### ML Datasets

#### Overview

An ML dataset is a collection of data that the machine learning model will learn from. The model learns patterns from the provided dataset and predicts values or patterns based on similar data, making the quality of the training dataset crucial to the outcomes of the ML model.

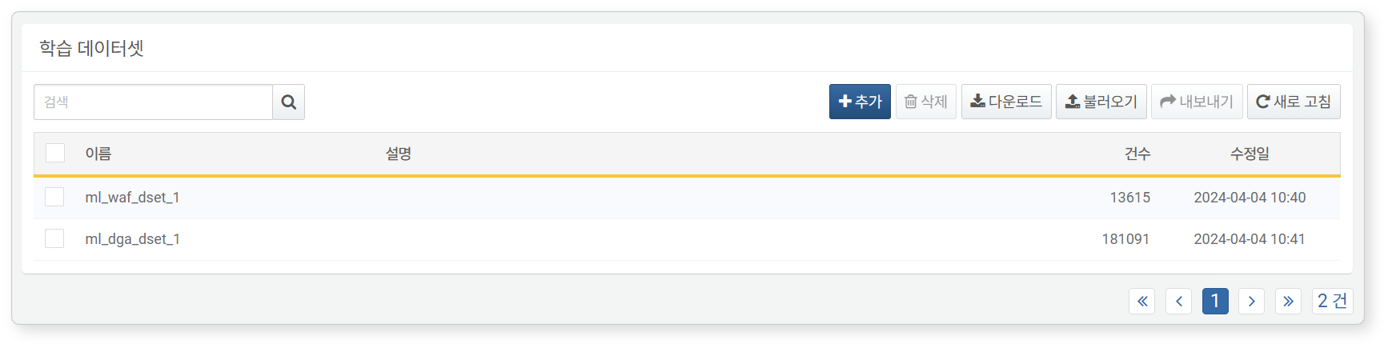
Considerations

The training dataset should ideally mirror the data generated in the environment where the ML model will be utilized. When creating the dataset, consider the following:

* **Inclusion of Diverse Information**: Collect and prepare data that encompasses various characteristics (time, location, user demographics, event types, success/failure, etc.) that may occur in the environment where the ML model will be applied. Utilizing data from real-world scenarios is recommended.
* **Target Variable Selection for Supervised Learning**: Supervised learning models, such as Random Forest, must include a target variable for prediction. When creating the training dataset, ensure that the target variable (the value to be predicted) is included.
* **Data Preprocessing**: Remove or adjust data errors and outliers as necessary, and normalize or scale feature values to prevent significant discrepancies.
* **Vectorization of Strings**: General strings must be converted into a numerical format for the ML model to learn. Use methods such as TF-IDF query commands for vectorization.
* **Quality and Quantity of Data**: Data must be accurate, complete, and consistent, with a sufficient volume to learn various patterns. Ensure that there are enough instances of anomalous events.
* **Overfitting Prevention**: Apply techniques to prevent overfitting, such as cross-validation and regularization, to avoid excessive specialization to the training data.
* **Changing Patterns**: Security threat patterns may evolve over time, necessitating periodic retraining of the ML model.

#### Viewing/Search for ML Datasets

You can view or search for the list of ML datasets under **Policies > ML Datasets**.



* **Name**: The name of the ML dataset
* **Description**: Additional information about the ML dataset
* **Count**: The number of data entries currently included in the ML dataset
* **Modification Date**: The date the ML dataset was created or last modified

To find a specific ML dataset in the list, use the search tool in the toolbar. The search tool will display datasets that include the entered term in the **Name** or **Description** fields. The search tool is case-insensitive.

Downloading the List

To download the list of ML datasets to your local PC, click **Download** in the toolbar and choose your desired file format.

Refreshing the List

To refresh the list of ML datasets with the latest information, click **Refresh** in the toolbar.

Export/Import

You can export or import created ML datasets as files. This can be useful for backing up and restoring your created ML datasets.

To export an ML dataset:

Select the checkbox for the row of the ML dataset you wish to export from the list.

Click **Export** in the toolbar.

In the **Export ML Dataset** dialog, set a name and click **OK**.

To import an ML dataset:

Click **Import** in the toolbar.

In the **Import ML Dataset** dialog, click **Select File** and choose the previously exported ML dataset file.

