Sumário

Apresentação

O rapidminer é uma ferramenta *open source* que fornece um ambiente visual para realização de processos de *data mining*. Apresenta várias funcionalidades que estão disponíveis sob a forma de operadores (caixas) que podem ser interconectadas, como por exemplo: testes e validações, seleção de atributos, tratamento de dados, classificação, associação e agrupamentos.

Instalação

Acesse o site `https://my.rapidminer.com/nexus/account/index.html#downloads`

Existem versões para Windows 32 e 64 bits, Mac (Apple) e Linux.

Click na versão desejada. Um programa executável será baixado. Execute-o e siga as instruções de instalação.

Sempre que usar o rapidminer será verificado se existe nova versão disponível. Siga as instruções para realizar a atualização de versão.

Obs.: os testes foram realizados no Windows 10 (64 bits).

Primeiros passos

Abrindo

Se a instalação deixou o ícone  na área de trabalho, click sobre ele para executar o rapidminer.

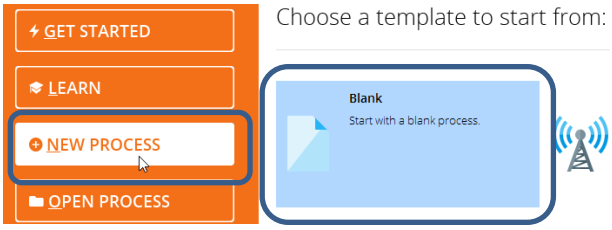
Caso contrário, click sobre “Iniciar” no Windows e selecione RapidMiner Studio.

Enquanto o programa estiver carregando a seguinte tela será exibida:

O rapidminer iniciará exibindo uma tela de entrada.

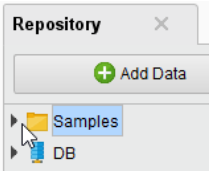
Novo projeto

Na tela de entrada click em “New Process” e em seguida em “Brank” para criar um projeto vazio.



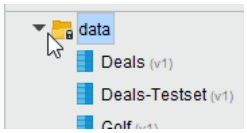
Adicionando dados

No painel “Repository” à esquerda, localize a base de dados desejada. Para esta introdução, click na seta à esquerda de “Samples” para abrir as opções logo abaixo.

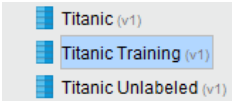


Observe que aparecem novas opções abaixo de “Samples” e a seta à esquerda passa a apontar para baixo.

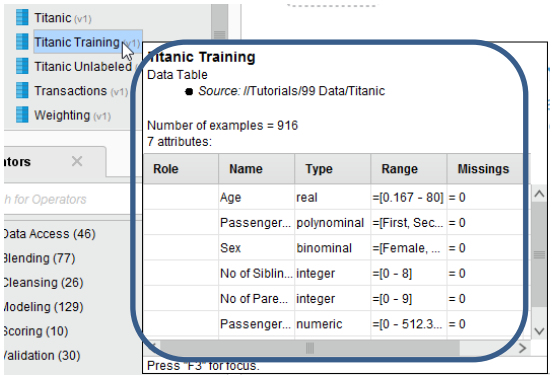
Repita o procedimento para abrir as opções abaixo de “Data”:



Role a barra de rolagem até exibir “Titanic Training”:



Observe que aparece uma descrição da base de dados quando o mouse permanece sobre ela:



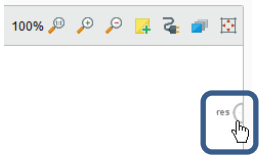
Click sobre “Titanic Training”, mantenha o mouse pressionado e arraste-o até o meio do painel “Process”.
Observe que o arquivo é adicionado como uma caixa denominada “Retrieve Titanic Training”:

Primeiros resultados

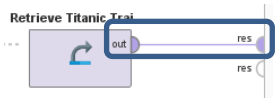
Click sobre o ícone “out” existente à direita da caixa “Retrieve Titanic Training”:

