**APRENDIZADO DE MÁQUINA**

***Curso de especialização em Bancos de Dados***

***Curso de especialização em Business Intelligence***

***Curso de especialização em Big Data***

**Prof. Hugo de Paula**

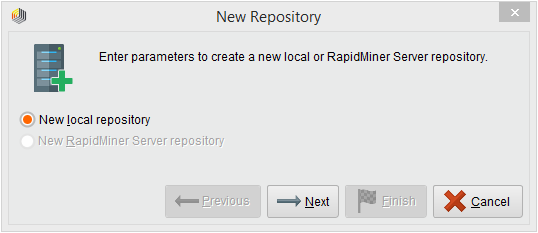
**ATIVIDADE EM LABORATÓRIO 1**

**PREPARAÇÃO DE DADOS**

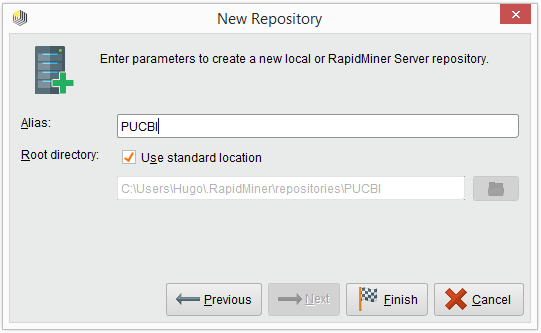
*Adaptado de North, Matthew A. Data Mining for the Masses. 2012.*

# PREPARAÇÃO DO *RAPIDMINER*

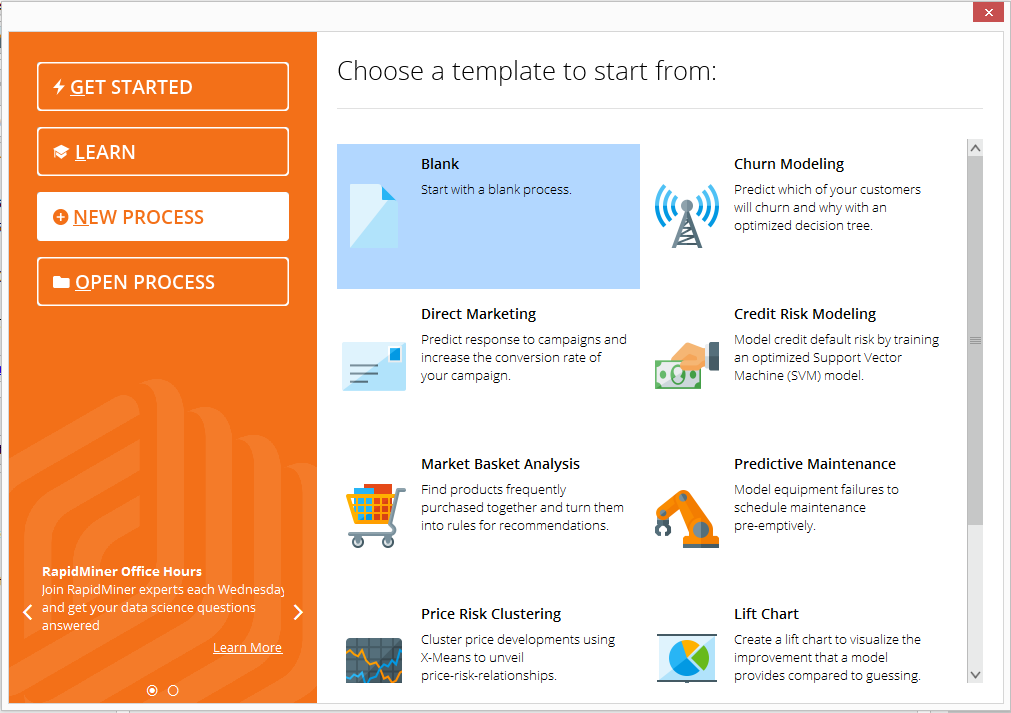
1. Abra o *RapidMiner* clicando no ícone da aplicação (rapidminer_frame_icon_24). A primeira vez que você abrir o *RapidMiner* ele irá solicitar a criação de um repositório inicial.
2. Vamos criar um repositório local.



