Orange
Data Analysis Workflow

2023

<lapti>
UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Conteúdo

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

Orange

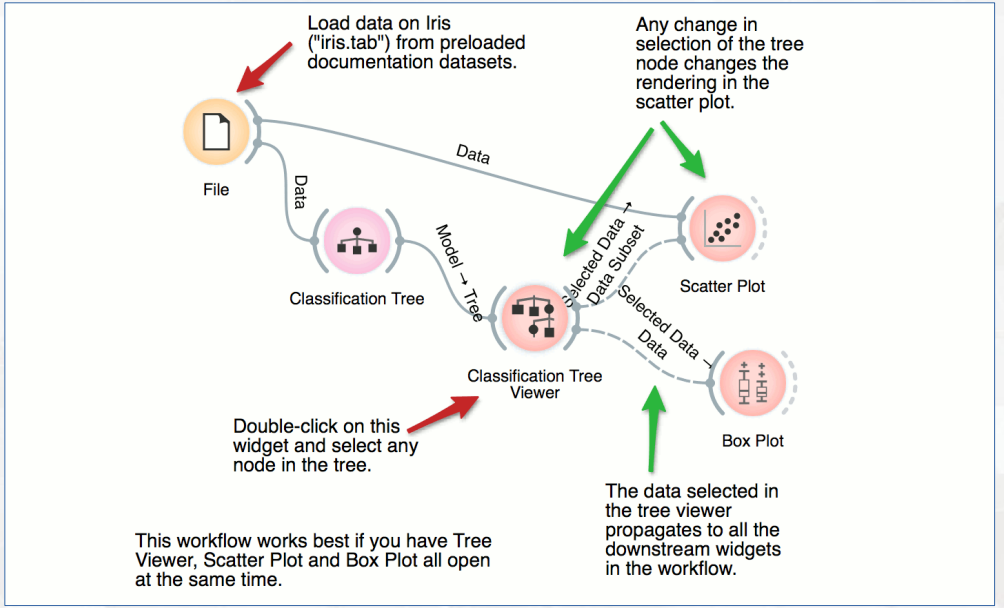
- Desenvolvido pela Universidade de Ljubljani
  - Eslovênia
  - https://www.uni-lj.si/university/
- Ferramenta de visualização e manipulação de dados
  - Data mining
  - Machine learning
  - Programação visual
    - Lema: “anyone can do data science!”
- Desenvolvida em Python, C, C++ e Cython
  - Utilização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

Orange

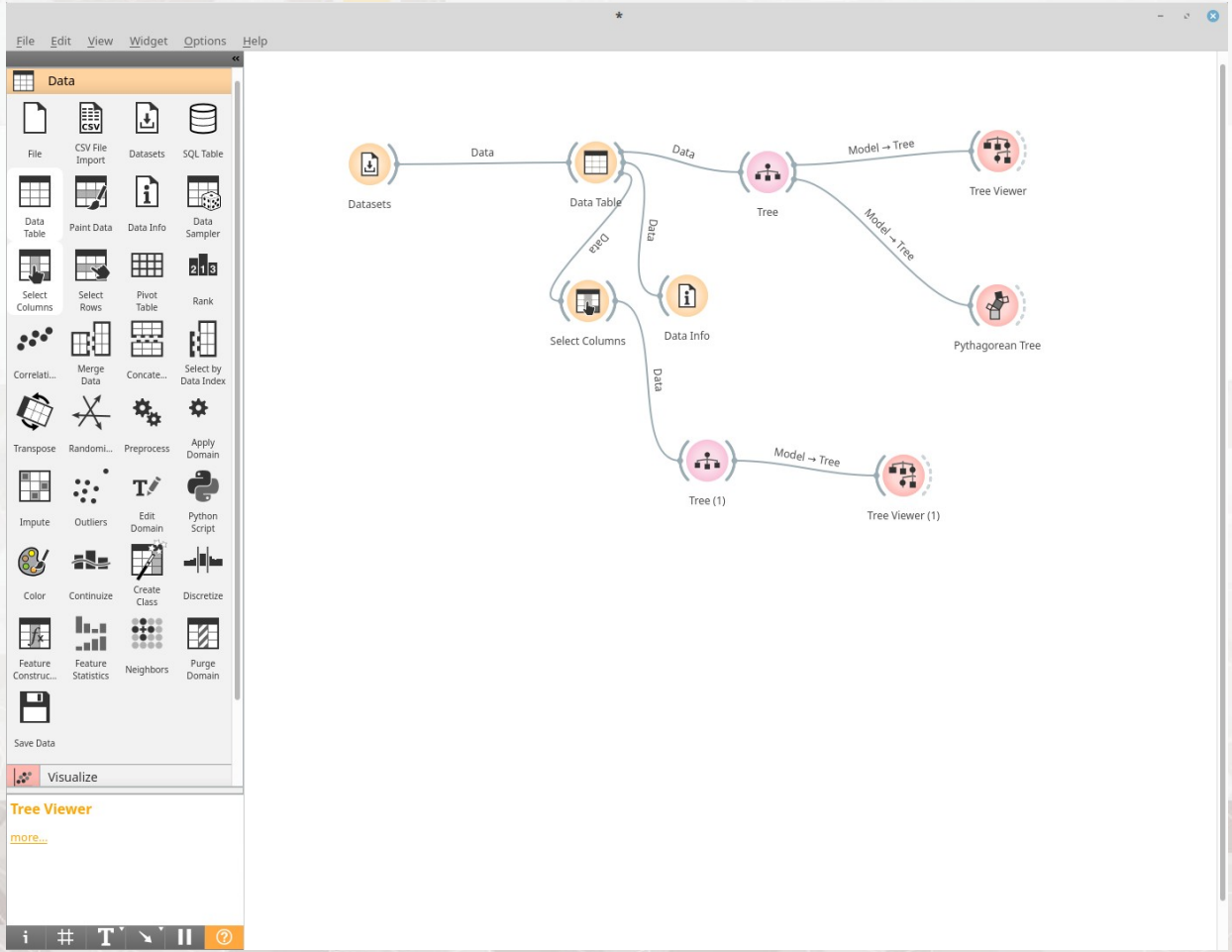
- Site principal
  - https://orangedatamining.com/
  - https://orange.biolab.si/
- Download e instalação
  - https://orangedatamining.com/download/
- Documentação
  - https://orangedatamining.com/docs/
  - Tutoriais e vídeos disponíveis

