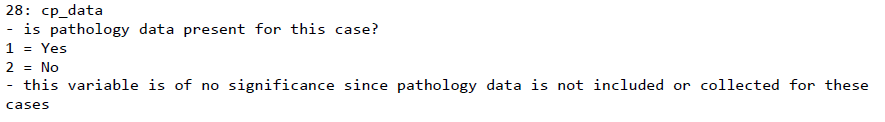
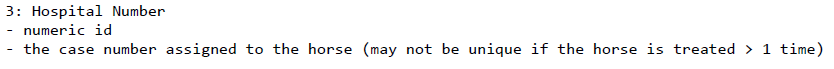
## Análise Exploratória

### Análise do dicionário de dados:

* ***cp\_data***: pode ser desconsiderado.

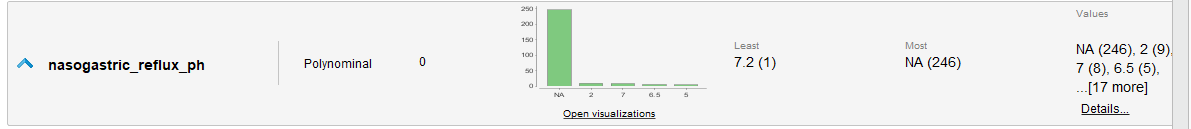


* ***hospital\_number***: identificador do atendimento no hospital, deve ser removido.

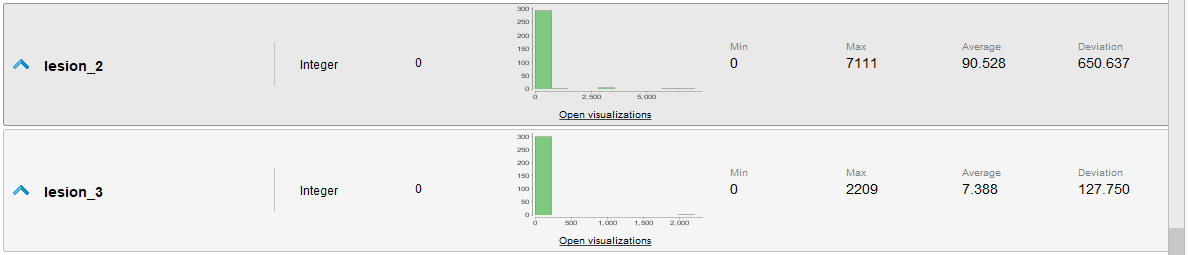


### Análise utilizando o RapidMiner:

* ***Cnasogastric\_reflux\_ph***: pode ser desconsiderado, pois tem 246 valores faltantes ‘NA’. O valor ‘NA’ foi reconhecido como valor válido e não como valor faltante pelo RapidMiner. Deve ser corrigido no pré-processamento.