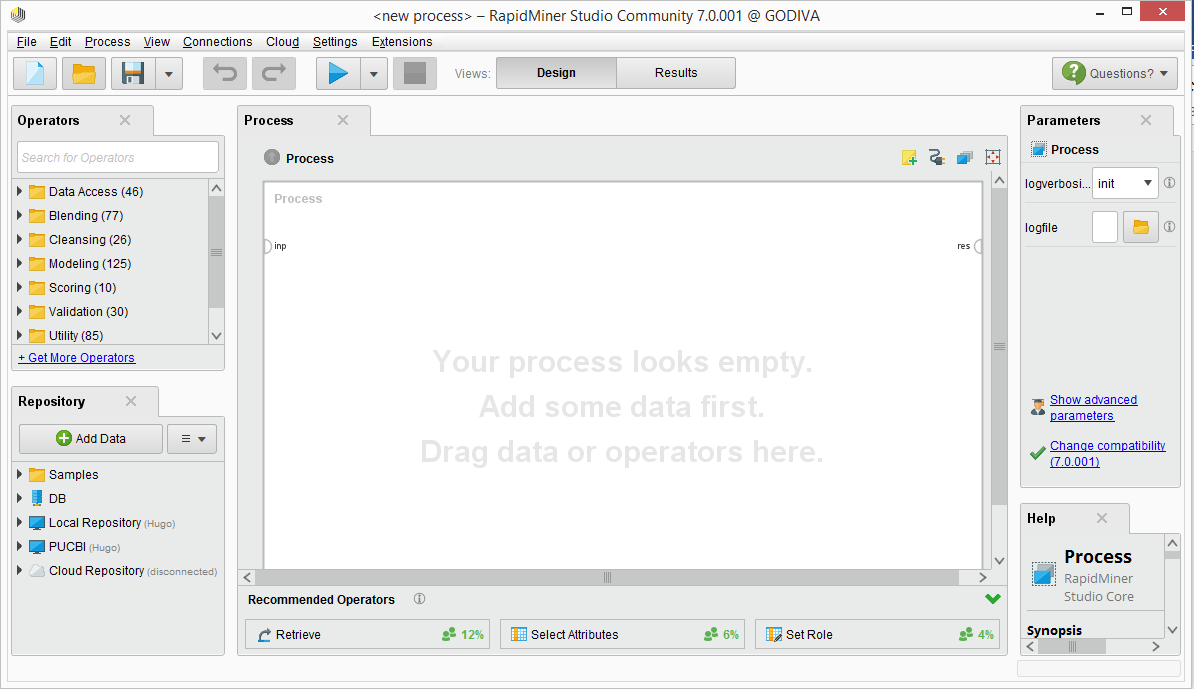
1. Vamos criar o repositório local PUCBI. Preste atenção à pasta de armazenamento.



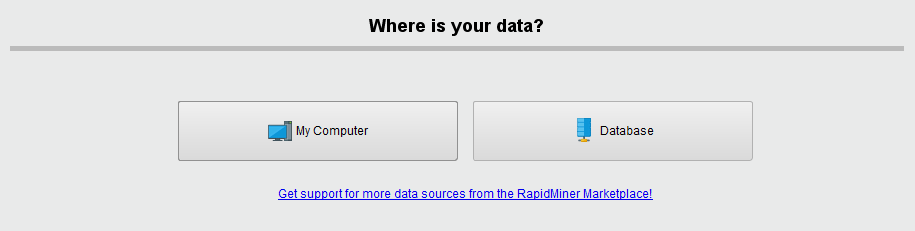
1. Uma vez que o repositório foi criado, vamos criar um novo processo de mineração de dados clicando no ícone ***New Process***. Selecione um projeto em branco (**Blank**).



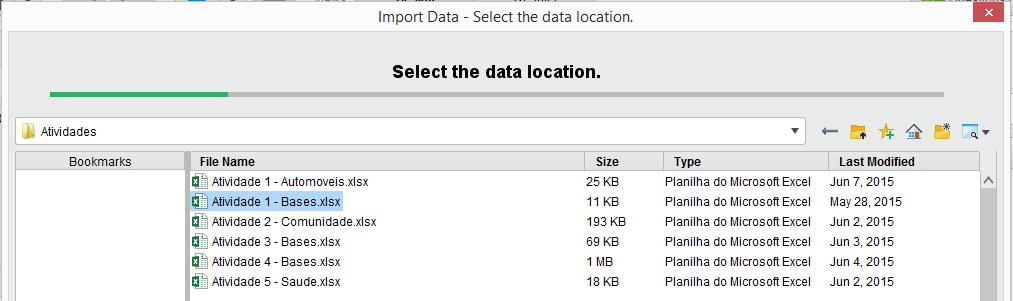
1. Deverá abrir uma janela com a perspectiva default do processo. À esquerda podem-se ver duas abas: ***Repositories*** e ***Operators***. A aba ***Repositories*** é o local onde você irá conectar suas bases de dados (*data* sets) que você deseja minerar. A área ***Operators*** é onde todas as ferramentas de mineração de dados estão localizadas. Os operadores são usados para manipular data sets e criar modelos. Na área ***Repositories*** será possível visualizar o repositório local **PUCBI** que nós criamos.



