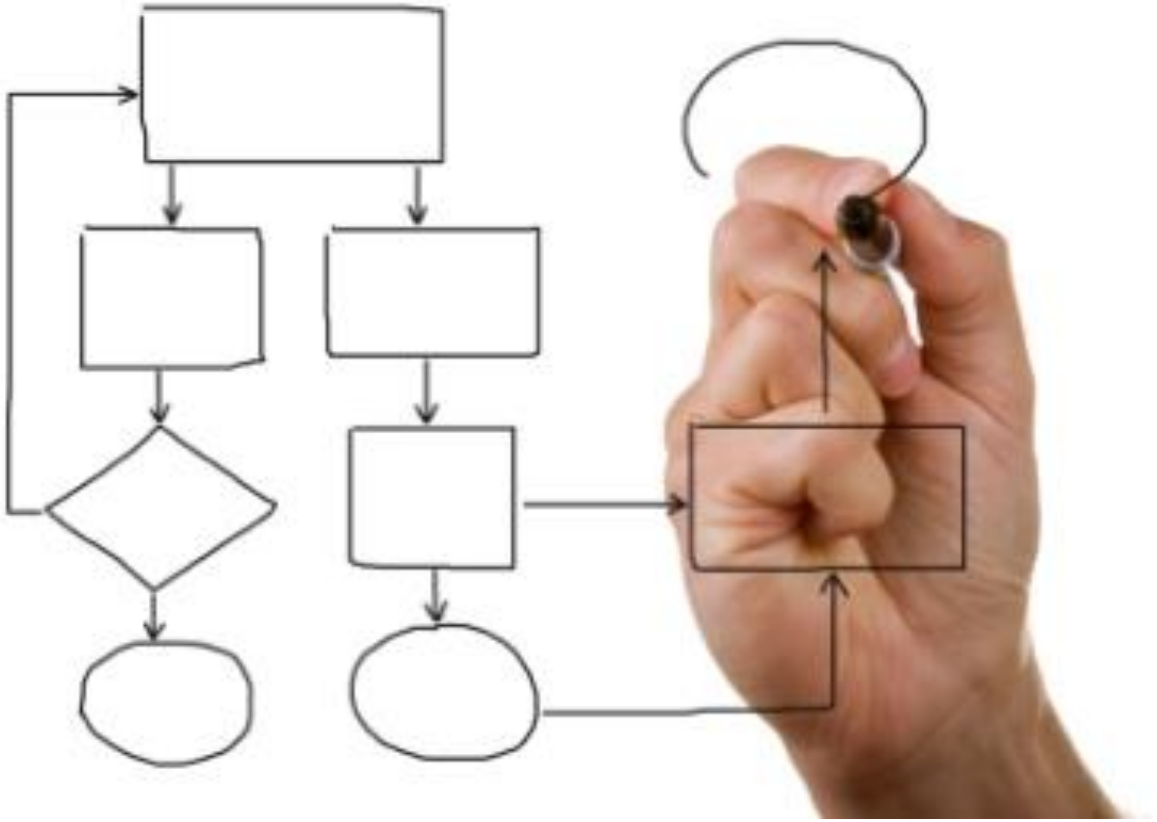


Business Intelligence

CRISP-DM: Data Mining e Modelos Preditivos

Prof. Leandro Guerra

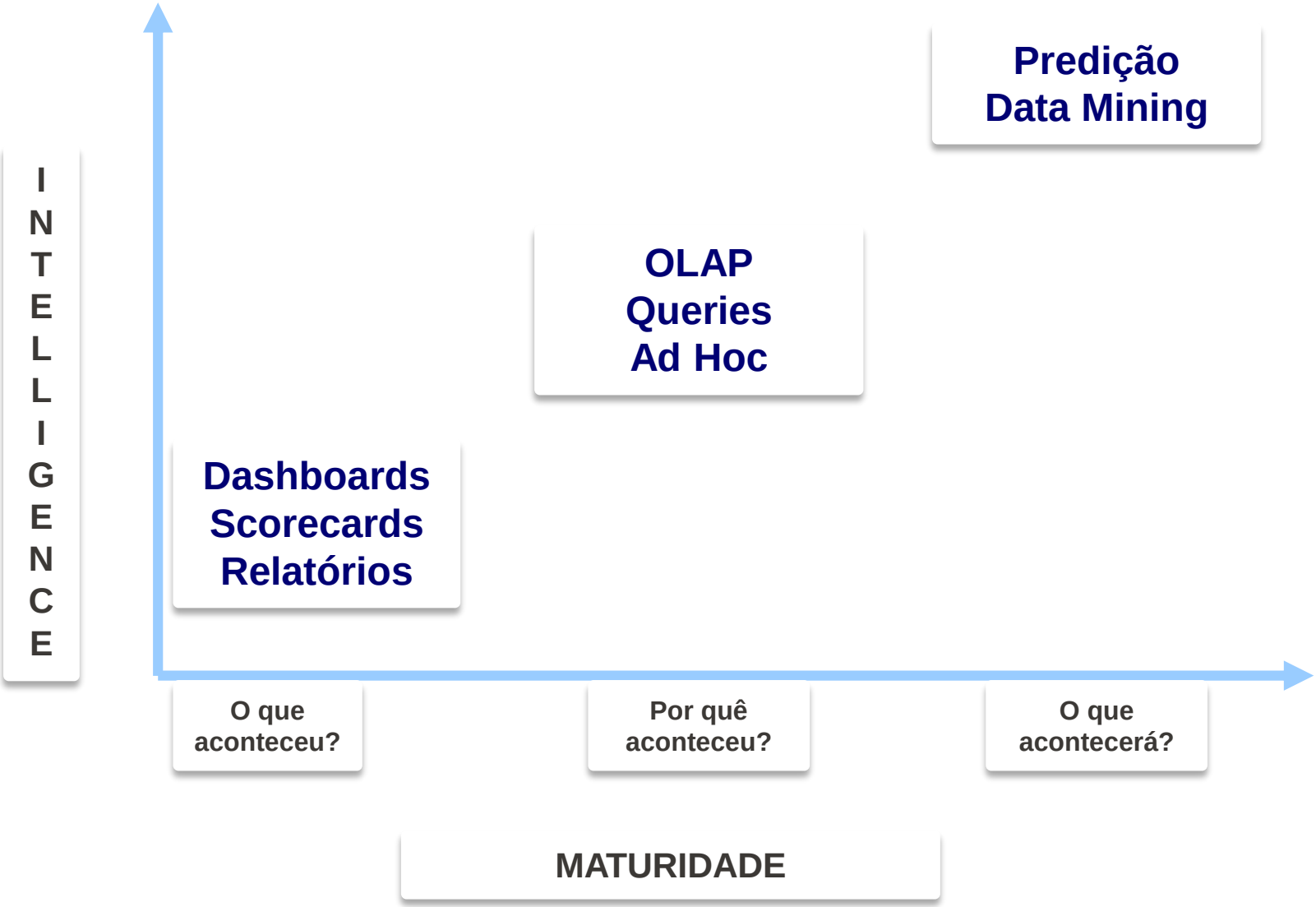
E-mail: leandro.guerra@outspokenmarket.com.br **IG: @leandrowar**

