2023年5月18日
データサイエンティスト集会 in VRC

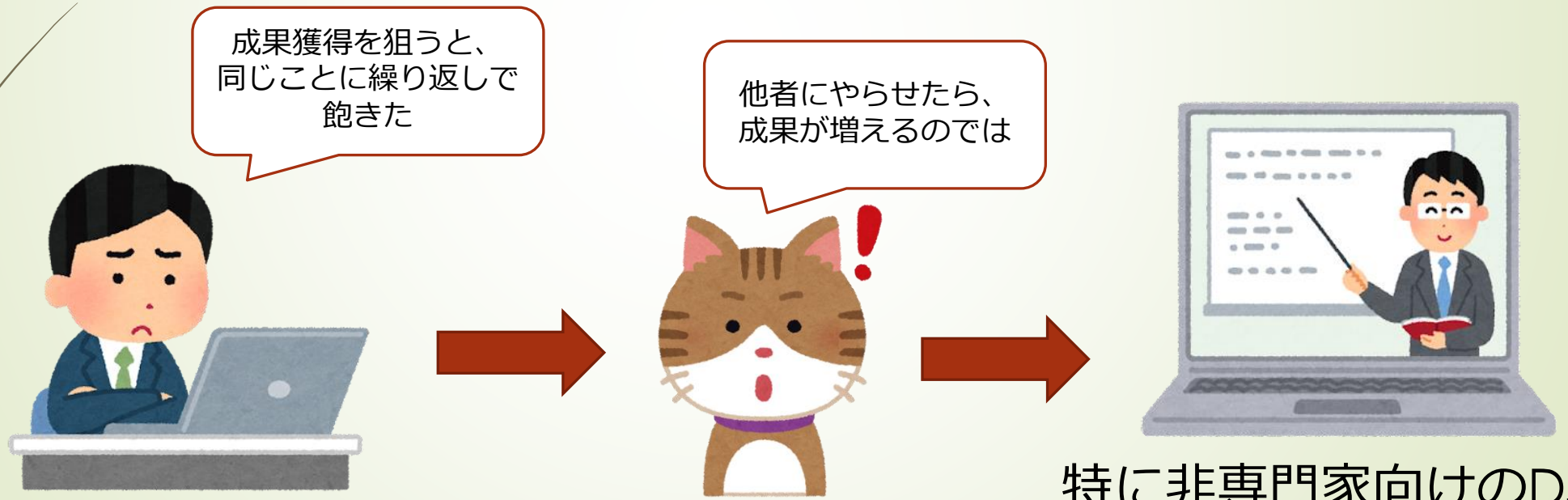
# １分で作って評価する機械学習モデル Orange Data Miningの紹介