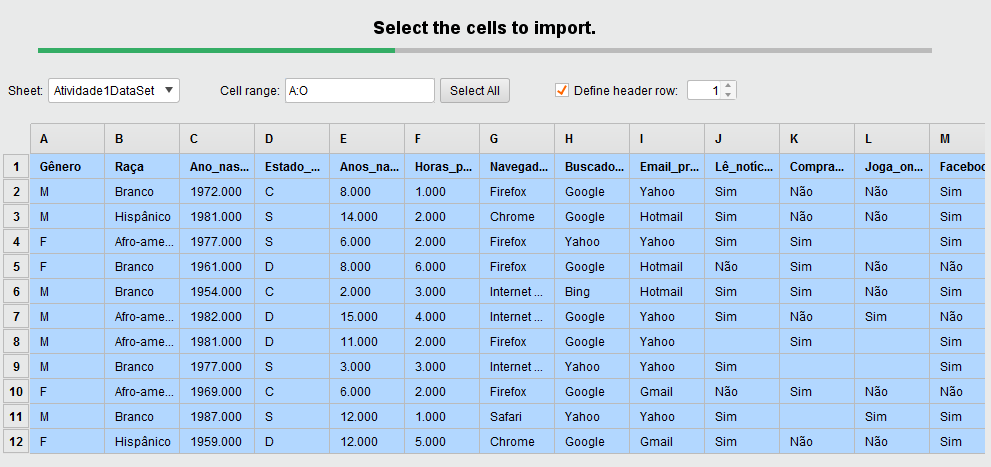
1. Vamos conectar a base de dados disponível no arquivo **Atividade 1 - Bases.xlsx**. Clique no botão ***Add Data*** (). Na pergunta “onde está o seu dado?” (*Where is your data?*) , selecione a opção ***My Computer***.



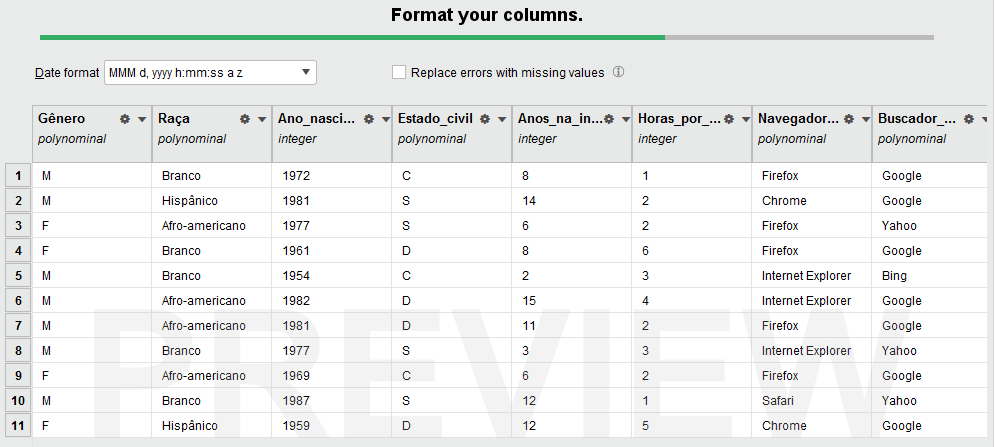
1. Navegue até a pasta onde está o arquivo.



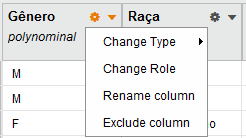
1. O ***RapidMiner*** solicita que seja selecionado a área dos dados (***Cell range***), e deixa marcado, por padrão, que a primeira linha possui os rótulos das colunas (***Define header row***). Se isso não aconteceu, é necessário corrigir.



1. Em seguida, o ***RapidMiner*** tenta adivinhar o tipo de dado de cada atributo. Os tipos de dados podem ser numéricos, textuais, datas, nominais, binomiais, polinomiais, etc.

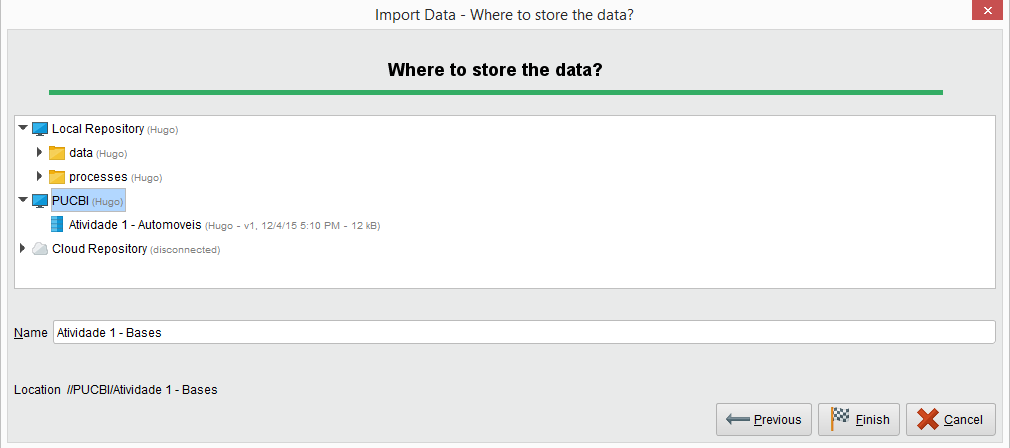


1. Cada atributo possui um papel (***Role***) no processo. Por padrão, as colunas são importadas com o papel em branco. Para alterar o papel, basta clicar em  e em ***Change Role***.

