### 학습 데이터셋

#### 개요

학습 데이터셋은 머신러닝 모델이 학습할 데이터의 집합입니다. 머신러닝 모델은 주어진 데이터셋의 패턴을 학습하고 유사한 데이터를 토대로 값이나 패턴을 예측하므로 학습 데이터셋의 품질이 머신러닝 모델의 결과를 좌우합니다.

고려사항

학습 데이터셋은 머신러닝 모델을 사용할 환경에서 생성되는 데이터와 최대한 동일한 것이 좋습니다. 데이터셋을 생성할 때 다음과 같은 사항들을 고려하세요.

**다양한 정보 포함**

학습 데이터셋은 머신러닝 모델을 활용할 환경에서 발생할 수 있는 다양한 특징(시간, 위치, 사용자 정부, 이벤트 유형, 성공/실패 등)이 포함된 데이터를 수집해서 준비하세요. 실제 환경에서 발생한 데이터를 활용하는 것이 좋습니다.

**지도학습 시 타겟 변수 선정**

랜덤포레스트와 같은 지도 학에 머신러닝 모델링은 예측 타겟 변수가 포함되어 있어야 합니다. 을 위한 학습 데이터셋 생성 시 타겟변수(학습예측 목표 값)를 포함하여 학습 데이터를 준비해야 합니다.

**데이터 전처리**

필요한 경우 데이터 오류, 극단적인 값 등을 제거하거나 조정하고, 특징값들이 큰 차이가 나지 않도록 정규화하거나 스케일링하세요.

**문자열의 벡터화**

일반 문자열은 머신러닝 모델이 학습하려면 수치화한 형태로 변환되어야 합니다. tfidf 쿼리 명령 등을 이용해 벡터화하세요.

**데이터의 품질과 양**

데이터는 정확성, 완전성, 일관성이 있어야 하며, 충분한 양의 데이터가 있어야 다양한 패턴을 학습할 수 있습니다. 특히 비정상 이벤트의 사례를 충분히 포함하도록 준비하세요.

**과적합 방지**

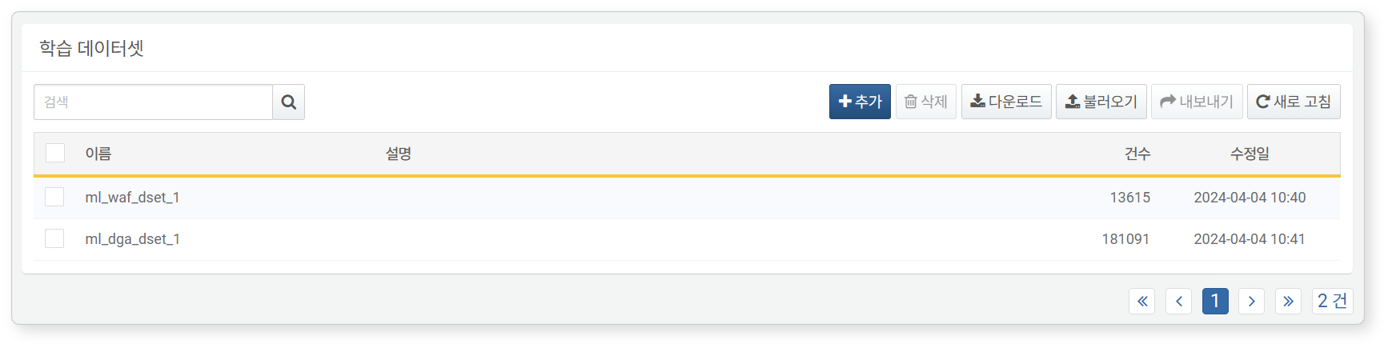
훈련 데이터에 너무 특화되지 않도록 과적합(overfitting)을 방지하기 위한 기법(교차 검증, 정규화 등)을 적용하세요.

**변경되는 패턴**

보안 위협 패턴은 시간이 지나면서 변화할 수 있으므로 머신러닝 모델을 주기적으로 재훈련해야 합니다.

