



# Algumas Ferramentas Aplicadas em Ciência de Dados

Agente Educacional

Sérgio M. Dias

# Agenda

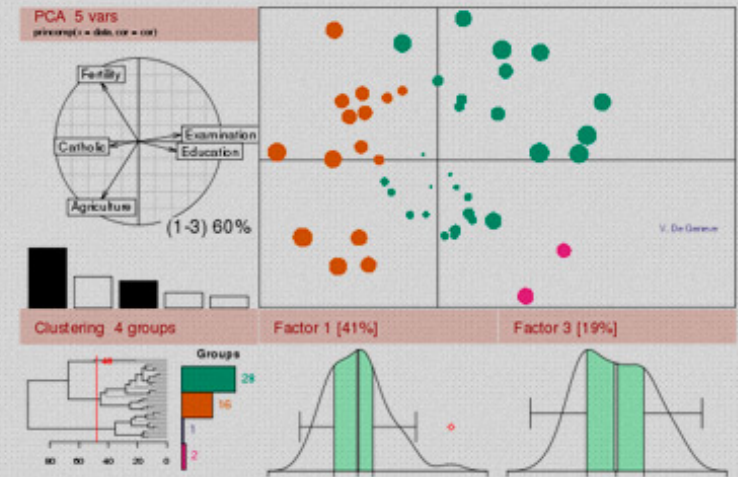
Ambiente R  
Ambiente Python  
SAS Enterprise Miner  
IBM SPSS  
KNIME  
RapidMiner





# Ambiente R

R é um ambiente gratuito e de código aberto que propicia excelente ambiente para análises estatísticas e com recursos gráficos de alta qualidade.



## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

### Vantagens:

- Código aberto com grande comunidade
- A maior variedade de técnicas
- Estado da arte em técnicas
- Constante evolução
- Disponível nas principais soluções comerciais (lista não exaustiva):





## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

## Interpretador R acessado pelo terminal

```
R version 3.1.0 (2014-04-10) -- "Spring Dance"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i686-pc-linux-gnu (32-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> █
```

## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

### Rstudio

Considerada a **mais poderosa IDE** para desenvolvimento em R, com licenças de código aberto e comerciais.





## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

## Rstudio Desktop:

- Disponível para Linux, Windows e Mac
- Syntax highlighting, code completion, indentação automática
- Execução de código R, ajuda e documentação integrados
- Gerenciamento de ambientes de trabalho
- Desenvolvimento de pacotes



## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

### Rstudio Server:

- Acesso via navegador Web
- Ferramentas administrativas
- Segurança
- Monitoramento de processos
- Gerenciamento de recursos
- Computação próxima dos dados

