



# *MLOps 101*

## *Episode 2: ML 생애주기 (1) 데이터 준비*

한석진  
마이크로소프트

---

## Episode 2

### ML 생애주기 (1)

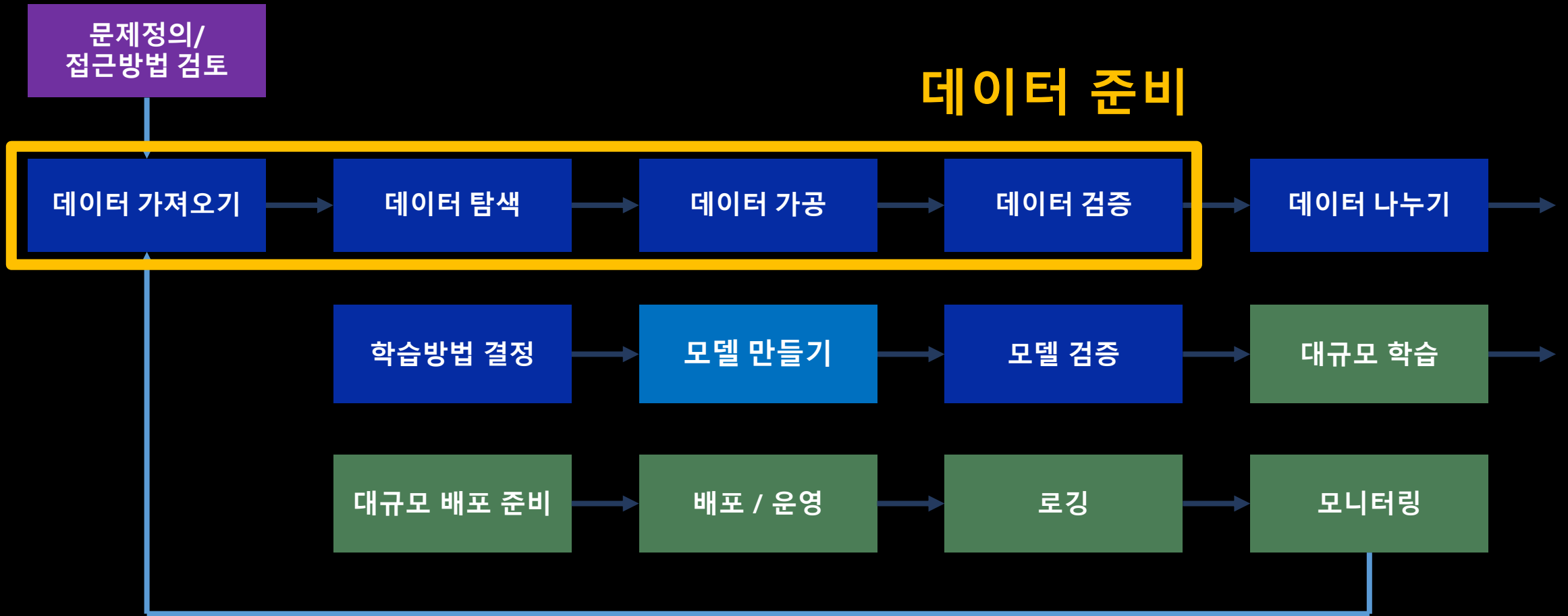
#### 데이터 준비

---

#### ML 생애주기 (1) 데이터 준비

- 문제/데이터 정의, 가설 수립
- 데이터 이동/연계
  - 데이터셋 공유 및 재사용 *DEMO*
- 데이터 탐색/가공
  - 데이터셋 및 주피터 노트북에서 탐색 *DEMO*
  - 데이터 레이블링 *DEMO*
  - Feature Importance 탐색 *DEMO*

# ML 생애주기



# 문제/데이터 정의, 가설 수립

문제정의/  
접근방법 검토

풀고 싶은 문제 정의 (제대로 된 질문하기)

