

**Universidade do Minho**  
Escola de Engenharia

# D escoberta do C onhecimento

José Machado  
Cristiana Neto



# ÁRVORES DE DECISÃO COM O RAPIDMINER

# CONTEXTO E PRESPECTIVA



O Ricardo trabalha para uma grande loja retalhista *online*. A sua empresa vai lançar um *eReader* da próxima geração em breve e eles querem maximizar a eficácia do seu *marketing*.

O Ricardo percebeu que algumas pessoas estavam mais ansiosas para adquirir o dispositivo da geração anterior, enquanto outras pareciam contentar-se em esperar para comprar o dispositivo eletrónico mais tarde.

Assim, ele questiona-se sobre o que faz algumas pessoas estarem motivadas a comprar o produto assim que ele sai, enquanto outras estão menos motivadas a adquirir o produto.

# CONTEXTO E PRESPECTIVA



A empresa onde o Ricardo trabalha também vende outros produtos, como livros (em papel e digitais), música e produtos eletrônicos de vários tipos. O Ricardo acredita que, ao extrair os dados dos clientes sobre o comportamento geral do consumidor no *site*, poderá descobrir quais os clientes que comprarão o novo *eReader* mais cedo, quais os que comprarão de seguida e quais os que comprarão mais tarde.

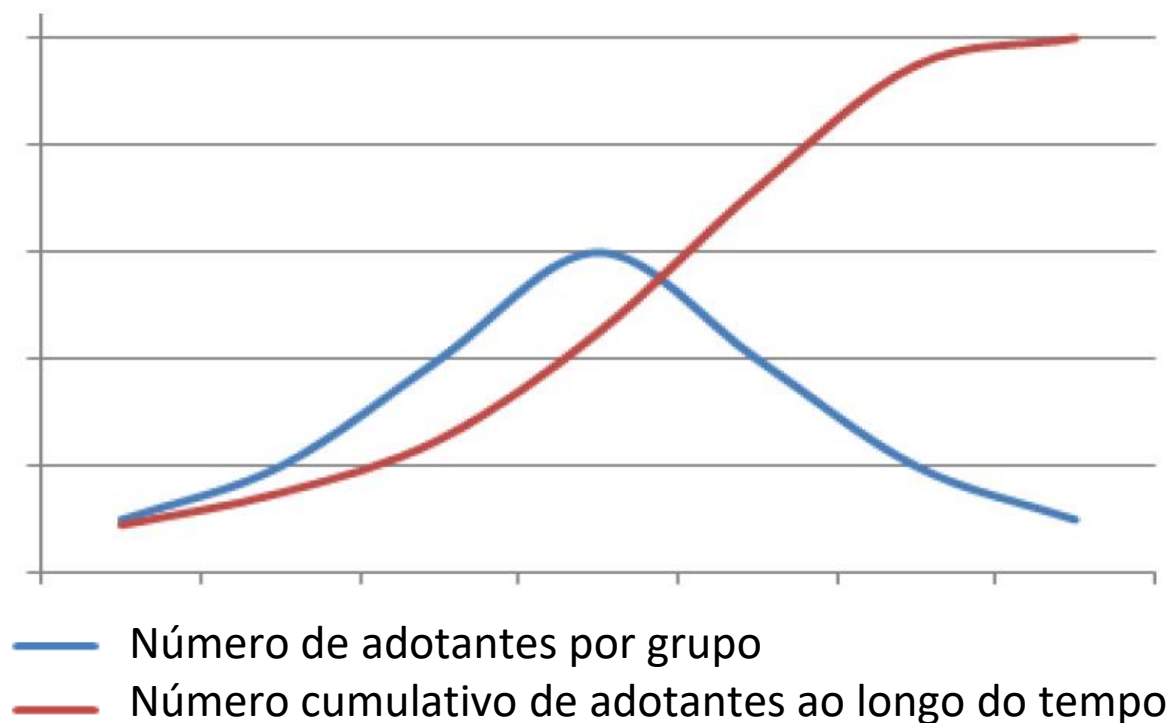
**O Data Mining pode ajudar o Ricardo a prever quando um cliente estará pronto a comprar o *eReader* de próxima geração, podendo assim direcionar seu *marketing* para as pessoas mais prontas para responder a anúncios e promoções.**

# BUSINESS UNDERSTANDING



O Ricardo quer também compreender como os comportamentos dos clientes no *site* da sua empresa poderão indicar o momento da compra do novo *eReader*.

## Teoria da Difusão da Inovação (Rogers, 1960s)

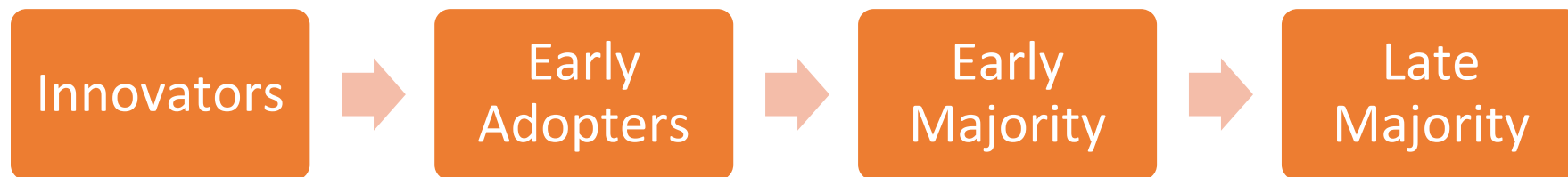


A adoção de uma nova tecnologia ou inovação tende a seguir uma curva em forma de 'S' começando com um grupo menor de clientes mais empreendedores e inovadores que adquirem a tecnologia no início, seguido por grupos maiores de adotantes médios (maioria dos adotantes), seguidos por grupos menores de adotantes tardios.

# BUSINESS UNDERSTANDING



Seguindo a teoria de Rogers, decidiu-se categorizar os clientes da empresa que eventualmente comprarão o novo *eReader* num dos quatro grupos:



O Ricardo espera que, ao observar a atividade dos clientes no *site* da empresa, seja possível antecipar aproximadamente quando cada pessoa terá maior probabilidade de comprar um *eReader*. O **Data Mining** pode ajudar o Ricardo a descobrir quais atividades são os melhores preditores da categoria na qual cada cliente se enquadra.

# DATA UNDERSTANDING



## ***Dataset de Treino***

Contém as atividades do *site* dos clientes que compraram o leitor da geração anterior da empresa e o momento em que eles compraram o leitor.

## ***Dataset de Teste***

Composto por atributos de clientes atuais que se espera que comprem o novo *eReader*.

O Ricardo espera descobrir em qual categoria cada cliente do *dataset* de teste se encaixará, com base nos perfis e no tempo de compra dos clientes do *dataset* de treino.

# DATA UNDERSTANDING



Os *datasets* apresentam os seguintes atributos:

- **User\_ID**: um identificador único e numérico atribuído a cada cliente que possui conta no *site* da empresa.
- **Gender**: o sexo do cliente. No *dataset*, é registado um 'M' para homem e 'F' para mulher. O operador Decision Tree pode manipular tipos de dados não numéricos.
- **Age**: a idade do cliente no momento em que os dados foram extraídos da base de dados do *site*.
- **Marital\_Status**: o estado civil do cliente. No *dataset*: casado -> M, solteiro -> S
- **Site\_Activity**: indicação do quão ativo cada cliente está no site da empresa (raramente, regular ou frequente).