ぶんちん

# 自己紹介　ぶんちん

VRChat初めて約２か月です

- 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。。。




特に非専門家向けのDS教育

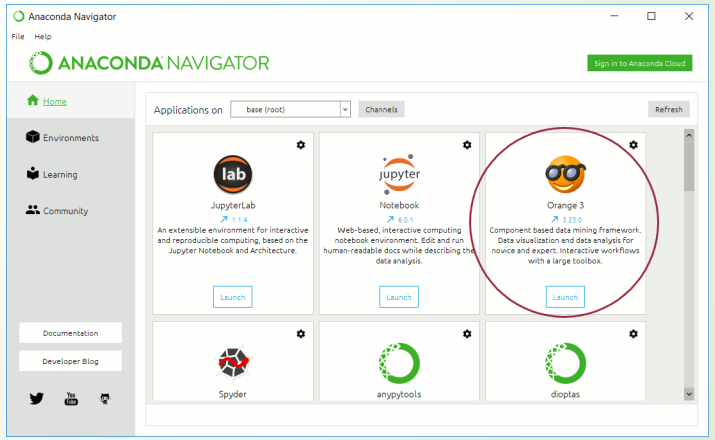
# Orange Data Mining

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**

Anaconda使っている人は
見たことあるのでは？

- 初心者は勉強に使おう！
- 専門家は手抜き・教育に使おう！

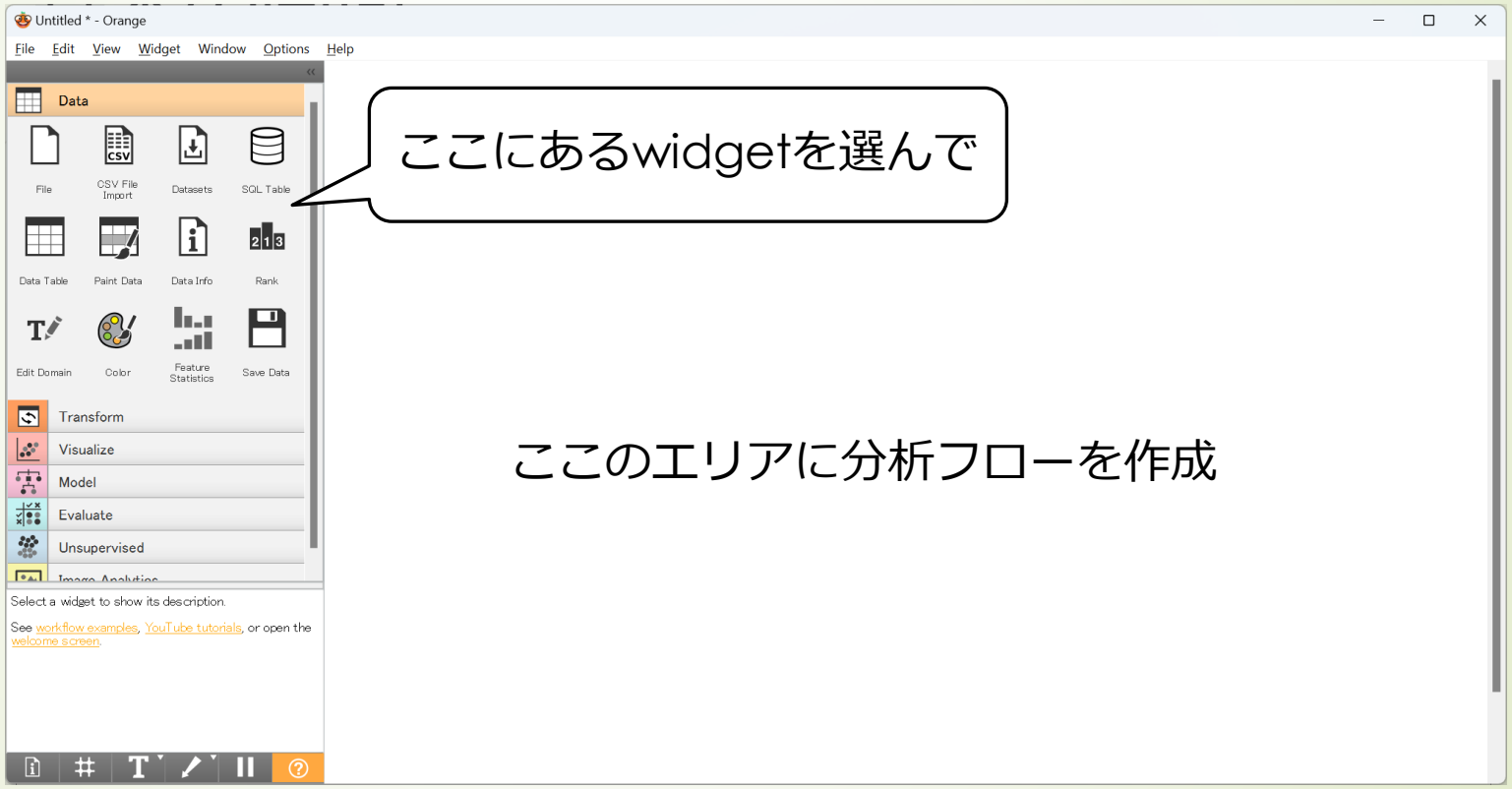