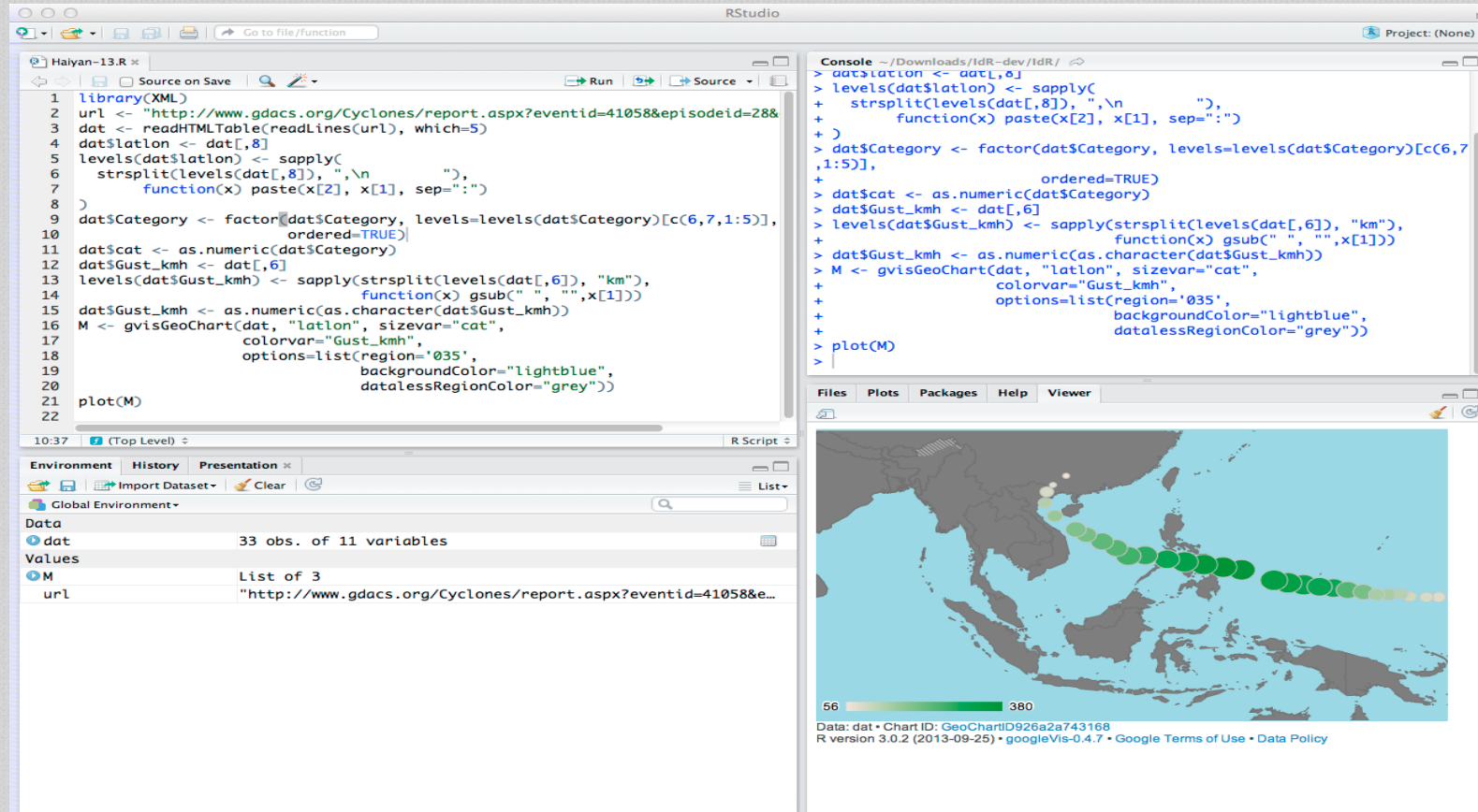




## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

### Rstudio



## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R

## Shiny

Pacote de R de **código aberto** que fornece uma estrutura elegante e poderosa para construção de aplicações web usando **R**.

**Transforma resultados** dos modelos **em aplicações web interativas** sem a necessidade de conhecer HTML, CSS ou JavaScript.

Implantação é recomendada no **Shiny Server**.