# DATA UNDERSTANDING



- **Browsed\_Electronics\_12Mo:** atributo do tipo Sim / Não, indicando se o cliente pesquisou ou não produtos eletrônicos no *site* da empresa no ano passado.
- **Bought\_Electronics\_12Mo:** atributo do tipo Sim / Não, indicando se o cliente comprou ou não um item eletrônico no *site* da empresa no ano passado.
- **Bought\_Digital\_Media\_18Mo:** atributo do tipo Sim / Não, indicando se o cliente comprou ou não alguma forma de mídia digital (como música MP3) no último ano e meio. Este atributo não inclui compras de livros digitais.
- **Bought\_Digital\_Books:** atributo do tipo Sim / Não, indicando se o cliente comprou ou não um livro digital desde sempre e não apenas no ano passado.

# DATA UNDERSTANDING



- **Payment\_Method:** indica o método como o cliente paga as suas compras. Nos casos em que o cliente efetuou o pagamento através de mais de uma maneira, é utilizado o modo ou método de pagamento mais frequente. Existem quatro opções:
  - Transferência bancária – pagamento via cheque eletrónico ou outra forma de transferência bancária diretamente do banco para a empresa.
  - Conta do *site* – o cliente configurou um cartão de crédito ou uma transferência eletrónica permanente de fundos na sua conta, para que as compras sejam cobradas diretamente na conta no momento da compra.
  - Cartão de crédito - a pessoa insere um número e uma autorização de cartão de crédito cada vez que compra algo no *site*.
  - Faturação mensal - a pessoa faz compras periodicamente e recebe uma fatura em papel ou eletrónica que paga mais tarde enviando um cheque ou através do sistema de pagamento do *site* da empresa.

# DATA UNDERSTANDING



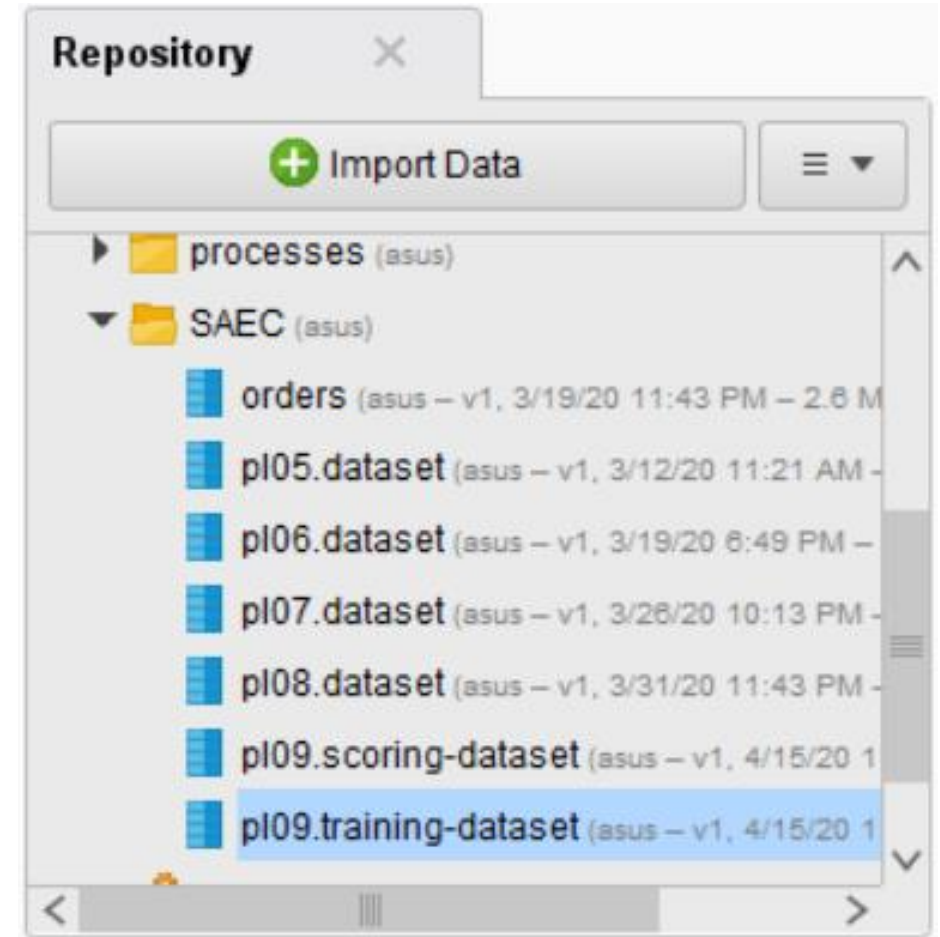
- **eReader\_Adoption:** este atributo existe apenas no *dataset* de treino e consiste em dados sobre clientes que compraram o *eReader* da geração anterior. Aqueles que compraram dentro de uma semana após o lançamento do produto são registrados neste atributo como "*Innovator*". Aqueles que compraram após a primeira semana, mas dentro da segunda ou terceira semana, são inscritos como "*Early Adopter*". Aqueles que compraram após três semanas, mas nos primeiros dois meses, são "*Early Majority*". Aqueles que compraram após os primeiros dois meses são "*Late Majority*". Este atributo servirá como a nossa *label* quando aplicarmos os dados de treino aos dados de teste.

# DATA PREPARATION



Download dos datasets: `pl09.training-dataset.csv`  
`pl09.scoring-dataset.csv`

1. Importe o *dataset* para o repositório do RapidMiner (Import Data -> My Computer).
1. Verifique a *view* dos resultados e inspecione os dados CSV importados. Não precisa de se preocupar com os tipos de dados do atributo porque o operador *Decision Tree* consegue manipular todos os tipos de dados.



# DATA PREPARATION



3. Ligue ambas as portas *out* às portas *res*, como mostrado na figura abaixo, e depois execute o modelo. Examine os dados e familiarize-se com os atributos apresentados na tabela.

The diagram on the left shows two 'Retrieve' blocks. The top block is labeled 'Retrieve pl09.trainin...' and the bottom block is labeled 'Retrieve pl09.scorin...'. Both blocks have an 'out' port on the right. Two lines connect these 'out' ports to two 'res' ports on the right. A large grey arrow points from this diagram to the screenshot of the software interface on the right.

The screenshot shows the 'Result History' window with two tabs: 'ExampleSet (Retrieve Scoring)' and 'ExampleSet (Retrieve Training)'. The 'ExampleSet (Retrieve Scoring)' tab is active. It shows a table of data with columns: User\_ID, Gender, Age, Marital\_Stat..., Website\_Ac..., Browsed\_El..., Bought\_Elec..., and Bo. The table contains 10 rows of data. The 'Data' tab is selected in the left sidebar.

User_ID	Gender	Age	Marital_Stat...	Website_Ac...	Browsed_El...	Bought_Elec...	Bo
56031	M	57	S	Regular	Yes	Yes	Ye
25913	F	51	M	Regular	Yes	Yes	No
19396	M	41	M	Seldom	Yes	Yes	Ye
93666	M	66	S	Regular	Yes	Yes	Ye
72282	F	31	S	Seldom	Yes	No	Ye
64466	M	68	M	Regular	Yes	Yes	Ye
76655	F	51	S	Seldom	Yes	No	No
48465	F	36	S	Frequent	Yes	No	Ye
19889	M	29	M	Regular	Yes	Yes	Ye
63570	M	61	M	Frequent	Yes	No	Ye

# DATA PREPARATION



Aparentemente não há dados inconsistentes nem *missing values*, contudo ainda há alguma preparação dos dados a fazer.

## 1º Atributo User\_ID

Serve apenas como identificador do cliente no *dataset* e portanto não deve ser incluído no modelo como uma variável independente.

