**APRENDIZADO DE MÁQUINA**

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

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

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

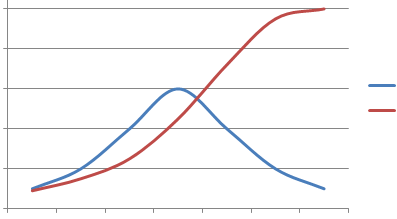
**Prof. Hugo de Paula**

**ATIVIDADE EM LABORATÓRIO 3**

**ÁRVORES DE DECISÃO**

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

# OBJETIVOS

1. O objetivo dessa atividade é classificar pessoas pela sua propensão a comprar um novo leitor digital (eReader), de modo a melhorar a efetividade de campanhas de marketing.
2. Para alcançar este objetivo, iremos utilizar o modelo de árvores de decisão para minerar hábitos de consumo de usuários de um site de e-commerce. O sociólogo Everett Rogers publicou, na década de 60, um trabalho que mostrava a adoção de novas tecnologias por consumidores. Ele identificou os seguintes grupos de consumidores: *inovators, early adopters, early majority e late majority*. Os dois primeiros grupos compreendem os usuários mais propensos a comprar uma nova tecnologia, enquanto os dois últimos eventualmente poderão comprar a tecnologia, se achar conveniente.

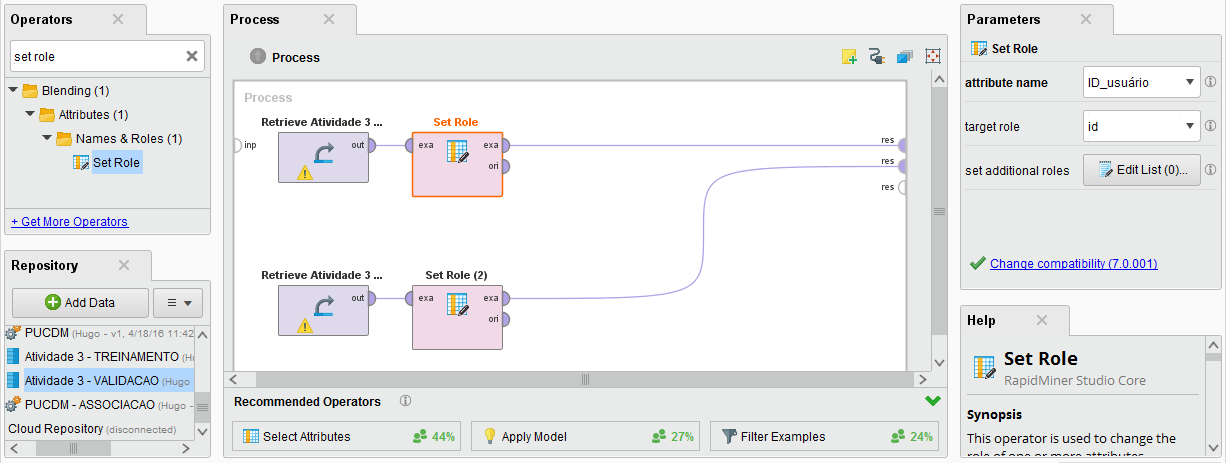
Número de *adopters* por grupo

Número cumulativo de *adopters* ao longo do tempo

1. Resumo dos dados:
   1. ***ID\_usuário***: um identificador numérico associado a cada pessoa que possui uma conte no site.
   2. ***Gênero***: M (masculino) ou F (feminino).
   3. ***Idade***: numérico.
   4. ***Estado\_civil***: C – casado, S – não casados (solteiros, viúvos, divorciados, etc.).
   5. ***Atividade\_no\_site***: Esporádico, Intermitente, Frequente.
   6. ***Pesquisou\_eletronicos\_12m***: Sim/Não indicando se é o usuário andou pesquisando eletrônicos no site nos últimos 12 meses.
   7. ***Comprou\_eletronicos\_12m***: Sim/Não indicando se é o usuário comprou eletrônicos no site nos últimos 12 meses.
   8. ***Pesquisou\_mídia\_digital\_18m***: Sim/Não indicando se é o usuário andou pesquisando mídia digital (como mp3) no site nos últimos 18 meses.
   9. ***Comprou\_mídia\_digital\_18m***: Sim/Não indicando se é o usuário comprou mídia digital (como mp3) no site nos últimos 18 meses.
   10. ***Forma\_pagamento***: Transferência; Website; Cartão e Boleto bancário.
   * ***Adoção\_eReader***: Esse atributo será o rótulo classificador (***label***)
     + ***Inovator***: quem comprou eReaders de geração anteriores até 1 semana após o lançamento.
     + ***Early Adopter***: quem comprou eReaders de geração anteriores entre 1 e 3 semanas após o lançamento.
     + ***Early Majority***: quem comprou eReaders de geração anteriores entre 3 semanas e 2 meses após o lançamento.
     + ***Late Majority***: quem comprou eReaders de geração anteriores após 2 meses do lançamento.