**Edições do Shiny Server:** *Open Source e Professional.*

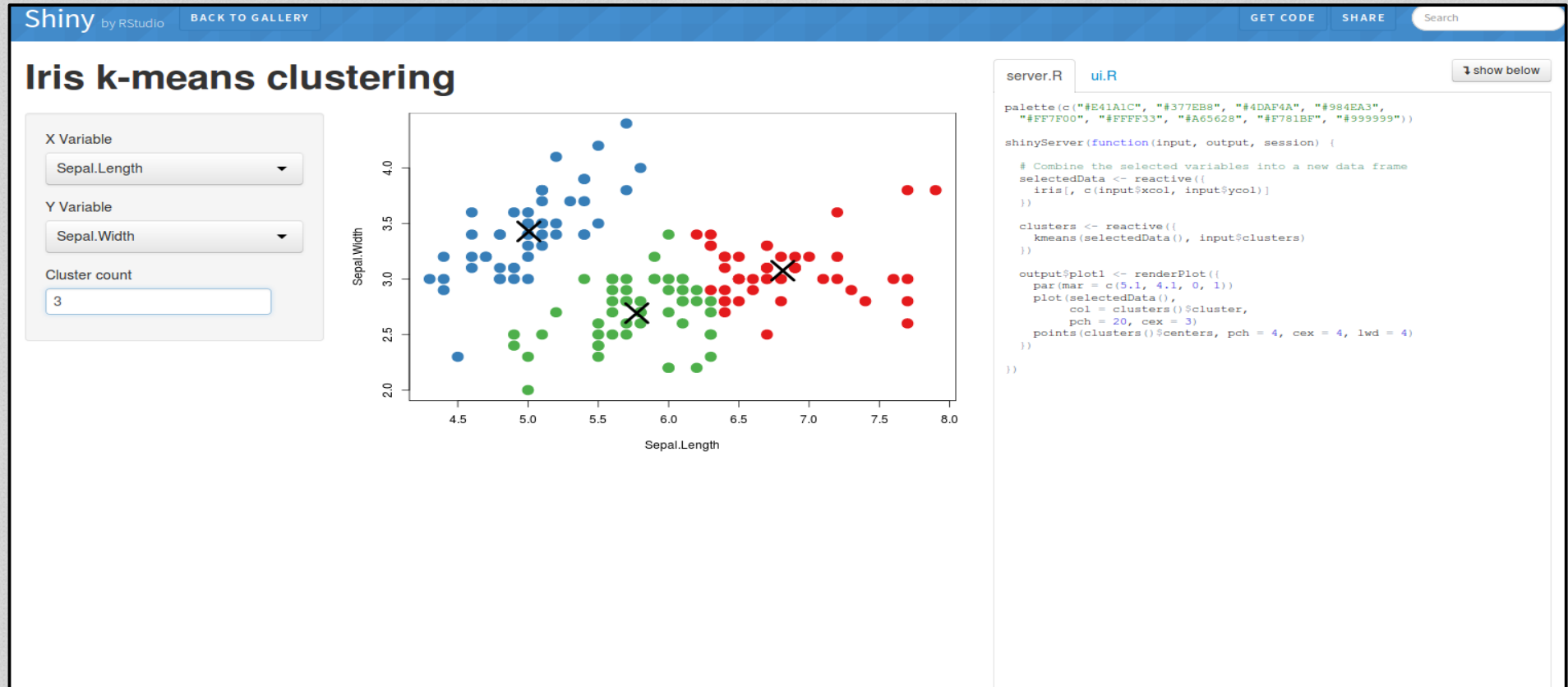
Fonte:  
<http://shiny.rstudio.com/>





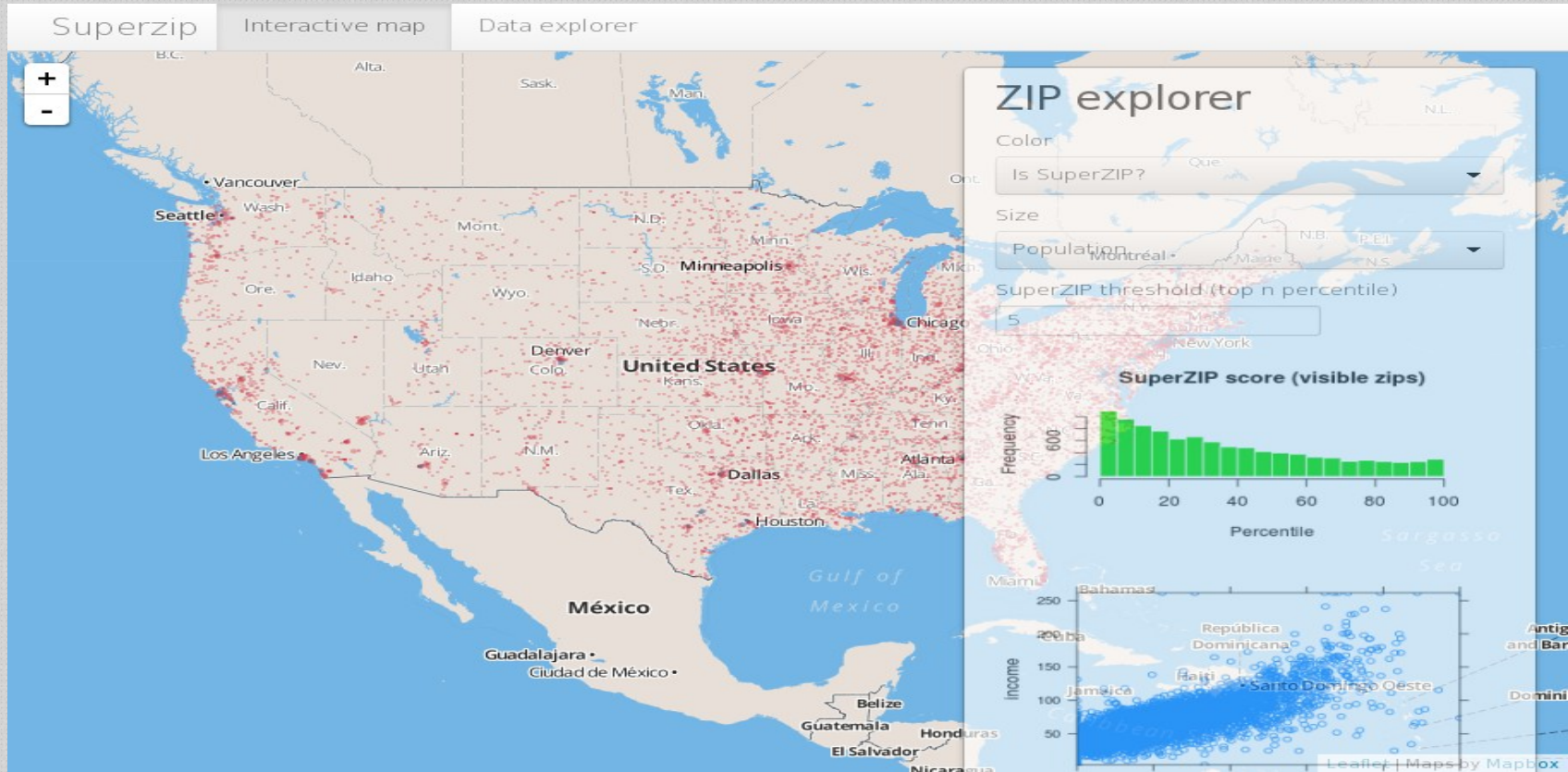
## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R  
Shiny



## Algumas Ferramentas Aplicadas em Ciência de Dados

Ambiente R  
Shiny





## Algumas Ferramentas Aplicadas em Ciência de Dados

# Ambiente Python

Python é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte.

Lançada por Guido van Rossum em 1991. Atualmente possui um modelo de desenvolvimento aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation.



# SAS Enterprise Miner

A empresa SAS possui uma **solução analítica** comercial com diversas implementações de algoritmos para **pré-processamento**, **mineração** e **visualização de dados**.