Uma linha será criada e se arrastará junto com o mouse:

Click sobre o ícone “res” existente no canto superior direito do painel “Process”:

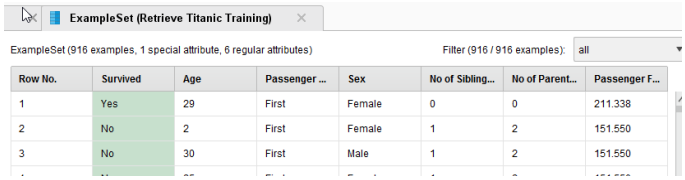


Será criada uma linha ligando os 2 ícones.

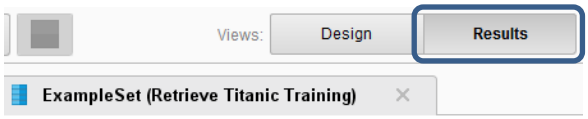


Click no botão “Run” na barra superior:

Aparecerá a janela “ExampleSet” com os dados da tabela:



Click na barra de rolagem para exibir mais dados.
Observe que o indicador “Results” ficou destacado:



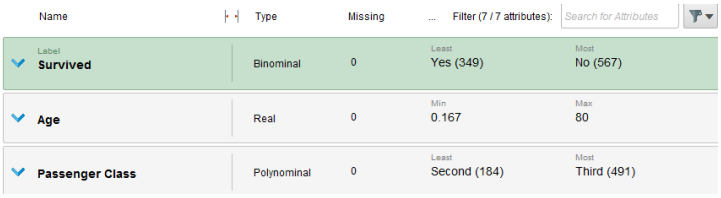
Analisando os dados

Click no botão “Statistics” na esquerda

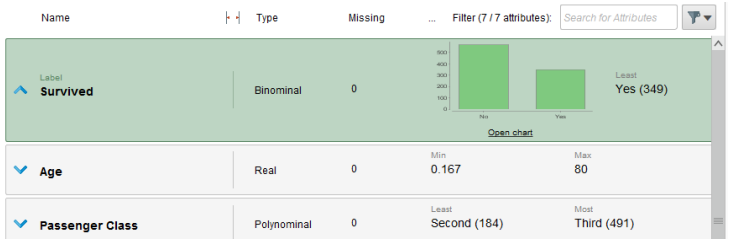
Será exibido um resumo estatístico dos dados como:

- binomiais: quantidade de cada.
- polinomial: quantidade de 2 valores e valores de todos.
- numérico (inteiro e real): mínimo, máximo e valor médio.

Em todos será exibida a quantidade de registros em que uma coluna não tenha valor (Missing).



Click sobre a primeira linha. Os valores são substituídos por um gráfico de distribuição dos valores.

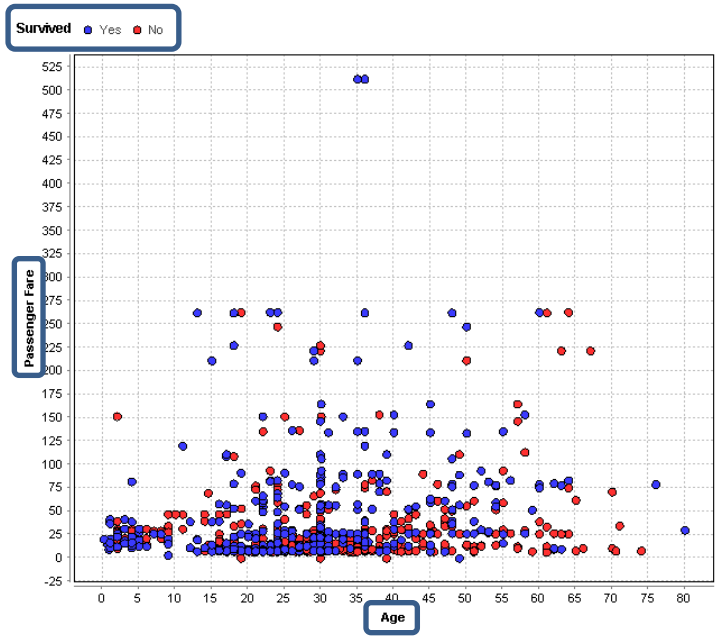


Click no botão “Charts” na esquerda.