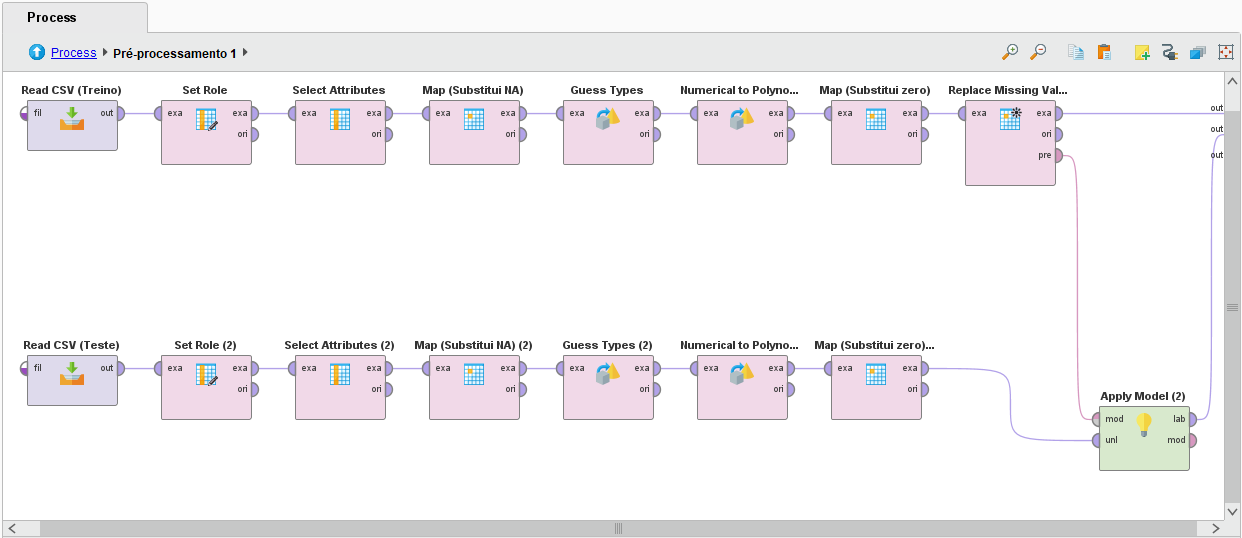


* ***lesion\_2 e lesion\_3***: podem ser removidos, pois têm muitos valores faltantes ‘0’. O valor ‘0’ foi reconhecido como valor válido e não como valor faltante pelo RapidMiner. Deve ser corrigido no pré-processamento para o atributo ***lesion\_1***.

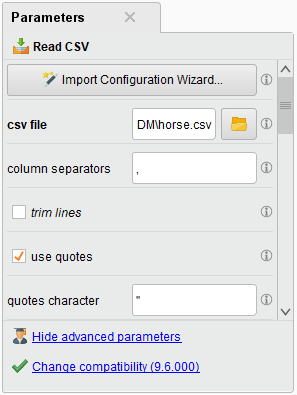


## Primeira abordagem para o Pré-processamento – Pré-processamento 1

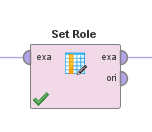
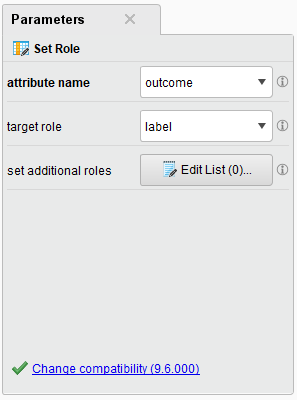
* Carregar a base.
* Definir a coluna ***outcome*** como *label.*
* Remover os atributos abaixo:
  + ***cp\_data***: atributo não relevante para o modelo.
  + ***hospital\_number***: identificação do atendimento.
  + ***lesion\_2, lesion\_3 e nasogastric\_reflux\_PH***: muitos valores faltantes.
* Substituir os valores ‘NA’ por valores faltantes.
* Corrigir o tipo dos atributos.
* Substituir os valores ‘zero’ do atributo ***lesion\_1*** por valores faltantes.
* Converter o atributo ***lesion\_1*** de inteiro para nominal, para que ao fazer o processamento de valores faltantes, seja utilizado o valor mais frequente e não a média.
* Substituir valores faltantes.