가설

데이터 확보

데이터셋	확보방안

데이터셋	확보방안

데이터셋	확보방안

데이터셋	확보방안

데이터셋	확보방안

비즈니스 임팩트 +  
고려할 사항

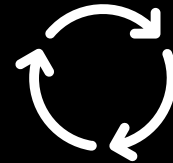
# 데이터 이동/연계

## 데이터 확보

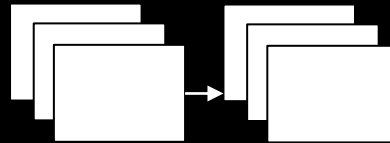
데이터셋	확보방안



## 데이터 수집

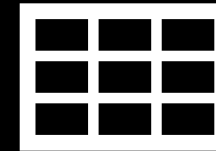


실시간



배치

## 데이터 연계



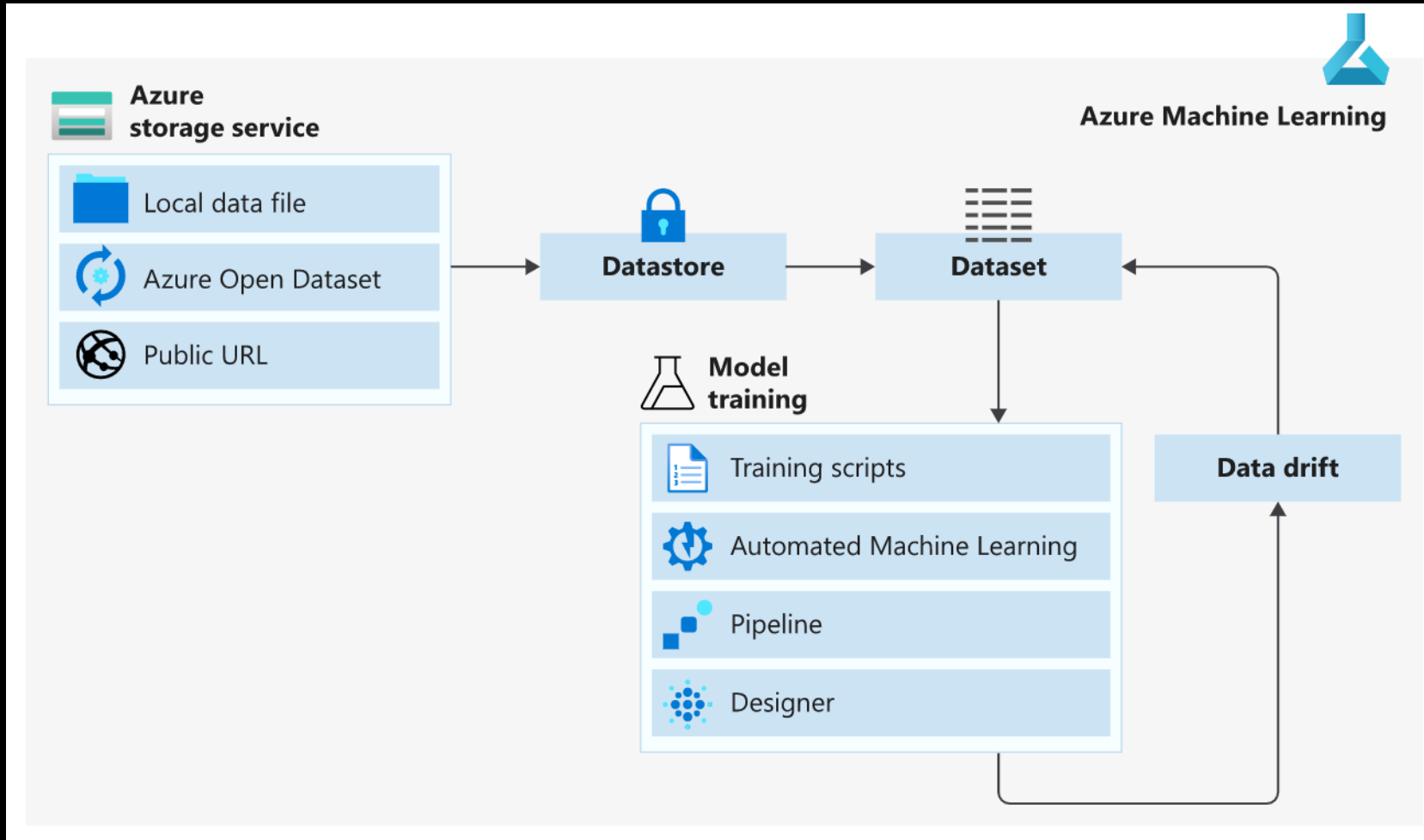
JOIN  
VLOOKUP



아버지 / NNG  
가 / JKS  
방 / NNG  
에 / JKB  
들어가 / VV  
신다 / EP+EC

갑툰튀,0,0,0,NNG,\*F,갑툰튀,\*\*\*\*\*  
강퇴,0,0,0,NNG,\*F,강퇴,\*\*\*\*\*  
개드립,0,0,0,NNG,\*T,개드립,\*\*\*\*\*  
갠소,0,0,0,NNG,\*F,갠소,\*\*\*\*\*  
고퀄,0,0,0,NNG,\*T,고퀄,\*\*\*\*\*  
광삭,0,0,0,NNG,\*T,광삭,\*\*\*\*\*  
광탈,0,0,0,NNG,\*T,광탈,\*\*\*\*\*

