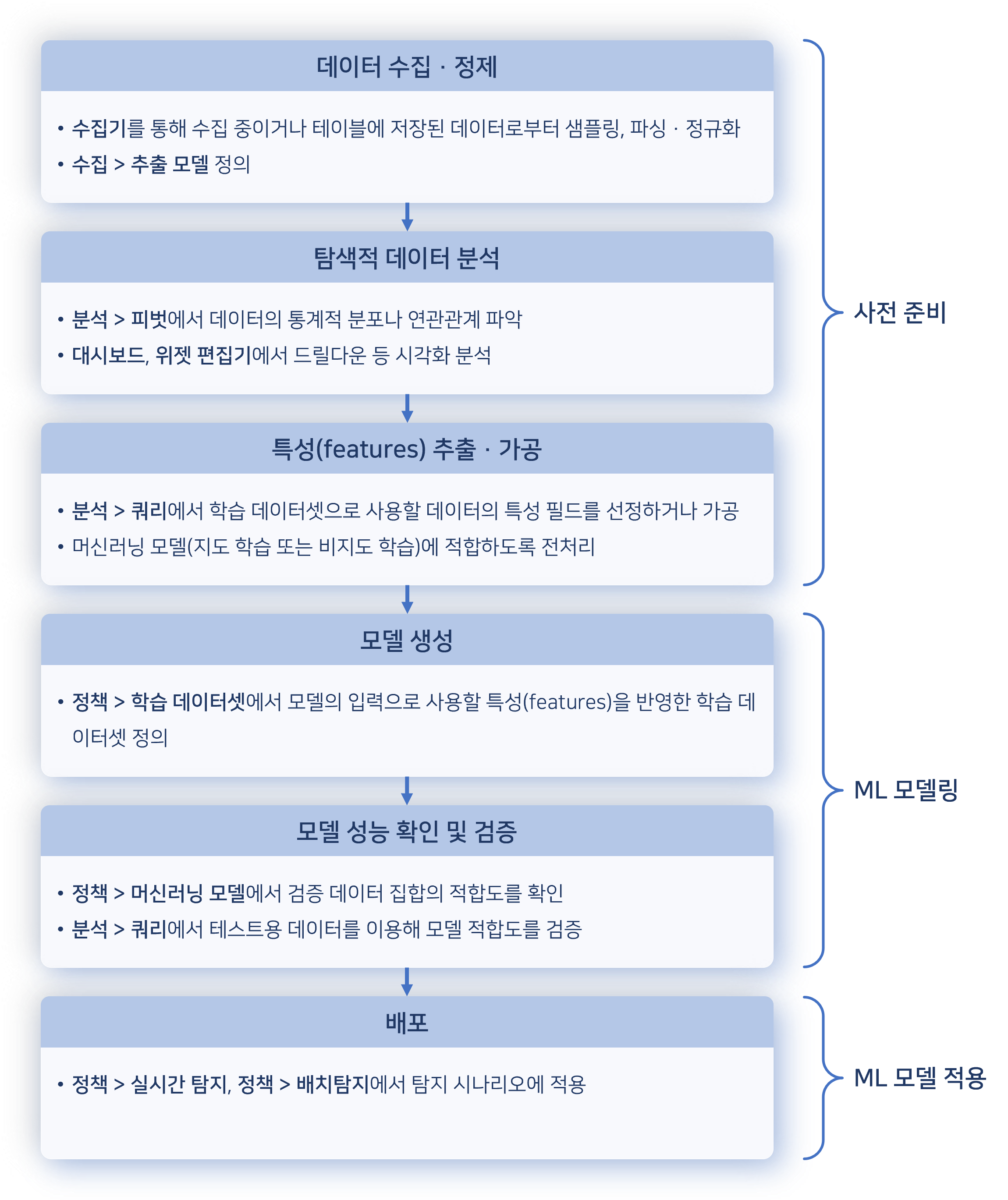
### Machine Learning Models

#### Overview

Machine learning models refer to a set of algorithms that analyze a given training dataset and make predictions or decisions based on that data. Machine learning is a technology that enables computing devices to learn without being explicitly programmed. The model learns patterns from the given data and performs predictions on new data.

Machine Learning Modeling Process

The overall steps to create and utilize machine learning models are as follows.



The data collection and cleansing stages involve gathering data from data sources and allowing users to analyze it through **Analysis > Query** and **Analysis > Pivot**, identifying statistical distributions or correlations. It is advisable to use numerical data for machine learning models. If field values are strings, they need to be converted into categorical numbers like 0, 1, 2, or transformed into numerical vectors using one-hot encoding. String values that cannot be categorized may be considered for TF-IDF vectorization, where TF stands for term frequency and IDF stands for inverse document frequency.