R – Árvore de Decisão

Gephi

Business Intelligence
Nível de Maturidade

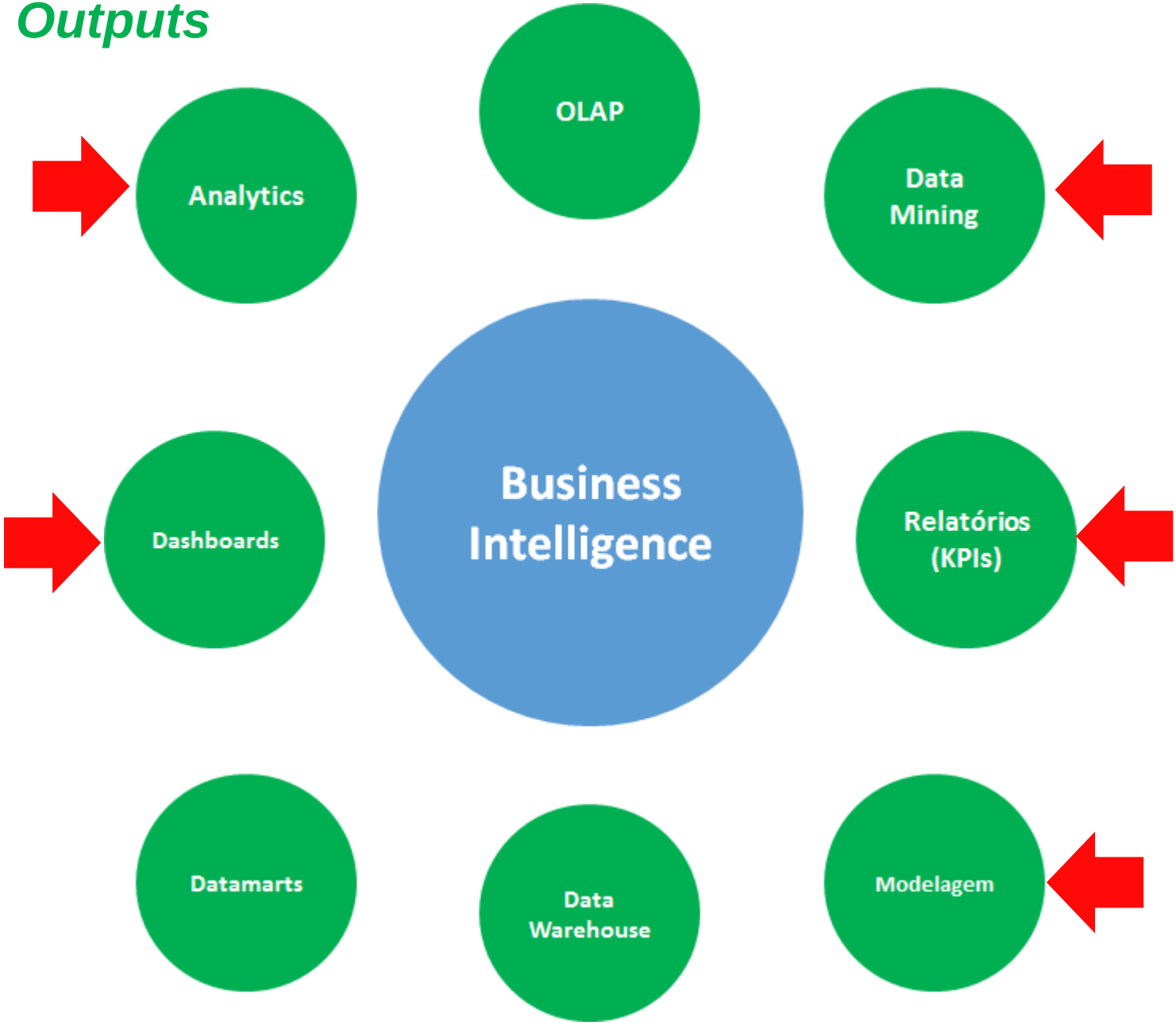
Relembrar é viver...



Business Intelligence

Relembrar é viver...