Será exibido um gráfico do tipo “Scalter”. Selecione:

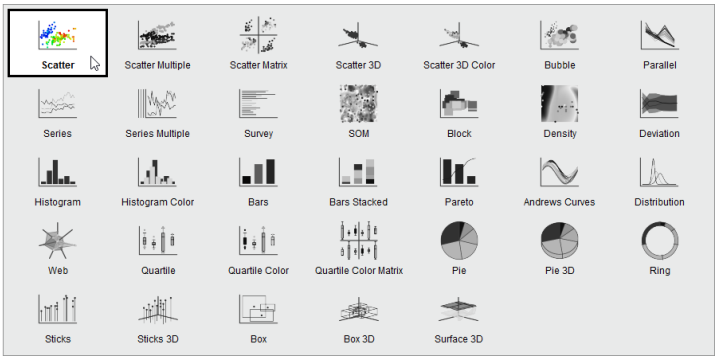
- x-Axis: Age
- Y-Axis: Passenger Fare
- Color Column: Survived

Observe o gráfico gerado, onde se vê a identificação dos campos usados nos eixos e do campo usado nos marcadores.



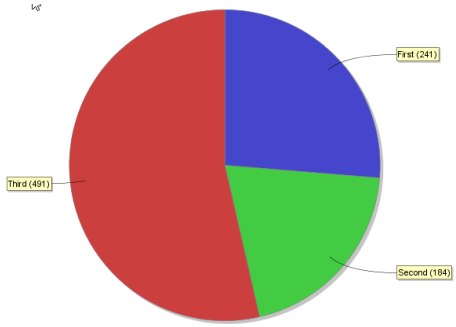
Click no botão abaixo de “Charter style”:

Será exibida uma lista de tipos de gráficos disponíveis como o tipo de gráfico atual destacado:



Selecione o tipo “Pie” (o gráfico fica colorido quando o mouse passa sobre ele):

Será exibido um gráfico de torta com os percentuais de cada valor do campo selecionado em “Group-By Column”:



Titanic Training

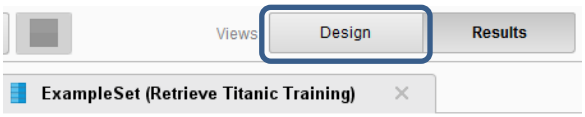
O arquivo “Titanic Training” contém dados de passageiros do Titanic e é formado pelos seguintes campos:

- Row No.: número da linha (na verdade informação adicionada pelo rapidminer)
- Survived: indicar se o passageiro sobreviveu (Yes, No)
- Age: idade do passageiro
- Passenger Class: classe em que o passageiro estava viajando (First, Second, Third)
- Sex: sexo do passageiro (Male, Female)
- No of Siblings or Spouses on Board: quantidade de irmãos e cônjuge do passageiro junto com ele (a bordo)
- No. of Parents or Children on Board: quantidade de parentes e crianças junto com o passageiro (a bordo)
- Passenger Fare: preço da passagem paga pelo passageiro

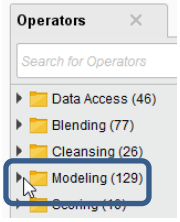
Estendendo o processo

Adicionar operação

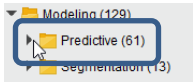
Click no indicador “Design” na parte superior da tela para voltar a exibir o painel “Process”.



Click na seta à esquerda de “Modeling” no painel “Operators” à esquerda para abrir as opções logo abaixo.

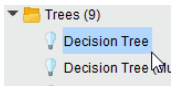


Click na seta à esquerda de “Predictive” para abrir as opções logo abaixo:



Click na seta à esquerda de “Trees” para abrir as opções logo abaixo:

Dê um **duplo-click** sobre a opção “Decision Tree”.