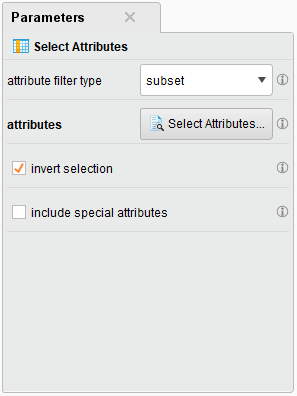
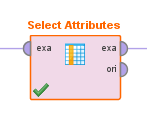
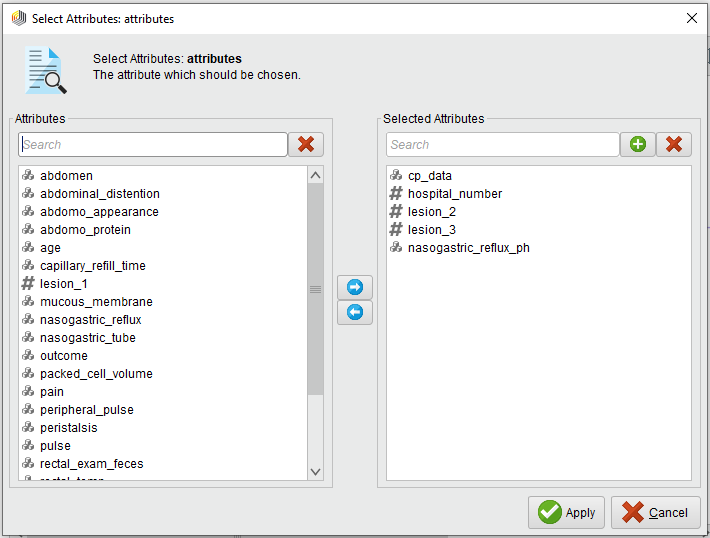
### Read.csv – Importar o arquivo ‘horse.csv’:

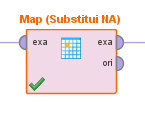
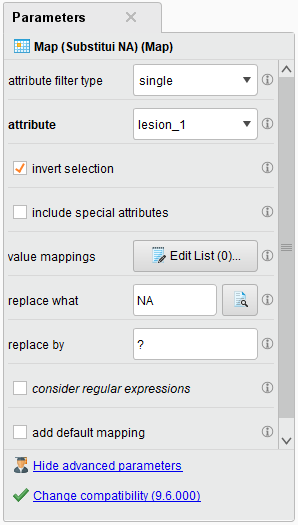
### Set Role – Definir a coluna ***outcome*** como *label:*

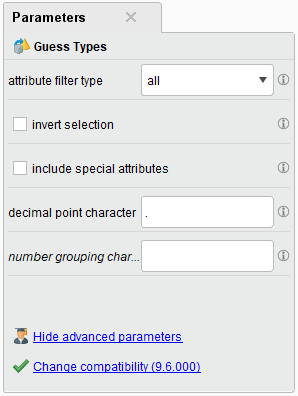
### Select Attributes – Remover atributos:

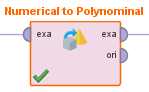
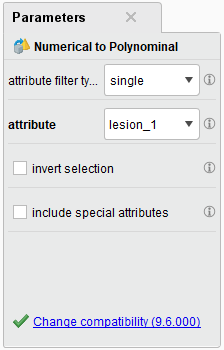
### Map – Substituir valores ‘NA’ por valores faltantes ‘?’:

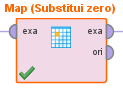
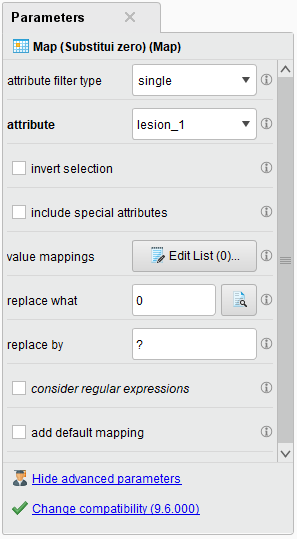
### Guess Types – Corrigir o tipo dos atributos

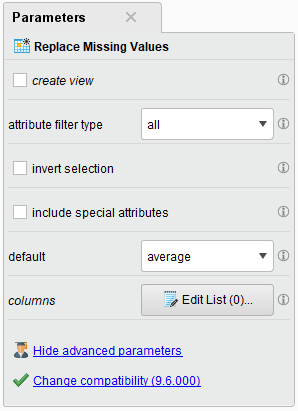
### Numerical to Polynominal – Converter o atributo “lesion\_1” de inteiro para nominal

