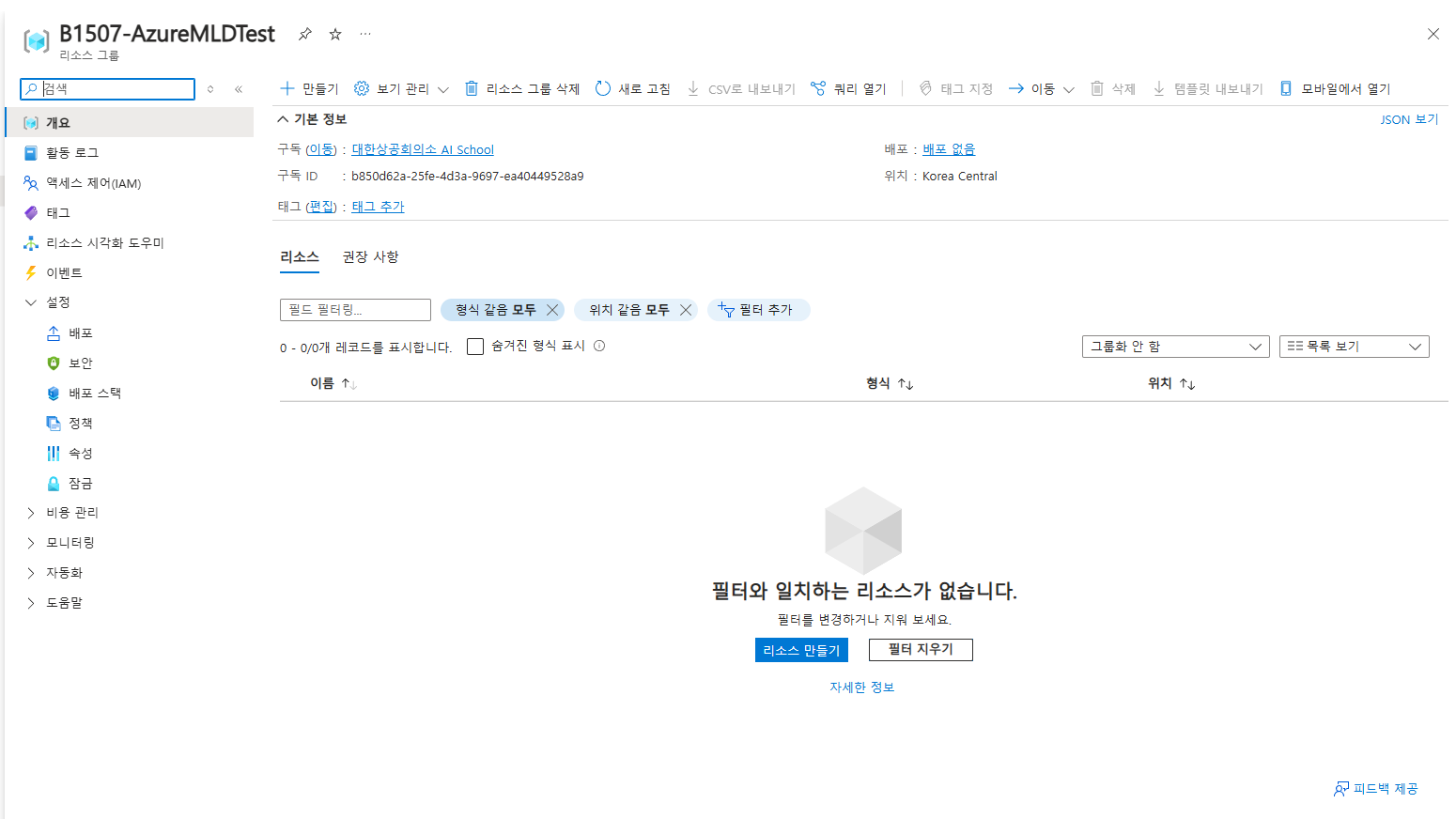


만들기 중

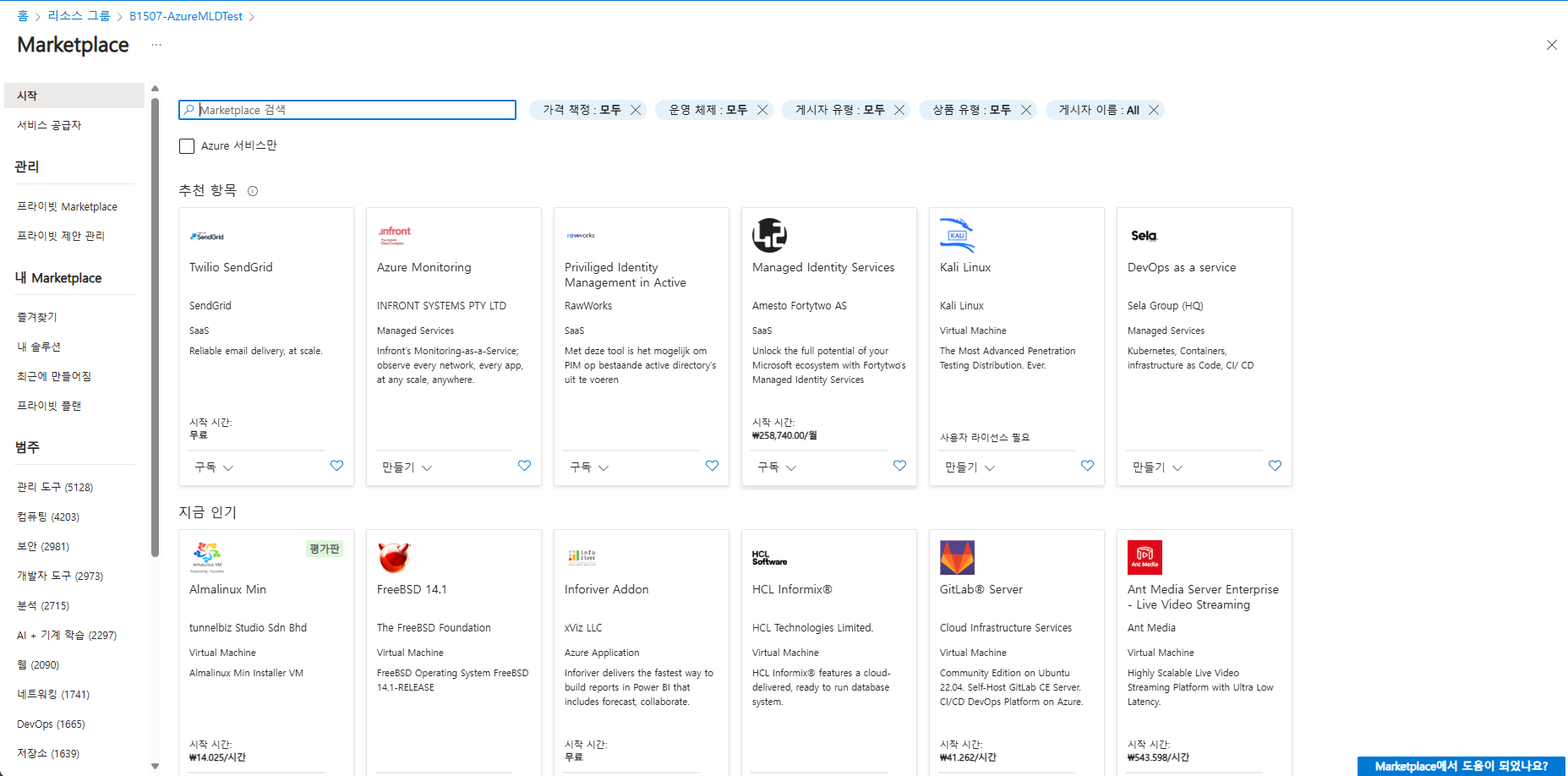


생성 후 확인

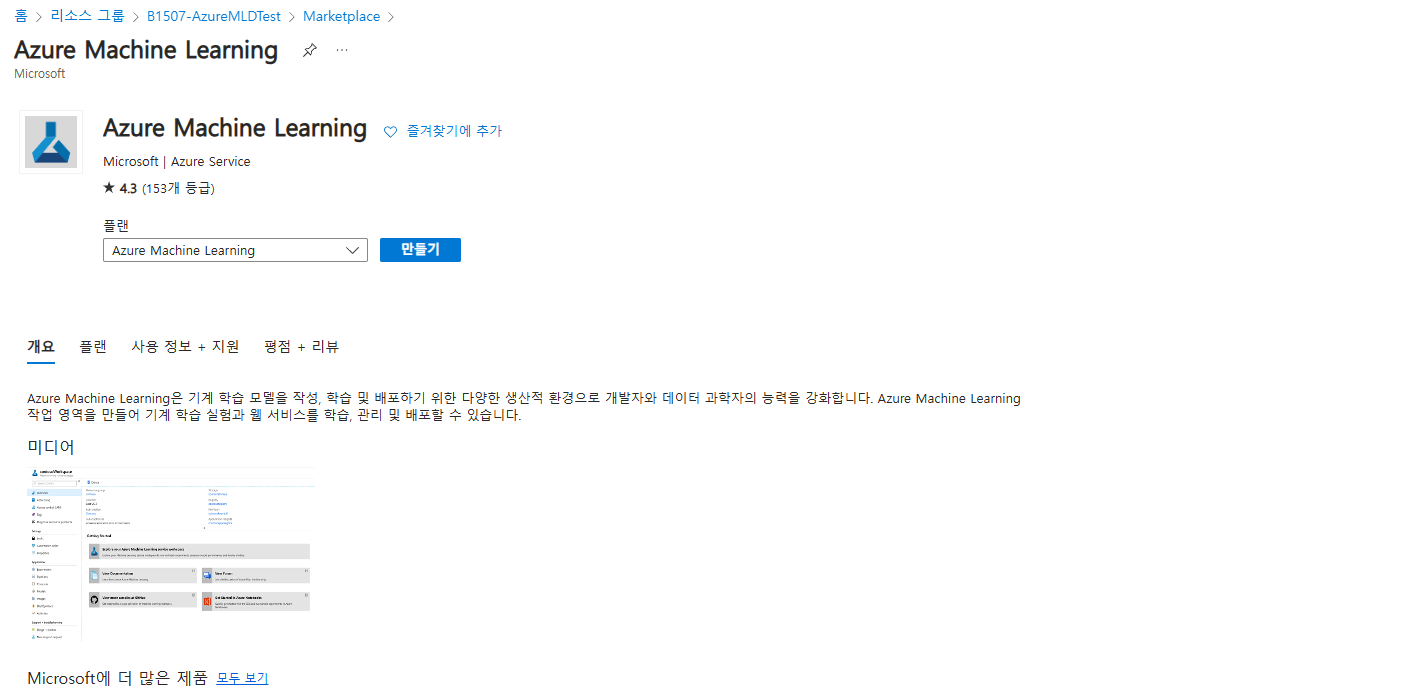




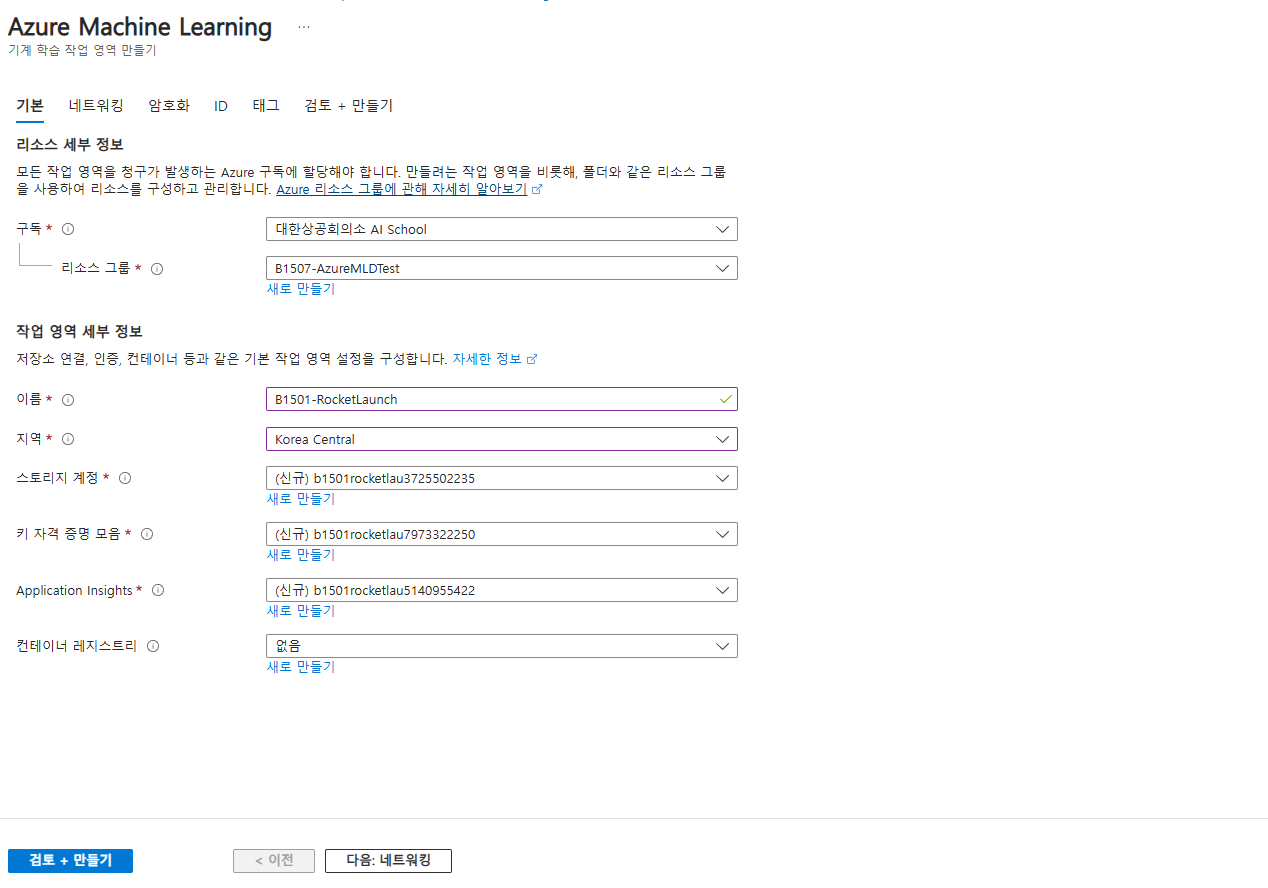
만들기 클릭



Azure Machine Learning 검색 후 선택

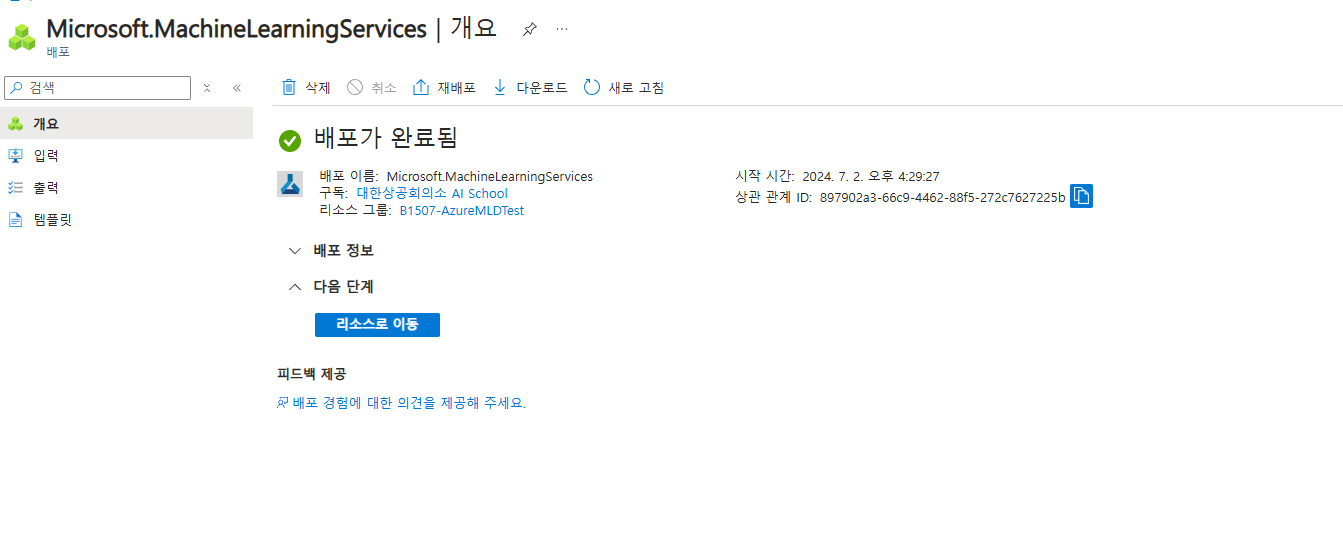


만들기

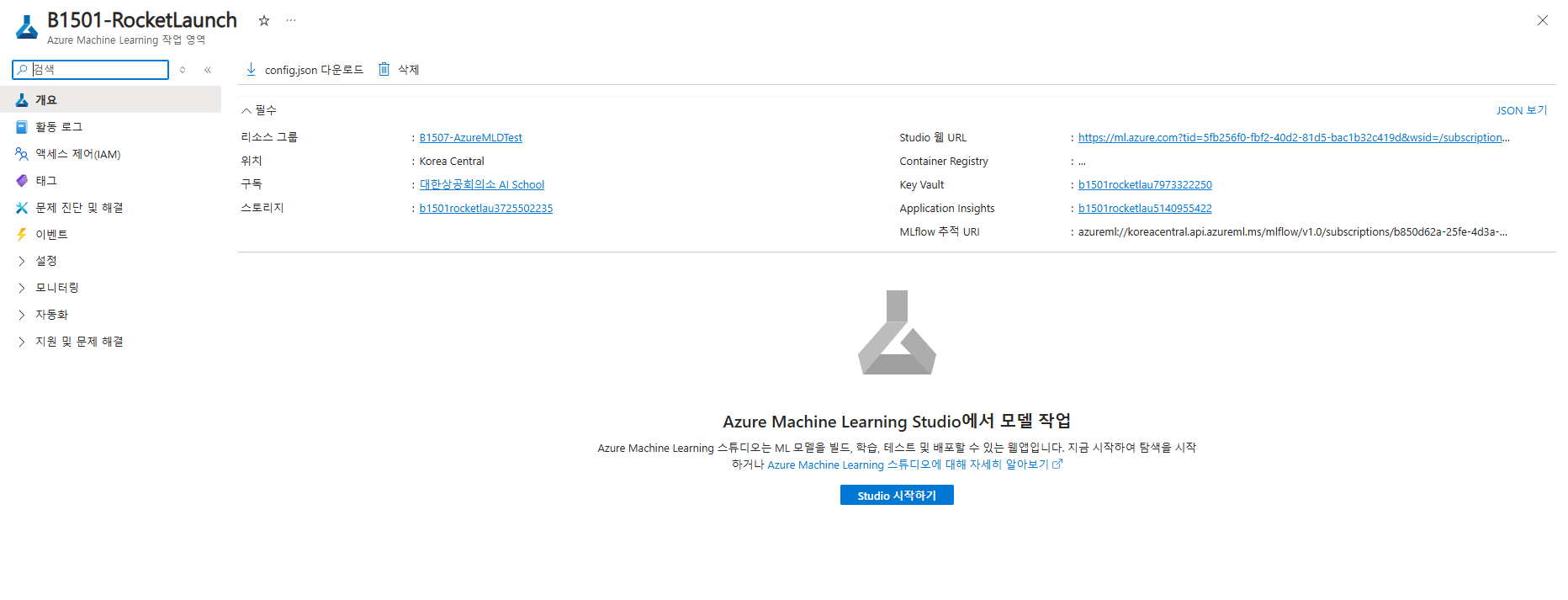




배포 완료 확인

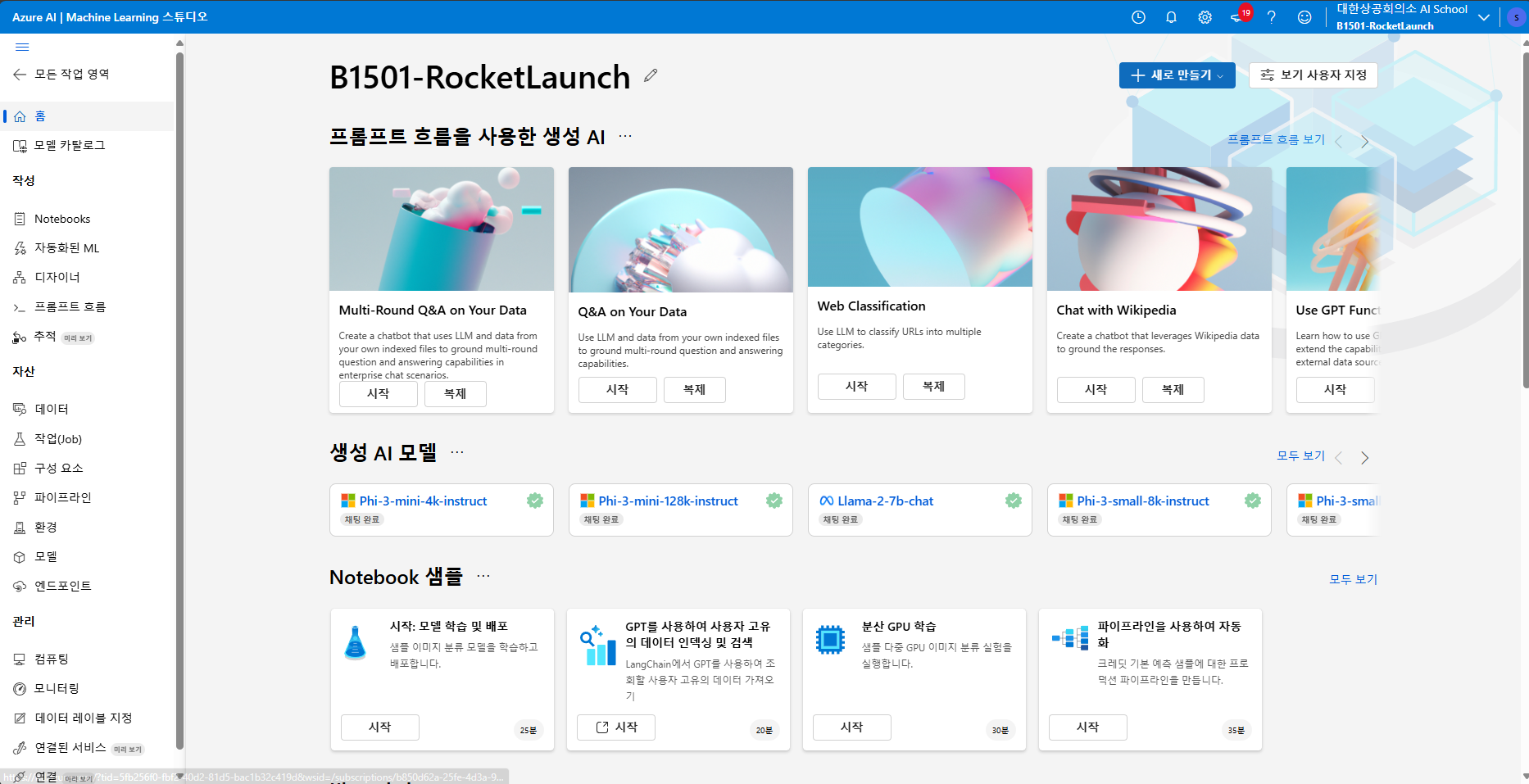


리소스로 이동 클릭

