UNIVERSITAS PELITA HARAPAN

Data Science

Introduction to Orange3

Training Series

Week 1

18 Januari 2024

Orange3

Prepared by I Made Murwantara, Universitas Pelita Harapan

Overview

1. What is Orange3
2. Resources and Installation
3. Visual Modelling
4. Dataset
5. Basic Module & Library
6. Add-on
7. First Try

What is Orange3

- University of Ljubljana, Slovenia
- https://orange.biolab.si/
- Orange Widgets and Canvas are based on Qt,
  - which is distributed under GPL 3.0 (as well as LGPL 2.1, see https://www.qt.io/licensing/
- Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research*14(Aug): 2349−2353.
- A visual programming tools for Data Science

Resources

- https://orangedatamining.com
- Windows & MacOS, Linux (via pip)
- Global Discussion via datascience.stackexchange.com
- Documents: http://file.biolab.si/notes/
- Blog: https://orangedatamining.com/blog/

Installation

https://orangedatamining.com/download/

Windows

Standalone installer (default)

Orange3-3.36.2-Miniconda-x86_64.exe (64 bit)

Can be used without administrative priviledges.

MacOS

Orange for Apple silicon

Orange3-3.36.2-Python3.9.12-arm64.dmg

Orange for Intel

Orange3-3.36.2-Python3.9.12.dmg

pip

Orange can also be installed from the Python Package Index. You may need additional system packages provided by your distribution.

```
pip install orange3
```

Installing from Source

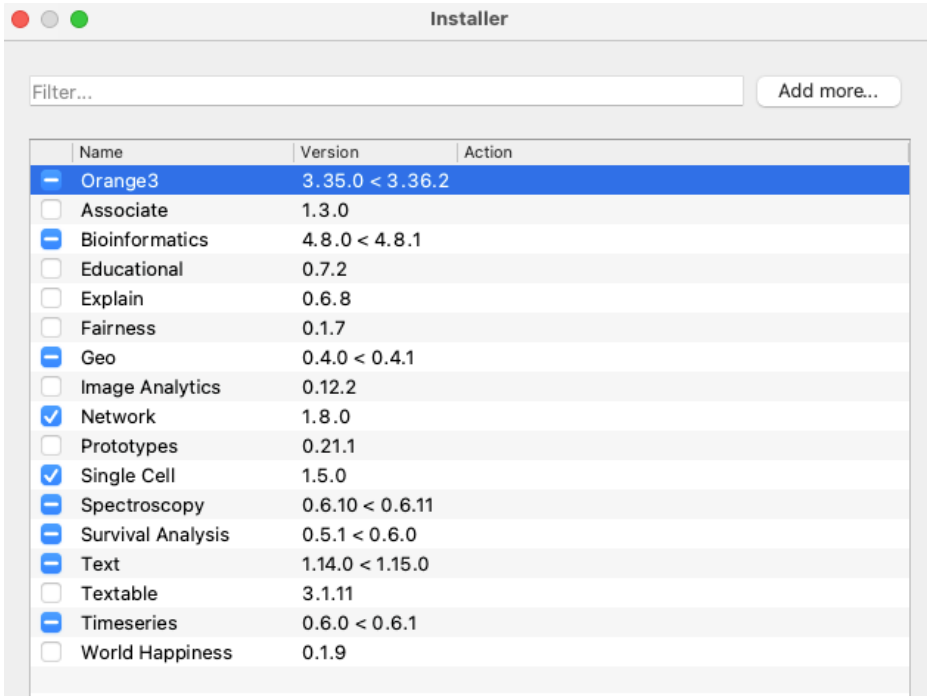
Clone our repository from GitHub or download the source code tarball. Then follow the instructions in README.md

To run Orange Canvas run

```
python -m Orange.canvas
```

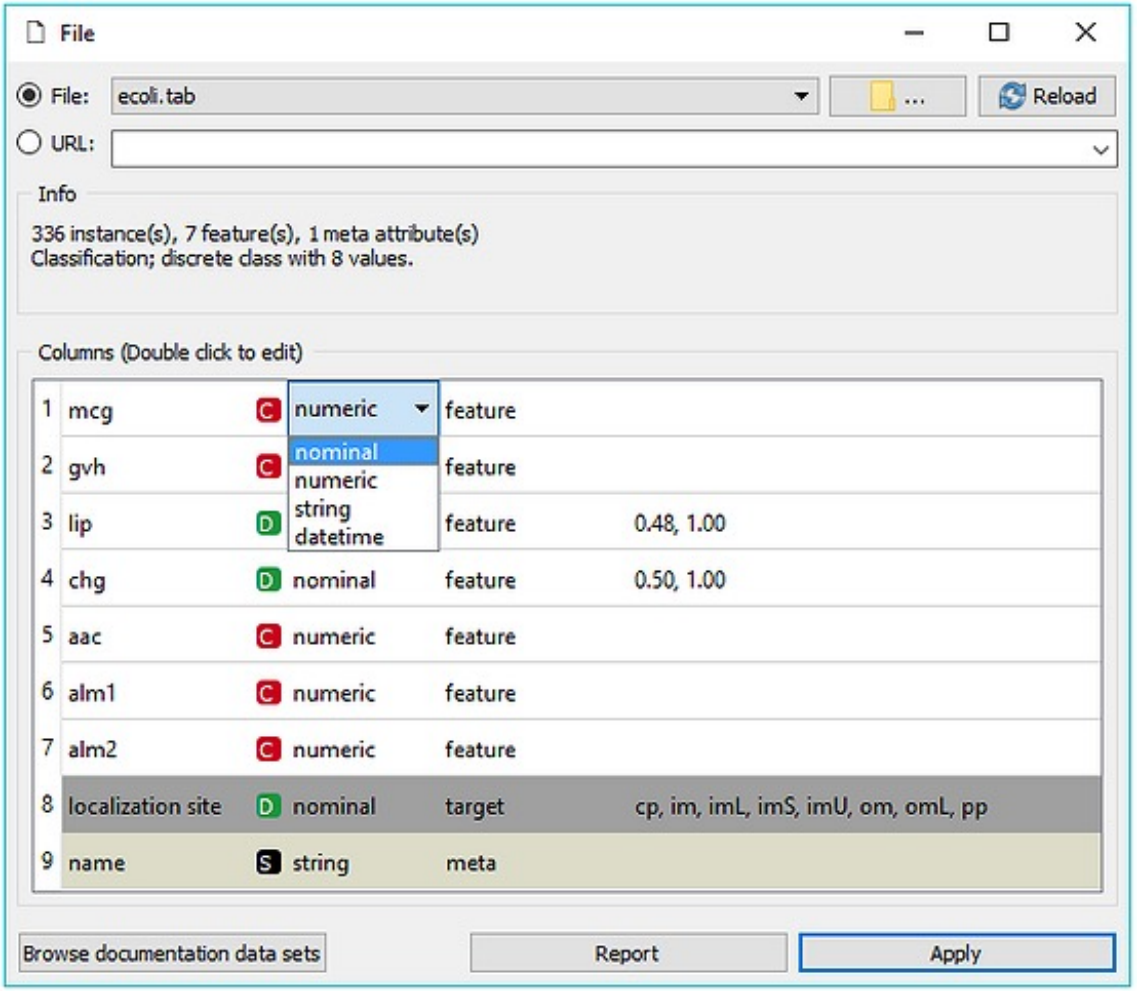
Add-On

1. Orange3-Associate
2. Orange3-Bioinformatics
3. Orange3-Educational
4. Orange3-Explain
5. Orange3-Fairness
6. Orange3-Geo
7. Orange3-ImageAnalytics
8. Orange3-Network
9. Orange3-Prototypes
10. Orange3-SingleCell
11. Orange-Spectroscopy
12. Orange3-Survival-Analysis
13. Orange3-Text
14. Orange3-Textable
15. Orange3-Timeseries
16. Orange3-WorldHappiness



