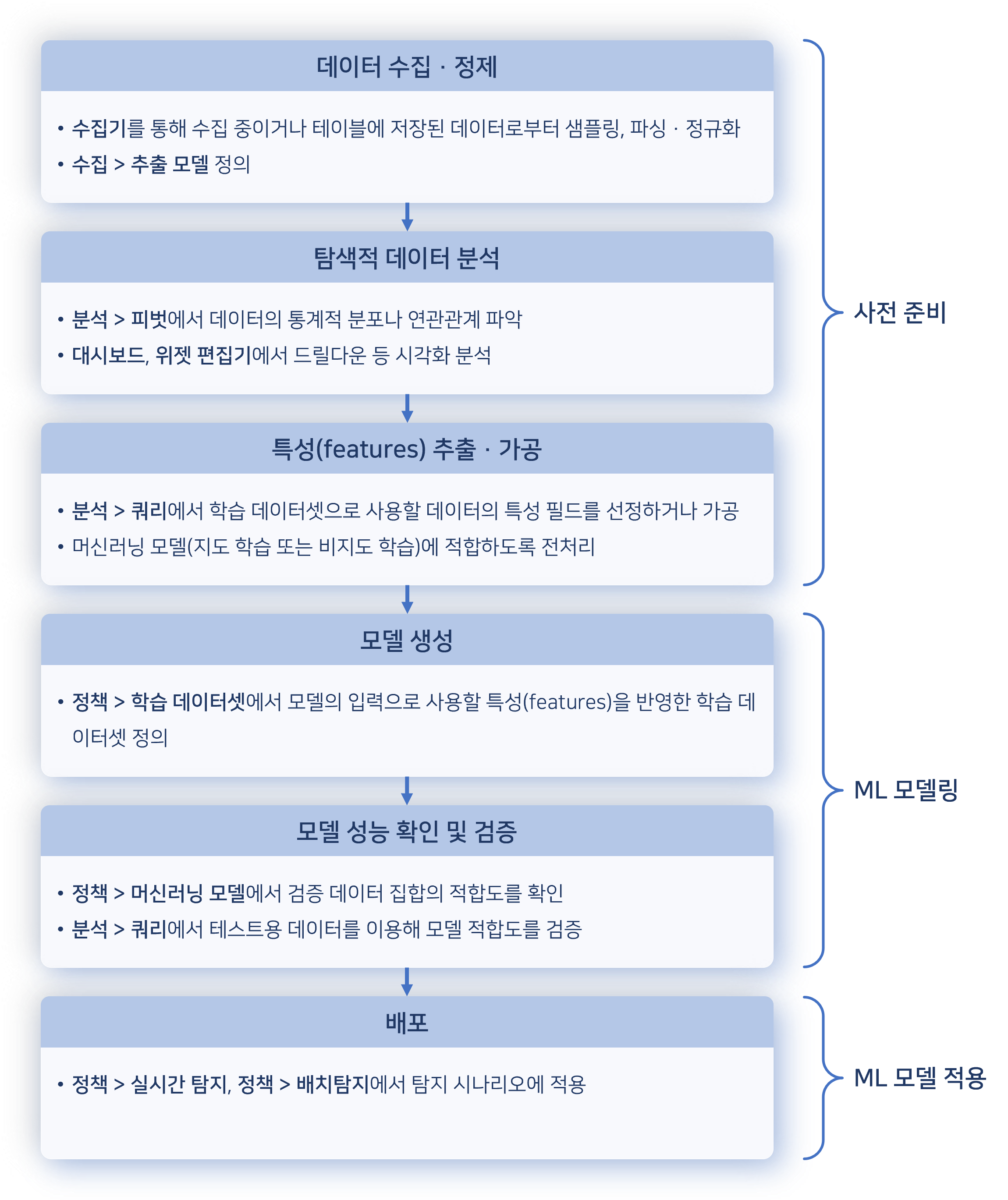
### 머신러닝 모델

#### 개요

머신러닝 모델은 주어진 학습 데이터셋을 분석하고, 그 데이터를 기반으로 예측이나 결정을 내리는 알고리즘의 집합을 의미합니다. 머신러닝은 컴퓨팅 기기가 명시적으로 프로그래밍되지 않고도 학습할 수 있도록 하는 기술입니다. 모델은 주어진 데이터에서 패턴을 학습하고, 이를 통해 새로운 데이터에 대한 예측을 수행합니다.

