PQM13N

Aplicação de Redes Neurais Artificiais para um problema de Regressão utilizando o software *Statistica® 7*

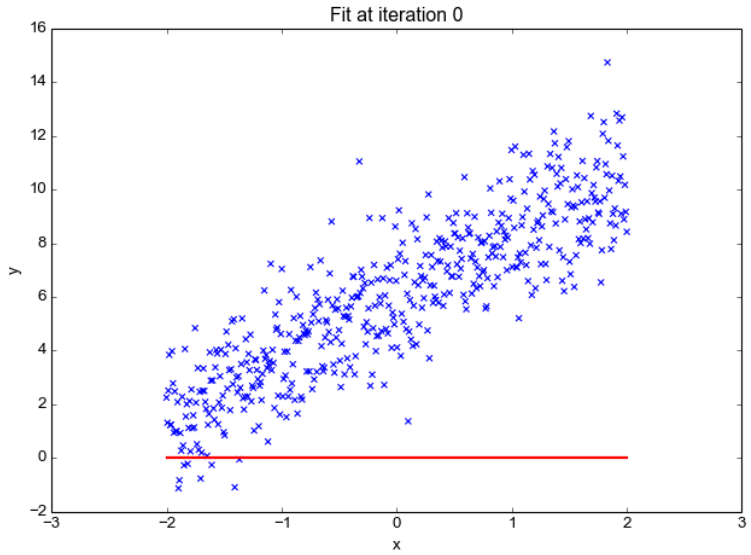
Aluno: Fabricio Alves de Almeida
Professor: Dr. Pedro Paulo Balestrassi

Instituto de IEPG
Engenharia de Produção e Gestão

Introdução

Contextualização:

Um problema de regressão pode ser definido como "a busca por aproximar uma determinada função de mapeamento ($F(x)$) das variáveis de entrada (x) para uma variável de saída contínua (y)".



Estudo de Caso

Para realizar a aplicação, utilizaremos dados reais do setor imobiliário, onde temos um histórico de preços e tamanho dos imóveis para uma determinada cidade dos Estados Unidos da América.

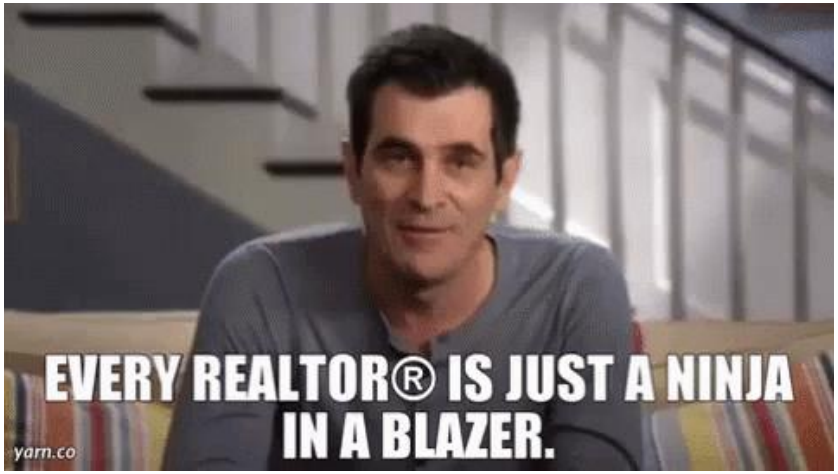
A partir desses dados, busca-se criar um modelo que possa prever o preço dos imóveis baseado em seu tamanho, favorecendo à estratégias de vendas.

Estudo de Caso

Deste modo, tem-se como variáveis:

- ***Preço***: variável contínua (*y*) de saída;
- ***Tamanho***: variável contínua (*x*) de entrada.

No total, tem-se 21.613 linhas de dados, no qual às 20 últimas serão utilizadas para validar a RNA.



Dados

Treinamento		
Dados	Preço [*US$*]	Tamanho [ft^2]
1	$ 221,900.00	1180
2	$ 538,000.00	2570
3	$ 180,000.00	770
4	$ 604,000.00	1960
5	$ 510,000.00	1680
6	$ 1,230,000.00	5420
7	$ 257,500.00	1715
8	$ 291,850.00	1060
...	...	...
21590	$ 347,500.00	2540
21591	$ 1,220,000.00	4910
21592	$ 572,000.00	2770
21593	$ 475,000.00	1190

Validação		
Dados	Preço [*US$*]	Tamanho [ft^2]
1	$ 1,090,000.00	4170
2	$ 350,000.00	2500
3	$ 520,000.00	1530
4	$ 679,950.00	3600
5	$ 1,580,000.00	3410
6	$ 541,800.00	3118
7	$ 810,000.00	3990
8	$ 1,540,000.00	4470
...	...	...
17	$ 400,000.00	2310
18	$ 402,101.00	1020
19	$ 400,000.00	1600
20	$ 325,000.00	1020

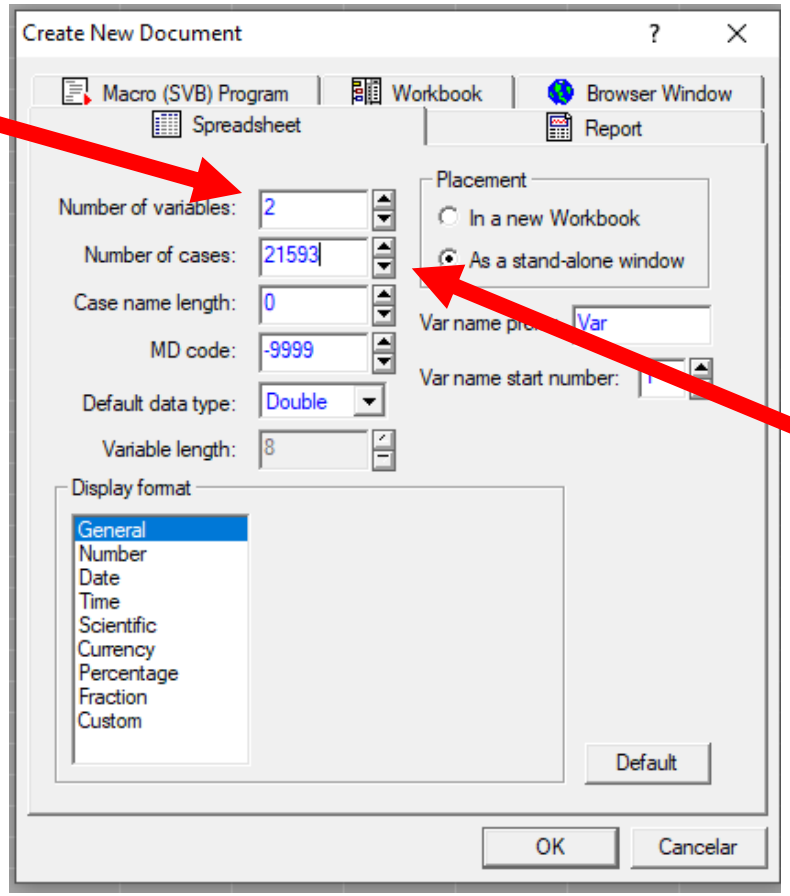
Aplicação

Inicialmente deve-se trazer as informações para o *Statistica*®.

Clicar em *"file > new"* ou utilizar o atalho *"Ctrl + N"*

Adicionar o número de variáveis dos dados.

Neste caso "*preço*" & "*tamanho*" ; $n = \mathbf{2}$



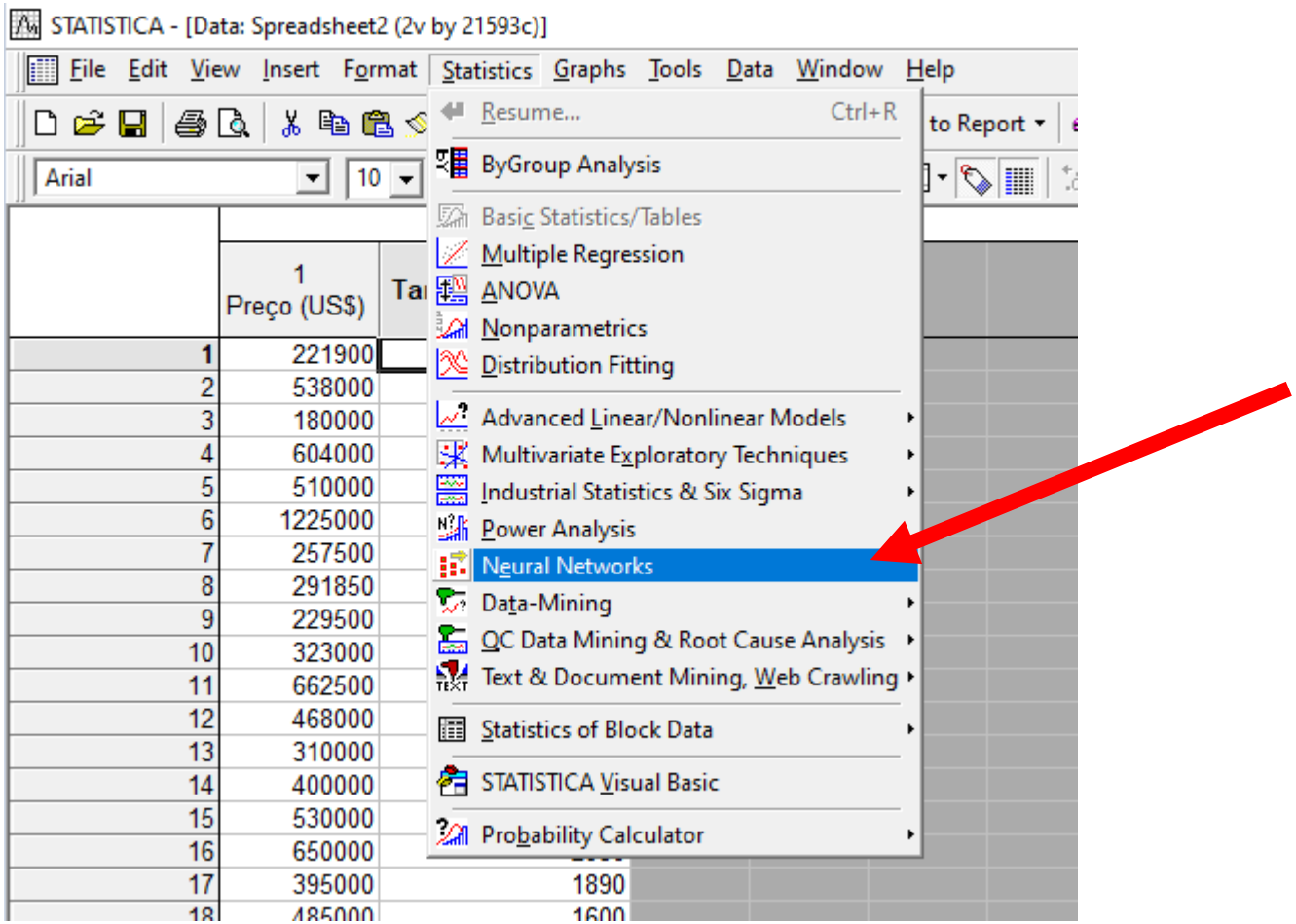
Adicionar o número de linhas do conjunto.