### Select Attributes

Remove-se o atributo

OU

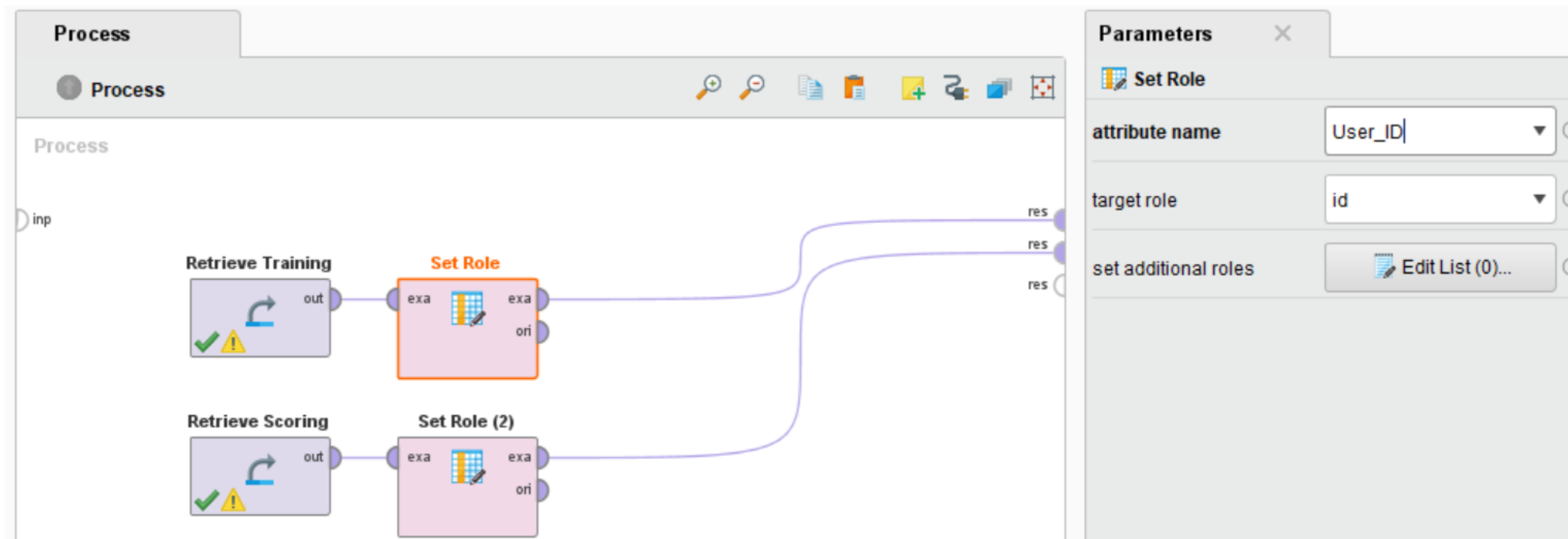
### Set Role

Nova maneira de lidar com  
um atributo não preditivo

# DATA PREPARATION



4. Localize e adicione dois operadores *Set Role* a cada um dos fluxos (treino e teste). Nos parâmetros do lado direito, defina a função do atributo User\_ID como 'id' (para os 2 operadores *Set Role*). Isto faz com que o atributo permaneça no *dataset*, mas que não seja considerado como um preditor para o atributo *label*.



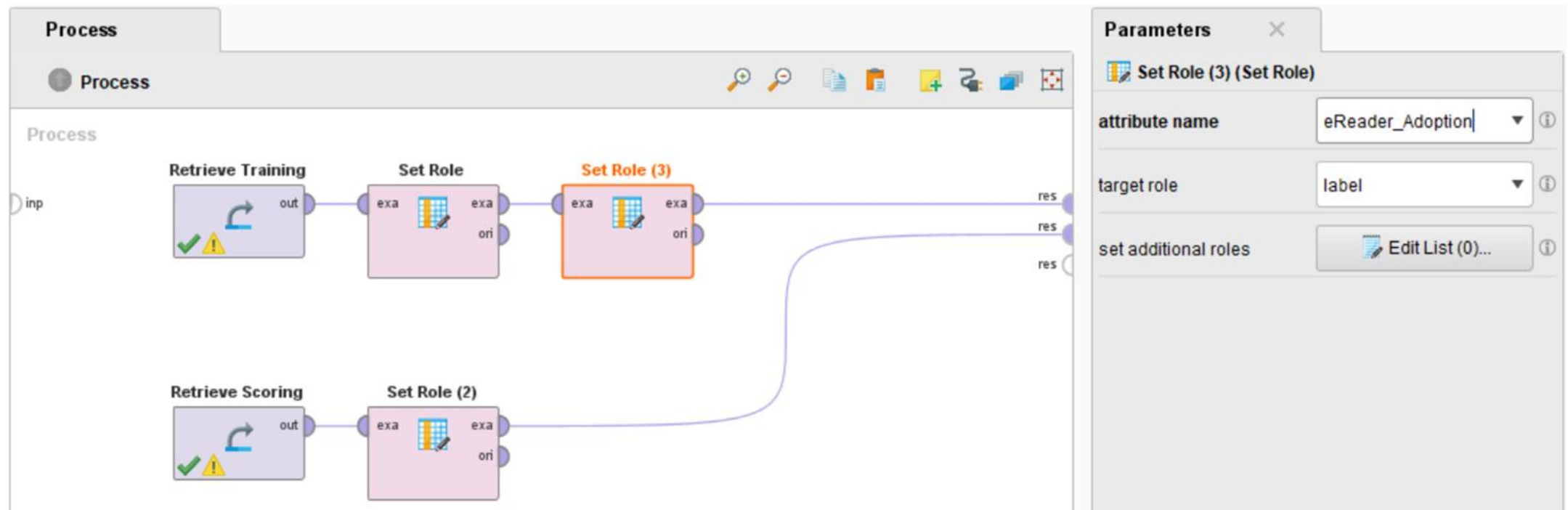
# DATA PREPARATION



## 2º Atributo *label*

Tal como nos outros modelos preditivos, é necessário definir o atributo label.