Users can prepare the data through these preliminary steps to construct a training dataset, which can then be used to create machine learning models.

Logpresso Sonar provides machine learning models that include both supervised and unsupervised learning.

Supervised Learning: Random Forest

The supervised learning model provided by Logpresso Sonar is the Random Forest model. The Random Forest model uses multiple decision trees to make predictions and is a type of ensemble learning. Compared to a typical decision tree, it is robust against overfitting and can reduce the bias and variance of individual models, thereby enhancing generalization capabilities. The Random Forest model excels in predictive performance and is useful for detecting anomalies in complex, multidimensional data such as logs.

Random Forest is primarily used to solve classification and regression problems, with the following prediction aggregation methods:

* Classification: The final prediction is selected as the class that appears most frequently among the predictions made by each tree.
* Regression: The final prediction is chosen by averaging the predicted values from each tree.

The procedure for using Random Forest is as follows:

**Preparing the Training Dataset**The fields that make up the training dataset are divided into feature fields and target variable fields.

* **Features**: Attributes of the data to be analyzed, corresponding to the log data fields. The collected log data falls under this category. For example, various attributes such as IP address, event occurrence time, and event type can be set as features using log data collected from the system.
* **Target Variable**: The data field that categorizes each data point, also known as the label. This is typically the field that the machine learning model needs to predict, expressed as numbers, boolean values, or strings depending on the data characteristics. The target variable values in the training dataset serve as guidelines for the model to learn and assist in classifying log patterns.

In supervised learning, the values of the target variable generally need to be converted into numerical values for ease of model training. Binary classification based on boolean values such as 1/0 or true/false is typical. Additionally, multi-class classification can be represented by numbers (e.g., 0, 1, 2) or strings (e.g., Normal, Warning, Critical), and regression analysis values may be expressed as real numbers.

**Model Training**The Random Forest model is trained using data that includes the target variable field. The trained model will then have the capability to predict the values of the target variable field for incoming new data.

**Prediction and Detection**The trained model is used to analyze incoming new log data in real-time and detect abnormal patterns. For example, if an unusually high number of login attempts occur from a specific IP address within a short time, the model will classify this as abnormal and issue a warning. This detection process enables rapid response.

Unsupervised Learning: Isolation Forest

The unsupervised learning model provided by Logpresso Sonar is the Isolation Forest model. The Isolation Forest model is an algorithm modified from the Random Forest model to detect anomalies or abnormal data. It is effectively used for anomaly detection problems such as detecting abnormal transactions, network intrusions, and credit card fraud. Unsupervised learning consists of a process where patterns or structures in the data are learned without given answers.

The training of the Isolation Forest involves calculating an anomaly score by averaging the isolation depth of individual trees (the number of splits required to isolate individual data points) and determining whether the data is anomalous based on this score. A lower average isolation depth indicates a higher likelihood of being an anomaly. The anomaly score is expressed as a value between 0 and 1, with values closer to 1 indicating a higher likelihood of being anomalous. Data points are classified as anomalies when they exceed a specific threshold.

The procedure for using the Isolation Forest is as follows:

**Preparing the Training Dataset**Since the Isolation Forest operates under an unsupervised learning approach, there is no need to pre-specify a target variable (normal/abnormal distinction) in the data. Instead, the training is conducted using the log data collected by the system as is.

**Model Training**The Isolation Forest is trained by repeatedly randomly splitting the given data samples to isolate each data point. Since abnormal data is generally rare, it can be easily isolated with fewer splits. Conversely, normal data requires more splits to be isolated, allowing for easier detection of abnormal data.