Neste caso:
$21.613 - 20 = \mathbf{21.593}$

Aplicação

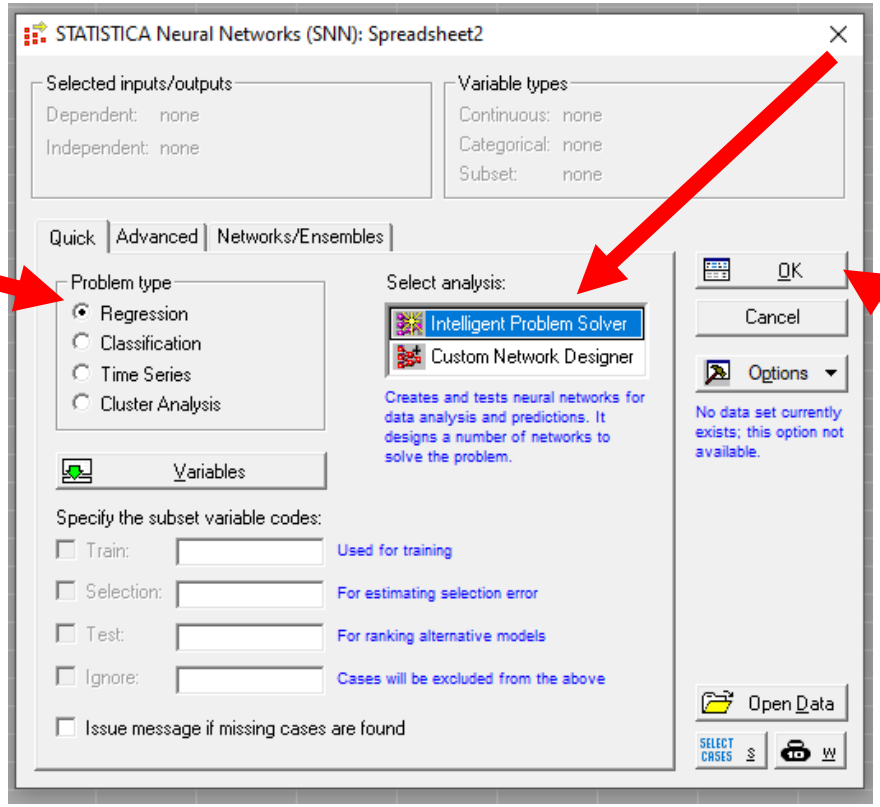
Em seguida, deve-se buscar o pacote de Redes Neurais.

Statistics > Neural Networks



Aplicação

A seguinte janela será aberta. Assim, deve-se selecionar o tipo do problema (neste caso: *Regressão*) e o tipo de análise, que manteremos o *IPS*, em que, o próprio *software* define a parametrização ótima para o estudo.

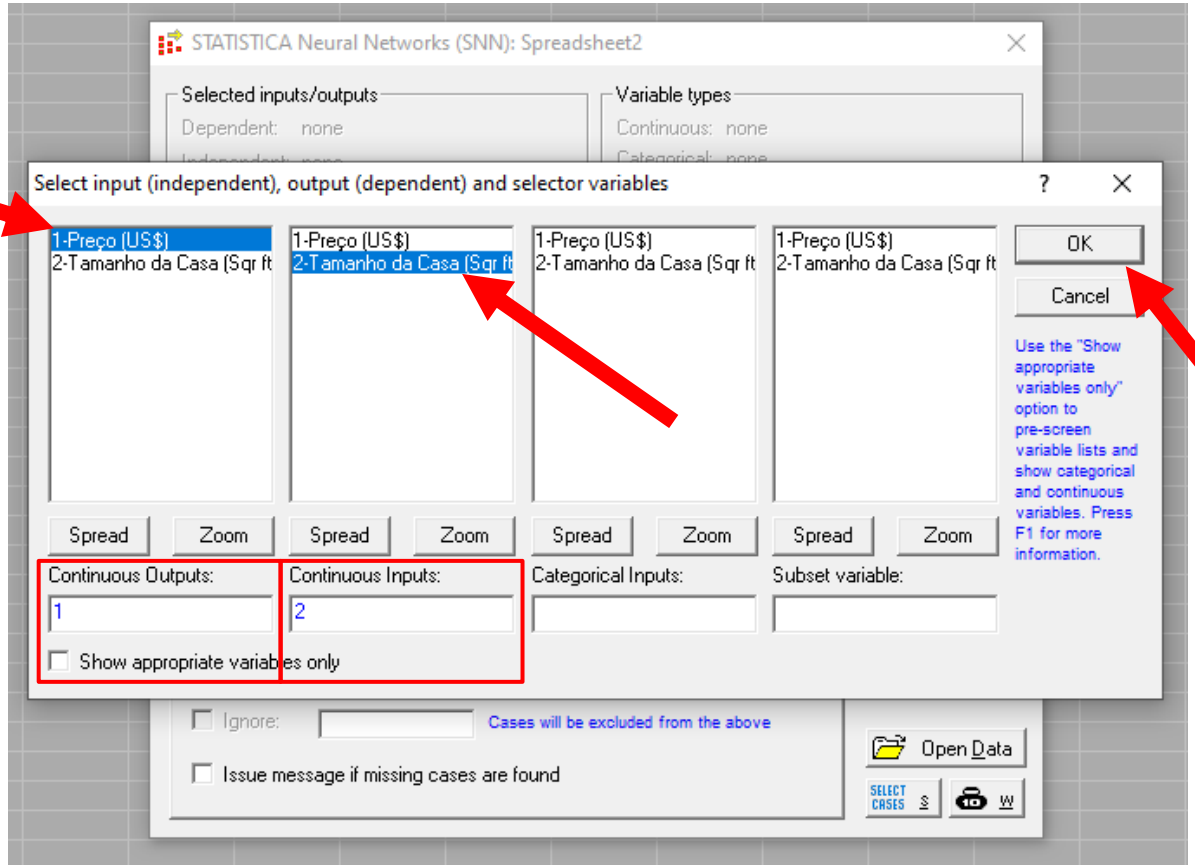


Após isso, deve-se clicar em OK para seguir na próxima etapa.

Aplicação

Em seguida, deve-se definir as variáveis de saída e entrada.

Para este caso: "*Preço*" é variável de saída (*y*) & "*Tamanho*" é a variável de entrada (*x*).