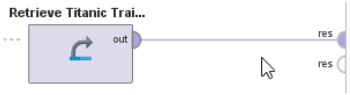


A caixa “Decision Tree” é adicionada no painel “Process”.

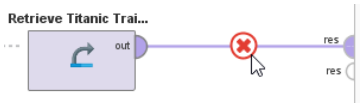
Obs.: outra forma de adicionar a caixa é clicar sobre a opção, arrastar sem soltar o mouse e soltar no local desejado; daqui para a frente só será mencionado “adicionar a opção no painel “Process”.

Excluindo uma ligação

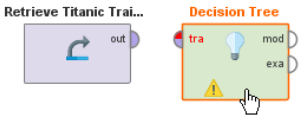
Posicione o mouse sobre a ligação existente saindo do ícone “out” da caixa “Retrieve Titanic Training”:



A ligação será destacada e aparecerá um “X” vermelho sobre ela. Click no “X” para excluir a ligação.

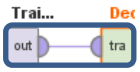


Click sobre a caixa “Decision Tree”, mantenha o mouse pressionado e arraste-a até o lado da caixa “Retrieve Titanic Training”.



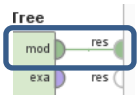
Obs.: esse passo não é necessário e serve somente para melhor visualização do projeto.

Click no ícone “out” da caixa “Retrieve Titanic Training” e em seguida no ícone “tra” na caixa “Decision Tree” para criar a ligação entre as duas caixas:



Obs.: também é possível ligar 2 ícones dando um click no primeiro, manter o mouse pressionado e soltando-o sobre o segundo.

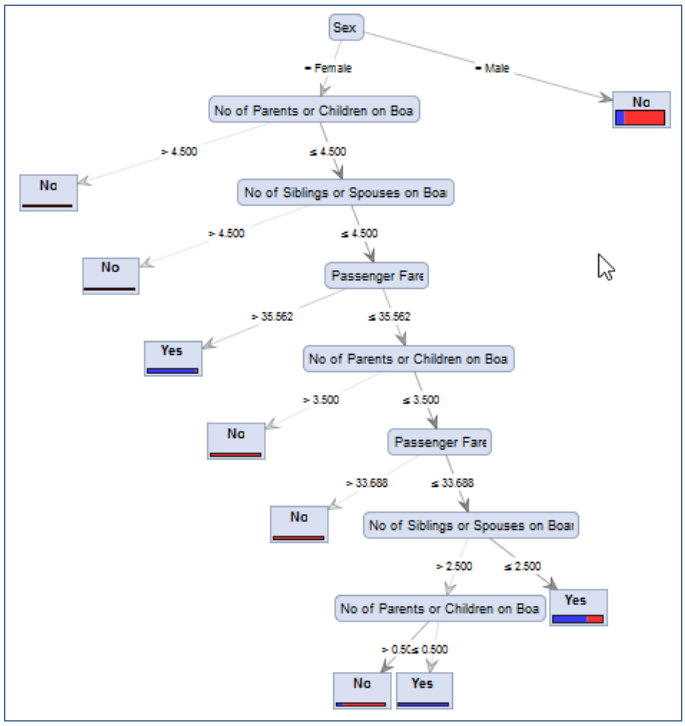
Repita o procedimento para ligar o ícone “mod” da caixa “Decision Tree” ao ícone “res” existente no canto superior direito do painel “Process”:



Exibindo os resultados

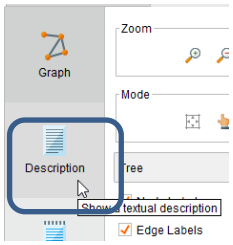
Click no botão “Run” na barra superior:

Será exibida a seguinte árvore na janela “Tree (Decision Tree)”:



Sugestão: click nas caixas “Node Labels” (descrição dos nós, isto é, as caixas) e “Edge Labels” (descrição das ligações entre os nós) para ver os resultados.