머신러닝 모델링 과정

머신러닝 모델을 만들고 활용하는 전체 단계는 다음과 같습니다.



데이터의 수집, 정제 단계는 데이터 원천으로부터 수집된 데이터를 가져와서 **분석 > 쿼리**, **분석 > 피벗**에서 사용자가 직접 분석하며 통계적 분포나 연관관계를 파악하는 과정입니다. 머신러닝 모델은 수치화된 데이터를 사용하는 것이 좋습니다. 필드 값이 문자열인 경우, 필드 값을 0, 1, 2와 같이 범주화된 숫자로 표현하는 방식으로 변환하거나, 원-핫 인코딩을 이용해 수치 벡터로 변환할 필요가 있습니다. 범주화해 처리할 수 없는 문자열 값은 TF-IDF 벡터화를 고려해볼 수 있습니다. TF는 단어 빈도(term frequency)를, IDF는 역문서 빈도(inverse document frequency)를 의미합니다.

사용자는 이러한 사전 준비 단계를 거쳐 준비된 데이터를 이용해 학습 데이터셋을 구성하고, 학습 데이터셋을 토대로 머신러닝 모델을 생성할 수 있습니다.

로그프레소 소나가 제공하는 머신러닝 모델은 지도 학습과 비지도 학습을 포함합니다.

지도 학습: 랜덤포레스트

로그프레소 소나가 제공하는 지도 학습 모델은 랜덤 포레스트(Random Forest) 모델입니다. 랜덤 포레스트 모델은 여러 개의 결정 트리를 사용하여 예측을 수행하는 앙상블 학습(ensemble learning)의 일종입니다. 일반적인 결정 트리에 비해 과적합(overfitting)에 강하고 개별 모델의 편향과 분산을 줄일 수 있어 일반화 능력을 향상시킬 수 있습니다. 랜덤 포레스트 모델은 예측 성능이 뛰어나 로그와 같이 복잡하고 다차원적인 데이터를 분석해 이상 행위를 탐지하는 데 유용합니다.

랜덤 포레스트는 주로 분류와 회귀 문제를 해결하는 데 사용되고, 예측 집계 방식은 다음과 같습니다.

* 분류: 각 트리가 예측한 클래스 중 가장 많이 나온 클래스를 최종 예측으로 선택합니다.
* 회귀: 각 트리의 예측값을 평균하여 최종 예측으로 선택합니다.

랜덤 포레스트의 사용 절차는 다음과 같습니다.

**학습 데이터셋 준비**

학습 데이터셋을 구성하는 필드들은 특징 필드와 타겟 변수 필드로 나뉩니다.

* **특징**: 분석할 데이터의 속성, 로그의 데이터 필드. 수집한 로그 데이터가 이에 해당합니다. 예를 들어 스템에서 수집된 로그 데이터를 사용하여 IP 주소, 이벤트 발생 시간, 이벤트 유형과 같은 다양한 속성을 특징으로 설정합니다.
* **타겟 변수**: 각 데이터가 속하는 범주가 되는 데이터 필드. 레이블이라고도 합니다. 일반적으로 머신러닝 모델이 값을 예측해야 할 대상 필드로, 데이터 특성에 따라 숫자, 불리언 값, 문자열 등으로 표현됩니다. 학습 데이터셋의 타겟 변수 값은 모델이 학습할 수 있는 지침이 되어, 로그의 패턴을 분류하는 데 도움을 줍니다.

지도 학습에서 타겟 변수의 값은 일반적으로 모델 학습에 용이하도록 숫자로 변환된 값으로 변환이 필요합니다. 1/0, true/false와 같은 불리언 값에 기반한 이진 분류가 대표적입니다. 그 외에 숫자(예: 0, 1, 2)나 문자열(예: Normal, Warning, Critical)로 표현되는 다중 클래스 분류, 실수형 값으로 표현되는 회귀 분석 값일 수도 있습니다.