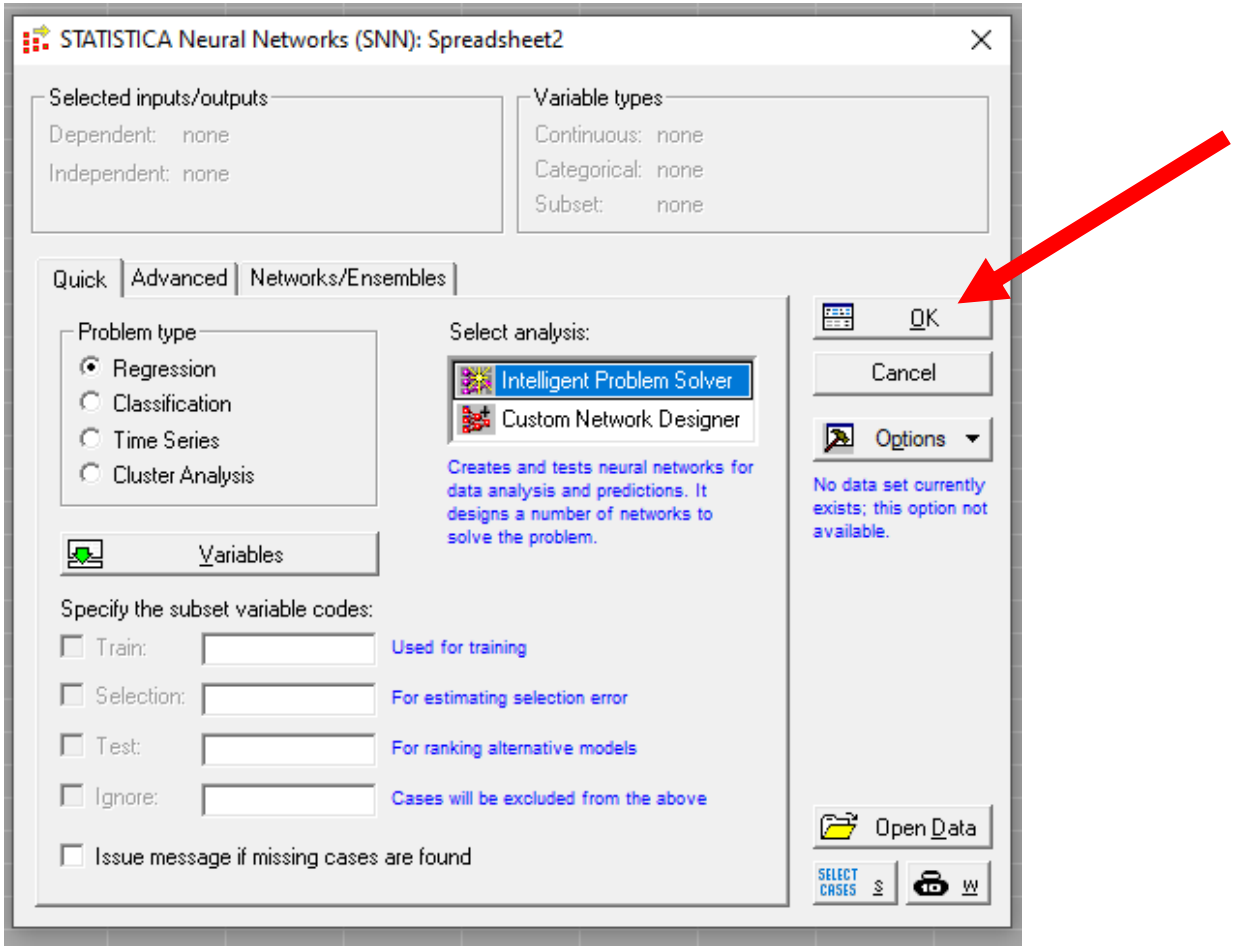


Após isso, deve-se clicar em OK para seguir na próxima etapa.

Aplicação

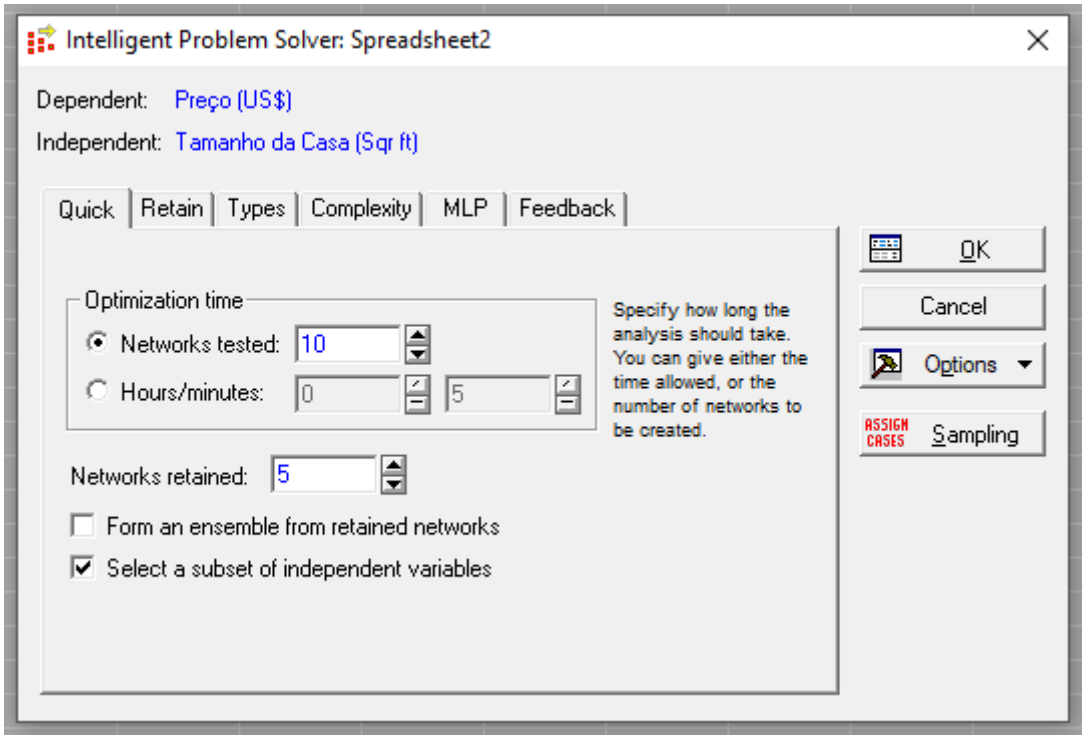
Após selecionar as variáveis do problema, o *software* volta a esta janela.

Para prosseguirmos, deve-se clicar novamente em "*OK*".



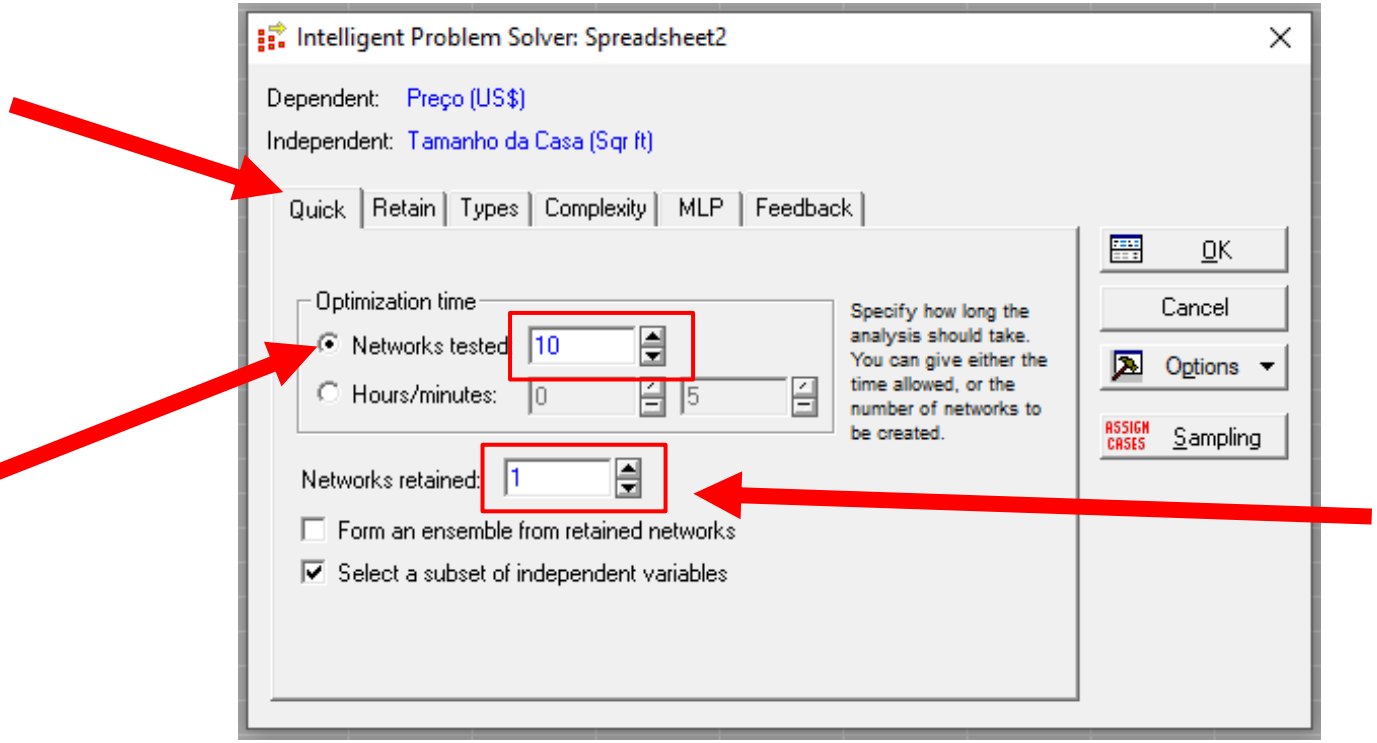
Aplicação

Agora, dentro do IPS, tem-se várias configurações de RNA para escolher, com abas distintas (*Quick*, *Retain*, *Types*, *Complexity*, *MPL* & *Feedback*). Trataremos delas separadamente a seguir.