Outputs

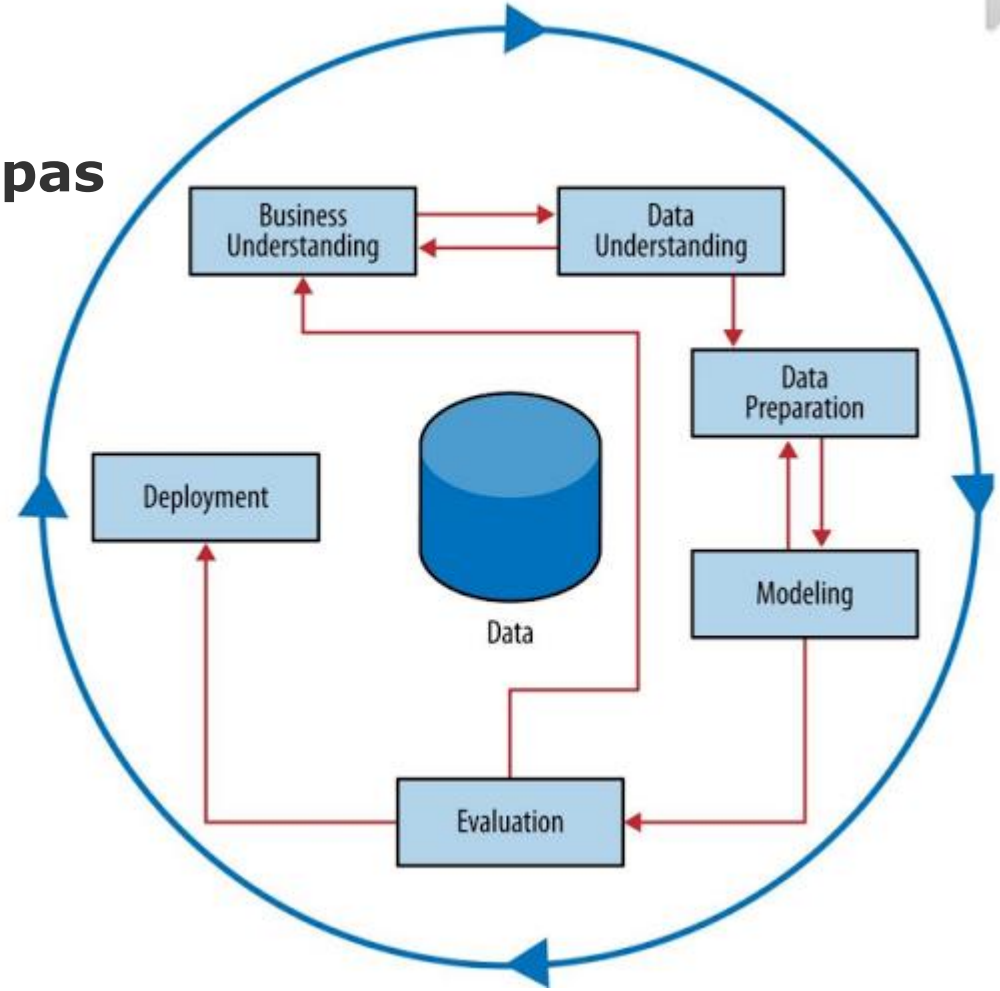


Business Intelligence
CRISP-DM

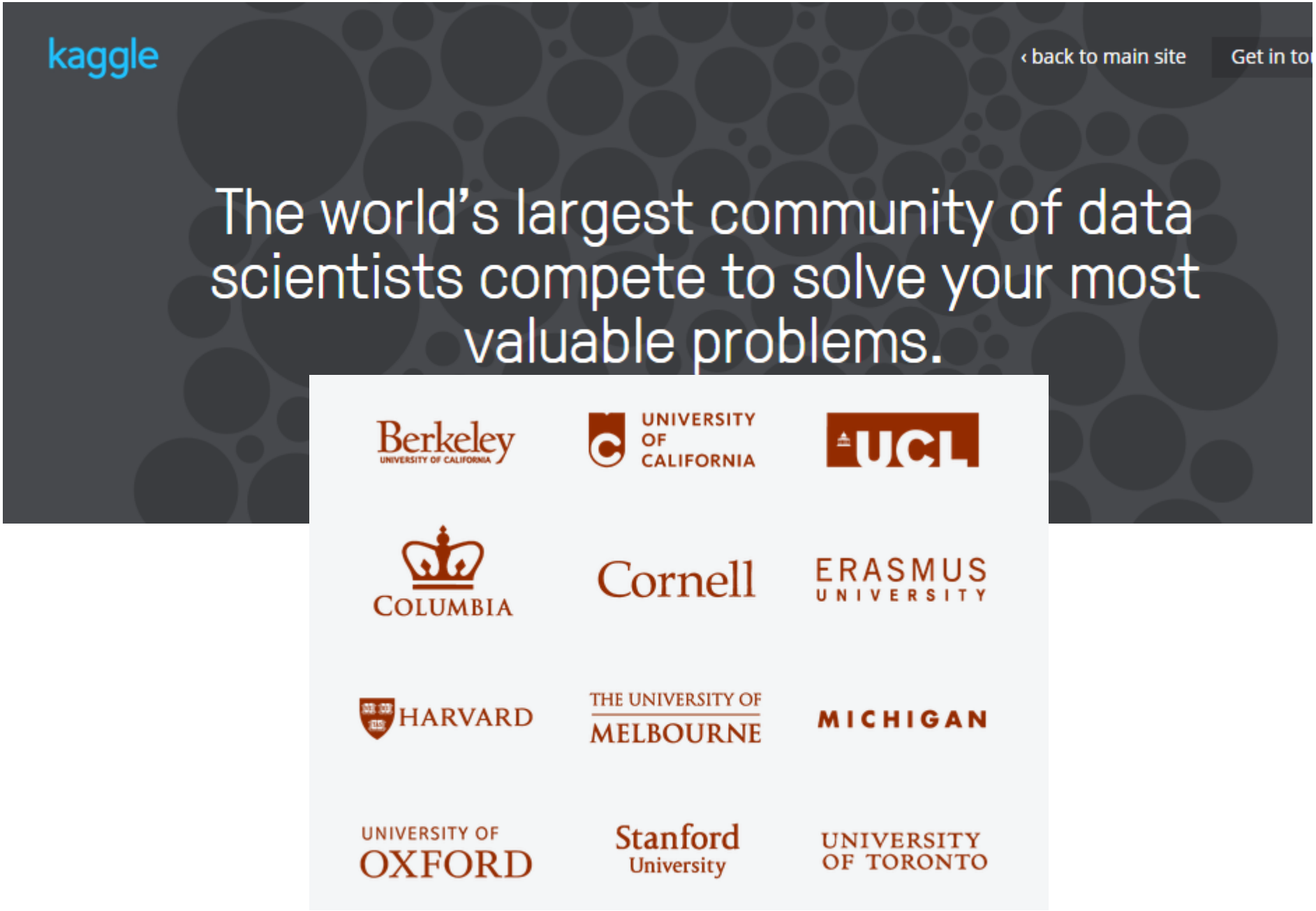
Relembrar é viver...