Screenshot – Dataset

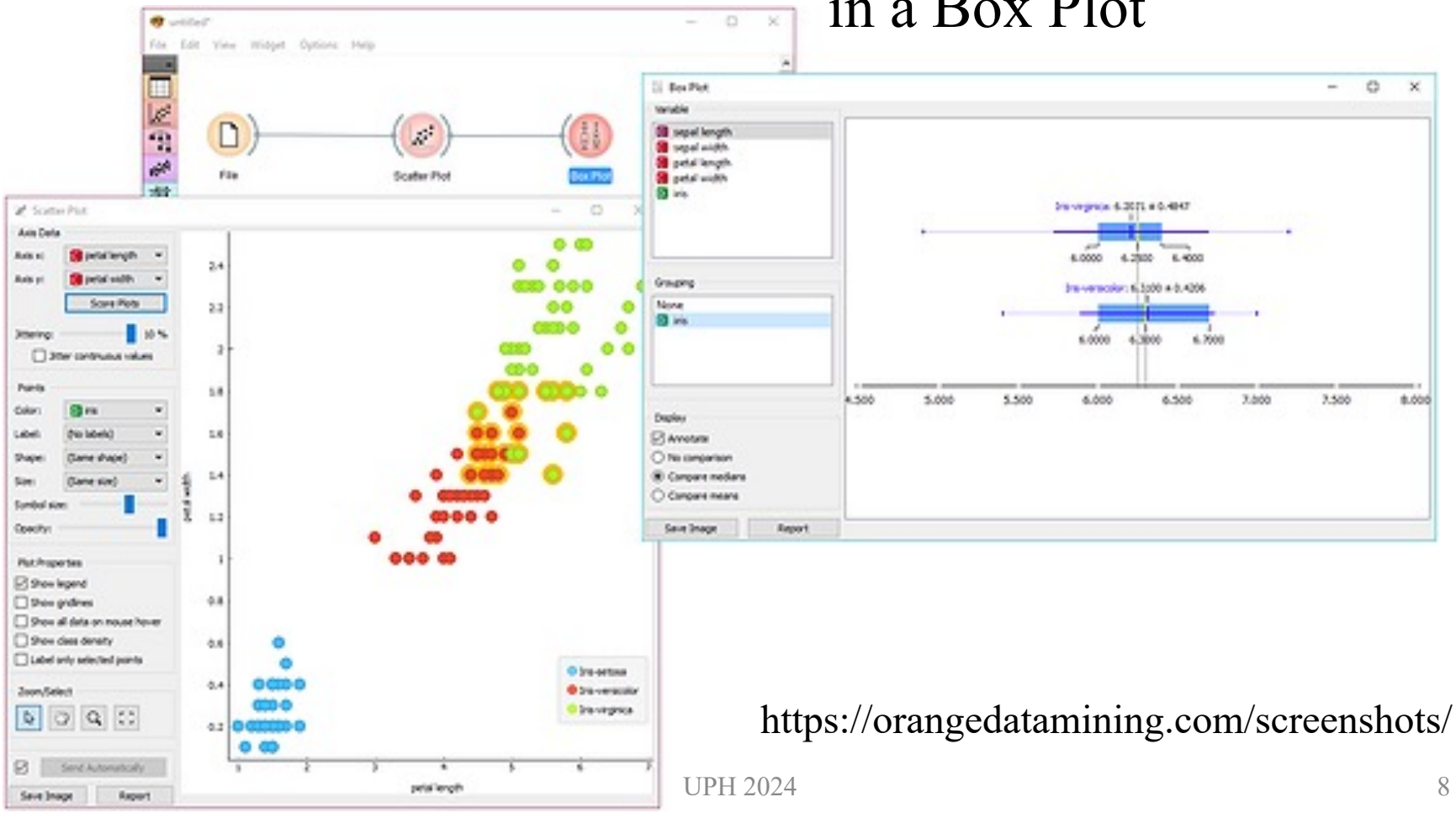
Load & Edit data in the file widget



https://orangedatamining.com/screenshots/

Screenshot – Interactive Visualisation

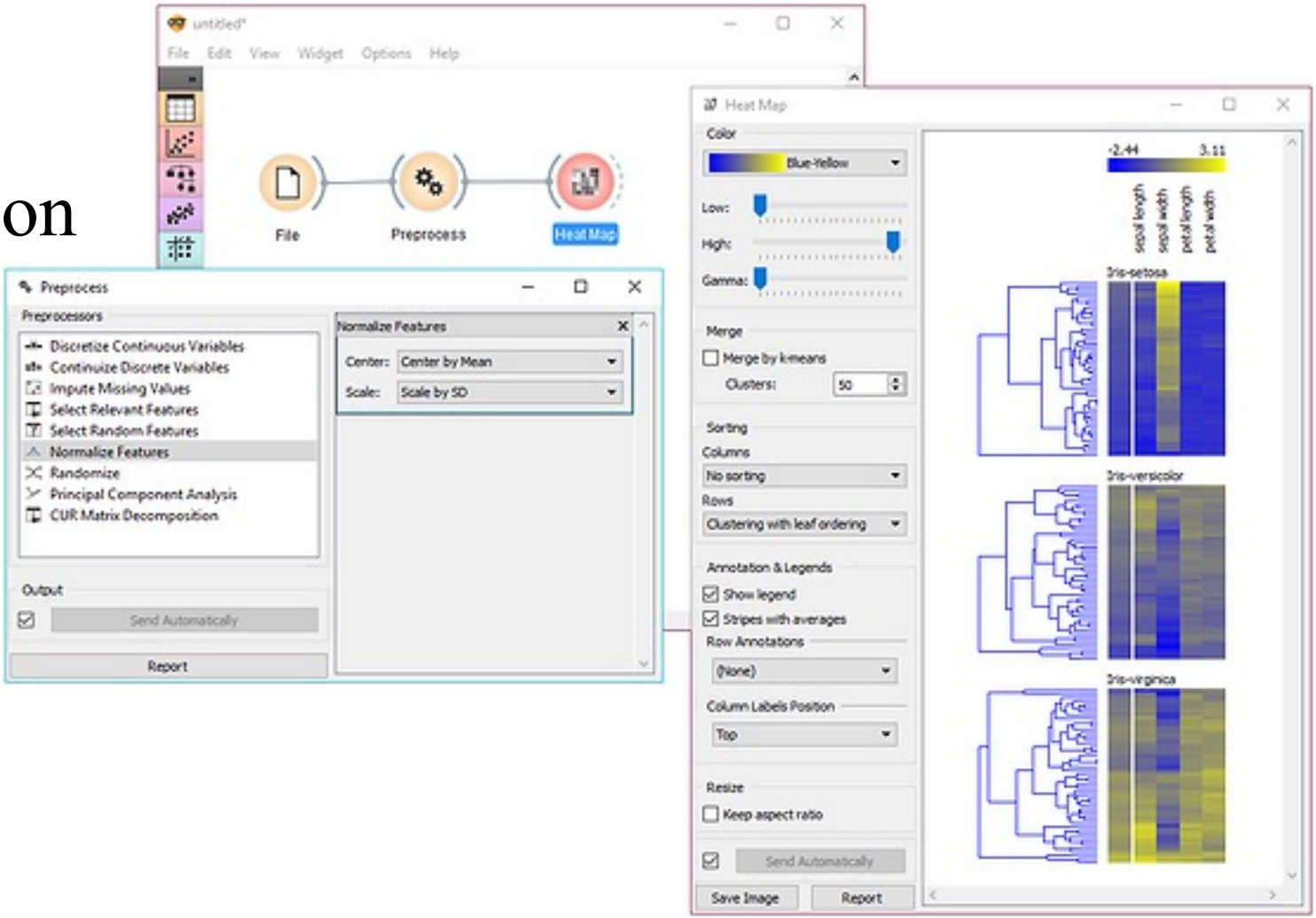
Data Selection in Scatter Plot visualized in a Box Plot



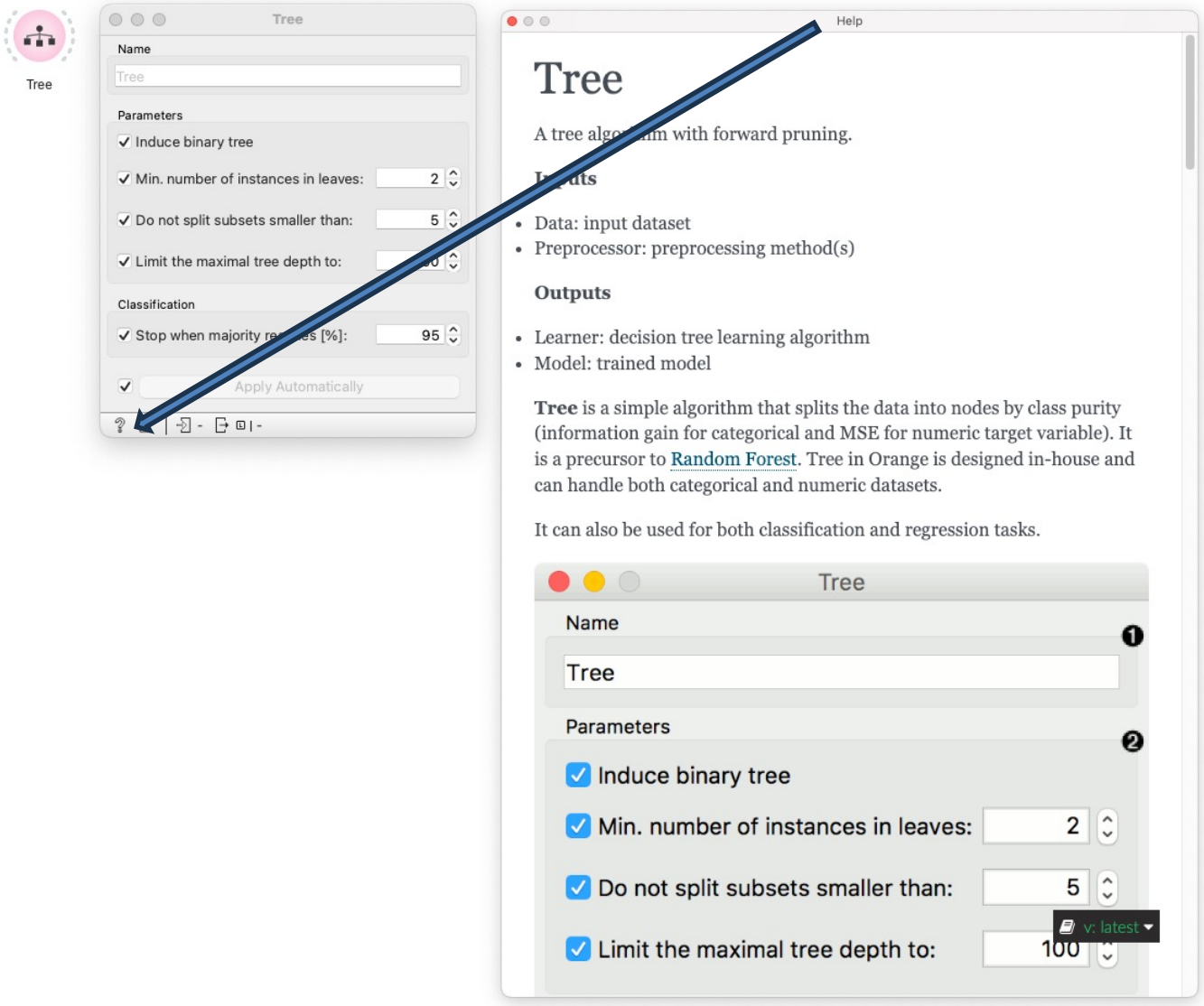
https://orangedatamining.com/screenshots/

Screenshot – Analytic Visualisation

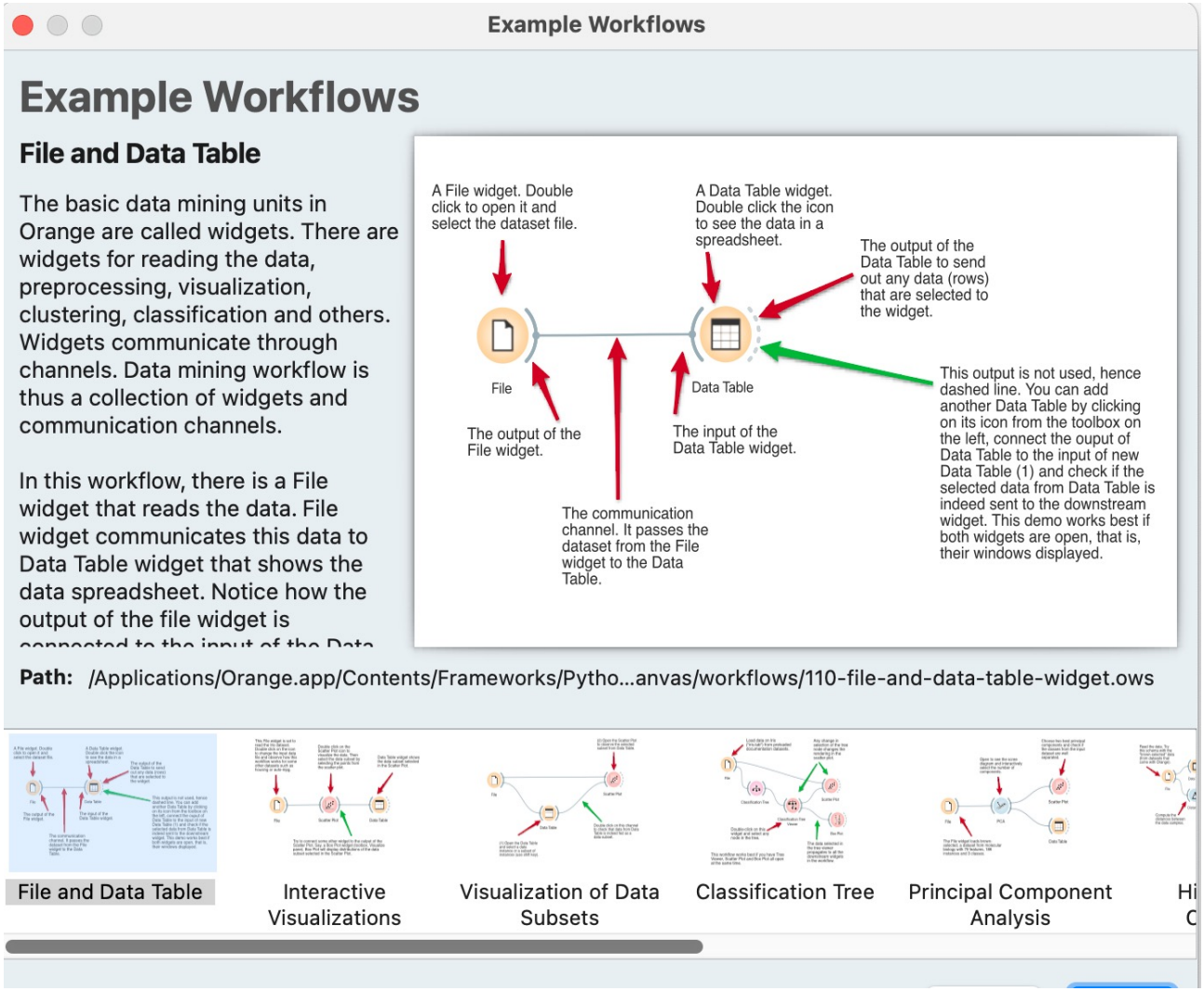
Heatmap Visualisation



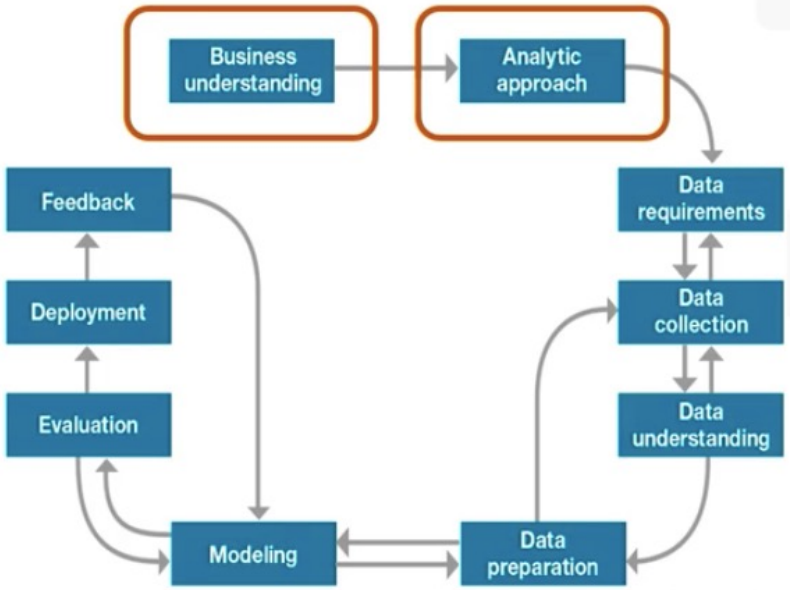
Help


Tree
Tree
Name
Tree
Parameters
Induce binary tree
Min. number of instances in leaves: 2
Do not split subsets smaller than: 5
Limit the maximal tree depth to:
Classification
Stop when majority reaches [%]: 95
Apply Automatically
Help
Tree
A tree algorithm with forward pruning.
Inputs
Data: input dataset
Preprocessor: preprocessing method(s)
Outputs
Learner: decision tree learning algorithm
Model: trained model
Tree is a simple algorithm that splits the data into nodes by class purity (information gain for categorical and MSE for numeric target variable). It is a precursor to Random Forest. Tree in Orange is designed in-house and can handle both categorical and numeric datasets.
It can also be used for both classification and regression tasks.
Tree
Name
Tree
Parameters
Induce binary tree
Min. number of instances in leaves: 2
Do not split subsets smaller than: 5
Limit the maximal tree depth to: 100
v: latest