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

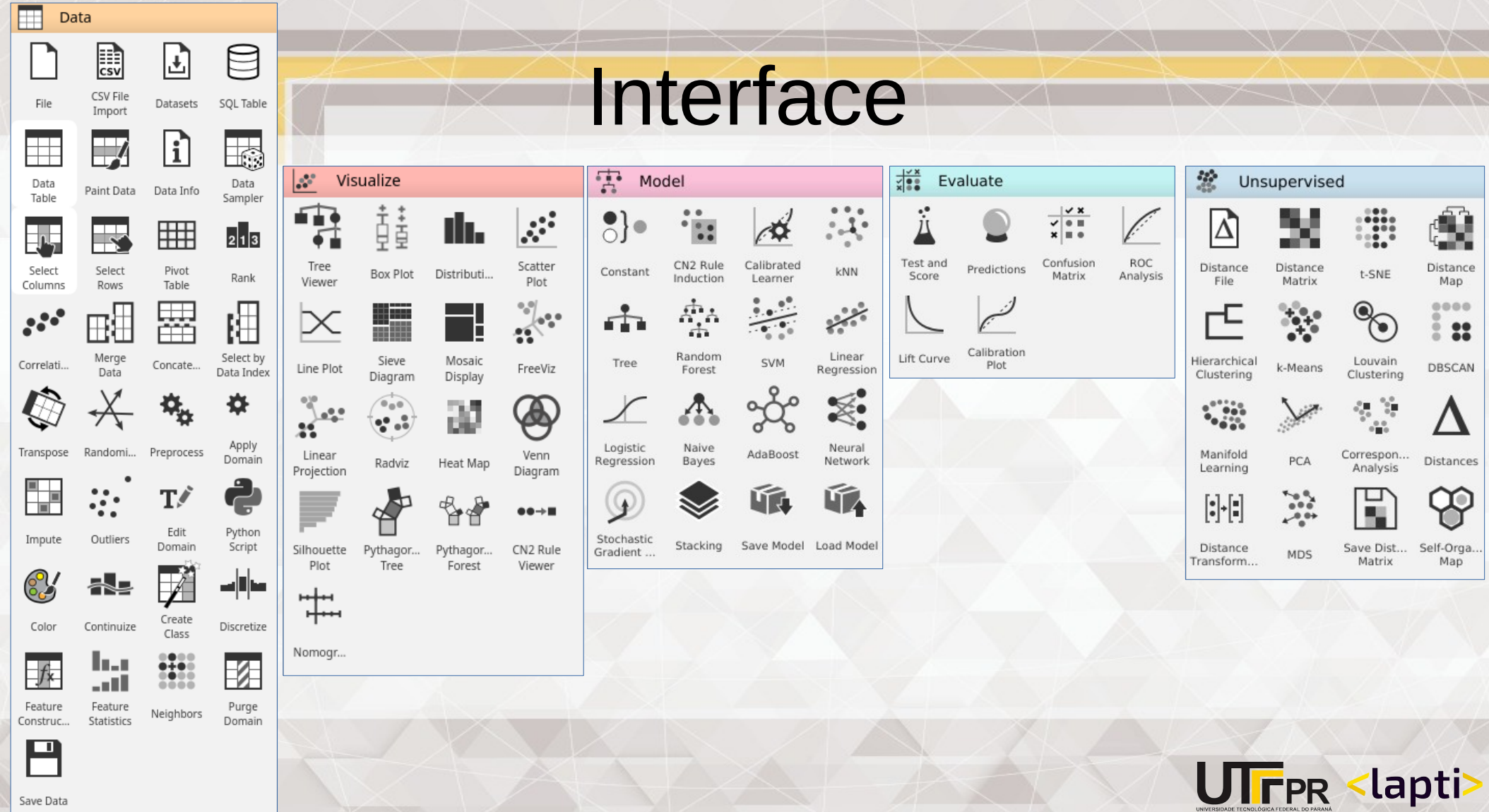


Interface



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

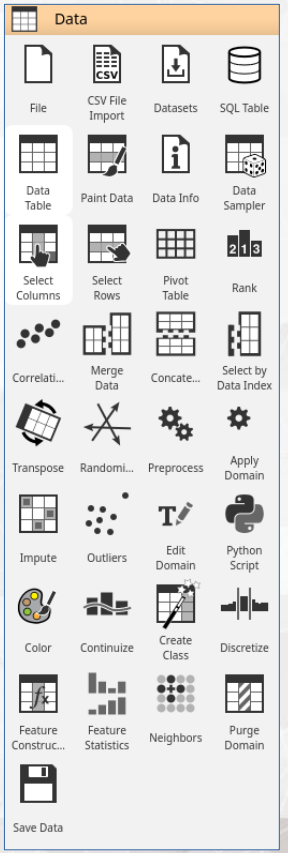
Interface



UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
<lapti>

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

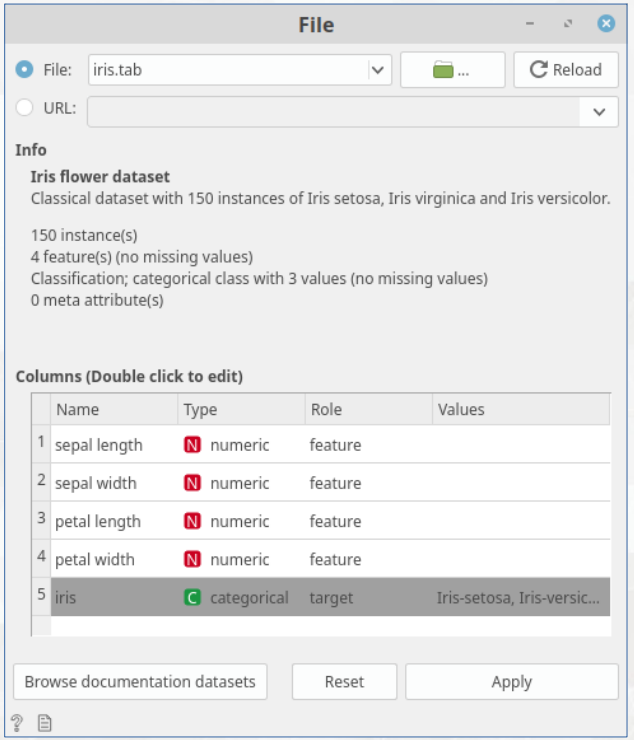
Modelos de ML e Data Mining

- 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

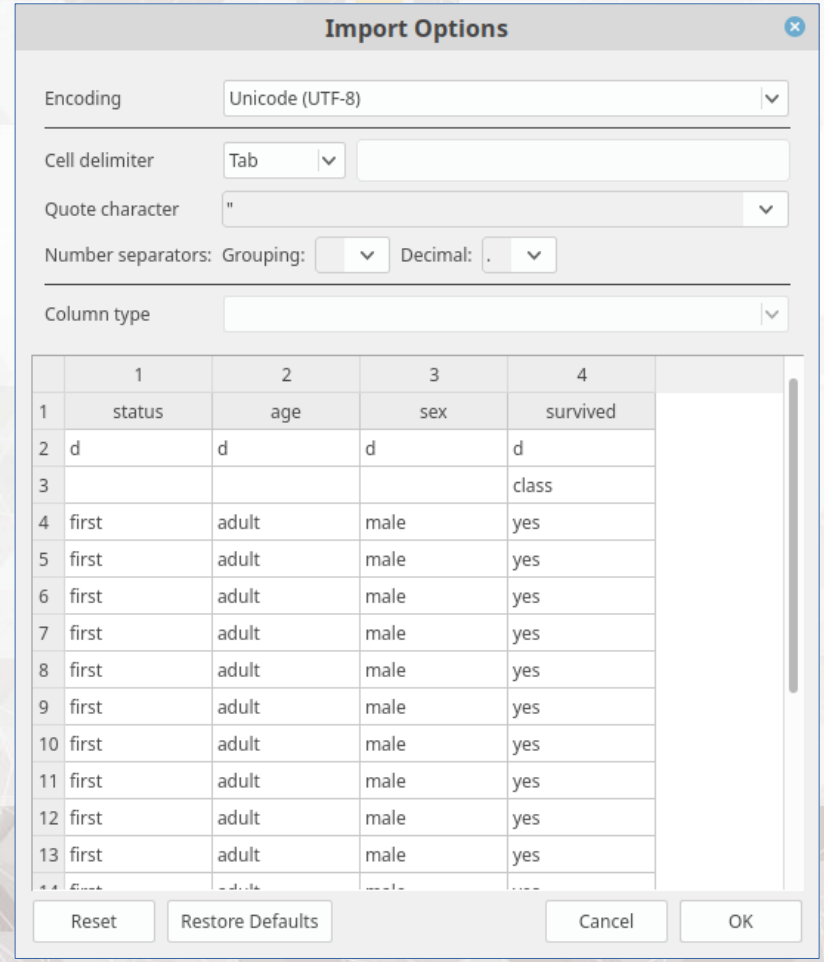
UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ <lapti>

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



Carregamento de dados



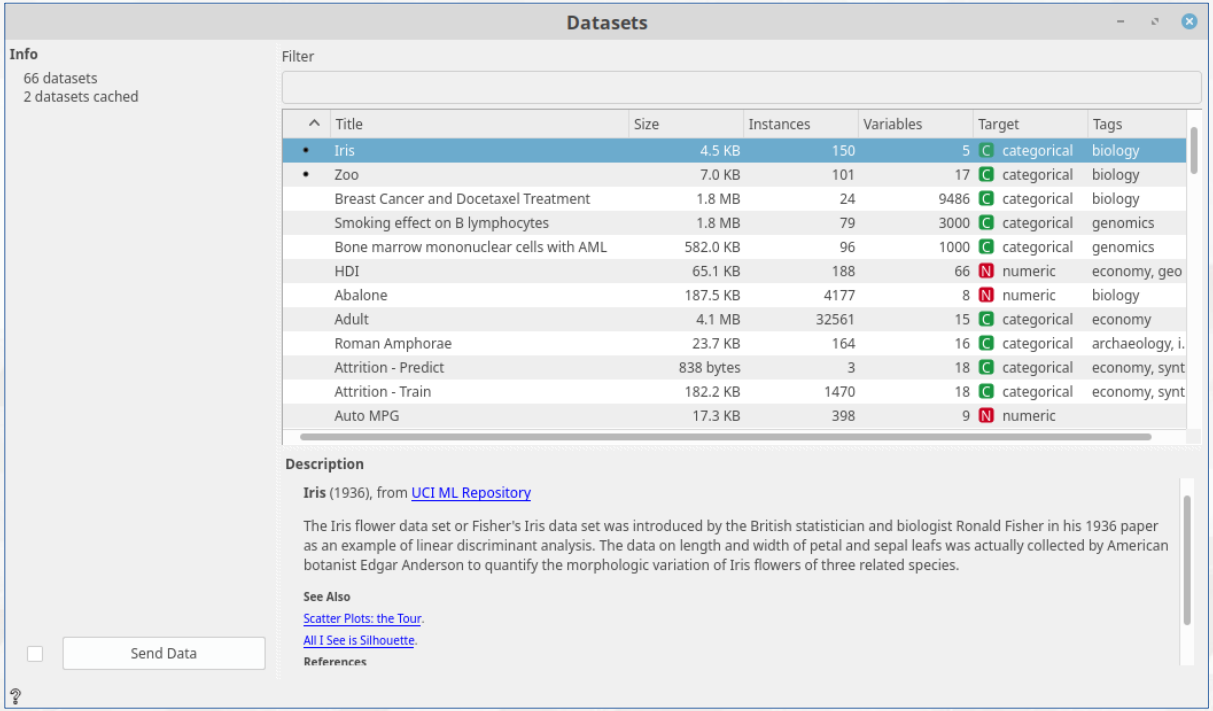
	1	2	3	4
1	status	age	sex	survived
2	d	d	d	d
3				class
4	first	adult	male	yes
5	first	adult	male	yes
6	first	adult	male	yes
7	first	adult	male	yes
8	first	adult	male	yes
9	first	adult	male	yes
10	first	adult	male	yes
11	first	adult	male	yes
12	first	adult	male	yes
13	first	adult	male	yes