Aplicação

Para a primeira aba "*Quick*", pode-se escolher a quantidade de redes testadas ou o tempo de teste, determinando o quanto a análise demorará. Baseado em testes preliminares, selecionaremos uma quantidade de *10 redes* para teste. *(tais configurações são opcionais)*

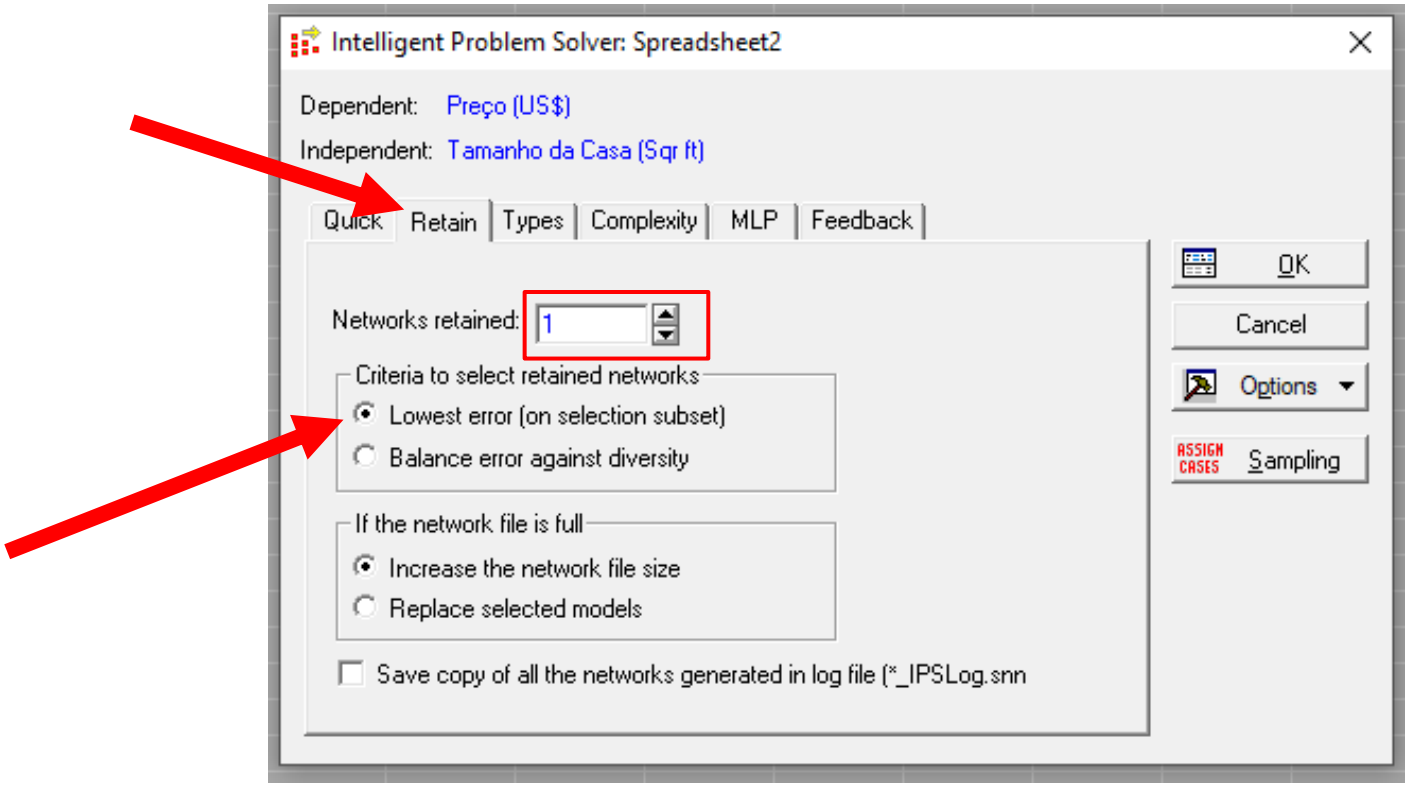


Ainda nesta aba, deve-se definir a quantidade de redes retidas.

Para nosso exemplo, iremos reter apenas ***UMA*** *rede*, ou seja, apenas a que apresentar melhor desempenho.

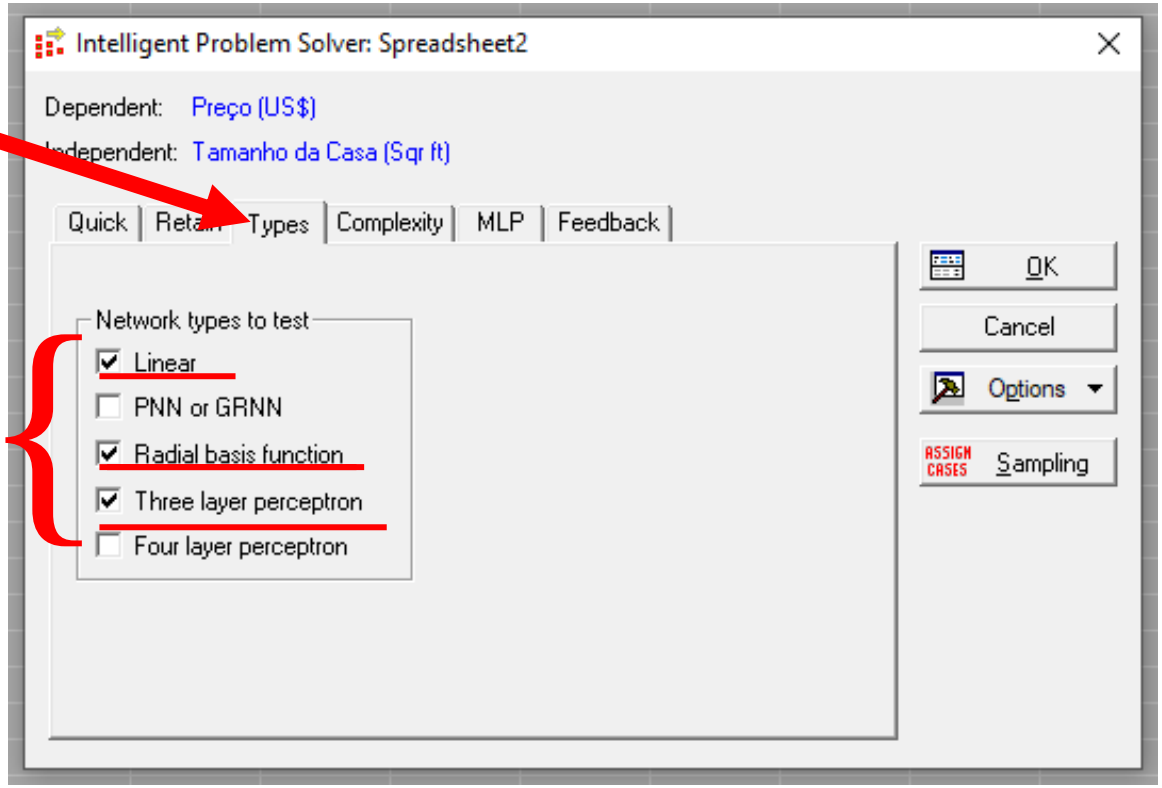
Aplicação

Na aba "*Retain*" já temos a informação de quantas redes serão retidas (neste caso, 1). Assim, em termos de parametrização, deve-se selecionar como critério de retenção a rede que apresentar o **menor erro**.



Aplicação

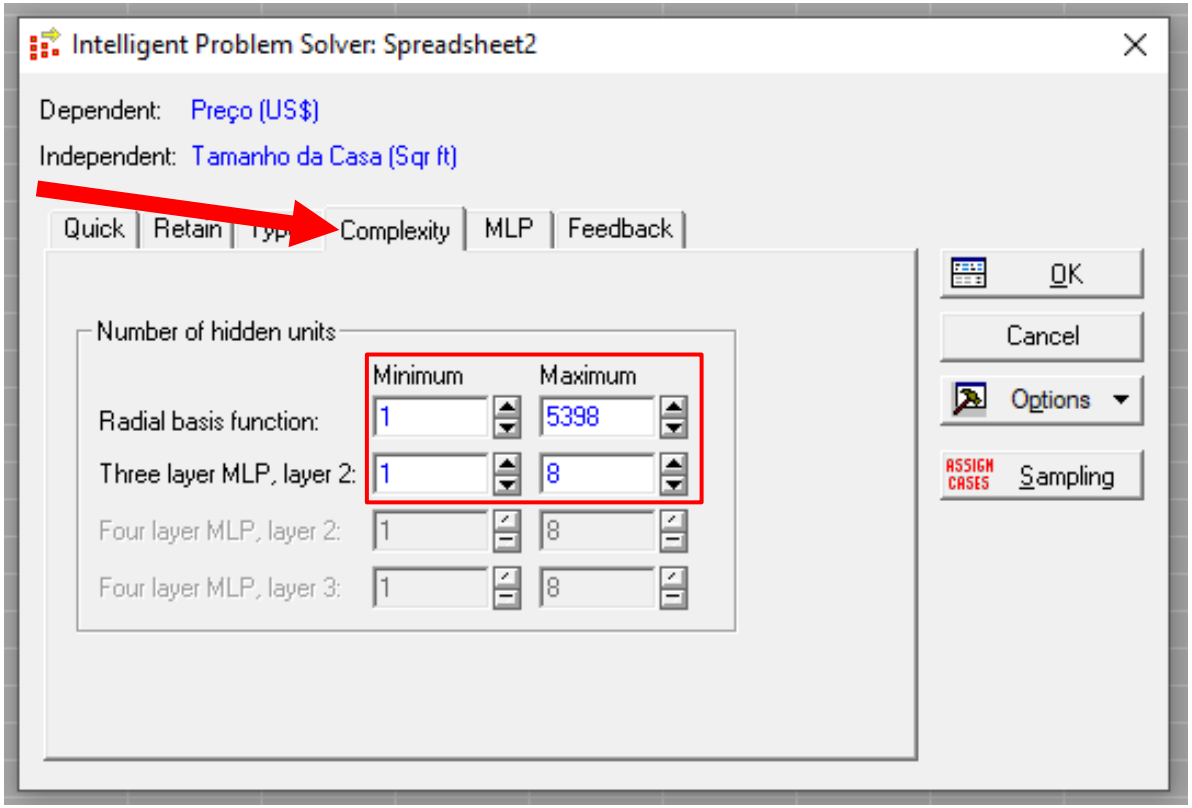
Na aba "*Types*" tem-se a opção de selecionar os tipos de redes que serão testadas. Usualmente, a configuração abaixo é a padrão do *software* e manteremos, para este caso, assim.