Example Workflows



Data Science Methodology



Business understanding

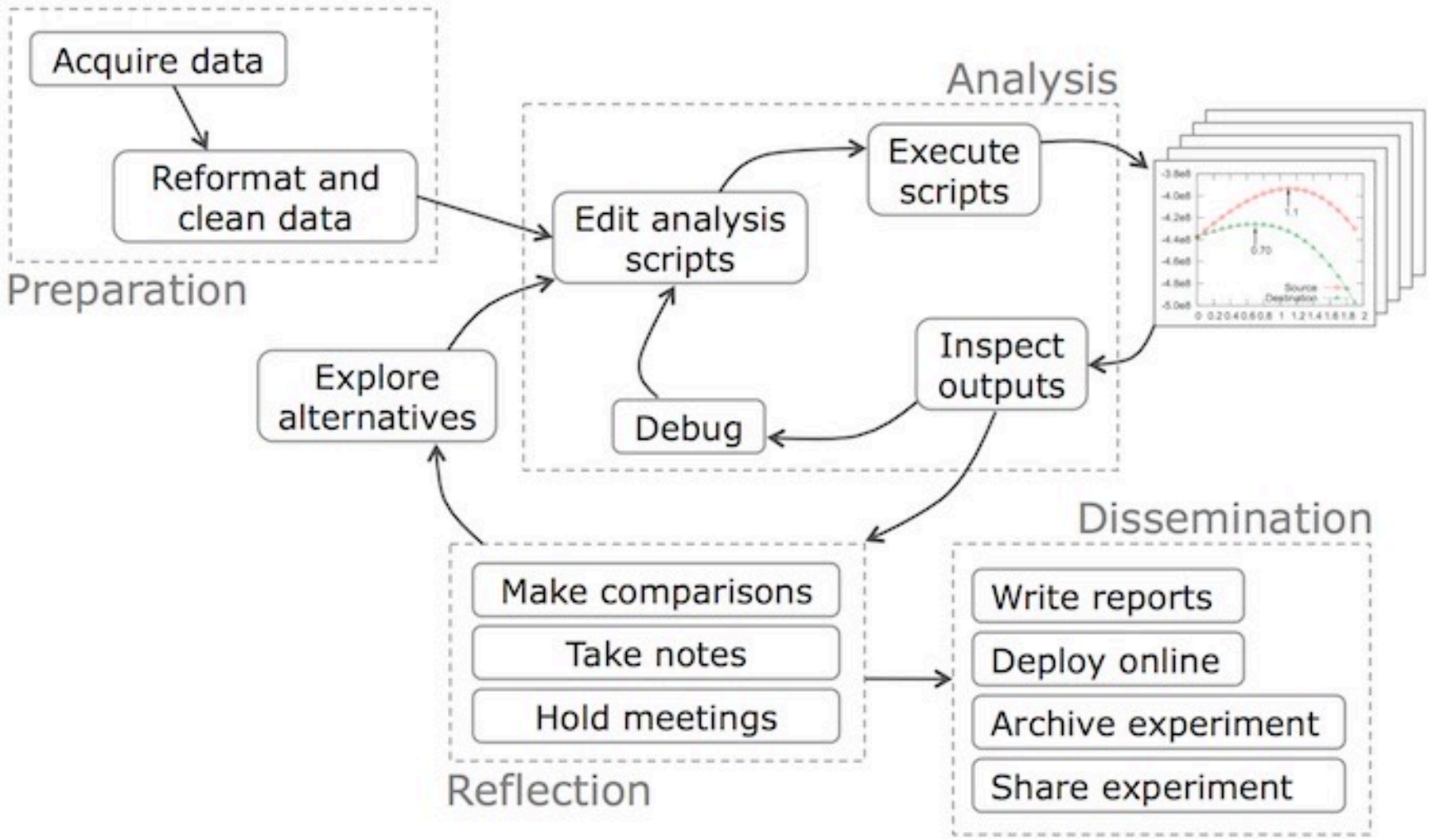
- *What is the problem that you are trying to solve?*

Analytic approach

- *How can you use data to answer the question?*

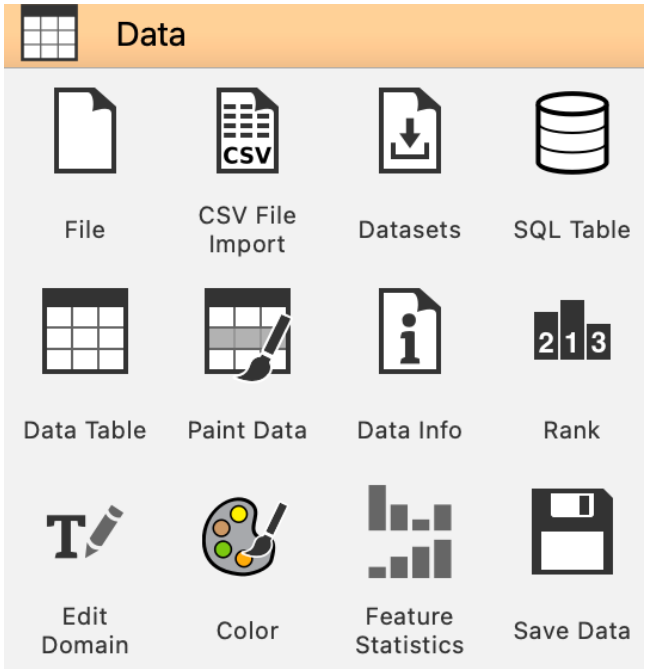
https://www.kaggle.com/code/xiuwenbo666/data-science-methodology

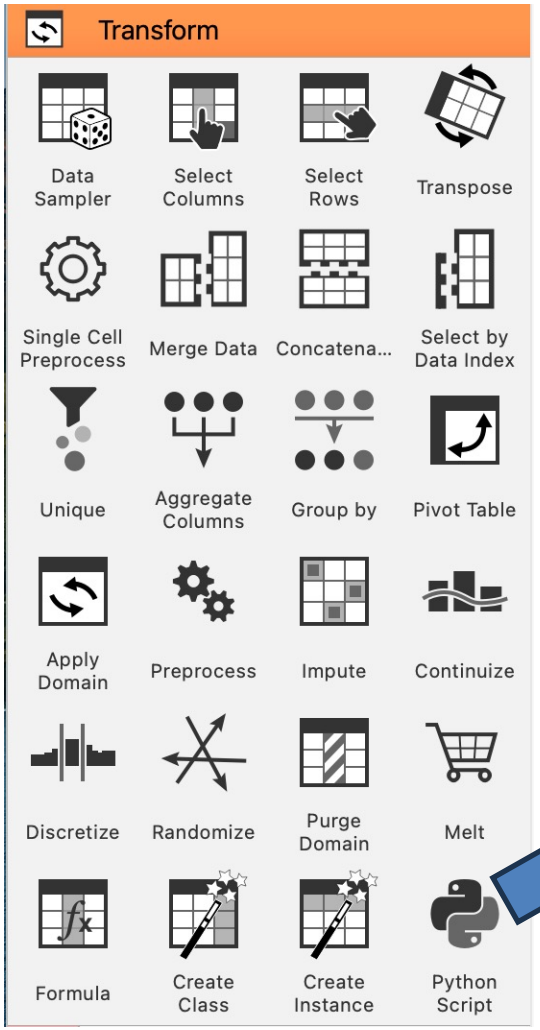
Data Science Workflow



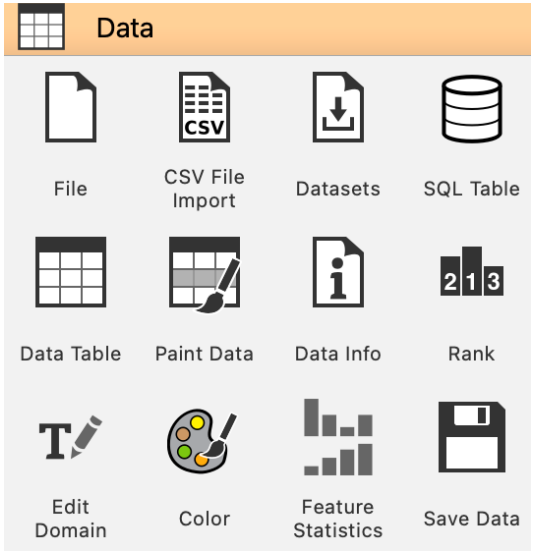
https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