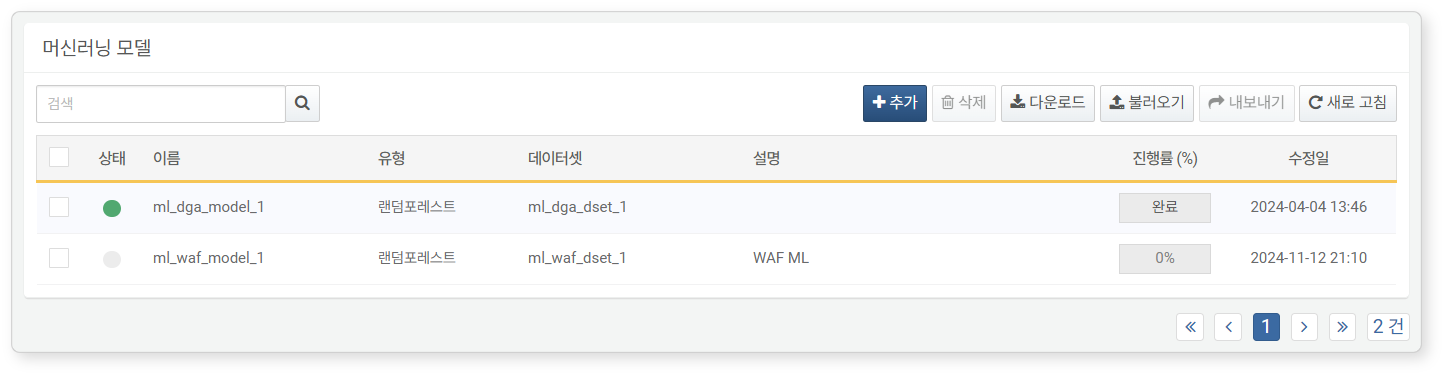
**Prediction and Detection**Once the training of the Isolation Forest model is complete, it analyzes new log data and detects abnormal patterns. The model assigns scores based on the degree to which each log is isolated, with higher scores indicating abnormality. For example, when analyzing normal network traffic and rare traffic patterns occurring at specific times, the Isolation Forest may classify an unusually high volume of traffic from a specific IP or access by unauthorized devices as abnormal. When such abnormal patterns are detected, the system can immediately send alerts and take action.

Preliminary Preparation

To use machine learning models, a training dataset is required. Please refer to [Training Dataset](https://docs.logpresso.comnull) to prepare the dataset.

#### Viewing/Search Machine Learning Model List

You can view or search the list of machine learning models under **Policies > Machine Learning Models**.



* **Status**: Indicates whether the machine learning model is available for use. If modeling has not been executed, it cannot be used in detection policies (green: available, gray: unavailable).
* **Name**: The name of the machine learning model.
* **Type**: The type of machine learning model.
* **Dataset**: The type of training dataset used for the machine learning model (Random Forest or Isolation Forest).
* **Description**: Additional information about the machine learning model.
* **Progress (%)**: The training progress. Machine learning models that are currently being trained display their progress as a percentage (%).
* **Modification Date**: The date the machine learning model was created or last modified.

To find a specific machine learning model in the list, use the search tool in the toolbar. The search tool will find and display machine learning models that include the entered term in their **Name** or **Description**. The search tool is case-insensitive.

Downloading the List

To download the machine learning model list as a file to your local PC, click **Download** in the toolbar and select the desired file format for download.

Refreshing the List

To refresh the machine learning model list with the latest information, click **Refresh** in the toolbar.

Exporting/Importing

You can export or import created machine learning models as files. This can be used for backing up and restoring created machine learning models.

To export a machine learning model:

Select the checkbox for the machine learning model row you wish to export from the machine learning model list.

Click **Export** in the toolbar.

In the **Export Machine Learning Model** dialog, set a name and click **OK**.

To import a machine learning model:

Click **Import** in the toolbar.

In the **Import Machine Learning Model** dialog, click **Select File** and choose the machine learning model file.