### Map - Substituir os valores “0” do atributo “lesion\_1” por valores faltantes “?” (missing values)

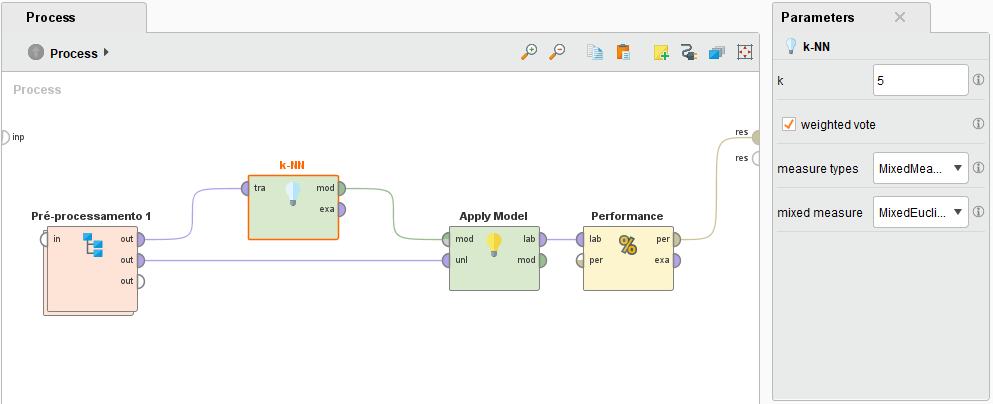
 

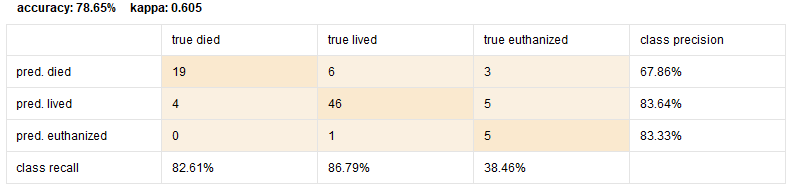
### Replace Missing Values – Substituir os valores numéricos pela média e os nominais pelo valor mais frequente:

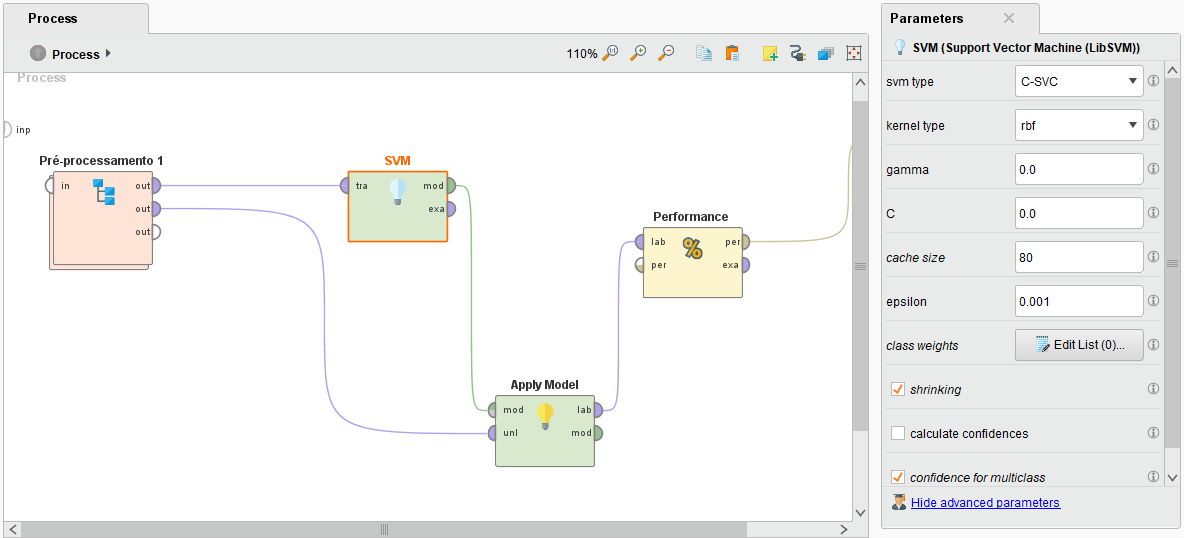
## Aplicação dos Modelos com configurações default, utilizando o Pré-processamento 1

