### 機械学習モデル

#### 概要

機械学習モデルとは、与えられたデータセットを分析し、予測や意思決定を行うアルゴリズムの集合体です。機械学習は、明示的なプログラミングなしにコンピュータシステムがデータから学習することを可能にします。モデルはデータ内のパターンを識別し、新しいデータに対して予測を行います。

機械学習モデリング

機械学習モデルの構築および活用プロセスは、いくつかの主要なステップで構成されます。



データ収集および前処理段階では、様々なソースからデータを収集し、精緻化します。ユーザーは **分析 > クエリ** および **分析 > ピボット** を通じて、統計的分布や相関関係を把握できます。

機械学習モデルは数値データで最も高い性能を発揮するため、テキストベースのフィールドは数値表現に変換する必要があります。フィールド値が文字列の場合、0、1、2などのカテゴリカルな数値に変換するか、ワンホットエンコーディングで数値ベクトルに変換します。カテゴリ化できない文字列値は、TF-IDFベクトル化（TFはTerm Frequency、IDFはInverse Document Frequency）を検討できます。

これらの前処理が完了したら、学習データセットを準備し、それに基づいて機械学習モデルを生成します。

ログプレッソ・ソナーは、教師あり学習と教師なし学習の両方をサポートする機械学習モデルを提供します。

教師あり学習：ランダムフォレスト

ログプレッソ・ソナーは教師あり学習用に **ランダムフォレスト** モデルを提供します。ランダムフォレストは複数の決定木を組み合わせて予測精度を向上させるアンサンブル学習手法です。単一の決定木と比較して過学習に強く、バイアスや分散を低減し、汎化性能を高めます。高い予測精度により、複雑かつ多次元なログデータの異常検知などに特に有用です。

ランダムフォレストは主に分類および回帰に利用され、予測の集約方法は以下の通りです。

* 分類：各決定木の予測結果のうち、最も多く選ばれたクラスが最終予測となります。
* 回帰：全ての決定木が予測した値の平均が最終予測となります。

ランダムフォレストの利用方法

(1) 学習データセットの準備データセットは特徴量（フィーチャー）とターゲット変数で構成されます。

* **特徴量**：収集したログデータから分析対象となる属性（例：IPアドレス、イベントタイムスタンプ、イベントタイプなど）。
* **ターゲット変数**：各データポイントを分類するデータフィールド（ラベルとも呼ばれる）。モデルが予測すべきフィールドであり、データの特性に応じて数値、ブール値、文字列で表現されます。学習データセット内のターゲット変数は、モデルがパターンを学習し、ログパターンを分類するための指標となります。

教師あり学習では、ターゲット変数を数値化してモデル学習を容易にします。例：

* 二値分類値：1/0（または true/false）
* 多クラス分類値：0、1、2（または Normal、Warning、Critical）
* 回帰分析値：実数値

(2) モデル学習ターゲット変数を含むデータセットでランダムフォレストモデルを学習させます。学習後は新規データに対してターゲット値を予測できます。

(3) 予測・検知学習済みモデルはリアルタイムのログデータを分析し、異常を検知します。例えば、あるIPアドレスが短時間に異常な数のログイン試行を行った場合、モデルはこれを疑わしいと判断し、アラートを発報します。これにより迅速な脅威検知と対応が可能となります。

教師なし学習：異常検知用アイソレーションフォレスト

ログプレッソ・ソナーの教師なし学習モデルは **アイソレーションフォレスト** アルゴリズムに基づいています。このモデルはランダムフォレストの変種で、データ内の異常や不規則性の検出に特化しています。アイソレーションフォレストは、不正取引の検出、ネットワーク侵入検知、クレジットカード不正利用検知などで広く利用されています。

教師なし学習は、事前にラベル付けされたデータなしでデータのパターンや構造を分析します。アイソレーションフォレストは、データポイントを分割して孤立させるまでの分割回数（アイソレーション深度）を測定します。複数の木で平均したアイソレーション深度から異常スコアを算出します。平均深度が低いほど異常である可能性が高く、異常スコアは0から1の範囲で表され、1に近いほど異常度が高いことを示します。あらかじめ設定した閾値を超えるデータポイントは異常と判定されます。

アイソレーションフォレストの利用方法

(1) 学習データセットの準備アイソレーションフォレストは教師なし学習モデルのため、ラベル（正常／異常）の事前定義は不要です。システムから収集したログデータをそのまま学習に使用します。

(2) モデル学習モデルはデータセットをランダムに分割しながらデータポイントを孤立させます。異常値は稀であり、少ない分割で孤立しやすい一方、正常データはより多くの分割を要します。これにより異常値の検出が容易になります。

