

2023年5月18日  
データサイエンティスト集会 in VRC

# 1分で作って評価する機械学習モデル Orange Data Miningの紹介

1

ぶんちん

# 自己紹介 ぶんちゃん

- ▶ 複合経営が特徴の企業（製造業）に所属
- ▶ データ分析担当者だったが。。。



VRChat初めて約2か月です



# Orange Data Mining

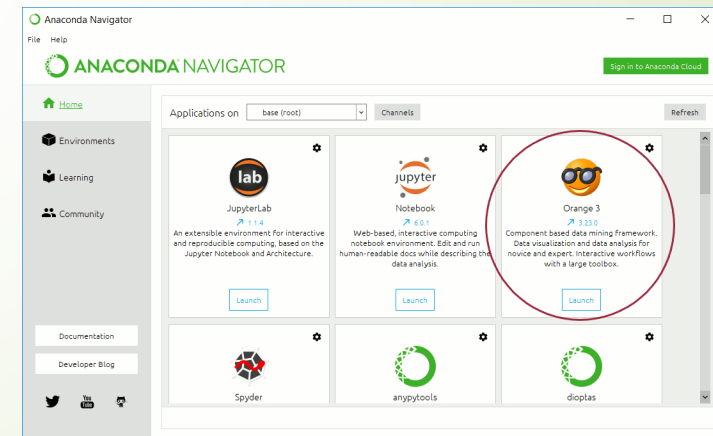
<https://orangedatamining.com/>

- ▶ ビジュアルプログラミング的にデータ分析や機械学習モデル作成・評価が可能
- ▶ 公式HPから入手すれば、企業でも無料で使用可能

## ▶ UIが素晴らしい

- ▶ 初心者は勉強に使おう！
- ▶ 専門家は手抜き・教育に使おう！

Anaconda使っている人は  
見たことあるのでは？

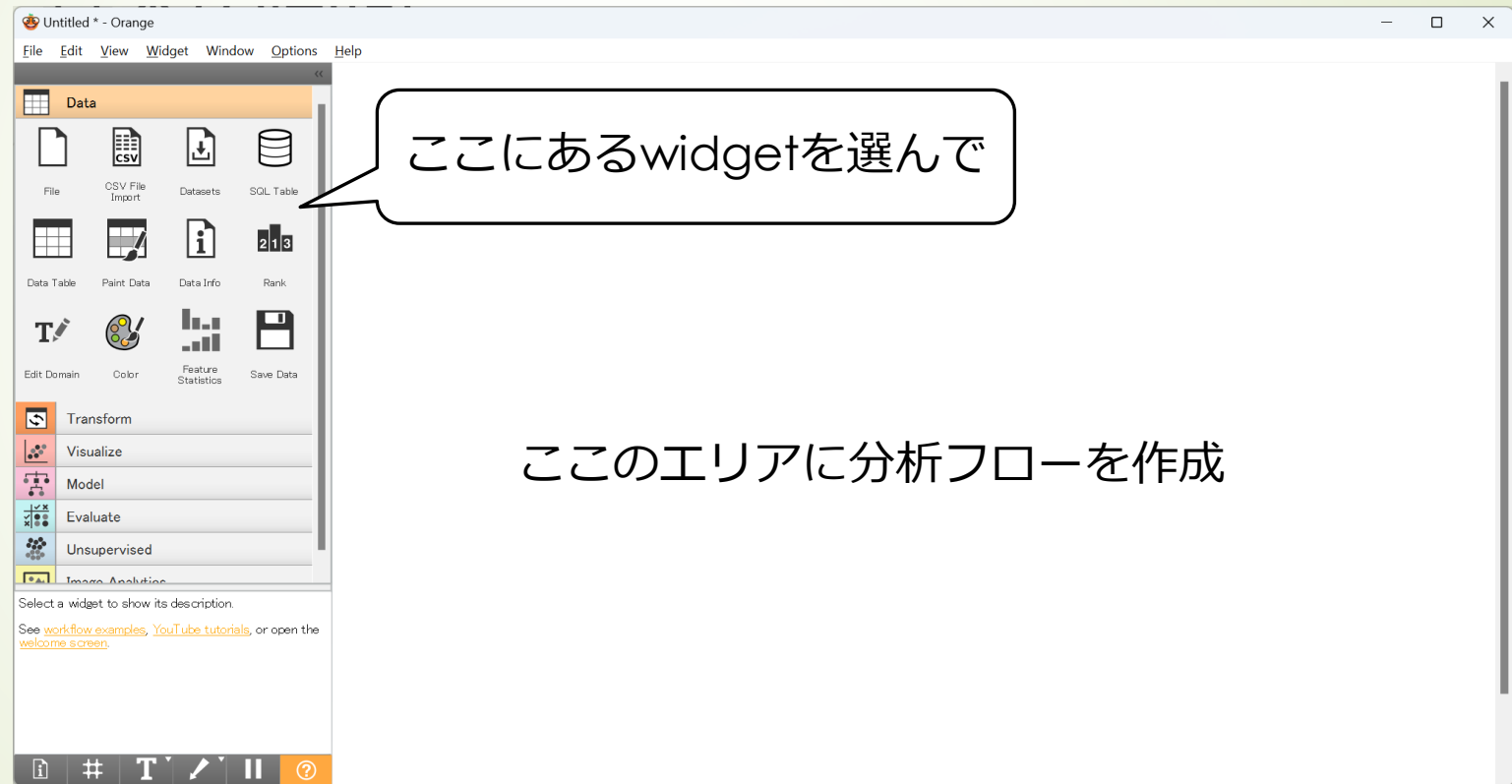


# 1分で機械学習モデルの作成・評価

- ▶ データ：Iris
- ▶ 目的変数：Iris（多クラス分類）
- ▶ アルゴリズム：Random Forest
- ▶ 評価方法：交差検定（ランダムサンプリング、5-fold）
- ▶ 評価指標：AUC, F1スコアなど諸々

機械学習の説明については省略します。

# 基本の画面



ここから1分でやっていきます。

# データの読み込み

The screenshot shows the Orange data mining software interface. On the left, the 'Data' widget is visible, with a red arrow pointing to the 'Datasets' icon. The 'Datasets' window is open, displaying a list of datasets. The 'Iris' dataset is selected and highlighted in blue. The window title is 'Irisデータを指定' (Specify Iris Data).

### Irisデータを指定

| Title                                 | Size      | Instances | Variables | Target      | Tags                          |
|---------------------------------------|-----------|-----------|-----------|-------------|-------------------------------|
| Iris                                  | 4.5 KB    | 150       | 5         | categorical | biology                       |
| Illegal waste dumpsites in Slovenia   | 2.8 MB    | 13165     | 25        | categorical | geo, timeseries, ecology      |
| Philadelphia Crime                    | 90.5 KB   | 9666      | 4         | categorical | criminology, time, geo        |