Ele é constituído de seis etapas

- Entendimento do Negócio
- Entendimento dos Dados
- Preparação dos Dados
- Modelagem
- Avaliação
- Entrega



Kaggle


kaggle
‹ back to main site
Get in to
The world's largest community of data scientists compete to solve your most valuable problems.
Berkeley
UNIVERSITY OF CALIFORNIA
UNIVERSITY OF CALIFORNIA
UCL
COLUMBIA
Cornell
ERASMUS UNIVERSITY
HARVARD
THE UNIVERSITY OF MELBOURNE
MICHIGAN
UNIVERSITY OF OXFORD
Stanford University
UNIVERSITY OF TORONTO

Kaggle
Titanic – Entendimento do Negócio

Knowledge • 2,051 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Thu 31 Dec 2015 (9 months to go)

Dashboard

Home
- Data
- Make a submission

Information

Competition Details » Get the Data » Make a submission

Predict survival on the Titanic (using Excel, Python, R, and Random Forests)

Somos os gerentes da empresa que construiu o Titanic!!!

Kaggle
Titanic – Entendimento do Negócio

Knowledge • 2,050 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Thu 31 Dec 2015 (9 months to go)

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - Frequently Asked Questio...
  - Further Reading / Watching
  - Getting Started With Excel
  - Getting Started With Pyth...
  - Getting Started With Pyth...
  - Getting Started With Ran...
  - New: Getting Started with R
  - Submission Instructions
- Forum
- Leaderboard
- Visualization

Submit to Titanic: Machine Learning from Disaster

Compete as myself

(You can always add team members later.)

Compete as a team

Kaggle
Titanic – Entendimento dos dados

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic
```

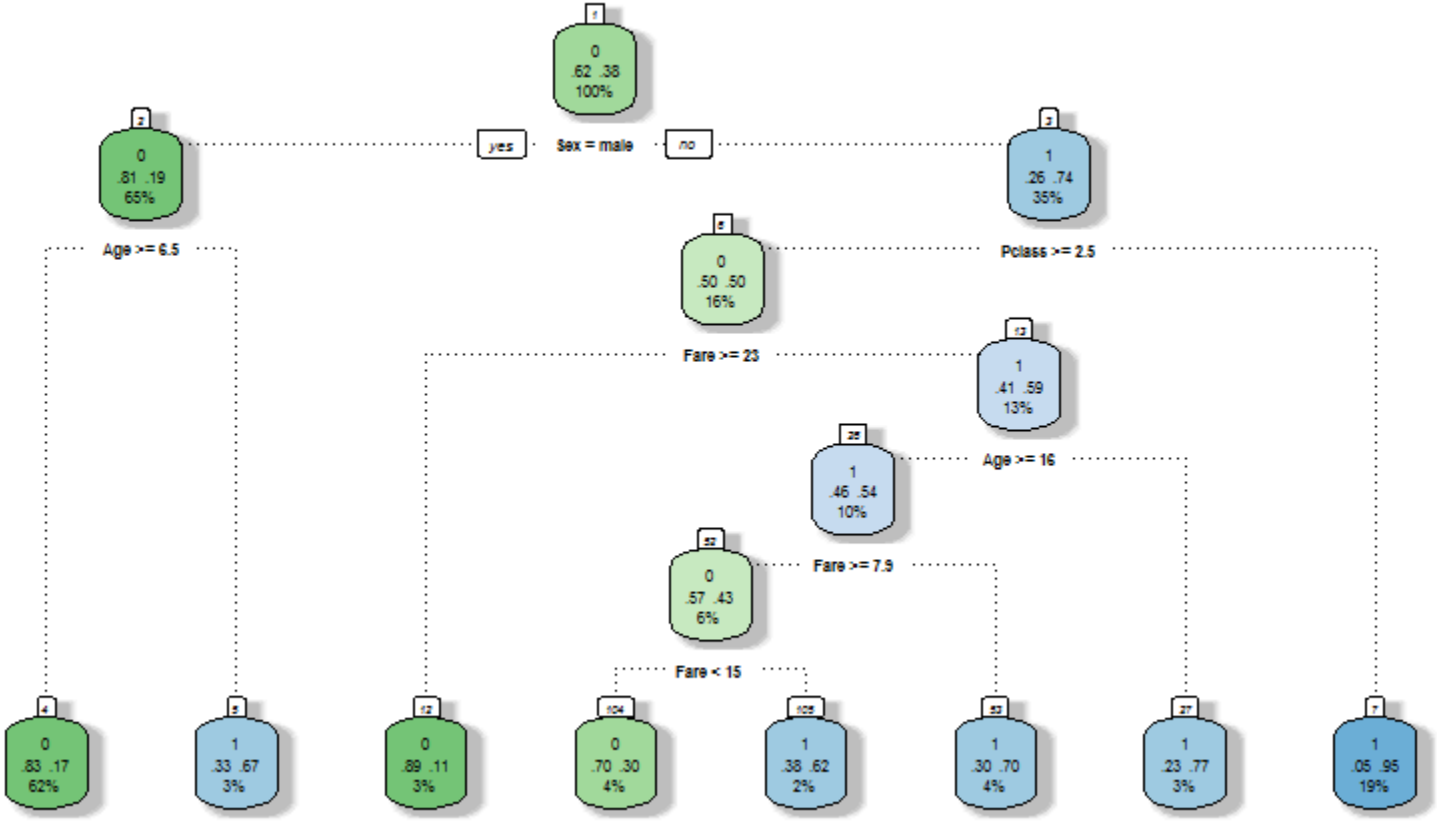
Kaggle

Titanic – Preparação dos dados e Modelagem

```
#Escolhendo o diretorio de trabalho
setwd("C:/Users/Leandro/Google Drive/FMU/Kaggle/TITANIC")