# １分で機械学習モデルの作成・評価

- 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

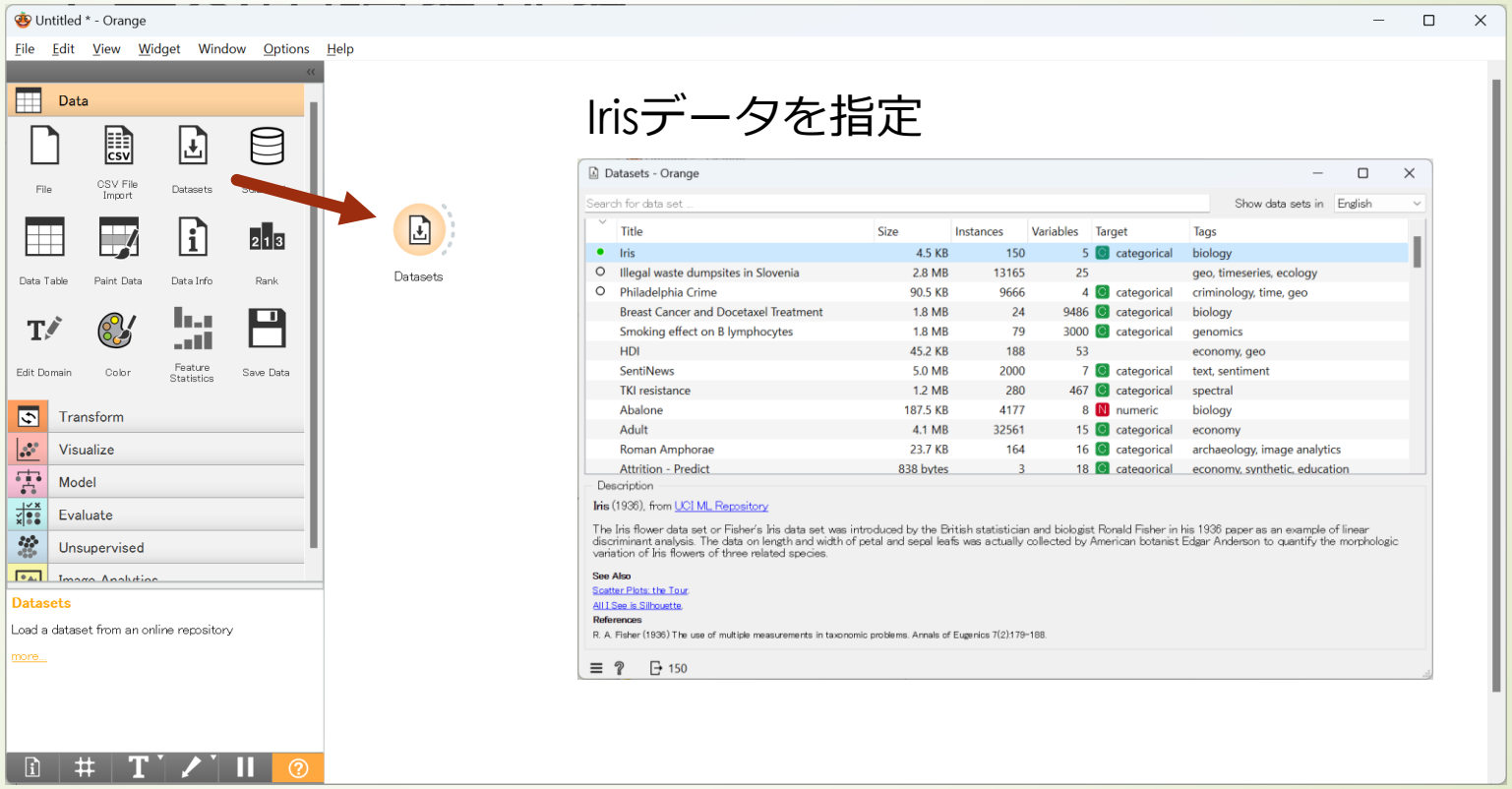
機械学習の説明については省略します。

# 基本の画面

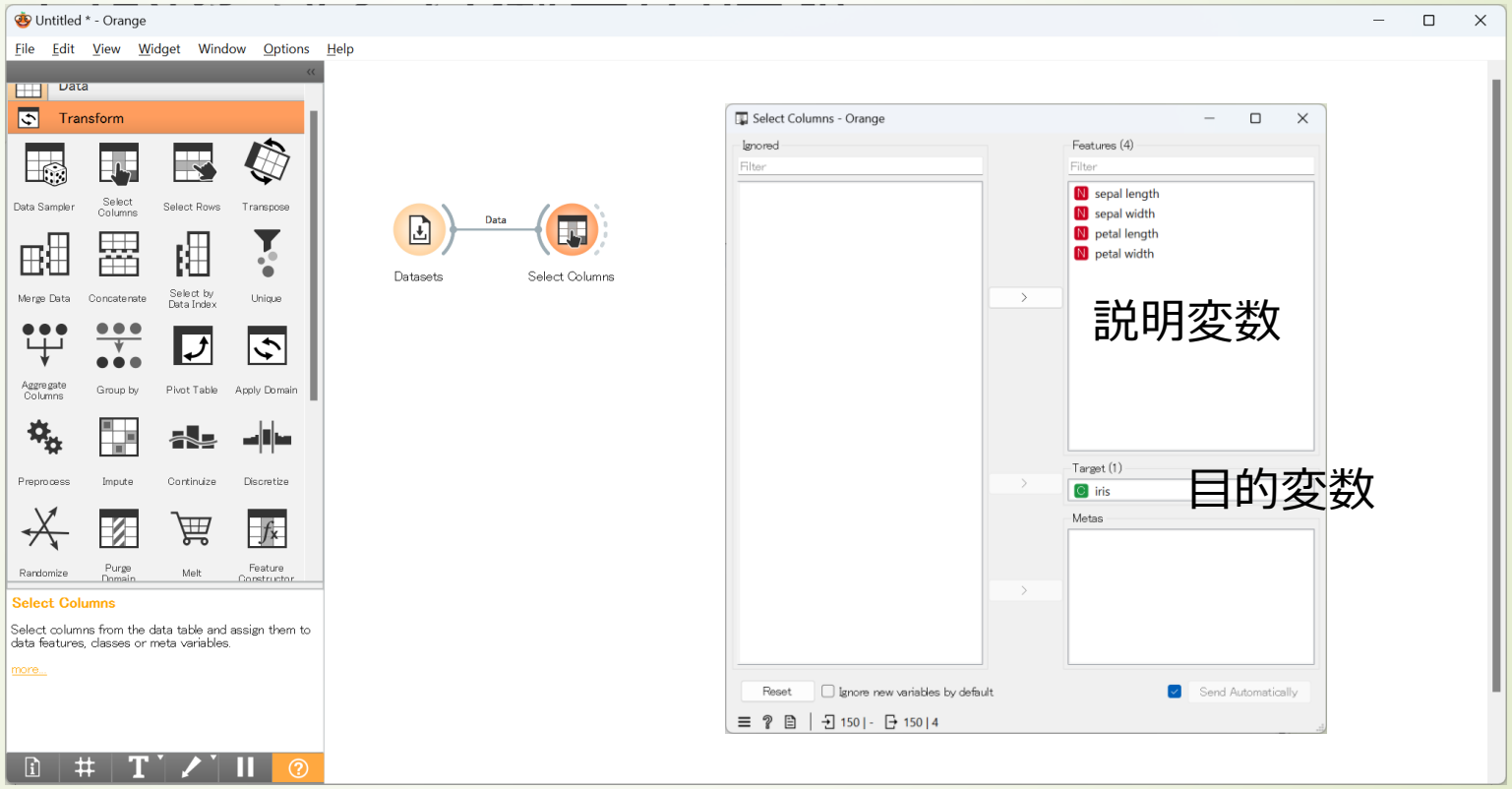



ここから１分でやっていきます。

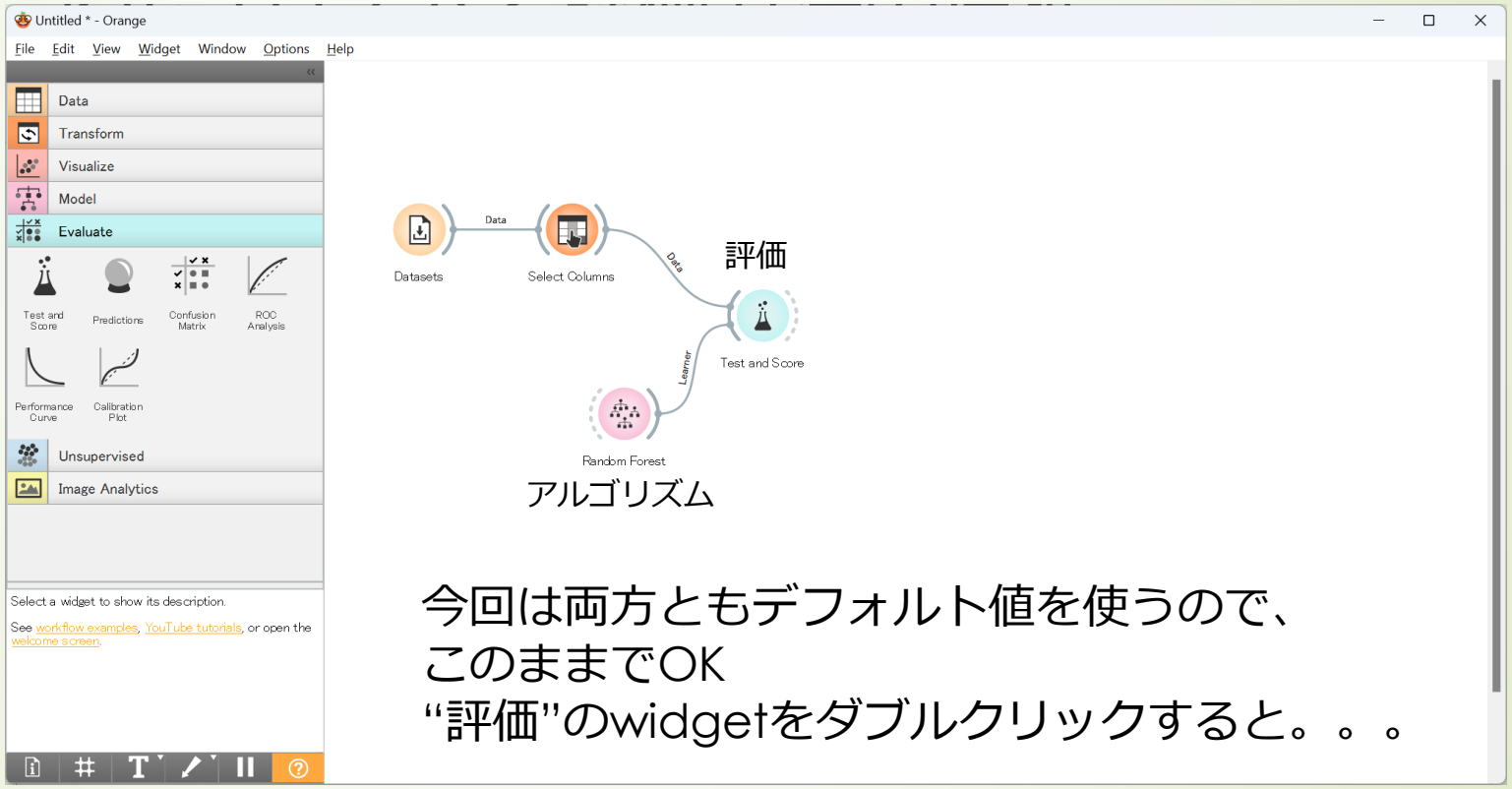
# データの読み込み


Untitled * - Orange
File Edit View Widget Window Options Help
Data
File
CSV File Import
Datasets
Data Table
Paint Data
Data Info
Rank
Edit Domain
Color
Feature Statistics
Save Data
Transform
Visualize
Model
Evaluate
Unsupervised
Datasets
Load a dataset from an online repository
more...
Datasets
Irisデータを指定
Datasets - Orange
Search for data set ...
Show data sets in English
Title Size Instances Variables Target Tags
Iris 4.5 KB 150 5 categorical biology
Illegal waste dumpsites in Slovenia 2.8 MB 13165 25 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 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 leafs 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.
150