| Breast Cancer and Docetaxel Treatment | 1.8 MB    | 24        | 9486      | categorical | biology                       |
| Smoking effect on B lymphocytes       | 1.8 MB    | 79        | 3000      | categorical | genomics                      |
| HDI                                   | 45.2 KB   | 188       | 53        | categorical | economy, geo                  |
| SentiNews                             | 5.0 MB    | 2000      | 7         | categorical | text, sentiment               |
| TKI resistance                        | 1.2 MB    | 280       | 467       | categorical | spectral                      |
| Abalone                               | 187.5 KB  | 4177      | 8         | numeric     | biology                       |
| Adult                                 | 4.1 MB    | 32561     | 15        | categorical | economy                       |
| Roman Amphorae                        | 23.7 KB   | 164       | 16        | categorical | archaeology, image analytics  |
| Attrition - Predict                   | 838 bytes | 3         | 18        | categorical | economy, synthetic, education |

Description  
Iris (1936), from [UCI ML Repository](#)  
The Iris flower data set or Fisher's Iris data set was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper as an example of linear discriminant analysis. The data on length and width of petal and sepal leaves was actually collected by American botanist Edgar Anderson to quantify the morphologic variation of Iris flowers of three related species.  
See Also  
[Scatter Plots: the Tour](#)  
[All I See is Silhouette](#)  
References  
R. A. Fisher (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179-188.