Veja que existe a opção da *perceptron* com mais camadas, contudo, muitos autores inferem que muitas camadas pode trazer problemas de *overfitting*.

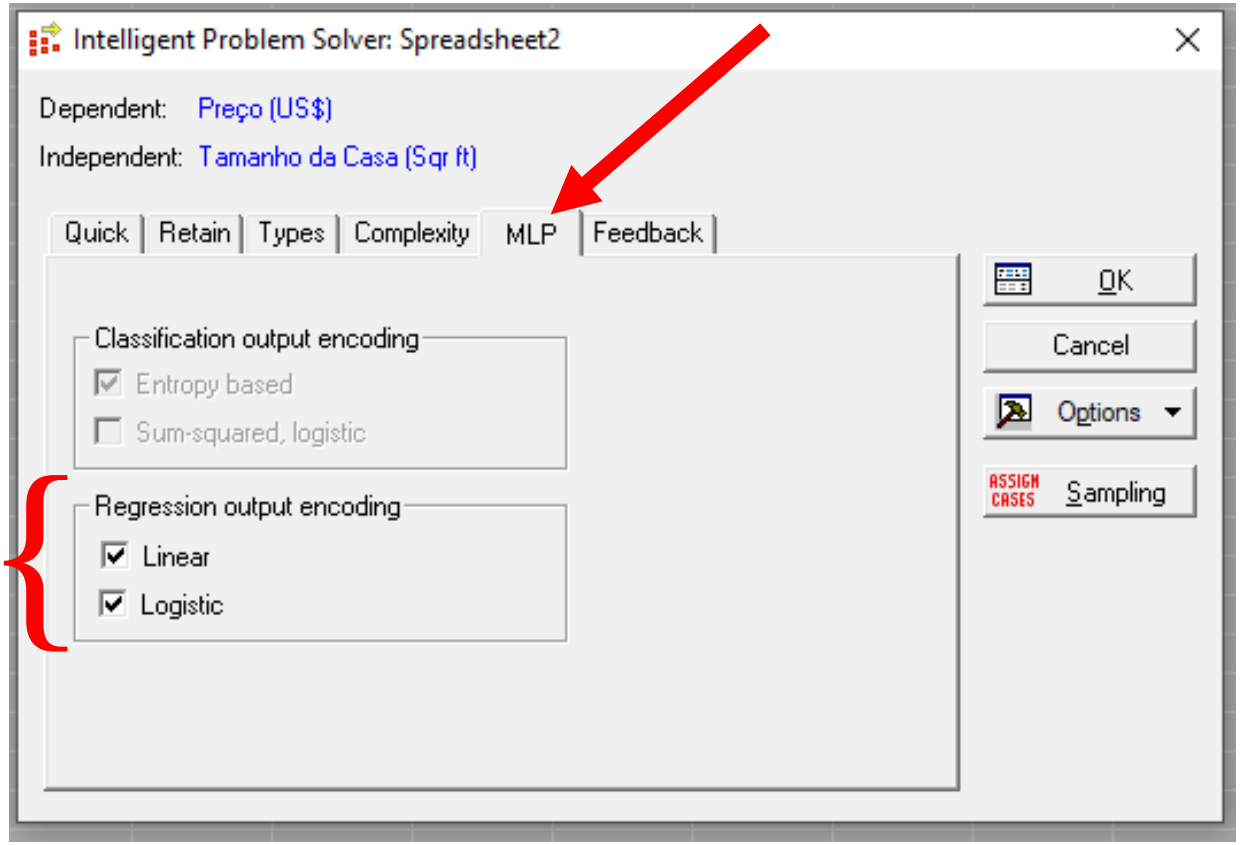
Aplicação

Na aba "*Complexity*" também manteremos a configuração padrão, calculada pelo próprio *software*, para este estudo de caso.



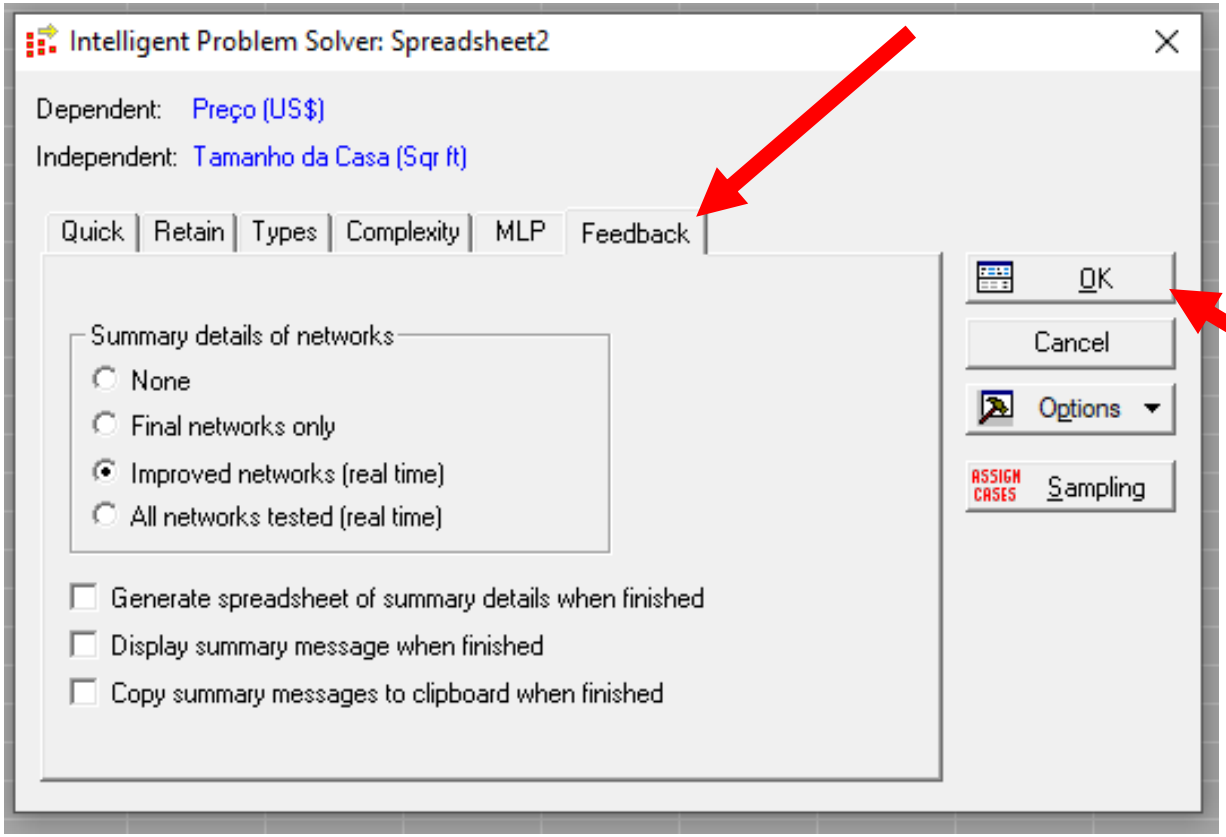
Aplicação

Em seguida, para a aba "*MLP*" tem-se a codificação da saída de regressão, onde é possível selecionar a opção linear e logística. Para nosso exemplo, iremos selecionar ambas opções.



Aplicação

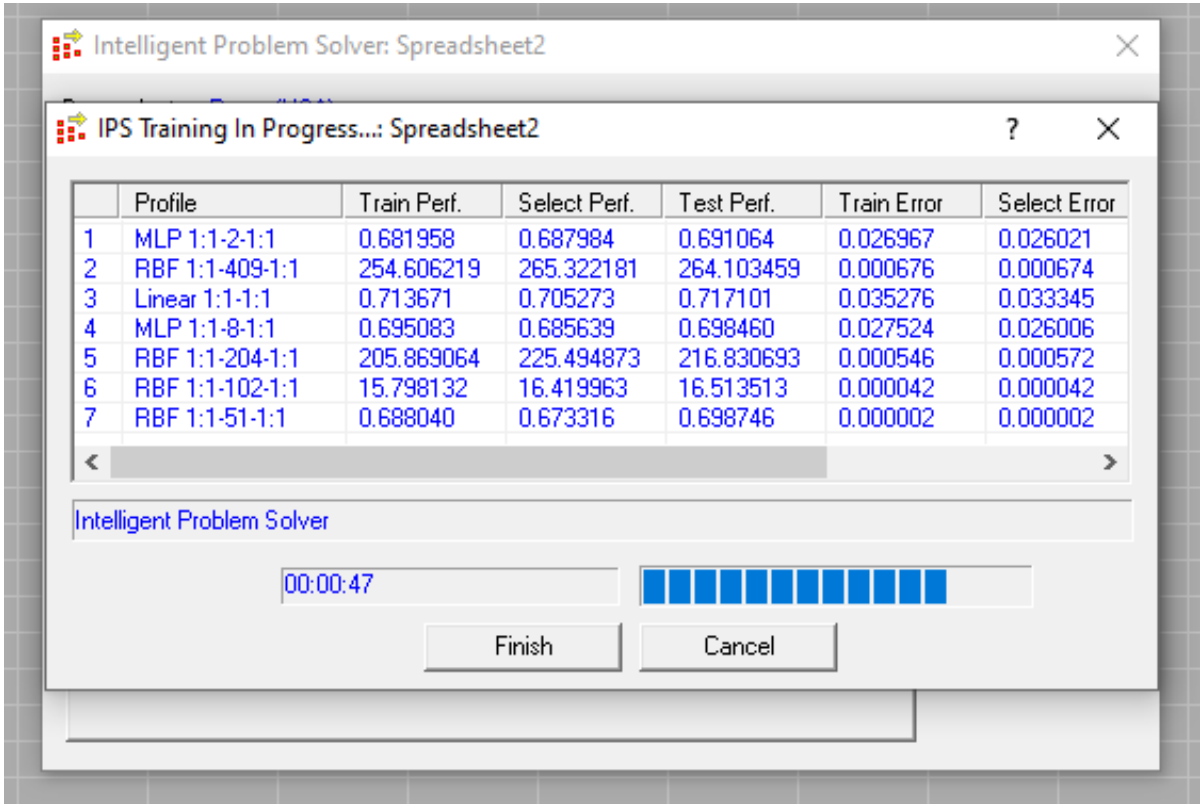
Por fim, na aba "*Feedback*" tem-se a opção de selecionar detalhes da rede e também do *software*. Neste caso, não iremos modificar as opções padrões.