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

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ <lapti>

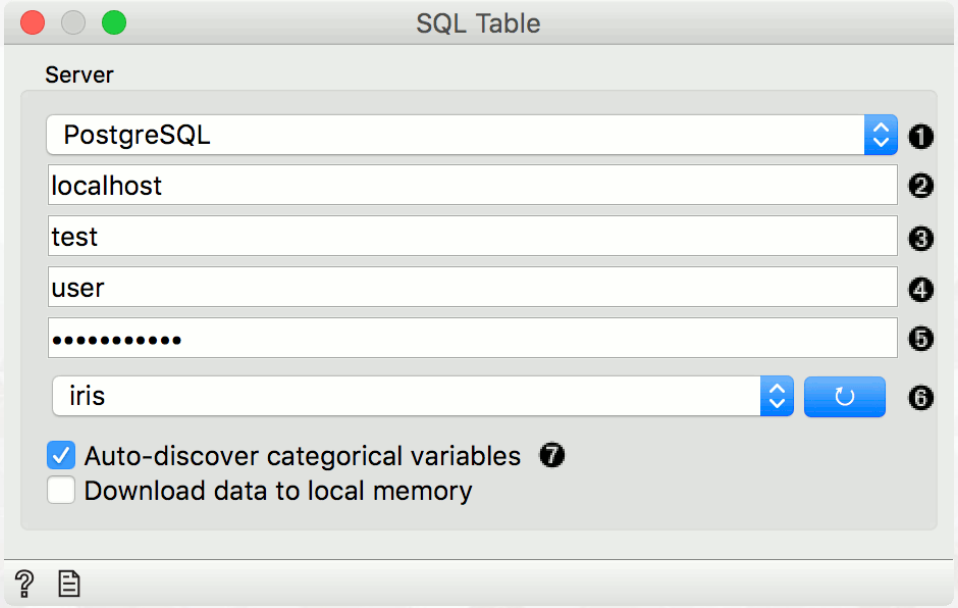
Carregamento de dados

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

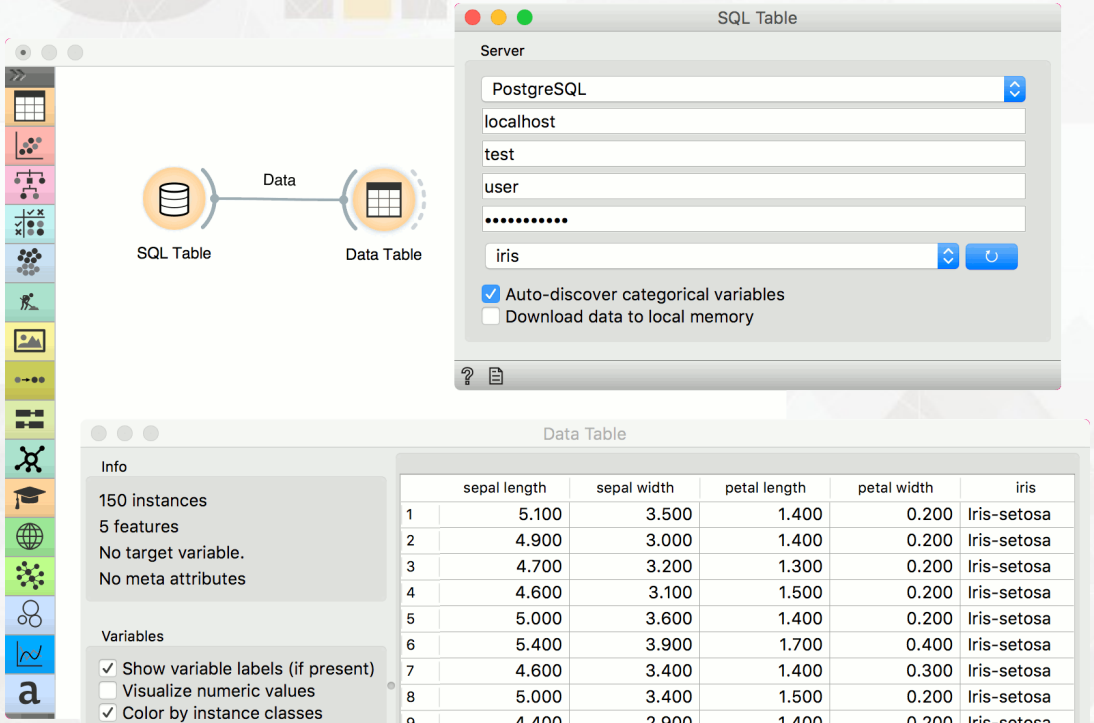


Carregamento de dados

- 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

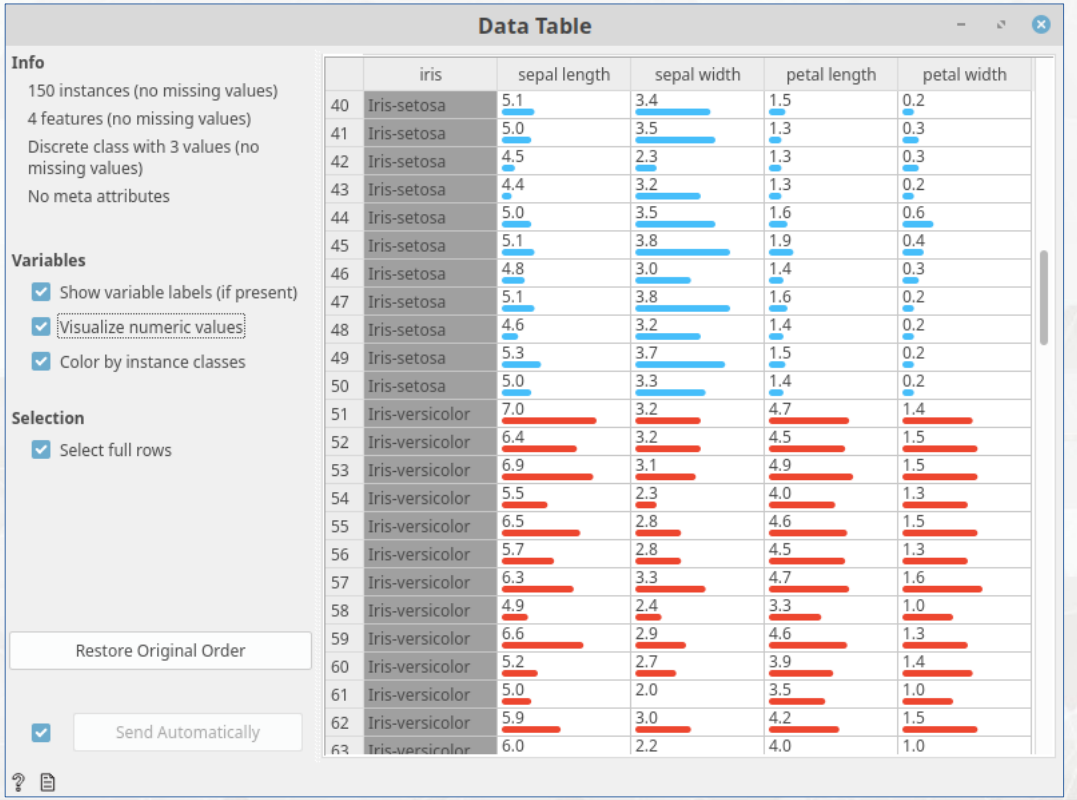


	sepal length	sepal width	petal length	petal width	iris
1	5.100	3.500	1.400	0.200	Iris-setosa
2	4.900	3.000	1.400	0.200	Iris-setosa
3	4.700	3.200	1.300	0.200	Iris-setosa
4	4.600	3.100	1.500	0.200	Iris-setosa
5	5.000	3.600	1.400	0.200	Iris-setosa
6	5.400	3.900	1.700	0.400	Iris-setosa
7	4.600	3.400	1.400	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

