



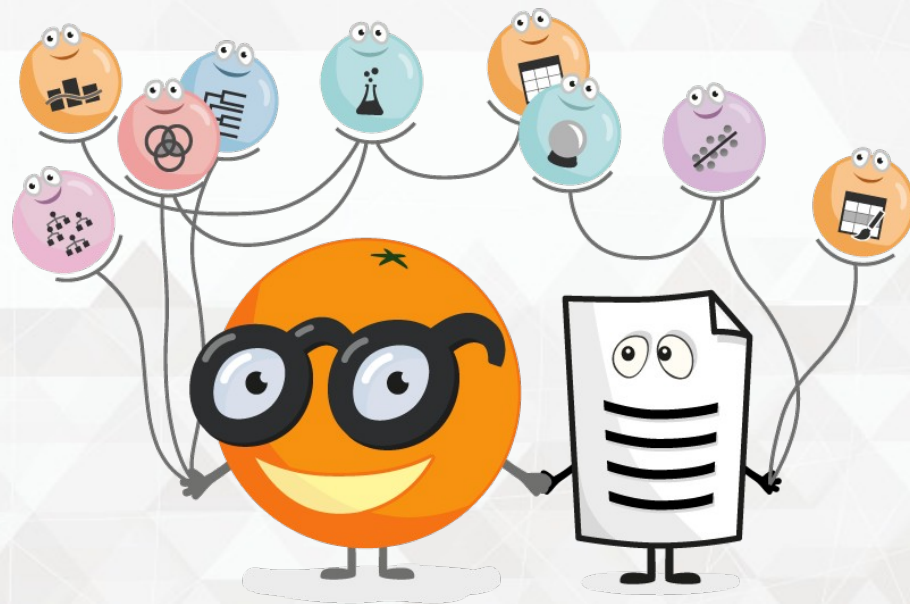
Orange

Data Analysis Workflow

2023

Conteúdo

- Orange data workflow
- Utilização do workflow
- Carregamento de dados
- Ferramentas de tratamento de dados



Orange

- Desenvolvido pela Universidade de Ljubljani
 - Eslovênia
 - <https://www.uni-lj.si/university/>
- Ferramenta de visualização e manipulação de dados
 - Data mining
 - Machine learning
 - Programação visual
 - Lema: “anyone can do data science!”
- Desenvolvida em Python, C, C++ e Cython
 - Utilização de bibliotecas convencionais
 - APIs próprias de scripting



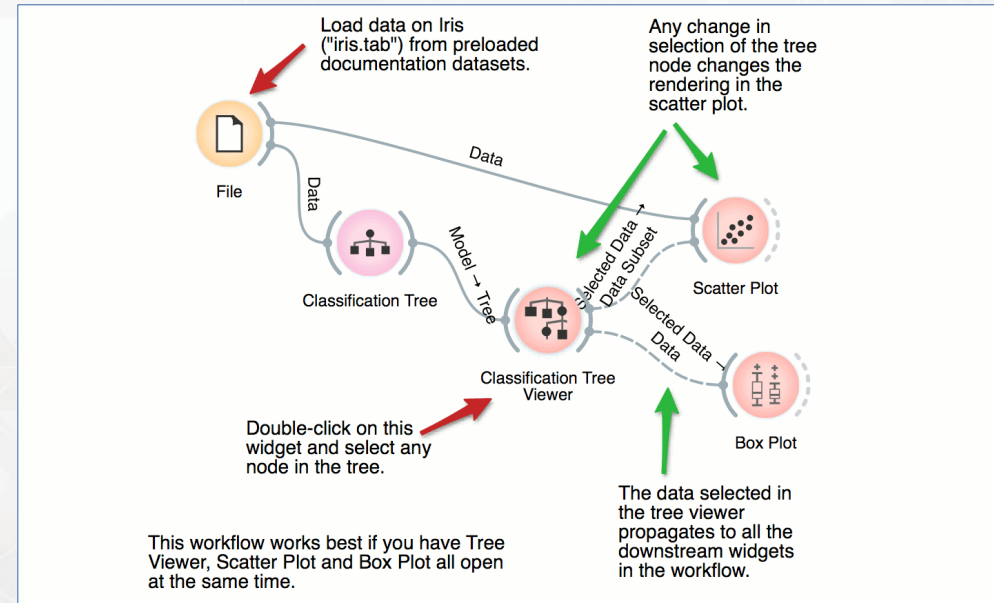
Orange

- Site principal
 - <https://orangedatamining.com/>
 - <https://orange.biolab.si/>
- Download e instalação
 - <https://orangedatamining.com/download/>
- Documentação
 - <https://orangedatamining.com/docs/>
 - Tutoriais e vídeos disponíveis