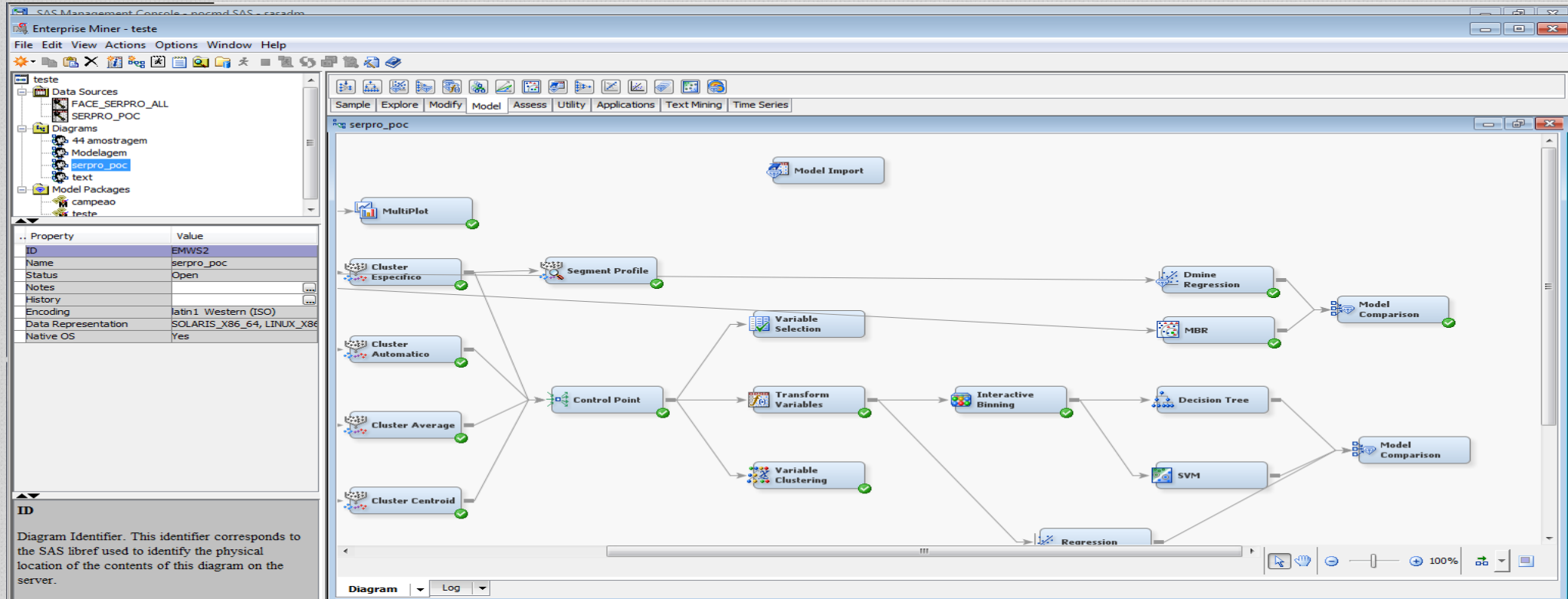
- SAS *Enterprise Guide*
- SAS *Metadata Bridge for Microstrategy*
- SAS *Enterprise Miner Server*
- SAS *Text Mining*
- SAS *Content Categorization*
- SAS *Sentiment Analysis*
- SAS *Text Miner Serve*
- SAS *Social Network Analysis*