Após isso, deve-se clicar em OK para realizar o treinamento da rede.

Resultados

Após clicar em "*OK*", como informado no slide anterior, o *software* começa a realizar o treinamento e apresenta os resultados prévios no *display*, similar a imagem abaixo:



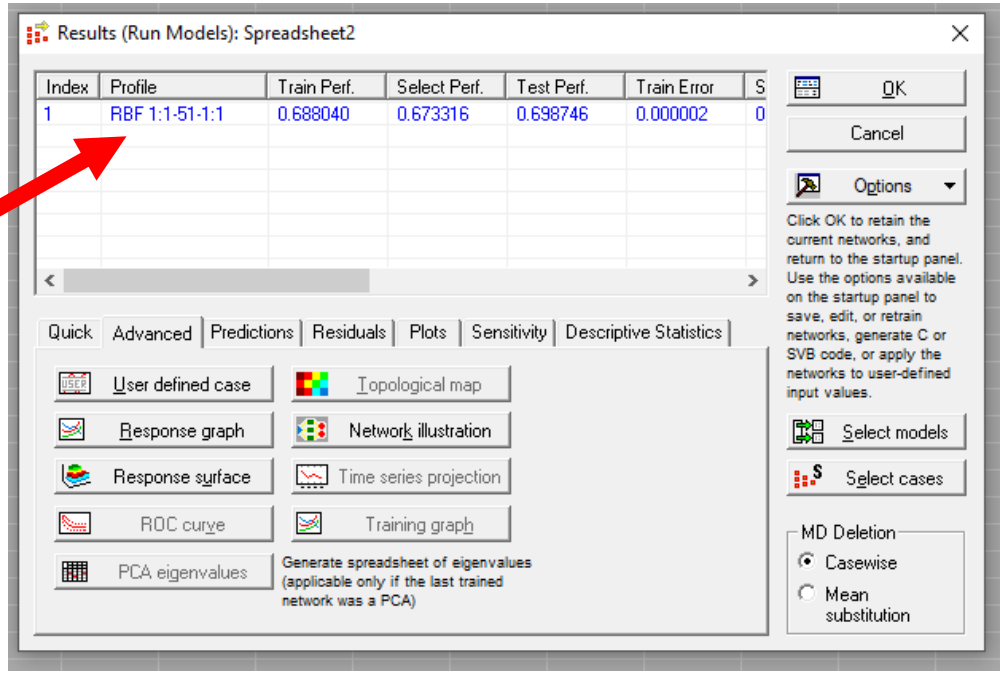
	Profile	Train Perf.	Select Perf.	Test Perf.	Train Error	Select Error
1	MLP 1:1-2-1:1	0.681958	0.687984	0.691064	0.026967	0.026021
2	RBF 1:1-409-1:1	254.606219	265.322181	264.103459	0.000676	0.000674
3	Linear 1:1-1:1	0.713671	0.705273	0.717101	0.035276	0.033345
4	MLP 1:1-8-1:1	0.695083	0.685639	0.698460	0.027524	0.026006
5	RBF 1:1-204-1:1	205.869064	225.494873	216.830693	0.000546	0.000572
6	RBF 1:1-102-1:1	15.798132	16.419963	16.513513	0.000042	0.000042
7	RBF 1:1-51-1:1	0.688040	0.673316	0.698746	0.000002	0.000002

Resultados

Na sequência, o *software* irá selecionar, de maneira automática, a melhor rede. Ou seja, a que apresentou o menor erro (conforme parametrizado nas etapas anteriores).

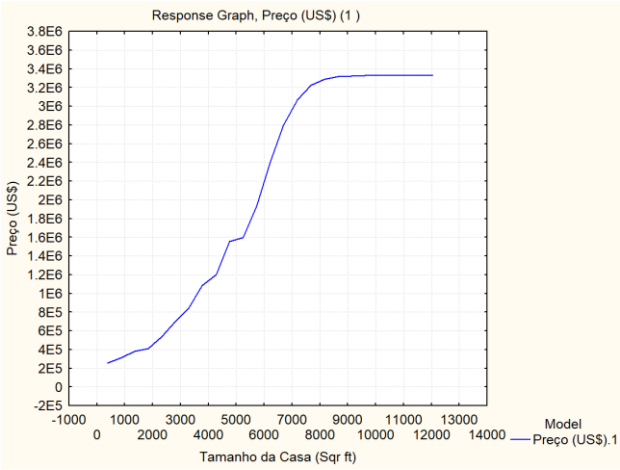
- Para este caso, verifica-se que a melhor rede foi a *"Radial Basis Function"*

RBF 1:-51-1:1



Resultados

Ainda nesta etapa, é possível verificar graficamente o resultado desses treinamentos, tais como:



Gráficos de superfície de resposta

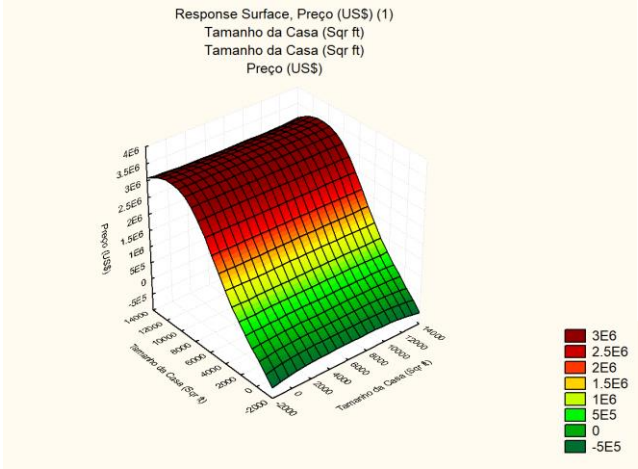
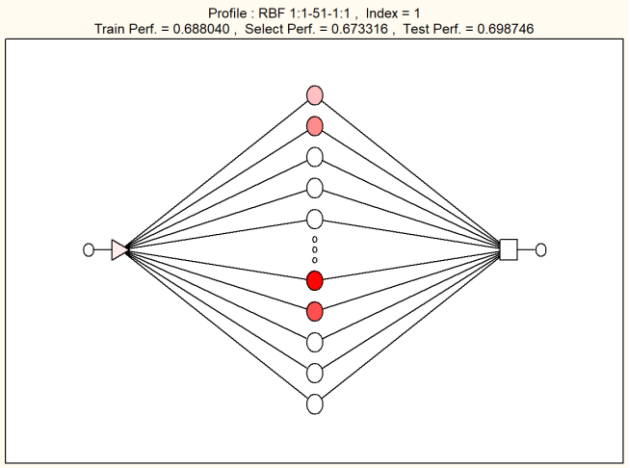
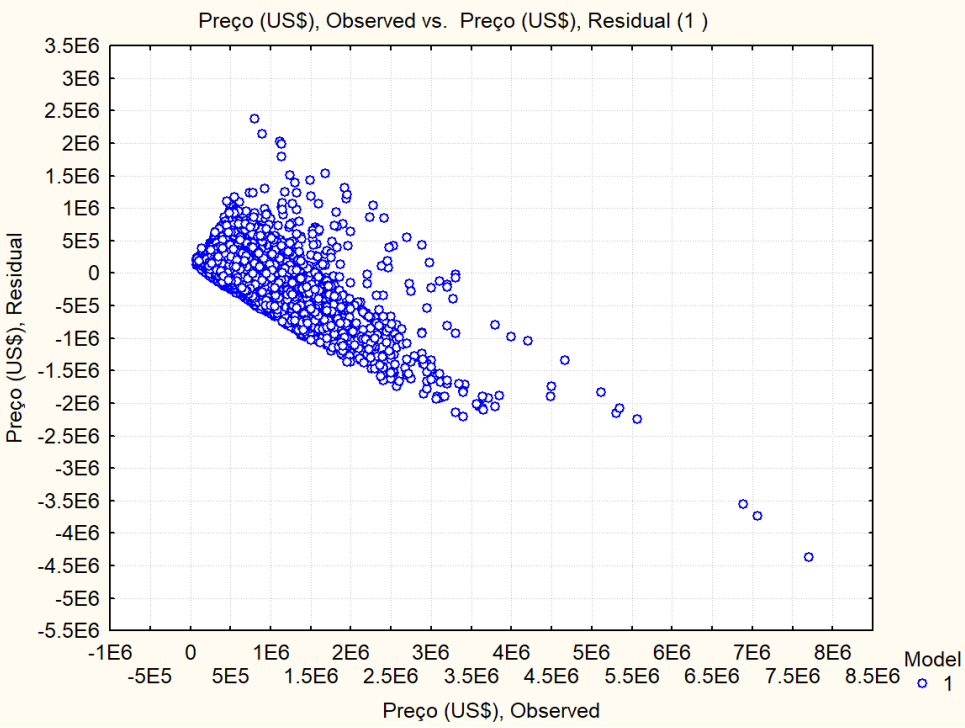