# 目的変数と説明変数の選択

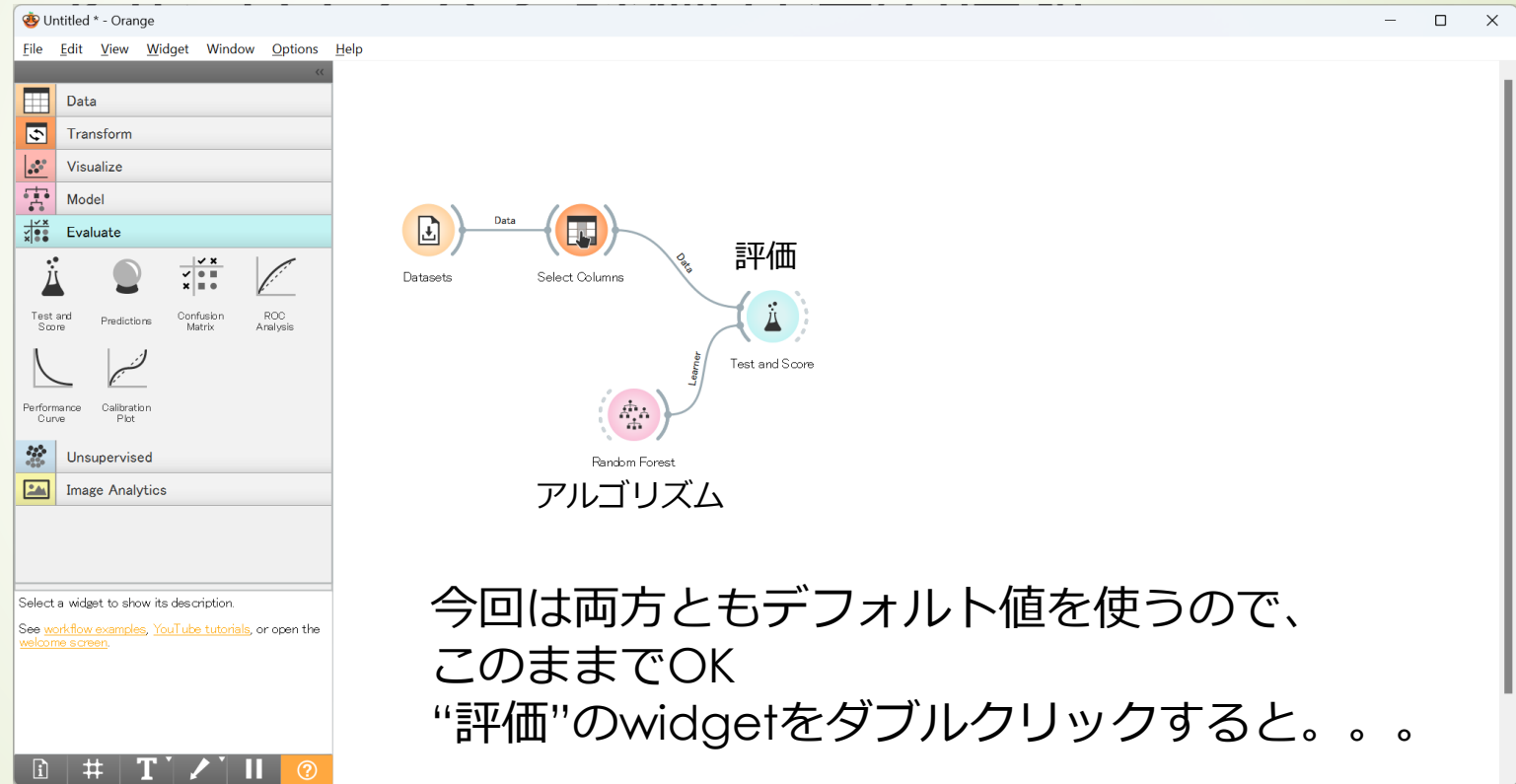
The screenshot displays the Orange data mining software interface. On the left, the 'Transform' widget palette is visible, with the 'Select Columns' widget selected. The main workspace shows a workflow with 'Datasets' and 'Select Columns' widgets. A 'Select Columns - Orange' dialog box is open, showing the configuration for the widget. The 'Features (4)' list includes 'sepal length', 'sepal width', 'petal length', and 'petal width'. The 'Target (1)' list includes 'iris'. The 'Metas' list is empty. The dialog box has a 'Reset' button, an 'Ignore new variables by default' checkbox, and a 'Send Automatically' checkbox. The text '説明変数' (Explanatory variables) is overlaid on the 'Features (4)' list, and '目的変数' (Target variable) is overlaid on the 'Target (1)' list.

