Relatório de Data Mining

Trabalho de conclusão da disciplina

Professora Manoela Kohler

Horse Colic Data Set

Rodrigo Vidigal

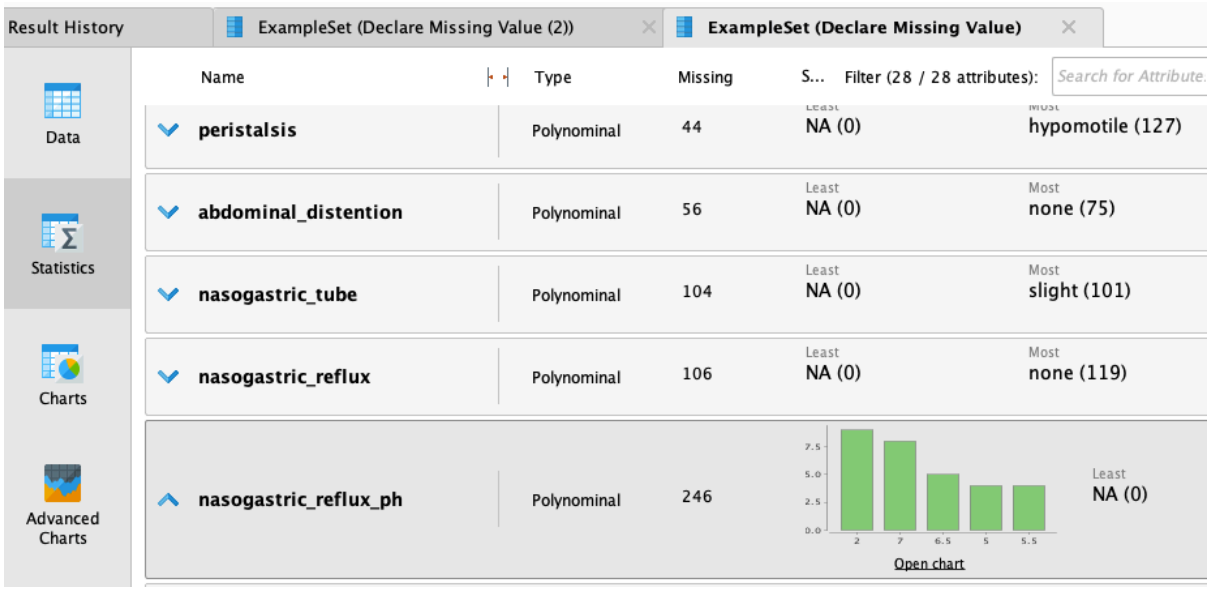
15/01/2019

Análise exploratória, missing values e atributos desnecessários

Uma primeira analisada nos dados para entender quais atributos manter no modelo

Após uma primeira olhada nos dados, o que me chamou a atenção foi a quantidade de atributos com um número expressivo de missing values, como o atributo *nasogastric_reflux_ph,* com 246 exemplos com valores faltantes dentro de um total de 299(inicialmente eles estavam como NA, mas com o operador

Declare Missing Values eu corrigi isso para que o RapidMiner os entendesse corretamente como Missing Values). Um atributo com tantos valores faltantes assim certamente não ajudaria muito o meu modelo, por isso exclui este atributo (*nasogastric_reflux_ph*).