# 데이터셋 공유 및 재사용



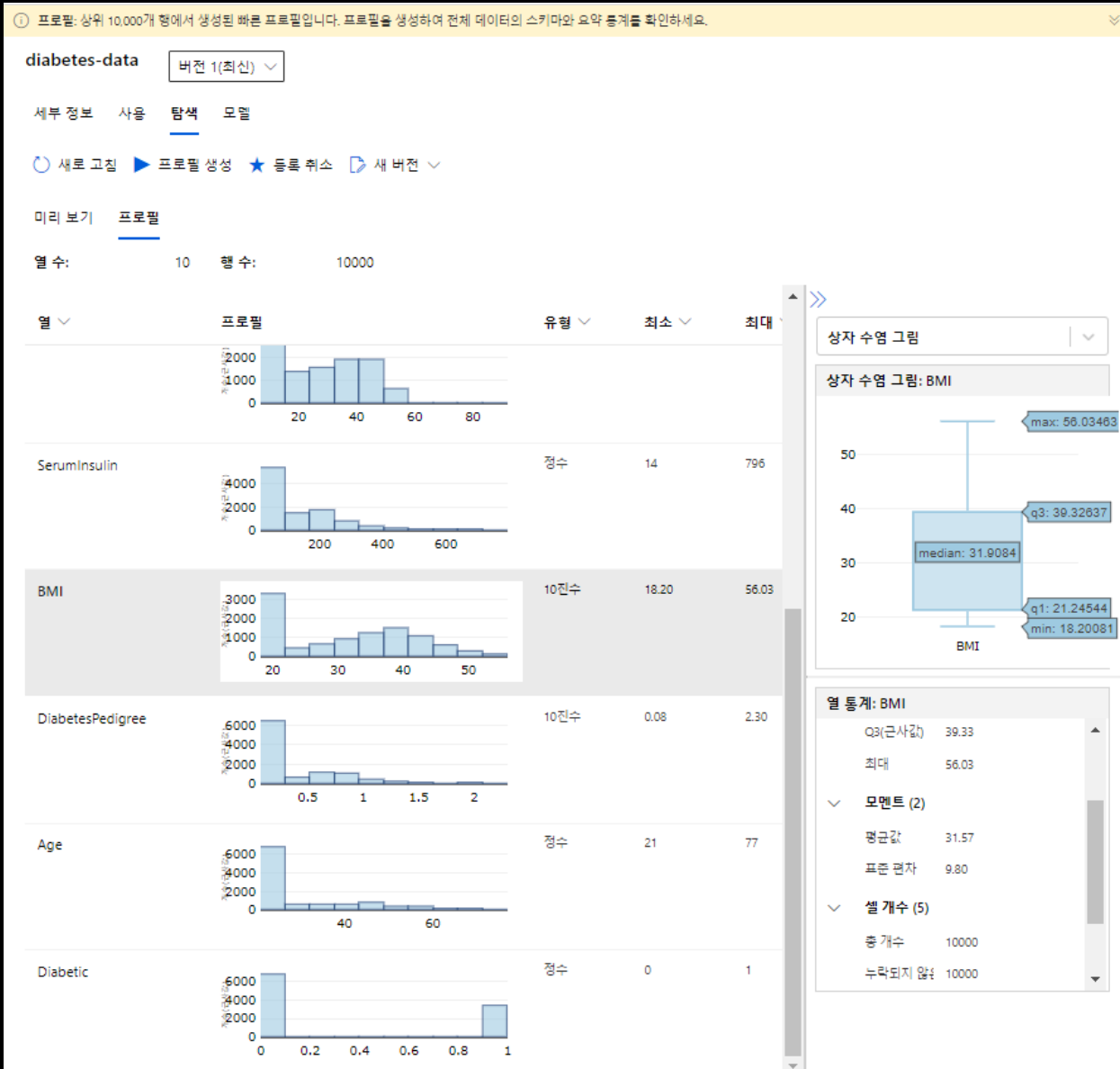
## 데이터셋 공유 및 재사용

- 버전관리
- 누적되는 데이터 중 특정 범위를 논리적으로 관리
- 어떤 실험에서 사용됐는지 연결 관리

## Data Drift

- 모니터링 단계에서 활용

# 데이터 탐색/가공



실행 157 실행 중

새로 고침 취소

세부 정보 데이터 가드 레일 모델 출력 + 로그 자식 실행 스냅샷

자동 기능화를 사용하도록 설정하면 자동화된 ML에서 데이터 보호책이 실행됩니다. 데이터 보호책은 입력 데이터를 대상으로 수행하는 검사 시퀀스로서 모델을 학습시키는 데 고품질 데이터가 사용되는지를 확인합니다.