説明変数

目的変数



# アルゴリズムと評価方法の選択



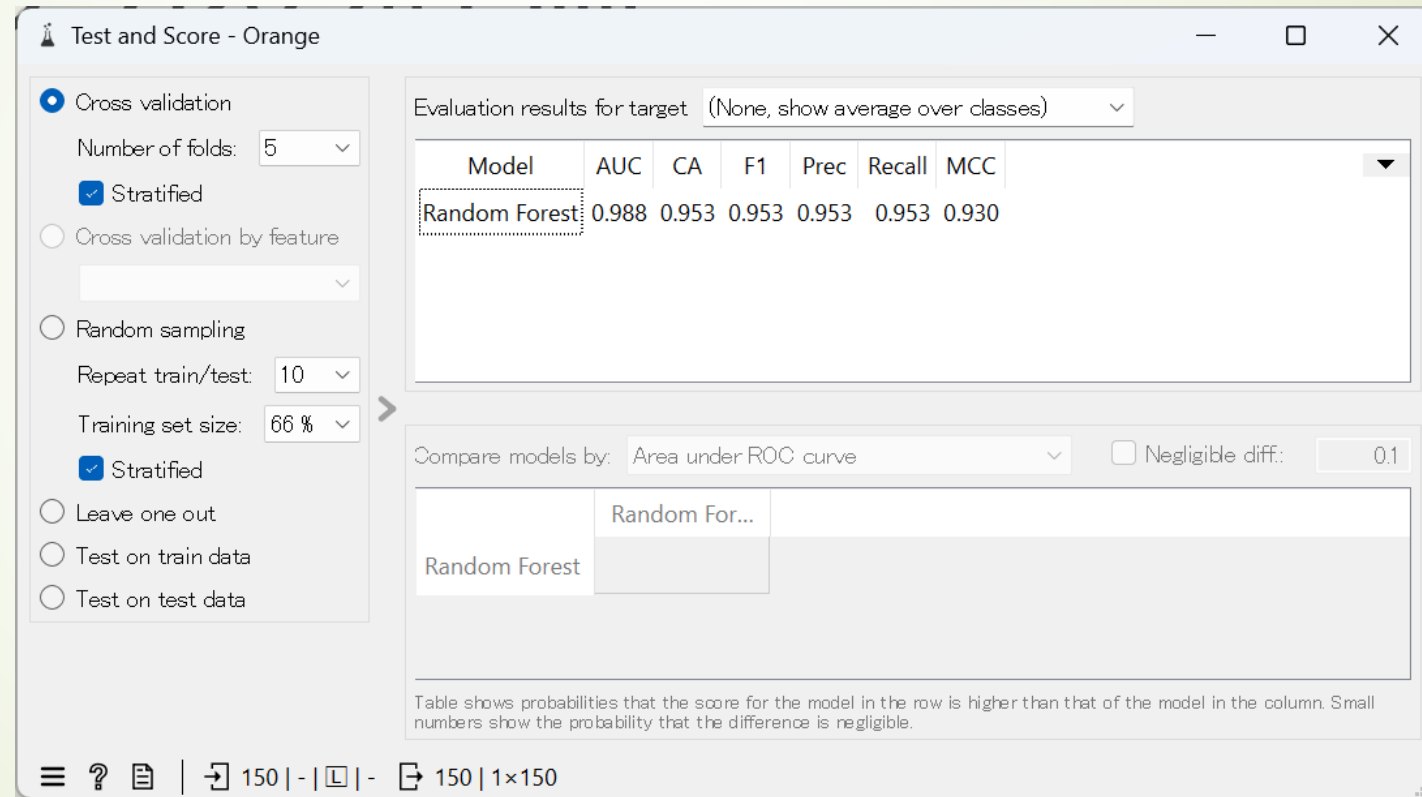
The screenshot shows the Orange data mining software interface. The left sidebar contains a widget palette with categories: Data, Transform, Visualize, Model, Evaluate, Unsupervised, and Image Analytics. The 'Evaluate' category is selected, showing widgets like Test and Score, Predictions, Confusion Matrix, ROC Analysis, Performance Curve, and Calibration Plot. The main workspace displays a workflow diagram with the following components and connections:

- Datasets** widget connected to **Select Columns** widget via a **Data** connection.
- Select Columns** widget connected to **Test and Score** widget via a **Data** connection.
- Random Forest** widget (under the **アルゴリズム** label) connected to **Test and Score** widget via a **Learner** connection.

The **Test and Score** widget is highlighted with a blue border and labeled **評価** (Evaluation). Below the workflow, the text reads: "今回は両方ともデフォルト値を使うので、このままでOK" (This time, since we use default values for both, it's OK as is) and "“評価”のwidgetをダブルクリックすると。。。" (When you double-click the "Evaluation" widget...).



# モデルの評価



Test and Score - Orange

Cross validation  
Number of folds: 5  
 Stratified  
 Cross validation by feature  
 Random sampling  
Repeat train/test: 10  
Training set size: 66 %  
 Stratified  
 Leave one out  
 Test on train data  
 Test on test data

Evaluation results for target (None, show average over classes)

| Model         | AUC   | CA    | F1    | Prec  | Recall | MCC   |
|---------------|-------|-------|-------|-------|--------|-------|
| Random Forest | 0.988 | 0.953 | 0.953 | 0.953 | 0.953  | 0.930 |

Compare models by: Area under ROC curve  
 Negligible diff.: 0.1

| Model         | Random For... |
|---------------|---------------|
| Random Forest |               |

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

150 | 150 | 1x150

ね、簡単でしょ？

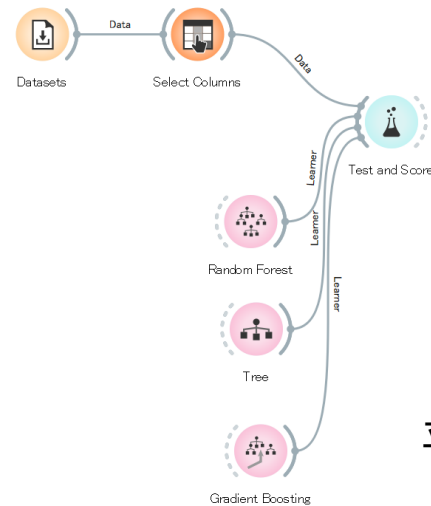
# 便利な機能① アルゴリズムの比較

Untitled \* - Orange  
File Edit View Widget Window Options Help

Data  
Transform  
Visualize  
Model

Constant CN2 Rule Induction Calibrated Learner K-NN  
Tree Random Forest Gradient Boosting SVM  
Linear Regression Logistic Regression Naive Bayes AdaBoost  
Curve Fit Neural Network Stochastic Gradient Descent Stacking

Test and Score  
Cross-validation accuracy estimation.  
[more...](#)