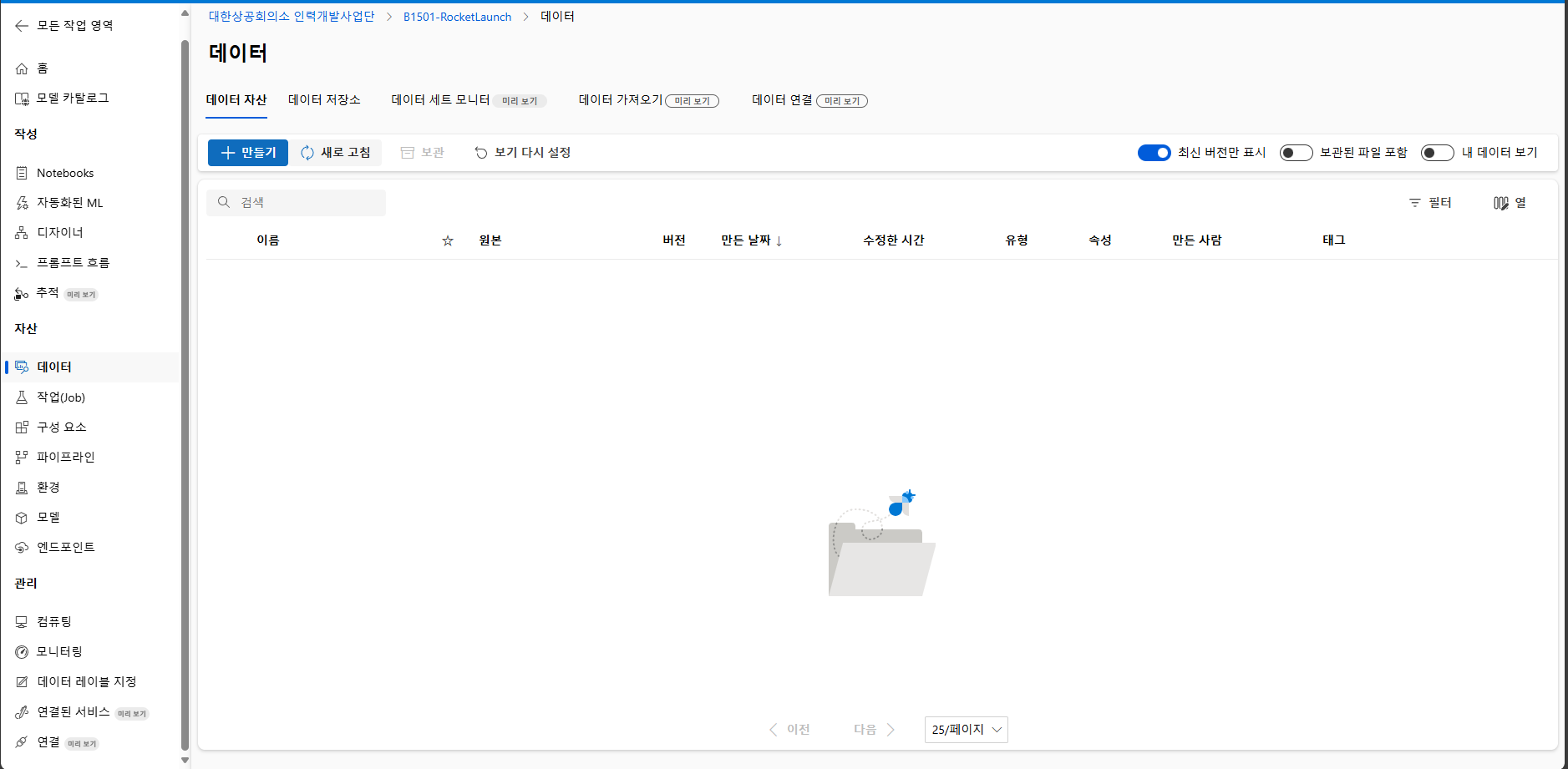


Studio 시작하기 클릭

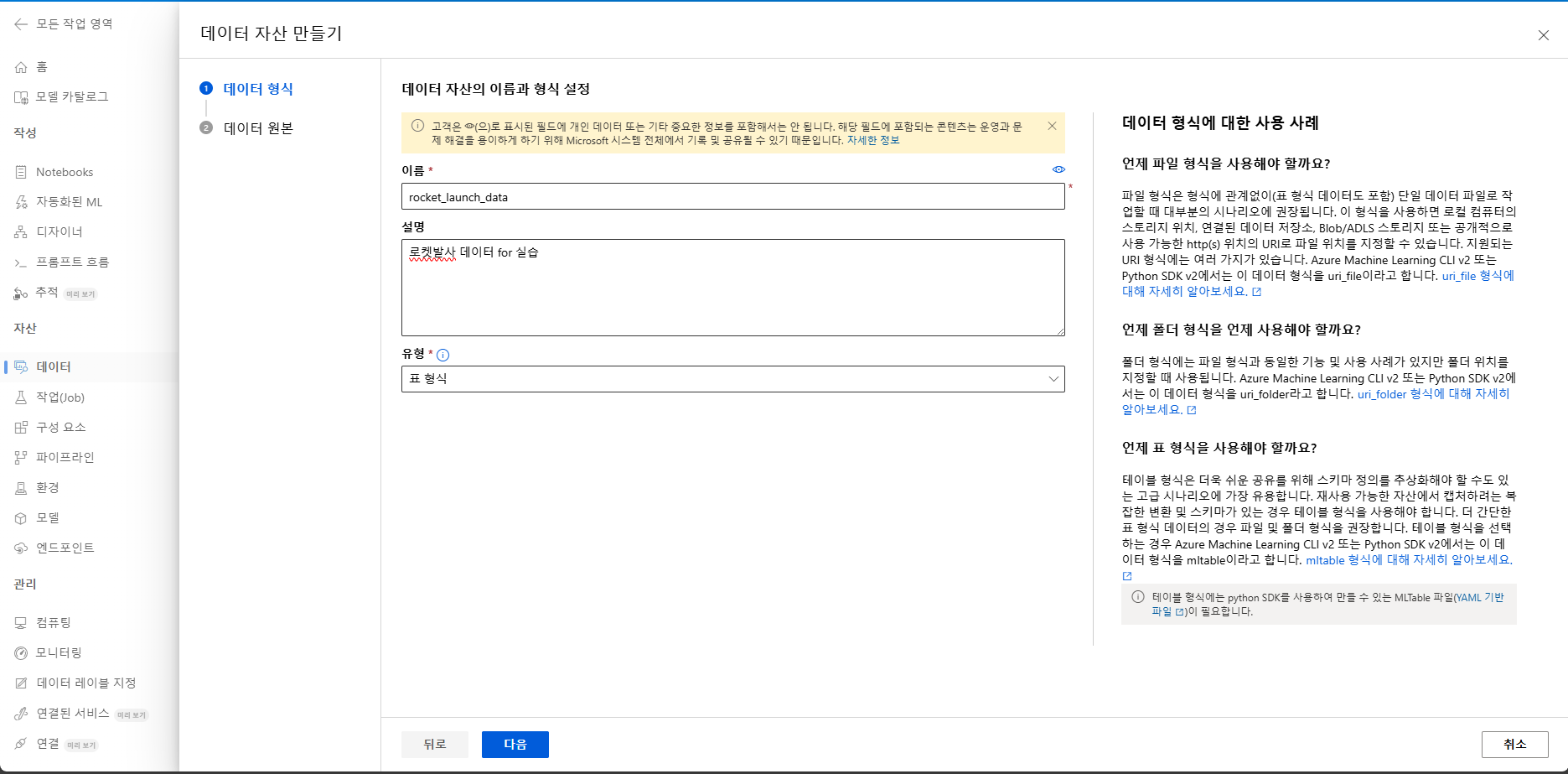




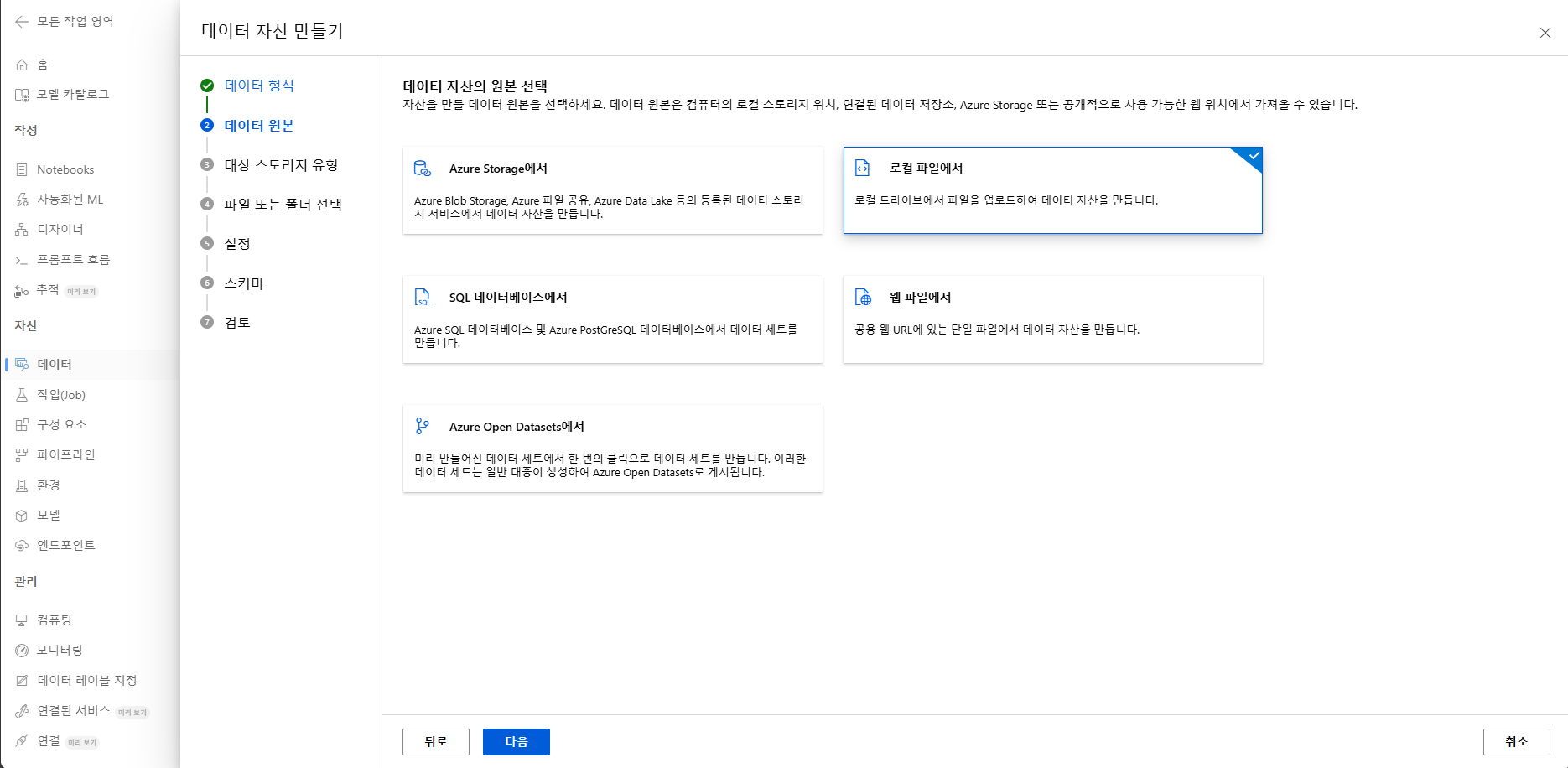
데이터



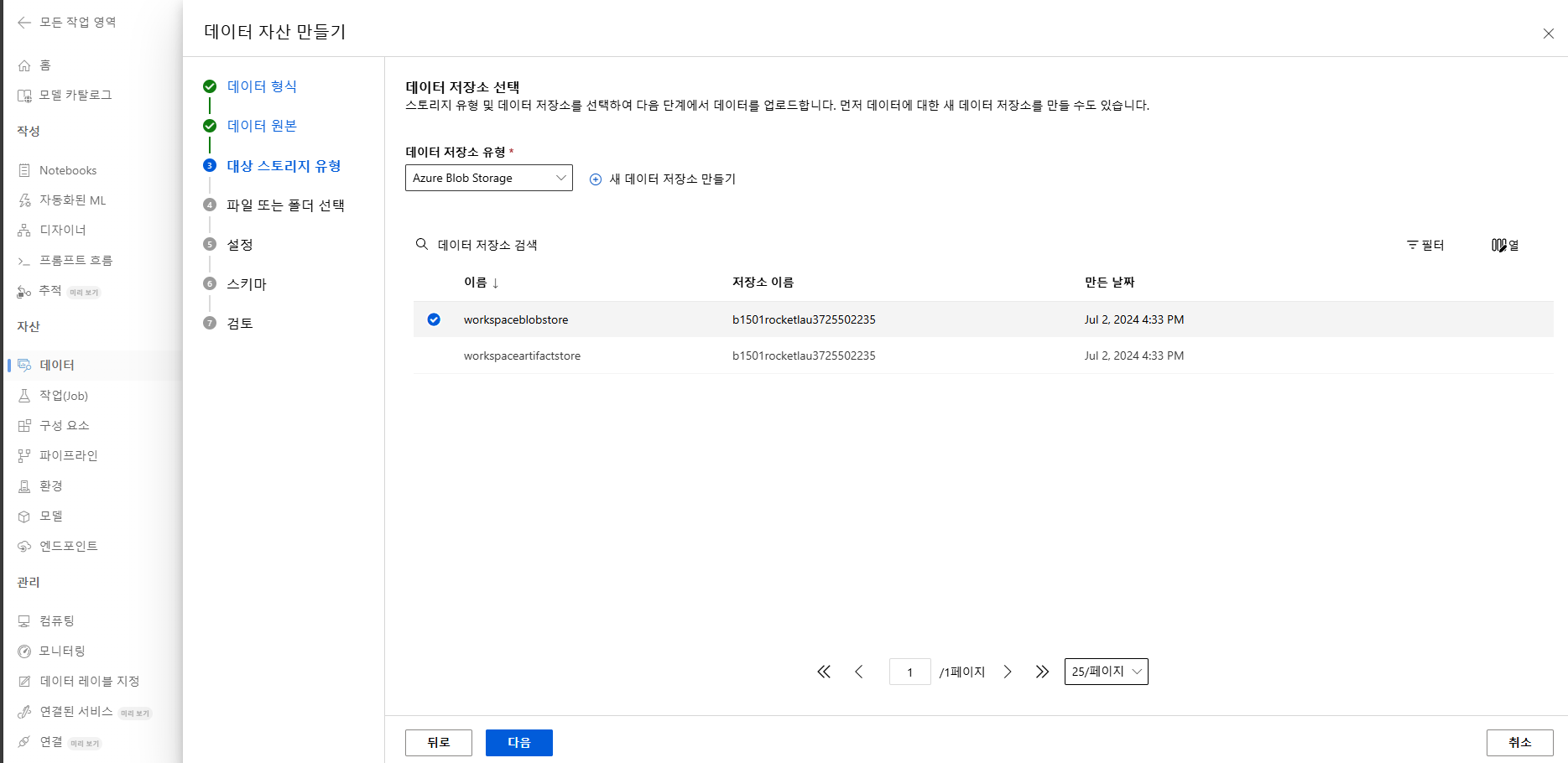
만들기 / 이름과 설명 작성 및 유형 선택



로컬 파일에서 선택



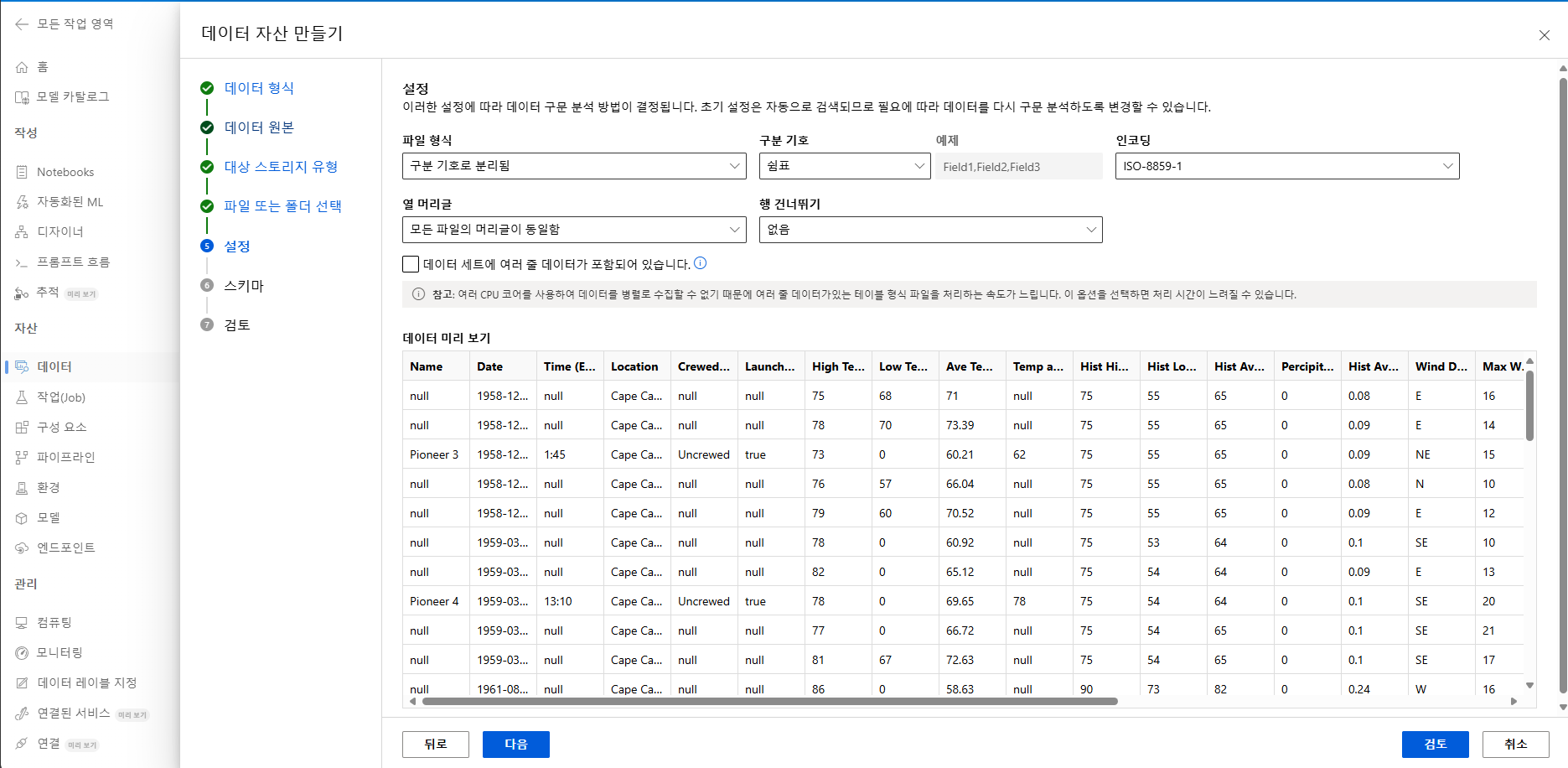
저장소 선택



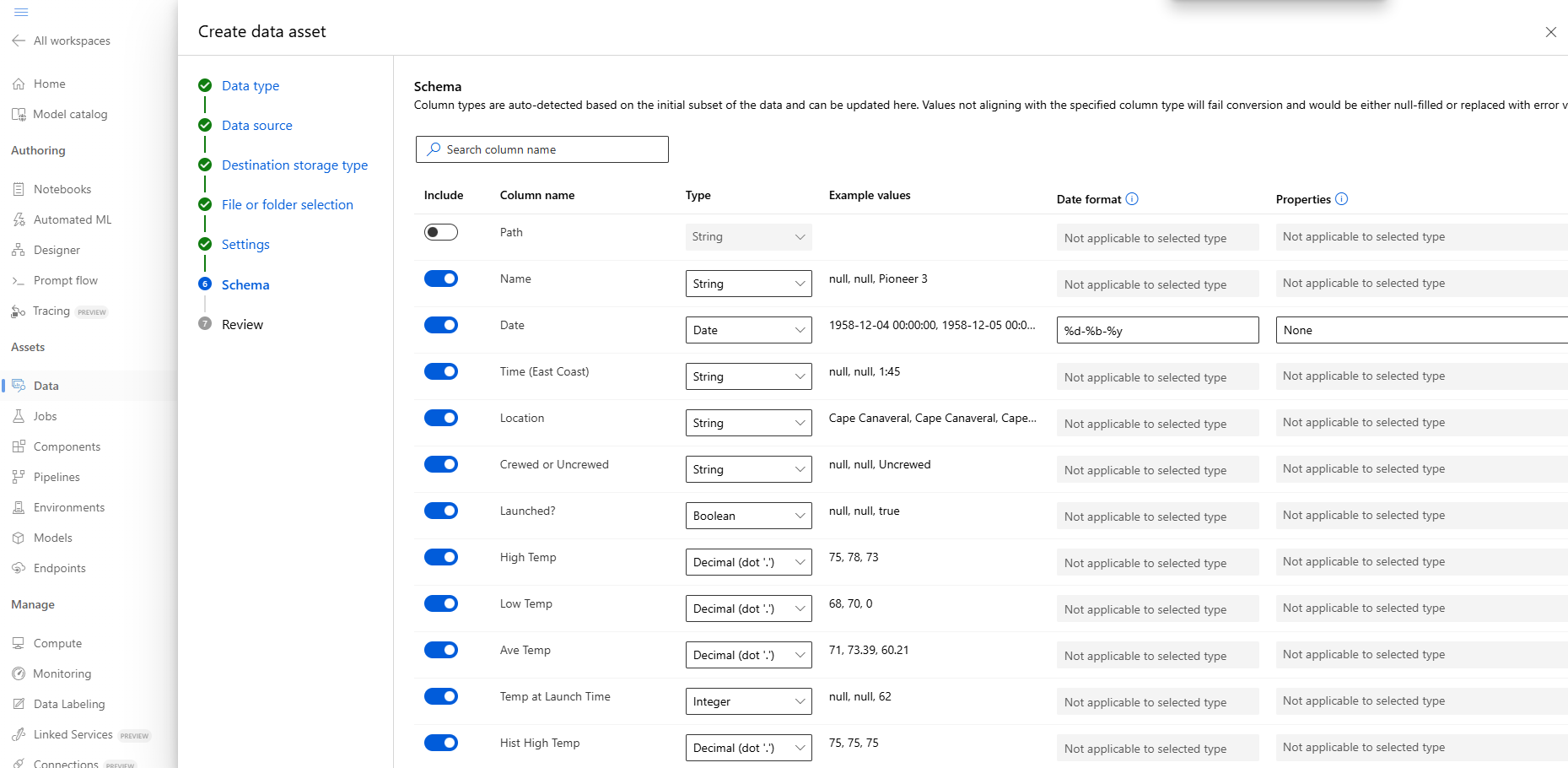
파일 불러오기



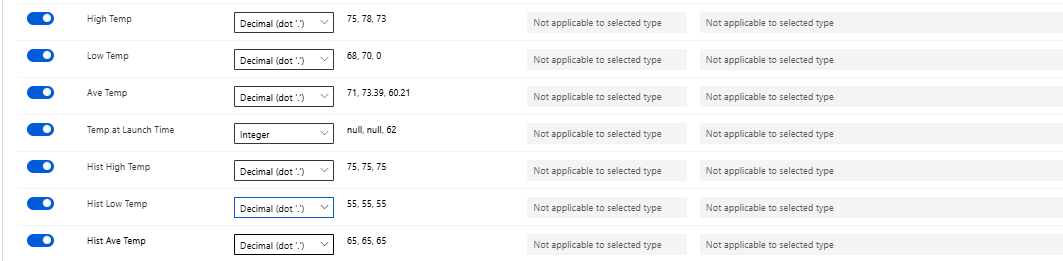
불러온 csv 데이터 확인