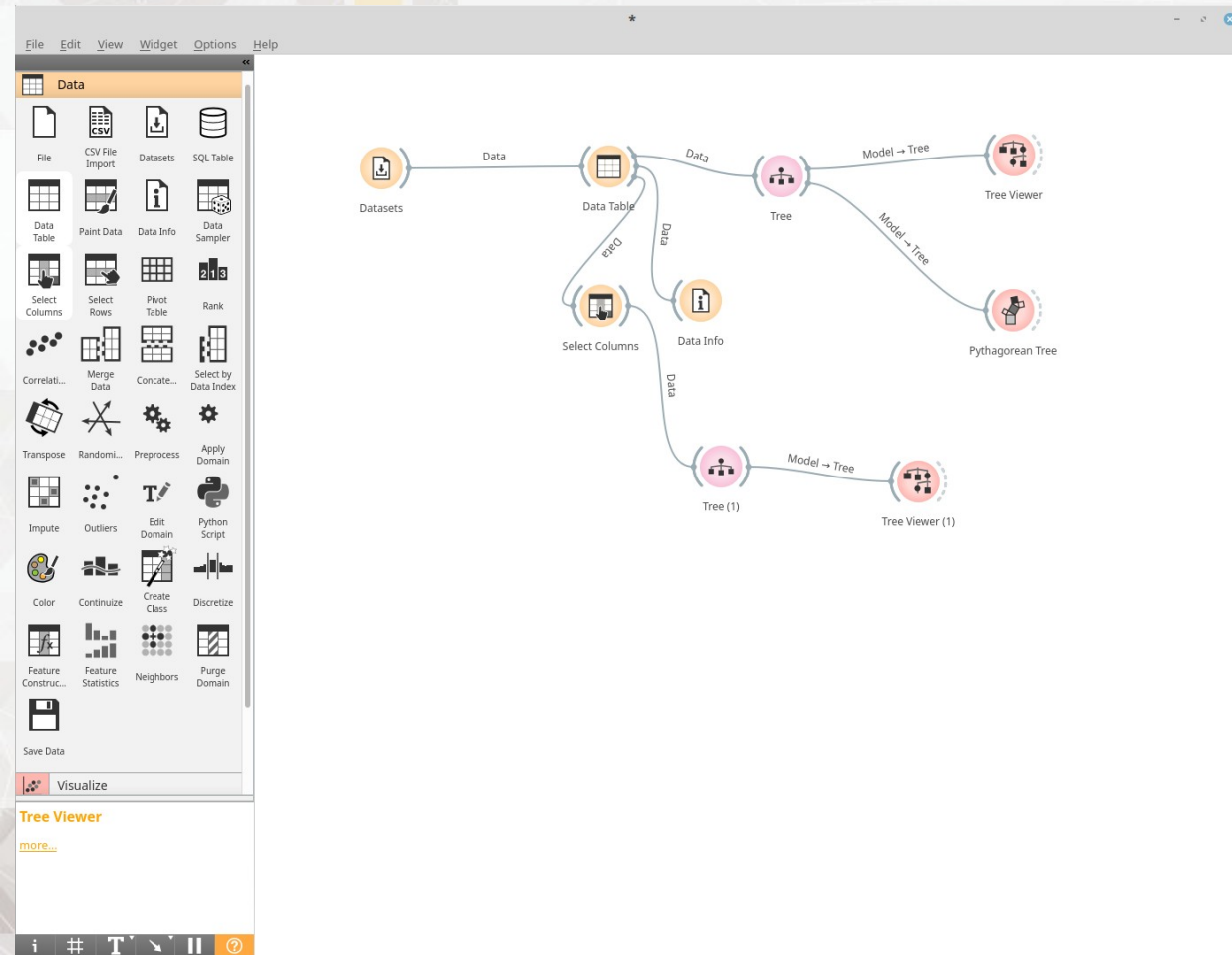


Orange - Workflows

- Organização do tratamento dos dados em fluxos
 - Sem necessidade de programação
 - Embora permita scripts
 - Tarefas definidas e discretas
 - Armazenamento temporário em memória
 - Estabelecimento de sequência de passos para atingir objetivo
- Adequado para ensaios, testes
 - E até mesmo projetos reais de porte que seja suportado
- Suporte a add-ons