**모델 훈련**

타겟 변수 필드가 포함된 데이터를 사용하여 랜덤포레스트 모델을 훈련시킵니다. 훈련된 모델은 이후 들어오는 새로운 데이터의 타겟 변수 필드의 값을 예측할 수 있는 능력을 갖추게 됩니다.

**예측 및 탐지**

훈련된 모델을 사용하여 실시간으로 들어오는 새로운 로그 데이터를 분석하고, 비정상적인 패턴을 탐지합니다. 예를 들어, 특정 IP 주소에서 짧은 시간 내에 비정상적으로 많은 로그인 시도가 발생하면, 모델은 이를 비정상으로 판단해 경고를 발송합니다. 이러한 탐지 과정을 통해 빠른 대응이 가능해집니다.

비지도 학습: 이상탐지 포레스트

로그프레소 소나의 비지도 학습 모델은 이상탐지 포레스트(Isolation Forest) 모델입니다. 이상탐지 포레스트 모델은 이상치(anomaly)나 비정상적인 데이터를 탐지하기 위해 랜덤 포레스트 모델을 변경한 알고리즘입니다. 이상탐지 포레스트는 비정상적인 트랜잭션, 네트워크 침입 탐지, 신용 카드 사기 탐지와 같은 이상탐지 문제에 유용하게 사용됩니다. 비지도 학습은 주어지는 정답 없이 데이터의 패턴이나 구조를 학습하는 과정으로 구성됩니다.

이상탐지 포레스트의 학습은 개별 트리의 고립 깊이(isolation depth, 개별 데이터 포인트를 고립시키기 위한 분할 횟수)를 평균하여 이상치 점수(anomaly score)를 계산하고, 이 점수를 기반으로 이상 여부를 판정하는 방식으로 이뤄집니다. 고립 깊이의 평균값이 낮을수록 이상치일 가능성이 높은 것으로 판단됩니다. 이상치 점수는 0 ~ 1 사이의 값으로 표현됩니다. 값이 1에 가까울수록 이상치일 가능성이 높습니다. 특정 임계치 이상일 때 이상치로 분류합니다.

이상탐지 포레스트의 사용 절차는 다음과 같습니다.

**학습 데이터셋 준비**

이상탐지 포레스트는 비지도 학습 방식이므로 데이터에 타겟 변수(정상/비정상 구분)를 미리 지정할 필요가 없습니다. 대신, 시스템에서 수집한 로그 데이터를 그대로 사용하여 훈련을 진행합니다.

**모델 훈련**

이상탐지 포레스트는 주어진 데이터 샘플을 반복적으로 무작위 분할하여 각 데이터 포인트를 격리하는 방식으로 훈련을 진행합니다. 비정상적인 데이터는 일반적으로 소수이기 때문에 적은 분할로도 쉽게 격리될 수 있습니다. 반대로, 정상 데이터는 더 많은 분할을 거쳐야 격리되므로 비정상 데이터는 적은 분할 횟수로도 쉽게 탐지할 수 있습니다.