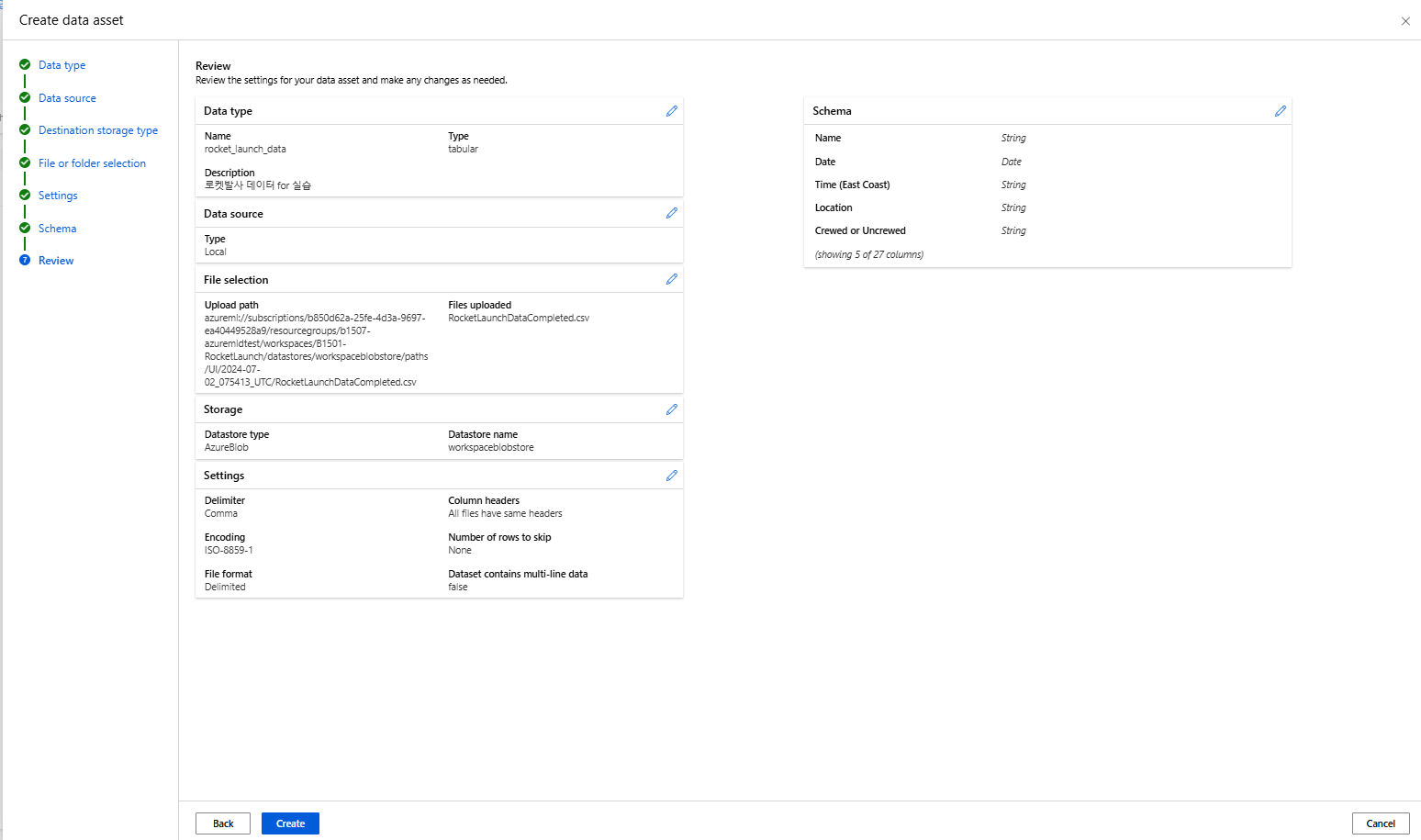


데이터 스키마 설정

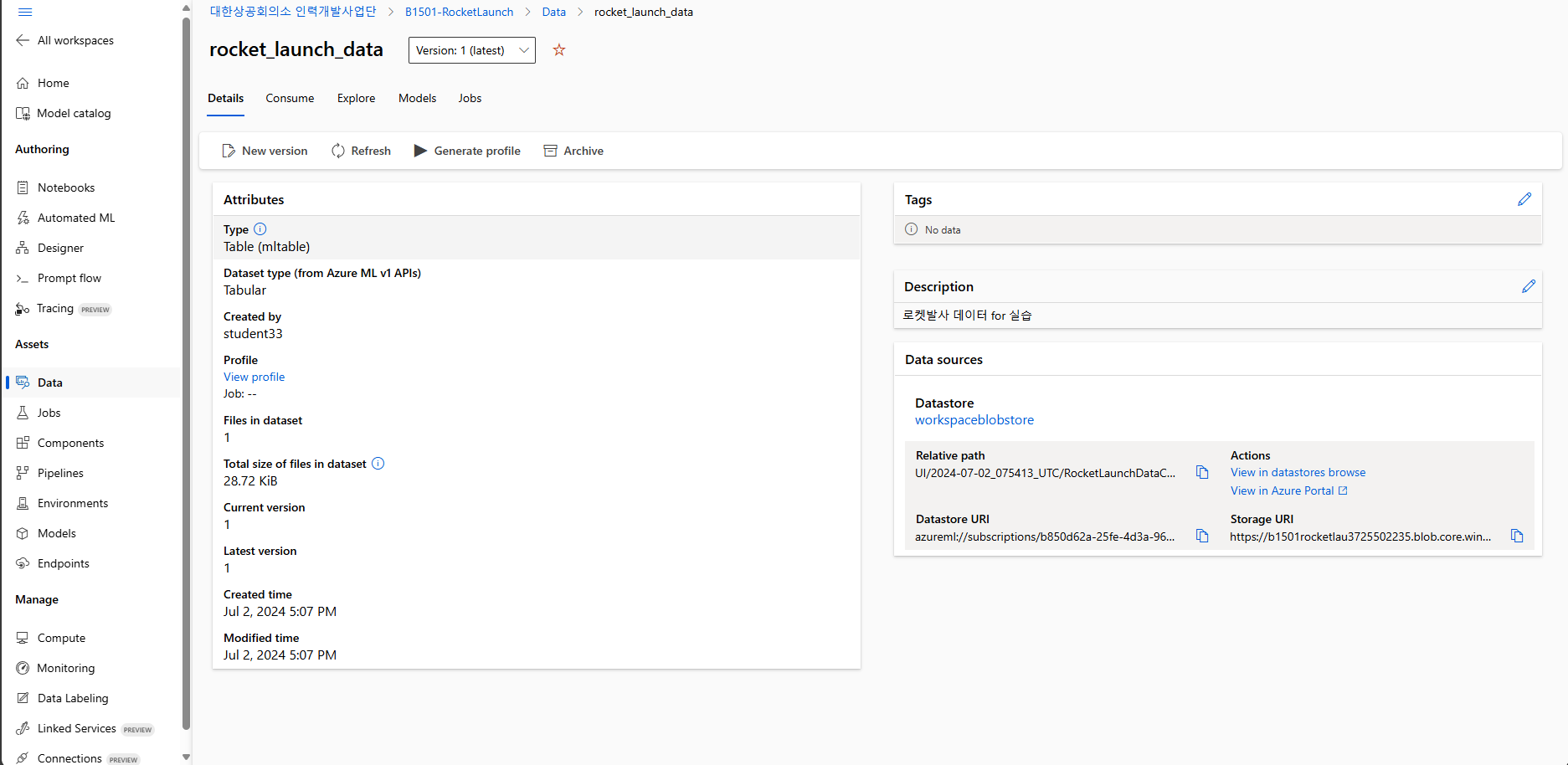


소수점으로 변경

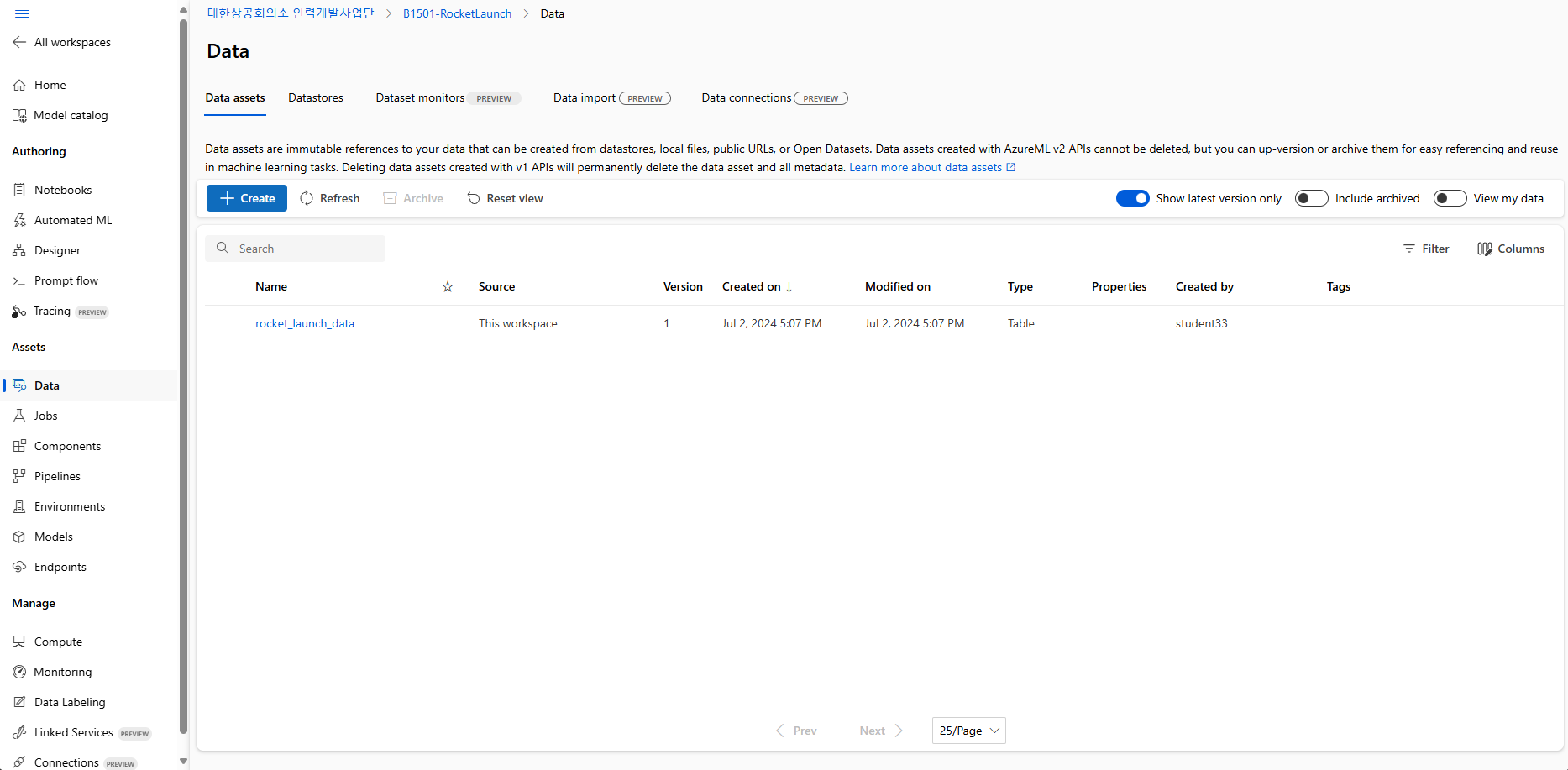




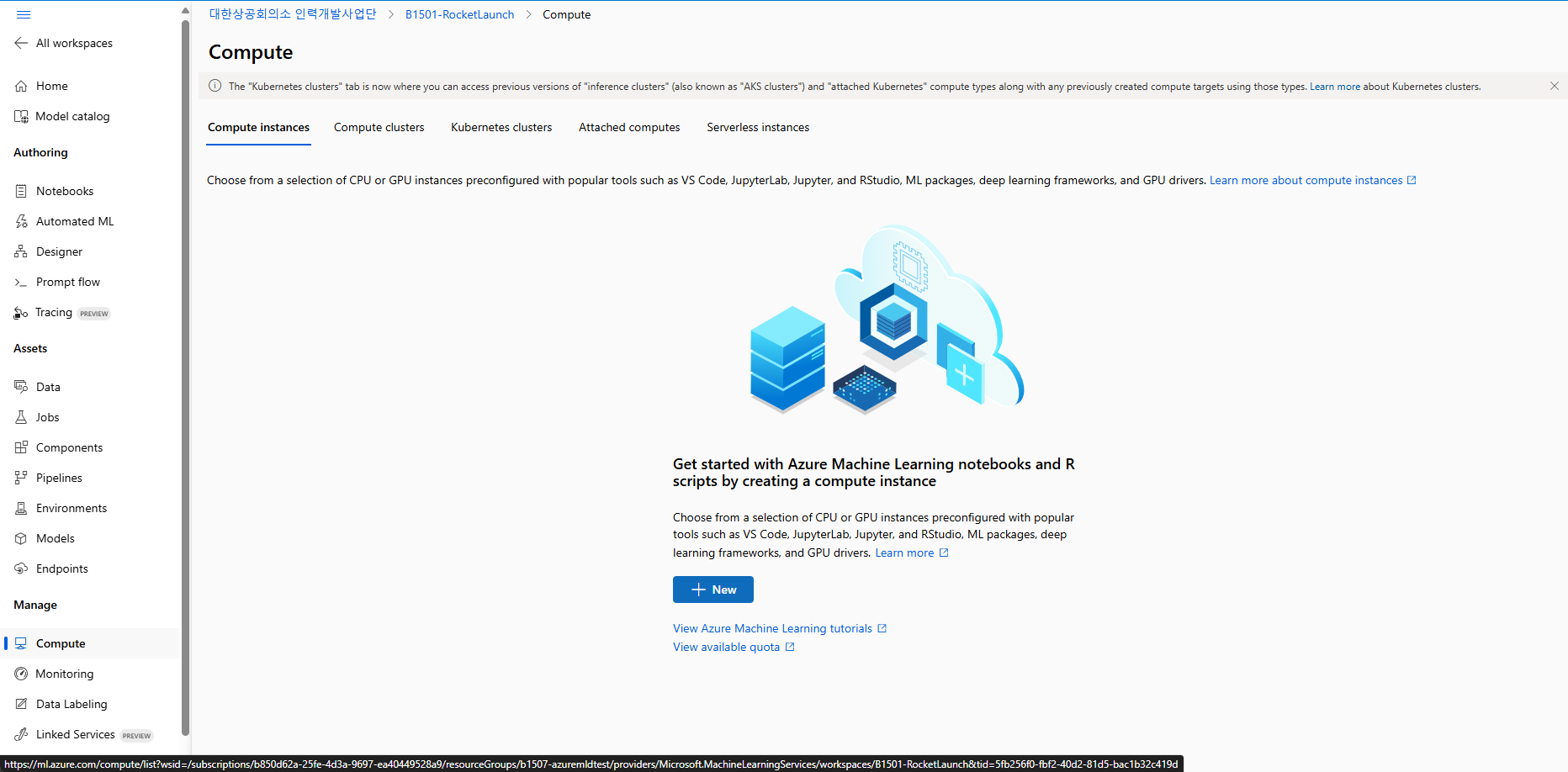
생성 된 설정들 확인



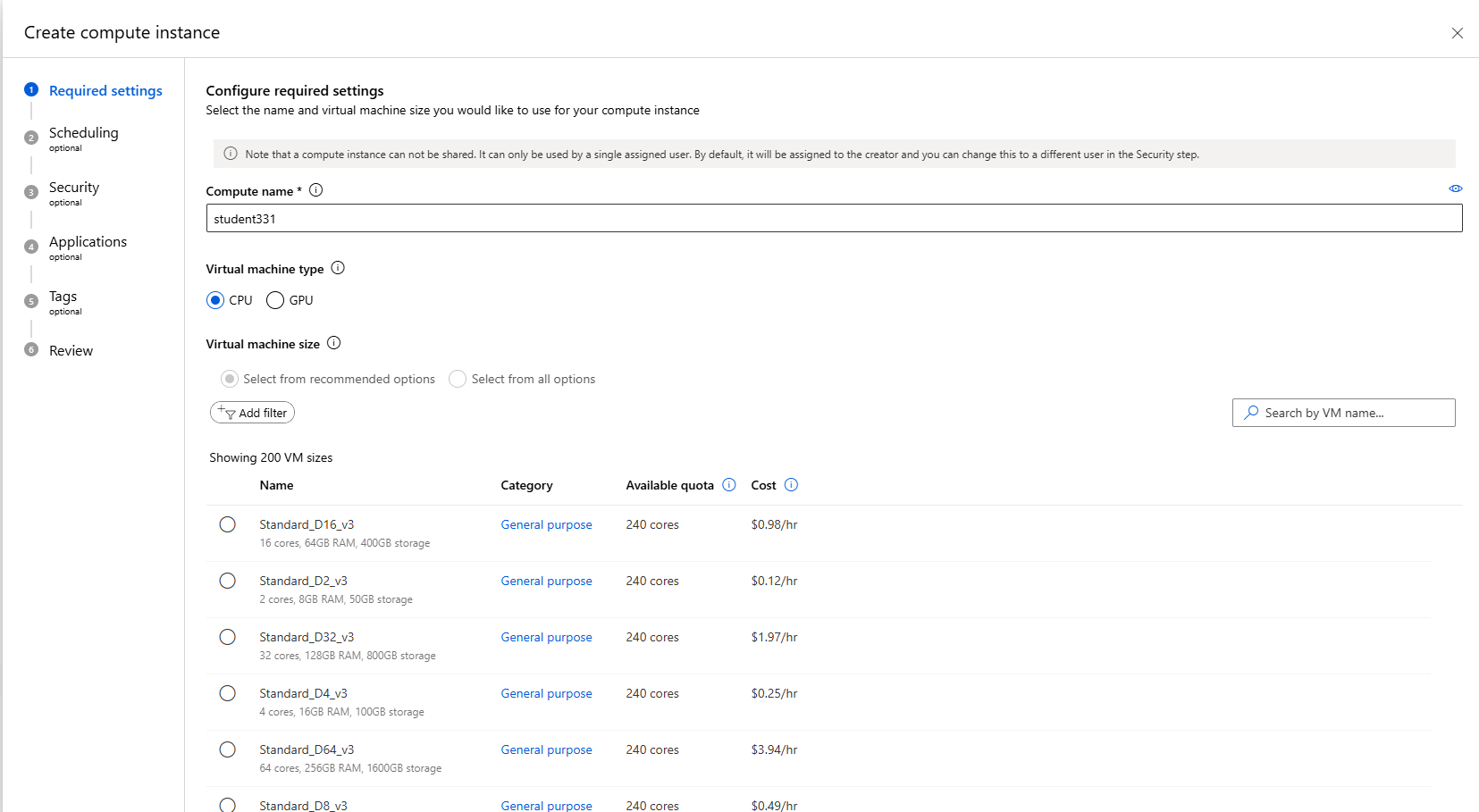
생성 확인



컴퓨터

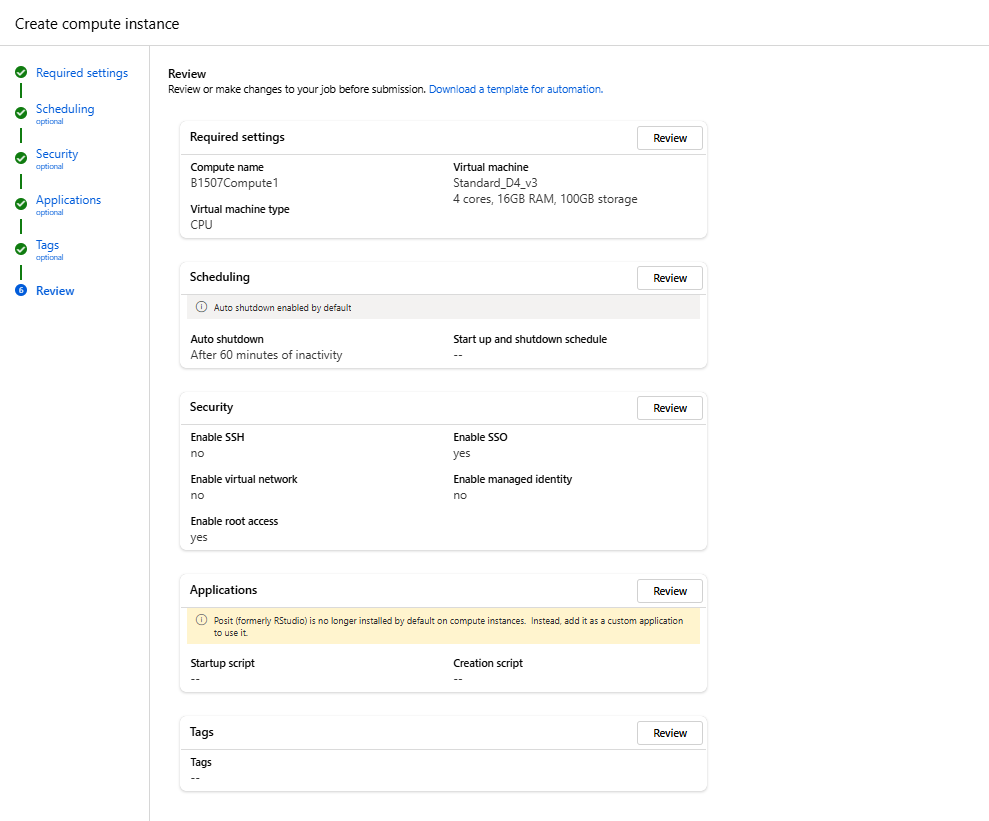


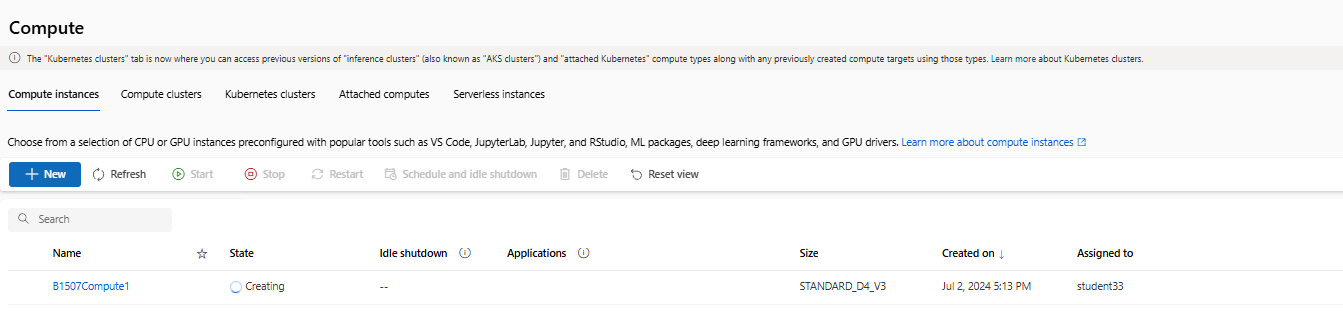
어떤 형식의 가상 컴퓨터를 만들지 설정



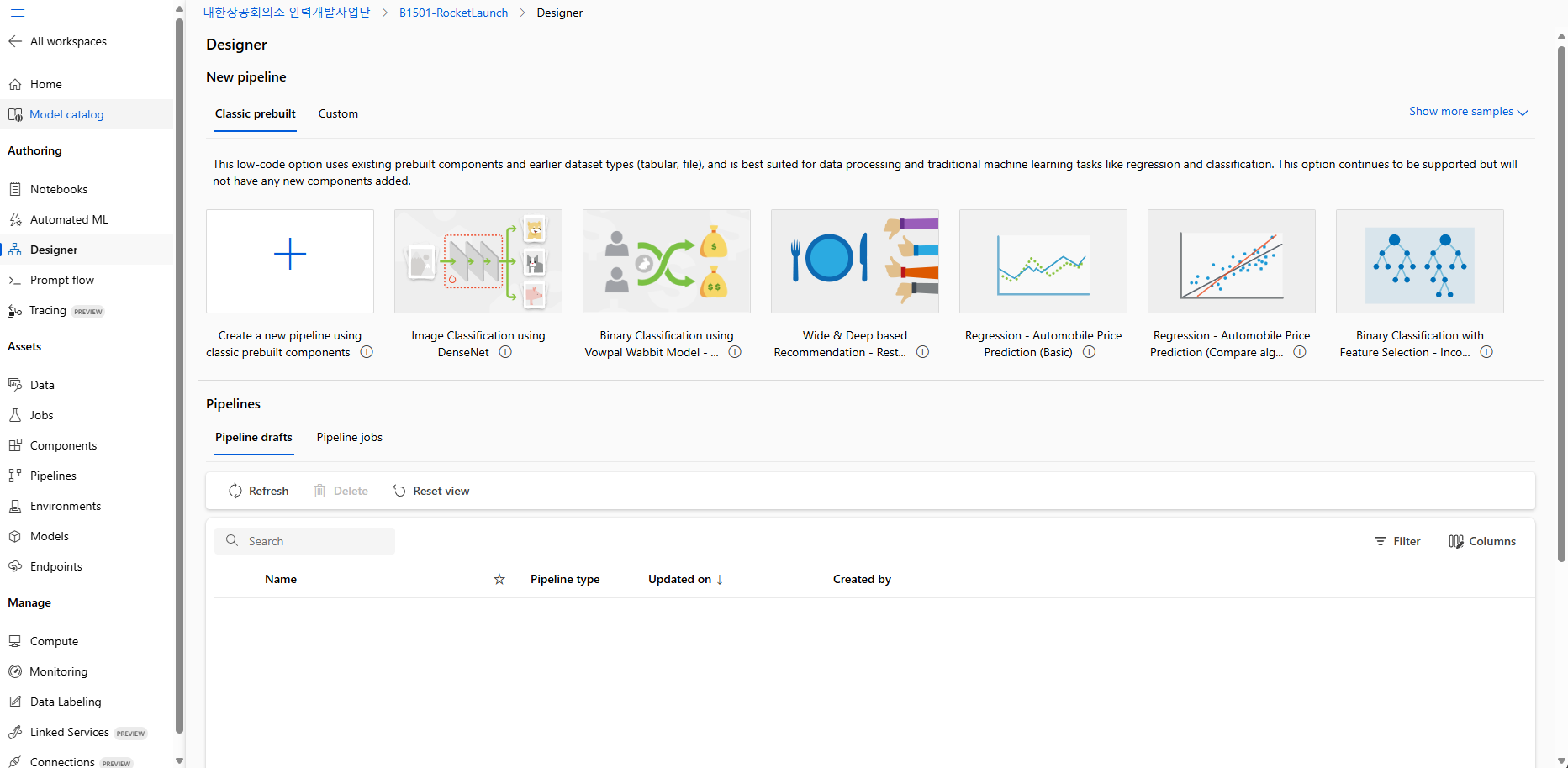