- SQL Table

Analisando dados

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



Manipulando dados

- Select Columns
  - Projeção

- Select Rows
  - Filtro

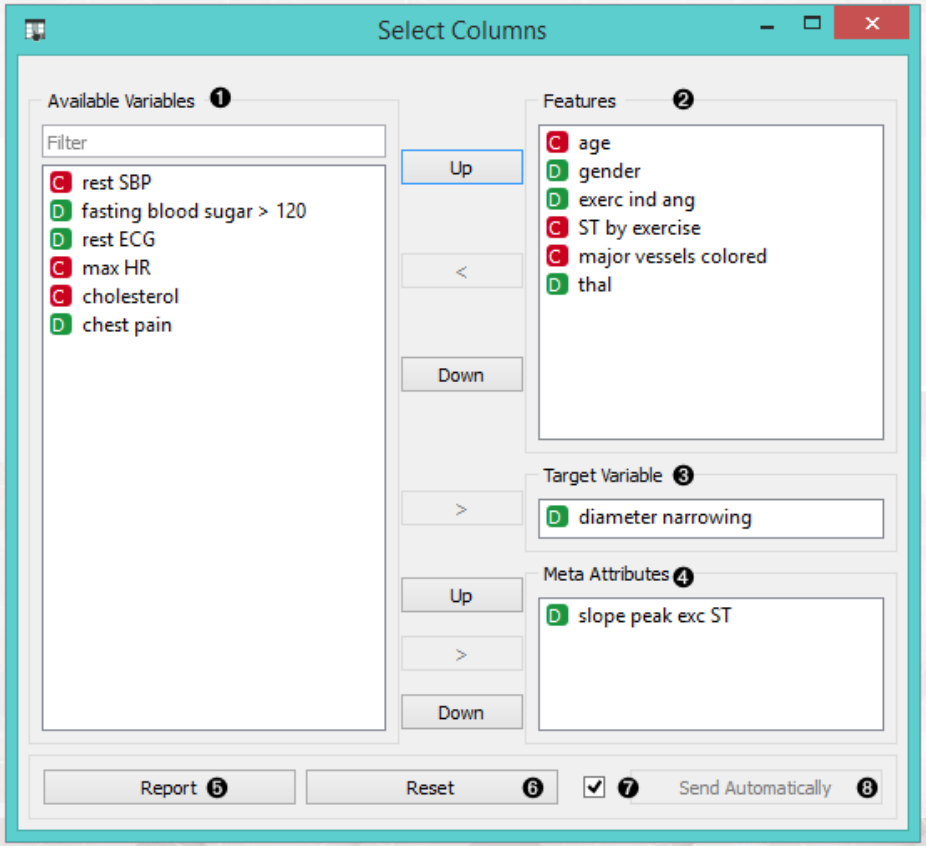
- Merge Data
  - Join

- Concatenate
  - Union ou Intersection

- Select by Data Index
  - Restaura label ou id por um índice

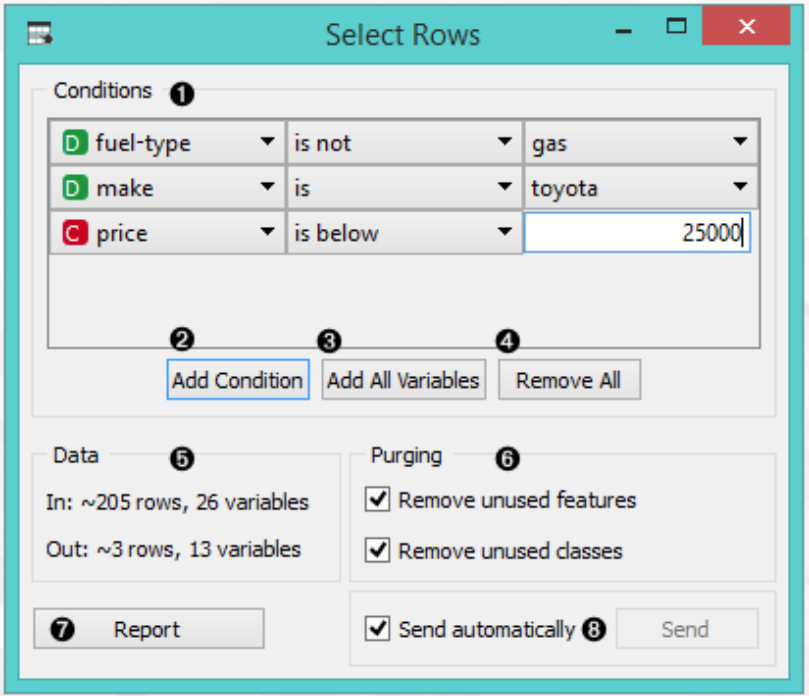
Manipulando dados

- Select Columns
  - Projeção
- Select Rows
  - Filtro
- Merge Data
  - Join
- Concatenate
  - Union ou Intersection
- Select by Data Index