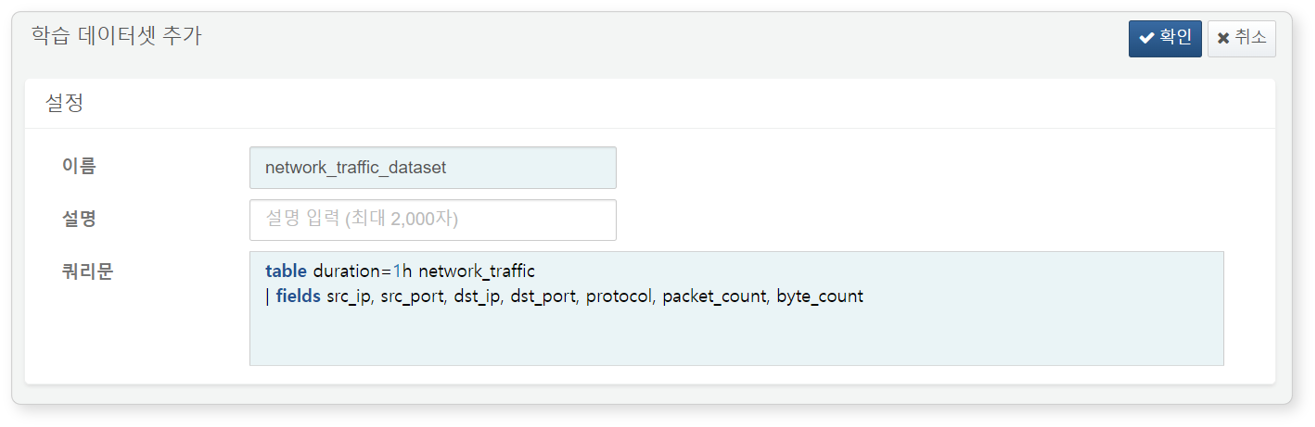
After selecting the file, click **OK**.

#### Adding an ML Dataset

To add an ML dataset:

Click **Add** in the toolbar under **Policies > ML Datasets**.

In the **Add ML Dataset** screen, enter or select the required values, then click **OK**.