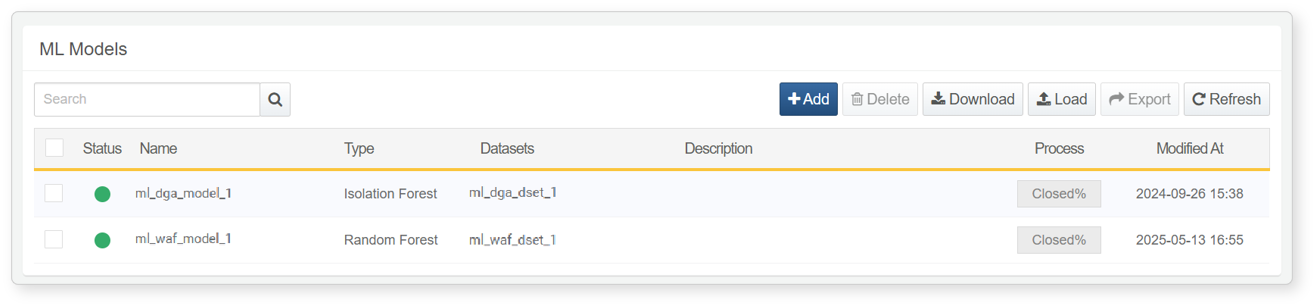
(3) 予測・検知学習後、アイソレーションフォレストモデルは新規ログデータを分析し、異常パターンを検出します。各ログエントリにはアイソレーション深度に基づく異常スコアが付与されます。例えば、ネットワークトラフィック分析では、特定IPからの異常な高トラフィックや不正なデバイス接続などを検知できます。異常が検出されると、システムは即座にアラートを発報し、迅速な対応が可能となります。

事前準備

機械学習モデルを利用するには、学習データセットが必要です。データセットの準備方法は[学習データセット](https://docs.logpresso.comnull)を参照してください。

#### 機械学習モデルの検索

**ポリシー > 機械学習モデル** で、機械学習モデルの一覧を閲覧・検索できます。



* **状態**：機械学習モデルの利用可否を示します。未学習の場合は検出ポリシーで利用できません（緑：利用可能、灰色：利用不可）。
* **名前**：機械学習モデルの名称。
* **タイプ**：機械学習モデルの種類。
* **データセット**：機械学習モデルに使用されたデータセットの種類（例：ランダムフォレストまたはアイソレーションフォレスト）。
* **説明**：機械学習モデルに関する補足情報。
* **進捗（%）**：モデル学習中は進捗率（パーセンテージ）で表示されます。
* **最終更新日**：モデルの作成または最終更新日時。

特定の機械学習モデルを検索するには、ツールバーの検索ツールを利用します。検索ツールは**名前**または**説明**に含まれるキーワードでモデルを検索します。大文字・小文字は区別されません。