  - Restaura label ou id por um índice



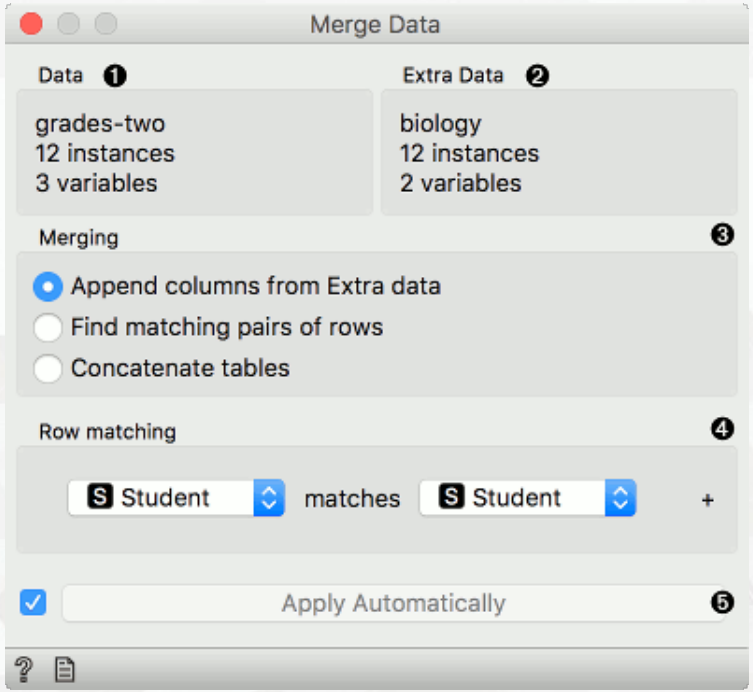
Manipulando dados

- Select Columns
  - Projeção
- Select Rows
  - Filtro
- Merge Data
  - Join
- Concatenate
  - Union ou Intersection
- Select by Data Index
  - Restaura label ou id por um índice



Manipulando dados

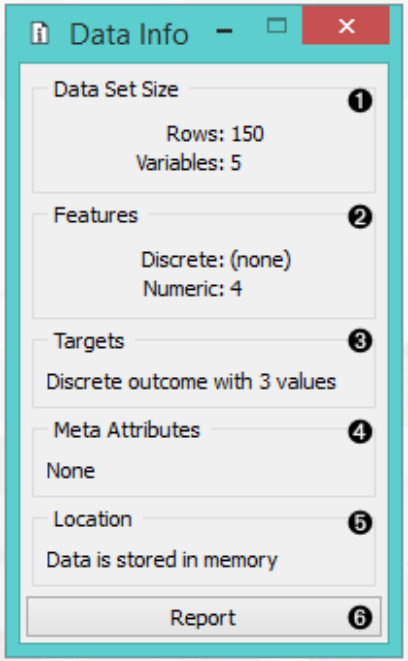
- Select Columns
  - Projeção
- Select Rows
  - Filtro
- Merge Data
  - Join
- Concatenate
  - Union ou Intersection
- Select by Data Index
  - Restaura label ou id por um índice