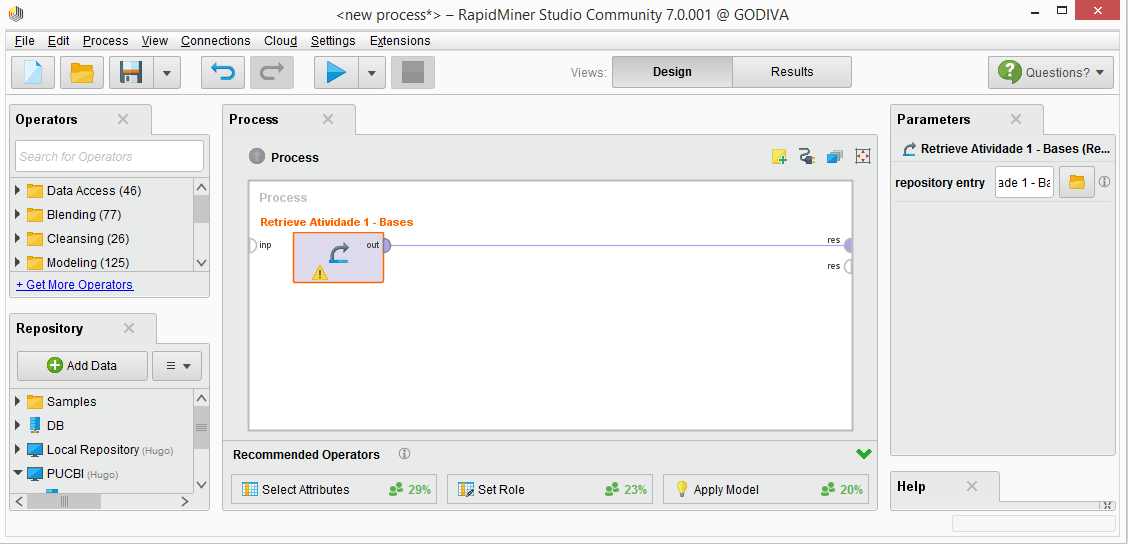


Como podemos definir esse papel posteriormente na tela principal do processo, vamos deixar como está. Os papeis podem ser rótulos, outliers, pesos, identificadores, cluster, etc. Nesta mesma opção podemos excluir um atributo. Os atributos também podem ser excluídos do modelo posteriormente. Vamos aceitar os valores default.

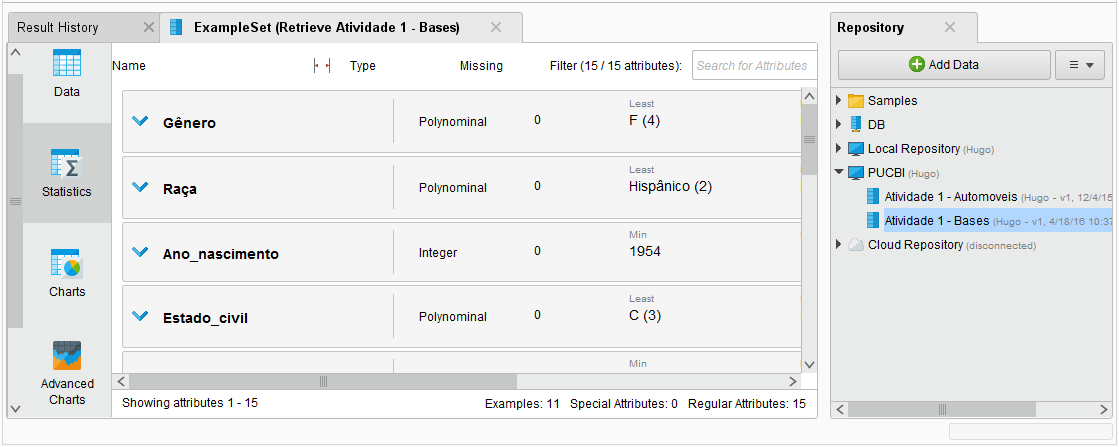
1. Finalmente, vamos selecionar o repositório de destino e definir o nome da base de dados como sendo Atividade1



1. O ***data set*** agora está disponível na aba ***Repositories***. Para usarmos, basta arrastá-lo para a janela principal do processo. Cada retângulo do ***RapidMiner*** é um ***operator***. O operador ***Retrieve*** carrega um ***data set*** e o disponibiliza para uso num processo de mineração. Os semicírculos que aparecem nos operadores são portas de entrada (***input ports***) ou portas de saída (***output ports***). Para criarmos uma conexão entre portas, basta clicar na porta de saída e arrastá-la até a porta de entrada. Na figura abaixo, nós conectamos a porta ***out*** na porta ***res***, que irá exibir o dado na ***Results View***. A ***Design View*** exibe a estrutura do processo, enquanto a ***Results View*** exibe os resultados da execução do processo de mineração. Para executar um processo de mineração, basta clicar no ícone ***play*** ().