# 目的変数と説明変数の選択

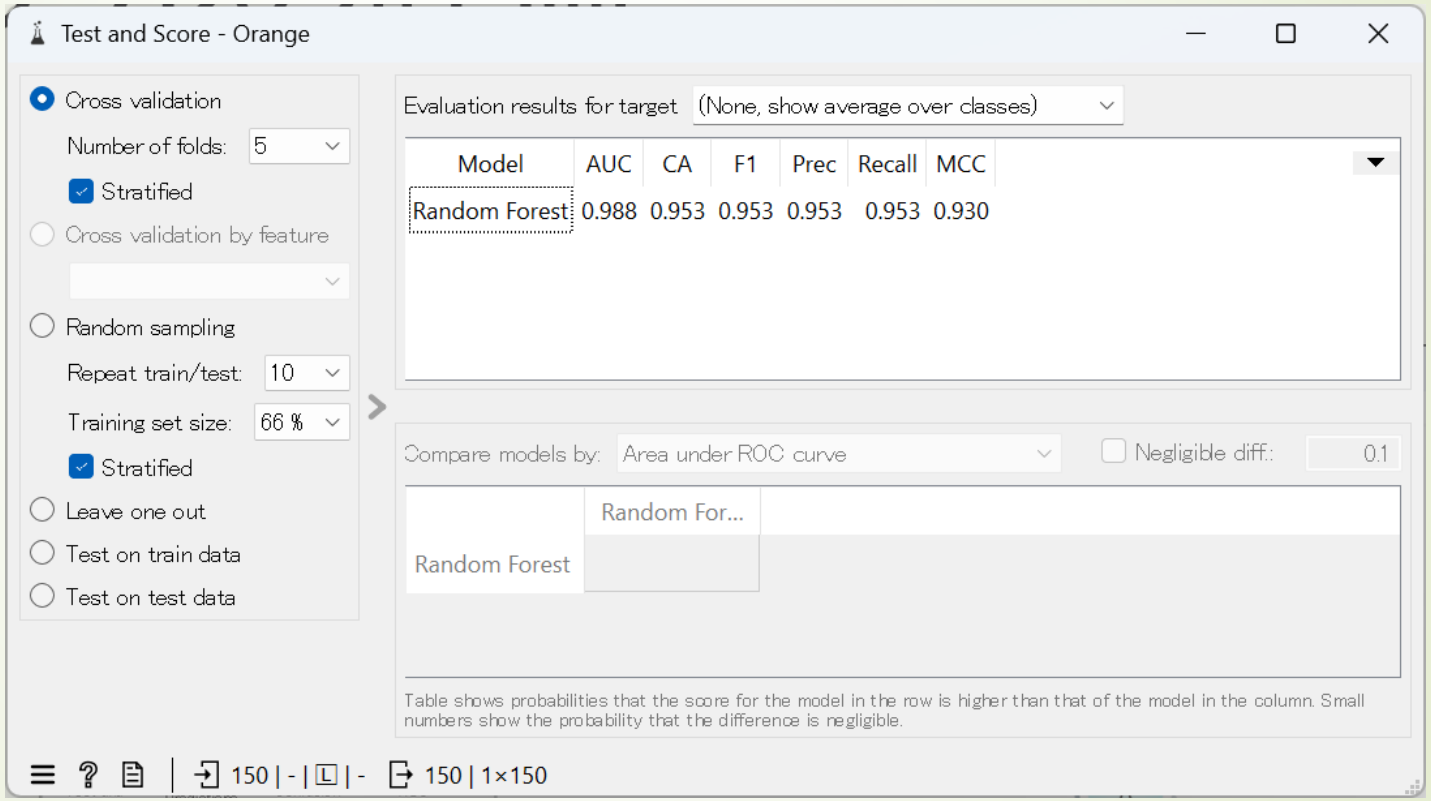


# アルゴリズムと評価方法の選択




今回は両方ともデフォルト値を使うので、
このままでOK
“評価”のwidgetをダブルクリックすると。。。

# モデルの評価

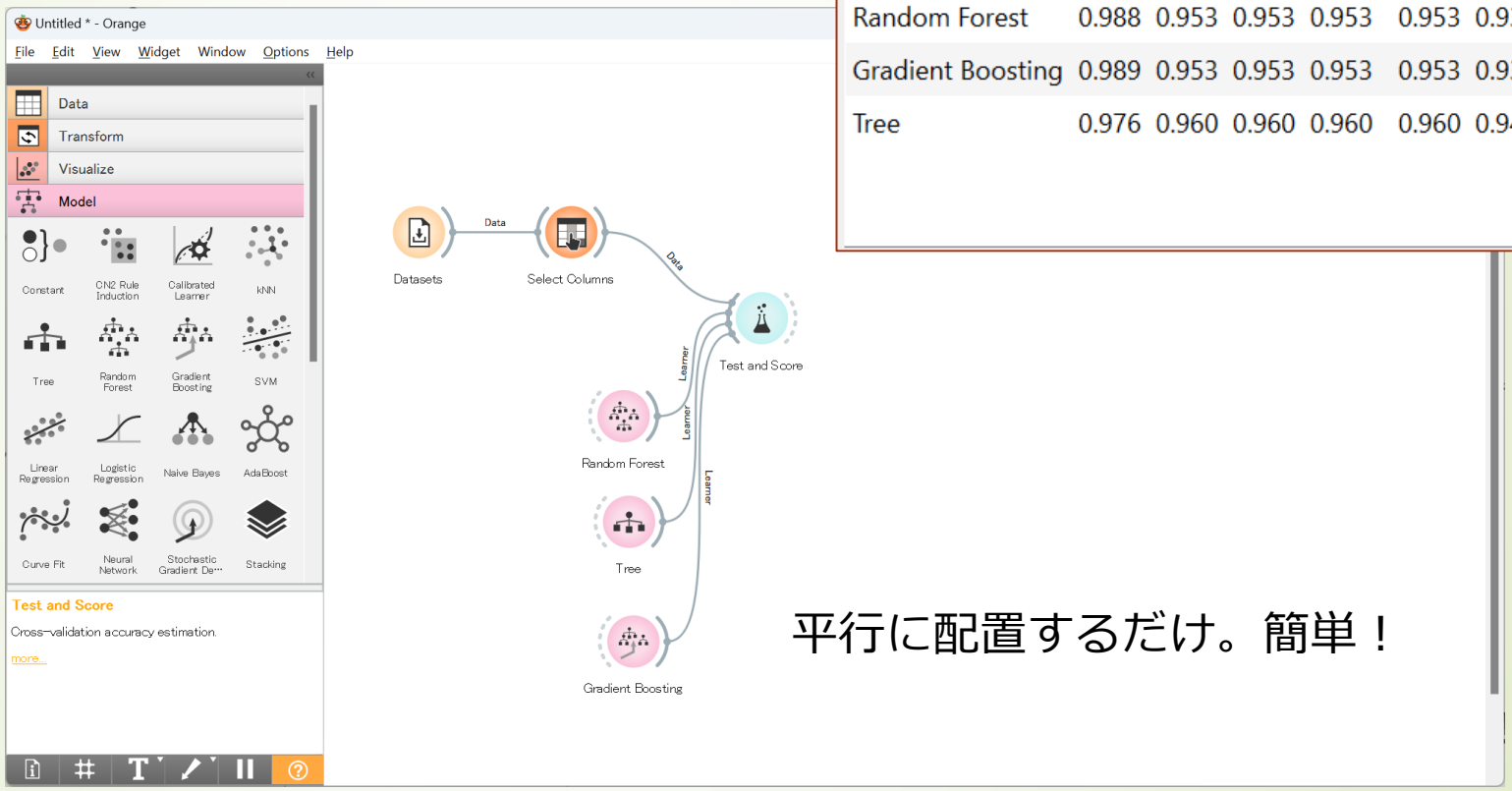