4코어 생성



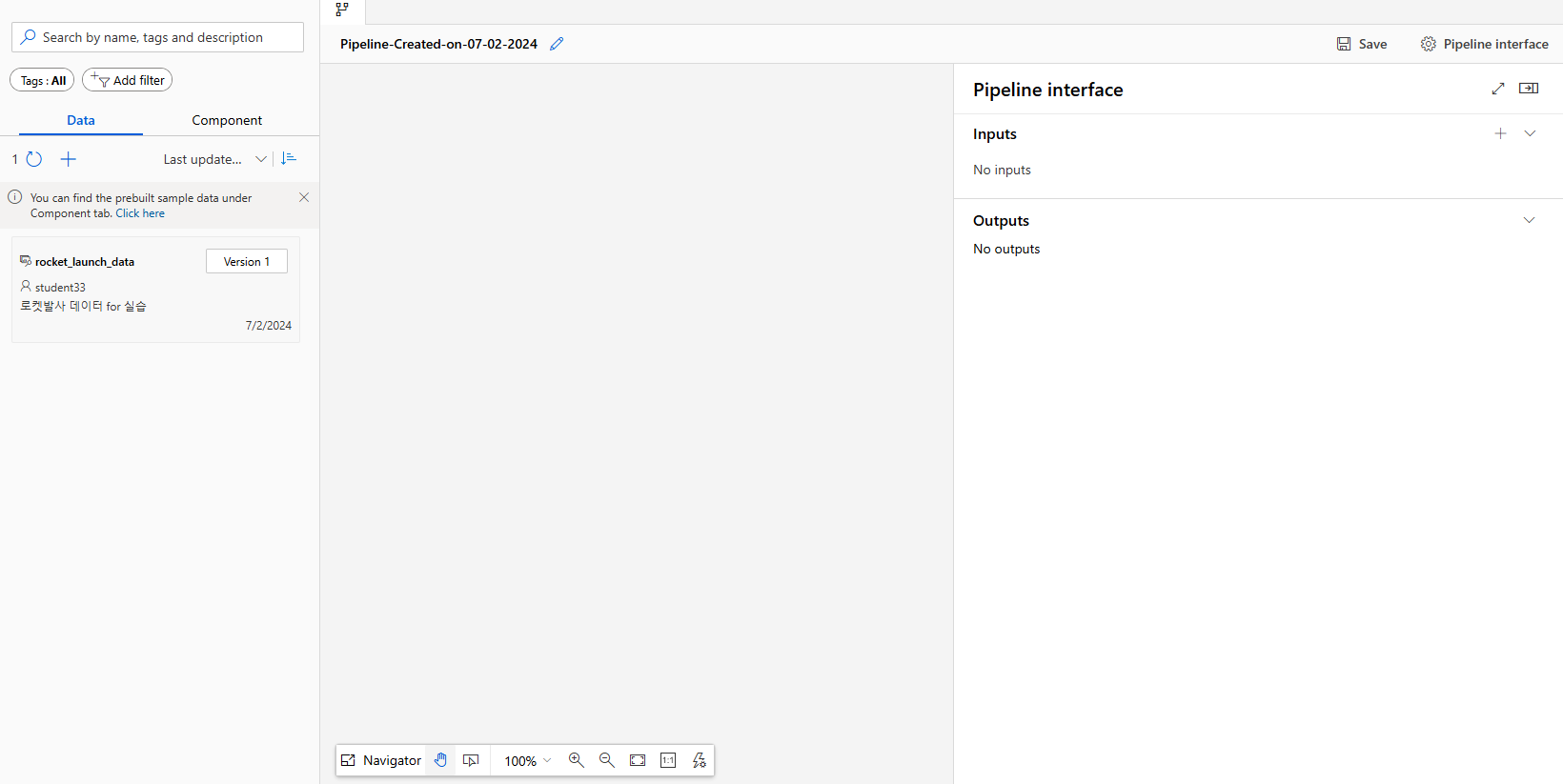




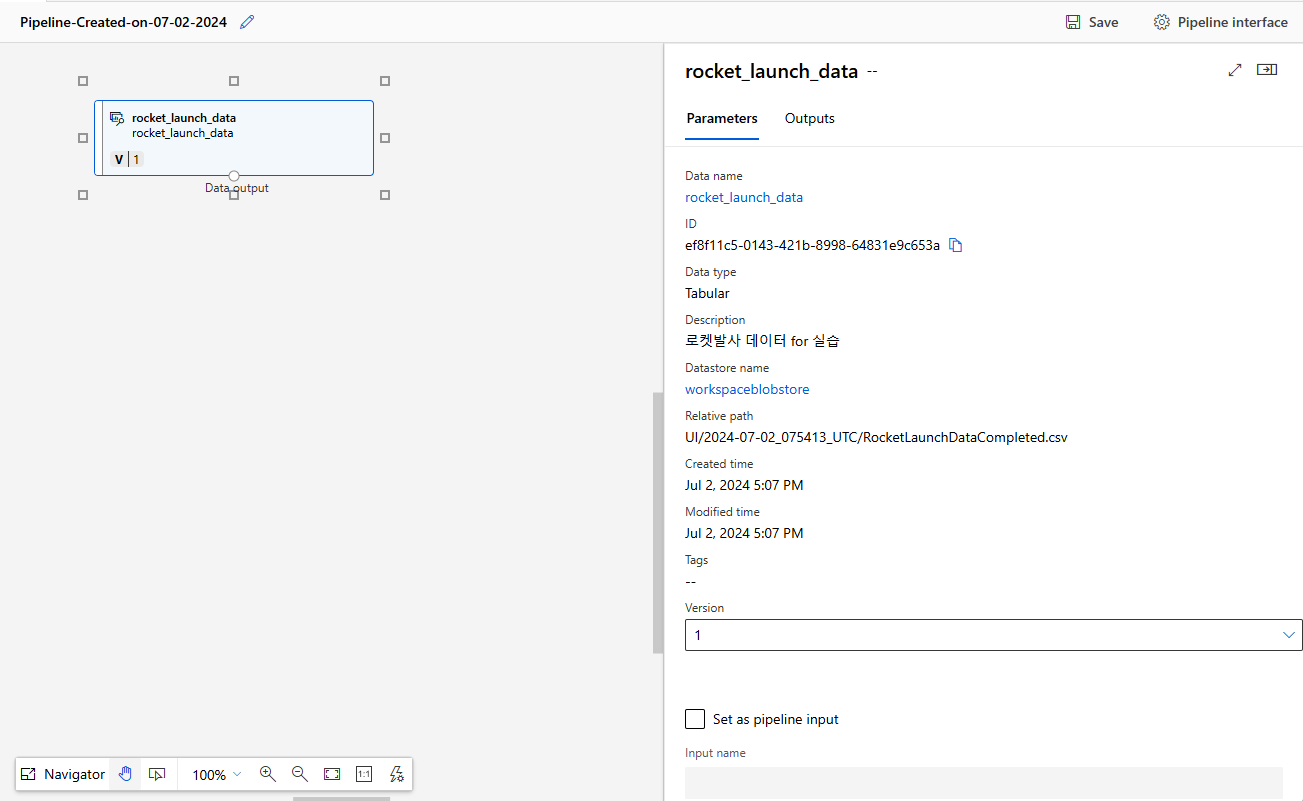
디자인



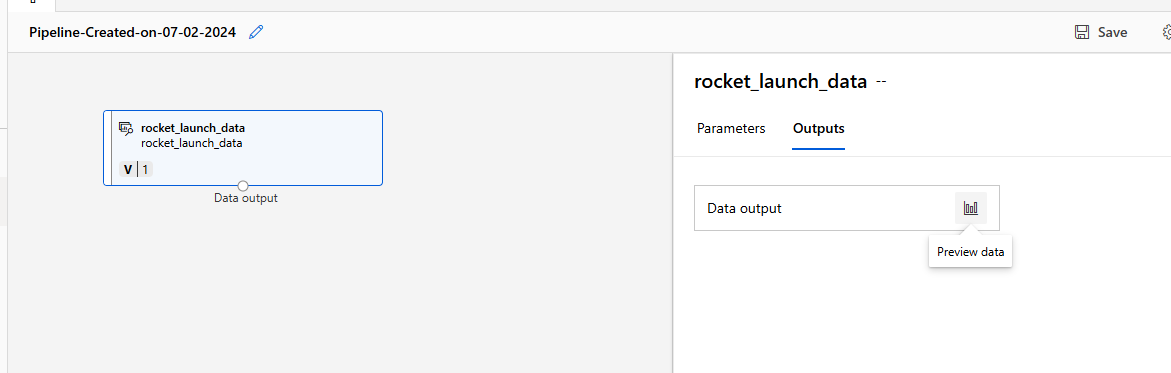
새로운 화면으로 생성

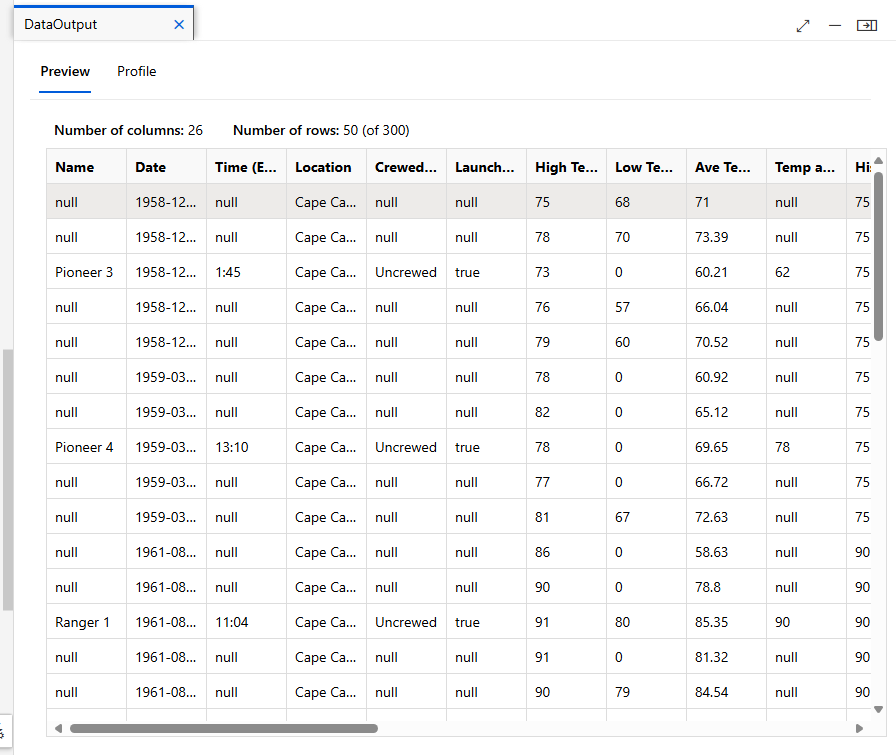


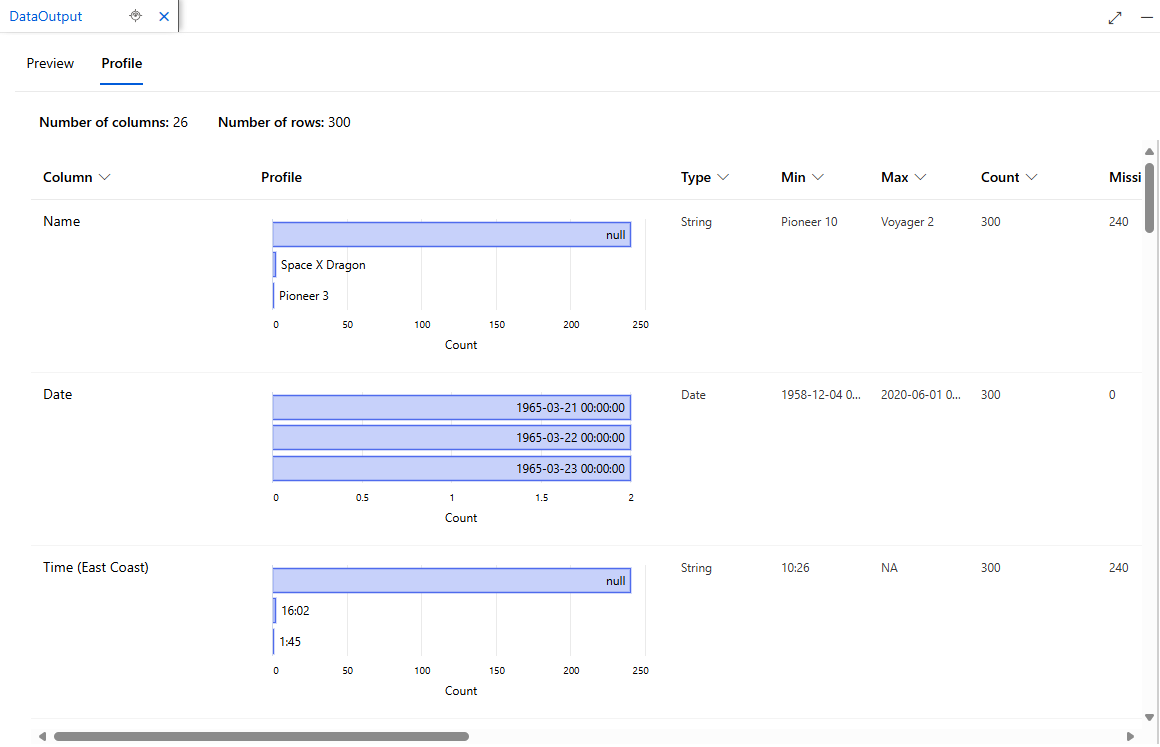
좌측에서 드래그 하여 유닛 놓기



Outputs 에서 그래프 모양 클릭

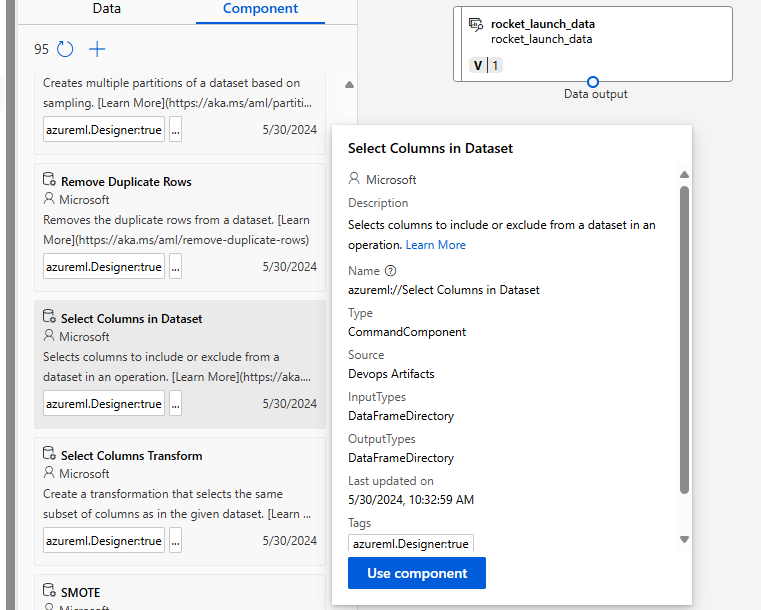


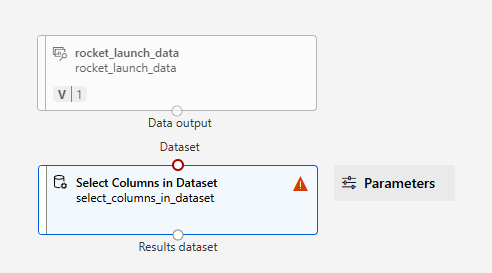