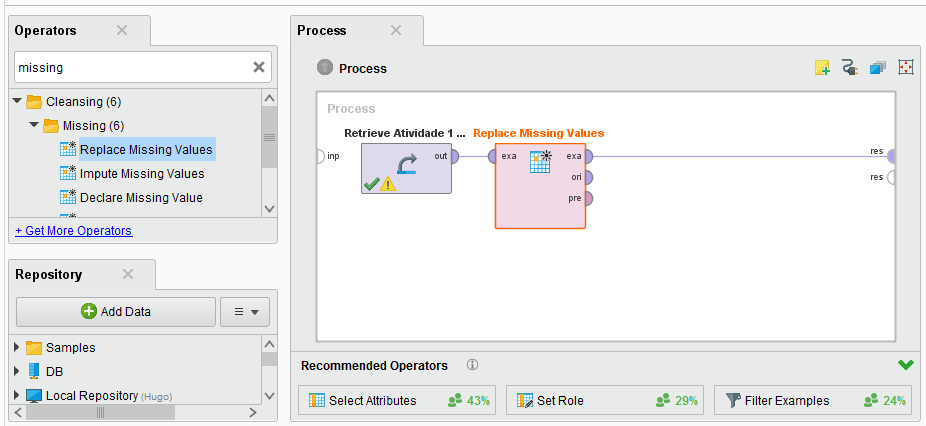


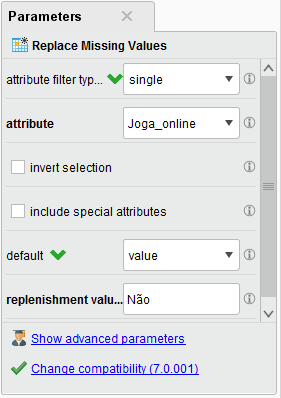
1. Na ***Results View*** é possível gerar diversas visualizações do ***example set***. Em ***Statistics*** são exibidas estatísticas básicas dos dados. No ***Charts*** e ***Advanced Charts*** é possível construir inúmeros tipos de gráficos com os atributos.

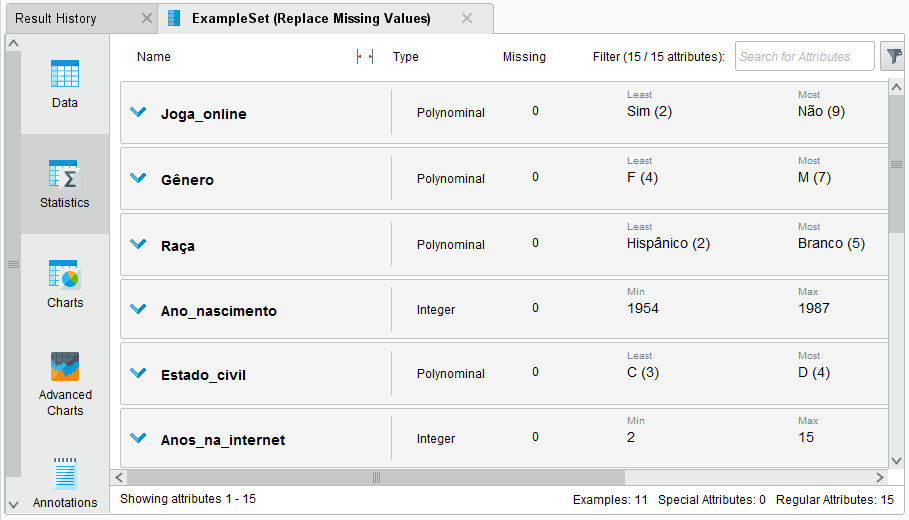


# PREPARAÇÃO DE DADOS: TRANSFORMAÇÃO DE DADOS

1. Nesta prática iremos aprender a tratar dados omissos. Dados omissos não são valores zero. Eles são dados faltantes. Dados omissos podem representar um valor nulo ou um problema.
   1. Por exemplo, em ***Statistics*** observa-se que o ***Outras\_redes\_sociais*** possui 7 ***Missing*** (omissos). Isso significa que o usuário não utiliza outra rede social, e o valor é válido.
   2. Por outro lado, no campo ***Joga\_online***, cuja resposta deve ser ***sim/não***, o valor omisso não tem significado.
2. Pela ***Statistics***, podemos ver que os campos ***Lê\_notícias***, ***Compra\_online***, ***Joga\_online***, ***Outras\_redes\_sociais*** possuem valores omissos.
3. Vamos supor, nessa atividade, que os valores omissos no caso de ***Joga\_online***, significam na verdade ***não***. Vamos voltar para a ***Design View***. Na aba ***Operators***, digite ***missing*** no campo ***Search***. Note que aparecerá ***Cleansing → Missing → Replace Missing Values***. Vamos adicionar esse operador ao nosso processo, conectando a saída da nossa base de dados na entrada do operador, e a saída do operador na saída do processo (isso pode ser feito manualmente, ou arrastando e soltando o operador sobre a linha já existente).