ね、簡単でしょ？

# 便利な機能①　アルゴリズムの比較

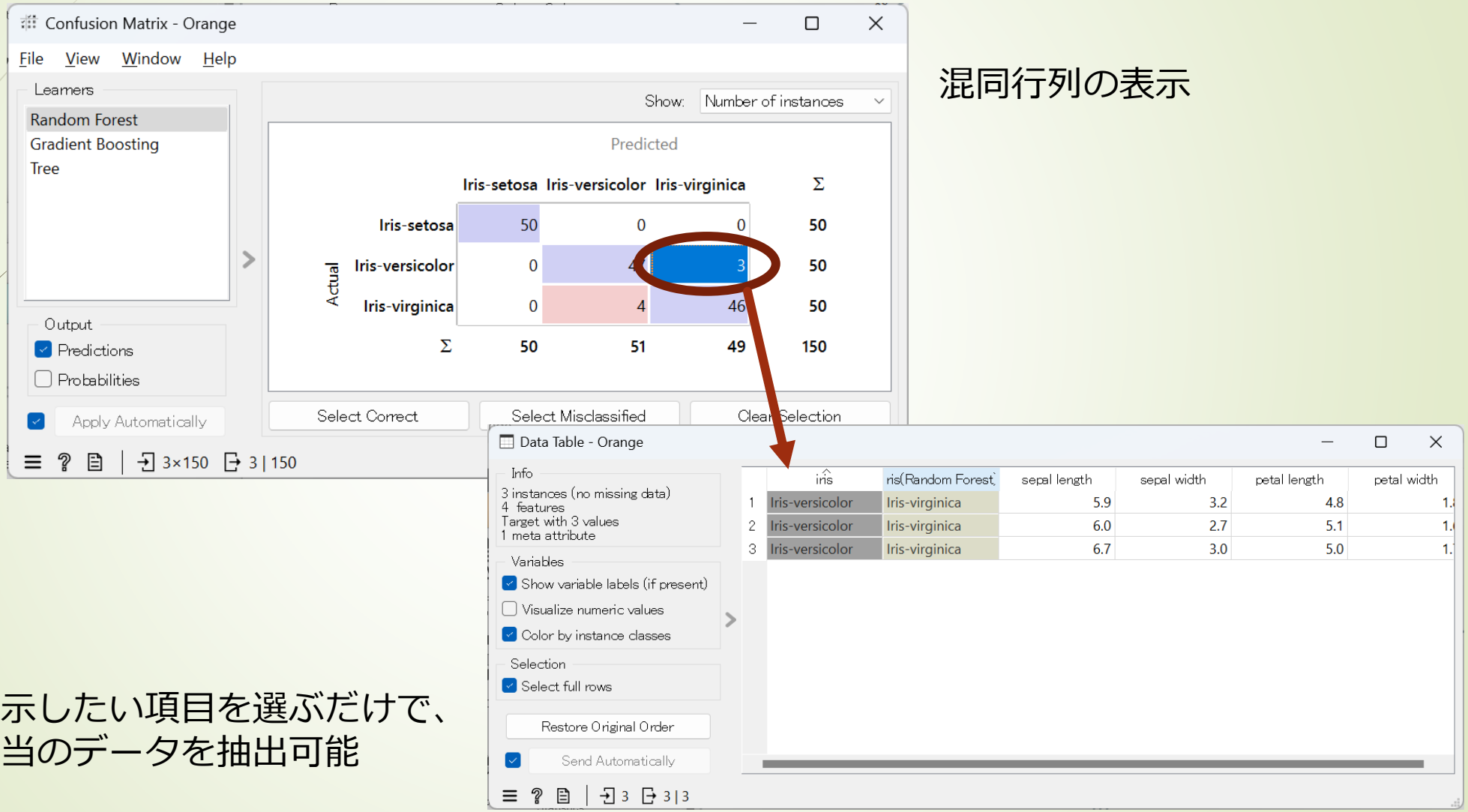
Evaluation results for target (None, show average over classes)

| 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

# 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


Twitter：@bunnchinn3

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。