유형	상태	설명
클래스 균형 검색	통과	입력이 분석되었으며, 학습 데이터에서 모든 클래스가 균형 상태입니다. <a href="#">불균형 데이터에 대해 자세히 알아보세요.</a>

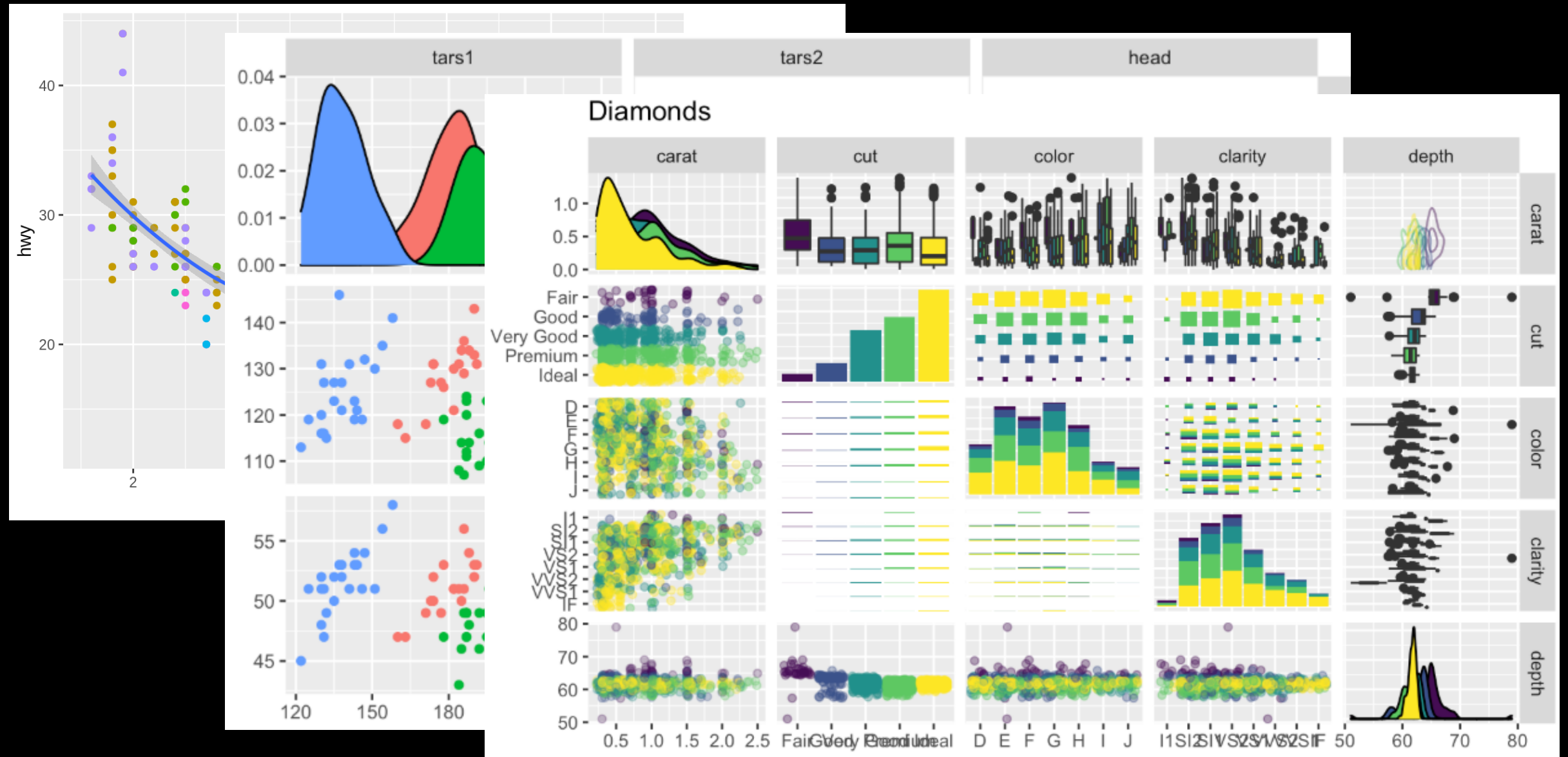
유형	상태	설명
기능 값 대체 누락	완료	학습 데이터에서 기능 값 누락이 감지되어 대체되었습니다. 값 누락이 예상되는 경우 실행이 완료되도록 합니다. 그렇지 않으면 현재 실행을 취소하고 스크립트를 사용하여 데이터 형식 및 비즈니스 요구 사항에 따라 더 적합할 수 있는 기능 값 누락 처리를 사용자 지정합니다. <a href="#">값 대체 누락에 대해 자세히 알아보세요.</a>

