



PL06 – RapidMiner: Classificação

AEC - Mestrado em Engenharia Biomédica

<https://hpeixoto.me/class/aec>

Plano de Aula – PL05



Classificação



Ficha Exercícios (fe05)



Classificação: Exemplo

Contexto e Perspectiva



O **Ricardo** trabalha para uma grande loja online.

A sua empresa vai lançar um **novo eReader** em breve e querem maximizar a efetividade do marketing que vão executar.

O Ricardo notou que alguns dos seus clientes estão mais ansiosos para comprar a versão anterior enquanto outros estão dispostos a esperar pela nova versão do gadget que ainda irá sair.

A questão que o Ricardo coloca é quais serão os fatores que motivam as pessoas a escolher comprar um equipamento assim que este sai no mercado, ou os fatores que fazem com que as pessoas prefiram esperar mais algum tempo para o comprar.

Contexto e Perspectiva



O **Ricardo** acredita que ao extrair os dados dos clientes relativos aos comportamentos gerais de consumo no site, ele será capaz de descobrir quais os clientes que comprarão o **novo eReader mais cedo**, quais os que comprarão a seguir, e quais os que comprarão mais tarde.

Ele espera que, ao prever quando um cliente estará pronto para comprar o eReader de próxima geração, seja capaz de apontar o seu marketing às pessoas mais preparadas para responder a anúncios e promoções.

Classificação

Business Understanding

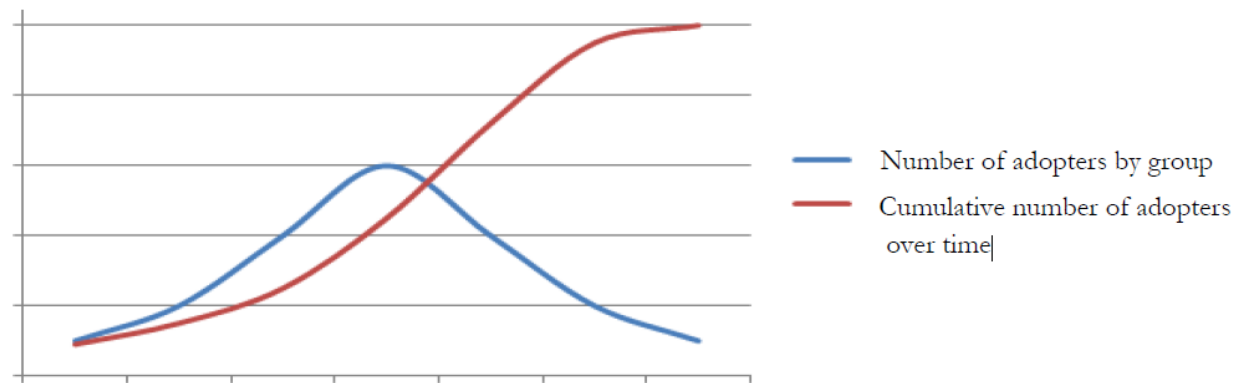


Figure 10-1. Everett Rogers' theory of adoption of new innovations.

Innovators
Early Adopters
Early Majority
Late Majority

Classificação

Contexto e Perspectiva



Ele espera que, ao observar a atividade dos clientes no site da empresa, possa antecipar aproximadamente quando cada pessoa terá mais probabilidades de comprar um eReader.

Ele sente que a pesquisa de dados pode ajudá-lo a descobrir quais as atividades que são os melhores preditores de que categoria um cliente se enquadrará.

Sabendo disto, pode condicionar o marketing a cada cliente para coincidir com a sua probabilidade de compra.

Classificação

Data Understanding

User_ID: Um identificador numérico e único de cada individuo.

Gender: O sexo do cliente, tal como identificado na sua conta de cliente. 'M' - masculino | 'F' - feminino.

Age: A idade da pessoa na altura em que os dados foram extraídos da base de dados do sítio web.

Marital_Status: O estado civil da pessoa, tal como registado na sua conta. "M" – casado | "S" – Solteiro(a), Divorciado(a), Viúvo(a).