Informações sobre dados

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

- Feature Statistics
  - Estatísticas básicas dos atributos



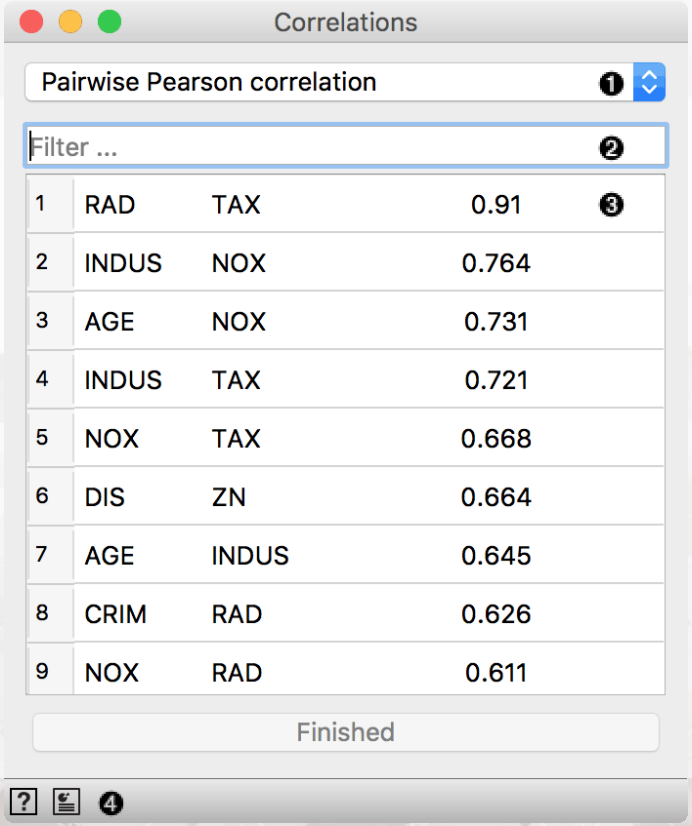
Informações sobre dados

- 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

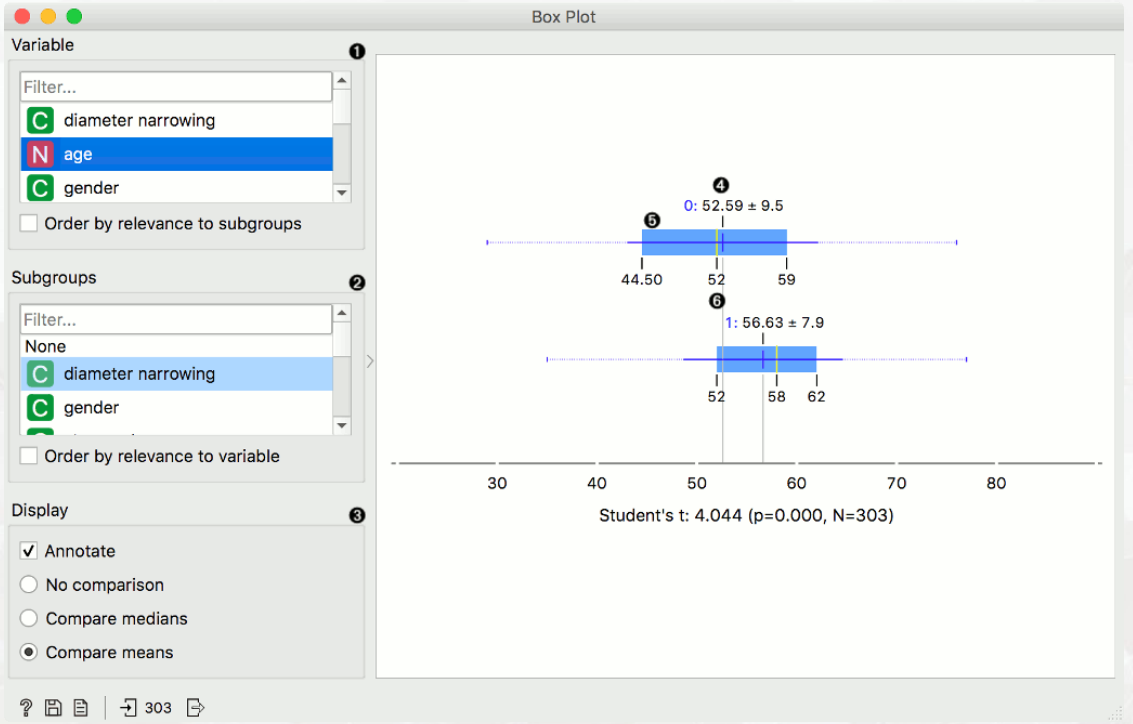
- Box Plot
  - Mostra distribuição dos valores de atributos
    - Mediana, média, desvio padrão, primeiro e terceiro quartis
    - Para atributos discretos, mostra a proporção relativa de instâncias
  - Útil para descoberta de anomalias em dados



Informações sobre dados

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

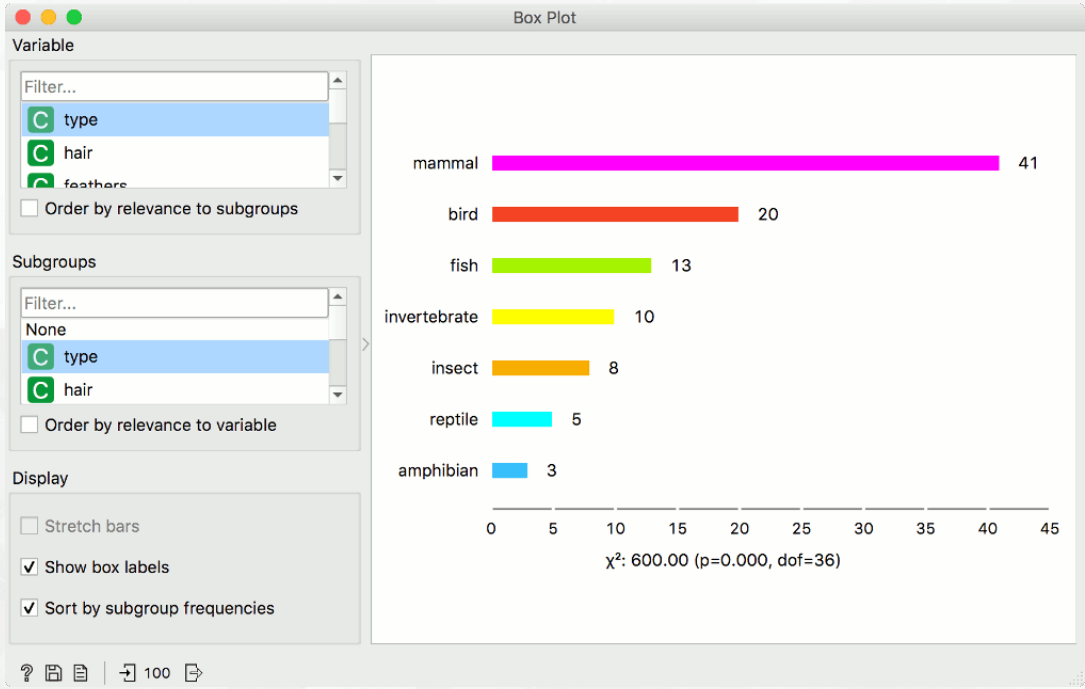
- Box Plot
  - Mostra distribuição dos valores de atributos
    - Mediana, média, desvio padrão, primeiro e terceiro quartis
    - Para atributos discretos, mostra a proporção relativa de instâncias
  - Útil para descoberta de anomalias em dados