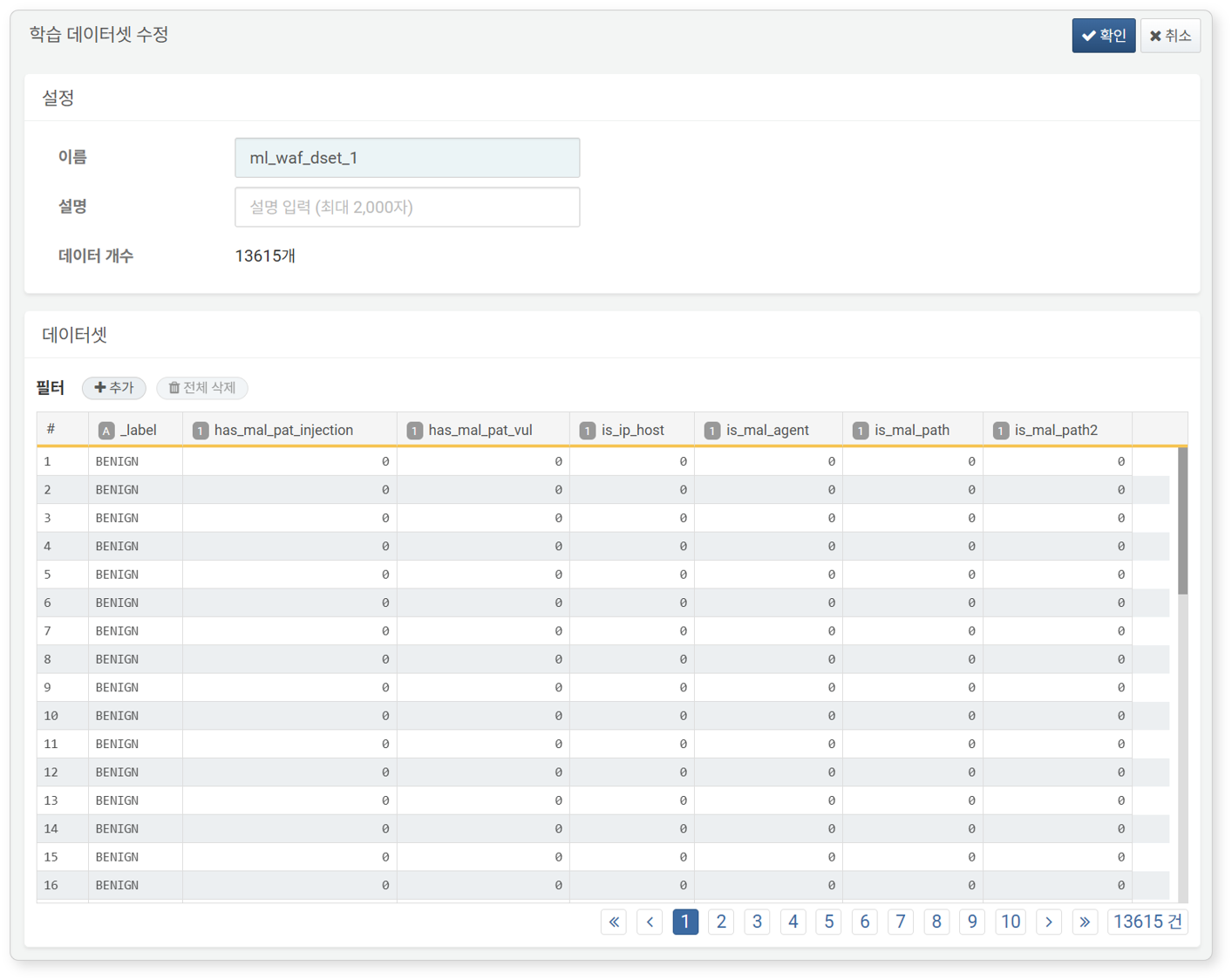


* **Name**: The name of the ML dataset (up to 50 characters)
* **Description**: A detailed description of the ML dataset (up to 2,000 characters)
* **Query**: The query used to create the ML dataset. The field values generated from the ML dataset will be used during modeling execution. You can set it to allow changes to field values if necessary (up to 10,000 characters).

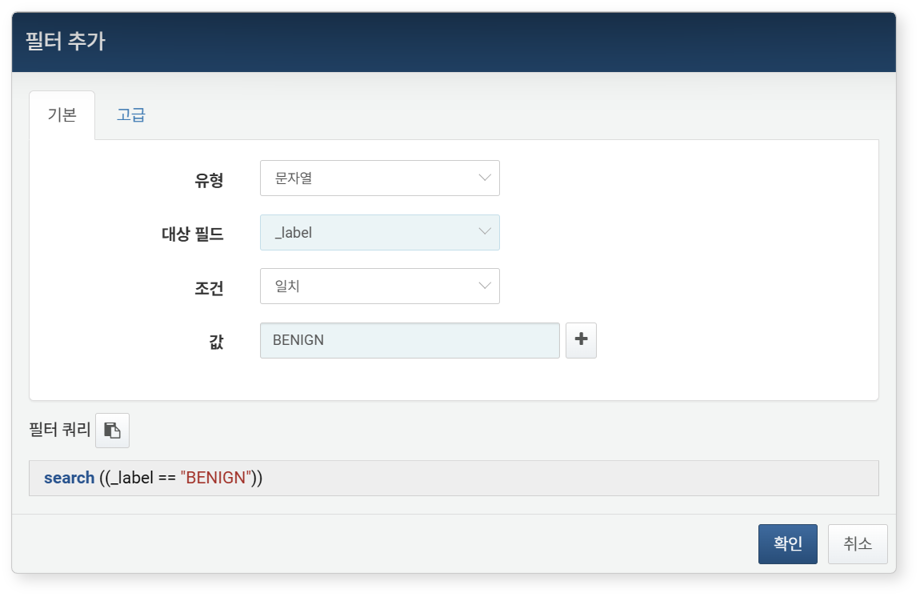
ML datasets with duplicate names cannot be added. Additionally, if the query cannot be executed due to an invalid query or specifying a non-existent data table, the ML dataset cannot be added.

#### Viewing an ML Dataset

Clicking the name of an ML dataset in the list allows you to view the entered data.



You can apply filters to view only the desired data. To add filtering conditions, click **Add** next to **Filter**.



#### Modifying an ML Dataset

To modify an ML dataset:

Click the name of the ML dataset you wish to view or modify in the list.

In the **Modify ML Dataset** screen, edit the information and click **OK**. The only property that can be modified is the **Description**.

#### Deleting an ML Dataset

To delete an ML dataset:

Select the checkbox for the row of the ML dataset you wish to delete from the list.

Click **Delete** in the toolbar.

In the **Delete ML Dataset** dialog, review the list of datasets to be deleted and click **Delete**. Click **Cancel** if you do not wish to delete.

ML datasets set for machine learning models cannot be deleted. To delete, first remove the dataset setting from the corresponding machine learning model, then proceed with the deletion.