+ 추가 세부 정보 보기

유형	상태	설명
높은 카디널리티 기능 검색	완료	높은 카디널리티 기능이 입력에서 검색되고 처리되었습니다. <a href="#">높은 카디널리티 기능 검색에 대해 자세히 알아보세요.</a>

+ 추가 세부 정보 보기

# 데이터 탐색/가공







# 데이터 레이블링

All Projects > Photo labels (Multilabel)

## Photo labels (Multilabel)

Instructions Tasks

☐ Select all (2 selected)

Tags

- ☒ Land
- ☐ Ocean
- ☐ Closeup
- ☒ Wideangle

Closeup >

Wideangle

Submit

All Projects > Photo Subject ID (Object Identification)

## Photo Subject ID (Object Identification)

Instructions Tasks













image-segmentation-cats-dogs

내보내기 레이블 데이터 새로 고침

대시보드 데이터 세부 정보

데이터 세트 미리 보기

레이블이 지정된 데이터

미디어	레이블
	 cat
	 dog dog
	 dog
	 dog
	 cat
	 cat cat cat

naver-sentiment-labeling

내보내기 레이블 데이터 새로 고침

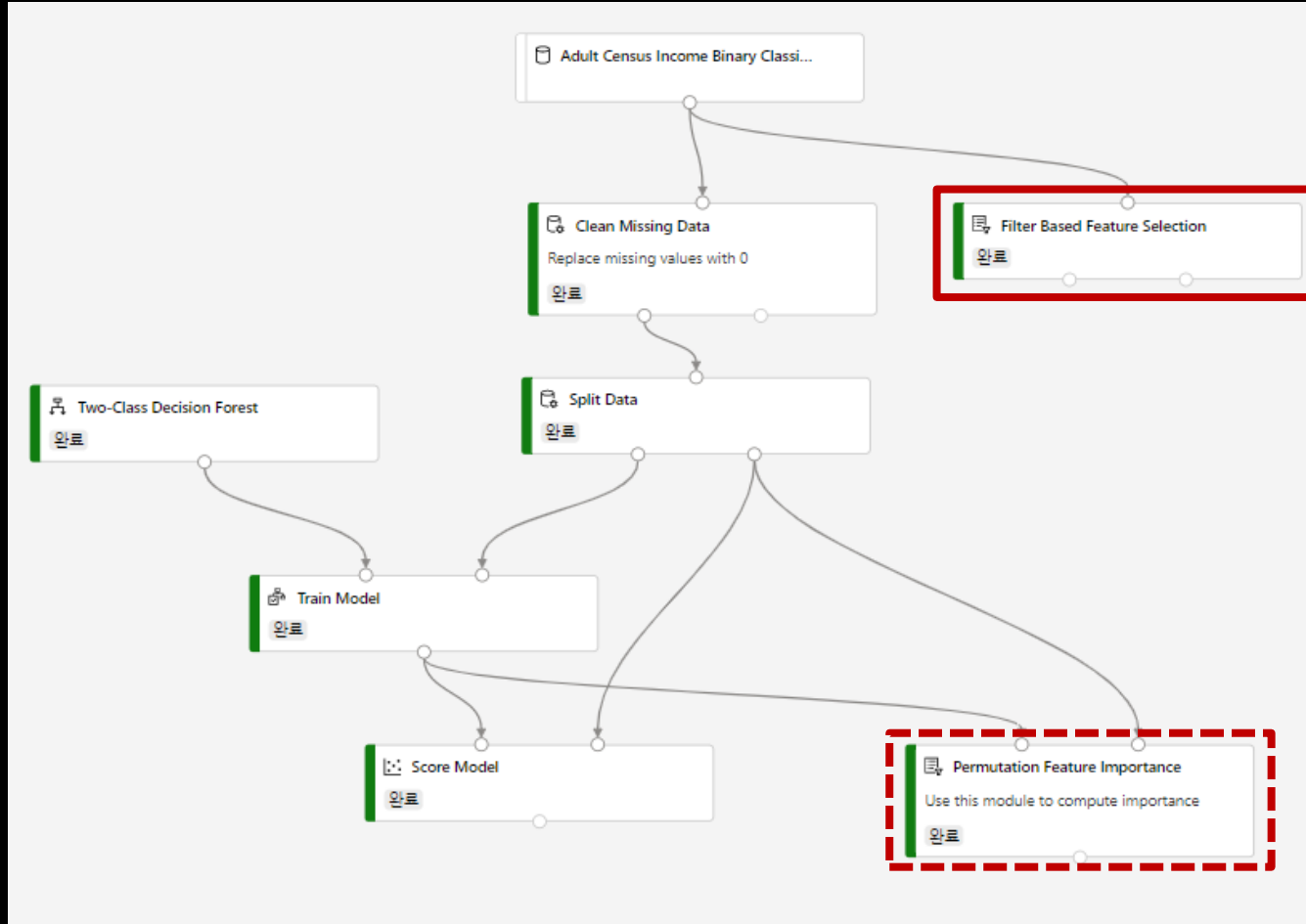
대시보드 데이터 세부 정보

데이터 세트 미리 보기

레이블이 지정된 데이터

텍스트	레이블
전 좋아요	positive
영화가 사람의 영혼을 어루만져 줄 수도 있군요 거친 세상사를 잠시 잊고 동화같은 영화에 행복...	positive
너무재밌었다그래서보는것을추천한다	positive
한국영화 흥행코드: 갈등-갈등-계~예속 갈등-확해-감동- 명점 10점 남발- 흥행 편하지 뭐...	negative
액션이 없는데도 재미 있는 몇안되는 영화	positive
출...포스터보고 초딩영화줄...오버연기조자 가법지 않구나	negative