Interface



- Painel de ferramentas
- Canvas de desenvolvimento
 - Drag-and-drop

Interface

Data

File

CSV File Import

Datasets

SQL Table

Data Table

Paint Data

Data Info

Data Sampler

Select Columns

Select Rows

Pivot Table

Rank

Correlation

Merge Data

Concatenate

Select by Data Index

Transpose

Randomize

Preprocess

Apply Domain

Impute

Outliers

Edit Domain

Python Script

Color

Continuize

Create Class

Discretize

Feature Construction

Feature Statistics

Neighbors

Purge Domain

Save Data

Visualize

Tree Viewer

Box Plot

Distribution

Scatter Plot

Line Plot

Sieve Diagram

Mosaic Display

FreeViz

Linear Projection

Radviz

Heat Map

Venn Diagram

Silhouette Plot

Pythagorean Tree

Pythagorean Forest

CN2 Rule Viewer

Nomogram

Model

Constant

CN2 Rule Induction

Calibrated Learner

kNN

Tree

Random Forest

SVM

Linear Regression

Logistic Regression

Naive Bayes

AdaBoost

Neural Network

Stochastic Gradient Descent

Stacking

Save Model

Load Model

Evaluate

Test and Score

Predictions

Confusion Matrix

ROC Analysis

Lift Curve

Calibration Plot

Unsupervised

Distance File

Distance Matrix

t-SNE

Distance Map

Hierarchical Clustering

k-Means

Louvain Clustering

DBSCAN

Manifold Learning

PCA

Correspondence Analysis

Distances

Distance Transformation

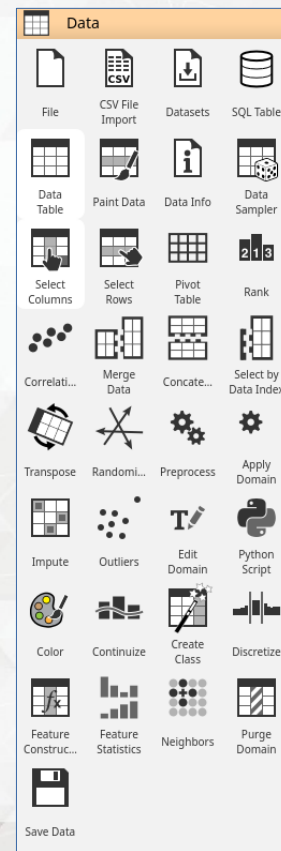
MDS

Save Distance Matrix

Self-Organizing Map

Carregamento e tratamento de dados

- Widgets para funções diversas
 - Ligação output-input
- Carregamento de dados
 - De arquivos, repositórios, rede, bancos relacionais
 - Somente PostgreSQL e SQL Server, por hora
- Manipulação de dados
 - Projeções, filtros, joins, particionamentos, transposições
 - Analisar custo de se fazer isso em uma ferramenta gráfica ou na preparação anterior dos dados
 - Amostragem de dados (sampling)
 - Útil para avaliação de modelos
 - Discretização, tratamento de classes e domínios
 - Informações sobre datasets



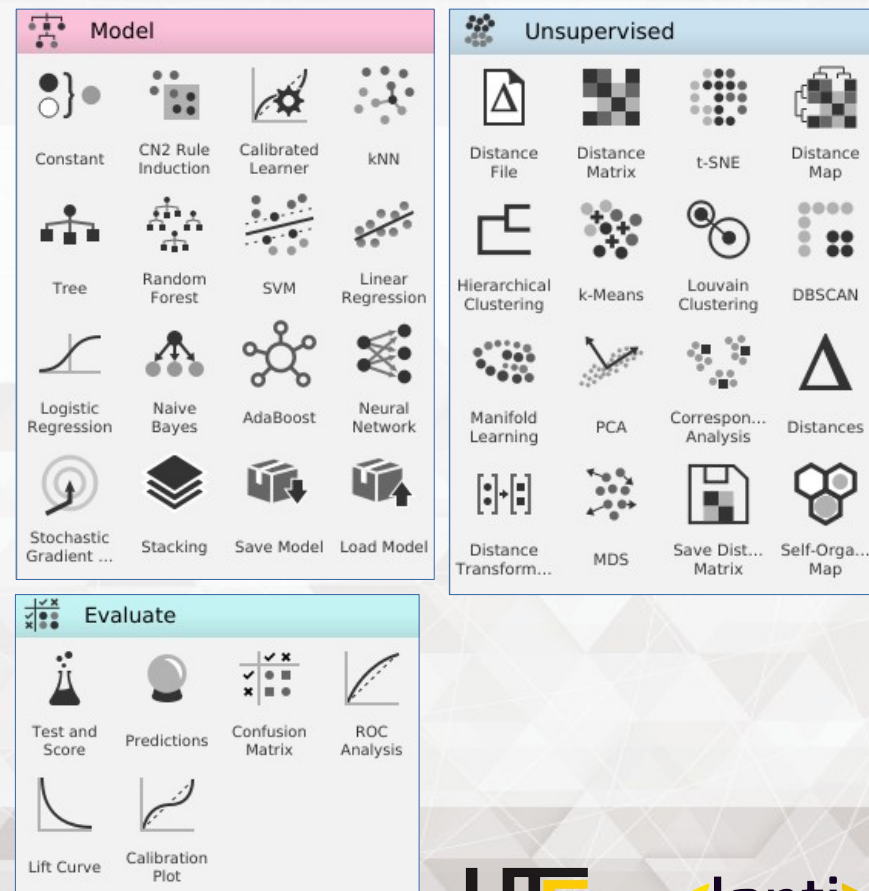
Visualização de dados

- Visualização em geral de dados
 - ScatterPlots, diagramas, distribuições,
- Visualização de resultados de modelos e análises
 - TreeViewer, PythagoraTree, Projections
- Análise de visualizações
 - FreeViz
 - Análise de correlações de maneira visual



Modelos de ML e Data Mining

- Modelos de Machine Learning
 - Supervisionados e não-supervisionados
- Modelos de Data Mining
- Tratamento e ciclo de vida de modelos
 - Modelos podem ser salvos em separado
- Avaliação de modelos



Carregamento de dados

- Widget File
 - CSVs, xlsx, txt ou URLs
- Orange importa qualquer arquivo com separações
 - Comma, tab
 - Planilhas e documentos Google Sheets
 - Nome de campos em cabeçalho
- Tipos de campos
 - Class (target), feature, meta
 - No carregamento pode ser definido como ignore e weight
 - Continuous (numeric), discrete (categorical), time, string (text)
- Tipos e papéis são definidos no widget File
 - Deduzidos por default
 - Redefinidos a cada campo



File

File: iris.tab ... Reload

URL: ...

Info

Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attribute(s)

Columns (Double click to edit)

| | Name | Type | Role | Values |
|---|--------------|-------------|---------|-----------------------------|
| 1 | sepal length | numeric | feature | |
| 2 | sepal width | numeric | feature | |
| 3 | petal length | numeric | feature | |
| 4 | petal width | numeric | feature | |
| 5 | iris | categorical | target | Iris-setosa, Iris-versic... |

Browse documentation datasets Reset Apply

? 📄

Carregamento de dados



- CSV File Import
 - Customização de leitura de CSVs
 - Delimitador, strings, números, etc
 - Encoding
 - DataFrame do Pandas

Import Options

Encoding:

Cell delimiter:

Quote character:

Number separators: Grouping: Decimal:

Column type:

| | 1 | 2 | 3 | 4 |
|----|--------|-------|------|----------|
| 1 | status | age | sex | survived |
| 2 | d | d | d | d |
| 3 | | | | class |
| 4 | first | adult | male | yes |
| 5 | first | adult | male | yes |
| 6 | first | adult | male | yes |
| 7 | first | adult | male | yes |
| 8 | first | adult | male | yes |
| 9 | first | adult | male | yes |
| 10 | first | adult | male | yes |
| 11 | first | adult | male | yes |
| 12 | first | adult | male | yes |
| 13 | first | adult | male | yes |

Reset Restore Defaults Cancel OK

Carregamento de dados

- Datasets
 - Datasets de repositórios online
 - Buscados e atualizados dinamicamente



Datasets

Info
66 datasets
2 datasets cached

Filter

| ^ | Title | Size | Instances | Variables | Target | Tags |
|---|--|-----------|-----------|-----------|--------|----------------------------|
| • | Iris | 4.5 KB | 150 | 5 | C | categoryal biology |
| • | Zoo | 7.0 KB | 101 | 17 | C | categoryal biology |
| | Breast Cancer and Docetaxel Treatment | 1.8 MB | 24 | 9486 | C | categoryal biology |
| | Smoking effect on B lymphocytes | 1.8 MB | 79 | 3000 | C | categoryal genomics |
| | Bone marrow mononuclear cells with AML | 582.0 KB | 96 | 1000 | C | categoryal genomics |
| | HDI | 65.1 KB | 188 | 66 | N | numeric economy, geo |
| | Abalone | 187.5 KB | 4177 | 8 | N | numeric biology |
| | Adult | 4.1 MB | 32561 | 15 | C | categoryal economy |
| | Roman Amphorae | 23.7 KB | 164 | 16 | C | categoryal archaeology, i. |
| | Attrition - Predict | 838 bytes | 3 | 18 | C | categoryal economy, synt |
| | Attrition - Train | 182.2 KB | 1470 | 18 | C | categoryal economy, synt |
| | Auto MPG | 17.3 KB | 398 | 9 | N | numeric |

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.](#)

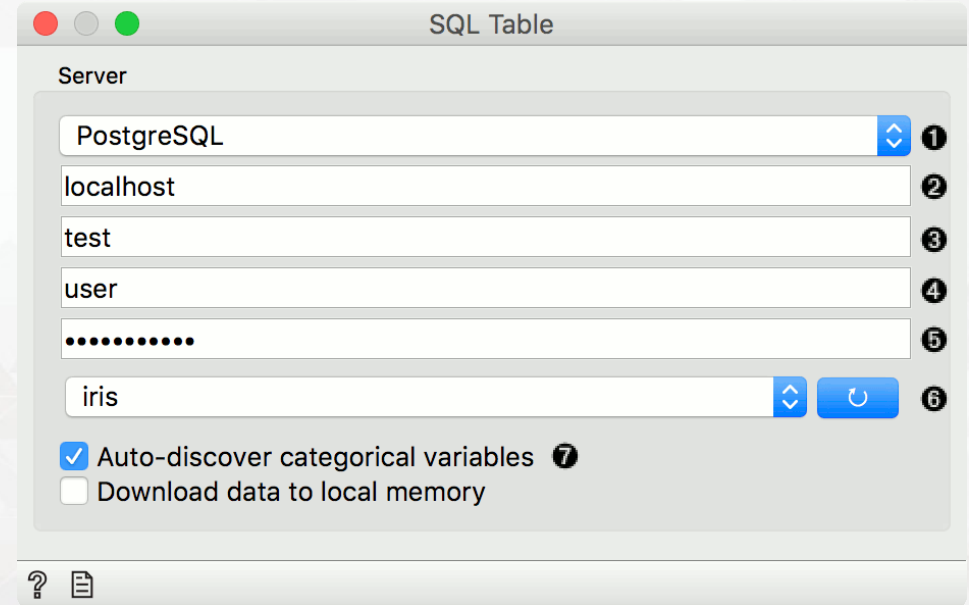
[All I See is Silhouette.](#)

References

☐ Send Data

Carregamento de dados

- SQL Table
 - Tabela de um banco de dados relacional
 - PostgreSQL e MS SQL Server
 - Precisa da instalação do backend
 - Clientes Python para os servidores
 - psycopg2 e pymssql



Carregamento de dados

- SQL Table

The screenshot displays the Orange3 data mining software interface. On the left, a vertical toolbar contains various icons for data sources, preprocessing, modeling, and evaluation. The main workspace shows two widgets: 'SQL Table' and 'Data Table', connected by a 'Data' link. The 'SQL Table' widget is configured with the following settings:

- Server: PostgreSQL
- localhost
- test
- user
-
- iris
- ☒ Auto-discover categorical variables
- ☐ Download data to local memory

The 'Data Table' widget displays the loaded data in a table format. The table has 16 rows and 6 columns: 'sepal length', 'sepal width', 'petal length', 'petal width', and 'iris'. The 'iris' column contains categorical values: 'Iris-setosa'.