Data Understanding

Website_Activity: Número de visitas ao site: "*Seldom*", "*Regular*" ou "*Frequent* "

Browsed_Electronics_12Mo: Sim/Não. Atividade nos últimos 12 meses.

Bought_Electronics_12Mo: Sim/Não. Comprou equipamentos nos últimos 12 meses.

Bought_Digital_Media_18Mo: Sim/Não. Comprou algum tipo de suporte digital, multimédia, MP3, Cds, etc..

Bought_Digital_Books: Sim/Não. Alguma vez comprou livros digitais. Este item está separado dos anteriores devido à sua possível preponderância!

Data Understanding

Payment_Method: Existem quatro opções:

Bank Transfer - pagamento através de cheque eletrónico ou outra forma de transferência bancária diretamente do banco para a empresa.

Website Account - o cliente criou um cartão de crédito ou transferência eletrónica permanente de fundos na sua conta, para que as compras sejam cobradas diretamente através da sua conta no momento da compra.

Credit Card - a pessoa insere um número de cartão de crédito e autorização cada vez que compra algo através do site.

Monthly Billing - a pessoa faz compras periodicamente e recebe uma fatura em papel ou eletrónica que paga mais tarde, quer pelo envio de um cheque ou através do sistema de pagamento do site da empresa.

Data Understanding

eReader_Adoption: Consiste em dados para clientes que adquiriram o eReader de geração anterior:

Innovator – Aqueles que compraram no prazo de uma semana após o lançamento;

Early Adopter – compraram após a primeira semana mas dentro da segunda ou terceira semana;

Early Majority – Compraram após 3 semanas e nos primeiros 2 meses;

Late Majority – Compraram após 2 meses.

Data Preparation

Download do data set: pl05-dataset.csv

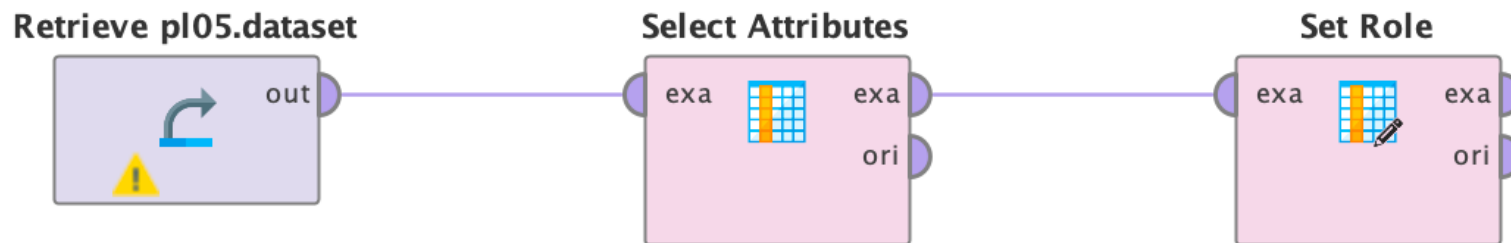
1. Importe o data set para o repositório RapidMiner
2. Avalie os dados importados

ExampleSet (//Local Repository/data/AEC/pl05.dataset)						
Name	Type	Missing	Statistics			
✓ User_ID	Integer	0	Min 1003	Max 9982	Average 5638.188	
✓ ⚠ Gender	Nominal	0	Least F (315)	Most M (346)	Values M (346), F (315)	
✓ ⚠ Age	Integer	0	Min 16	Max 66	Average 42.794	
✓ ⚠ Marital_Status	Nominal	0	Least S (280)	Most M (381)	Values M (381), S (280)	
✓ Website_Activity	Nominal	0	Least Frequent (54)	Most Seldom (424)	Values Seldom (424), Regular (183), ...[1 more]	
✓ Browsed_Electronics_12Mo	Nominal	0	Least No (48)	Most Yes (613)	Values Yes (613), No (48)	
✓ Bought_Electronics_12Mo	Nominal	0	Least No (322)	Most Yes (339)	Values Yes (339), No (322)	
✓ Bought_Digital_Media_18Mo	Nominal	0	Least No (136)	Most Yes (525)	Values Yes (525), No (136)	
✓ Bought_Digital_Books	Nominal	0	Least Yes (297)	Most No (364)	Values No (364), Yes (297)	
✓ Payment_Method	Nominal	0	Least Monthly Billing (93)	Most Website Account (235)	Values Website Account (235), Bank Transfer (229), ...[2 more]	
✓ eReader_Adoption	Nominal	0	Least Innovator (98)	Most Early Adopter (205)	Values Early Adopter (205), Early Majority (186), ...[2 more]	