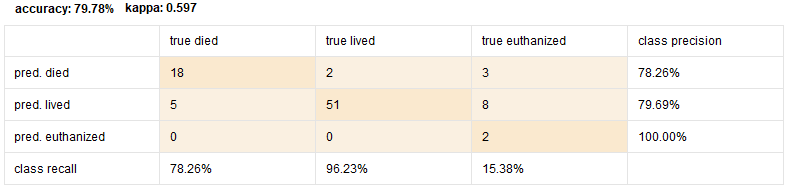
### k-NN (k-nearest neighbors)



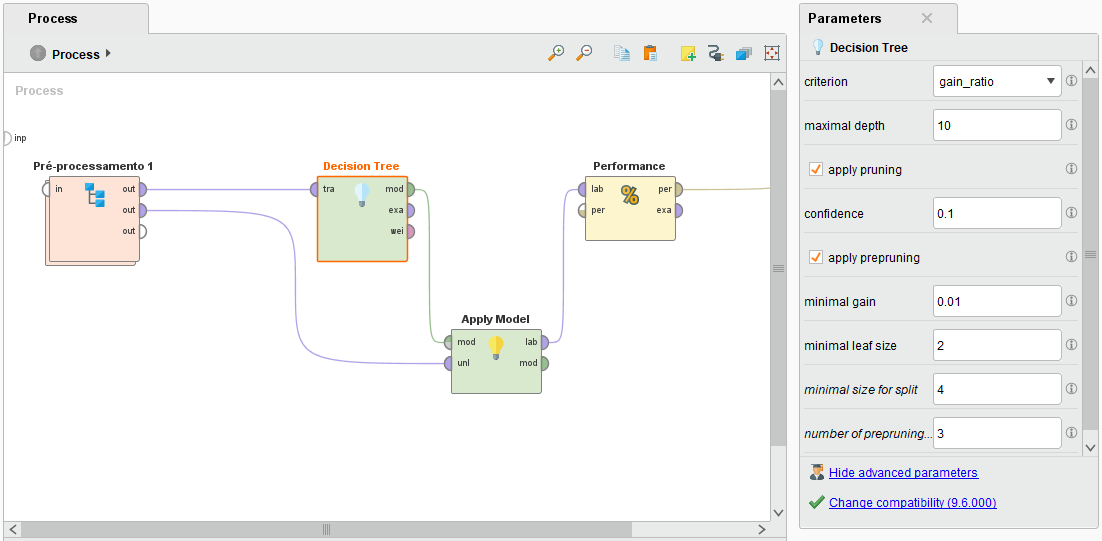


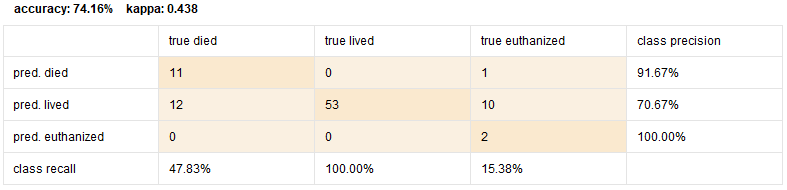
### SVM (support vector machine) – LibSVM (multiclass classification)