**예측 및 탐지**

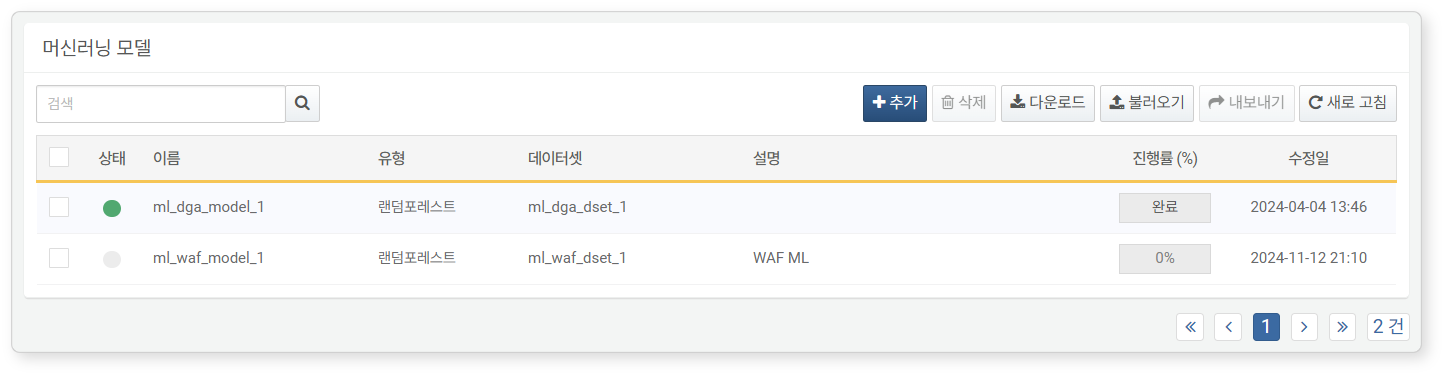
훈련이 완료된 이상탐지 포레스트 모델은 새로운 로그 데이터를 분석하고, 비정상적인 패턴을 탐지합니다. 모델은 각 로그가 격리되는 정도에 따라 점수를 부여하고, 점수가 높은 데이터는 비정상으로 간주합니다. 예를 들어, 정상적인 네트워크 트래픽과 특정 시간대에 발생하는 드문 트래픽 패턴을 분석할 때, 이상탐지 포레스트는 비정상적으로 많은 트래픽이 특정 IP로부터 발생하거나 허가되지 않은 장치가 접근하는 상황을 비정상으로 판단합니다. 이렇게 비정상 패턴이 탐지되면, 시스템은 즉시 알림을 보내고 대응을 할 수 있습니다.

사전 준비

머신러닝 모델을 사용하려면 학습 데이터셋이 필요합니다. [학습 데이터셋](https://docs.logpresso.com/ko/sonar/4.0/ui/section-ml-input)을 읽고 데이터셋을 준비하세요.

#### 머신러닝 모델 목록 조회/검색

**정책 > 머신러닝 모델**에서 머신러닝 모델 목록을 조회하거나 검색할 수 있습니다.



* **상태**: 머신러닝 모델을 사용할 수 있는 상태를 표시합니다. 모델링이 실행되지 않은 상태에서는 탐지 정책에 사용할 수 없습니다(녹색: 사용 가능, 회색: 사용 불가)
* **이름**: 머신러닝 모델 이름
* **유형**: 머신러닝 모델 유형
* **데이터셋**: 머신러닝 모델에 사용된 학습 데이터셋 유형(랜덤포레스트 또는 이상탐지 포레스트)
* **설명**: 머신러닝 모델에 대한 부가적인 정보
* **진행률(%)**: 학습 진행률. 학습 중인 머신러닝 모델은 진행률을 백분률(%)로 나타냅니다.
* **수정일**: 머신러닝 모델이 생성된 날짜, 또는 마지막으로 수정된 날짜

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

목록 다운로드

머신러닝 모델 목록을 로컬 PC에 파일로 다운로드하려면 도구 모음에서 **다운로드**를 클릭한 후 원하는 파일 형식을 선택하여 다운로드하세요.

목록 새로 고침

머신러닝 모델의 목록을 최신 정보로 조회하려면 도구 모음에서 **새로고침**을 클릭하세요.

내보내기/불러오기

생성한 머신러닝 모델을 파일로 내보내거나 불러올 수 있습니다. 생성한 머신러닝 모델을 백업하고 복원할 경우 사용할 수 있습니다.

머신러닝 모델을 내보내려면,

머신러닝 모델 목록에서 내보낼 머신러닝 모델 행의 체크박스를 선택하세요.

도구 모음에서 **내보내기**를 클릭하세요.

**머신러닝 모델 내보내기** 대화상자에서 이름을 설정하고 **확인**을 클릭하세요.

머신러닝 모델을 불러오려면,

도구 모음에서 **불러오기**를 클릭하세요.