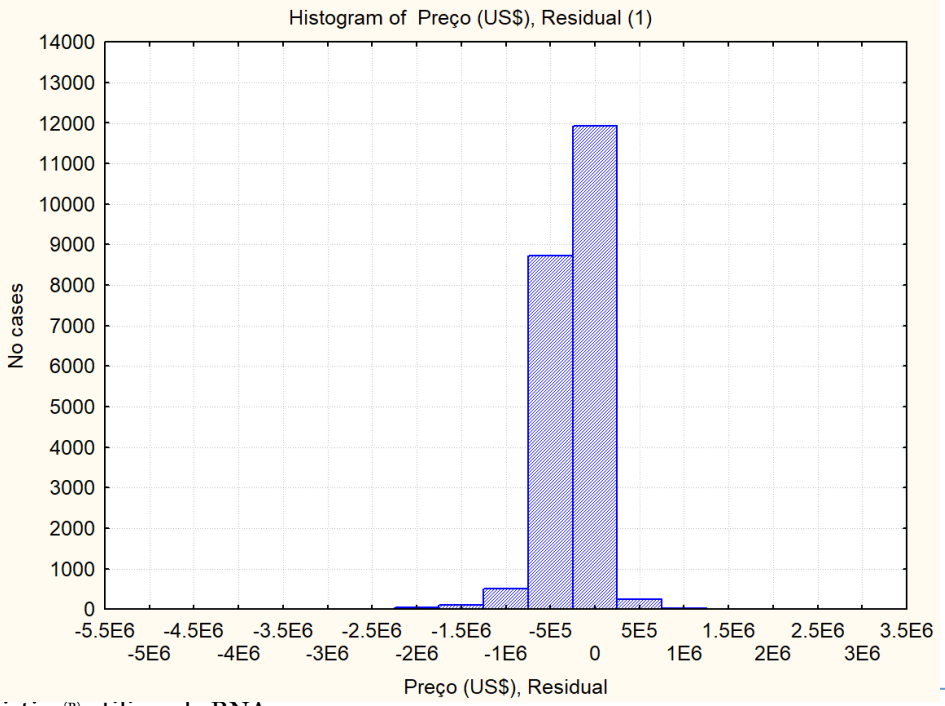
Ilustração da RNA





Observado x Resíduo

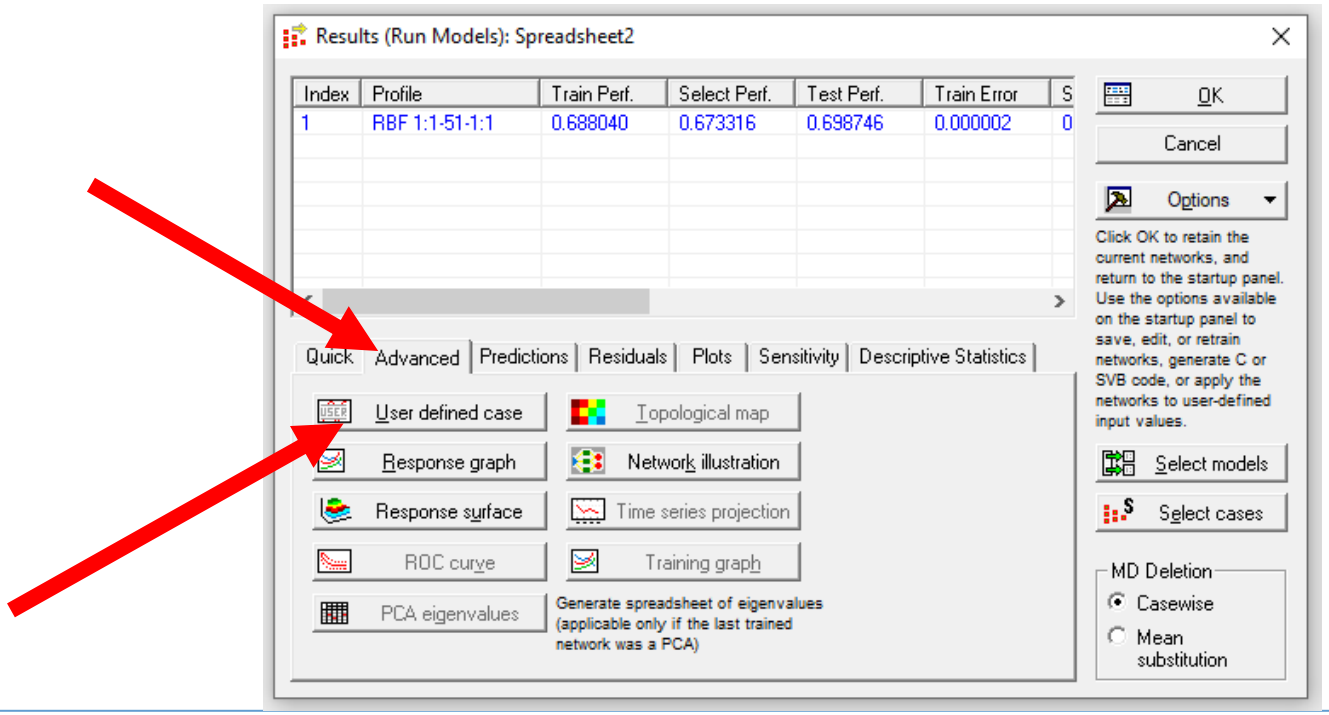
Histograma dos Resíduos



Validação

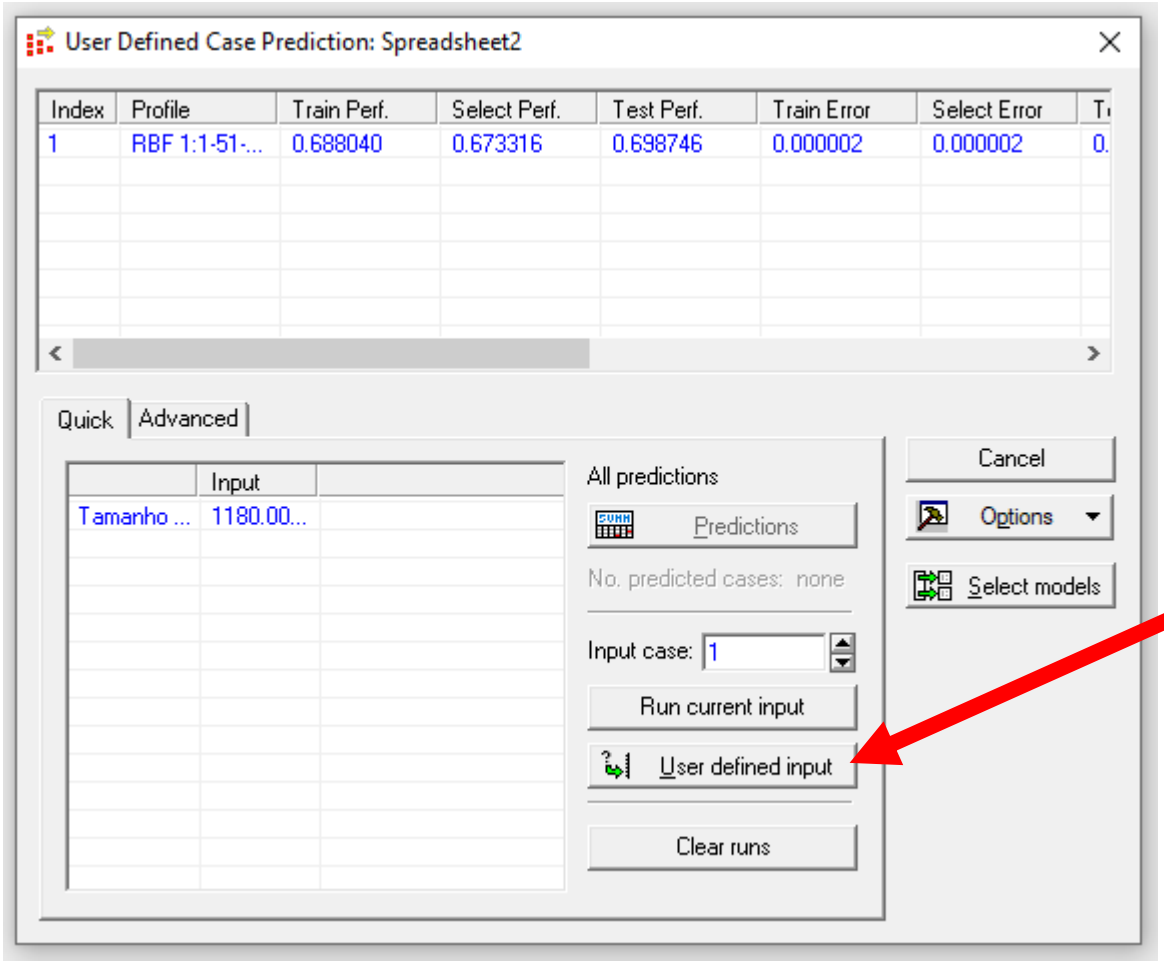
Para testar se o modelo está adequado, iremos utilizar as 20 linhas removidas no começo da aplicação, no qual já conhecemos o resultado. Assim, será possível verificar se a rede conseguirá prever adequadamente os valores aproximados.