Click na opção “Description” no painel à esquerda:



Será exibida a árvore sob a forma textual:

Tree

```
Sex = Female
|   No of Parents or Children on Board > 4.500: No {Yes=0, No=4}
|   No of Parents or Children on Board ≤ 4.500
|   |   No of Siblings or Spouses on Board > 4.500: No {Yes=0, No=2}
|   |   No of Siblings or Spouses on Board ≤ 4.500
|   |   |   Passenger Fare > 35.562: Yes {Yes=101, No=3}
|   |   |   Passenger Fare ≤ 35.562
|   |   |   |   No of Parents or Children on Board > 3.500: No {Yes=0, No=3}
|   |   |   |   No of Parents or Children on Board ≤ 3.500
|   |   |   |   |   Passenger Fare > 33.688: No {Yes=0, No=2}
|   |   |   |   |   Passenger Fare ≤ 33.688
|   |   |   |   |   |   No of Siblings or Spouses on Board > 2.500
|   |   |   |   |   |   |   No of Parents or Children on Board > 0.500: No {Yes=2, No=11}
|   |   |   |   |   |   |   No of Parents or Children on Board ≤ 0.500: Yes {Yes=2, No=0}
|   |   |   |   |   |   No of Siblings or Spouses on Board ≤ 2.500: Yes {Yes=134, No=58}
Sex = Male: No {Yes=110, No=484}
```

Análise

A que conclusão podemos chegar observando a árvore criada a partir dos dados “Titanic Training”?

Pode-se concluir, por exemplo, que sobreviver após o naufrágio não foi uma questão de sorte!

```
Sex = Female
|  No of Parents or Children on Board > 4.500: No {Yes=0, No=4}
|  No of Parents or Children on Board ≤ 4.500
|  |  No of Siblings or Spouses on Board > 4.500: No {Yes=0, No=2}
|  |  No of Siblings or Spouses on Board ≤ 4.500
|  |  |  Passenger Fare > 35.562: Yes {Yes=101, No=3}
|  |  |  Passenger Fare ≤ 35.562
|  |  |  |  No of Parents or Children on Board > 3.500: No {Yes=0, No=3}
|  |  |  |  No of Parents or Children on Board ≤ 3.500
|  |  |  |  |  Passenger Fare > 33.688: No {Yes=0, No=2}
|  |  |  |  |  Passenger Fare ≤ 33.688
|  |  |  |  |  |  No of Siblings or Spouses on Board > 2.500
|  |  |  |  |  |  |  No of Parents or Children on Board > 0.500: No {Yes=2, No=11}
|  |  |  |  |  |  |  No of Parents or Children on Board ≤ 0.500: Yes {Yes=2, No=0}
|  |  |  |  |  |  No of Siblings or Spouses on Board ≤ 2.500: Yes {Yes=134, No=58}
Sex = Male: No {Yes=110, No=484}
```

Um passageiro do sexo feminino, com uma pequena família teve maiores chances de sobreviver, ainda mais se tivesse pago mais caro pela passagem, pois, entre os 916 passageiros analisados, sendo 594 homens e 322 mulheres:

- apenas 110 (18.5%) dos homens sobreviveram
- sobreviveram 239 (74.2%) de mulheres, sendo que não sobreviveram:
    - 4 que tinham mais que 4 (4.5) parentes ou crianças junto.
    - 2 que tinham menos que 5 (4.5) parentes ou crianças junto E mais que 4 (4.5) irmãos ou cônjuge junto
    - 3 que tinham menos que 5 (4.5) parentes ou crianças junto E menos que 5 (4.5) irmãos ou cônjuge junto E pagou mais que 35.562 (em média) pela passagem.
    - 3 que tinham menos que 5 (4.5) parentes ou crianças junto E menos que 5 (4.5) irmãos ou cônjuge junto E pagou menos que 35.562 (em média) pela passagem E tinham mais que 3 (3.5 em média), portanto exatamente 4 (uma vez que tinham menos 5!) parentes ou crianças junto.
    - 2 que tinham menos que 5 (4.5) parentes ou crianças junto E menos que 5 (4.5) irmãos ou cônjuge junto E pagou menos que 35.562 (em média) pela passagem E tinham menos que 4 (3.5 em média) parentes ou crianças junto E pagou mais que 33.688 (em média) – mas menos que 35.562! – pela passagem.
    - 11 que tinham **mais que 2 (2.5 em média)** e menos que 5 (4.5) parentes ou crianças junto E menos que 5 (4.5) irmãos ou cônjuge junto E **pagou menos que 33.688 (em média) pela passagem** E tinham menos que 4 (3.5 em média) parentes ou crianças junto.
    - 58 que tinham **menos que 3 (2.5) parentes ou crianças junto** E menos que 5 (4.5) irmãos ou cônjuge junto E pagou menos que 35.562 (em média) pela passagem E tinham menos que 4 (3.5 em média) parentes ou crianças junto E pagou mais que 33.688 (em média) – mas menos que 35.562! – pela passagem.

Salvando e recuperando um processo

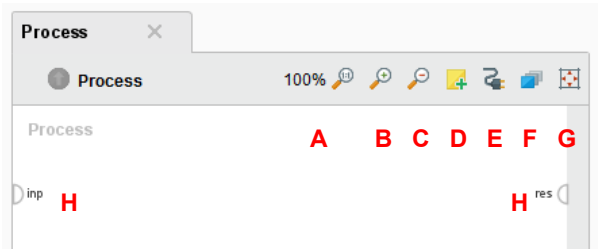
Para salvar um processo, click no botão  ou pressione **Ctrl+S** ou selecione “File” no menu e em seguida “Save Process”. Será exibida a janela “Repository Browser” para seleção do repositório e definição do nome como o processo será gravado. Em seguida click em “OK”.

Para recuperar um processo gravado, click no botão  ou pressione **Ctrl+O** ou selecione “File” no menu e em seguida Open Process”.Será exibida a janela de entrada do rapidminer, onde existe uma lista com os últimos processos gravados. Caso o processo desejado esteja nessa lista, basta clicar sobre ele para abri-lo. Em caso conrário, click no botão “Browse repository” para exibir a janela com todos os processos gravados. Procure-o numa das pastas disponíveis e click sobre ele para abri-lo.

Elementos

Processo

Toda operação no rapidminer é feita num processo no painel “Process”.



Essa janela possui os seguintes elementos:

A. Indicador do zoom atual; click para voltar ao normal (100%)
B. Botão para aumentar zoom
C. Botão para diminuir zoom
D. Botão para adicionar uma caixa de comentário; o mesmo que dar um duplo-click numa área vazia do painel “Process”
E. Botão para tentar conectar automaticamente as portas sem conectores
F. Botão para exibir/ocultar indicador da ordem de execução das caixas do processo
G. ???
H. Portas do processo

Portas

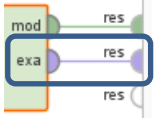
Os operadores (caixas) possuem portas que podem ser conectadas entre si. Mas não podemos conectar quaisquer portas, pois as portas estão associadas a tipos específicos de dados e o dado que sai de uma porta tem que ser compatível com o dado aceito pela porta de entrada do outro operador.

A tabela a seguir contém uma lista de portas mais usadas:

Porta	Significado	Descrição
exa	Example set	Conjunto de dados de exemplo.
For	Formula	Fórmula resultante.
Inp	Input	Porta de entrada; pode receber vários objetos.
Ite	Item sets	Conjunto de itens frequentes.
Lef	Left	Entrada de dados que formará o lado esquerdo de um “join”.
mod	Model	Modelo gerado pelo operador.
Ori	Original	Conjunto de dados originais (sem alteração).
Out	Output	Porta de saída.
Res	Result set	Conjunto de dados de saída.
Rig	Right	Entrada de dados que formará o lado direito de um “join”.
Rul	Rules	Regras de associação geradas a partir de um conjunto de itens frequentes.
Tra	Training	Conjunto de dados de treinamento de um modelo;