## Algumas Ferramentas Aplicadas em Ciência de Dados

SAS Enterprise Miner



## Algumas Ferramentas Aplicadas em Ciência de Dados

## IBM SPSS

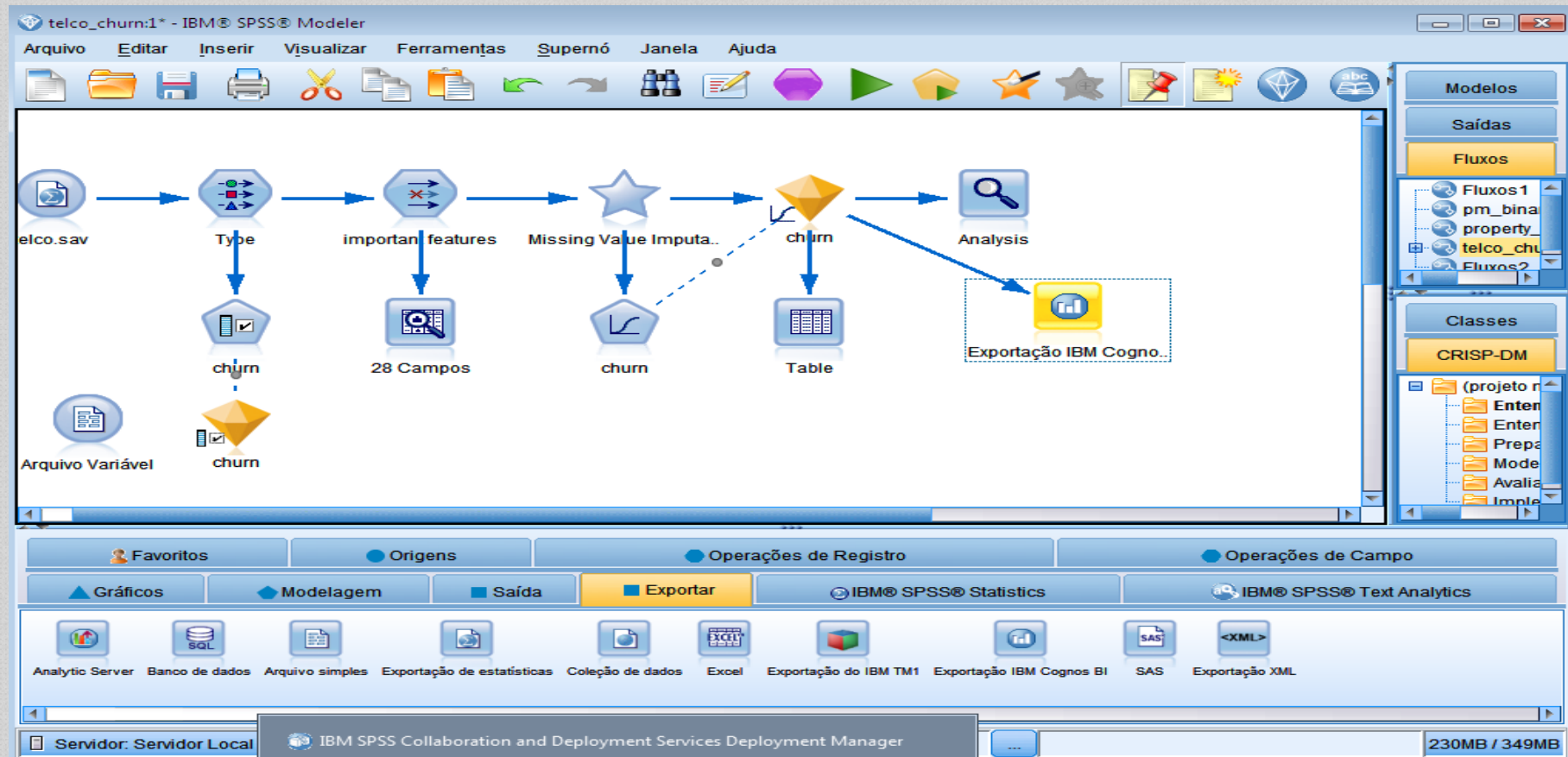


- Solução analítica comercial da IBM
- Possui dois módulos principais ligados a mineração de dados, o *SPSS Statistics* e o *SPSS Modeler*
- Para tarefas administrativas existe o módulo *SPSS Collaboration and Deployment Services*
- Na Web existe a possibilidade de modelagem rápida e rasa usando o *SPSS Modeler Advantage* e *Rules Management*
- O *Statistics* possui uma grande variedade de técnicas para modelagem estatística, mas sem nenhum compromisso com processo de MD



## Algumas Ferramentas Aplicadas em Ciência de Dados

IBM SPSS



## Algumas Ferramentas Aplicadas em Ciência de Dados

# KNIME

KNIME é uma plataforma de análise de dados que permite análises estatísticas e de mineração de dados.

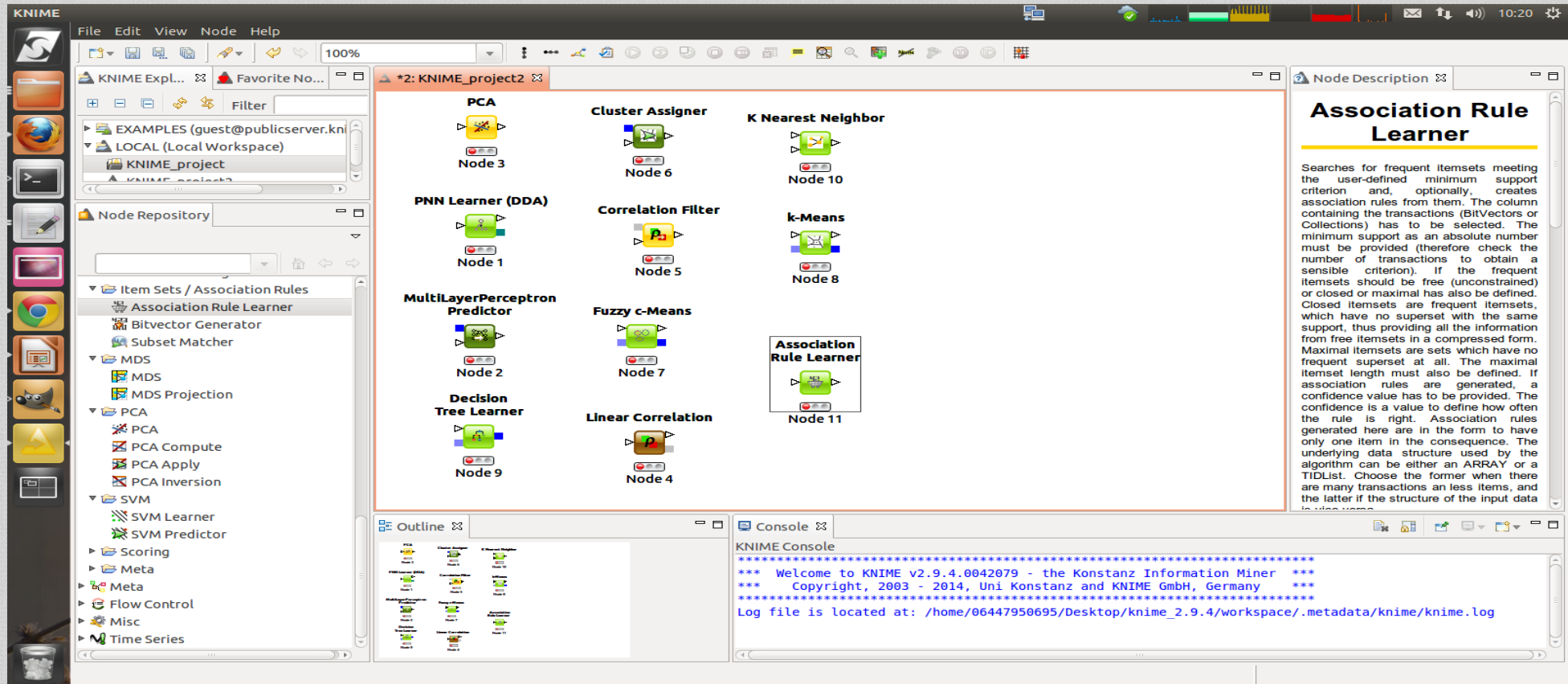
- Ambientes Desktop e Server
- Construção de fluxos de ETL e MD
- Disponibilidade para Linux, Windows e Mac
- Grande conjunto de técnicas
- Classificação, regressão, SVM, etc.