#Carregando as bases de treinamento e teste
library(data.table)
?data.table

treinamento<-data.table(read.csv("train.csv"))
validacao<-data.table(read.csv("test.csv"))
```

```
#######################################
### Construindo a árvore de decisão ###
#######################################

arvore1 <- rpart(Survived ~ Pclass + Sex + Age + Fare,
                 data=treinamento, method="class")
fancyRpartPlot(arvore1)

#Fazendo a predição
predicao1 <- predict(arvore1, validacao, type = "class")
avaliacao1 <- data.frame(PassengerId = validacao$PassengerId,
                         Survived = predicao1)
write.csv(avaliacao1, file = "avaliacao1.csv", row.names = FALSE)
```

Kaggle
Titanic – Árvore de Decisão 1



Rattle 2015-mar-29 12:36:03 Leandro

Kaggle
Titanic – Avaliação 1

Knowledge • 2,050 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Thu 31 Dec 2015 (9 months to go)

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - Frequently Asked Questio...
  - Further Reading / Watching
  - Getting Started With Excel
  - Getting Started With Pyth...
  - Getting Started With Pyth...
  - Getting Started With Ran...
  - New: Getting Started with R
  - Submission Instructions
- Forum
- Leaderboard
- Visualization

Competition Details » Get the Data » Make a submission

Make a submission

You have 10 entries today. This resets 8.1 hours from now (00:00 UTC).

Click or drop your submission here

Enter a brief description of this submission here.

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive if you prefer.

# of Predictions
We expect the solution file to have 418 predictions. The file should have a header row. Please see the sample submission file on the data page for an example of a valid submission.

Kaggle
Titanic – Avaliação 1

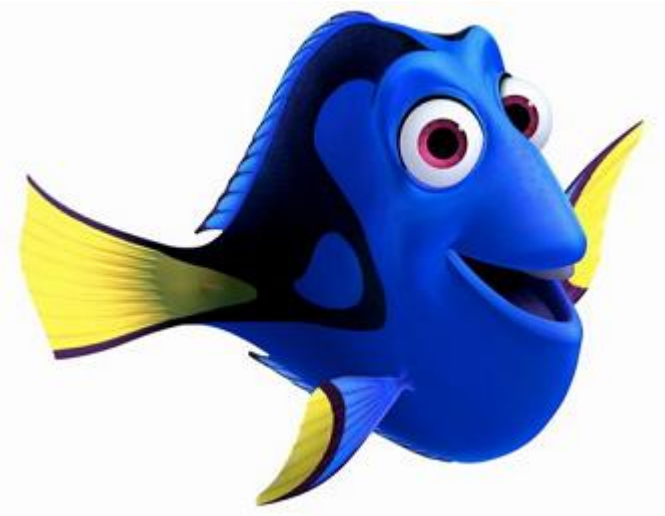
873	new	Leandro Guerra	0.78469	1	Sun, 29 Mar 2015 15:55:40

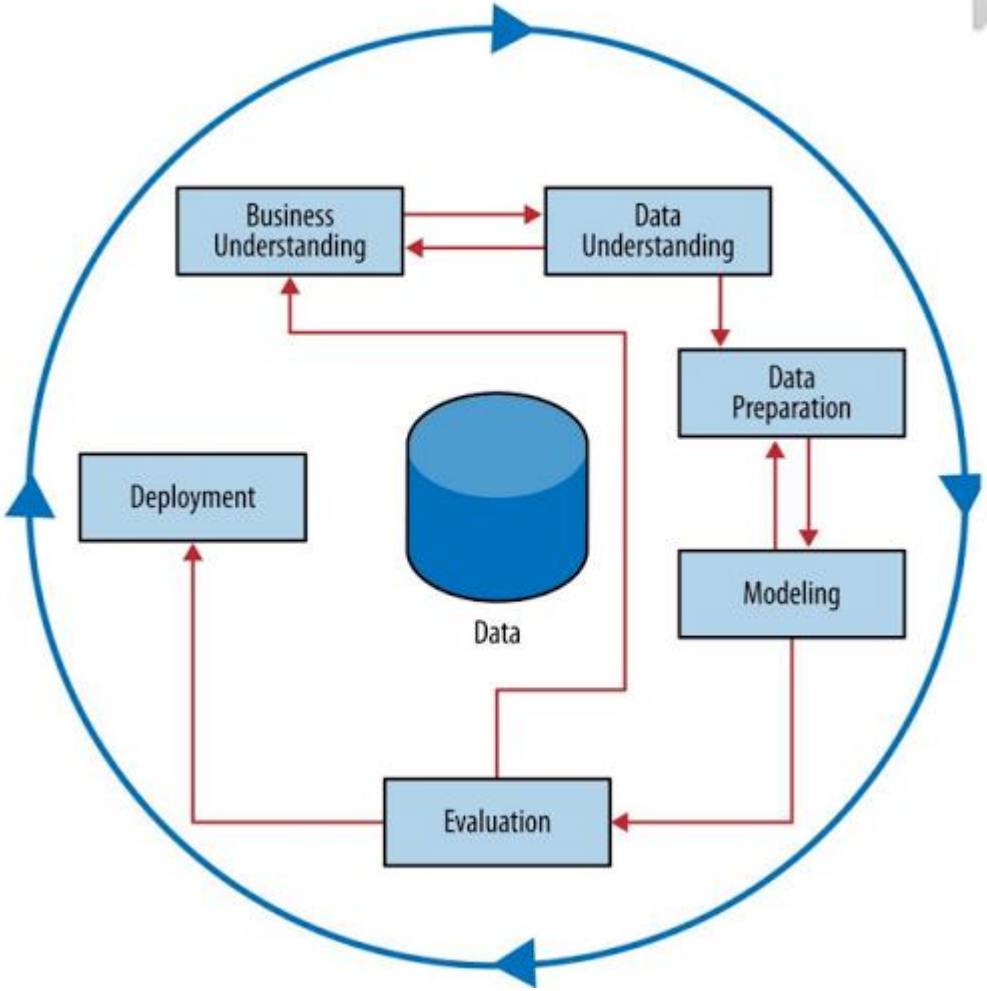
Your Best Entry ↑
Congratulations on making your first submission!

Tweet this!

É a única forma de fazer?
E se nossa etapa de preparação fosse melhor?

Kaggle
Titanic – Lembrando do Ciclo





Repare nas setas duplas ao longo de todo o fluxo!

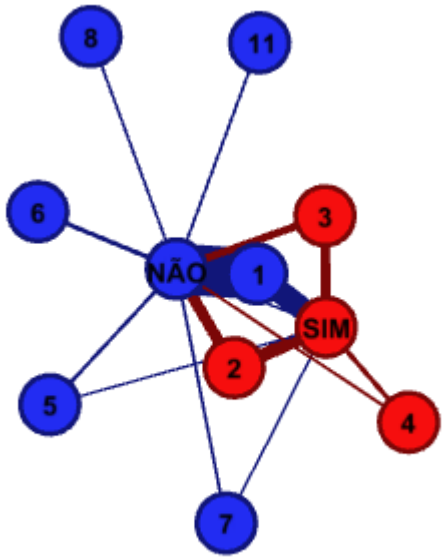
Kaggle
Titanic – Preparação dos dados e Modelagem
Loop 2

Será que o tamanho de uma família é um problema?

E o local de embarque?

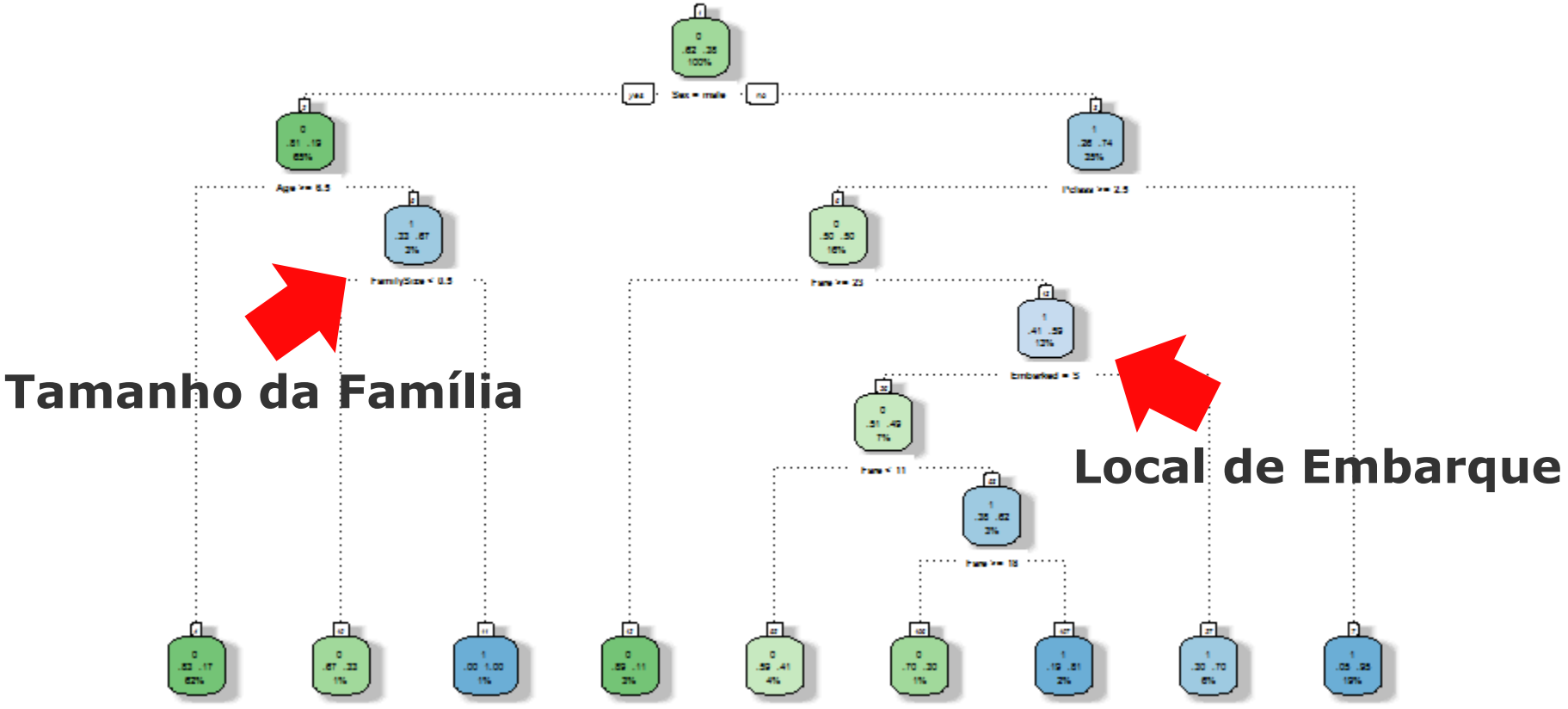
Gephi em ação!