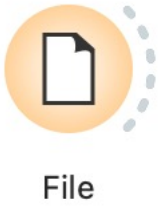
Data Engineering

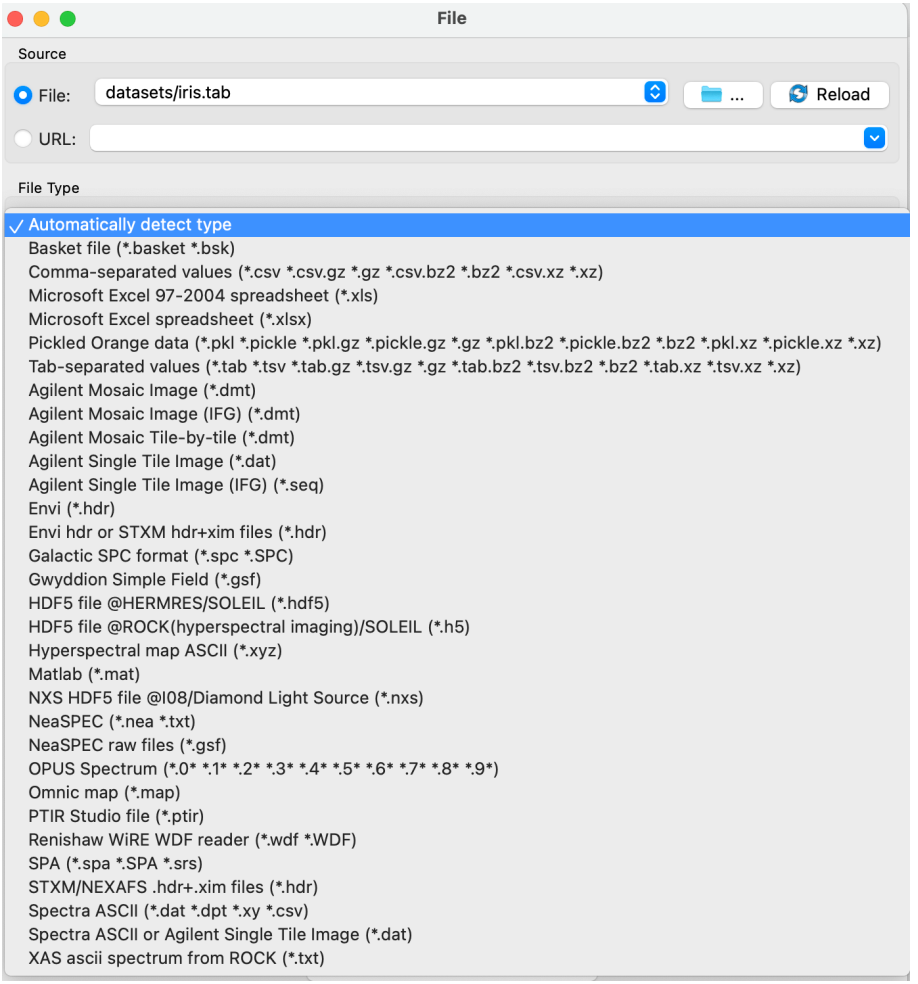


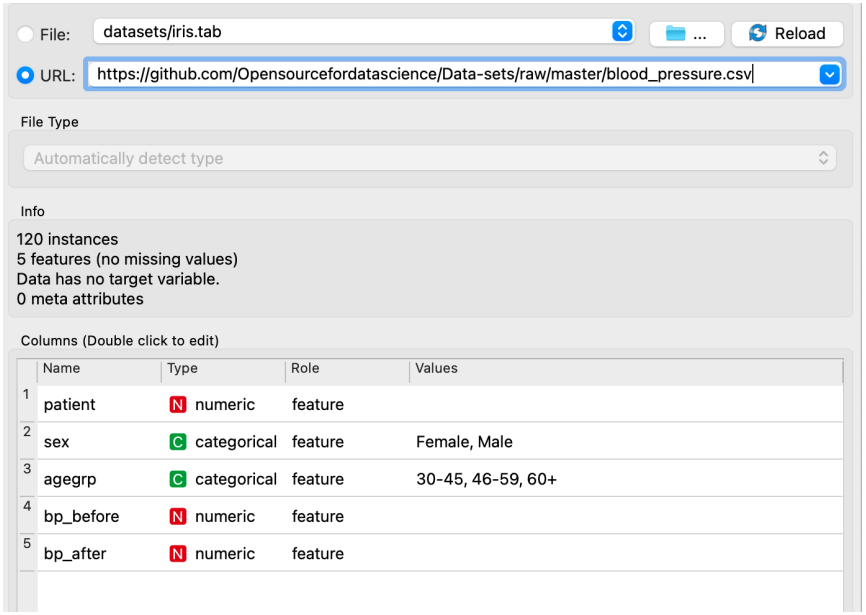


Data - File

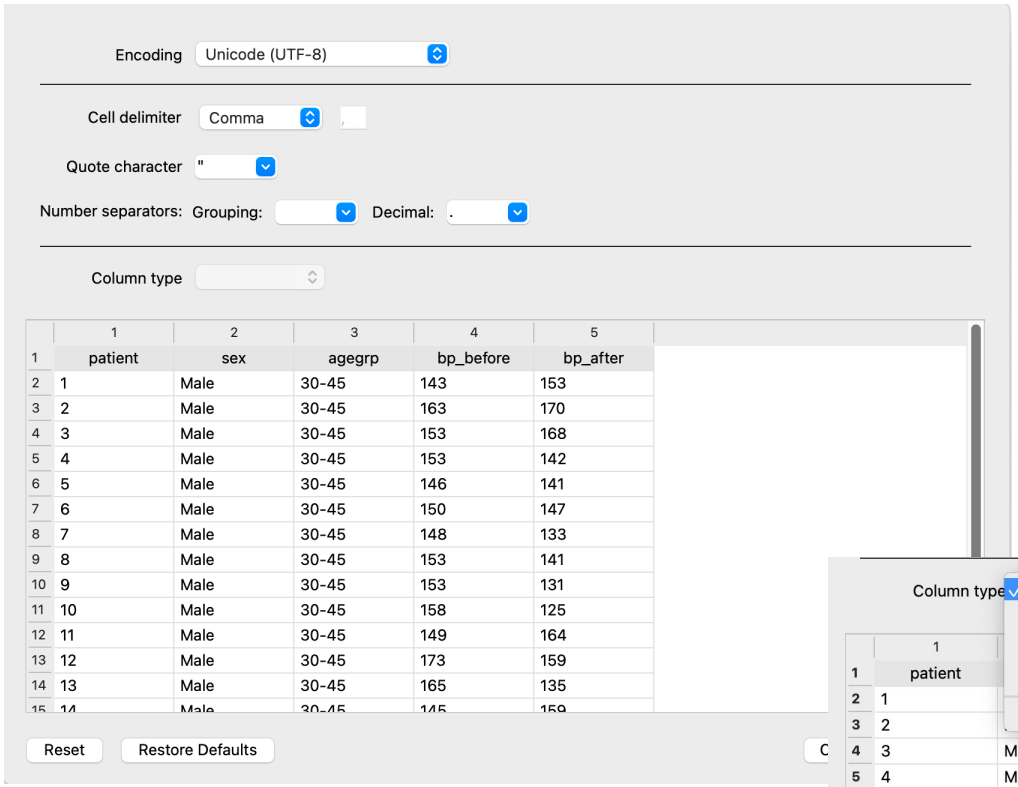

Data
File
CSV File Import
Datasets
SQL Table
Data Table
Paint Data
Data Info
Rank
Edit Domain
Color
Feature Statistics
Save Data


File


File
Source
File: datasets/iris.tab
...
Reload
URL:
File Type
Automatically detect type
Basket file (*.basket *.bsk)
Comma-separated values (*.csv *.csv.gz *.gz *.csv.bz2 *.bz2 *.csv.xz *.xz)
Microsoft Excel 97-2004 spreadsheet (*.xls)
Microsoft Excel spreadsheet (*.xlsx)
Pickled Orange data (*.pkl *.pickle *.pkl.gz *.pickle.gz *.gz *.pkl.bz2 *.pickle.bz2 *.bz2 *.pkl.xz *.pickle.xz *.xz)
Tab-separated values (*.tab *.tsv *.tab.gz *.tsv.gz *.gz *.tab.bz2 *.tsv.bz2 *.bz2 *.tab.xz *.tsv.xz *.xz)
Agilent Mosaic Image (*.dmt)
Agilent Mosaic Image (IFG) (*.dmt)
Agilent Mosaic Tile-by-tile (*.dmt)
Agilent Single Tile Image (*.dat)
Agilent Single Tile Image (IFG) (*.seq)
Envi (*.hdr)
Envi hdr or STXM hdr+xim files (*.hdr)
Galactic SPC format (*.spc *.SPC)
Gwyddion Simple Field (*.gsf)
HDF5 file @HERMRES/SOLEIL (*.hdf5)
HDF5 file @ROCK(hyperspectral imaging)/SOLEIL (*.h5)
Hyperspectral map ASCII (*.xyz)
Matlab (*.mat)
NXS HDF5 file @I08/Diamond Light Source (*.nxs)
NeaSPEC (*.nea *.txt)
NeaSPEC raw files (*.gsf)
OPUS Spectrum (*.0* *.1* *.2* *.3* *.4* *.5* *.6* *.7* *.8* *.9*)
Omnic map (*.map)
PTIR Studio file (*.ptir)
Renishaw WiRE WDF reader (*.wdf *.WDF)
SPA (*.spa *.SPA *.srs)
STXM/NEXAFS .hdr+.xim files (*.hdr)
Spectra ASCII (*.dat *.dpt *.xy *.csv)
Spectra ASCII or Agilent Single Tile Image (*.dat)
XAS ascii spectrum from ROCK (*.txt)

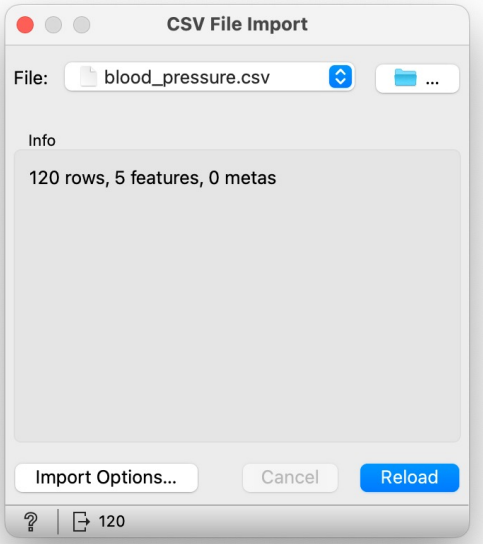

File: datasets/iris.tab
...
Reload
URL: https://github.com/Opensourcefordatascience/Data-sets/raw/master/blood_pressure.csv
File Type
Automatically detect type
Info
120 instances
5 features (no missing values)
Data has no target variable.
0 meta attributes
Columns (Double click to edit)
Name Type Role Values
1 patient N numeric feature
2 sex C categorical feature Female, Male
3 agegrp C categorical feature 30-45, 46-59, 60+
4 bp_before N numeric feature
5 bp_after N numeric feature

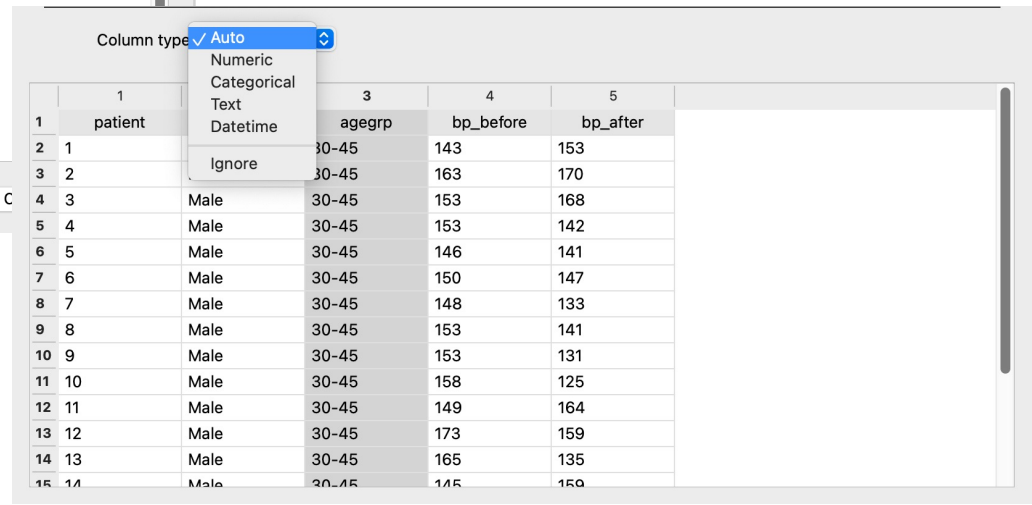
Data - CSV



	1	2	3	4	5
1	patient	sex	agegrp	bp_before	bp_after
2	1	Male	30-45	143	153
3	2	Male	30-45	163	170
4	3	Male	30-45	153	168
5	4	Male	30-45	153	142
6	5	Male	30-45	146	141
7	6	Male	30-45	150	147
8	7	Male	30-45	148	133
9	8	Male	30-45	153	141
10	9	Male	30-45	153	131
11	10	Male	30-45	158	125
12	11	Male	30-45	149	164
13	12	Male	30-45	173	159
14	13	Male	30-45	165	135
15	14	Male	30-45	145	159

CSV File Import





	1	2	3	4	5
1	patient		agegrp	bp_before	bp_after
2	1		30-45	143	153
3	2		30-45	163	170
4	3	Male	30-45	153	168
5	4	Male	30-45	153	142
6	5	Male	30-45	146	141
7	6	Male	30-45	150	147
8	7	Male	30-45	148	133
9	8	Male	30-45	153	141
10	9	Male	30-45	153	131
11	10	Male	30-45	158	125
12	11	Male	30-45	149	164
13	12	Male	30-45	173	159
14	13	Male	30-45	165	135
15	14	Male	30-45	145	159

Datasets

Title	Size	Instances	Variables	Target	Tags
BBC3	2.6 MB	1407	3	C categorical	text, classification, news
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	C categorical	biology
Smoking effect on B lymphocytes	1.8 MB	79	3000	C categorical	genomics
HDI	45.2 KB	188	53		economy, geo
ParlaMint	1.7 MB	1000	17	C categorical	text, classification, time, politics
SentiNews	5.0 MB	2000	7	C categorical	text, sentiment
TKI resistance	1.2 MB	280	467	C categorical	spectral
Abalone	187.5 KB	4177	8	N numeric	biology
Adult	4.1 MB	32561	15	C categorical	economy, fairness
Roman Amphorae	23.7 KB	164	16	C categorical	archaeology, image analytics
Attrition - Predict	838 bytes	3	18	C categorical	economy, synthetic, education
Attrition - Train	182.2 KB	1470	18	C categorical	economy, synthetic
Auto MPG	17.3 KB	398	9	N numeric	
Bank Marketing	466.1 KB	4119	20	C categorical	economy
Banking Crises	31.3 KB	211	73		time, economy

Description

Breast Cancer and Docetaxel Treatment (2006), from NCBI

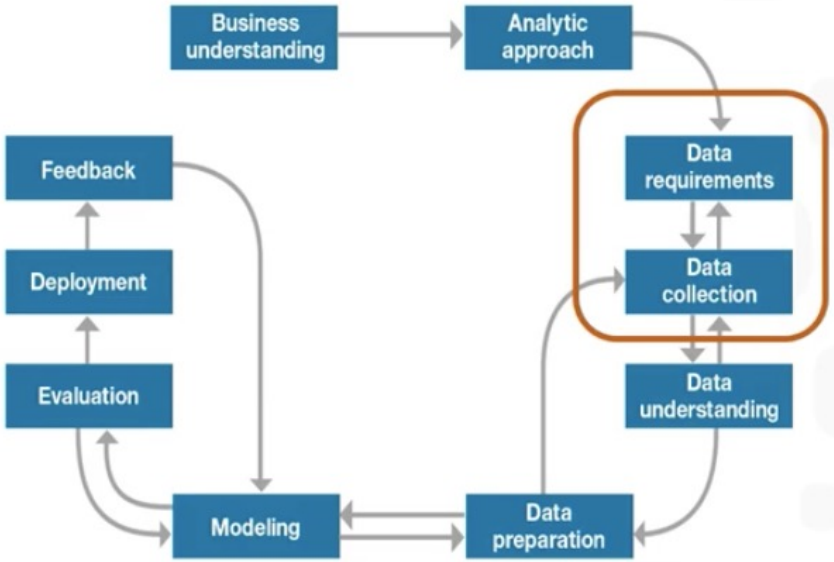
Breast cancer core biopsies taken from patients found to be resistant (greater than 25% residual tumor volume) or sensitive (less than 25% residual tumor volume) to docetaxel treatment.

References

Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG et al. (2005) Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. J Clin Oncol, 23(6): 1169-77.