5. Adicione um operador *Set Role* ao fluxo de treino e defina o atributo “eReader\_Adoption” como ‘label’.

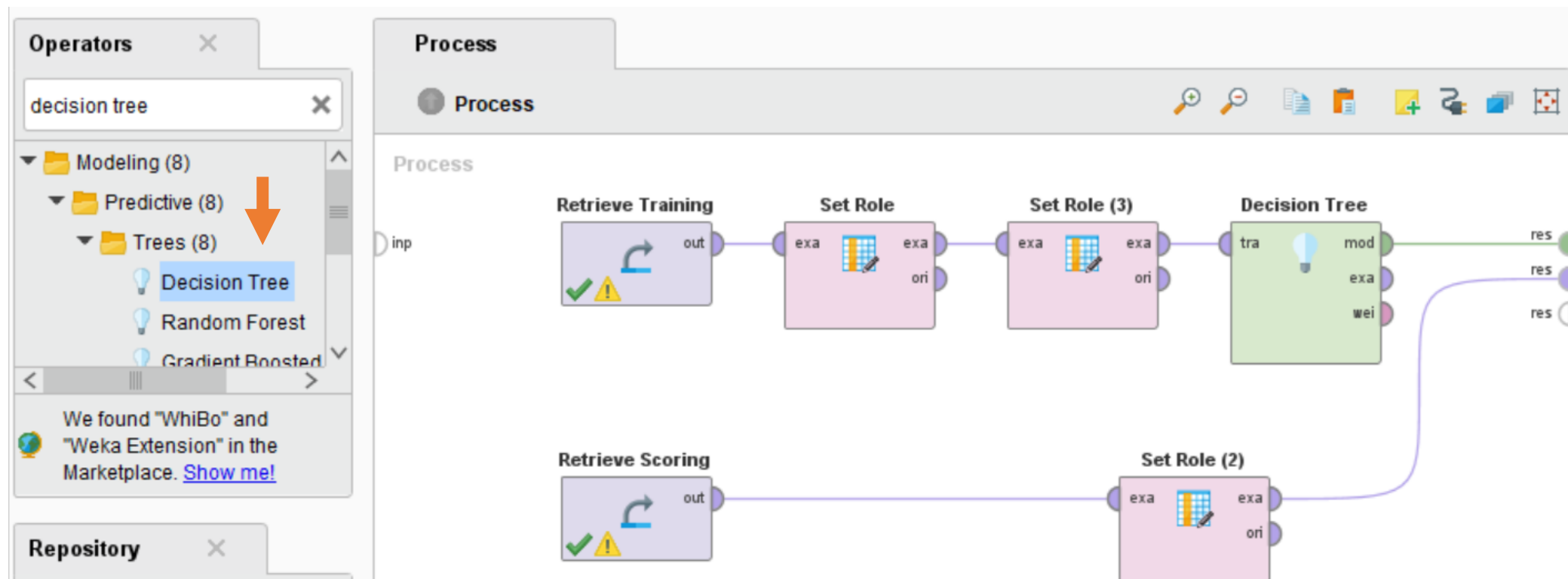




# DATA PREPARATION



6. Em seguida, pesquise nos Operadores por “Decision Tree”. Selecione o operador básico da Árvore de Decisão e adicione-o ao seu fluxo de treino.



# DATA PREPARATION



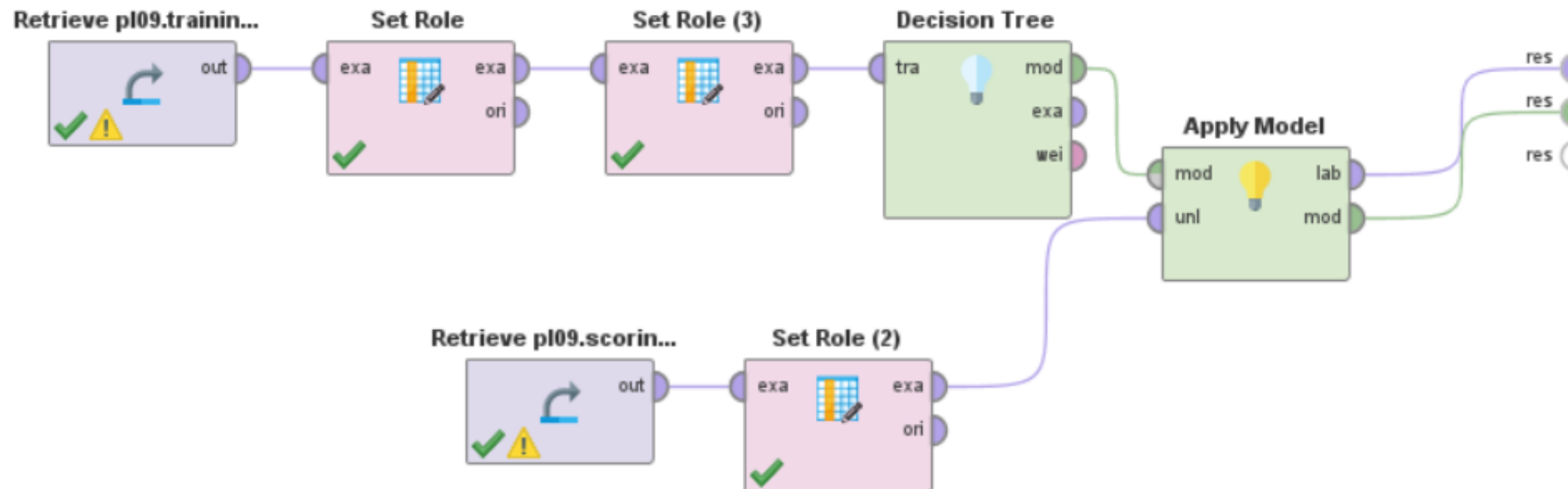
7. Execute o modelo e mude para o separador *Tree (Decision Tree)* na perspectiva de resultados. Conseguirá ver a nossa árvore preliminar, constituída por **nodos** (retângulos totalmente cinzentos) e **folhas** (retângulos cinzentos com uma linha colorida no fundo).

Os nodos são atributos que servem como bons preditores para o atributo *label*. As folhas são os pontos finais que nos mostram a distribuição de categorias do nosso atributo *label* que seguem o ramo da árvore até o ponto dessa folha.

# MODELING



1. Mude para a perspectiva de *Design*. No separador Operadores procure o operador 'Apply Model' e arraste-o para a janela do processo, juntando os fluxos de *training* e *scoring*. Certifique-se de que tanto as portas lab como as portas mod estão ligadas às portas res, de modo a gerar os resultados desejados.



# MODELING



2. Execute o modelo. Clique no separador 'ExampleSet' ao lado do separador 'Tree'. A árvore criada foi aplicada aos dados de teste. Como resultado, foram criados pelo RapidMiner os atributos de confiança, juntamente com um atributo de previsão.

✓ Id <b>User_ID</b>	Integer	0	Min 10153	Max 99694	Average 54647.074
✓ Prediction <b>prediction(eReader_Adoption)</b>	Polynomial	0	Least Innovator (37)	Most Early Adopter (153)	Values Early Adopter (153), Late Majority (14)
✓ Confidence_Early Majority <b>confidence(Early Majority)</b>	Real	0	Min 0	Max 1	Average 0.287
✓ Confidence_Late Majority <b>confidence(Late Majority)</b>	Real	0	Min 0	Max 1	Average 0.294
✓ Confidence_Early Adopter <b>confidence(Early Adopter)</b>	Real	0	Min 0	Max 1	Average 0.288
✓ Confidence_Innovator <b>confidence(Innovator)</b>	Real	0	Min 0	Max 1	Average 0.131
✓ Gender	Polynomial	0	Least F (221)	Most M (252)	Values M (252), F (221)

# MODELING



3. Mude para a opção 'Data View' onde é apresentada a previsão para o grupo de adoção de cada cliente, juntamente com as percentagens de confiança para cada previsão. Existem quatro atributos de confiança, correspondentes aos quatro valores possíveis na etiqueta (eReader\_Adoption).

Row No.	User_ID	prediction(e...	confidence(...	confidence(...	confidence(...	confidence(l...	Gender	Age	Marital_Stat...	Website_A
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57	S	Regular
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51	M	Regular
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41	M	Seldom
4	93666	Early Majority	1	0	0	0	M	66	S	Regular
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31	S	Seldom
6	64466	Early Majority	0.750	0.250	0	0	M	68	M	Regular
7	76655	Late Majority	0.065	0.879	0.056	0	F	51	S	Seldom
8	48465	Innovator	0	0.111	0	0.889	F	36	S	Frequent
9	19889	Late Majority	0	0.500	0.500	0	M	29	M	Regular
10	63570	Early Majority	1	0	0	0	M	61	M	Frequent
11	63239	Early Adopter	0.273	0.045	0.545	0.136	M	47	S	Regular
12	67603	Early Majority	0.950	0	0	0.050	F	62	S	Regular

# MODELING



## Como é que se interpretam estes valores?

As percentagens de confiança somam um total de 100% e medem o quão confiantes estamos de que a previsão se vai tornar realidade. A previsão corresponde à categoria que produziu a maior percentagem de confiança.