```
#########################################
### Construindo a árvore de decisão 2 ###
#########################################

#Cria a variável FamilySize: 1-Pequena 0-Grande
treinamento$FamilySize <- treinamento$SibSp + treinamento$Parch + 1
validacao$FamilySize <- validacao$SibSp + validacao$Parch + 1

i <- 0
for (i in 1:length(treinamento$Survived)) {
  if (treinamento$FamilySize[i] < 4) {
    treinamento$FamilySize[i] <- 1
  } else {
    treinamento$FamilySize[i] <- 0
  }
}

i <- 0
for (i in 1:length(validacao$PassengerId)) {
  if (validacao$FamilySize[i] < 4) {
    validacao$FamilySize[i] <- 1
  } else {
    validacao$FamilySize[i] <- 0
  }
}
```



Kaggle

Titanic – Árvore de Decisão 2



Rattle 2015-mar-29 18:02:55 Leandro

Kaggle
Titanic – Avaliação 2

871	new	Leandro Guerra	0.78469	3	Sun, 29 Mar 2015 19:31:47 (-3.6h)

Your Best Entry ↑
Your submission scored **0.78469**, which is not an improvement of your best score. Keep trying!

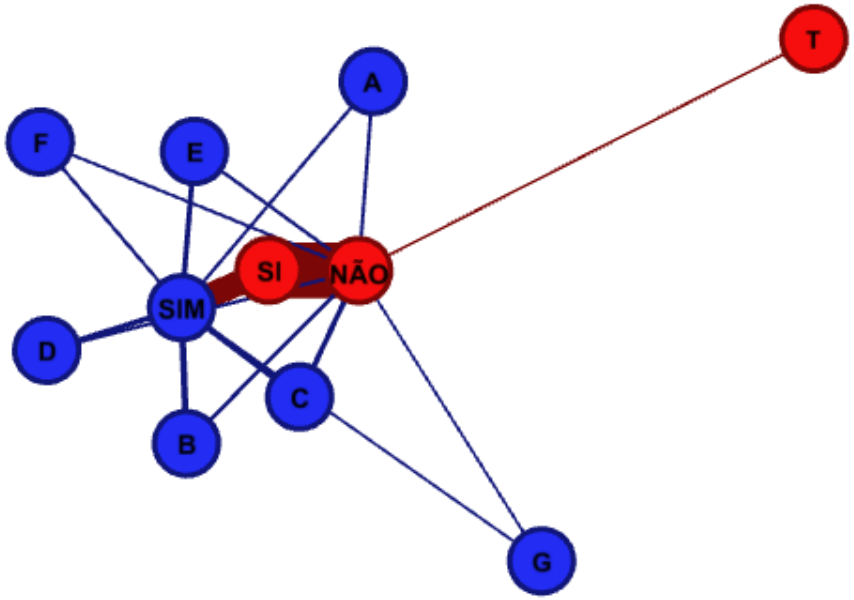
Subimos 2 posições!

Conseguimos melhorar?

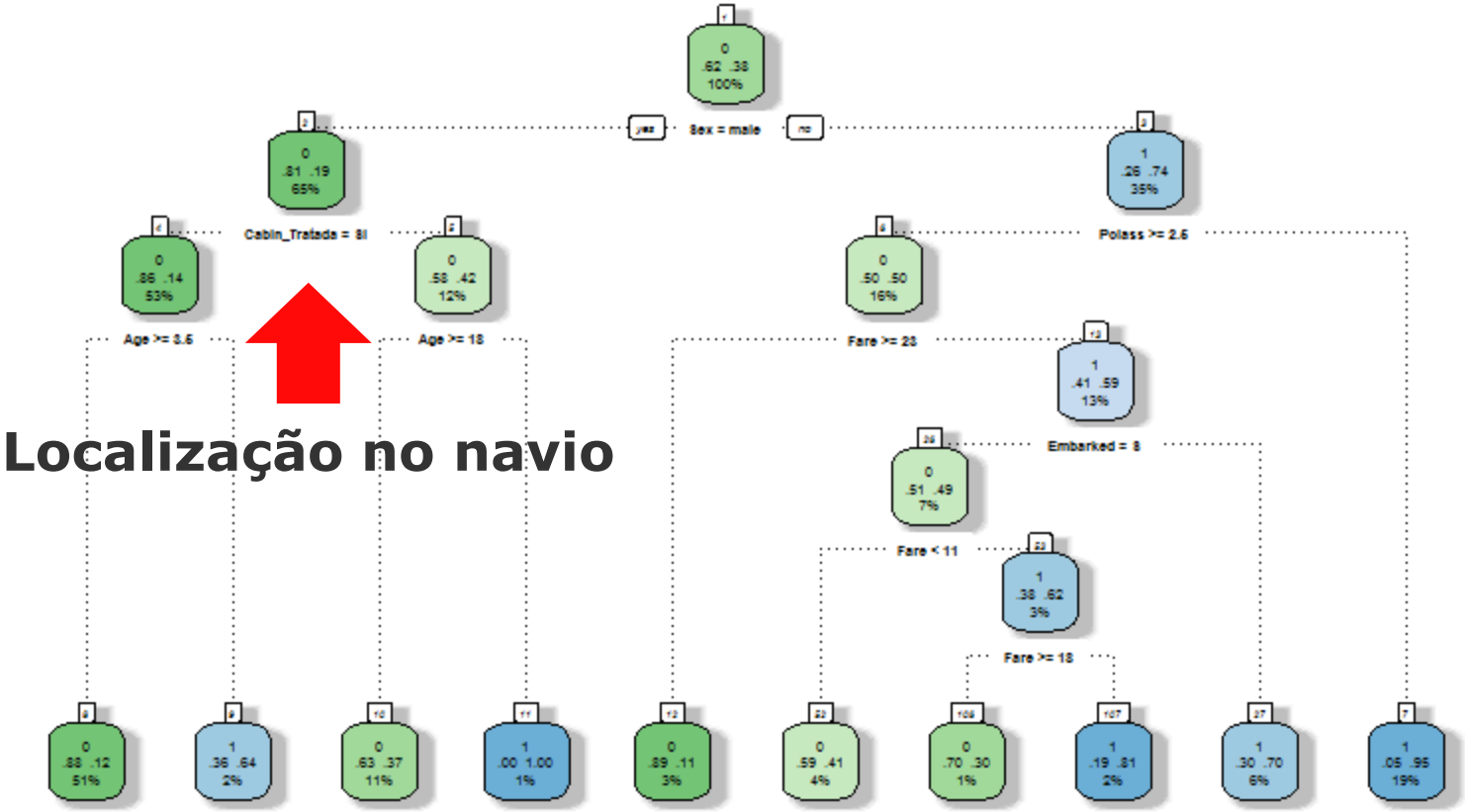
Kaggle
Titanic – Preparação dos dados e Modelagem
Loop 3

Será que onde a pessoa estava no navio era importante?