Data Preparation

3. Como estamos perante um problema de aprendizagem supervisionada, temos de seleccionar os atributos e atribuir o papel de label (class) ao atributo que pretendemos determinar.

- Avalie que atributos deve retirar do modelo?
- Qual a nossa classe?

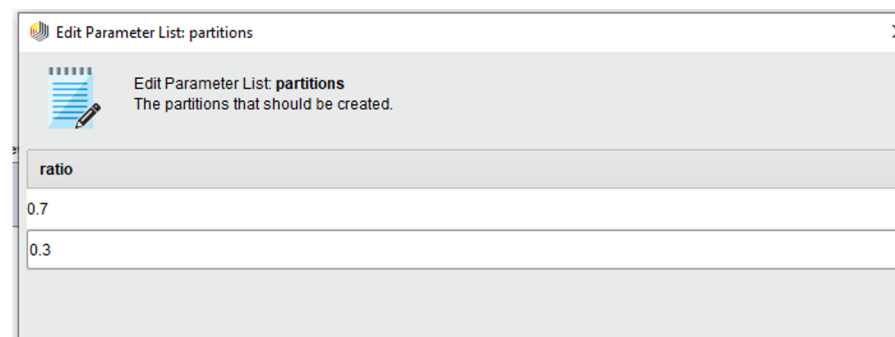
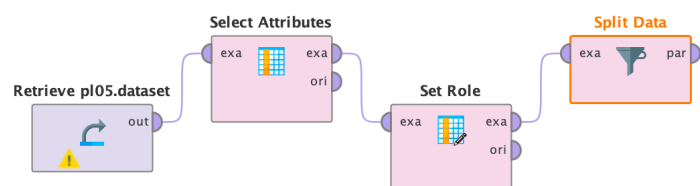


Classificação

Data Preparation

4. O próximo passo é definir os dois conjuntos de treino e teste. Para tal usamos o operador “split data”.

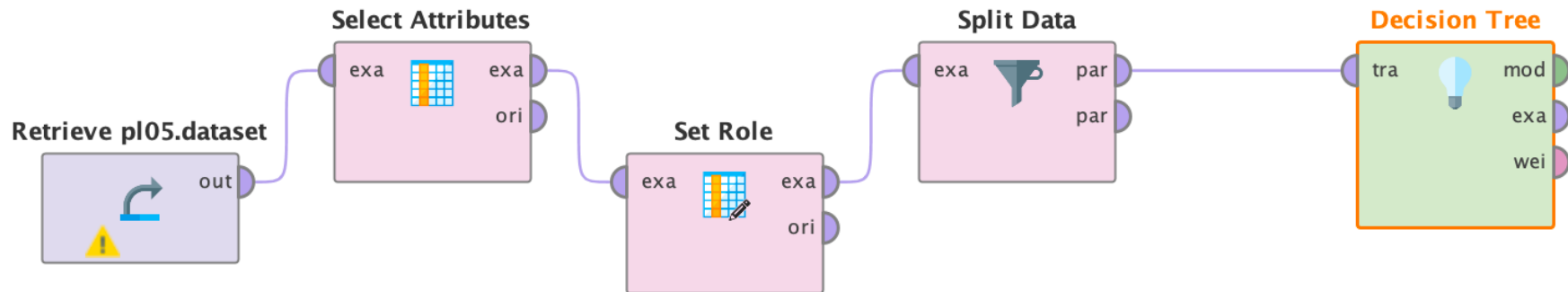
Neste operador podemos definir as percentagens a utilizar, neste caso usando as definições base. Para a definição das percentagens vamos usar 70% para treinar, 30% para testar.