Atributo nasogastric_reflux_ph: 246 missing values

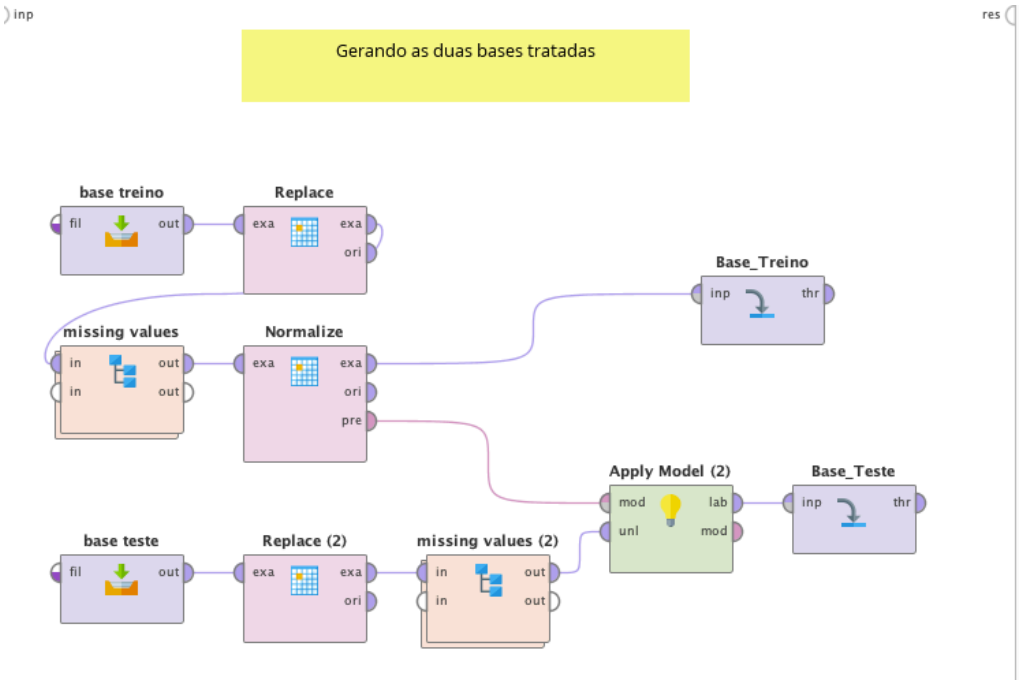
Excluí também os atributos com muitos exemplos com o mesmo valor, como o *lesion 2* que continha apenas 7 exemplos diferentes de 0 e o *lesion 3* com apenas 1 exemplo diferente de 0.

O atributo *cp_data* foi excluído pois no próprio dicionário do *dataset* consta que ele não tem importância.

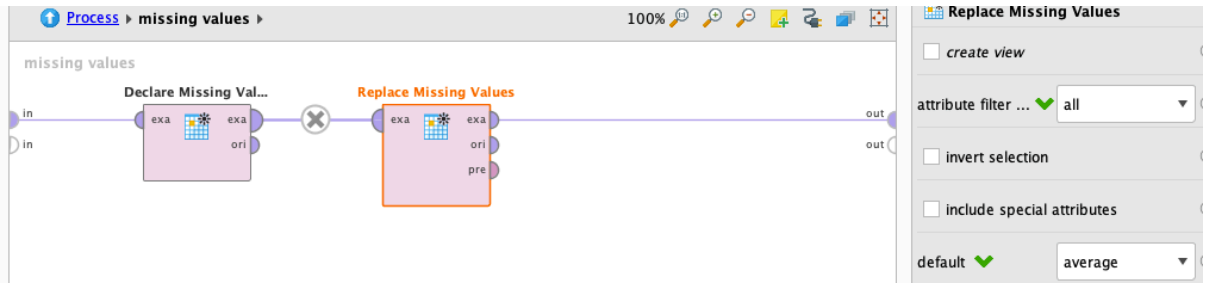
```
28: cp_data
- is pathology data present for this case?
1 = Yes
2 = No
- this variable is of no significance since pathology data is not included or collected for these
cases
```

Balanceamento dos Dados e processo para gerar as duas bases tratadas

Seguindo com o trabalho, fiz um processo no RapidMiner para gerar as duas bases, a de treino e a de teste, e salvar num repositório com o operador *Store*. Dentro deste processo eu normalizei os dados para o modelo não privilegiar nenhum atributo específico. Outra coisa que fiz nesta etapa do trabalho, foi substituir a classe *Euthanized* por *Died* com o operador *Replace*, pois caso eu fosse utilizar algum modelo de aprendizado que suportasse apenas um rótulo com duas classes(o SVM, por exemplo), eu não teria problemas quanto a isso. E os exemplos com o valor *NA* foram declarados como Missing Values e substituídos pela média do atributo em questão.



Abaixo, o subprocesso de Missing Vlaues.



Reavaliação dos atributos pelo peso por Qui Quadrado e ganho de informação

Já com as duas bases tratadas, a de treino e a de teste, eu resolvi utilizar dois métodos para entender a relevância de cada atributo em relação ao rótulo a ser classificado, e posteriormente após uma análise desses resultados, decidir quais atributos eliminar para finalmente aplicar os algoritmos de aprendizado.