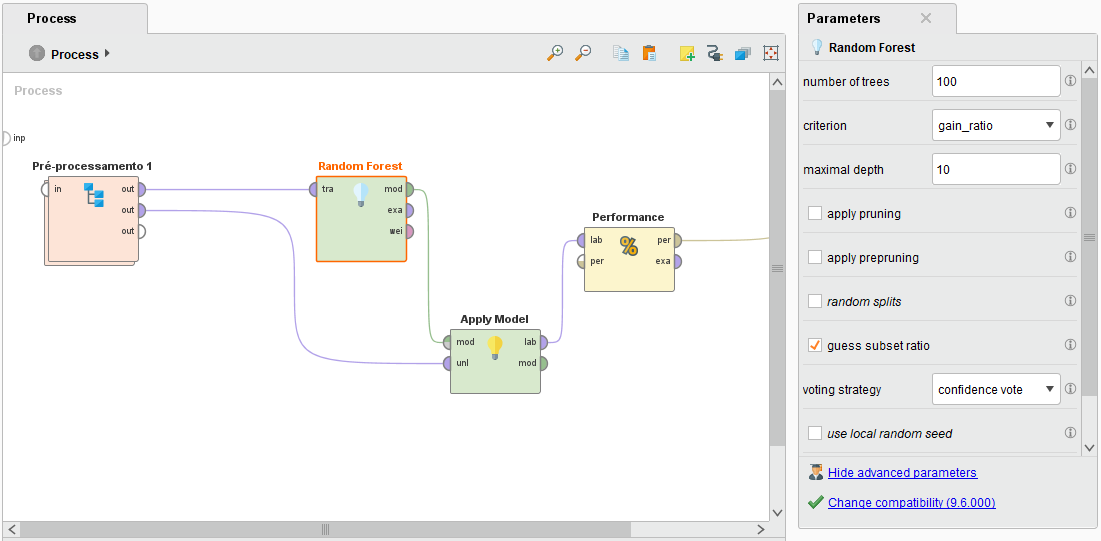


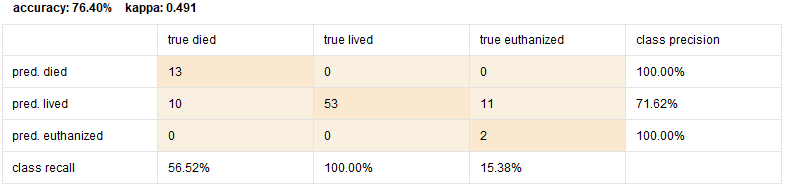
### Decision Tree





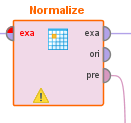
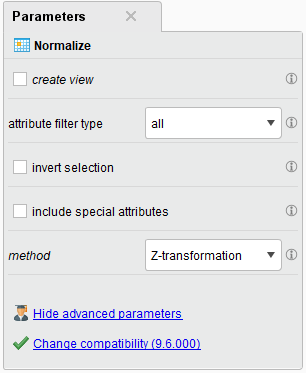
### Random Forest