Methodology

From Requirements to Collection



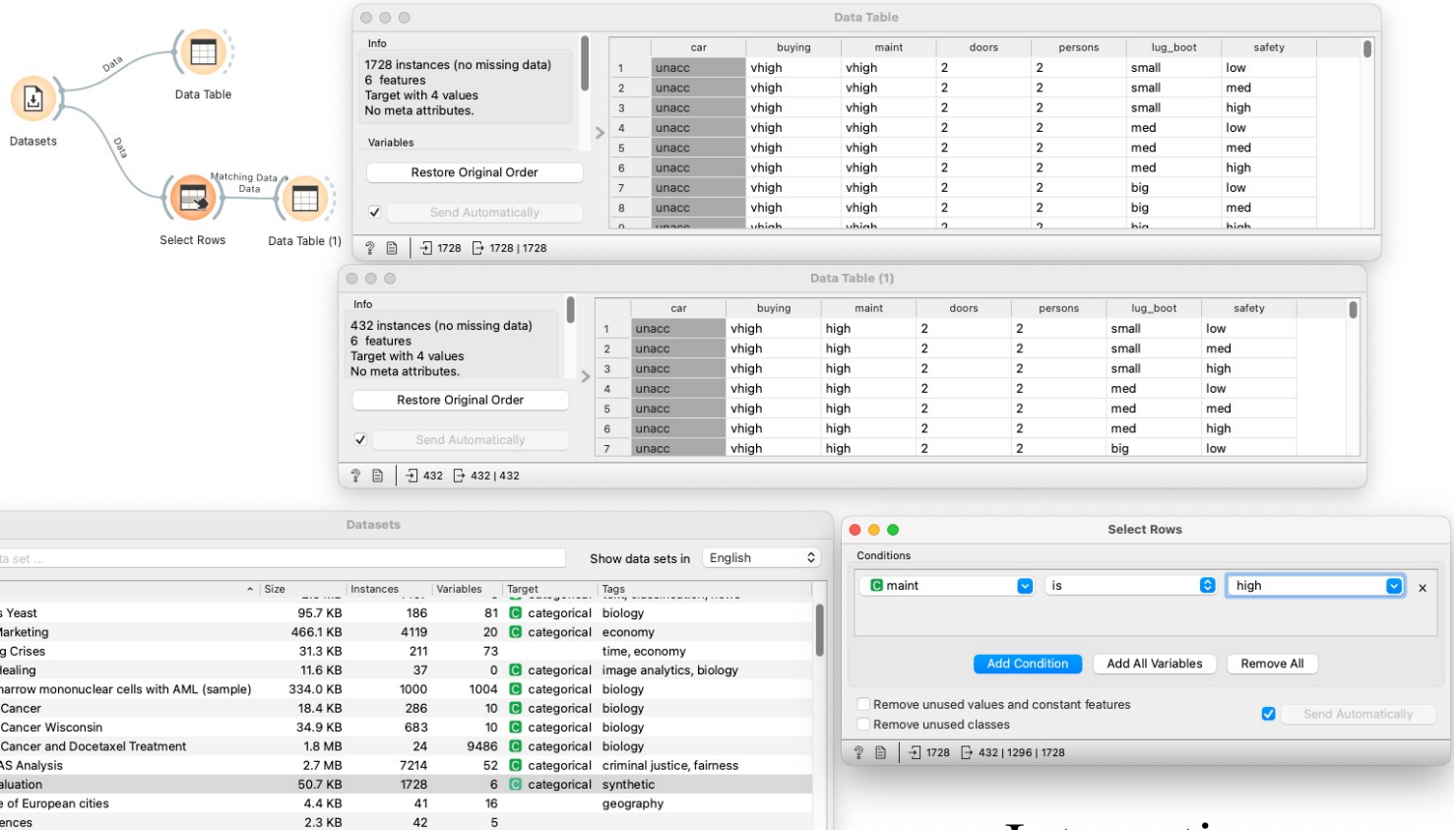
Data Requirements

- *What are data requirements?*

Data Collection

- *What occurs during data collection?*

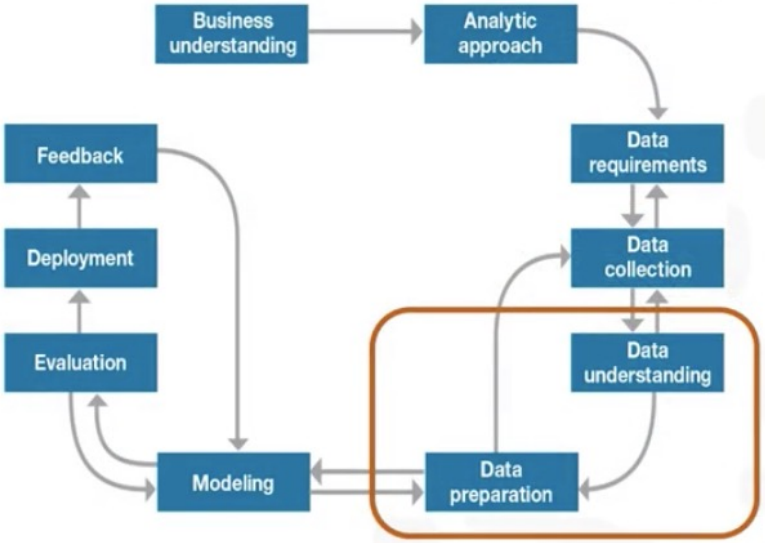
Data Engineering (Cont.)



- Interactive
- Visual
- Simplified

Data Understanding

From Understanding to Preparation



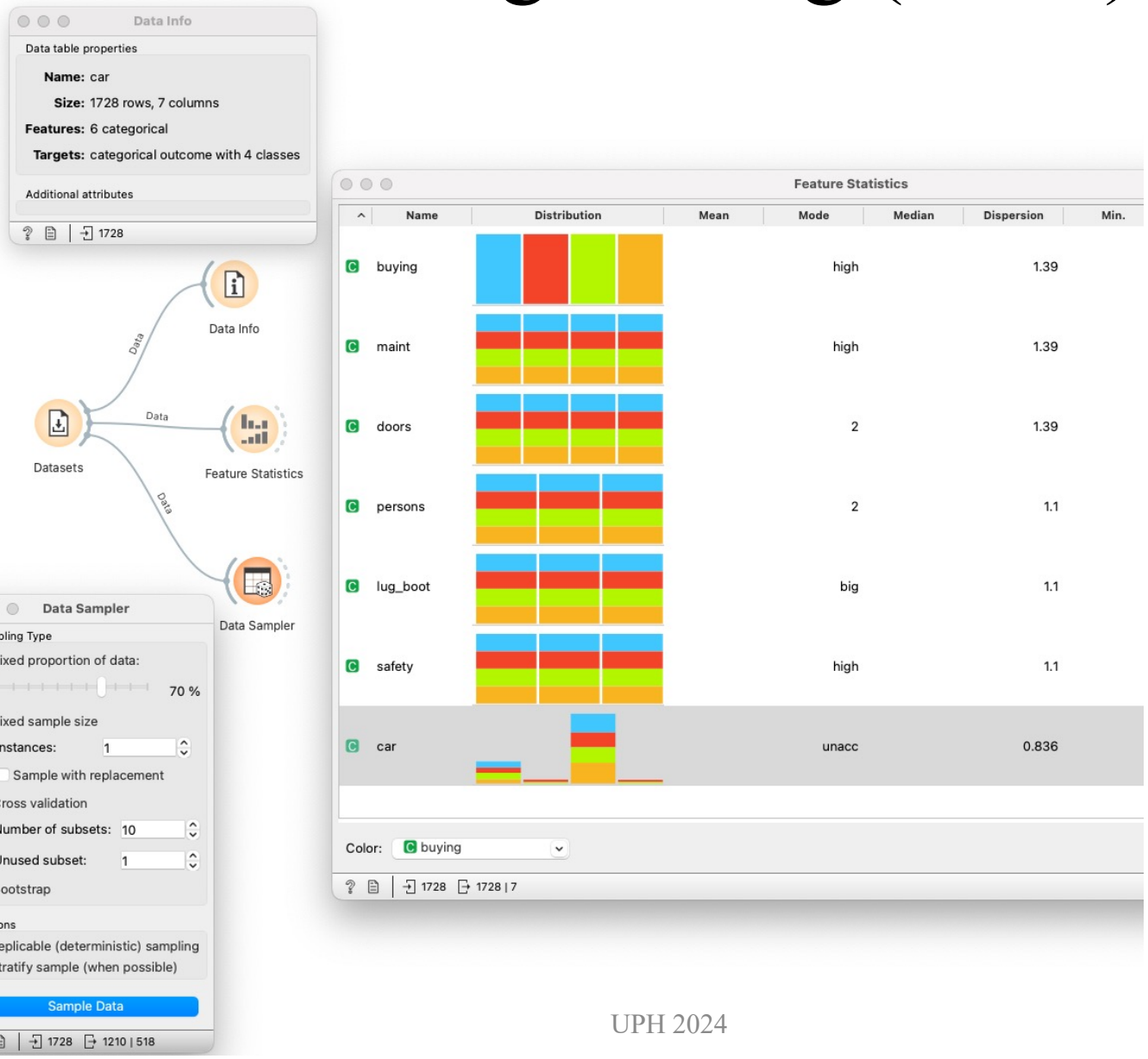
Data understanding

- *What does it mean to "prepare" or "clean" data?*

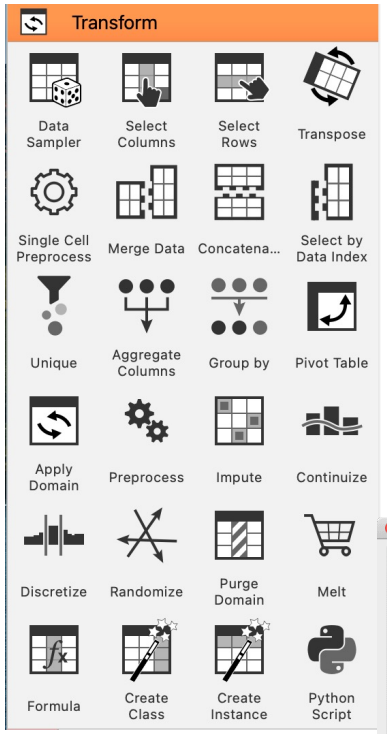
Data preparation

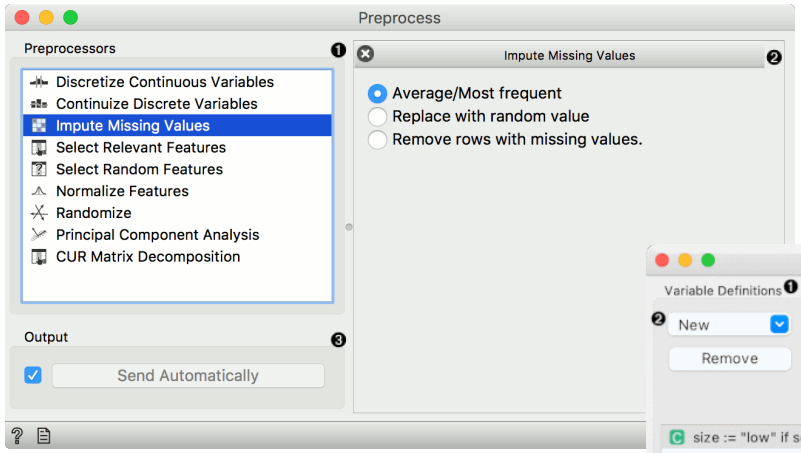
- *What are ways in which data is prepared?*

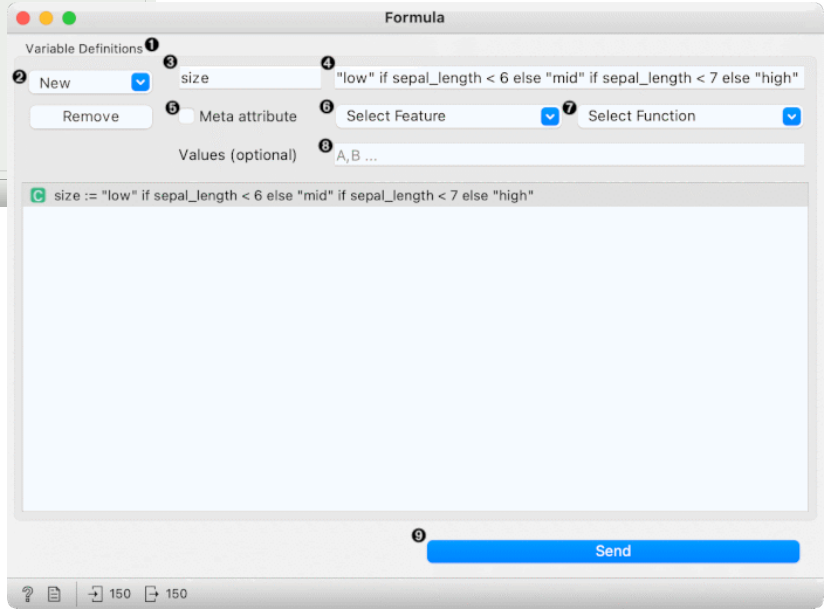
Data Engineering (Cont.)

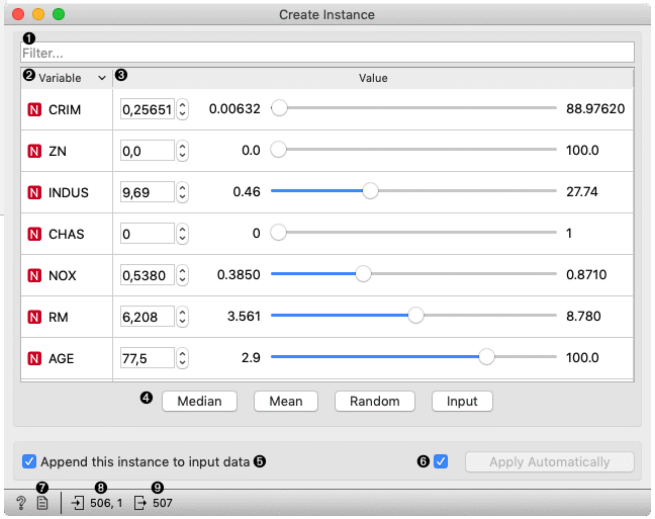


Data Engineering (Cont.)