## Algumas Ferramentas Aplicadas em Ciência de Dados

KNIME



The screenshot displays the KNIME software interface. The main workspace shows a workflow canvas with several nodes connected by arrows, representing a data processing pipeline. The nodes include:

- PCA** (Node 3)
- Cluster Assigner** (Node 6)
- K Nearest Neighbor** (Node 10)
- PNN Learner (DDA)** (Node 1)
- Correlation Filter** (Node 5)
- k-Means** (Node 8)
- MultiLayerPerceptron Predictor** (Node 2)
- Fuzzy c-Means** (Node 7)
- Association Rule Learner** (Node 11)
- Decision Tree Learner** (Node 9)
- Linear Correlation** (Node 4)

The left sidebar contains a 'Node Repository' with various categories of nodes, including 'Item Sets / Association Rules', 'MDS', 'PCA', 'SVM', 'Scoring', 'Meta', 'Flow Control', and 'Misc'. The right sidebar shows the 'Node Description' for the 'Association Rule Learner' node, which includes a detailed explanation of its functionality:

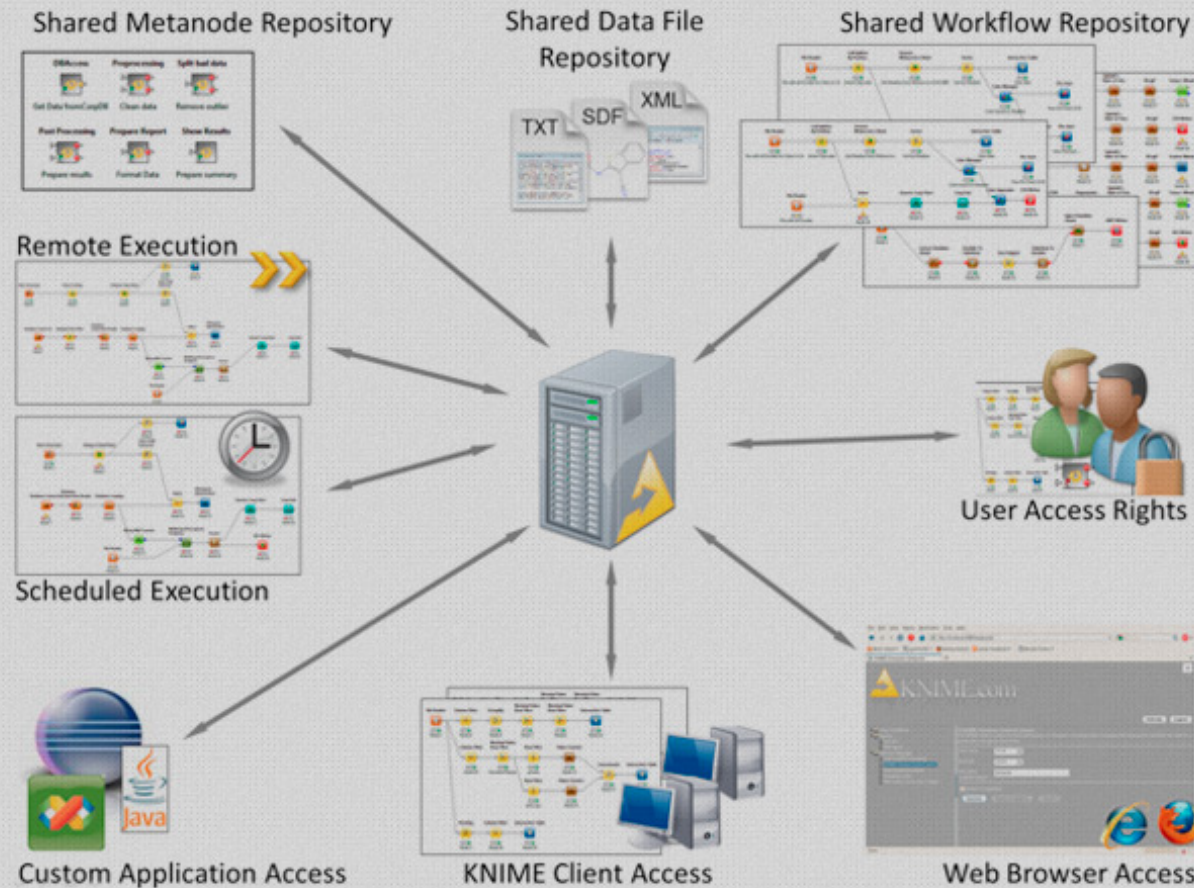
**Association Rule Learner**

Searches for frequent itemsets meeting the user-defined minimum support criterion and, optionally, creates association rules from them. The column containing the transactions (BitVectors or Collections) has to be selected. The minimum support as an absolute number must be provided (therefore check the number of transactions to obtain a sensible criterion). If the frequent itemsets should be free (unconstrained) or closed or maximal has also to be defined. Closed itemsets are frequent itemsets, which have no superset with the same support, thus providing all the information from free itemsets in a compressed form. Maximal itemsets are sets which have no frequent superset at all. The maximal itemset length must also be defined. If association rules are generated, a confidence value has to be provided. The confidence is a value to define how often the rule is right. Association rules generated here are in the form to have only one item in the consequence. The underlying data structure used by the algorithm can be either an ARRAY or a TIDList. Choose the former when there are many transactions an less items, and the latter if the structure of the input data is like name.

The bottom of the interface shows an 'Outline' view of the workflow and a 'Console' window displaying the KNIME welcome message and the location of the log file.

## Algumas Ferramentas Aplicadas em Ciência de Dados

KNIME



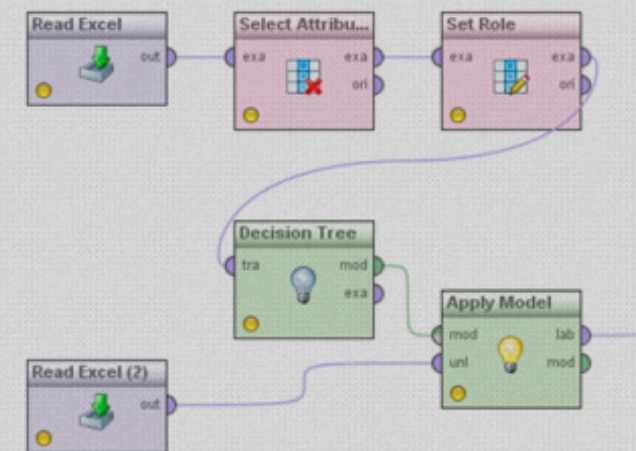


## Algumas Ferramentas Aplicadas em Ciência de Dados

# RapidMiner

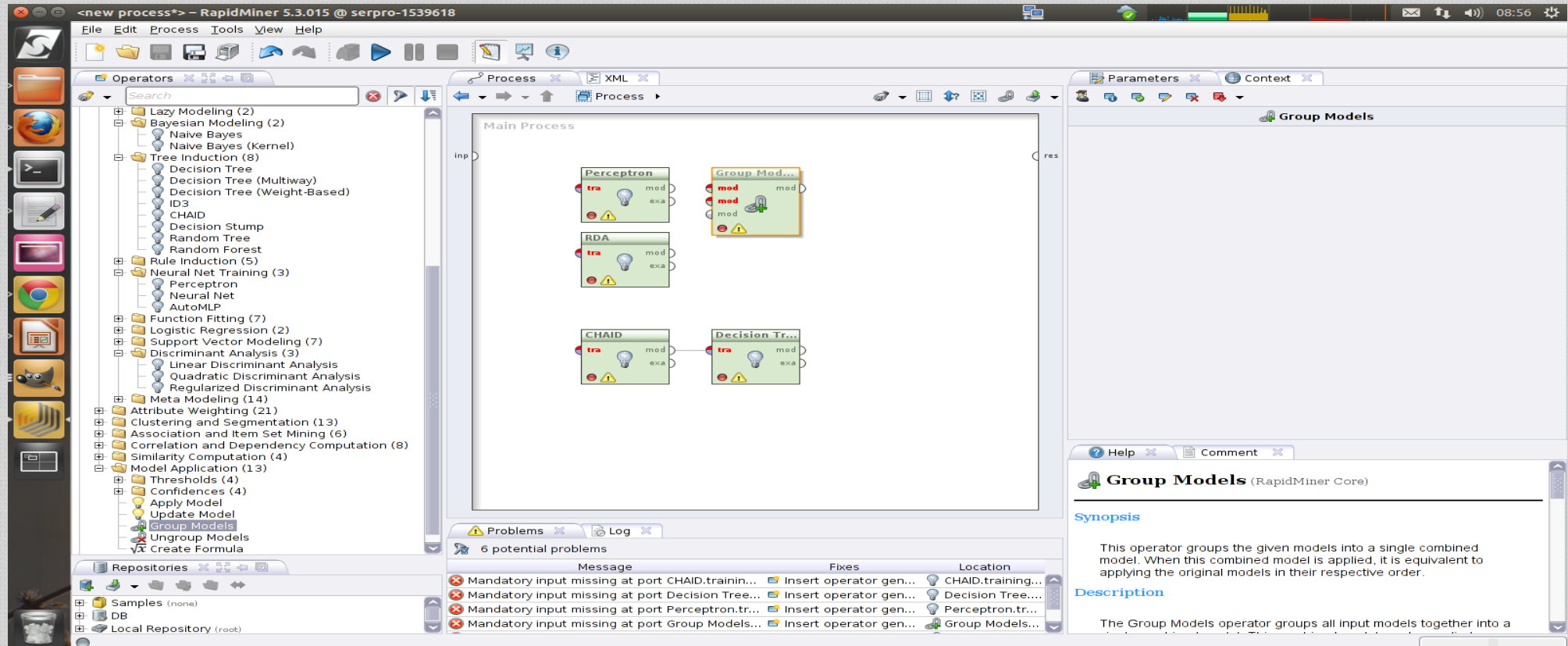
- Ambiente gráfico para aplicações de MD
- Ambientes *Desktop* e *Server*
- Construção de fluxos de ETL e MD
- Desktop para Linux, Windows e Mac.
- Boa disponibilidade de técnicas:
  - Classificação
  - Regressão
  - Agrupamento
  - SVM, redes neurais...