#### 학습 데이터셋 목록 조회/검색

**정책 > 학습 데이터셋**에서 학습 데이터셋 목록을 조회하거나 검색할 수 있습니다.



* **이름**: 학습 데이터셋 이름
* **설명**: 학습 데이터셋에 대한 부가적인 정보
* **건수**: 현재 학습 데이터셋에 포함된 데이터의 개수
* **수정일**: 학습 데이터셋이 생성된 날짜, 또는 마지막으로 수정된 날짜

학습 데이터셋 목록에서 특정 학습 데이터셋을 찾으려면 도구 모음에 있는 검색 도구를 사용하세요. 검색 도구는 입력한 단어가 **이름**, **설명**에 포함된 학습 데이터셋을 찾아서 보여줍니다. 검색 도구는 대소문자를 구분하지 않습니다.

목록 다운로드

학습 데이터셋 목록을 로컬 PC에 파일로 다운로드하려면 도구모음에서 **다운로드**를 클릭하여 원하는 파일 형태로 다운로드하세요.

목록 새로 고침

학습 데이터셋의 목록을 최신 정보로 조회하려면 도구모음에서 **새로 고침**을 클릭하세요.

내보내기/불러오기

생성한 학습 데이터셋을 파일로 내보내거나 불러올 수 있습니다. 생성한 학습 데이터셋을 백업하고 복원할 경우 사용할 수 있습니다.

학습 데이터셋을 내보내려면,

학습 데이터셋 목록에서 내보낼 학습 데이터셋 행의 체크박스를 선택하세요.

도구모음에서 **내보내기**를 클릭하세요.

**학습 데이터셋 내보내기** 대화상자에서 이름을 설정하고 **확인**을 클릭하세요.

머신러닝 모델을 불러오려면,

도구모음에서 **불러오기**를 클릭하세요.

**학습 데이터셋 불러오기** 대화상자에서 **파일선택**을 클릭한 후 기존에 내보내기로 저장했던 학습 데이터셋 파일을 선택하세요.

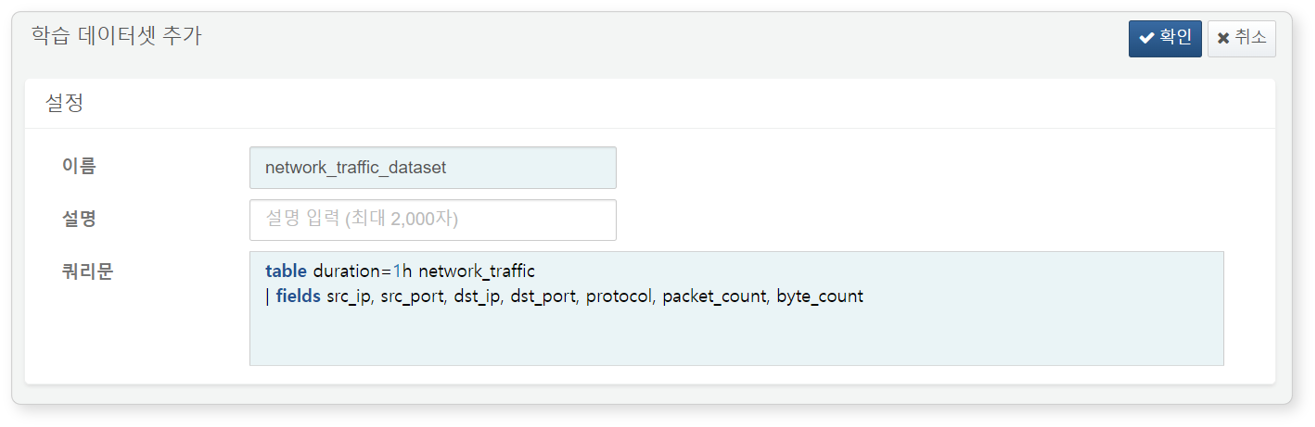
파일을 선택한 다음, **확인**을 클릭하세요.