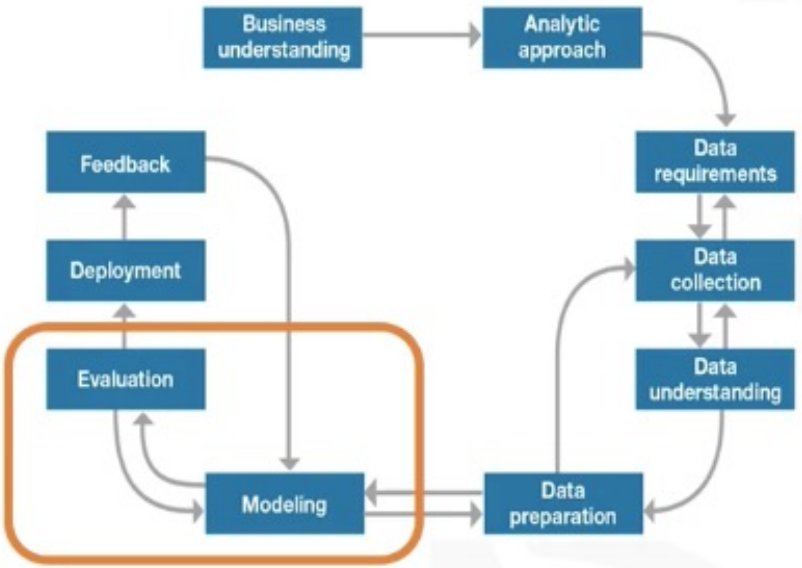






Modeling

From Modeling to Evaluation



Modeling

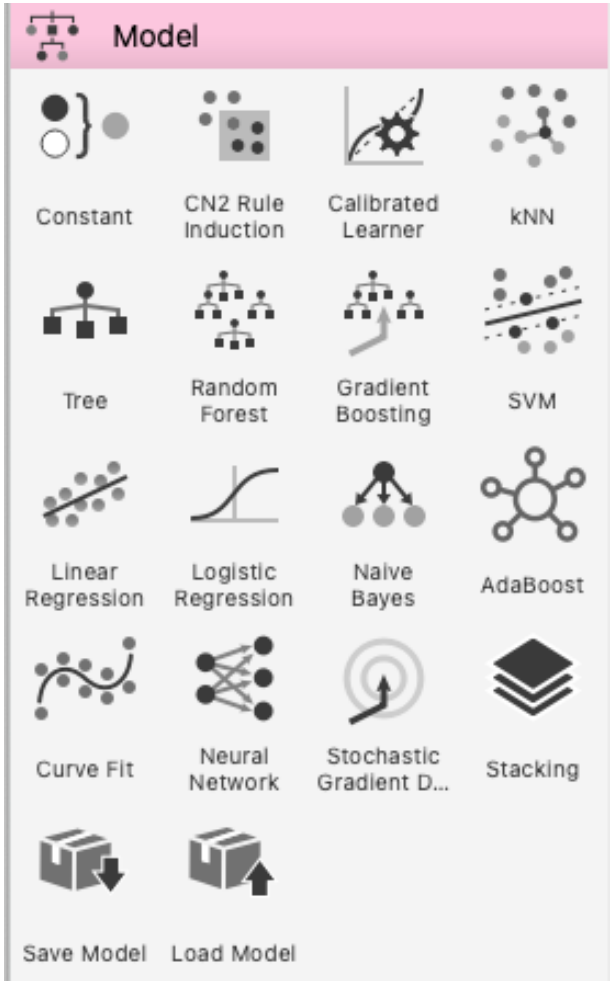
- *In what way can the data be visualized to get to the answer that is required?*

Evaluation

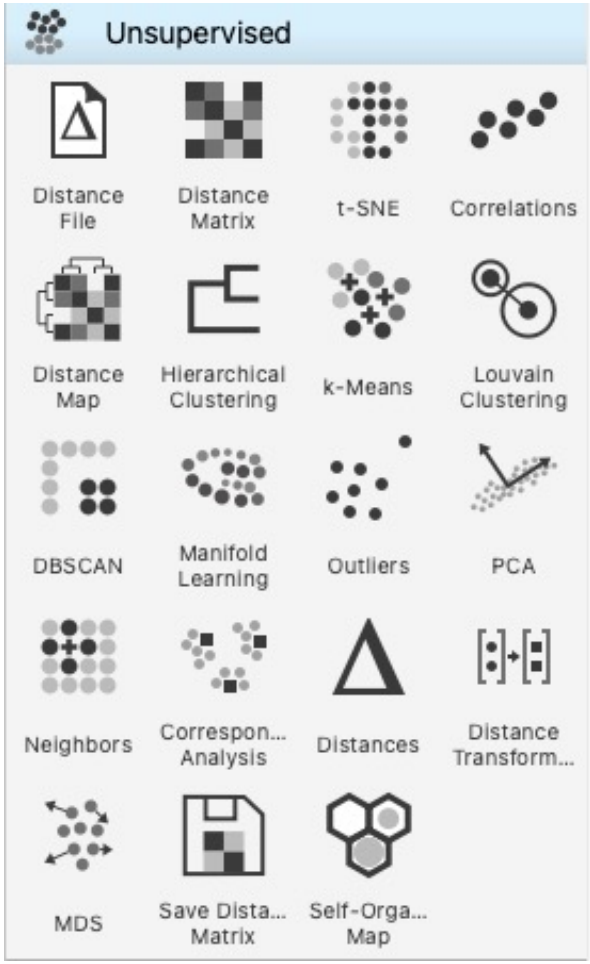
- *Does the model used really answer the initial question or does it need to be adjusted?*