내일이 기대되는`	positive
클라라블라고화신본거아닌데	negative

# Feature Importance 탐색



### Filter Based Feature Selection

매개 변수    출력 + 로그    세부 정보    메트릭    자식 실행    이미지    스냅샷

☒ Operate on feature columns only

Target column ② \*

열 이름: income

Number of desired features ② \*

5

Feature scoring method ② \*

PearsonCorrelation

출력 설정

## Filter Based Feature Selection 결과 시각화

Features	Filtered dataset					
형 ②	열 ②					
1	15					
income	education-num	age	hours-per-week	capital-gain	capital-loss	fnlwgt
1	0.335154	0.234037	0.229689	0.223329	0.150526	0.009463

---

## Episode 2

### ML 생애주기 (1)

### 데이터 준비

---

#### ML 생애주기 (1) 데이터 준비

- 문제/데이터 정의, 가설 수립
- 데이터 이동/연계
  - 데이터셋 공유 및 재사용 *DEMO*
- 데이터 탐색/가공
  - 데이터셋 및 주피터 노트북에서 탐색 *DEMO*
  - 데이터 레이블링 *DEMO*
  - Feature Importance 탐색 *DEMO*

# {다음 시간에는}

---

## Episode 3 ML 생애주기 (2) 실험/학습

---

### ML 생애주기 (2) 실험/학습

- 실험, 모델 학습/최적화/비교평가
- 실험 추적관리
  - 데이터셋, 코드, 환경, 모델, 서빙 추적 *DEMO*
- 자동화된 ML (Automated ML)
  - 자동화된 ML 엿보기 *DEMO*
- 모델의 검증: 예측성능, 처리성능
  - 예측 성능 *DEMO*
  - 처리 성능 *DEMO*