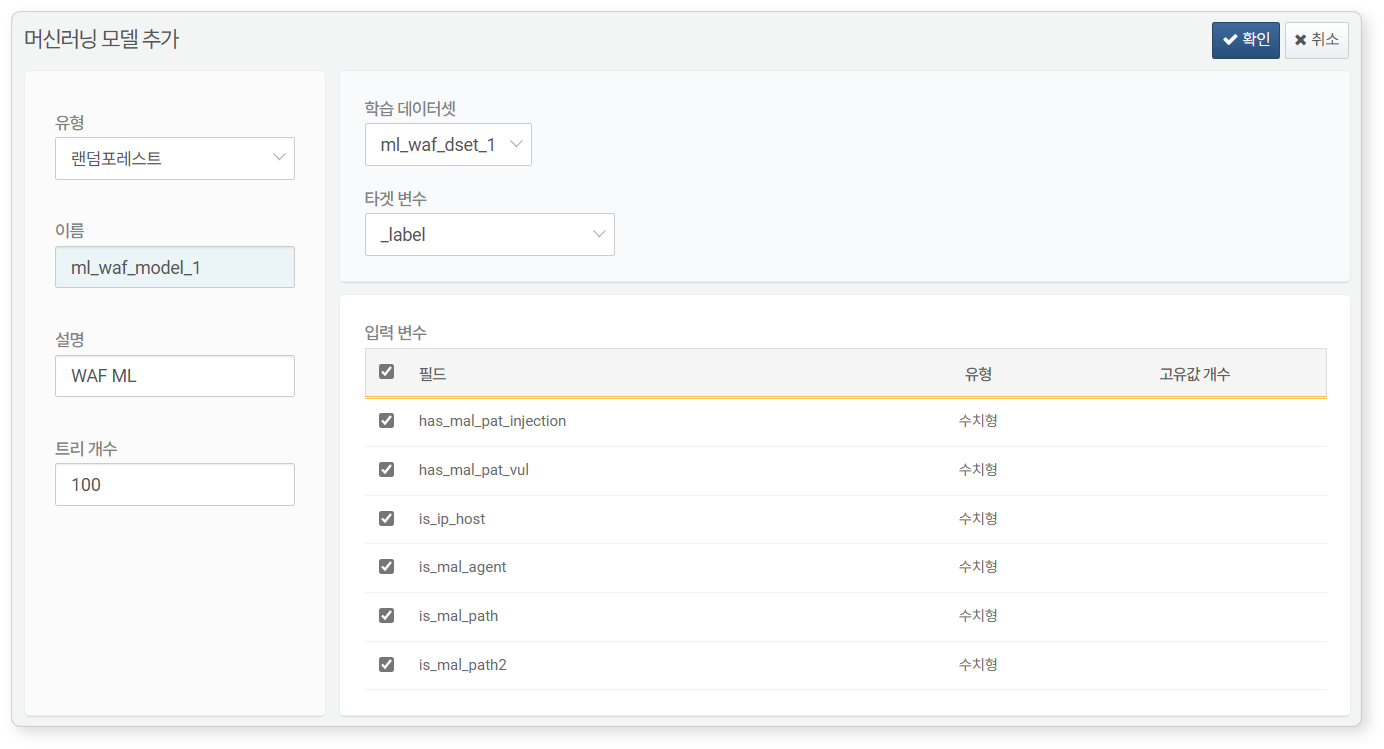
After selecting the file, click **OK**.

#### Adding a Machine Learning Model

To add a machine learning model:

Click **Add** in the toolbar under **Policies > Machine Learning Models**. A training dataset is required for machine learning modeling.

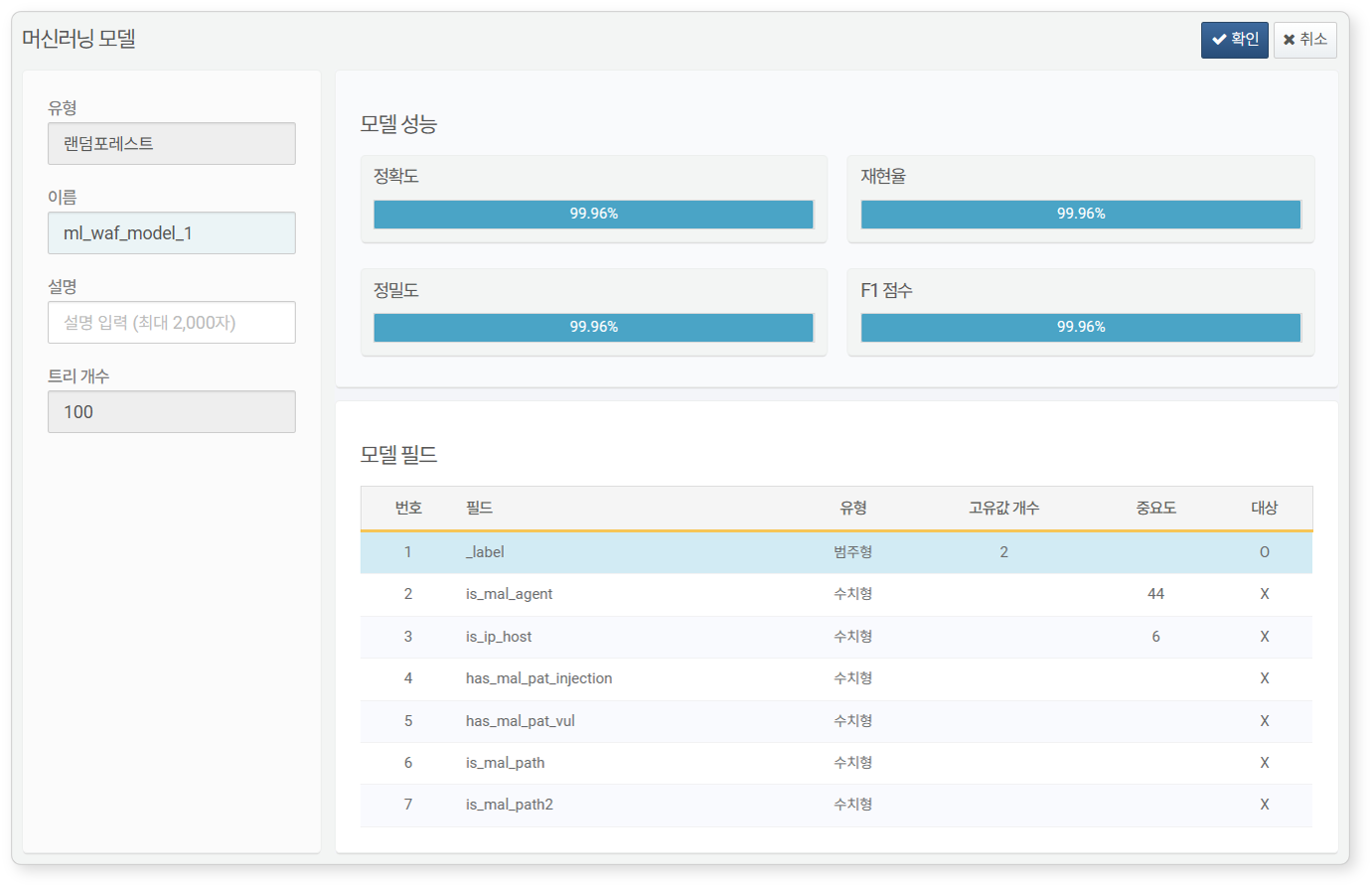
In the **Add Machine Learning Model** screen, enter or select the necessary values and click **OK**.