Modeling in Orange3

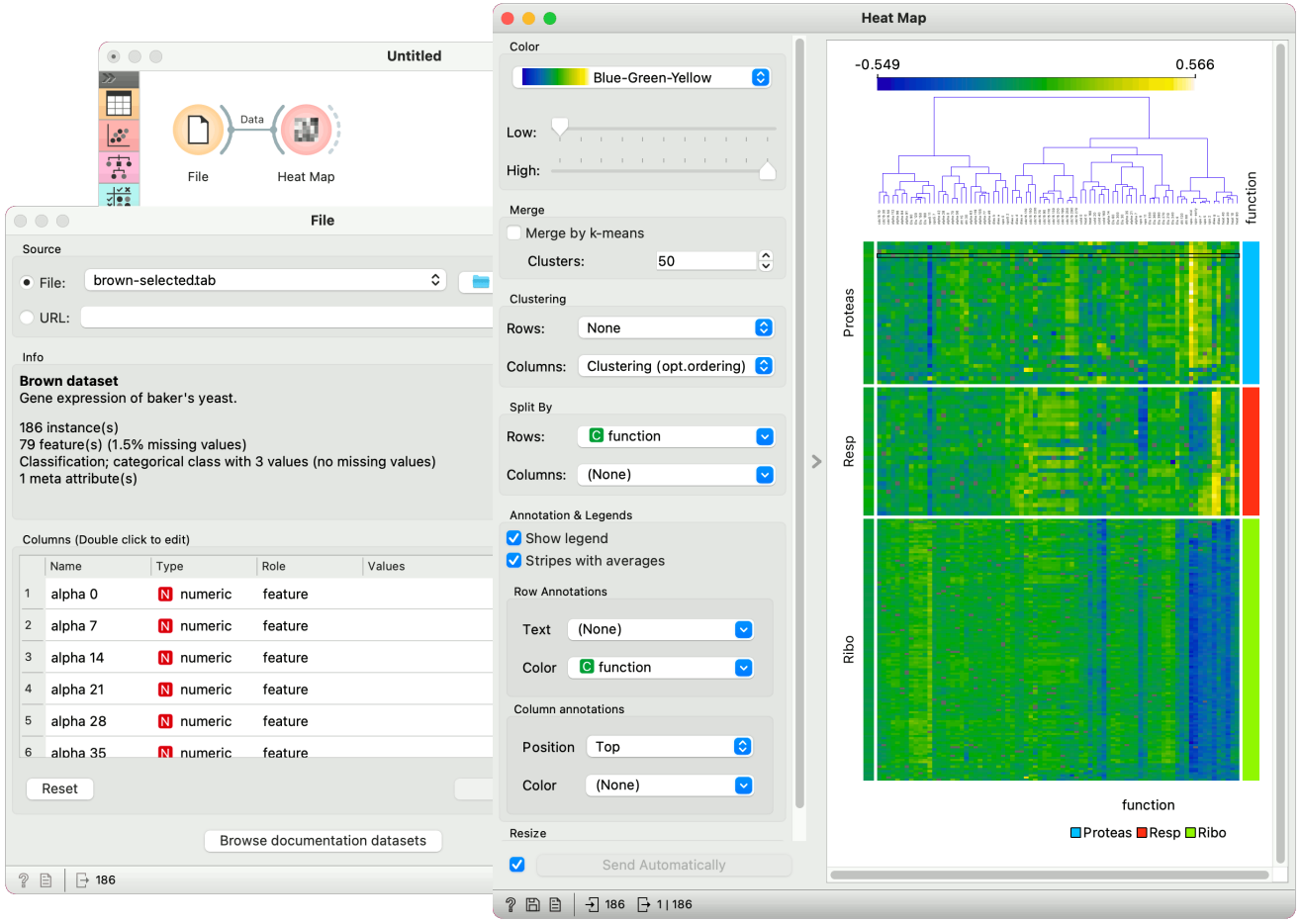
Supervised



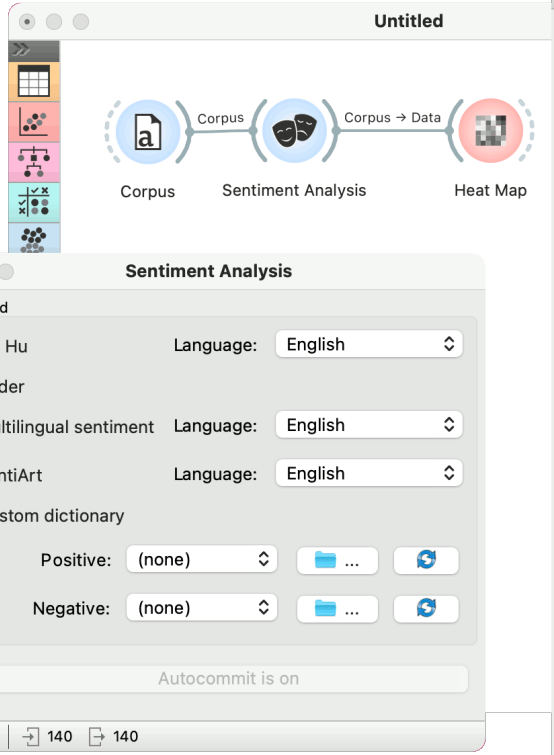
UnSupervised

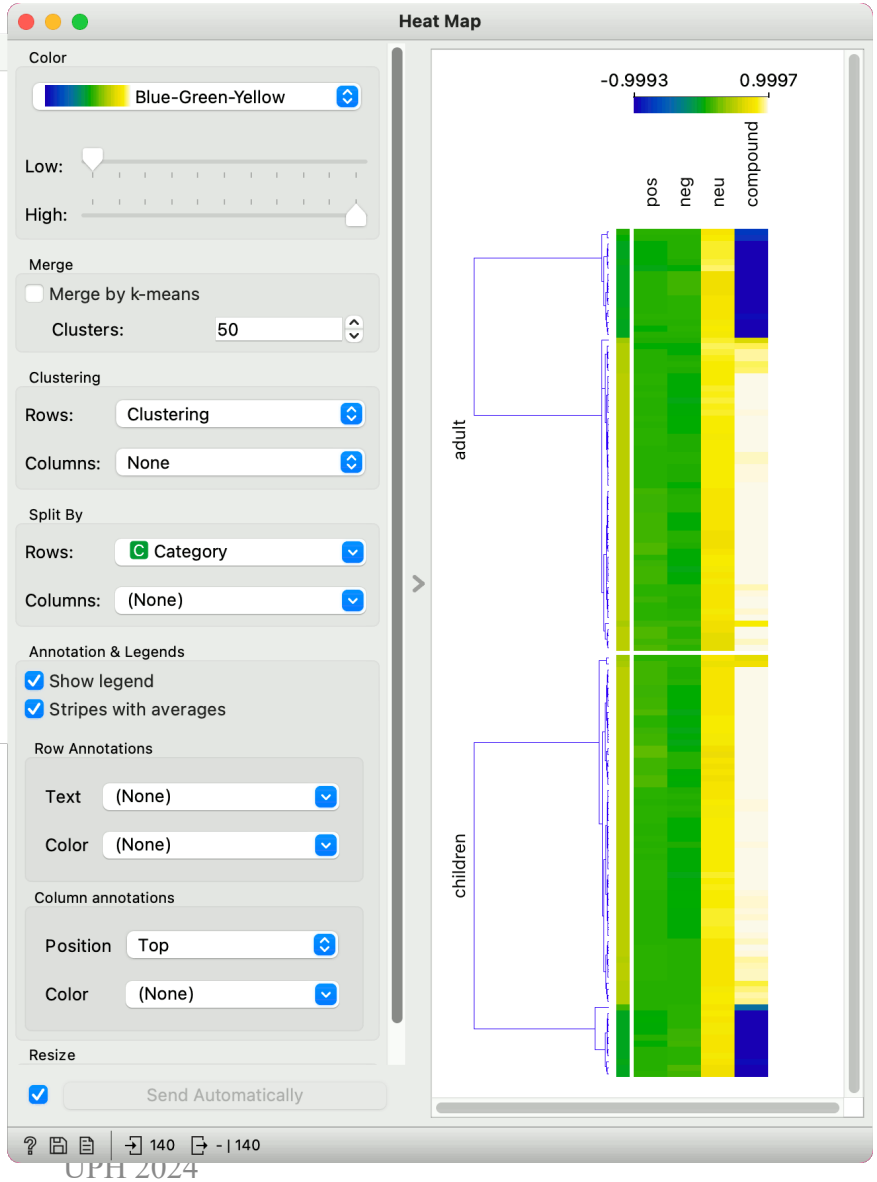


Visualization

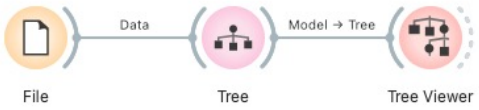


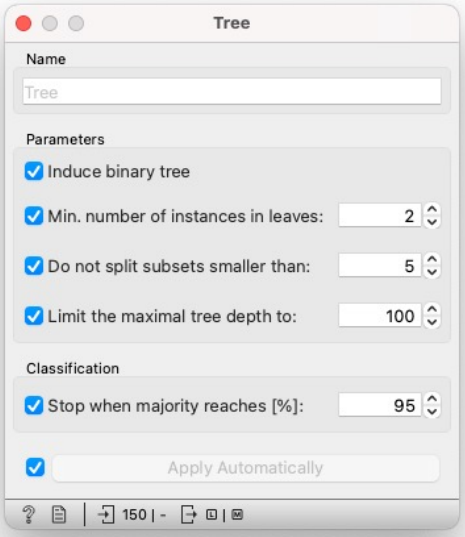
Visualization

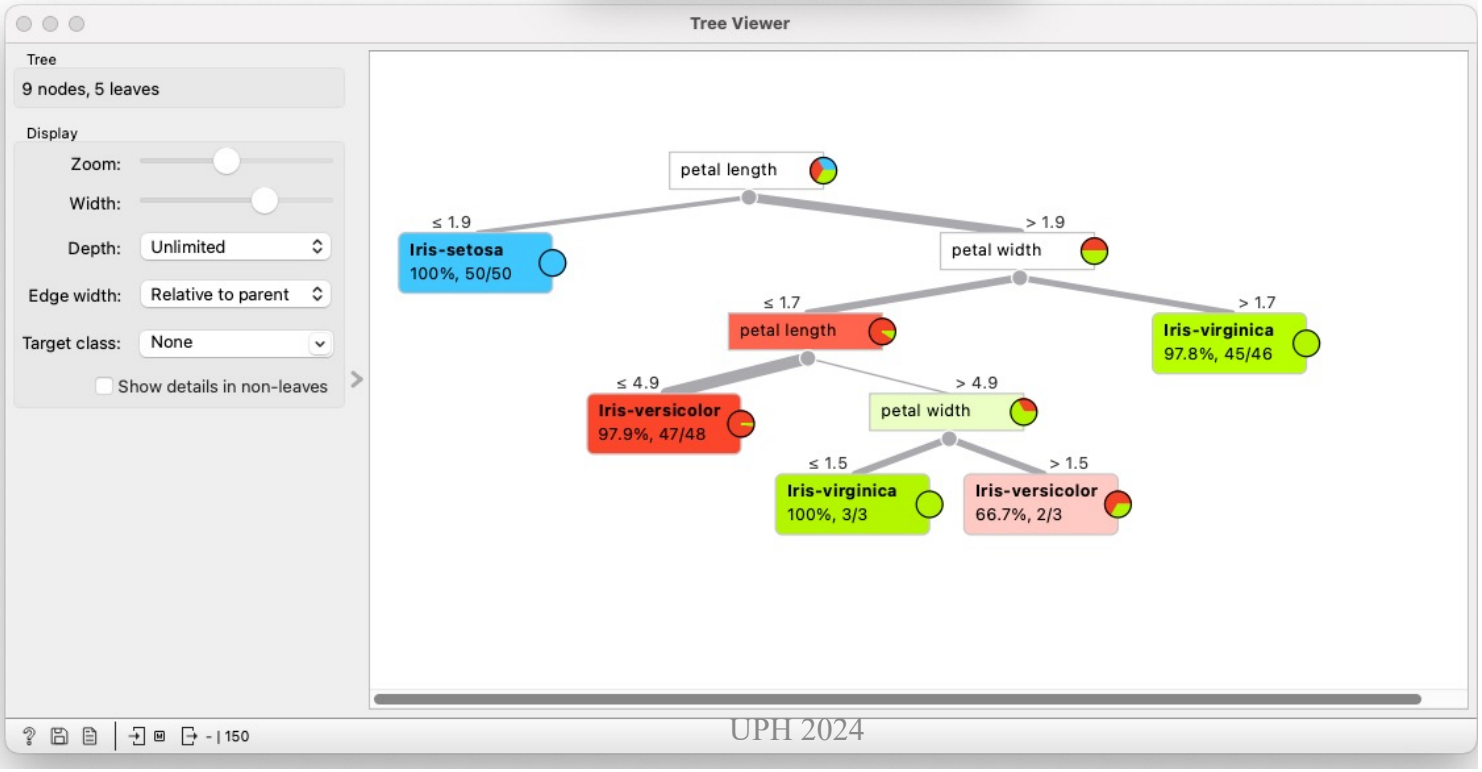




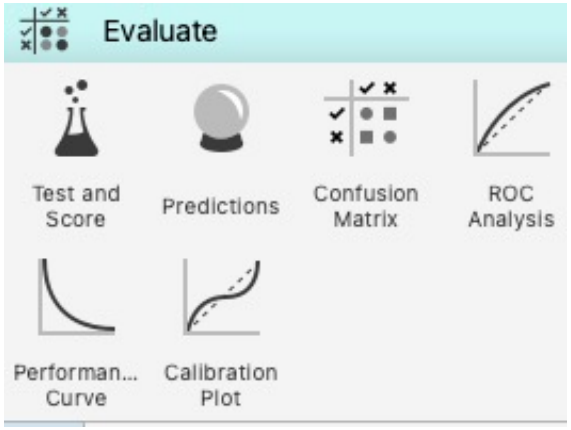
Tree

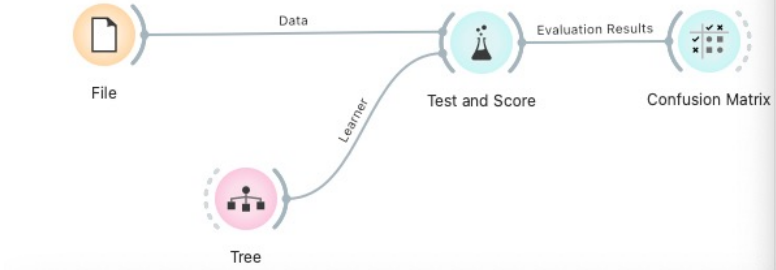

Data
Model → Tree
File
Tree
Tree Viewer

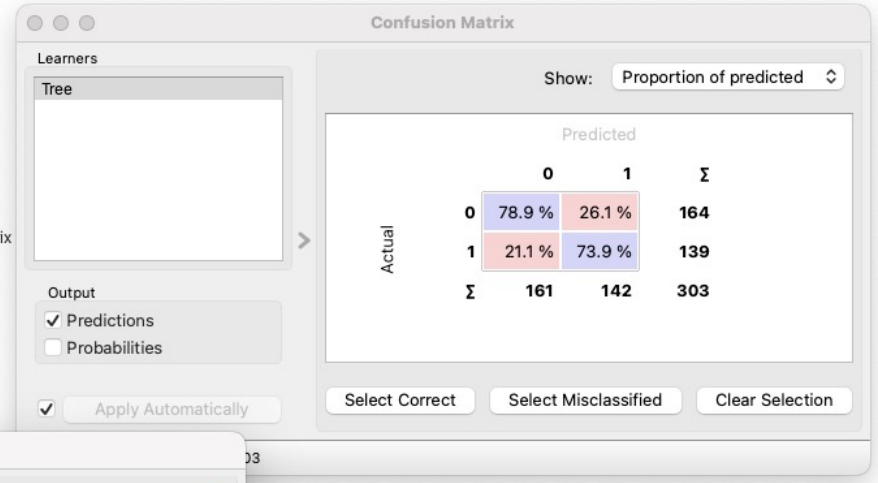

Tree
Name
Tree
Parameters
Induce binary tree
Min. number of instances in leaves: 2
Do not split subsets smaller than: 5
Limit the maximal tree depth to: 100
Classification
Stop when majority reaches [%]: 95
Apply Automatically
150 | -

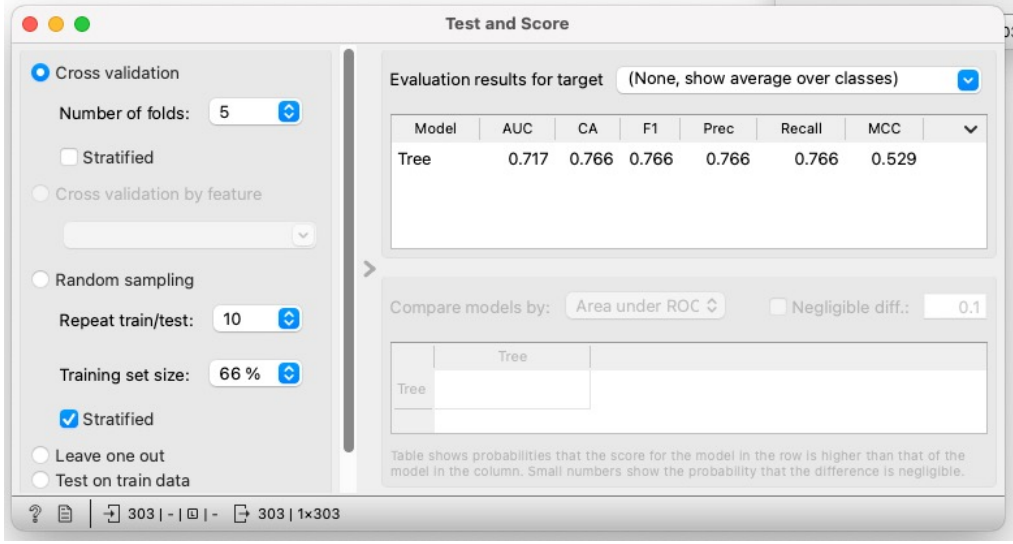

Tree Viewer
Tree
9 nodes, 5 leaves
Display
Zoom:
Width:
Depth: Unlimited
Edge width: Relative to parent
Target class: None
Show details in non-leaves
petal length
≤ 1.9
Iris-setosa
100%, 50/50
> 1.9
petal width
≤ 1.7
petal length
> 1.7
Iris-virginica
97.8%, 45/46
≤ 4.9
Iris-versicolor
97.9%, 47/48
> 4.9
petal width
≤ 1.5
Iris-virginica
100%, 3/3
> 1.5
Iris-versicolor
66.7%, 2/3
- | 150

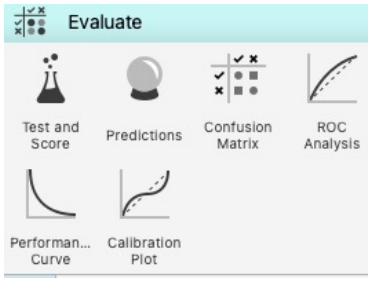
Evaluate


Evaluate
Test and Score
Predictions
Confusion Matrix
ROC Analysis
Performan... Curve
Calibration Plot