Row No.	User_ID	prediction(eReader_Adoption)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
5	72282	Late Majority	0.061	0.879	0.030	0.030
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.065	0.879	0.056	0

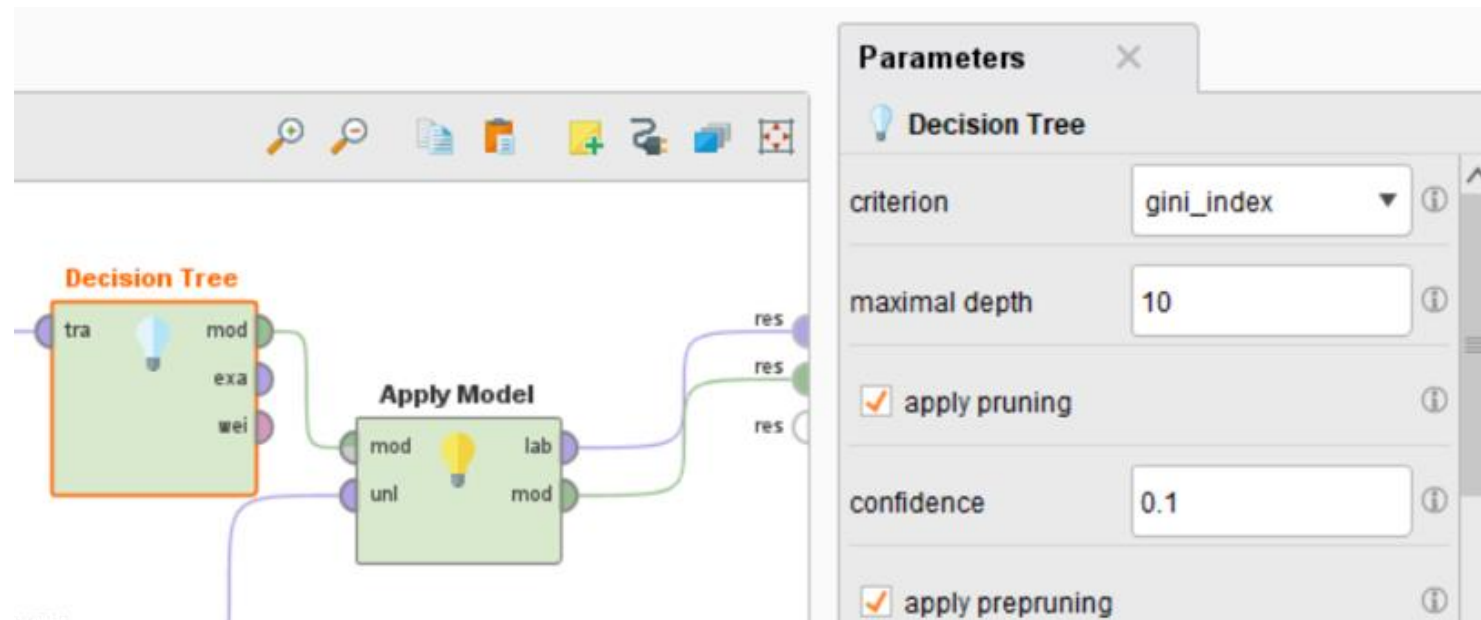
O RapidMiner está bastante (mas não 100%) convencido de que a pessoa 64466 (linha 6) vai ser um membro da *'early majority'* (75%). Apesar de alguma incerteza, o RapidMiner está completamente convencido de que esta pessoa não vai ser um *'early adopter'* (0%) nem um membro da *'innovator'* (0%).

# MODELING



Lembre-se que o CRISP-DM é cíclico por natureza, e que em algumas técnicas de modelação, especialmente naquelas com dados menos estruturados, algumas tentativa-erro podem revelar padrões mais interessantes nos dados.

4. Volte para a perspectiva de *Design*, clique no operador 'Decision Tree' e altere o parâmetro 'criterion' para 'gini\_index'. Guarde o processo com o nome "pl09-classification" e execute o modelo.



# EVALUATION



Através da análise dos resultados vemos que a árvore possui ainda mais detalhe ao usar o critério `gini_index`.

Poderíamos modificar ainda mais a árvore voltando ao separador de *Design* e alterando o número mínimo de itens para formar um nó (*minimal size for split*) ou o tamanho mínimo para uma folha (*minimal leaf size*).

Mesmo com os valores *default* para esses parâmetros, podemos ver que o algoritmo de Gini por si só é mais sensível do que o algoritmo Gain Ratio na identificação de nós e folhas.



# EVALUATION



1. Mude para o separador 'ExampleSet' e escolha a opção 'Data View'. A alteração do algoritmo subjacente à árvore mudou, em alguns casos, a nossa confiança na previsão.

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Innovator	0.333	0	0	0.667	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.750	0.250	0	0	M	68
7	76655	Late Majority	0.333	0.667	0	0	F	51
8	48465	Innovator	0	0.250	0	0.750	F	36
9	19889	Early Majority	0.500	0	0.500	0	M	29
10	63570	Early Majority	1	0	0	0	M	61
11	63239	Early Majority	0.667	0	0.167	0.167	M	47
12	67603	Early Majority	0.917	0	0.042	0.042	F	62

# EVALUATION



Tomemos como exemplo o cliente da linha 2 (ID 25913). Segundo o critério *Gain Ratio*, este cliente foi calculado como tendo pelo menos alguma percentagem de probabilidade de aterrar em qualquer uma das quatro categorias de adoptantes. Existia 54,5% de certeza de que ele seria um *early adopter*, mas quase 27,3% de certeza de que ele também poderia vir a ser um membro da *early majority*.

## *Gain Ratio*

Row No.	User_ID	prediction(...)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51

## *Gini Index*

Row No.	User_ID	prediction(e...)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51

O Ricardo terá de decidir durante a fase de implementação a qual das categorias o cliente pertence. Mas talvez usando o critério *Gini Index*, seja possível ajudá-lo a decidir.

# EVALUATION



De acordo com o critério *Gini Index*, este cliente tem 87,5% de hipóteses de ser um *early adopter* e apenas 12,5% de ser um *innovator*. Note que as probabilidades de ele se tornar parte da *early majority* e da *late majority* baixaram para zero.

Embora o cliente 25913 possa não estar no topo da lista do Ricardo quando a implementação for lançada, ele provavelmente será posicionado mais alto do que seria se estivesse sob o critério *Gain Ratio*.

## *Gain Ratio*

Row No.	User_ID	prediction(...)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.071	0	0.500	0.429	M	57
2	25913	Early Adopter	0.273	0.045	0.545	0.136	F	51

## *Gini Index*

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51



# EVALUATION

Note que, embora o critério *Gini Index* tenha mudado algumas das previsões, não afectou todas. Verifique novamente o ID da pessoa 64466. Não há diferença nas previsões desta pessoa sob qualquer um dos algoritmos. Por vezes o nível de confiança numa previsão através de uma árvore de decisão é tão elevado que um algoritmo subjacente mais sensível não altera em nada os valores dessa previsão.