Data Table Output:

| | sepal length | sepal width | petal length | petal width | iris |
|----|--------------|-------------|--------------|-------------|-------------|
| 1 | 5.100 | 3.500 | 1.400 | 0.200 | Iris-setosa |
| 2 | 4.900 | 3.000 | 1.400 | 0.200 | Iris-setosa |
| 3 | 4.700 | 3.200 | 1.300 | 0.200 | Iris-setosa |
| 4 | 4.600 | 3.100 | 1.500 | 0.200 | Iris-setosa |
| 5 | 5.000 | 3.600 | 1.400 | 0.200 | Iris-setosa |
| 6 | 5.400 | 3.900 | 1.700 | 0.400 | Iris-setosa |
| 7 | 4.600 | 3.400 | 1.400 | 0.300 | Iris-setosa |
| 8 | 5.000 | 3.400 | 1.500 | 0.200 | Iris-setosa |
| 9 | 4.400 | 2.900 | 1.400 | 0.200 | Iris-setosa |
| 10 | 4.900 | 3.100 | 1.500 | 0.100 | Iris-setosa |
| 11 | 5.400 | 3.700 | 1.500 | 0.200 | Iris-setosa |
| 12 | 4.800 | 3.400 | 1.600 | 0.200 | Iris-setosa |
| 13 | 4.800 | 3.000 | 1.400 | 0.100 | Iris-setosa |
| 14 | 4.300 | 3.000 | 1.100 | 0.100 | Iris-setosa |
| 15 | 5.800 | 4.000 | 1.200 | 0.200 | Iris-setosa |
| 16 | 5.700 | 4.400 | 1.500 | 0.400 | Iris-setosa |

The 'Data Table' widget also includes a sidebar with the following information:

- Info:** 150 instances, 5 features, No target variable, No meta attributes.
- Variables:** ☒ Show variable labels (if present), ☐ Visualize numeric values, ☒ Color by instance classes.
- Selection:** ☒ Select full rows.
- Buttons:** Restore Original Order, Send Automatically (checked).

Analizando dados

- Data Table

- Visualização dos dados de uma fonte
 - Mostra propriedades avançadas
 - Labels, colorização
- Pode ser usada como fonte para outros widgets de dados



Data Table

Info

- 150 instances (no missing values)
- 4 features (no missing values)
- Discrete class with 3 values (no missing values)
- No meta attributes

Variables

- ☒ Show variable labels (if present)
- ☒ Visualize numeric values
- ☒ Color by instance classes

Selection

- ☒ Select full rows

Restore Original Order

☒ Send Automatically

| | iris | sepal length | sepal width | petal length | petal width |
|----|-----------------|--------------|-------------|--------------|-------------|
| 40 | Iris-setosa | 5.1 | 3.4 | 1.5 | 0.2 |
| 41 | Iris-setosa | 5.0 | 3.5 | 1.3 | 0.3 |
| 42 | Iris-setosa | 4.5 | 2.3 | 1.3 | 0.3 |
| 43 | Iris-setosa | 4.4 | 3.2 | 1.3 | 0.2 |
| 44 | Iris-setosa | 5.0 | 3.5 | 1.6 | 0.6 |
| 45 | Iris-setosa | 5.1 | 3.8 | 1.9 | 0.4 |
| 46 | Iris-setosa | 4.8 | 3.0 | 1.4 | 0.3 |
| 47 | Iris-setosa | 5.1 | 3.8 | 1.6 | 0.2 |
| 48 | Iris-setosa | 4.6 | 3.2 | 1.4 | 0.2 |
| 49 | Iris-setosa | 5.3 | 3.7 | 1.5 | 0.2 |
| 50 | Iris-setosa | 5.0 | 3.3 | 1.4 | 0.2 |
| 51 | Iris-versicolor | 7.0 | 3.2 | 4.7 | 1.4 |
| 52 | Iris-versicolor | 6.4 | 3.2 | 4.5 | 1.5 |
| 53 | Iris-versicolor | 6.9 | 3.1 | 4.9 | 1.5 |
| 54 | Iris-versicolor | 5.5 | 2.3 | 4.0 | 1.3 |
| 55 | Iris-versicolor | 6.5 | 2.8 | 4.6 | 1.5 |
| 56 | Iris-versicolor | 5.7 | 2.8 | 4.5 | 1.3 |
| 57 | Iris-versicolor | 6.3 | 3.3 | 4.7 | 1.6 |
| 58 | Iris-versicolor | 4.9 | 2.4 | 3.3 | 1.0 |
| 59 | Iris-versicolor | 6.6 | 2.9 | 4.6 | 1.3 |
| 60 | Iris-versicolor | 5.2 | 2.7 | 3.9 | 1.4 |
| 61 | Iris-versicolor | 5.0 | 2.0 | 3.5 | 1.0 |
| 62 | Iris-versicolor | 5.9 | 3.0 | 4.2 | 1.5 |
| 63 | Iris-versicolor | 6.0 | 2.2 | 4.0 | 1.0 |