Resultados

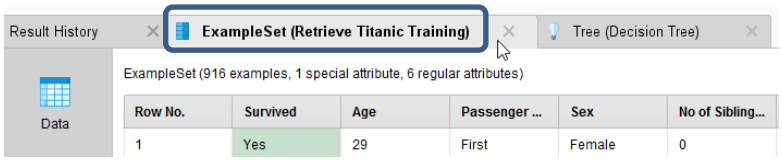
No painel “Process” conecte a saída “exa” da caixa “Decision Tree” à porta “res” livre existente no canto superior direito.



Obs.: uma nova porta “res” é criada no painel “Process” sempre que uma porta “res” é utilizada.

Click no botão “Run” na barra superior:

Desta vez aparecerá a janela “Example Set” (com os dados de entrada da caixa “Decision Tree”) além da janela “Tree (Decision Tree)” que já aparecia antes.

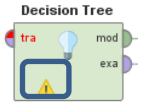


Obs.: a aba “Result History” exibe uma lista de registro histórico dos processos executados.

Erros

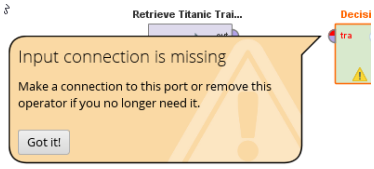
Exclua a ligação entre a porta “out” da caixa “Retrive” e a porta “tra” da caixa “Decision Tree”.

Um triângulo amarelo aparecerá na caixa “Decision Tree”, informando que existe um problema.



Click no botão “Run” na barra superior:

Aparecerá uma caixa de mensagem informando a ocorrência de um problema.



Observe a mensagem e então click no botão “Got it!”. Refaça a conexão para restasbelecer a condição normal do processo.

Janelas

O **rapidminer** trabalha painéis que podem ser exibidos numa janela ou numa aba de janela.

As janelas são organizadas em 2 visões: “Design” e “Results”. Para exibir a visão “Design” selecione “View” no menu, em seguida “Views” e “Design” ou pressione a tecla **F8** ou click no botão “Design” na parte central da tela, no alto (ao lado de “Views:”). Para exibir a visão “Results” selecione “View” no menu, em seguida “Views” e “Results” ou pressione a tecla **F9** ou click no botão “Results” na parte central da tela, no alto (ao lado de “Views:”)

A tela inicial exibe a visão “Design” com os painéis “Repository” no alto à esquerda, “Operators” abaixo à esquerda, “Process” no centro, “Parameters” no alto à direita e “Help” abaixo à direita, cada uma numa janela distinta.

Selecione “View” no menu e depois deixe o mouse sobre “Show Panel” para ver os painéis disponíveis. Se desejar exibir um painel, basta clicar sobre ele na lista sendo exibida.

Para adicionar um painel como aba junto com outro painel na mesma janela, basta criar sobre o título do painel e arrastá-lo sobre o outro painel. Ambos virarão abas da mesma janela. Ao arrastar uma janela, ela pode ser deixada entre duas outras janelas (ambas serão redimensionadas automaticamente) ou simplesmente deixada “flutuando” sobre as demais.

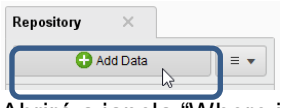
Para ocultar um painel, basta clicar sobre o “X” exibido ao lado do título do painel.

É possível redimensionar os painéis, bastando posicionar o mouse no seu limiar com outro painel (o cursor muda de formato para uma seta bidirecional) e arrastar até o tamanho desejado.

Selecione “View” no menu e em seguida em “Restore Default Views” para voltar a exibir as janelas como são exibidas como padrão pelo **rapidminer**.

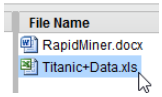
Importando dados

Click na opção “Add Data” no painel “Repository” à esquerda.