## *Gain Ratio*

Row No.	User_ID	prediction(eReader_Adoption)	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.065	0.879	0.056	0

## *Gini Index*

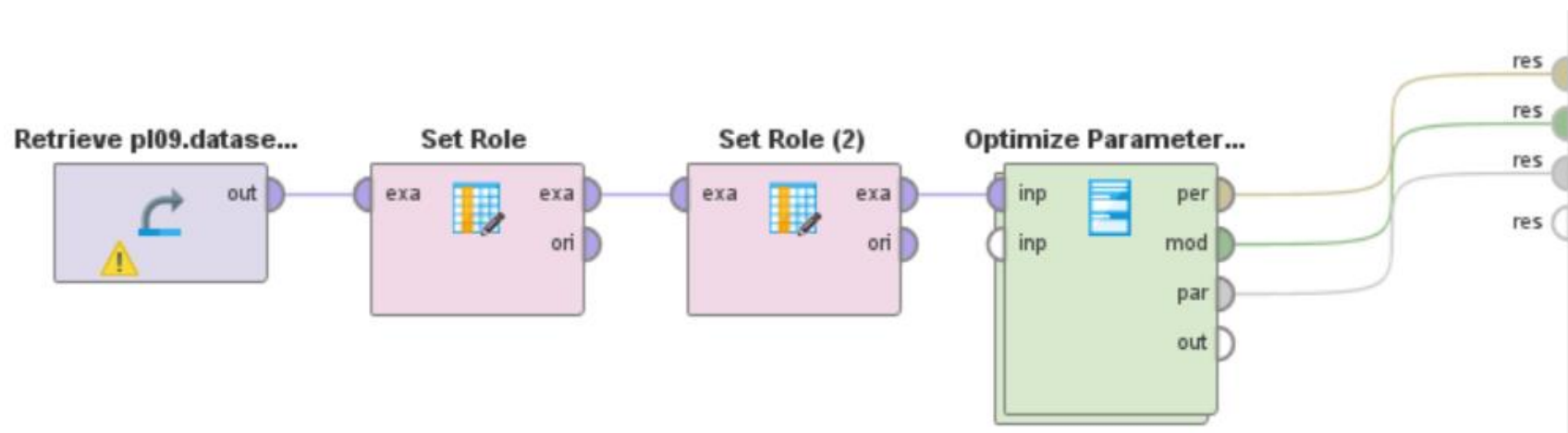
Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)
6	64466	Early Majority	0.750	0.250	0	0
7	76655	Late Majority	0.333	0.667	0	0

# EVALUATION



Após esta abordagem inicial é importante perceber que provavelmente os parâmetros que foram definidos no operador *Decision Tree* não são os mais indicados para alcançar o melhor resultado possível. Assim, é importante tentar encontrar os melhores valores que os parâmetros podem ter de forma a maximizar a *performance* do modelo.

2. Crie um novo processo (*OptimizeParameters*) e arraste o *dataset* de treino para esse processo. Volte a utilizar os dois operadores *Set Role* (um para o ID e um para o atributo *label*), tal como anteriormente. Procure o operador *Optimize Parameters (Grid)* e arraste-o para o processo. Ligue as 3 primeiras portas deste operador às portas *res*.

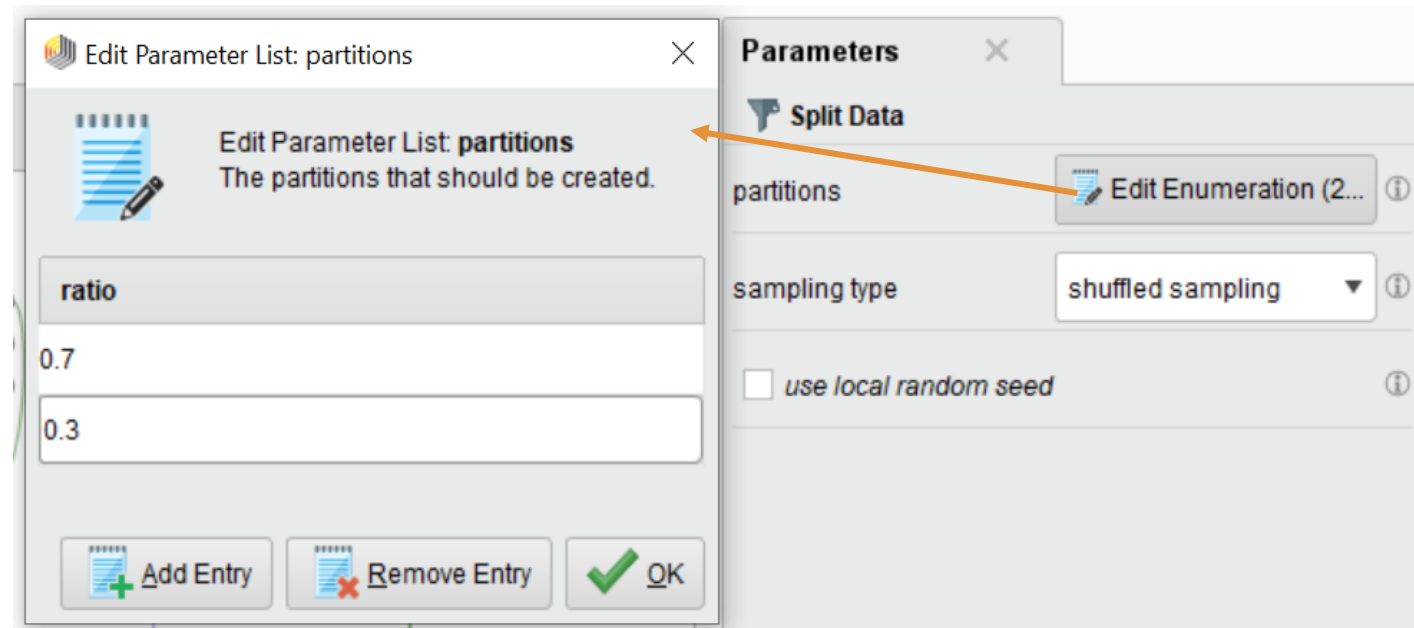


# EVALUATION



O operador *Optimize Parameters (Grid)* é um operador do tipo *nested*. Ele executa o subprocesso que incorpora para todas as combinações de valores dos parâmetros selecionados e, em seguida, devolve os valores ideais dos parâmetros. Falta agora incorporar o subprocesso que queremos repetir, dentro do nosso operador de otimização, ou seja, a nossa classificação com *Decision Tree*.