**머신러닝 모델 불러오기** 대화상자에서 **파일 선택**을 클릭한 후 머신러닝 모델 파일을 선택하세요.

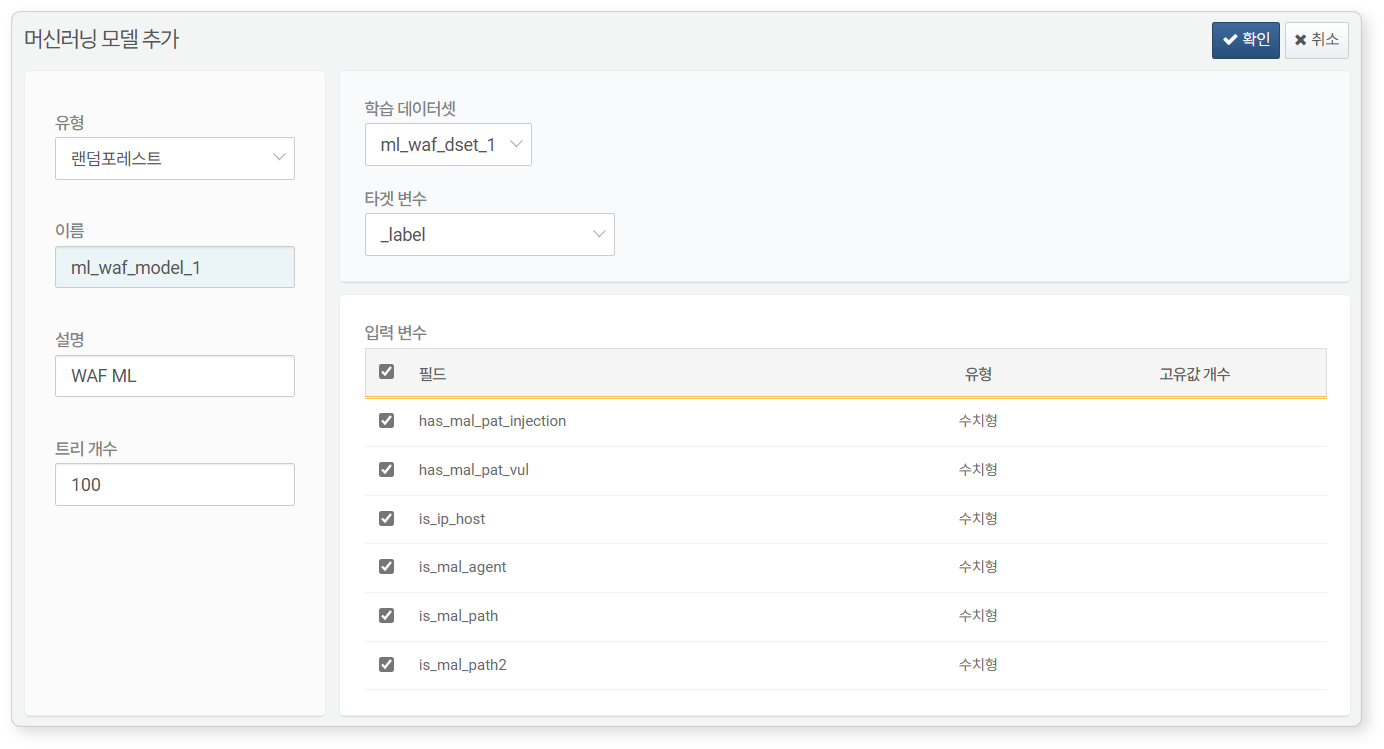
파일을 선택한 다음, **확인**을 클릭하세요.