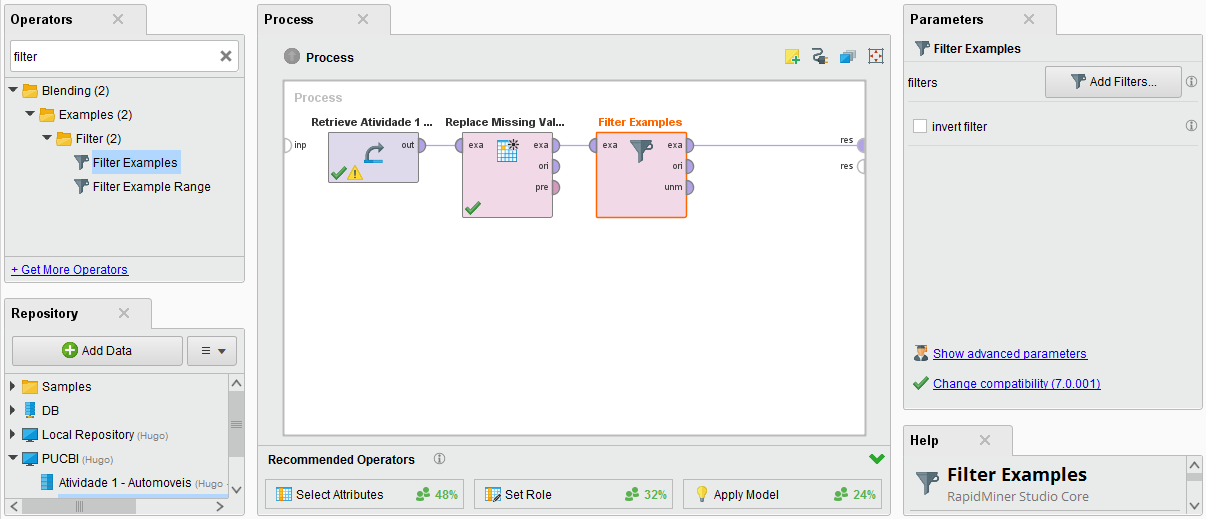


1. Na aba ***Parameters***, à direita, poderemos configurar o operador. Vamos mudar os valores omissos de ***Joga\_online*** para ***não***. Devemos definir 4 parâmetros:   
   ***atribute filter type:*** ***single*** – desejamos alterar um só atributo.  
   ***atribute: Joga\_online*** – seleciona o atributo a ser manipulado.  
   ***default: value*** – define que a transformação será feita digitando um valor. Podemos usar diversas funções de transformação como média, moda (valor mais frequente, entre outros).   
   ***replenishment value: Não*** – define que o valor a ser inserido nos campos omissos será o valor ***Não***.
2. Abaixo da aba ***Parameters*** existe uma aba ***Help*** que é um ótimo guia de referência rápida aos parâmetros do operador selecionado e seus possíveis valores.
3. Ao executarmos o processo, iremos obter o resultado a seguir. Perceba que não há mais valores omissos em ***Joga\_online***, e que agora temos 9 valores ***Não*** e 2 valores ***Sim***.

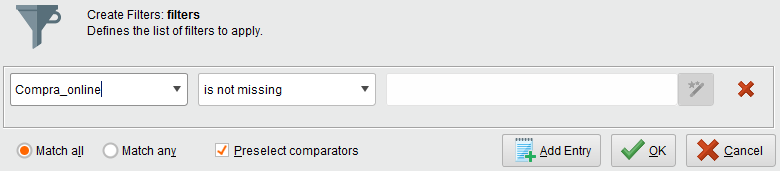


# PREPARAÇÃO DE DADOS: REDUÇÃO DE DADOS

1. Suponha agora que queremos eliminar registros com valores ***nulos*** no campo ***Compra\_online***. Digite a palavra ***filter*** no campo ***Search*** da aba operadores.