Manipulando dados

- Select Columns

- Projeção



- Select Rows

- Filtro



- Merge Data

- Join



- Concatenate

- Union ou Intersection



- Select by Data Index

- Restaura label ou id por um índice



Manipulando dados

- Select Columns

- Projeção



- Select Rows

- Filtro



- Merge Data

- Join



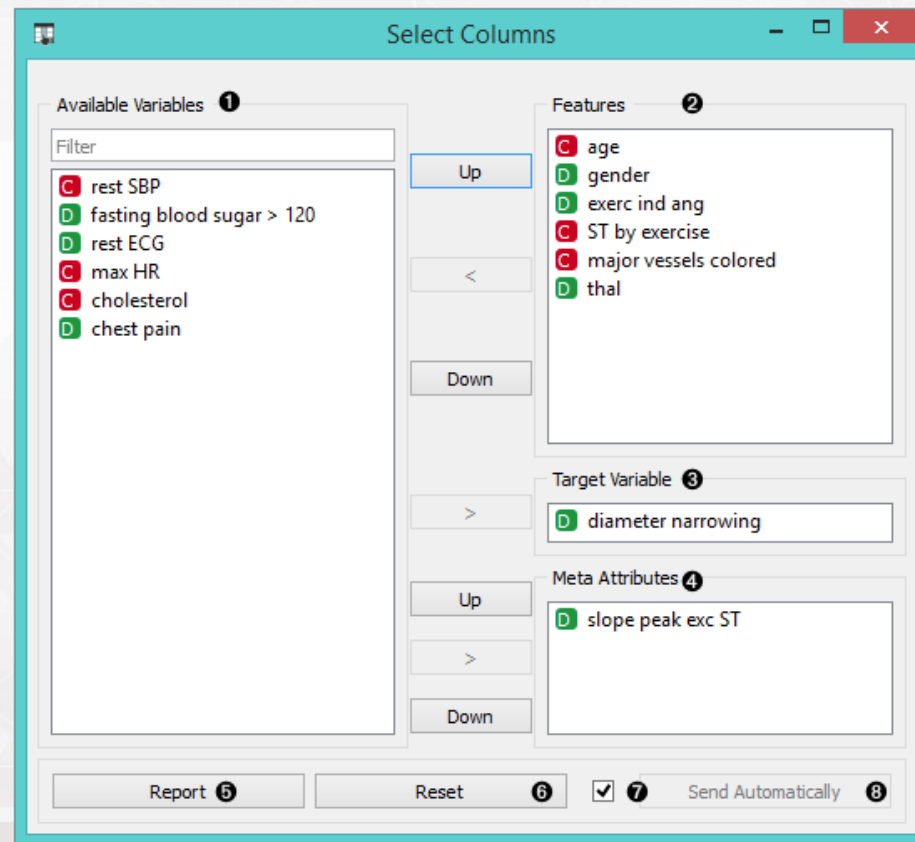
- Concatenate

- Union ou Intersection



- Select by Data Index

- Restaura label ou id por um índice



Manipulando dados

- Select Columns

- Projeção



- Select Rows

- Filtro



- Merge Data

- Join



- Concatenate

- Union ou Intersection



- Select by Data Index

- Restaura label ou id por um índice



The screenshot shows a 'Select Rows' dialog box with the following sections:

- Conditions 1:** A table with three rows of conditions.

| | | |
|-----------|----------|--------|
| fuel-type | is not | gas |
| make | is | toyota |
| price | is below | 25000 |
- Buttons:** 'Add Condition' (labeled 2), 'Add All Variables' (labeled 3), and 'Remove All' (labeled 4).
- Data 5:** 'In: ~205 rows, 26 variables' and 'Out: ~3 rows, 13 variables'.
- Purging 6:** Two checked options: 'Remove unused features' and 'Remove unused classes'.
- Buttons:** 'Report' (labeled 7) and 'Send' (labeled 8, with 'Send automatically' checked).