Fonte:  
<http://rapidminer.com/>



## Algumas Ferramentas Aplicadas em Ciência de Dados

RapidMiner



The screenshot displays the RapidMiner 5.3.015 interface. The main window shows a workflow diagram titled "Main Process" with the following operators: "Perceptron", "RDA", "CHAID", and "Group Mod...". The "Group Mod..." operator is highlighted, indicating it is the selected operator. The left sidebar shows a tree view of the "Operators" panel, with "Group Models" selected under the "Model Application" category. The bottom panel shows a "Problems" list with 6 potential problems, including "Mandatory input missing at port CHAID.trainin...", "Mandatory input missing at port Decision Tree...", "Mandatory input missing at port Perceptron.tr...", and "Mandatory input missing at port Group Models...". The right sidebar shows the "Group Models" operator details, including a "Synopsis" and a "Description".

**Group Models (RapidMiner Core)**

**Synopsis**

This operator groups the given models into a single combined model. When this combined model is applied, it is equivalent to applying the original models in their respective order.

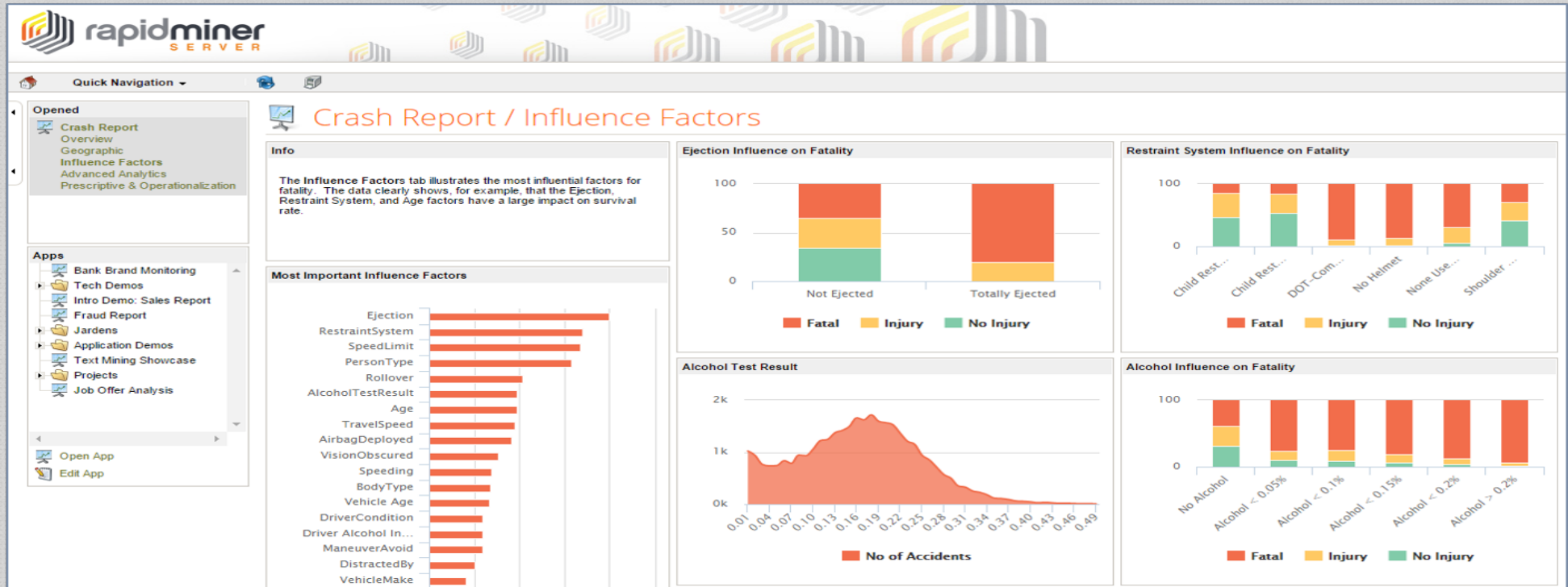
**Description**

The Group Models operator groups all input models together into a



## Algumas Ferramentas Aplicadas em Ciência de Dados

RapidMiner





# Obrigado!

Agente Educacional

Sérgio M. Dias

*sergio.dias@serpro.gov.br | #31 6539*

*Demais agentes educacionais sobre o assunto*

*Marcelo Pita | marcelo.pita@serpro.gov.br | #81 8794*

*Gustavo Torres | gustavo.gamatorres@serpro.gov.br | #31 6950*