#### 머신러닝 모델 추가

머신러닝 모델을 추가하려면,

**정책 > 머신러닝 모델**에서 도구 모음에 있는 **추가**를 클릭하세요. 머신러닝 모델링에는 학습 데이터셋이 필요합니다.

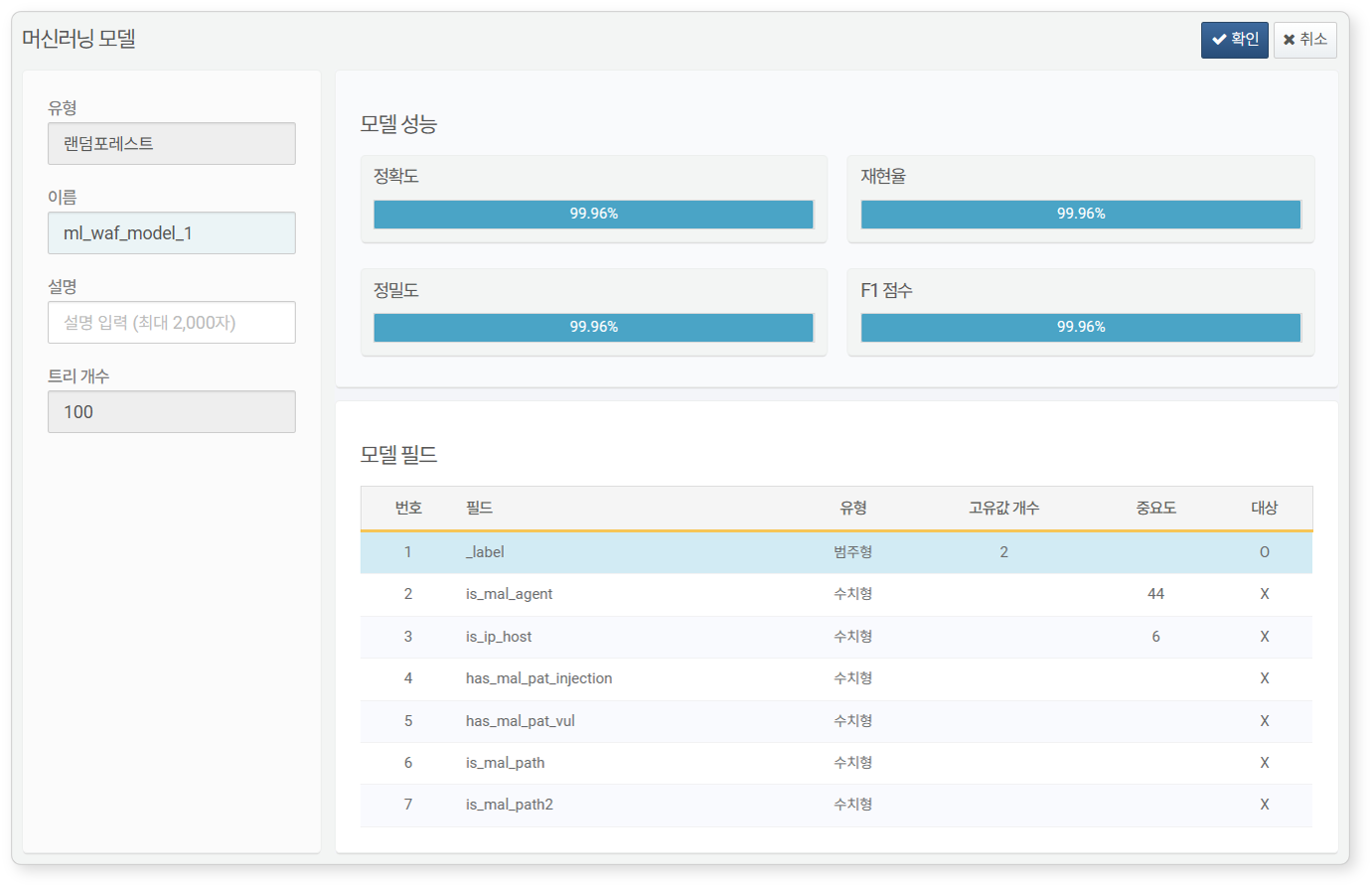
**머신러닝 모델 추가** 화면에서 필요한 값을 입력하거나 선택한 다음, **확인**을 클릭하세요.