#### 학습 데이터셋 추가

학습 데이터셋을 추가하려면,

**정책 > 학습 데이터셋**에서 도구 모음에 있는 **추가**를 클릭하세요.

**학습 데이터셋 추가** 화면에서 필요한 값을 입력하거나 선택한 다음, **확인**을 클릭하세요.

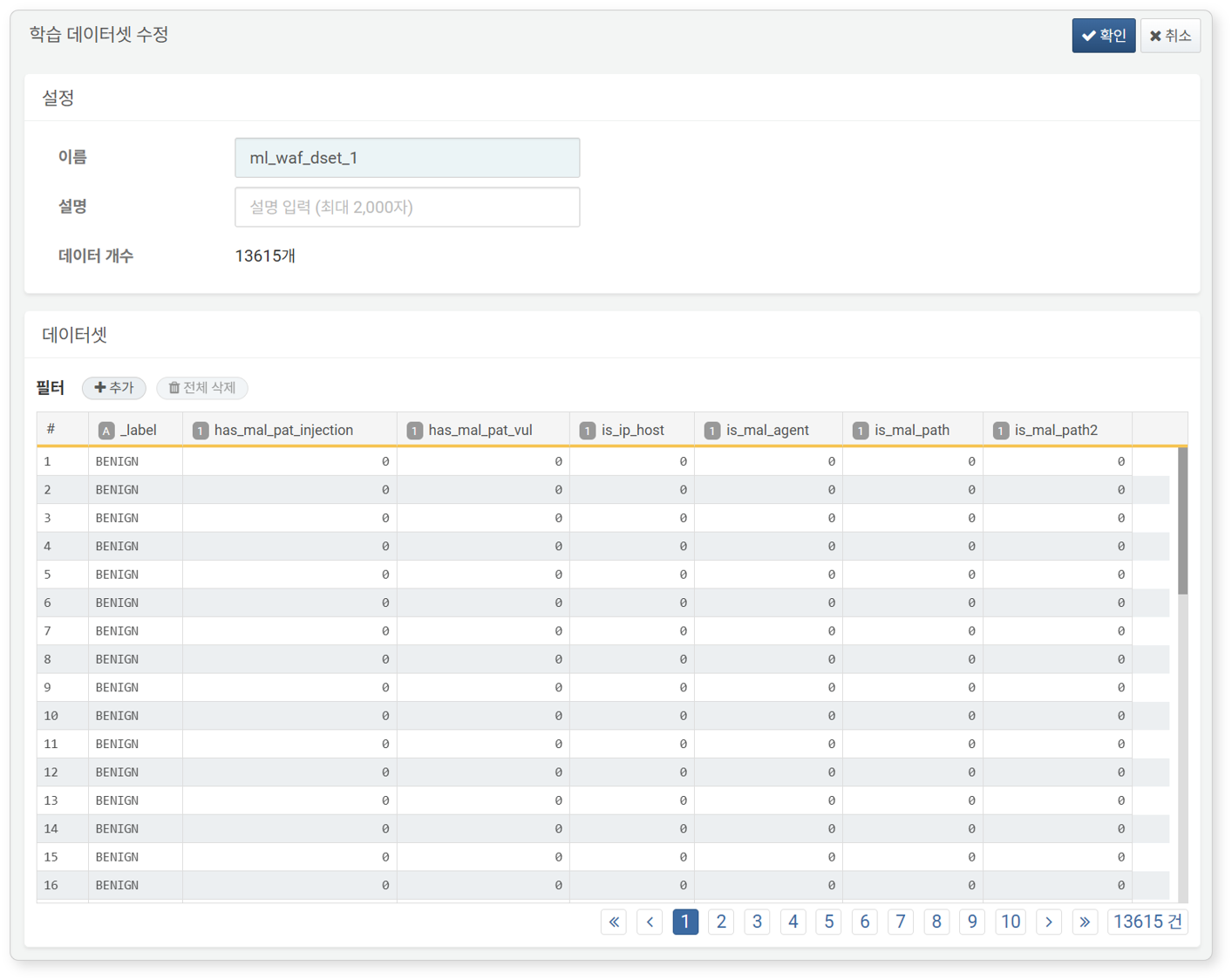


* **이름**: 학습 데이터셋 이름(최대 50자)
* **설명**: 학습 데이터셋의 상세 설명(최대 2,000자)
* **쿼리문**: 학습 데이터셋을 생성하는 데 사용할 쿼리문. 모델링 실행 시 학습 데이터셋에서 생성한 필드 값이 사용됩니다. 필요할 경우 필드값을 변경하도록 설정할 수 있습니다. (최대 10,000자)