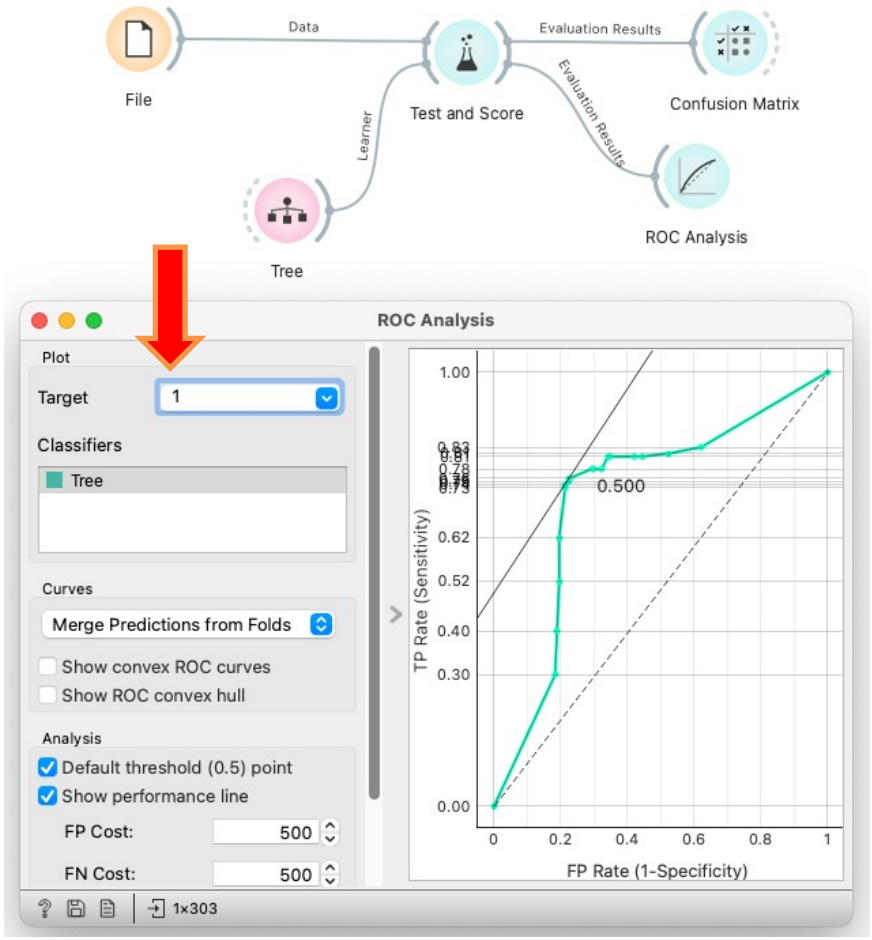

Data
Evaluation Results
File
Test and Score
Confusion Matrix
Learner
Tree

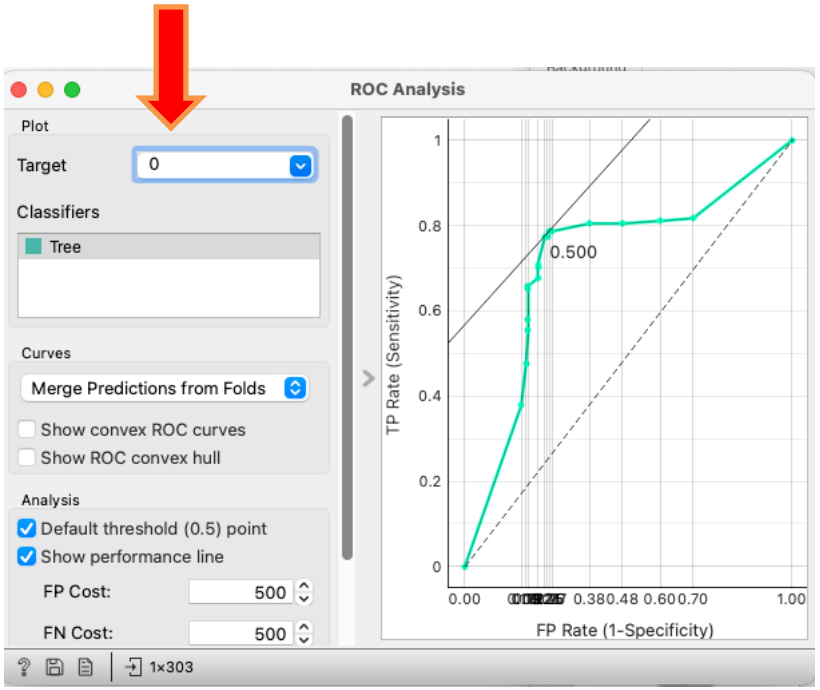

Confusion Matrix
Learners
Tree
Show:
Proportion of predicted
Predicted
0
1
Σ
Actual
0
78.9 %
26.1 %
164
1
21.1 %
73.9 %
139
Σ
161
142
303
Output
Predictions
Probabilities
Apply Automatically
Select Correct
Select Misclassified
Clear Selection


Test and Score
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
Evaluation results for target
(None, show average over classes)
Model
AUC
CA
F1
Prec
Recall
MCC
Tree
0.717
0.766
0.766
0.766
0.766
0.529
Compare models by:
Area under ROC
Negligible diff.:
0.1
Tree
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.
303 | - | - 303 | 1×303



Evaluate

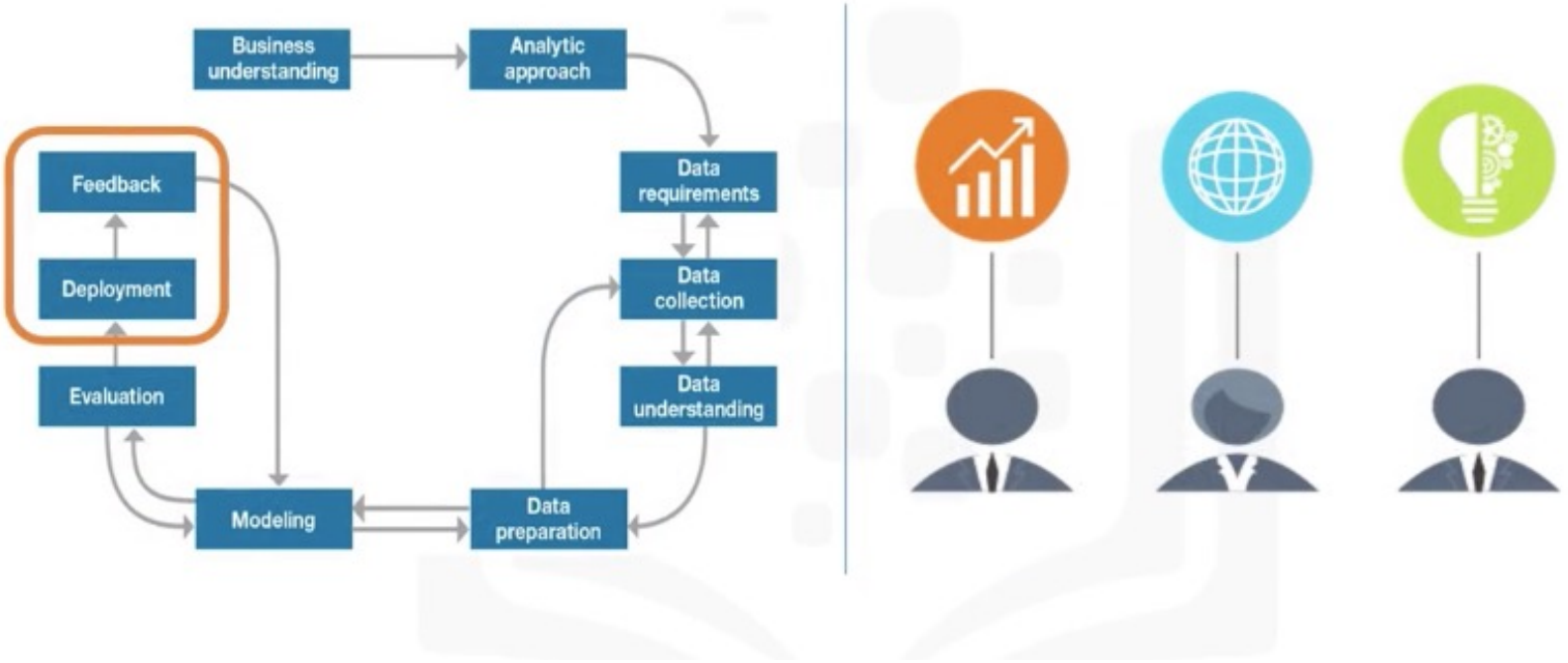




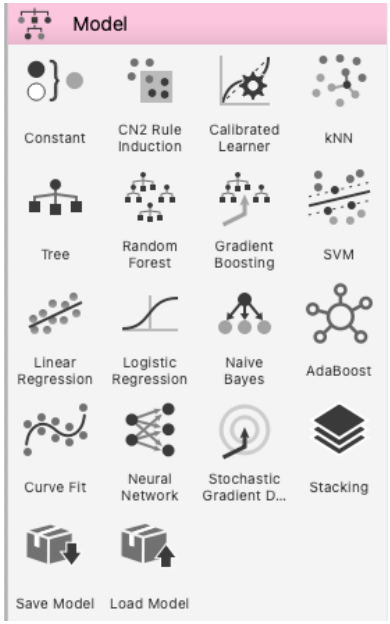
Interactive

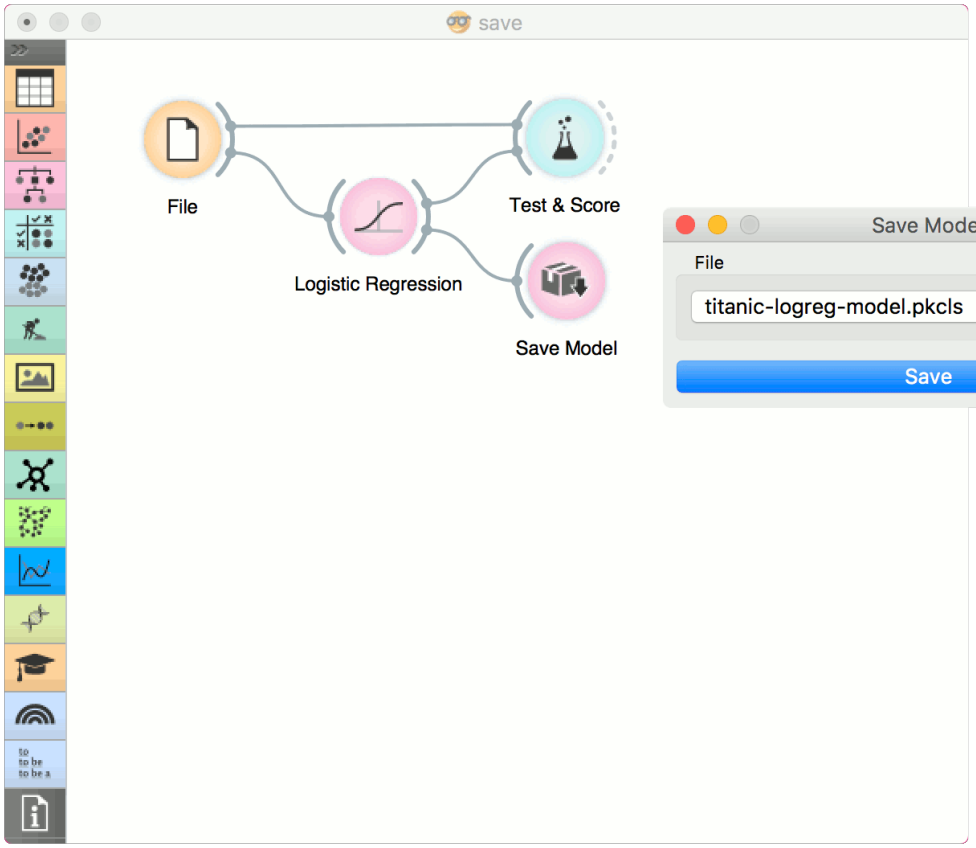
Deployment

From Deployment to Feedback



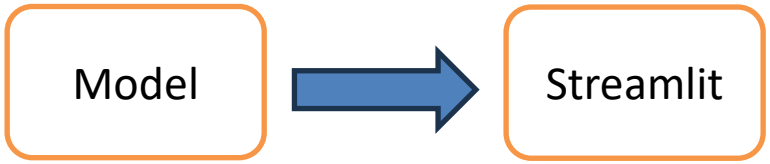
Deployment

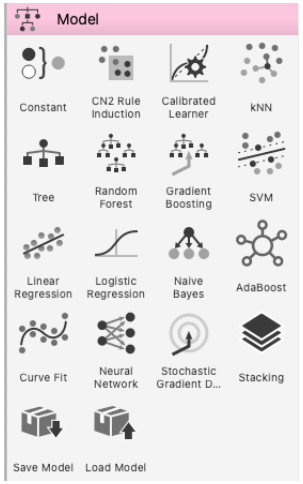

Model
Constant
CN2 Rule Induction
Calibrated Learner
kNN
Tree
Random Forest
Gradient Boosting
SVM
Linear Regression
Logistic Regression
Naive Bayes
AdaBoost
Curve Fit
Neural Network
Stochastic Gradient D...
Stacking
Save Model
Load Model

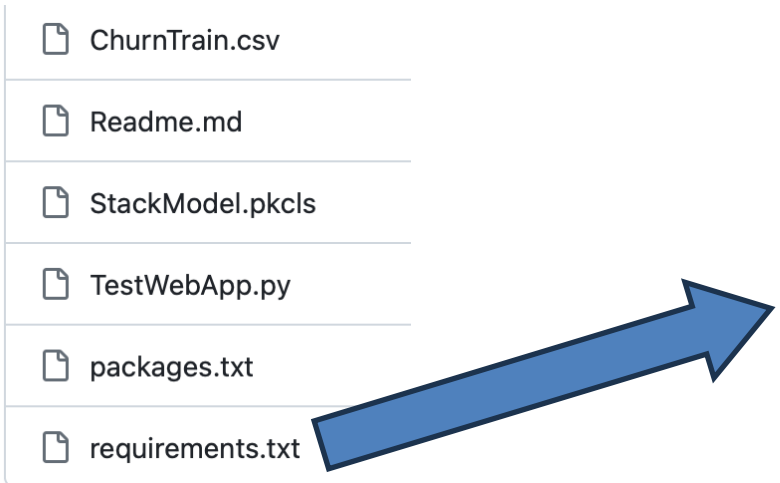

save
File
Test & Score
Logistic Regression
Save Model

Save Model
File
titanic-logreg-model.pkcls
...
Save

Deployment








```
1  streamlit==1.1.0
2  pandas==1.3.4
3  numpy==1.20.3
4  Orange3==3.30.2
5  xgboost==1.5.0
```

Example:
https://github.com/TonciG/Telecom-Customer-Churn---Web-App

Any Question ?