* **유형**: 머신러닝 모델 유형. 예측 분석이 목적이면 랜덤포레스트, 이상탐지가 목적이면 이상탐지 포레스트를 선택하세요.
* **이름**: 쿼리에서 참조할 머신러닝 모델의 고유 이름(최대 50자)
* **설명**: 머신러닝 모델에 대한 설명(최대 2,000자)
* **트리 개수**: 머신러닝 모델의 의사결정 트리 개수. 트리가 많을수록 안정적인 결과를 얻을 수 있지만, 모델링과 추론에 소요되는 시간도 증가합니다. 성능과 효율성을 고려하여 적절하게 설정하세요. (기본값: 100, 범위: 1~500)
* **학습 데이터셋**: 머신러닝 모델을 생성하는 데 사용할 학습 데이터셋
* **타겟 변수**: (**유형**이 랜덤포레스트일 때) 머신러닝 모델이 값을 예측할 데이터 필드. 예를 들어 파일 접근에 대한 이상 탐지 모델링을 생성할 경우 파일 접근의 정상/비정상 여부의 필드를 타겟 변수로 설정할 수 있습니다.
* **입력 변수**: 학습 데이터셋의 전체 입력 변수 중에서 모델링 생성 시 학습에 사용할 필드 목록을 선택

#### 머신러닝 모델 조회

학습이 완료된 후 머신러닝 모델 이름을 클릭하면 성능 지표와 모델 필드를 확인할 수 있습니다.



**성능 지표**

성능 지표는 랜덤 포레스트 모델인 경우에만 보여줍니다.

* **정확도**: 모델이 올바르게 예측한 샘플의 비율. 균질하지 않은 데이터를 이용하는 경우, 정확도는 신뢰하기 어려우므로 주의하세요.
* **재현율**: 실제 양성 데이터 중 모델이 양성으로 예측한 비율. 양성 예측의 정확성을 평가하는 지표. 거짓 긍정(false positive)를 줄이고자 할 때 중요한 지표입니다.
* **정밀도**: 모델이 **양성**으로 예측한 데이터 중 실제로 양성인 비율. 거짓 부정(false negative)를 줄이고자 할 때 중요한 지표입니다.
* **F1 점수**: 정밀도와 재현율의 조화 평균. 정밀도와 재현율이 비슷할수록 점수가 높고, 한쪽이 높거나 낮을 경우 F1 점수도 낮아집니다.

F1 점수는 정밀도와 재현율이 모두 좋아야만 높게 나오기 때문에, 단 하나의 지표를 봐야한다면 F1 점수를 참고하시면 됩니다. 다중 클래스 분류인 경우에는 테스트 집합의 분류 값 분포가 지표에 가중치로 반영됩니다.

**모델 필드**

머신러닝 모델이 학습한 특징 필드, 타겟 변수 필드와 관련 정보를 보여줍니다.

#### 머신러닝 모델 수정

머신러닝 모델을 수정하려면,

머신러닝 모델 목록에서 수정할 머신러닝 모델 이름을 클릭하세요.

**머신러닝 모델 수정** 화면에서 정보를 수정한 후 **확인**을 클릭하세요. 수정할 수 있는 속성은 **이름**과 **설명**입니다. 다른 속성은 수정할 수 없습니다.

#### 머신러닝 모델 활용

추가한 머신러닝 모델은 로그프레소 쿼리에서 [rforest](https://docs.logpresso.com/ko/sonar/4.0/ui/ko/query/rforest-command), [anomalies](https://docs.logpresso.com/ko/sonar/4.0/ui/ko/query/anomalies-command) 명령으로 사용할 수 있습니다. 입력 필드에 랜덤 포레스트를 적용하려면 rforest 명령을, 입력 필드에 이상탐지 포레스트 모델을 적용하려면 anomalies 명령을 사용하세요.

rforest, anomalies 명령을 이용해 머신러닝 모델을 적용하려면 입력 필드는 모두 머신러닝 모델 필드와 일치해야 합니다. 머신러닝 명령어 구문을 입력하기 전에 먼저 입력 필드를 일치시킬 수 있도록 데이터를 가공하세요.

#### 머신러닝 모델 삭제

머신러닝 모델을 삭제하려면,

머신러닝 모델 목록에서 삭제할 머신러닝 모델 행의 체크박스를 선택하세요.

도구 모음에서 **삭제**를 클릭하세요.

**머신러닝 모델 삭제** 대화상자에서 삭제할 머신러닝 모델 목록을 확인한 후 **삭제**를 클릭하세요. 삭제하지 않으려면 **취소**를 클릭하세요.