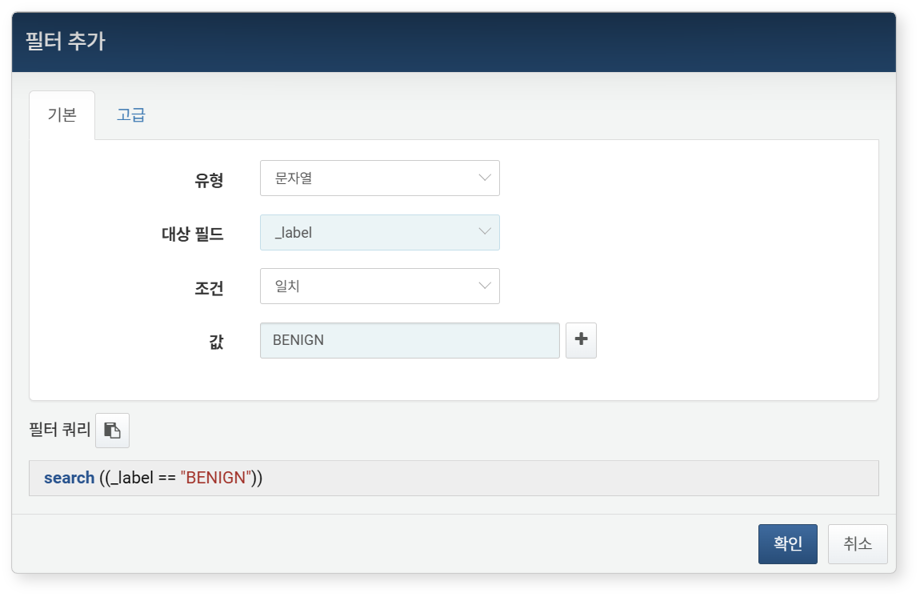
중복된 이름의 학습 데이터셋은 추가할 수 없습니다. 또한 올바르지 않은 쿼리문이나 존재하지 않는 데이터 테이블 지정 등으로 쿼리를 실행할 수 없는 경우, 학습 데이터셋을 추가할 수 없습니다.

#### 학습 데이터셋 조회

학습 데이터셋 목록에서 학습 데이터셋의 이름을 클릭하면 입력된 데이터를 확인할 수 있습니다.



필터를 적용해 원하는 데이터만 조회할 수 있습니다. 필터링 조건을 추가하려면 **필터** 우측에 있는 **추가**를 클릭하세요.



#### 학습 데이터셋 수정

학습 데이터셋을 수정하려면,

학습 데이터셋 목록에서 조회하거나 수정할 학습 데이터셋의 이름을 클릭하세요.

**학습 데이터셋 수정** 화면에서 정보를 수정한 후 **확인**을 클릭하세요. 수정할 수 있는 속성은 **설명**뿐입니다.

#### 학습 데이터셋 삭제

학습 데이터셋을 삭제하려면,

학습 데이터셋 목록에서 삭제할 학습 데이터셋 행의 체크박스를 선택하세요.

도구모음에서 **삭제**를 클릭하세요.

**학습 데이터셋 삭제** 대화상자에서 삭제할 학습 데이터셋 목록을 확인한 후 **삭제**를 클릭하세요. 삭제하지 않으려면 **취소**를 클릭하세요.

머신러닝 모델에 설정된 학습 데이터셋의 경우 삭제할 수 없습니다. 삭제하려면 해당 머신러닝 모델에서 학습 데이터셋 설정을 해제한 후 삭제하세요.