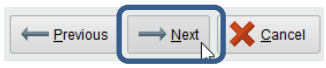


Abrirá a janela “Where is your data?”. Click no botão “My Computer”.

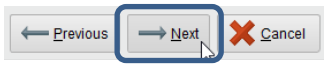
Abrirá a janela “Select the data location”. Selecione a pasta (diretório) onde se localiza o arquivo “Titanic+Data.xls”



Click no botão “Next” na parte inferior da janela.



Aparecerá a janela “Select the cells to import” com os dados existentes no arquivo, que podem então ser visualizados. Click em “Next” para o próximo passo, pois não será necessário alterar nenhum parâmetro nessa tela.

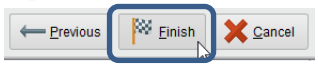


Obs.: a opção “Sheet” permite selecionar a aba, caso exista mais de uma aba na planilha (arquivo); a opção “Cell Range” permite definir o intervalo dos dados que será lido – clique no botão “Select All” para selecionar todas as células (default); marque a opção “Define header row” caso exista uma ou mais linhas de cabeçalho dos dados (que serão ignoradas como dados); defina na opção “Define header row” a quantidade de linhas que serão tratadas como cabeçalho dos dados (normalmente = 1).

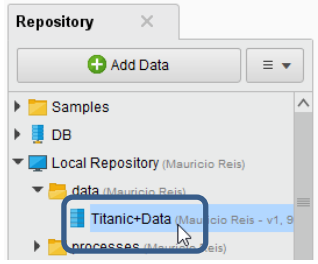
Será exibida a janela “Format your columns”.

O **rapidminer** indica se existe algum problema com os dados. Observe a mensagem “no problems.” no canto inferior da janela. Click em “Next” para o próximo passo, pois não será necessário alterar nenhum parâmetro nessa tela.

Será exibida a janela “Where to tore the data?” onde deve ser definido o local interno onde o **rapidminer** armezenará o arquivo. Selecione “Data”. Click em “Finish” para encerrar a importação e voltar à tela principal do **rapidminer**.

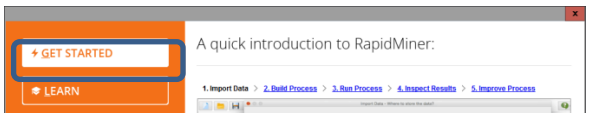


Observe que o arquivo “Titanic+Data.xls” aparece no painel Repository” abaixo de “Local Repository / Data”,

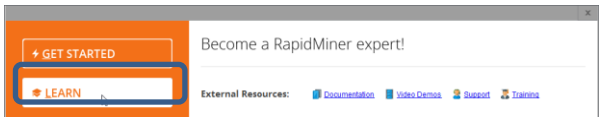


Tutoriais

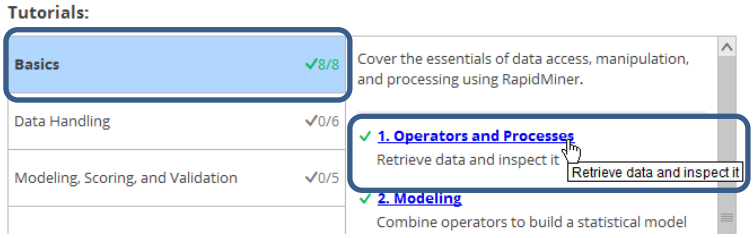
O **rapidminer** possui vários tutoriais (vídeos em inglês) que podem ser acessados clicando no botão “Get Started” exibido na tela de entrada:



Click no botão “Learn” para exibir uma tela onde são apresentados vários tutoriais passo-a-passo.



Selecione o tipo de tutorial e click nas opções mais à direita.



Esses tutoriais constam de exemplos práticos passo-a-passo que são feitos na própria ferramenta.

Essa tela pode ser exibida ou a qualquer momento pelo menu `File/New Process` ou pressionando **Ctrl-N**.

Para um tutorial em português veja o link `https://prezi.com/-yo8qjamdbbq/rapidminer-aprenda-a-usar`.