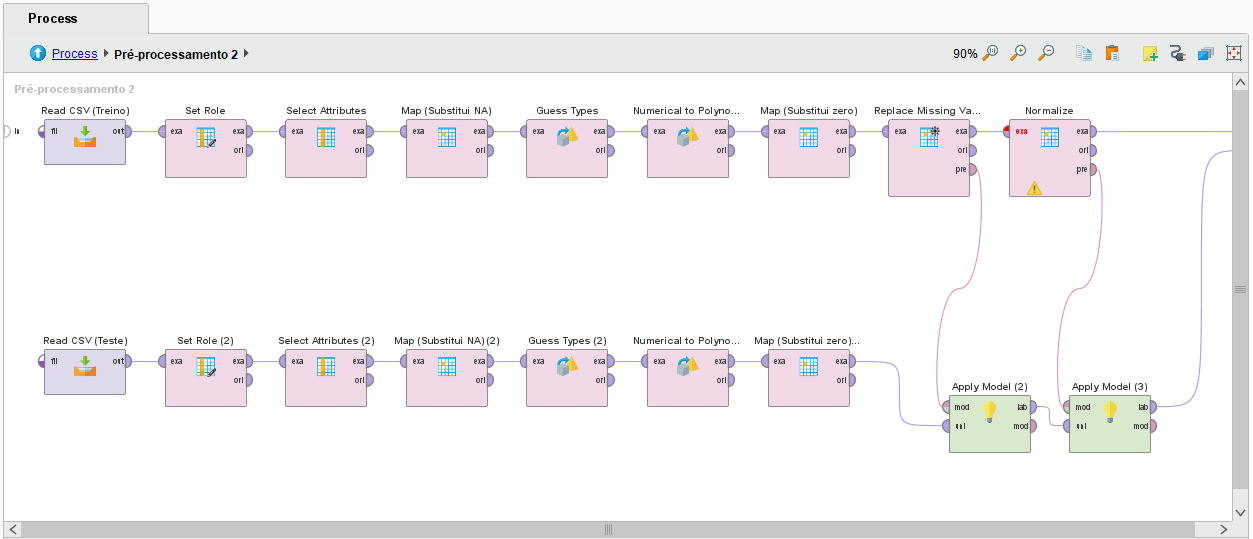




## Segunda abordagem para o Pré-processamento – Pré-processamento 2

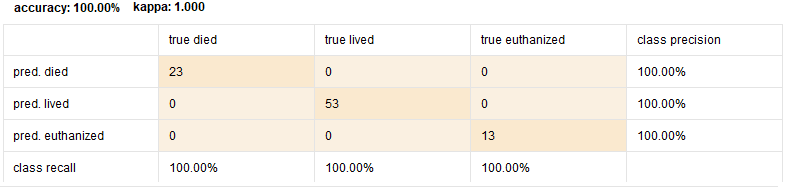
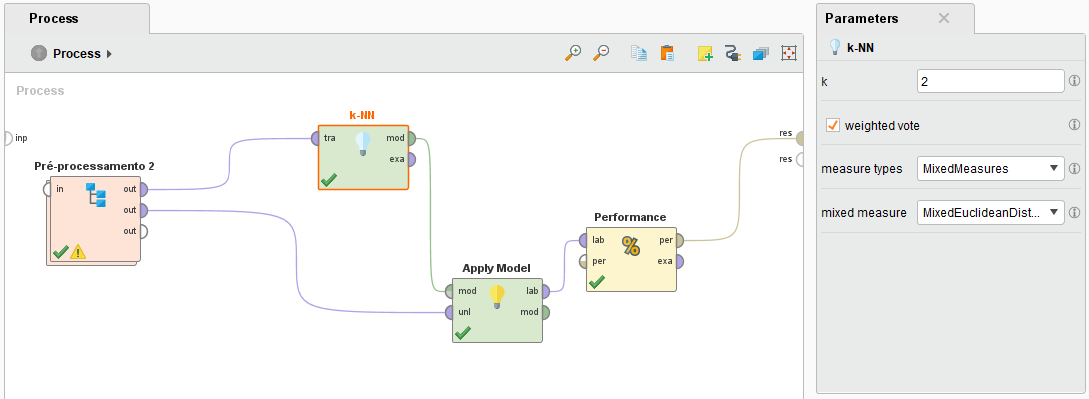
Todos os modelos apresentam valores baixos para a acurácia e o kappa, mostrando que é preciso melhorar o pré-processamento dos dados. Será incluída uma etapa de normalização da base.