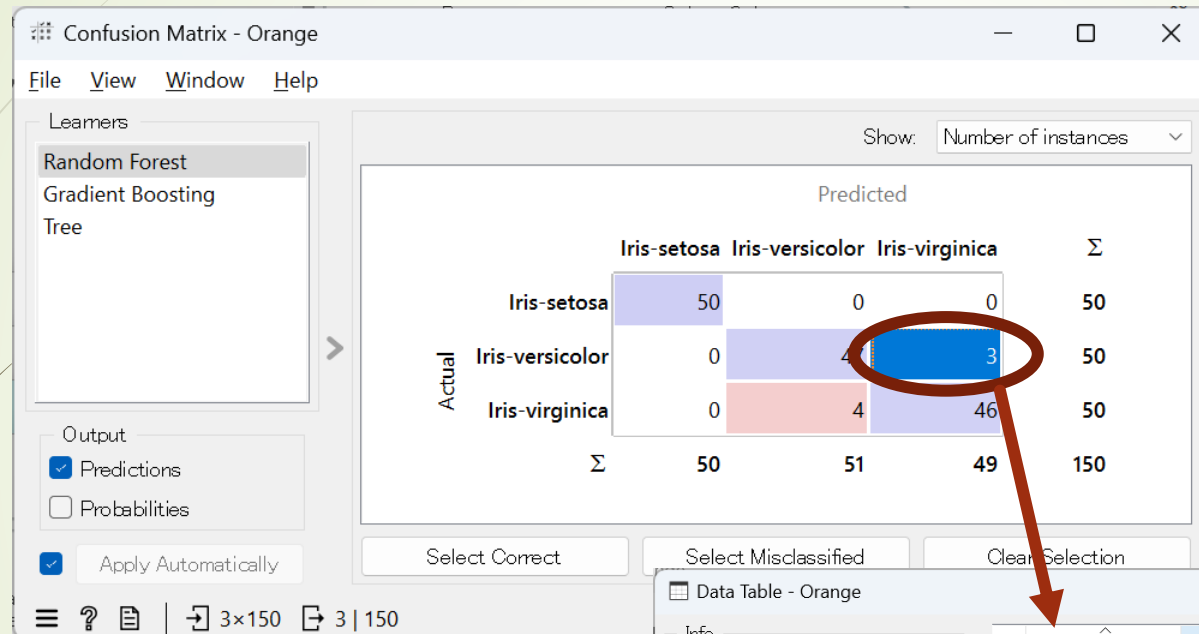


Evaluation results for target (None, show average over classes) ▼

| Model             | AUC   | CA    | F1    | Prec  | Recall | MCC   |
|-------------------|-------|-------|-------|-------|--------|-------|
| Random Forest     | 0.988 | 0.953 | 0.953 | 0.953 | 0.953  | 0.930 |
| Gradient Boosting | 0.989 | 0.953 | 0.953 | 0.953 | 0.953  | 0.930 |
| Tree              | 0.976 | 0.960 | 0.960 | 0.960 | 0.960  | 0.940 |

平行に配置するだけ。簡単！

## 便利な機能② GUIでのデータ抽出



混同行列の表示

表示したい項目を選ぶだけで、  
該当のデータを抽出可能

Data Table - Orange

Info: 3 instances (no missing data), 4 features, Target with 3 values, 1 meta attribute

Variables: Show variable labels (if present), Visualize numeric values, Color by instance classes

Selection: Select full rows

Buttons: Restore Original Order, Send Automatically

|   | $\hat{y}$       | $y$            | sepal length | sepal width | petal length | petal width |
|---|-----------------|----------------|--------------|-------------|--------------|-------------|
| 1 | Iris-versicolor | Iris-virginica | 5.9          | 3.2         | 4.8          | 1.4         |
| 2 | Iris-versicolor | Iris-virginica | 6.0          | 2.7         | 5.1          | 1.4         |
| 3 | Iris-versicolor | Iris-virginica | 6.7          | 3.0         | 5.0          | 1.4         |

# ご清聴、ありがとうございました。

他にも話したいネタがたくさんあります

- ▶ 数式の前に知っておきたい、実務における機械学習の考え方
- ▶ 実務で使える Orange Data Mining の便利な機能
- ▶ 組織の基礎レベルを上げる R AnalyticFlowの紹介
- ▶ 意識低い系のデータ分析プロジェクトの進め方
- ▶ 構造で考えるDSプロジェクトの課題の選び方
- ▶ 研修って実際に役に立つの？DS"技能"教育のススメ など

今後もLTでいろんなお話をしていきたいです。内容によっては連続講座になるかも。  
お気軽にお声がけください。



Twitter : @bunnchinn3