機械学習モデル一覧のダウンロード

機械学習モデル一覧をローカルPCにファイルとしてダウンロードするには、ツールバーの**ダウンロード**をクリックし、希望するファイル形式を選択します。

機械学習モデル一覧の更新

最新の機械学習モデル一覧を表示するには、ツールバーの**更新**をクリックします。

機械学習モデルのインポート／エクスポート

機械学習モデルをファイルとしてエクスポートまたはインポートできます。バックアップや復元に便利です。

機械学習モデルをエクスポートするには：

エクスポートしたいモデル行のチェックボックスを選択します。

ツールバーの**エクスポート**をクリックします。

**機械学習モデルのエクスポート**ダイアログで名前を設定し、**OK**をクリックします。

機械学習モデルをインポートするには：

ツールバーの**インポート**をクリックします。

**機械学習モデルのインポート**ダイアログで**ファイル選択**をクリックし、モデルファイルを選択します。

ファイル選択後、**OK**をクリックします。

#### 機械学習モデルの追加

機械学習モデルを追加するには：

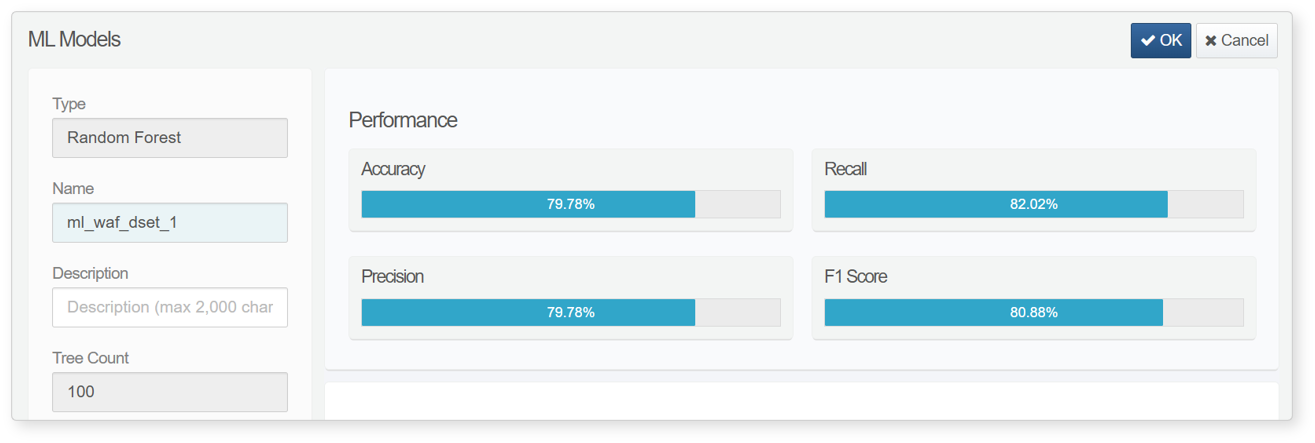
**ポリシー > 機械学習モデル**でツールバーの**追加**をクリックします。機械学習モデリングには学習データセットが必要です。

**新規機械学習モデル**画面で必要事項を入力または選択し、**OK**をクリックします。

* **タイプ**：機械学習モデルの種類を選択します。予測分析にはランダムフォレスト、異常検知にはアイソレーションフォレストを選択します。
* **名前**：クエリで参照される一意のモデル名（最大50文字）。
* **説明**：モデルの説明（最大2,000文字）。
* **ツリー数**：モデル内の決定木の数。ツリー数が多いほど結果は安定しますが、学習・推論時間が増加します。性能と効率を考慮して適切に設定してください（デフォルト：100、範囲：1～500）。
* **学習データセット**：モデル作成に使用する学習データセット。
* **ターゲット変数**：（ランダムフォレストモデルの場合）モデルが予測するデータフィールド。例：ファイルアクセスの異常検知モデルを作成する場合、各アクセスイベントの正常／異常状態をターゲット変数として定義できます。
* **入力変数**：学習データセットからモデル作成に使用するフィールドを選択します。

#### 機械学習モデルの閲覧

モデル学習が完了したら、モデル名をクリックして性能指標やモデルフィールドを確認できます。



* 性能指標：ランダムフォレストモデルのみ表示されます。
* **正解率（Accuracy）**：モデルが正しく予測したサンプルの割合。異種データでは信頼性が低下する場合があります。
* **再現率（Recall）**：実際に正例であるデータのうち、モデルが正例と予測した割合。偽陽性の削減に重要な指標です。
* **適合率（Precision）**：モデルが正例と予測したデータのうち、実際に正例であった割合。偽陰性の削減に重要な指標です。
* **F1スコア**：**適合率**と**再現率**の調和平均。両者のバランスが取れているほどスコアが高くなり、どちらかに偏ると低下します。

F1スコアは適合率と再現率の両方が重要な場合に最適な指標であり、バランスの取れた評価を提供します。多クラス分類では、テストセット内の分類値の分布も指標に反映されます。

* モデルフィールド：モデルが学習した特徴量およびターゲット変数のフィールドを表示します。

#### 機械学習モデルの編集

機械学習モデルを編集するには：

編集したいモデル名を一覧からクリックします。

**機械学習モデルの編集**画面で情報を更新し、**OK**をクリックします。**名前**と**説明**のみ編集可能で、その他の属性は変更できません。

#### 機械学習モデルの利用

追加した機械学習モデルは、ログプレッソクエリで [rforest](https://docs.logpresso.comnull) または [anomalies](https://docs.logpresso.comnull) コマンドを使って利用できます。ランダムフォレストモデルを入力フィールドに適用するには rforest コマンド、アイソレーションフォレストモデルには anomalies コマンドを使用します。

rforest または anomalies コマンドで機械学習モデルを適用する際は、入力フィールドがモデルのフィールドと一致している必要があります。コマンド利用前にデータのフィールド構成を確認してください。

#### 機械学習モデルの削除

機械学習モデルを削除するには：

削除したいモデルのチェックボックスを一覧から選択します。

ツールバーの**削除**をクリックします。

**機械学習モデルの削除**ダイアログで削除対象モデルを確認し、**削除**をクリックします。削除をキャンセルする場合は**キャンセル**をクリックします。