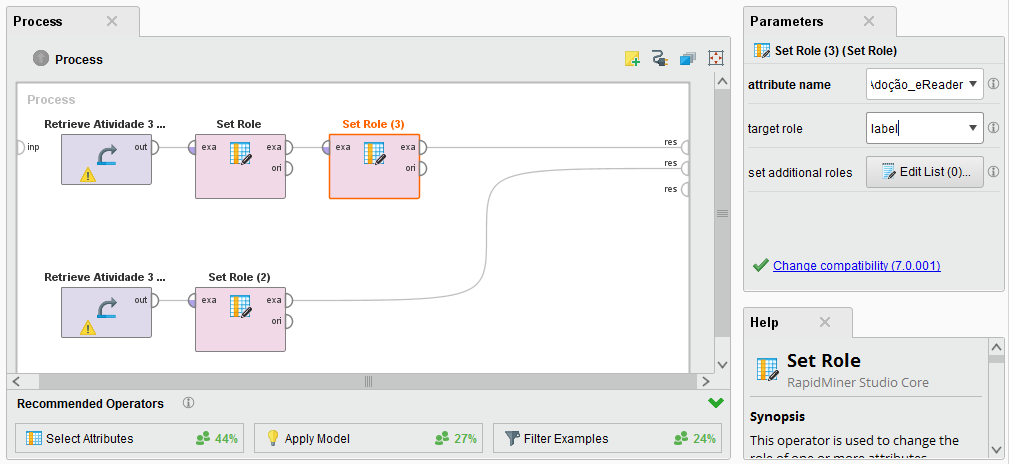
# PREPARAÇÃO DOS DADOS

1. Serão usadas duas bases de dados disponíveis em duas planilhas do arquivo ***Atividade 3 - Bases.xlsx***. A primeira planilha se chama ***TREINAMENTO*** e contém os registros já classificados. A segunda planilha se chama ***VALIDAÇÃO***, e contém novos registros sem classes definidas para podermos avaliar o modelo. Você deve importar as duas bases de uma vez.
2. Arraste as duas bases para a área principal, e modifique o nome para ***TREINAMENTO*** e ***VALIDAÇÃO***. Analise os dados no ***Statistics***. Você perceberá que, aparentemente, não há valores omissos nem inconsistentes, mas ainda temos que preparar os dados.
3. O campo ***ID\_usuário*** não tem relação com os consumidores, propriamente dito. Ao invés de removermos o atributo, vamos usar um outro operador para alterar o seu papel no modelo. Procure o operador ***Blending → Attributes → Names & Roles → Set Role***. Adicione o operador duas vezes, uma para cada base. Na aba ***Parameters*** modifique o papel de ***ID\_usuário*** para ***id***.



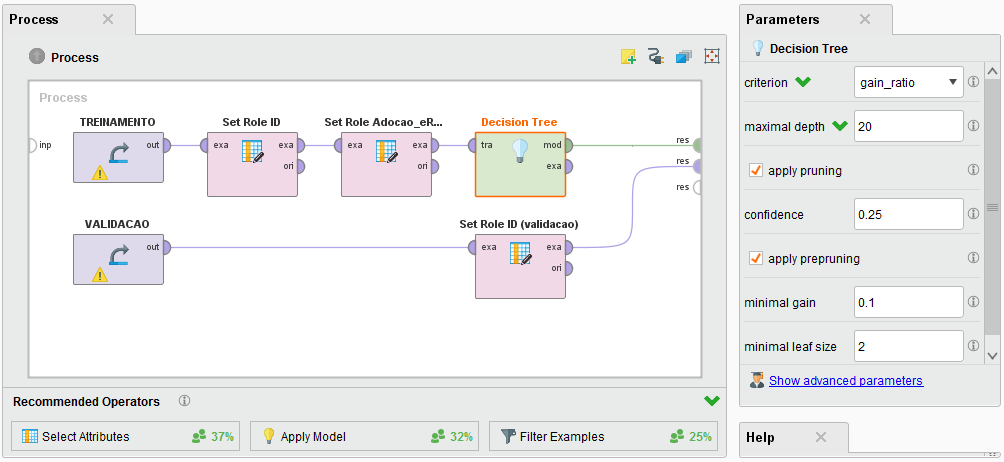
* 1. A vantagem de utilizar ***Set Role*** ao invés de ***Remove Attribute*** é que fica mais fácil identificar o usuário após a classificação.