Ganho info treino

attribute	weight
nasogastric_tube	0.002
age	0.002
hospital_number	0.010
surgery	0.021
rectal_temp	0.022
total_protein	0.026
respiratory_rate	0.030
nasogastric_reflux	0.031
rectal_exam_feces	0.042
abdomen	0.044
abdomo_appearance	0.049
surgical_lesion	0.068
capillary_refill_time	0.073
abdominal_distention	0.075
peristalsis	0.075
peripheral_pulse	0.084
temp_of_extremities	0.097
mucous_membrane	0.109

Ganho info teste

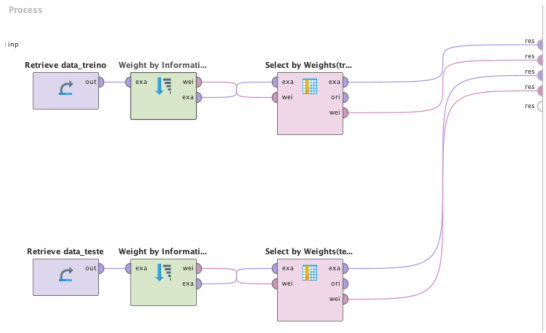
attribute	weight
hospital_number	0
nasogastric_tube	0.012
surgery	0.014
age	0.014
abdomo_appearance	0.023
total_protein	0.023
peristalsis	0.026
rectal_exam_feces	0.026
respiratory_rate	0.027
rectal_temp	0.032
abdomen	0.060
nasogastric_reflux	0.066
surgical_lesion	0.068
capillary_refill_time	0.076
mucous_membrane	0.115
peripheral_pulse	0.126
temp_of_extremities	0.134
abdomo_protein	0.140

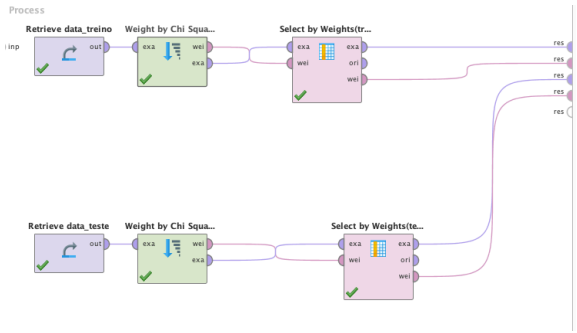
Qui Quadrado treino

attribute	weight
nasogastric_tube	0.788
age	0.984
hospital_number	1.080
surgery	8.564
nasogastric_reflux	13.104
rectal_temp	15.185
abdomen	16.315
rectal_exam_feces	16.814
respiratory_rate	18.205
abdomo_appearance	20.125
total_protein	20.258
surgical_lesion	26.707
peristalsis	29.129
capillary_refill_time	30.360
abdominal_distention	31.032
peripheral_pulse	34.443
temp_of_extremities	37.484
abdomo_protein	41.728

Qui Quadrado teste

attribute	weight
nasogastric_tube	1.497
surgery	1.690
age	1.788
abdomo_appearance	2.620
peristalsis	2.933
rectal_exam_feces	3.169
abdomen	5.918
total_protein	5.981
respiratory_rate	6.922
surgical_lesion	8.058
nasogastric_reflux	8.177
capillary_refill_time	9.386
rectal_temp	9.760
abdomo_protein	12.675
peripheral_pulse	13.789
mucous_membrane	13.969
temp_of_extremities	14.810
hospital_number	15.254



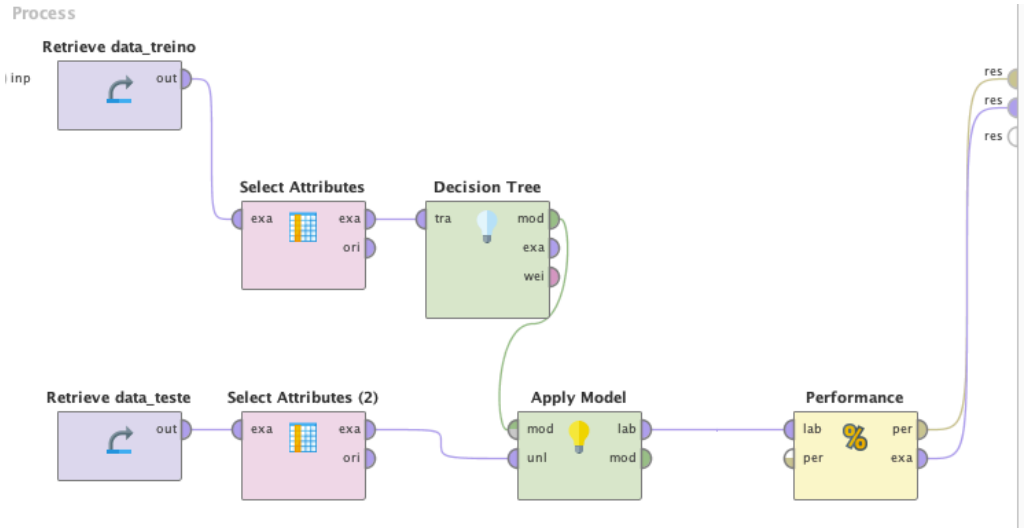


Avaliando os resultados, decidi excluir os atributos: nasogastric_tube, hospital_number e age.