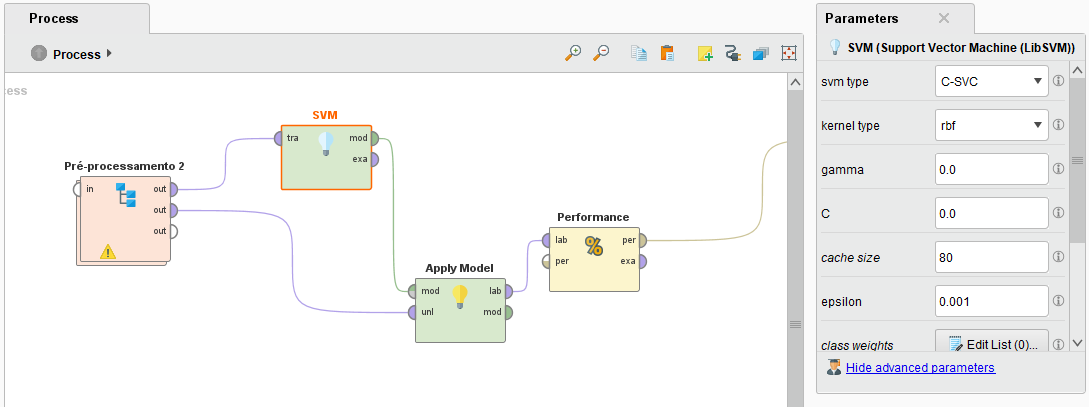


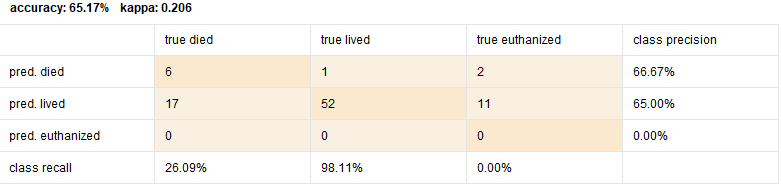
## Aplicação dos algoritmos utilizando o Pré-processamento 2

### K-NN (k-nearest neighbors) – k = 2

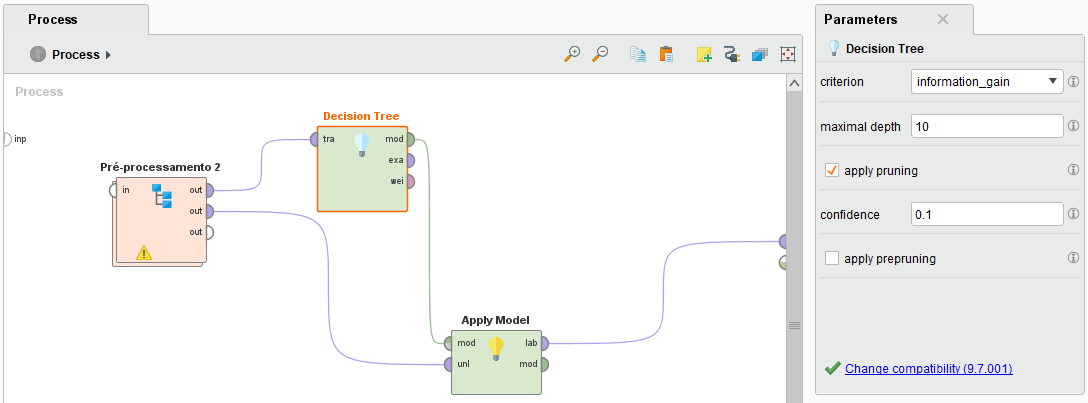


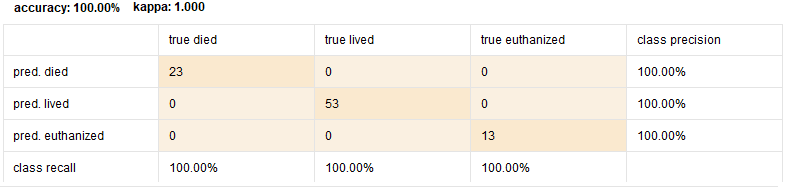
### SVM (support vector machine) – LibSVM (multiclass classification)





### Decision Tree - critério “information\_gain” e sem “prepruning”





### Random Forest