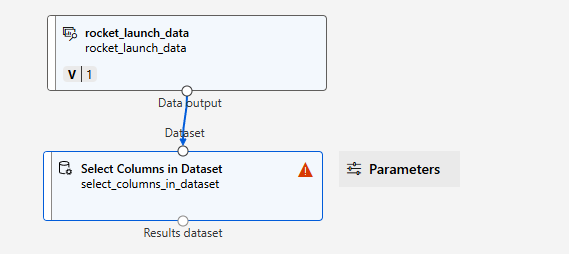
Data transformation에서 select columns in dataset 선택



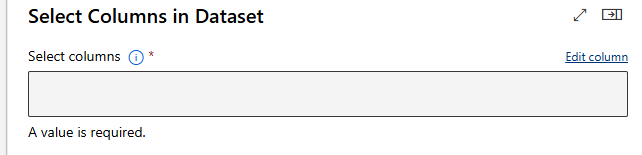




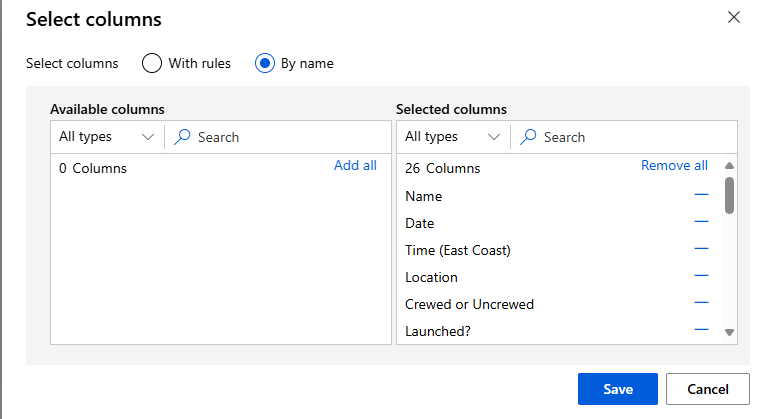
선 연결



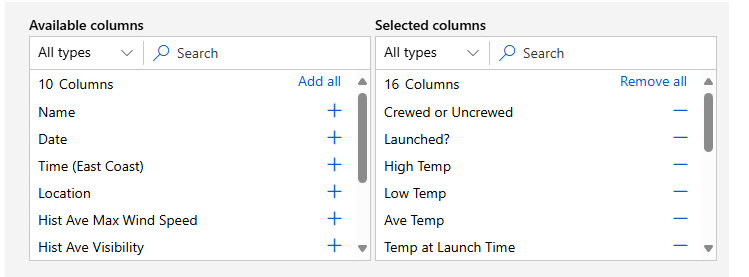
Edit column 선택

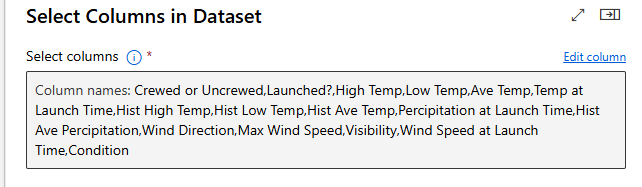


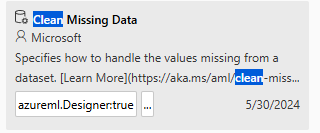
By name에서 Add all 선택

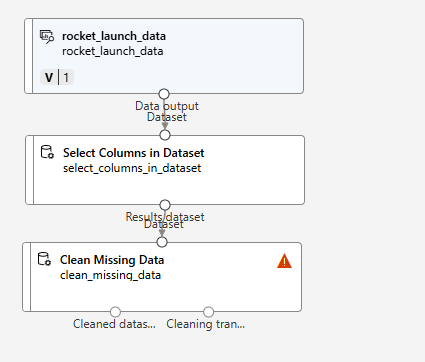