1. A classe destino deve estar em um atributo do tipo rótulo (***label***). Vamos usar novamente o operador ***Set Role*** para definir o campo ***Adoção\_eReader*** como do tipo ***label***.



# MODELAGEM

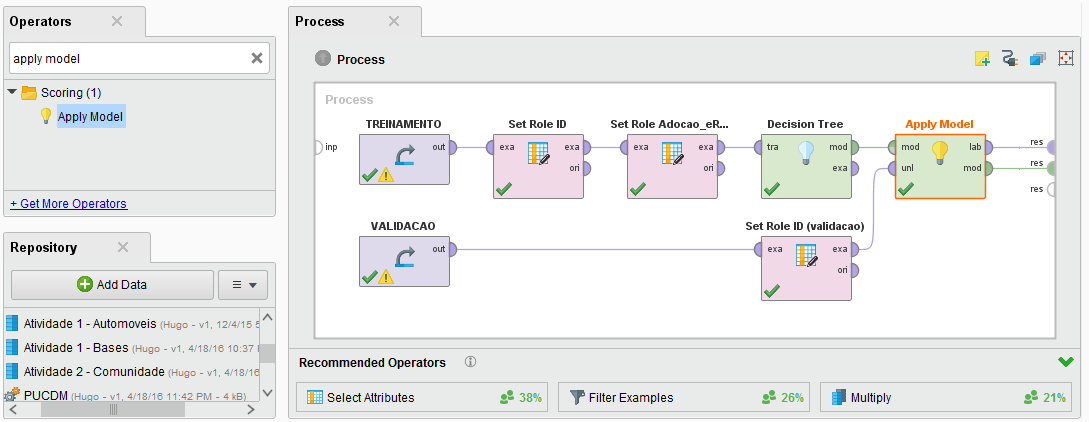
1. Vamos construir a árvore de decisão para os dados de treinamento. Utilize o operador ***Modelling → Predictive → Trees → Decision Tree***. Verifique a aba ***Parameters***. O critério de particionamento está definido como a taxa de ganho, e existem algumas limitações quanto ao menor tamanho de um nó.
2. Desabilite as opções ***apply prunning*** e ***apply preprunning***.



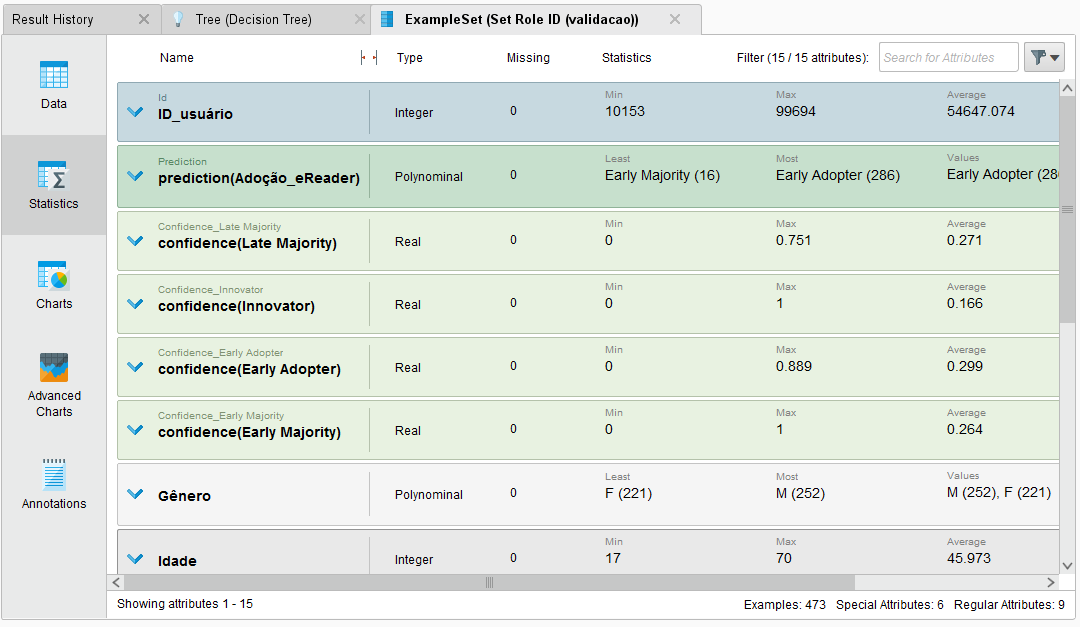
1. Execute o processo e selecione a aba ***Tree (Decision Tree)***, na perspectiva de resultados. Os nós ovais em cinza representam os atributos. Os ramos representam os valores do atributo pai e os nós-folha quadrados representam as classes. Pode-se utilizar o operador ***Picking Mode*** () para personalizar a exibição da árvore.