Para isso, deve-se ir na aba *"Advanced"* e clicar em *"User defined case"*



Validação

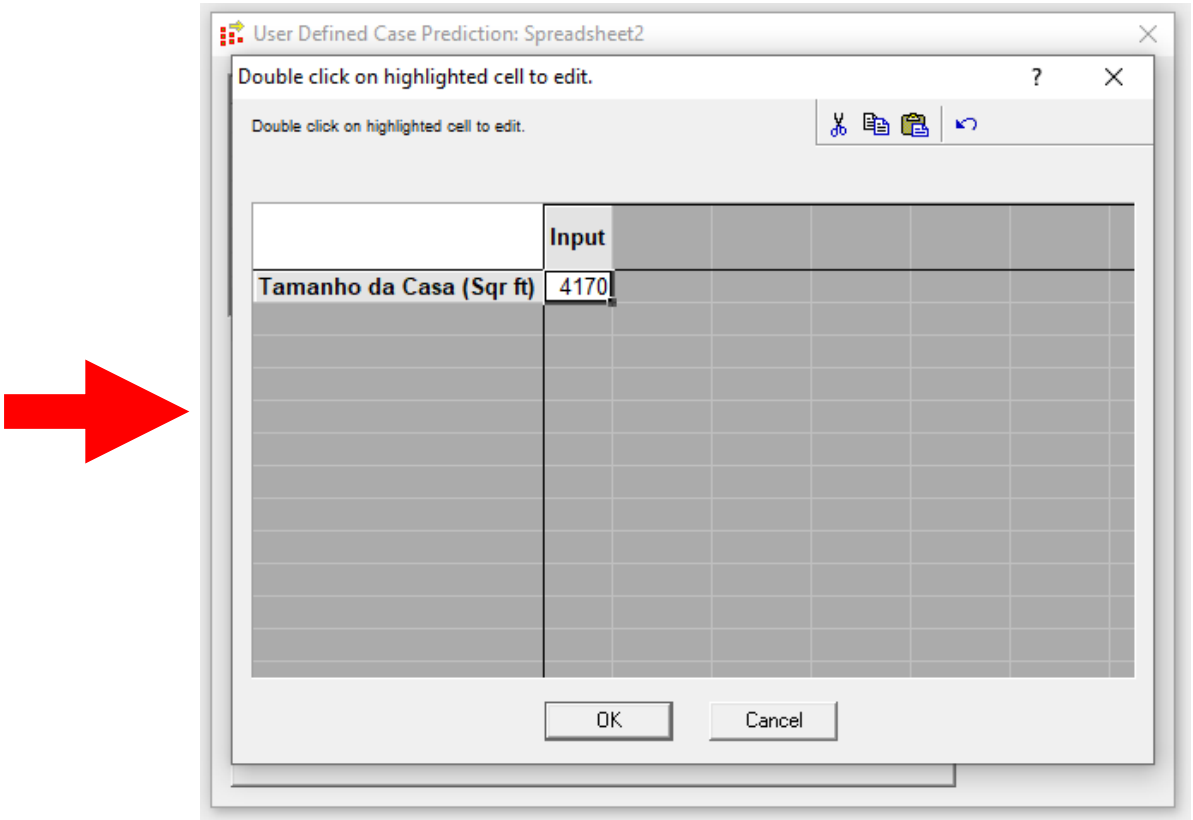
Uma nova janela irá abrir, no qual iremos clicar em "*User defined input*" para adicionar os valores da variável *x* (*Tamanho*) para prever o preço (*y*) pela RNA.



Validação

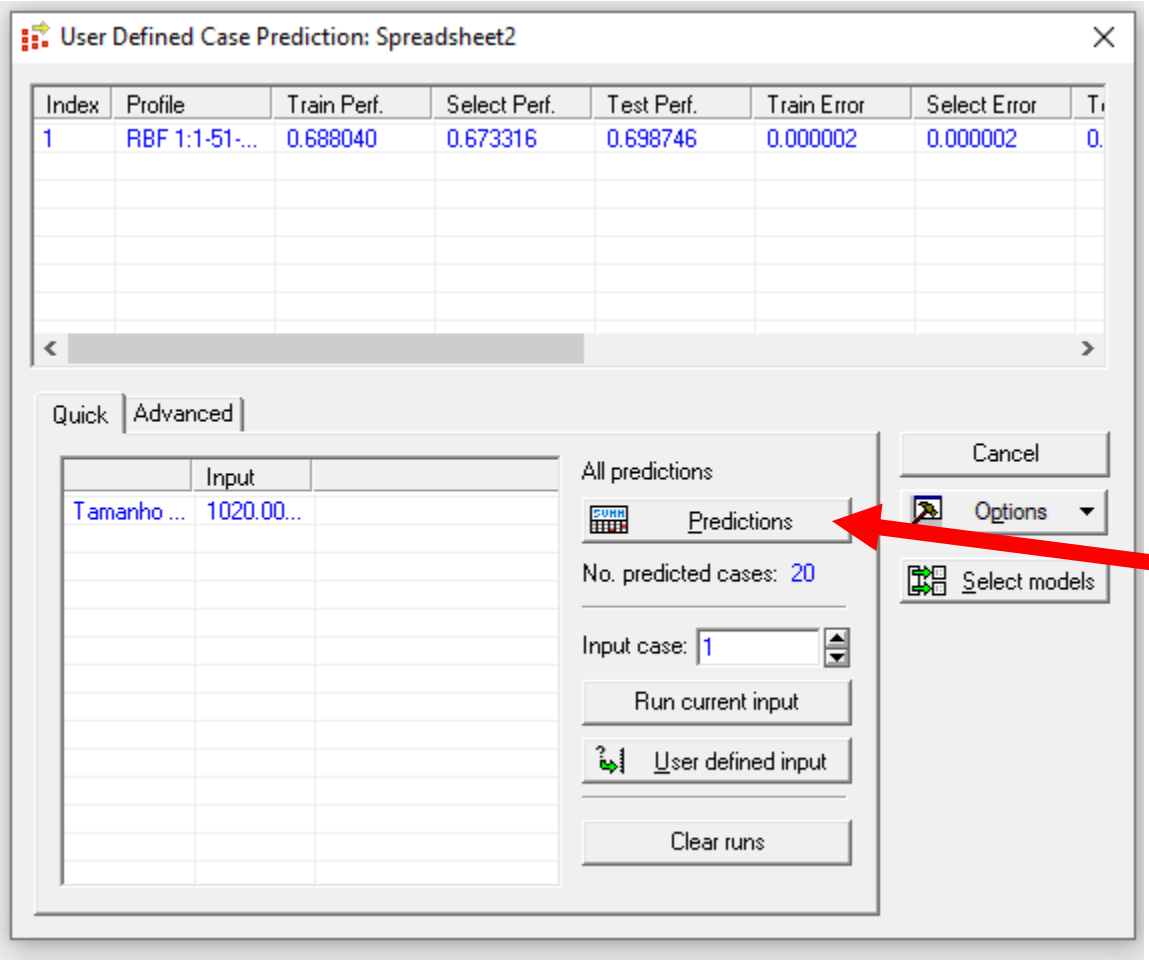
Assim, teremos uma nova planilha, onde é necessário computar os 20 valores de "*Tamanho*".

Validação		
Dados	Preço [*US$*]	Tamanho [ft^2]
1	$ 1,090,000.00	4170
2	$ 350,000.00	2500
3	$ 520,000.00	1530
4	$ 679,950.00	3600
5	$ 1,580,000.00	3410
6	$ 541,800.00	3118
7	$ 810,000.00	3990
8	$ 1,540,000.00	4470
...	...	...
17	$ 400,000.00	2310
18	$ 402,101.00	1020
19	$ 400,000.00	1600
20	$ 325,000.00	1020



Validação

Após adicionar todas as 20 variáveis de entrada (*Tamanho*), deve-se clicar em "*Predictions*".



Comparação via MAPE

O *Statistica*® retornará as 20 predições de preços realizadas pela RNA, com base nas informações de entrada, neste caso, o *Tamanho* (variável *x*)

	User Defined Case Prediction, (1)
	Preço (US$).1
1	1134988
2	593208
3	422671
4	939120
5	942063
6	784956
7	1166745
8	1427818
9	378622
10	401764
11	512765
12	395865
13	601675
14	954476
15	375589
16	422671
17	576272
18	311883
19	408545
20	311883

A partir destes valores é possível verificar o erro, comparando com os resultados de preço original. Para isso, pode-se utilizar o indicador conhecido como o "Erro Percentual Absoluto Médio" (*MAPE - Mean absolute percentage error*)

$$MAPE(\%) = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{y_t - \widehat{y_t}}{y_t}\right| \times 100$$

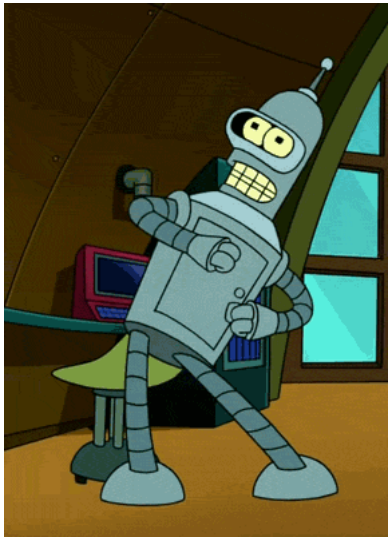
Conclusão

Para esses resultados, tem-se que a RNA apresentou um *MAPE* de *24,61%*.

Este resultado não é o ideal, pois queremos o erro com o menor valor possível.

Contudo, ao realizar uma regressão simples com o mesmo conjunto de dados, tem-se um *MAPE* de *30,13%*.

Ou seja, o erro percentual absoluto médio da RNA do tipo "*Radial Basis Function*" **é menor**, para um problema de regressão, ao comparar com uma métrica de regressão usual, como os mínimos quadrados ordinários.



Obrigado!