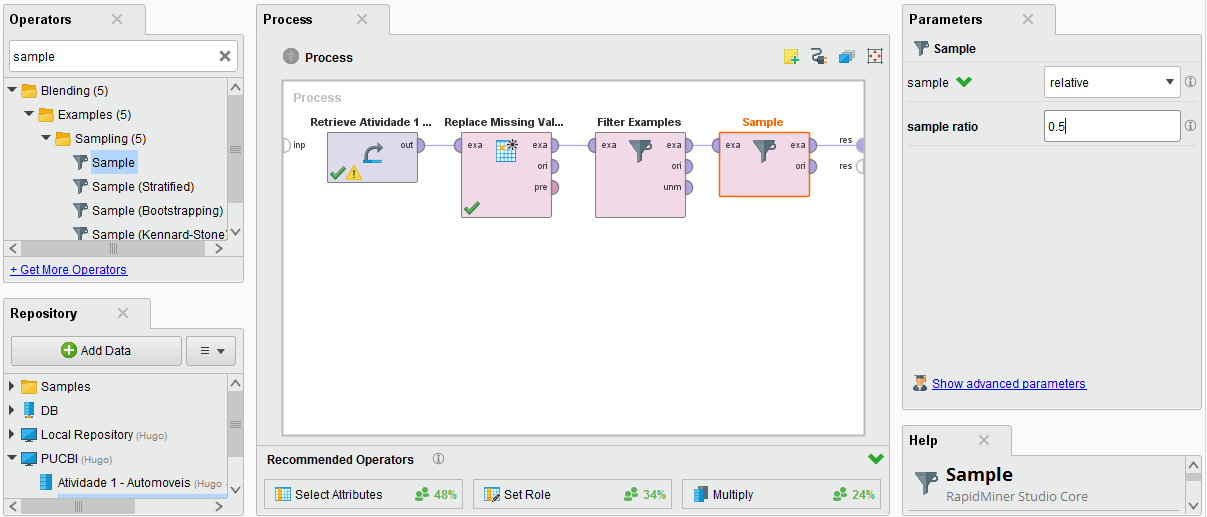


1. Então selecione o operador ***Blending → Examples → Filter Examples***. Selecione ***Add Filters...*** e configure o seguinte filtro.



# PREPARAÇÃO DE DADOS: AMOSTRAR DADOS

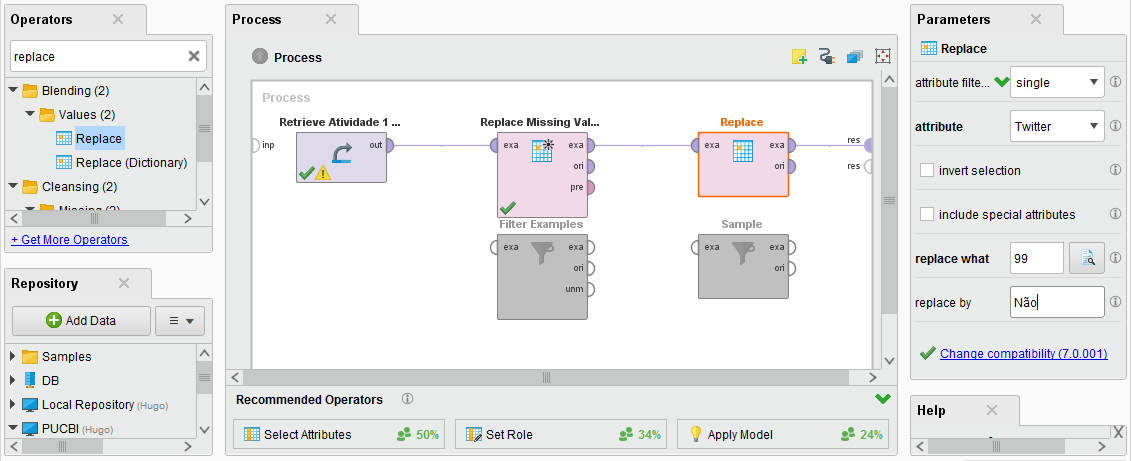
1. Ao eliminarmos registros com atributos omissos em atributos chave, nós reduzimos a nossa base de dados, tornando o processamento mais rápido e eficiente, sobre dados mais confiáveis. Outra forma de reduzir o tempo de processamento é utilizando a amostragem. Amostragem é especialmente útil quando desejamos apenas testar algumas alternativas de mineração de dados sobre um conjunto de dados menor, antes de escolher o algoritmo que será aplicado a toda a base de dados. Vamos digitar no campo ***Search*** a palavra ***Sample*** (amostrar) e selecionamos o operador ***Blending → Examples → Sample***.



1. Configure o parâmetro ***sample*** para ***relative***, e o parâmetro ***sample ratio*** para ***0.5***, significando que você deseja manter apenas 50% da base de dados.
2. Ao executar o processo você verá que sua base foi reduzida a apenas 4 registros.

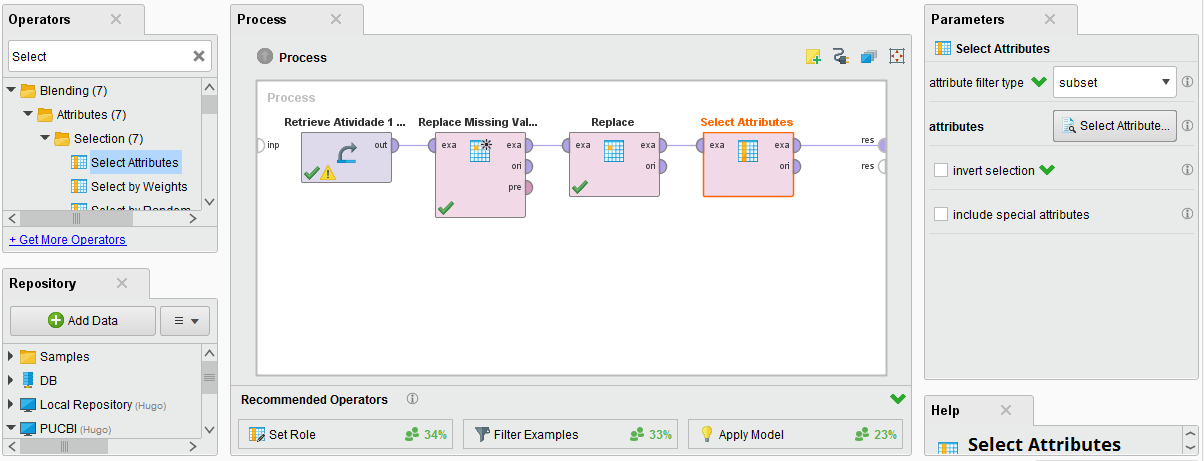
# PREPARAÇÃO DE DADOS: DADOS INCONSISTENTES