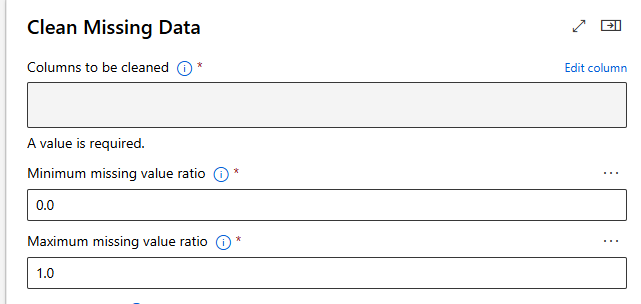
필요없는 것 선택해서 빼내기

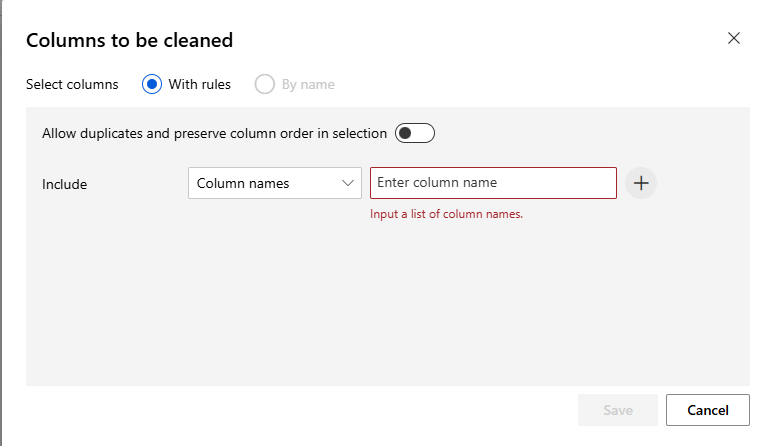


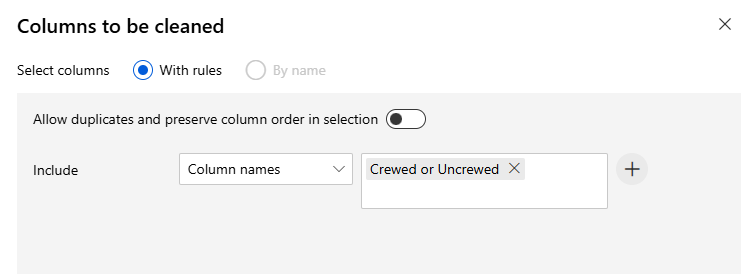


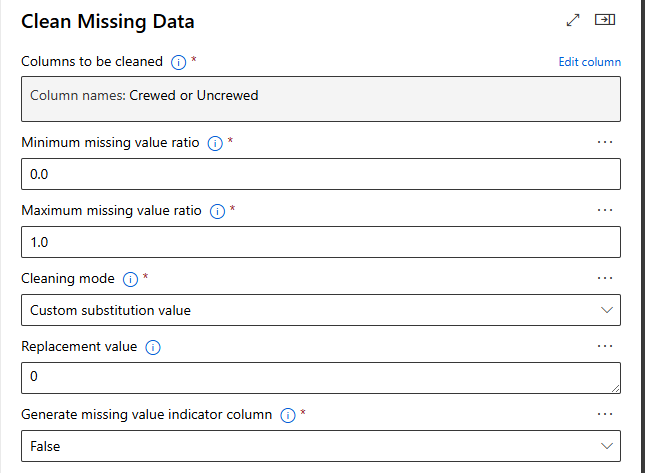


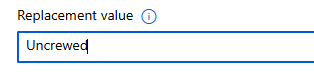


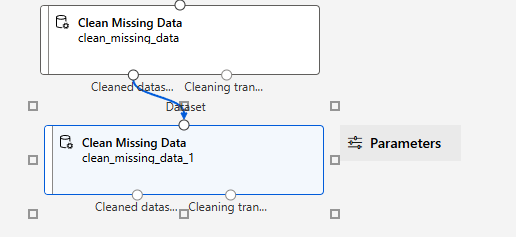


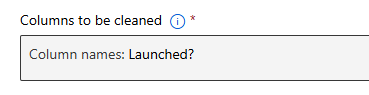




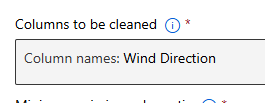
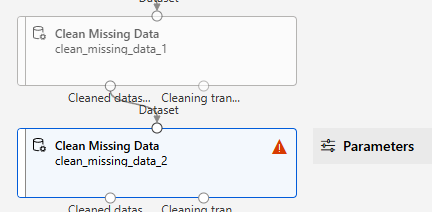