Testando diferentes modelos

Testei diferentes modelos utilizando os seguintes algoritmos de classificação: Decision Tree, KNN, Naive Bayes, SVM e Random Forest. Abaixo, seguem as imagens dos processos no RapidMiner e suas respectivas acurácias:

Decision Tree:



accuracy: 76.40%

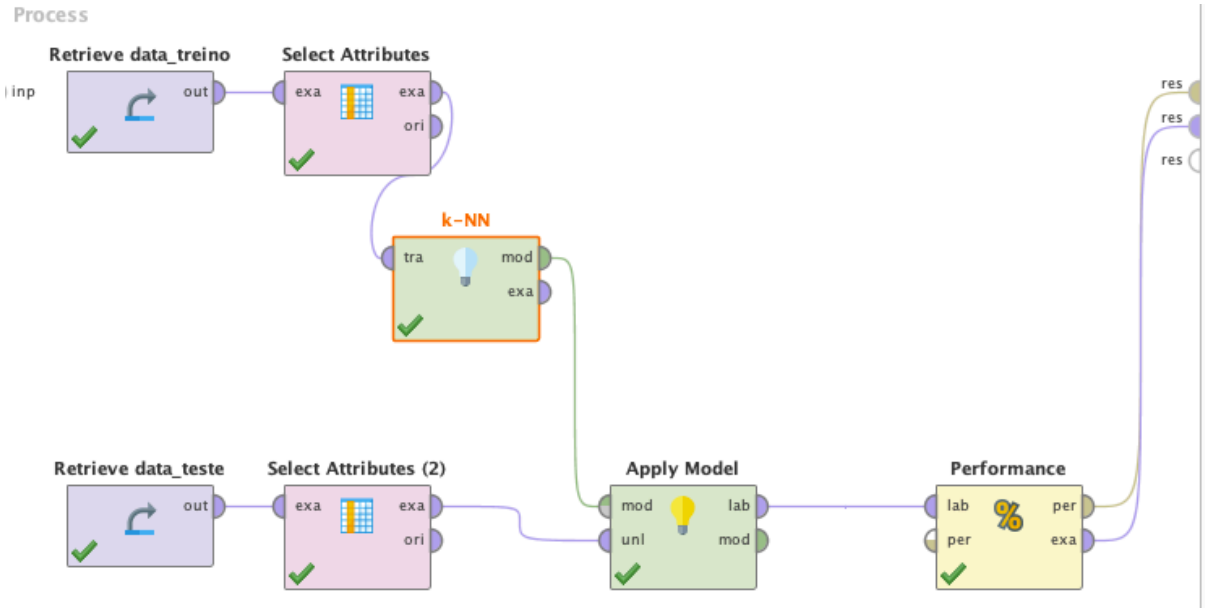
	true died	true lived	class precision
pred. died	15	0	100.00%
pred. lived	21	53	71.62%
class recall		100.00%	

pred. lived

Kappa: 0.460

KNN:

(Com K = 3)