Classificação

Modeling

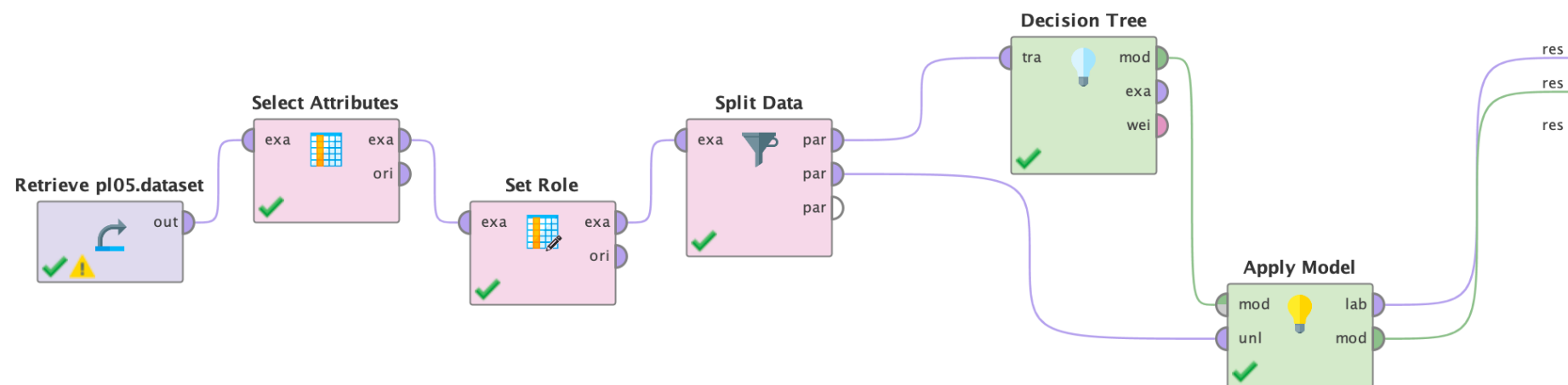
5. Encontrar o operador 'Decision Trees' e arraste-o para a janela do processo. Associe este operador ao fluxo de treino, como mostrado na figura abaixo.