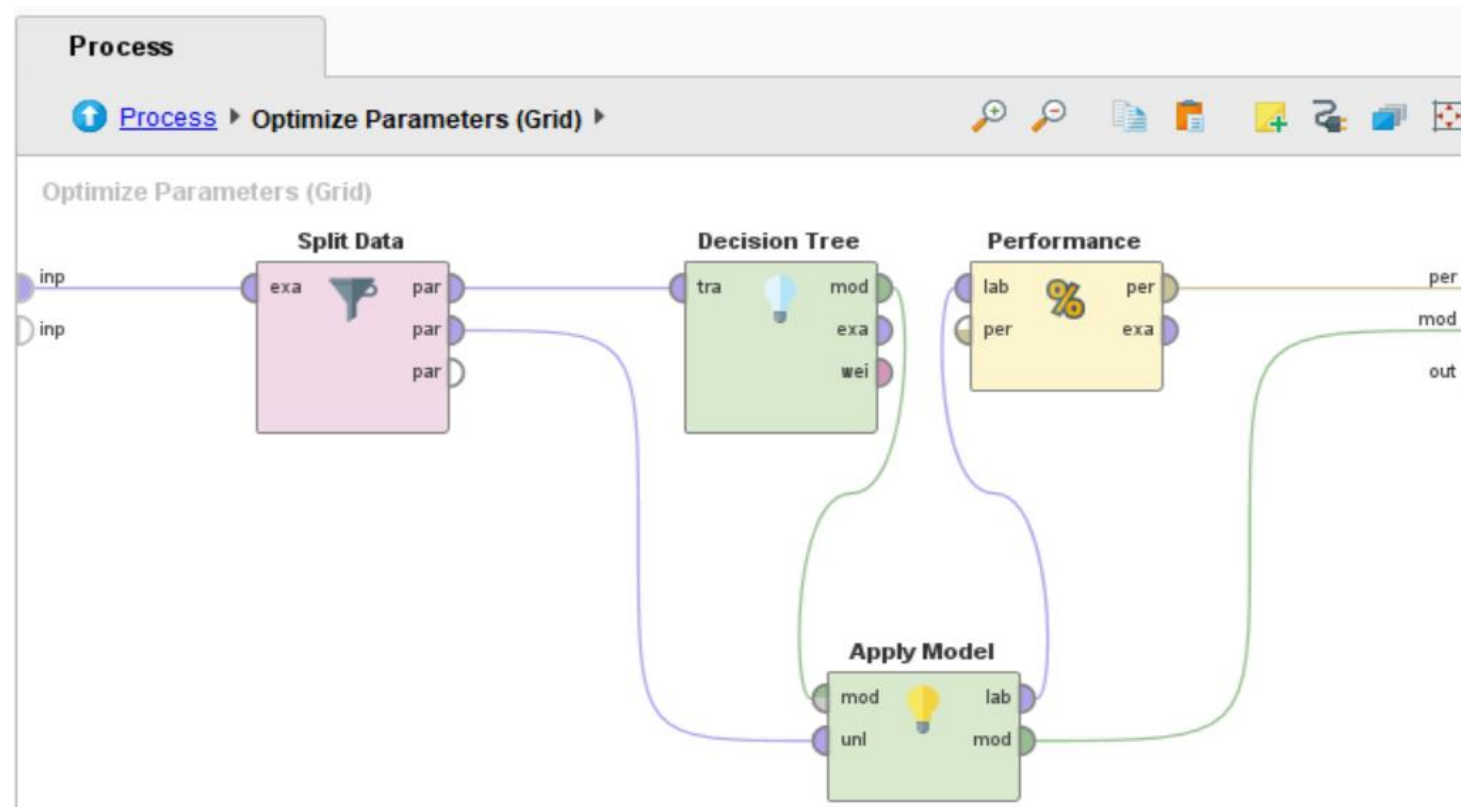
3. Clique duas vezes no operador *Optimize Parameters (Grid)*. Uma nova janela de subprocesso será aberta. Comece o subprocesso com um operador Split Data, uma vez que para este caso será necessário dividir o *dataset*, para que posteriormente seja possível avaliar a *accuracy* do modelo. Defina os parâmetros tal como na imagem



# EVALUATION



4. De seguida, adicione os operadores *Decision Tree* e *Apply Model* tal como representado na figura. Desta vez, será adicionado um operador *Performance* que irá permitir avaliar estatisticamente a performance do modelo de classificação. Por *default*, esta avaliação será feita através da *accuracy*.



# EVALUATION



Agora é necessário indicar no operador de otimização quais os parâmetros que queremos otimizar, neste caso, os parâmetros associados à árvore de decisão.

5. Volte para trás clicando em “Process” no barra do processo.

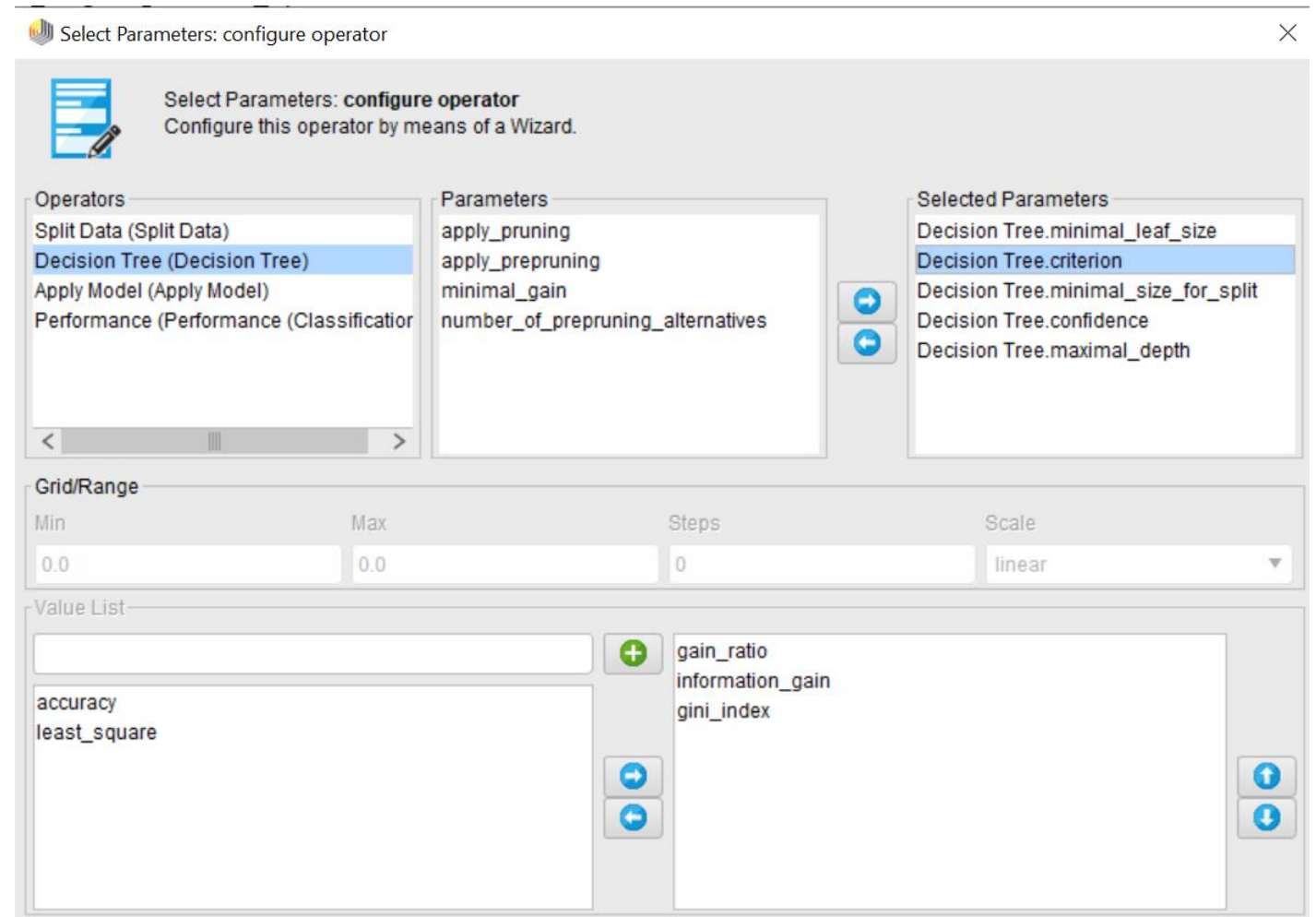


6. Clique uma vez sobre o operador *Optimize Parameters (Grid)* e no painel dos parâmetros, do lado direito, clique em *Edit Parameters Settings*.



# EVALUATION

É aberta uma nova janela onde é possível escolher os parâmetros a otimizar. Vamos primeiramente escolher o *Decision Tree* na lista de *Operators* e posteriormente escolher os *Parameters* pretendidos, enviando-os para a lista do lado direito. É importante ter em atenção que relativamente ao parâmetro *criterion* é necessário retirar os valores *accuracy* e *least\_square* da lista, pois não se adequam ao nosso modelo.



# EVALUATION



7. Finalmente, corra o modelo. Note que quantos mais parâmetros de otimização forem escolhidos, mais lento será a correr. Na janela *Results* verá vários separadores. O separador *ParameterSet* apresenta o melhor resultado (*accuracy*) obtido durante todas as iterações que foram feitas, e quais os valores dos parâmetros utilizados para obter esse resultado. No separador *Optimize Parameters* são apresentadas as iterações feitas para cada parâmetro.