* **Type**: The type of machine learning model. Select Random Forest for predictive analysis or Isolation Forest for anomaly detection.
* **Name**: A unique name for the machine learning model to be referenced in queries (up to 50 characters).
* **Description**: A description of the machine learning model (up to 2,000 characters).
* **Number of Trees**: The number of decision trees in the machine learning model. More trees yield more stable results, but also increase the time required for modeling and inference. Set appropriately considering performance and efficiency (default: 100, range: 1-500).
* **Training Dataset**: The training dataset to be used for creating the machine learning model.
* **Target Variable**: (When **Type** is Random Forest) The data field that the machine learning model will predict. For example, when creating an anomaly detection model for file access, the field indicating whether the file access is normal or abnormal can be set as the target variable.
* **Input Variables**: Select the list of fields from the entire input variables of the training dataset to be used for modeling.

#### Viewing a Machine Learning Model

After training is complete, clicking on the machine learning model name will allow you to view performance metrics and model fields.



**Performance Metrics**Performance metrics are displayed only for Random Forest models.

* **Accuracy**: The ratio of samples that the model predicted correctly. Be cautious when using accuracy as a measure with heterogeneous data, as it may not be reliable.
* **Recall**: The ratio of actual positive data that the model predicted as positive. This metric evaluates the accuracy of positive predictions and is important when aiming to reduce false positives.
* **Precision**: The ratio of data predicted as **positive** that is actually positive. This metric is important when aiming to reduce false negatives.
* **F1 Score**: The harmonic mean of precision and recall. The score is higher when precision and recall are similar, and lower when one is high or low.

The F1 score is high only when both precision and recall are good, so if you need to refer to a single metric, consider the F1 score. In the case of multi-class classification, the distribution of classification values in the test set is reflected as weights in the metrics.

**Model Fields**Displays the feature fields that the machine learning model has learned, the target variable fields, and related information.

#### Modifying a Machine Learning Model

To modify a machine learning model:

Click on the name of the machine learning model you wish to modify in the machine learning model list.

In the **Modify Machine Learning Model** screen, edit the information and click **OK**. The properties that can be modified are **Name** and **Description**. Other properties cannot be modified.

#### Utilizing a Machine Learning Model

The added machine learning models can be used in Logpresso queries with the [rforest](https://docs.logpresso.comnull) and [anomalies](https://docs.logpresso.comnull) commands. To apply Random Forest to input fields, use the rforest command, and to apply the Isolation Forest model to input fields, use the anomalies command.

To apply machine learning models using the rforest and anomalies commands, all input fields must match the machine learning model fields. Before entering the machine learning command syntax, ensure that the data is processed to match the input fields.

#### Deleting a Machine Learning Model

To delete a machine learning model:

Select the checkbox for the machine learning model row you wish to delete from the machine learning model list.

Click **Delete** in the toolbar.

In the **Delete Machine Learning Model** dialog, review the list of machine learning models to be deleted and click **Delete**. Click **Cancel** if you do not wish to delete.