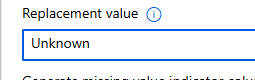


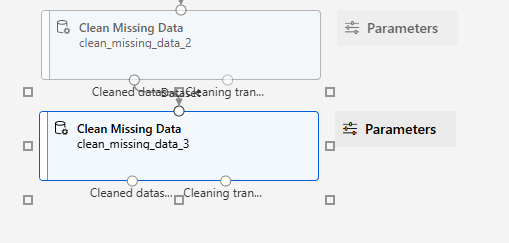


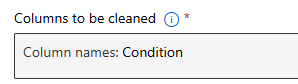


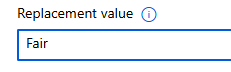


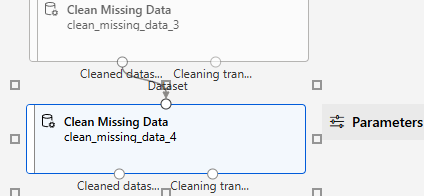


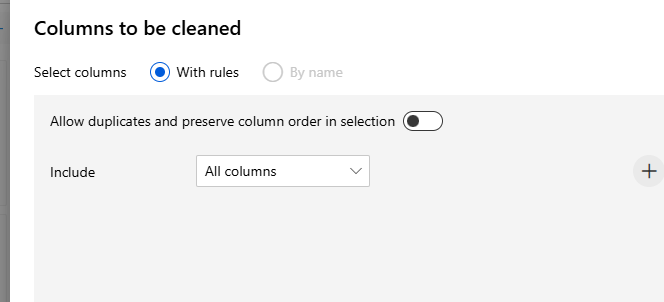




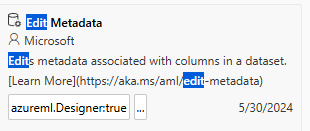




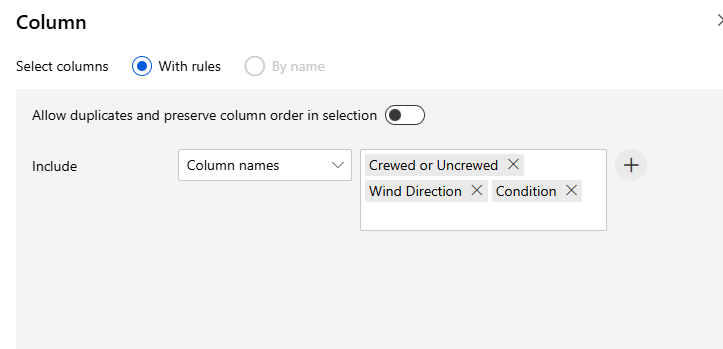


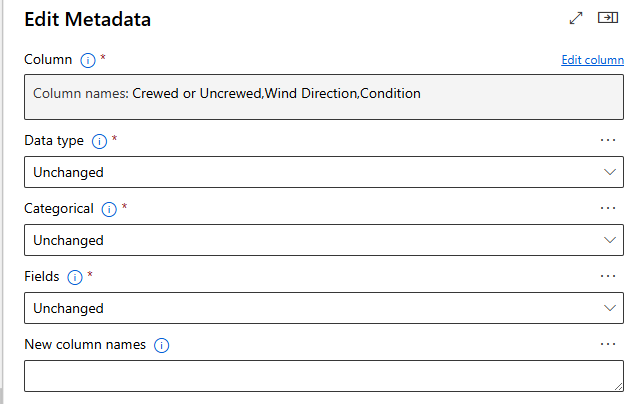


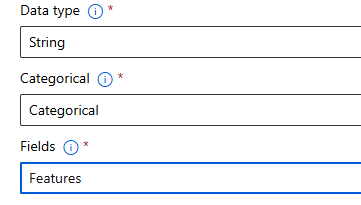


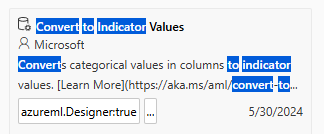


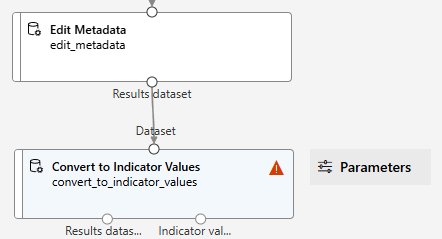


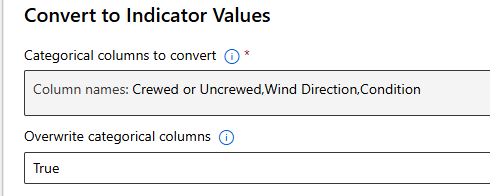










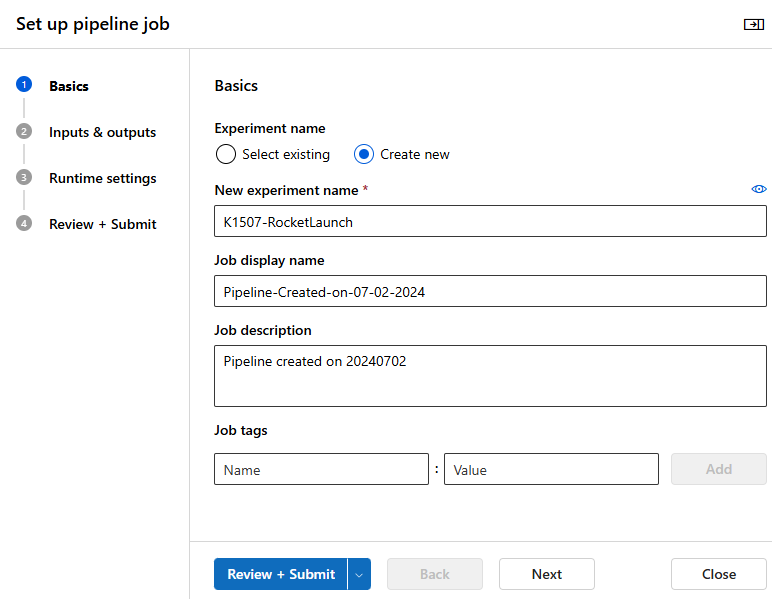
  
Save 클릭

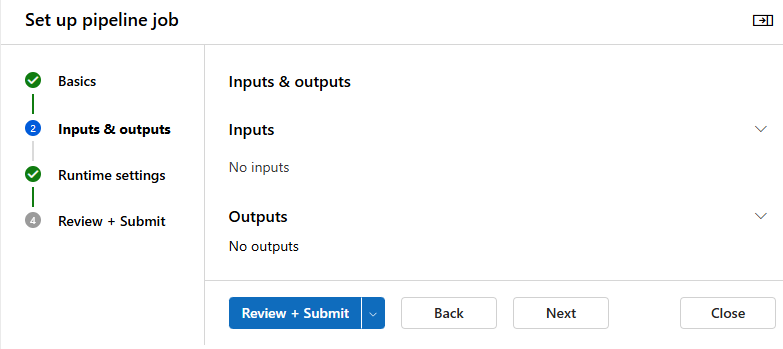


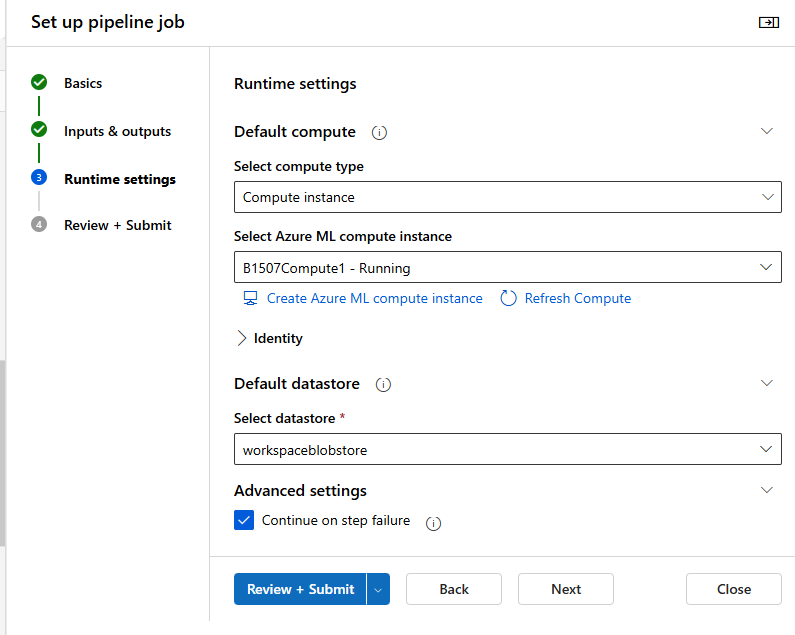
Configure & Submit 선택

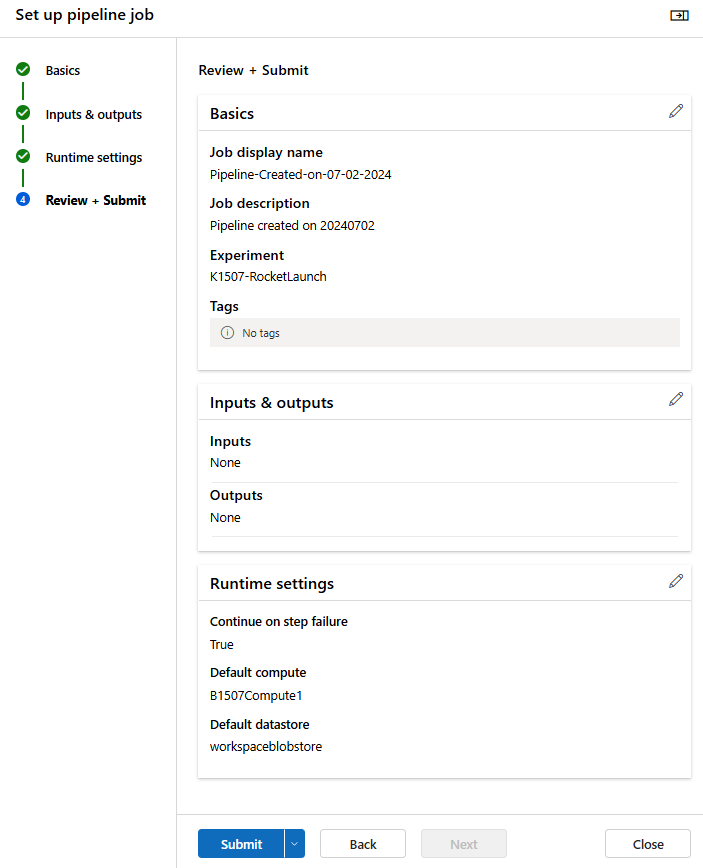


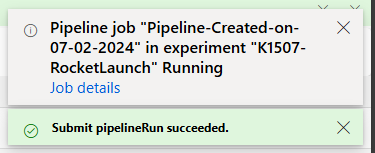
처음 생성하는 것이니 Create new 선택 후 설정

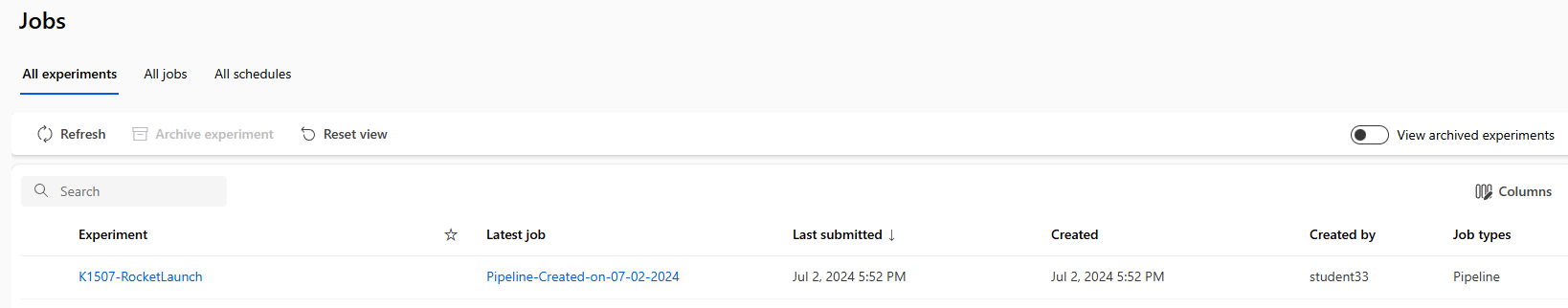






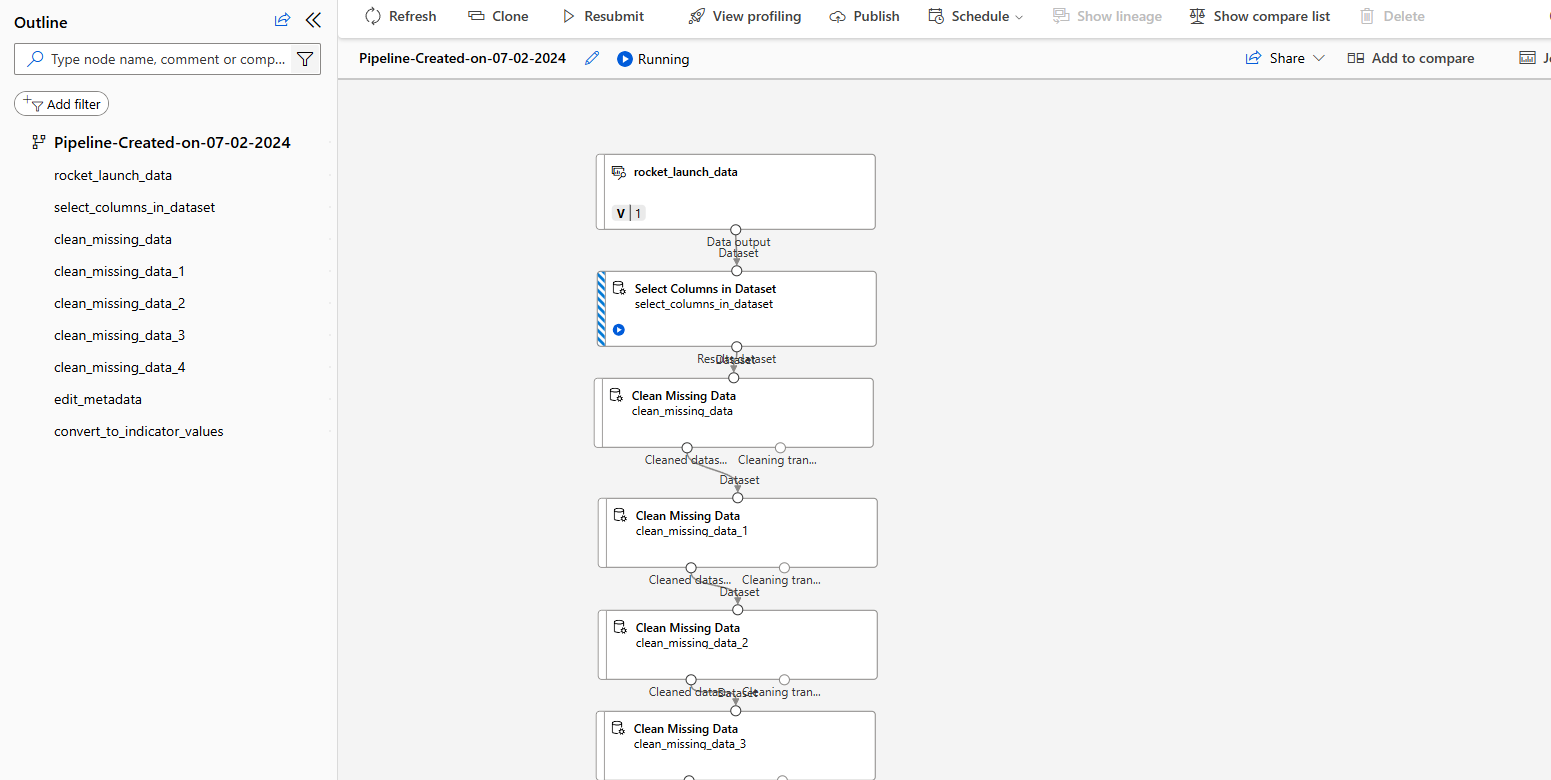