Manipulando dados

- Select Columns

- Projeção



- Select Rows

- Filtro



- Merge Data

- Join



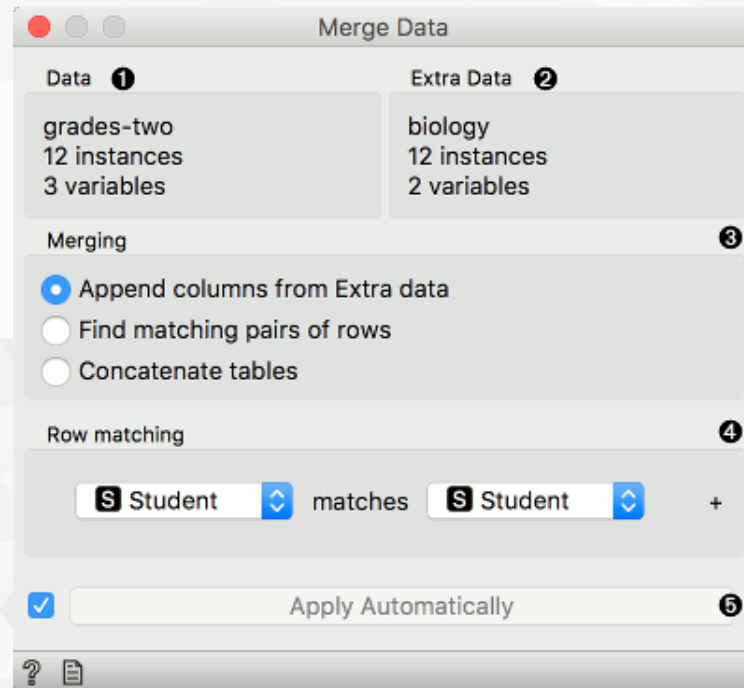
- Concatenate

- Union ou Intersection



- Select by Data Index

- Restaura label ou id por um índice



Informações sobre dados



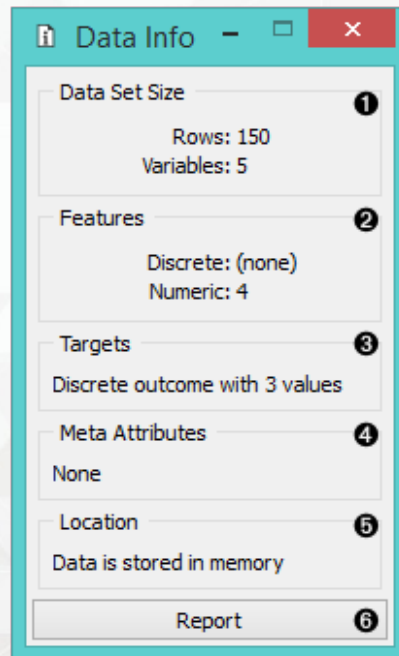
- Data Info

- Informações sobre o dataset selecionado
 - Tamanho, features, localização



- Feature Statistics

- Estatísticas básicas dos atributos



Informações sobre dados



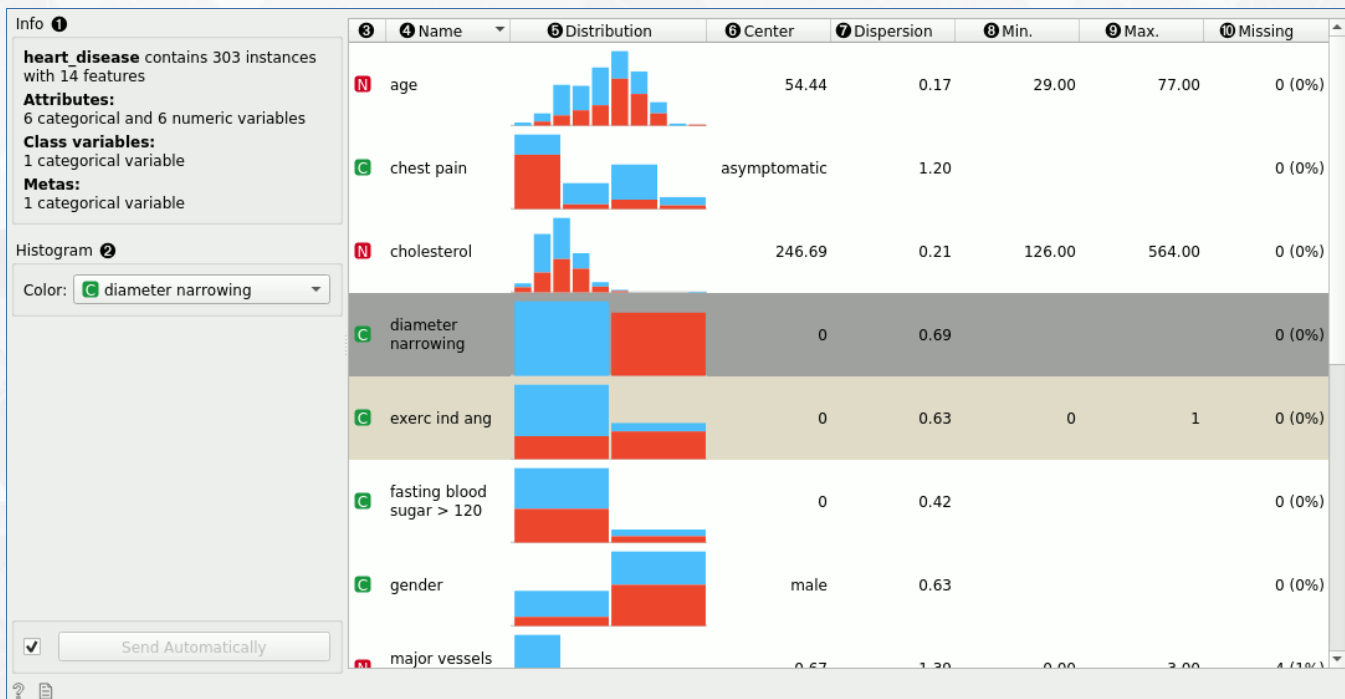
- Data Info

- Informações sobre o dataset selecionado
 - Tamanho, features, localização



- Feature Statistics

- Estatísticas básicas dos atributos



Informações sobre dados



- Correlations

- Calcula a correlação entre todos os pares de atributos
- Correlações de Pearson e Spearman
 - Correlação linear e dependência estatística

- Box Plot



- Mostra distribuição dos valores de atributos
 - Mediana, média, desvio padrão, primeiro e terceiro quartis
 - Para atributos discretos, mostra a proporção relativa de instâncias
- Útil para descoberta de anomalias em dados

| Pairwise Pearson correlation | | | |
|------------------------------|-------|-------|-------|
| Filter ... | | | |
| 1 | RAD | TAX | 0.91 |
| 2 | INDUS | NOX | 0.764 |
| 3 | AGE | NOX | 0.731 |
| 4 | INDUS | TAX | 0.721 |
| 5 | NOX | TAX | 0.668 |
| 6 | DIS | ZN | 0.664 |
| 7 | AGE | INDUS | 0.645 |
| 8 | CRIM | RAD | 0.626 |
| 9 | NOX | RAD | 0.611 |