Classificação

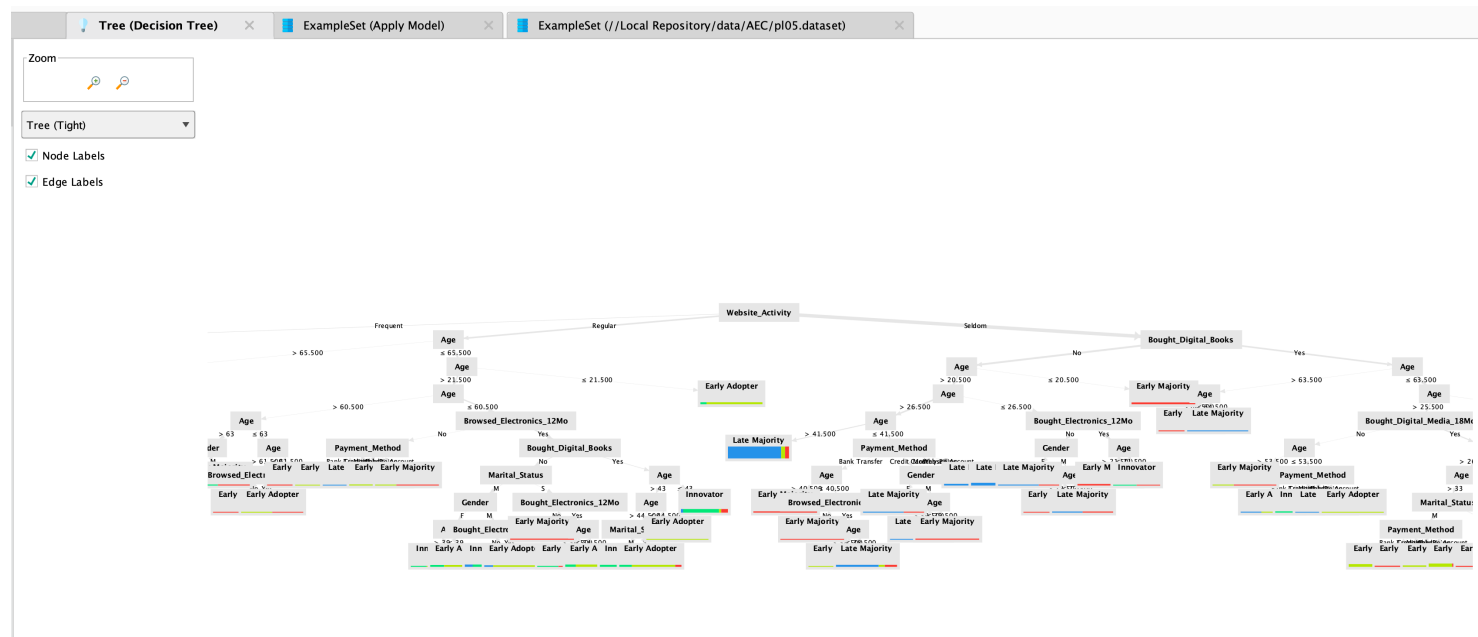
Modeling

6. O passo seguinte na modelação é usar um operador do tipo ‘Apply Model’ para ligar o fluxo de treino ao fluxo de teste. Procure este operador e arraste-o para a janela do processo. Certifique-se de conectar as portas lab e mod às portas res como ilustrado na figura. De seguida devemos colocar também a criação do modelo como modelo de saída para procedermos depois à avaliação.



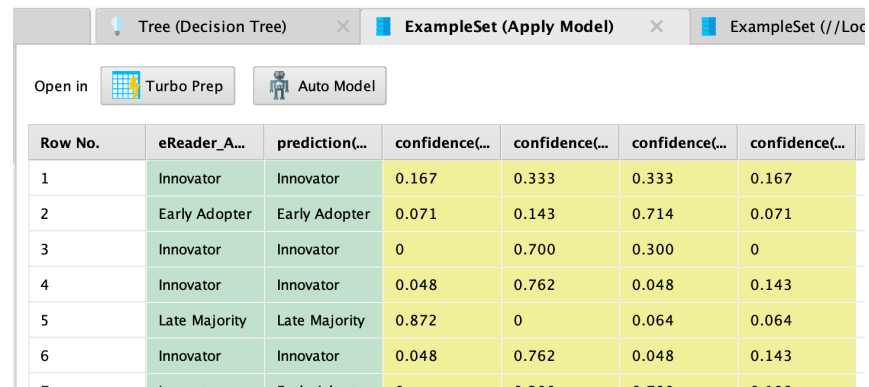
Evaluation

Corra o modelo. O facto de existirem duas saídas do operador 'Apply Model' conectadas às portas res, resultará em dois separadores na perspectiva de resultados. Vamos examinar primeiro o separador Tree(Decision Tree).



Evaluation

Ainda nos resultados, no separador ExampleSet, selecione a opção Data View. Podemos observar que o RapidMiner previu o custo de cada casapara que o Pedro possa perceber como se deve movimentar no mercado.



The screenshot shows the RapidMiner interface with the 'ExampleSet (Apply Model)' tab selected. The table displays the results of a decision tree model, including the predicted class and confidence scores for each row.