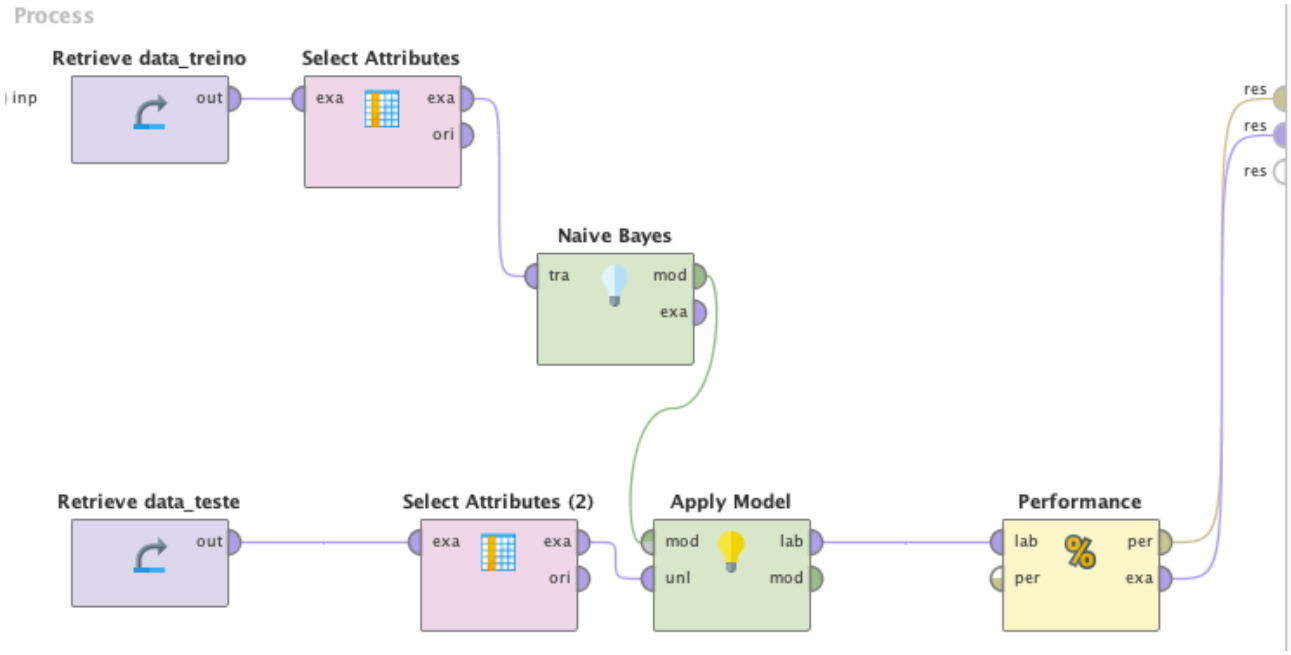
accuracy: 95.51%

	true died	true lived	class precision
pred. died	33	1	97.06%
pred. lived	3	52	94.55%
class recall	91.67%	98.11%	

true died

Kappa: 0.906

Naive Bayes:



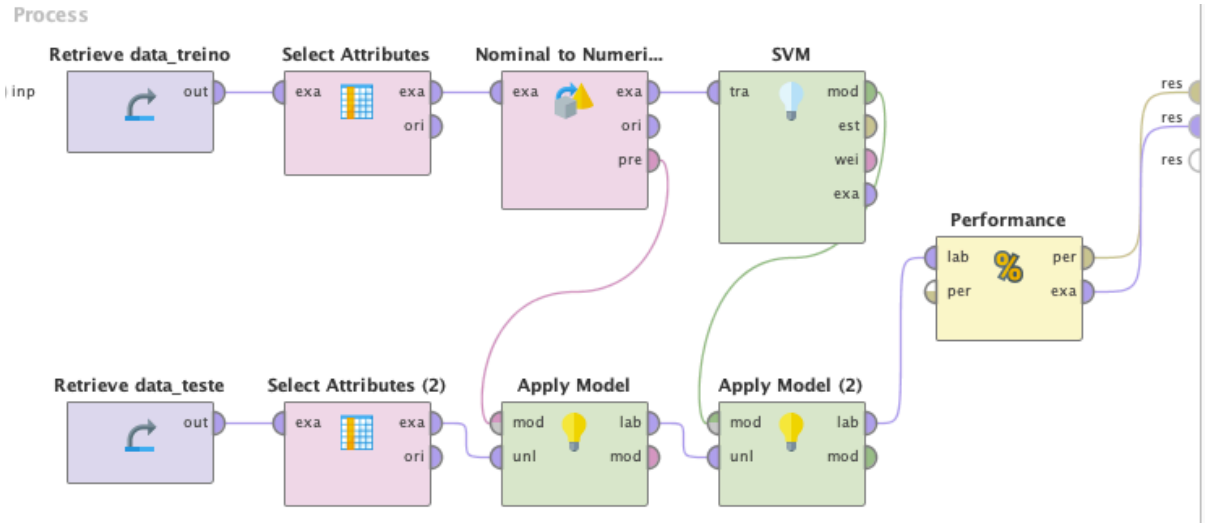
accuracy: 86.52%

	true died	true lived	class precision
pred. died	30	6	83.33%
pred. lived	6	47	88.68%
class recall	83.33%	88.68%	

Kappa: 0.720

SVM:

No SVM eu tive de transformar os atributos nominais para numéricos com o operador *Nominal to Numerical*.