Informações sobre dados

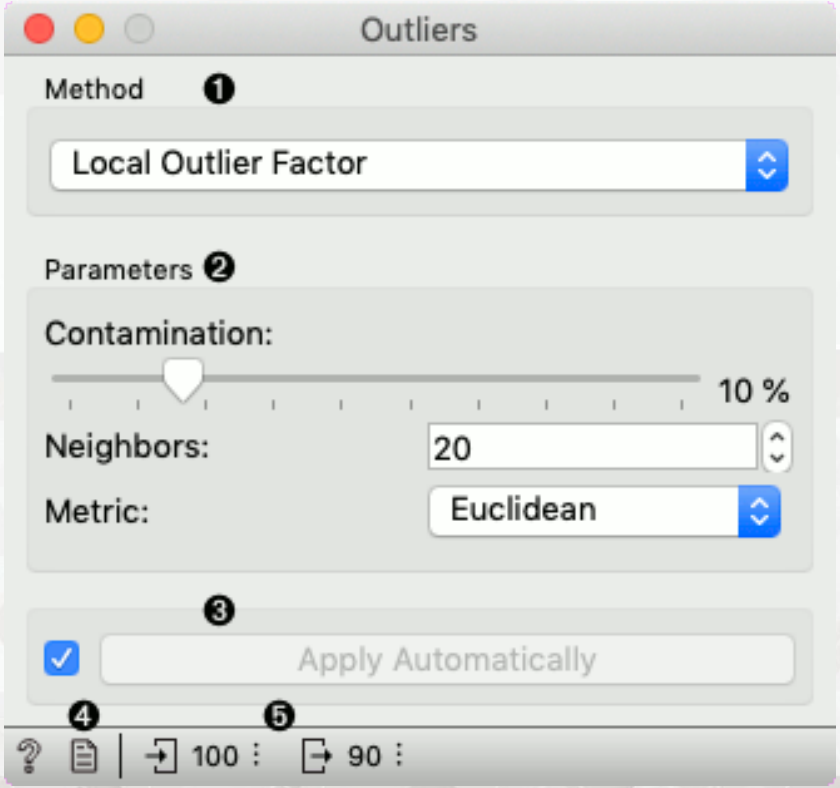
- Correlations
  - Calcula a correlação entre todos os pares de atributos
  - Correlações de Pearson e Spearman
    - Correlação linear e dependência estatística

- 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



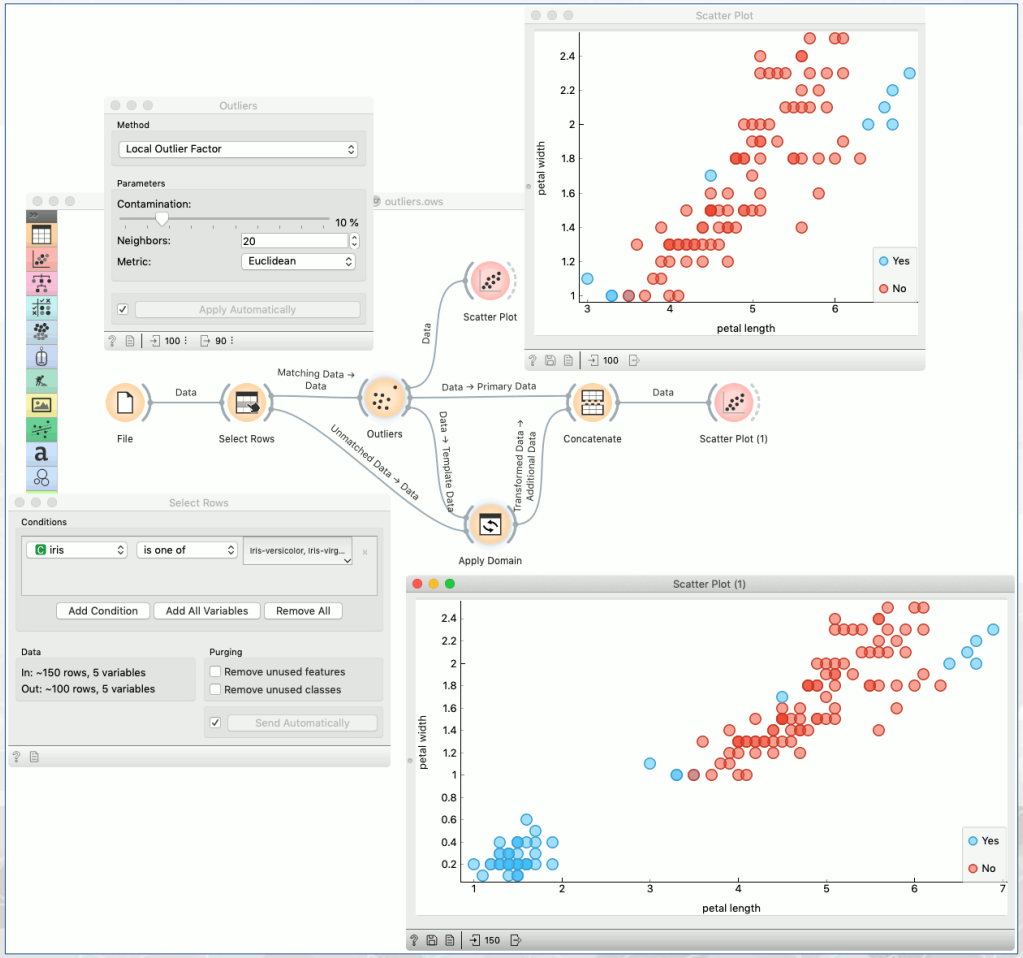
Informações sobre dados

- 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