## ParameterSet

Parameter set:

Performance:

PerformanceVector [

-----accuracy: 74.40%

ConfusionMatrix:

True:	Early Majority	Late Majority	Early Adopter	Innovator
Early Majority:	59	5	21	5
Late Majority:	3	62	4	2
Early Adopter:	9	1	46	9
Innovator:	1	1	3	19

```
]
Decision Tree.criterion = gain_ratio
Decision Tree.minimal_size_for_split = 21
Decision Tree.maximal_depth = 39
Decision Tree.confidence = 0.25000005
```

Optimize Parameters (Grid) (3993 rows, 6 columns)

iteration	Decision Tree.criterion	Decision Tree.minimal_size_...	Decision Tree.maximal_...	Decision Tree.confidence	accuracy
1001	information_gain	31	80	0.100	0.700
501	gini_index	11	39	0.050	0.632
1002	gini_index	31	80	0.100	0.632
1	gain_ratio	1	-1	0.000	0.636
502	gain_ratio	21	39	0.050	0.688
1003	gain_ratio	41	80	0.100	0.652
503	information_gain	21	39	0.050	0.664
2	information_gain	1	-1	0.000	0.588
1004	information_gain	41	80	0.100	0.672
504	gini_index	21	39	0.050	0.684
1005	gini_index	41	80	0.100	0.656
505	gain_ratio	31	39	0.050	0.696
1006	gain_ratio	51	80	0.100	0.688
3	gini_index	1	-1	0.000	0.588

# EVALUATION



Agora que descobrimos os valores otimizados dos parâmetros do operador *Decision Tree*, podemos voltar ao processo anterior (*pl09-classification*) para tentar obter melhores resultados na classificação do *dataset* de teste.

9. De volta ao processo (*pl09-classification*), substitua os valores dos parâmetros *criterion*, *minimal\_size\_for\_split*, *maximal\_depth* e *confidence* pelos valores encontrados.

The screenshot displays the Orange3 data mining software interface. On the left, a workflow is visible with a 'Decision Tree' widget (orange border) connected to an 'Apply Model' widget (green border). The 'Decision Tree' widget has inputs 'tra', 'mod', 'exa', and 'wei'. The 'Apply Model' widget has inputs 'mod', 'unl', and 'lab'. A 'le (2)' widget is also present. On the right, the 'Parameters' panel for the 'Decision Tree' widget is open, showing the following settings:

Parameter	Value
criterion	gain_ratio
maximal depth	39
apply pruning	<input checked="" type="checkbox"/>
confidence	0.25
apply prepruning	<input checked="" type="checkbox"/>
minimal gain	0.01
minimal leaf size	1
minimal size for split	21
number of prepruning alternati...	3

# EVALUATION



Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Early Adopter	0.200	0	0.600	0.200	M	57
2	25913	Early Adopter	0	0	0.875	0.125	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Innovator	0.333	0	0	0.667	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.750	0.250	0	0	M	68
7	76655	Late Majority	0.333	0.667	0	0	F	51
8	48465	Innovator	0	0.250	0	0.750	F	36

Row No.	User_ID	prediction(e...	confidence(Early Majority)	confidence(Late Majority)	confidence(Early Adopter)	confidence(Innovator)	Gender	Age
1	56031	Innovator	0	0	0.357	0.643	M	57
2	25913	Early Adopter	0	0	0.800	0.200	F	51
3	19396	Late Majority	0.061	0.879	0.030	0.030	M	41
4	93666	Early Majority	1	0	0	0	M	66
5	72282	Late Majority	0.061	0.879	0.030	0.030	F	31
6	64466	Early Majority	0.815	0.111	0	0.074	M	68
7	76655	Late Majority	0.063	0.874	0.049	0.014	F	51
8	48465	Innovator	0	0.111	0	0.889	F	36

# EVALUATION



Com a actualização dos valores dos parâmetros da árvore de decisão de acordo com os dados encontrados, verificou-se, como seria de esperar, um aumento na confiança das previsões efetuadas pela árvore.

Isto vai de encontro à lógica existente por detrás da metodologia CRISP-DM que defende que o processo de *Data Mining* é cíclico, sendo possível voltar atrás as vezes que forem necessárias para refazer e reajustar o modelo final de forma a obter resultados satisfatórios.

Com estes resultados, o Ricardo possui agora as informações e conhecimento necessários para atingir os seus objetivos.

# DEPLOYMENT



O objetivo do Ricardo é descobrir quais os clientes que se espera que comprem o novo *eReader* e em que prazo, com base no último lançamento de leitor digital da empresa.

A árvore de decisão permitiu-lhe prever isso e determinar até que ponto as previsões são fiáveis. O Ricardo foi também capaz de determinar quais os atributos que têm maior poder preditivo na adoção do *eReader*.

**Mas como é que o Ricardo pode utilizar este novo conhecimento encontrado?**

A resposta mais simples e directa é que ele tem agora uma lista de clientes e os seus prováveis prazos de adoção para o novo *eReader*. Estes clientes são identificáveis pelo User\_ID, através do qual o Ricardo pode iniciar um processo de *marketing* alvo que seja oportuno e relevante para cada indivíduo.

# DEPLOYMENT



Aqueles que têm maior probabilidade de comprar o produto no início (***early adopter***) podem ser contactados e encorajados a comprar assim que o novo produto sair e podem até querer a opção de pré-encomendar o novo dispositivo.

Aqueles que são menos prováveis (***early majority***) podem precisar de alguma persuasão, talvez uma oferta ou um desconto noutro produto com a compra do novo *eReader*.

Os menos prováveis (***late majority***), podem ser alvo de *marketing* de forma passiva, ou talvez não o sejam de todo, se os orçamentos de *marketing* forem apertados e o dinheiro tiver de ser gasto a incentivar os clientes mais prováveis a comprar.

Por outro lado, talvez seja necessário muito pouco *marketing* para os ***innovators***, uma vez que se prevê que estes sejam os mais propensos a comprar o *eReader* em primeiro lugar.

# DEPLOYMENT



O Ricardo tem agora uma árvore que lhe mostra quais os atributos mais importantes para determinar a probabilidade de compra para cada grupo.

As novas campanhas de *marketing* podem então utilizar esta informação para se focarem no aumento do nível de actividade no *site*, ou na associação de produtos electrónicos que estão em desconto no *site* da empresa com os *eReaders*.

Este tipo de promoções intercategóricas podem ser ainda mais aperfeiçoadas para atrair compradores de um sexo específico ou de uma determinada faixa etária.

Com esta análise de Data Mining, o Ricardo possui agora conhecimentos novos e ricos que o vão ajudar a promover o *eReader* de próxima geração.



# RESUMO



As **árvores de decisão** são excelentes modelos de **previsão** quando o atributo alvo é de **natureza categórica** e quando o conjunto de dados é de **tipos mistos**.

As árvores de decisão apresentam a vantagem de lidar de forma eficaz com **atributos que têm valores em falta ou que são inconsistentes** que não são tratados - as árvores de decisão funcionam em torno desses dados e geram resultados úteis.

As árvores de decisão são constituídas por **nós** e **folhas**, representando os **melhores atributos de previsão** num conjunto de dados. Estes nós e folhas conduzem a percentagens de confiança baseadas nos atributos do conjunto de dados de treino (*training*), podendo depois ser aplicadas a dados de teste estruturados de forma semelhante, de modo a gerar previsões para as observações de teste (*scoring*).

As árvores de decisão dizem-nos **qual é a previsão, quão confiantes podemos estar na previsão e como chegámos à previsão**. A parte "como chegámos" é mostrada numa representação gráfica da árvore.