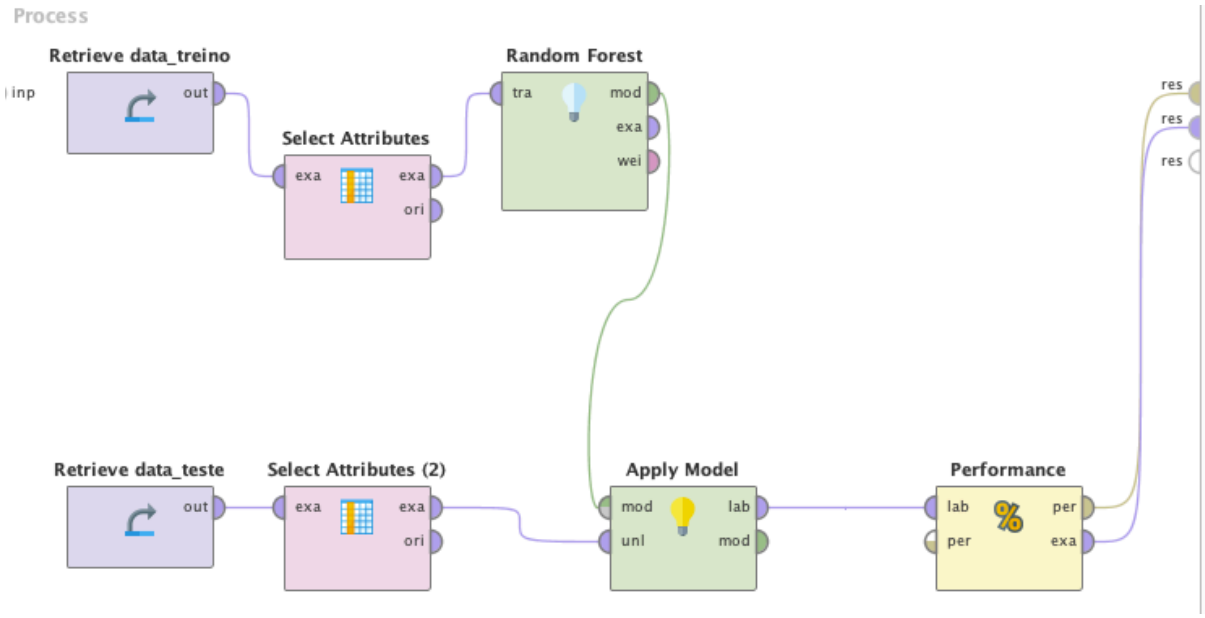


accuracy: 87.64%

	true died	true lived	class precision
pred. died	29	4	87.88%
pred. lived	7	49	87.50%
class recall	80.56%	92.45%	

Kappa: 0.740

Random Forest:



accuracy: 100.00%

	true died	true lived	class precision
pred. died	36	0	100.00%
pred. lived	0	53	100.00%
class recall	100.00%	100.00%	

Kappa: 1.000

Conclusão

O modelo com a melhor acurácia foi o Random Forest com 100% de precisão e Kappa de 1.000. O KNN, quando coloquei o número de K igual a 2 ele obteve o mesmo resultado, mas mantive K=3 pois o resultado já estava bastante satisfatório. O que obteve o pior rendimento foi o Decision Tree com acurácia de 76,40 % e Kappa de 0.460.

Considerações finais

O trabalho foi uma jornada de muito aprendizado. Primeiro pois revi o material de todas as aulas e li um livro sobre o tema, o "Data Science para Negócios", livro que me ajudou muito a entender um pouco mais dos modelos e como eles funcionam. Confesso que queria muito ter feito o trabalho em Python(tenho estudado Python pois começo agora no dia 16 um curso na Udacity de Data Science que tem como pré requisito esta linguagem e algumas de suas bibliotecas como NumPy, Pandas, MatplotLib e SciKitLearn), mas deixei para começar o trabalho apenas no final de dezembro, então utilizei o RapidMiner para me poupar tempo, visto que demoraria mais para realizá-lo em Python. Mas ainda assim, Data Mining é uma disciplina que muito me instiga a aprender mais, por uma curiosidade minha nata, e pelas infindáveis aplicabilidades que vislumbro neste nosso mundo contemporâneo que se produz cada vez mais e mais dados. Sou muito grato por tudo que venho aprendendo neste curso. Muito obrigado por tudo Manoela.