```
#######################################
### Construindo a árvore de decisão 3 ###
#######################################

#Tratando a variável Cabin
treinamento$Cabin_Tratada <- 0
i <- 0
for (i in 1:length(treinamento$Survived)) {
  if (treinamento$Cabin[i] == treinamento$Cabin[1]) {
    treinamento$Cabin_Tratada[i] <- "SI"
  } else {
    treinamento$Cabin_Tratada[i] <- "CA"
  }
}
treinamento$Cabin_Tratada <- as.factor(treinamento$Cabin_Tratada)

validacao$Cabin_Tratada <- 0
i <- 0
for (i in 1:length(validacao$PassengerId)) {
  if (validacao$Cabin[i] == validacao$Cabin[1]) {
    validacao$Cabin_Tratada[i] <- "SI"
  } else {
    validacao$Cabin_Tratada[i] <- "CA"
  }
}
validacao$Cabin_Tratada <- as.factor(validacao$Cabin_Tratada)
```



Kaggle
Titanic – Árvore de Decisão 3



Rattle 2015-mar-29 18:12:38 Leandro

Kaggle
Titanic – Avaliação 3

641	new	Leandro Guerra	0.78947	4	Sun, 29 Mar 2015 20:33:05

Your Best Entry ↑
You improved on your best score by 0.00478.

You just moved up 230 positions on the leaderboard. Tweet this!

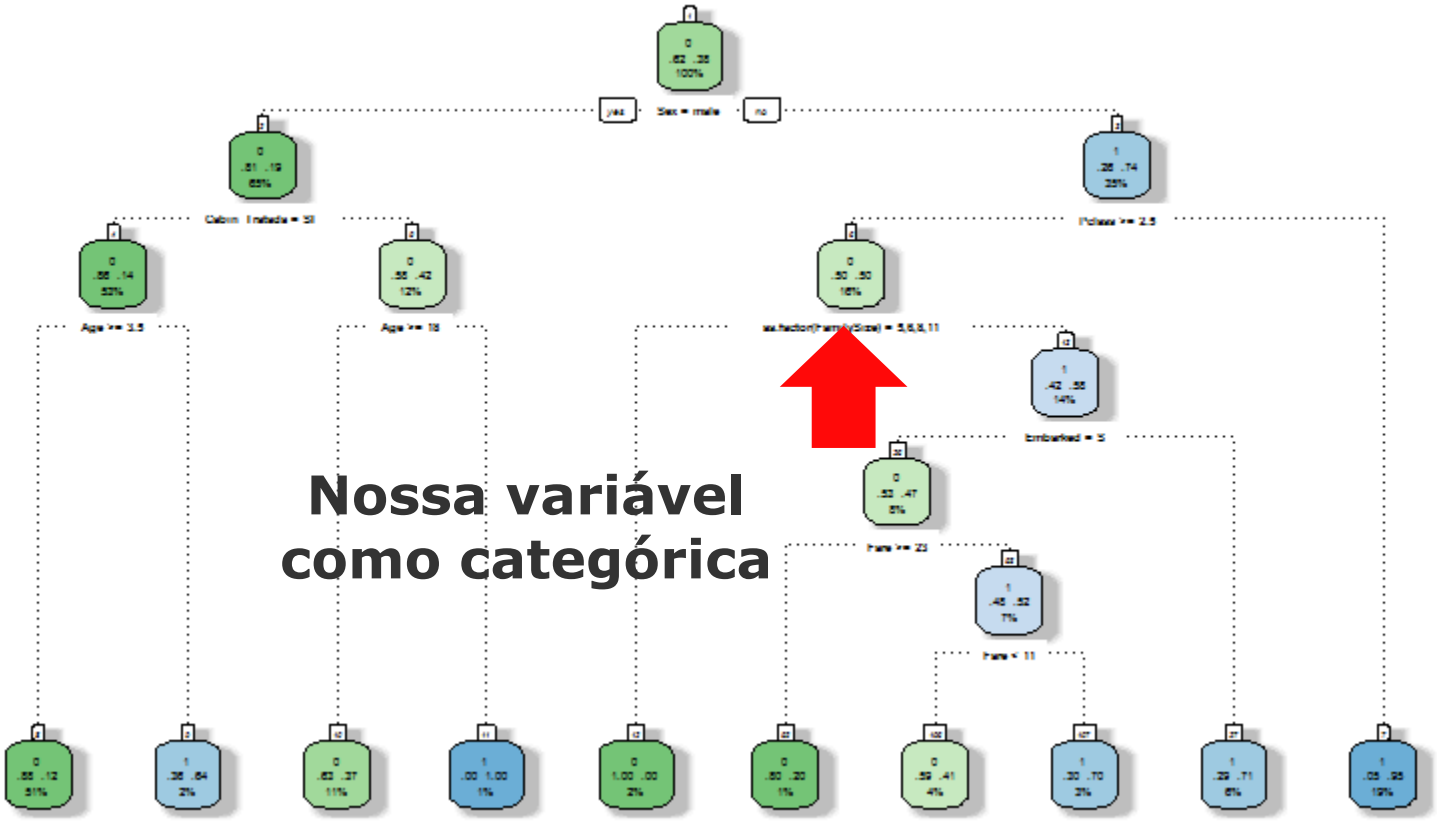
Subimos 230 posições!

Kaggle
Titanic – Preparação dos dados e Modelagem
Loop 4

E a forma como uma variável é classificada?

```
#########################################
### Construindo a árvore de decisão 4 ###
#########################################
arvore4 <- rpart(Survived ~ Pclass + Sex + Age + Fare +
                   Embarked + as.factor(FamilySize) + Cabin_Tratada,
                 data=treinamento, method="class")
fancyRpartPlot(arvore4)

#Fazendo a predição
predicao4 <- predict(arvore4, validacao, type = "class")
avaliacao4 <- data.frame(PassengerId = validacao$PassengerId,
                         Survived = predicao4)
write.csv(avaliacao4, file = "avaliacao4.csv", row.names = FALSE)
```

Kaggle
Titanic – Árvore de Decisão 4



Rattle 2015-mar-29 18:43:38 Leandro

Kaggle
Titanic – Avaliação 4

535	new	Leandro Guerra	0.79426	5	Sun, 29 Mar 2015 21:45:11

Your Best Entry ↑
You improved on your best score by 0.00478.

You just moved up 105 positions on the leaderboard. Tweet this!

Subimos mais 105 posições!

No total, subimos 338 posições!

Business Intelligence