Finished

Informações sobre dados



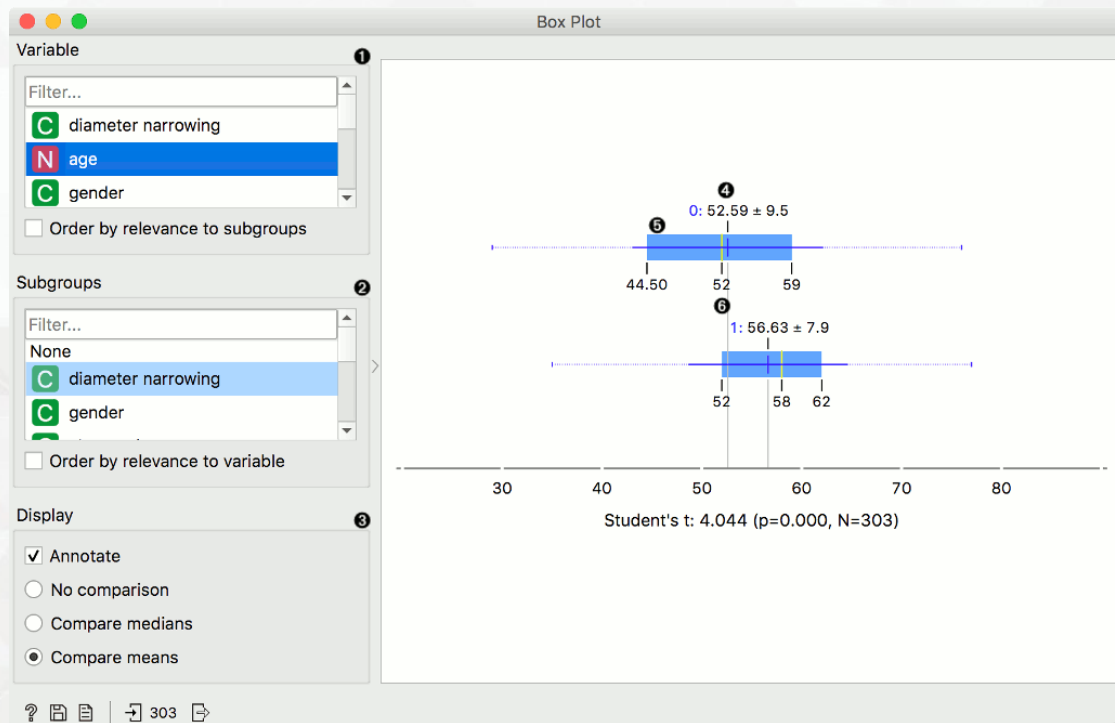
- Correlations

- Calcula a correlação entre todos os pares de atributos
- Correlações de Pearson e Spearman
 - Correlação linear e dependência estatística



- Box Plot

- Mostra distribuição dos valores de atributos
 - Mediana, média, desvio padrão, primeiro e terceiro quartis
 - Para atributos discretos, mostra a proporção relativa de instâncias
- Útil para descoberta de anomalias em dados



Informações sobre dados



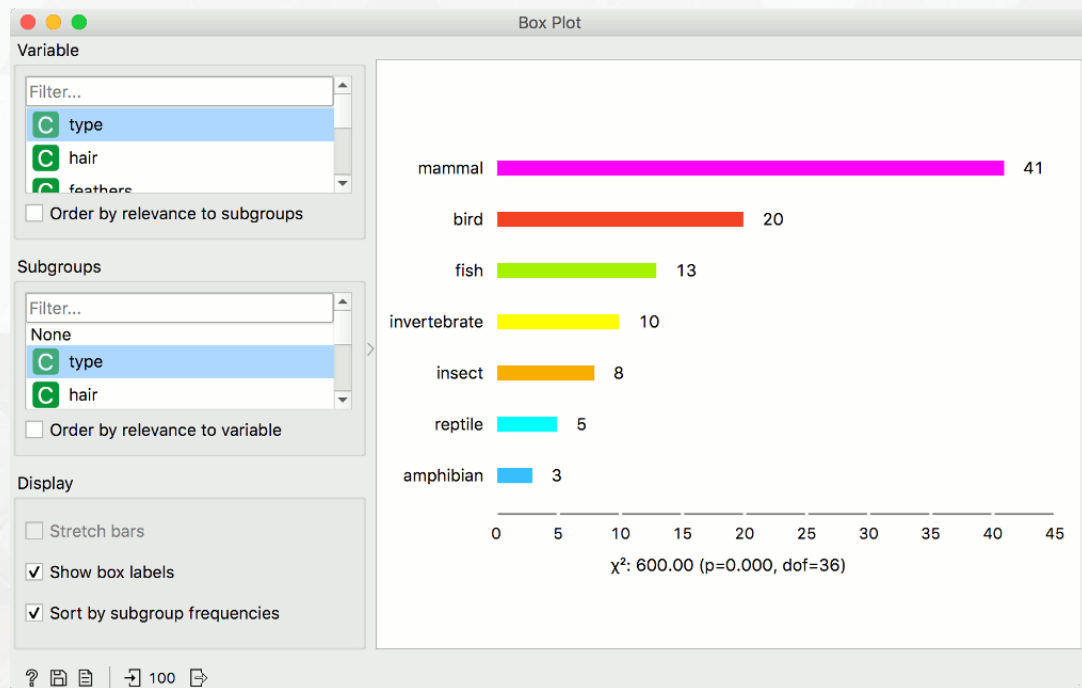
- Correlations

- Calcula a correlação entre todos os pares de atributos
- Correlações de Pearson e Spearman
 - Correlação linear e dependência estatística



- Box Plot

- Mostra distribuição dos valores de atributos
 - Mediana, média, desvio padrão, primeiro e terceiro quartis
 - Para atributos discretos, mostra a proporção relativa de instâncias
- Útil para descoberta de anomalias em dados



Informações sobre dados

- Outliers



- Detecção de outliers

- One Class SVM, Covariance Estimator, Local Outlier Factor, Isolation Forest

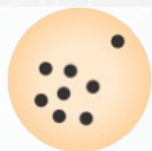
- Usa DataTables para analisar a saída

- Outliers, Inliers



Informações sobre dados

- Outliers



- Detecção de outliers

- One Class SVM, Covariance Estimator, Local Outlier Factor, Isolation Forest

- Usa DataTables para analisar a saída

- Outliers, Inliers





Obrigado

leandro@utfpr.edu.br

<http://lapti.ct.utfpr.edu.br>