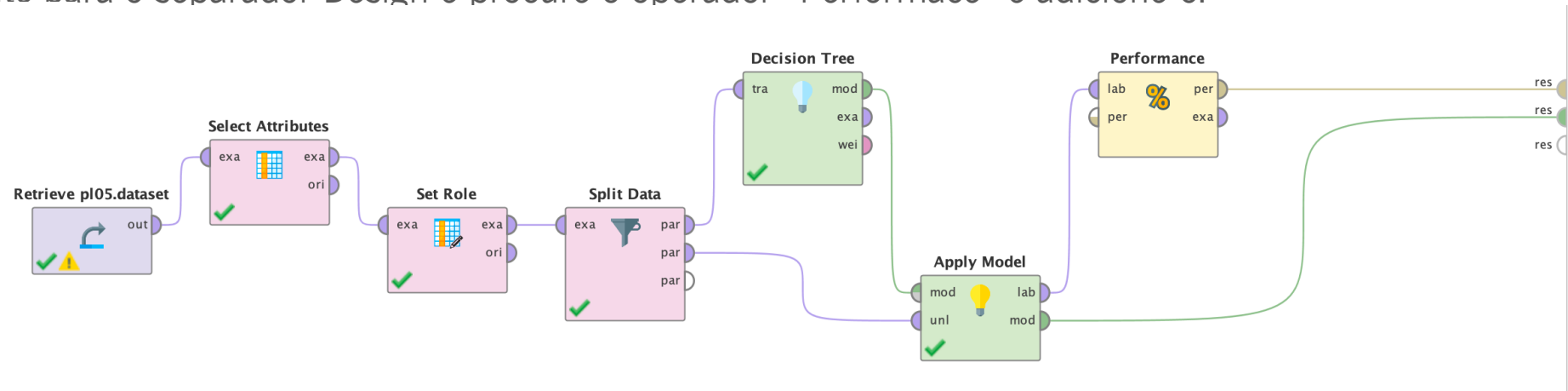
Row No.	eReader_A...	prediction(...)	confidence(...)	confidence(...)	confidence(...)	confidence(...)
1	Innovator	Innovator	0.167	0.333	0.333	0.167
2	Early Adopter	Early Adopter	0.071	0.143	0.714	0.071
3	Innovator	Innovator	0	0.700	0.300	0
4	Innovator	Innovator	0.048	0.762	0.048	0.143
5	Late Majority	Late Majority	0.872	0	0.064	0.064
6	Innovator	Innovator	0.048	0.762	0.048	0.143
7	Innovator	Early Adopter	0	0.300	0.700	0.100

Evaluation

O Ricardo tem agora uma previsão para o tipo de consumidor, mas como será que o modelo se comportou? Será que o modelo estará à altura da exigência do Ricardo para garantir que a adequação da estrutura de marketing?

Como validar esse comportamento?

1) Volte para o separador Design e procure o operador “Performace” e adicione-o.



Evaluation

Accuracy e Correlation Matrix

☒ Table View ☐ Plot View

accuracy: 56.57%

	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	44	2	6	10	70.97%
pred. Innovator	3	12	10	6	38.71%
pred. Early Adopter	3	11	38	22	51.35%
pred. Early Majority	2	4	7	18	58.06%
class recall	84.62%	41.38%	62.30%	32.14%	

Regressão Linear

Modelos

Avaliar a utilização de outro algoritmos de classificação:

- Naive Bayes
- Random Forest
- W-J48
- XGBoost (xgboost extension 0.1.3)

Classificação

Evaluation

iq.opengenus.org		Predicted Class	
		NO	YES
Actual Class	NO	True Negative (TN)	False Positive (FP)
	YES	False Negative (FN)	True Positive (TP)

$$\text{Accuracy} = (TP+TN)/(TP+FP+FN+TN)$$

$$\text{Precision} = TP/(TP+FP)$$

$$\text{Recall} = TP/(TP+FN)$$

$$\text{Specificity} = TN / (TN + FP)$$

Regressão Linear

Resumo

A classificação é um modelo preditivo que usa conjunto de treino e teste para classificar instâncias..

As árvores de decisão são bons modelos de previsão quando o atributo alvo é categórico em natureza, e quando o conjunto de dados é de tipos mistos.

As árvores de decisão são feitas de nós e folhas (ligadas por setas de ramos etiquetadas), representando os melhores atributos de previsão num conjunto de dados.

Estes nós e folhas levam a percentagens de confiança baseadas nos atributos reais em o conjunto de dados de formação, e pode então ser aplicado aos dados de pontuação estruturados de forma semelhante, por ordem para gerar previsões para as observações de pontuação.

Ficha de Exercícios 05



PL06 – RapidMiner: Classificação

AEC - Mestrado em Engenharia Biomédica

<https://hpeixoto.me/class/aec>