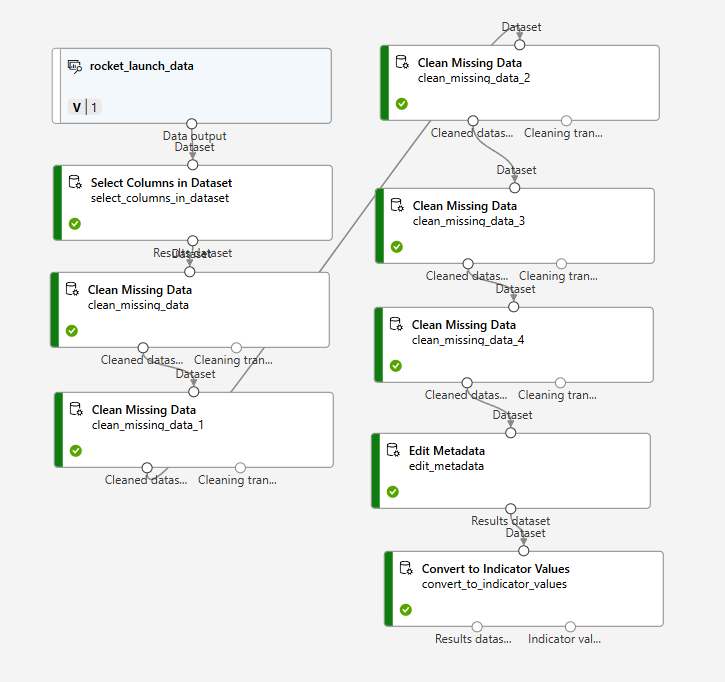




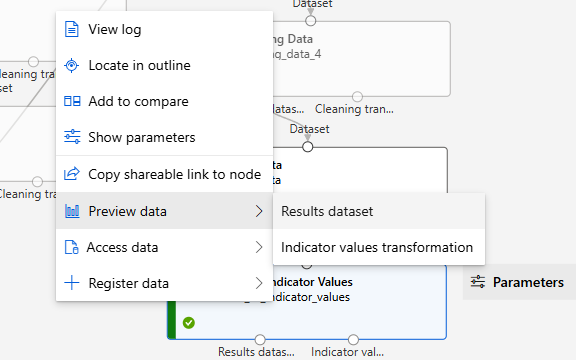
초록/파란 빗금은 현재 진행 중인 상태

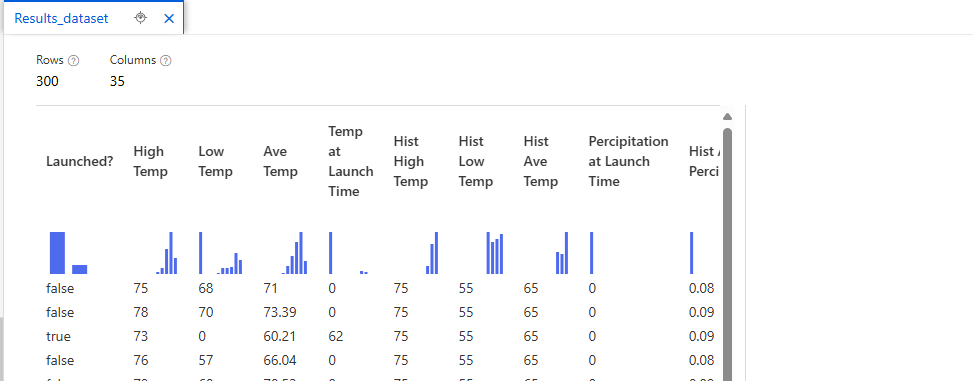
초록은 오류 없이 완료

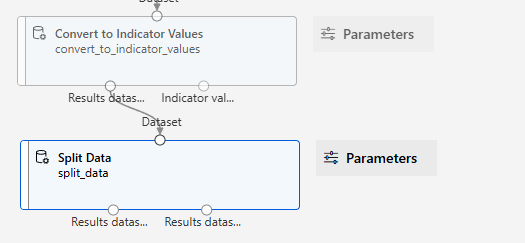


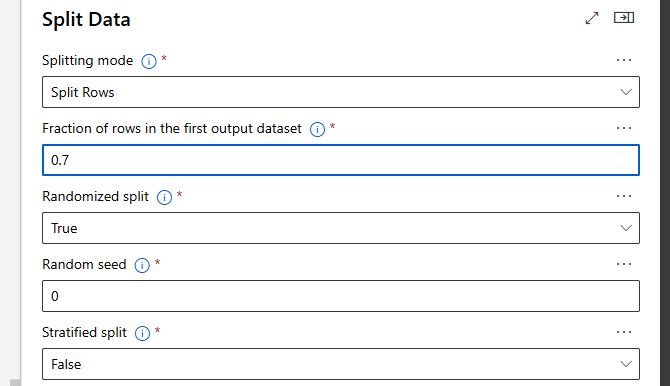


우클릭 후 preview data -> results dataset 선택

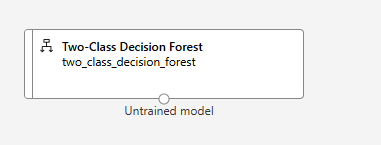


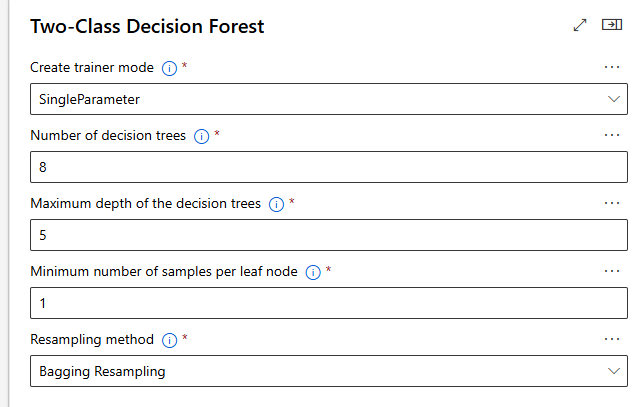


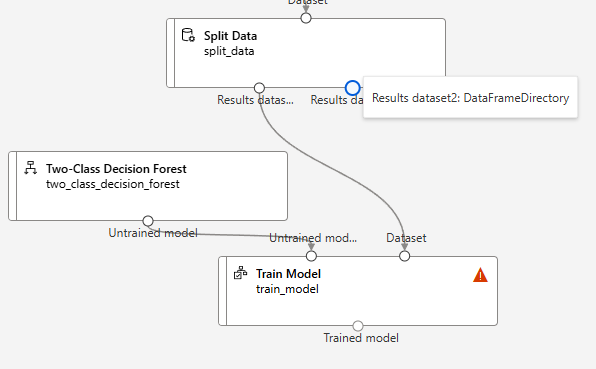




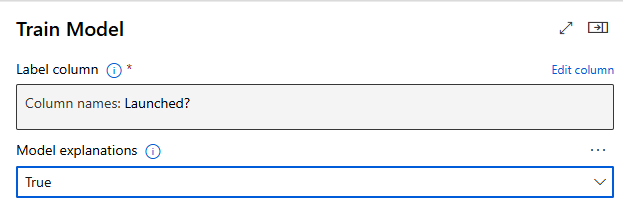
랜덤 포레스트를 사용하기 위해 배치

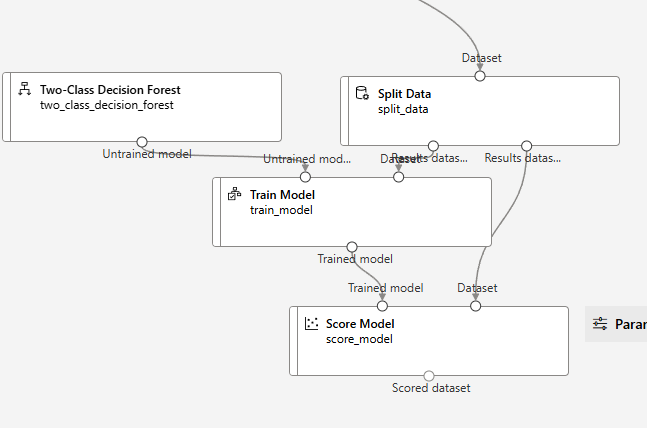


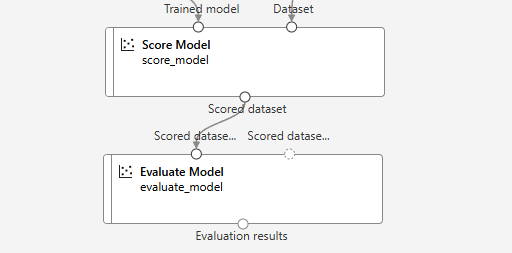




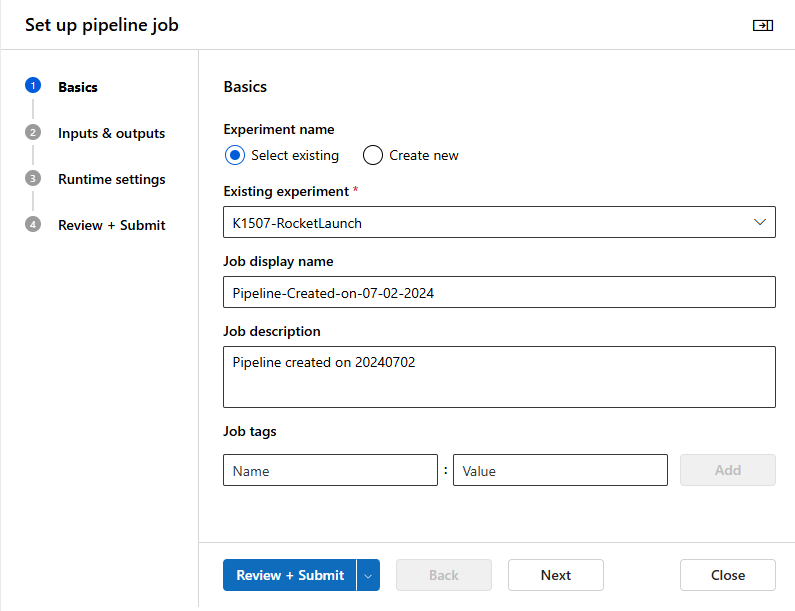
두 가지 모듈을 이어주기 위해 배치

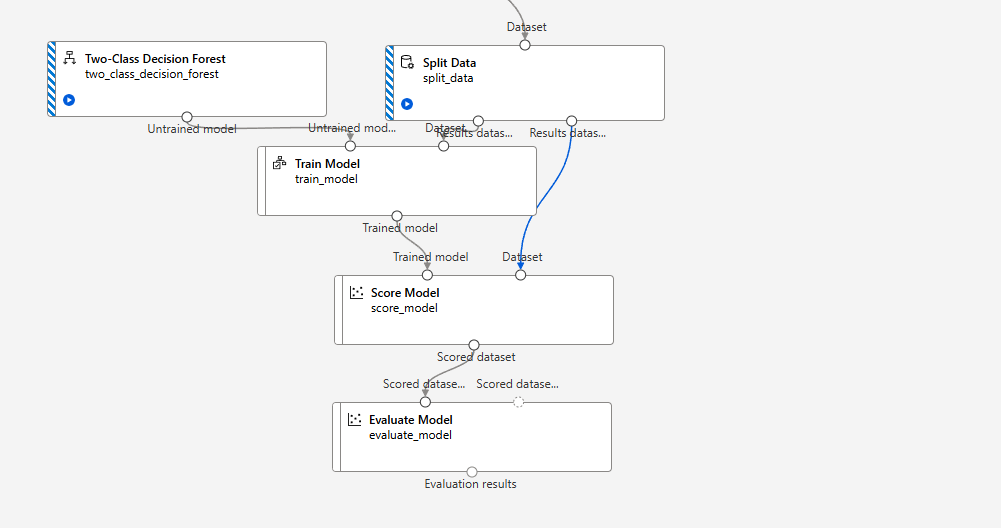




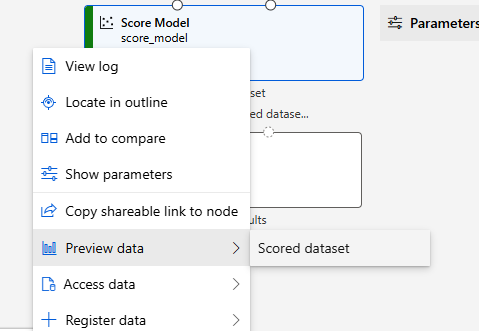


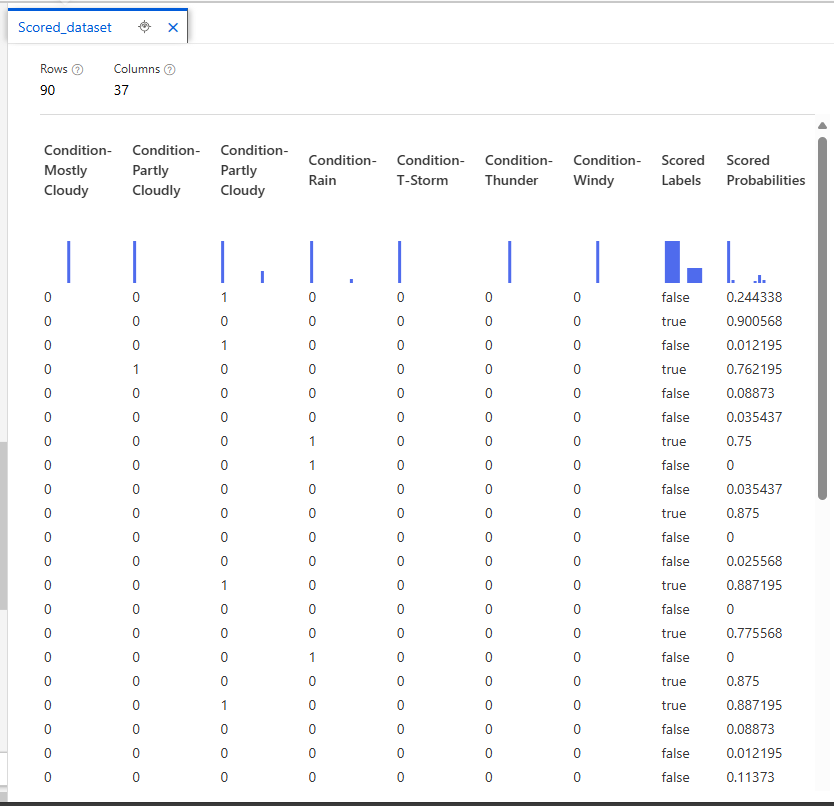
재설정 / 이번에는 만들었던 것에 다시 업로드 하는 형식이기에 Select existing로 실행





Scored labels -> Preview data -> Scored dataset 눌러 확인 가능





Evaluate Model -> Preview data -> Evaluation results 눌러 정보 확인