1. Dados inconsistentes são diferentes de dados omissos. Eles acontecem quando um valor existe na base de dados, mas não tem significado ou não é válido. Veja o parâmetro ***Twitter***, por exemplo. Ele é nitidamente um tributo ***Binominal***, mas se tornou ***Polynominal*** devido a um registro com valor igual a ***99***.
2. Localize o operador ***Blending → Values → Replace*** na aba ***Operators***. Nesse exemplo, eu desabilitei os operadores ***Filter Examples*** e ***Sample***, para voltar a ter a base completa. Como o valor ***Não*** é a maioria da base de dados (80%), então acredito que seria seguro substituir o valor ***99.0*** pelo valor ***Não***.

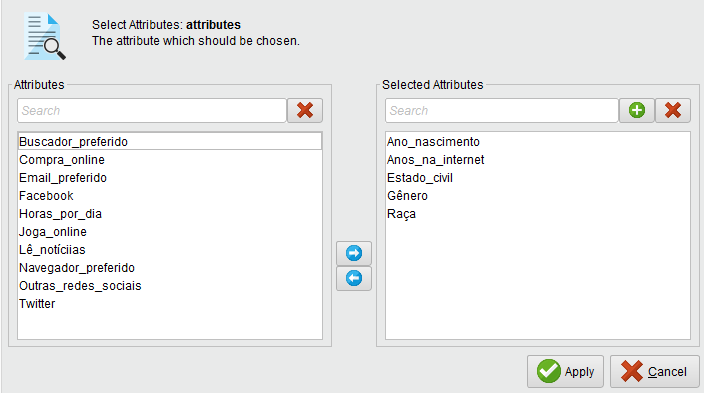


# PREPARAÇÃO DE DADOS: REDUÇÃO DE ATRIBUTOS

1. A redução de atributos é útil quando desejamos remover um atributo irrelevante para responder uma pergunta específica sem removê-lo da base de dados. As boas práticas de mineração de dados sugerem que todos os dados sejam importados, e os atributos sejam removidos na medida em que eles vão se mostrando desinteressantes ou redundantes.
2. Vamos localizar o operador ***Blending →Attributes →Selection →Select Attributes***. Iremos definir o parâmetro ***Attribute filter type*** como sendo ***subset***. Isso nos permitirá escolher quais campos remover.



1. Suponha que que iremos estudar a demografia dos usuários de internet. Por isso, desejamos selecionar os atributos: ***Ano\_nascimento, Gênero, Estado\_civil, Raça*** e ***Anos\_na\_internet***.

****

# EXERCÍCIO

1. Carregue a base de dados ***Atividade 1 - Automoveis.xlsx***, e realize a preparação dos dados para que eles sejam consistentes. Você irá encontrar:
   1. Valores omissos que poderão ser registros ou atributos.
   2. Valores omissos que poderão ser estimados.
   3. Valores inconsistentes.
   4. Problemas de formatação de dados.

Você poderá corrigir valores omissos e inconsistentes da forma que julgar procedente. Elimine campos caso ache que algum campo é redundante ou desinteressante. Explique sua estratégia. Quando é melhor substituir valores? Quando é melhor remover o registro? Quando podemos estimar o valor pelos demais?

1. Mostre os campos inconsistentes. Mostre como você detectou as inconsistências (na análise exploratória dos dados).
2. Devido a problemas de formatação entre a língua inglesa e portuguesa, alguns campos numéricos foram erroneamente detectados como campos polinomiais. Aprenda a utilizar o operador ***Parse Numbers*** para resolver este problema.
3. Escreva um documento mostrando o processo realizado, explicando os operadores e seus parâmetros, e mostrando a tabela resultante.
   1. Tente se limitar a uma ou duas páginas.
   2. Utilize imagens da tela do ***RapidMiner*** para mostrar os operadores, os dados antes e após o processamento.
   3. Este documento não precisa de capa e nem ser formatado em nenhum formato específico de trabalho acadêmico (estilo ABNT).
   4. Basta ter o título **ATIVIDADE 1**, e conter o nome do autor do trabalho.