# APLICANDO O MODELO

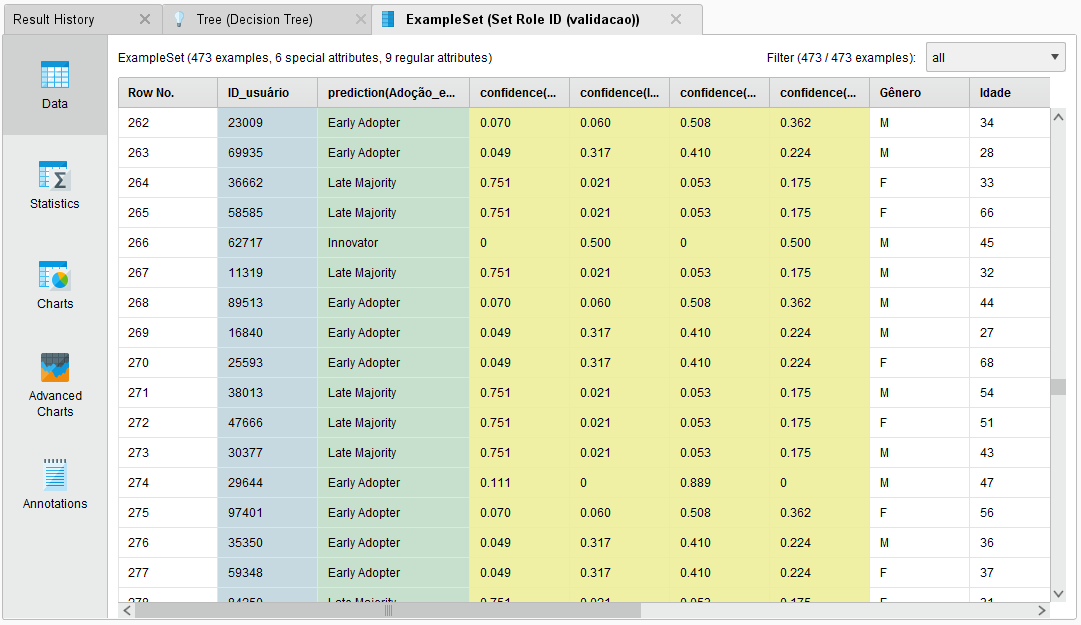
1. Vamos testar o modelo. Selecione o operador ***Scoring → Apply Model***. Conecte o modelo gerado pelo operador ***Decision Tree*** na entrada ***mod*** do operador ***Apply Model*** e, então, conecte a base de validação na entrada ***unl*** (***unlabelled data***).



1. Execute o processo. Você verá a base de validação com cinco novos campos (em rosa): ***confidence\_Late Majority, confidence\_Innovator, confidence\_Early Adopter, confidence\_Early Majority, prediction***.



1. O campo ***prediction*** define a classe em que o registro foi alocado. Os campos de confiança mostram o grau de certeza da classificação.



# EXERCÍCIO

1. Lembrando-se que o processo de *Data Science* sugere que **o aprendizado de máquina é um processo iterativo**, volte ao modelo e **altere os valores dos diversos parâmetros da árvore**.
   1. Altere o critério de seleção de atributos para ***gini\_index*** e ***information gain*** e veja o resultado, mantendo a pode desabilitada.
   2. Para evitar o *overfitting*, habilite as opções ***apply prunning*** e ***apply preprunning***. O operador ***apply prunning*** irá eliminar galhos da árvore que não produzirem uma confiança mínima. O operador ***apply preprunning*** irá interromper a geração de nós caso não produza um ganho mínimo de informação, ou o nó não represente uma quantidade mínima de registros. Altere valores e simplifique a árvore de decisão para uma complexidade que julgar aceitável.
   3. Baseado nesses valores, tente interpretar os resultados. Por exemplo, tente entender qual o potencial de alcance daquela regra na população. Quantos registros ela está representando. Analise o nó folha e identifique se as “confusões” geradas por cada regra podem ajudar a diferenciar uma regra boa de uma regra ruim.
2. Escreva um documento mostrando o processo realizado, explicando os operadores e seus parâmetros, e mostrando as regras resultantes.
   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 3**, e conter o nome do autor do trabalho.