**APRENDIZADO DE MÁQUINA**

***Curso de especialização em Bancos de Dados***

***Curso de especialização em Business Intelligence***

***Curso de especialização em Big Data***

**Prof. Hugo de Paula**

**ATIVIDADE EM LABORATÓRIO 2**

**REGRAS DE ASSOCIAÇÃO**

*Adaptado de North, Matthew A. Data Mining for the Masses. 2012.*

# OBJETIVOS

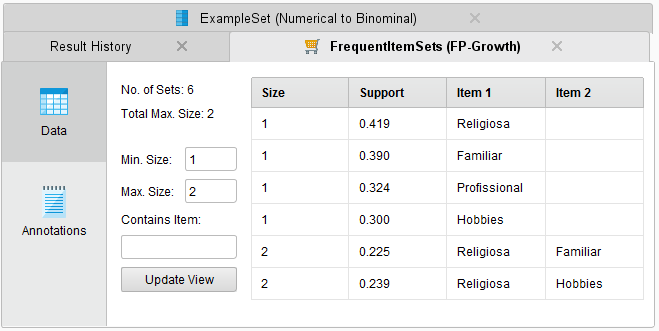
1. O objetivo dessa atividade é encontrar pessoas em organizações comunitárias de uma cidade que possuam algum tipo de conexão com as outras, para que elas possam, no futuro, trabalharem de forma cooperativa.
2. Para alcançar este objetivo, iremos utilizar o modelo de regras de associação. Então nosso objetivo torna-se encontrar regras de associação que modelem ligações entre tipos de organizações comunitárias.
3. Resumo dos dados:
   1. ***Tempo decorrido***: tempo gasto respondendo o questionário.
   2. ***Tempo\_na\_comunidade***: curto for (0-2 anos), médio (3-9 anos), ou longo (10+ anos).
   3. ***Gênero***: M (masculino) ou F (feminino).
   4. ***Trabalhando***: Sim/Não. Se o respondente possui um emprego remunerado.
   5. ***Idade***: numérico.
   6. ***Familiar***: Sim/Não indicando se é membro de uma organização orientada à família como grupos de apadrinhamento.
   7. ***Hobbies***: Sim/Não indicando se é membro de uma organização orientada hobby como rádio amador, aeromodelismo, motocicleta, etc.
   8. ***Clubes***: Sim/Não indicando se é membro de uma organização social como o Lions Club ou Rotary Club, etc.
   9. ***Política***: Sim/Não indicando se é membro de uma organização política como partidos políticos, grupos de lobby, etc.
   10. ***Profissional***: Sim/Não indicando se é membro de uma organização profissional como a OAB, o CRM, CRA, etc.
   11. ***Religiosa***: Sim/Não indicando se é membro de uma organização religiosa como uma igreja.
   12. ***Grupo de suporte***: Sim/Não indicando se é membro de uma organização de suporte à comunidade, como os Alcóolicos Anônimos, LBV, Médicos sem Fronteiras, etc.

# PREPARAÇÃO DOS DADOS

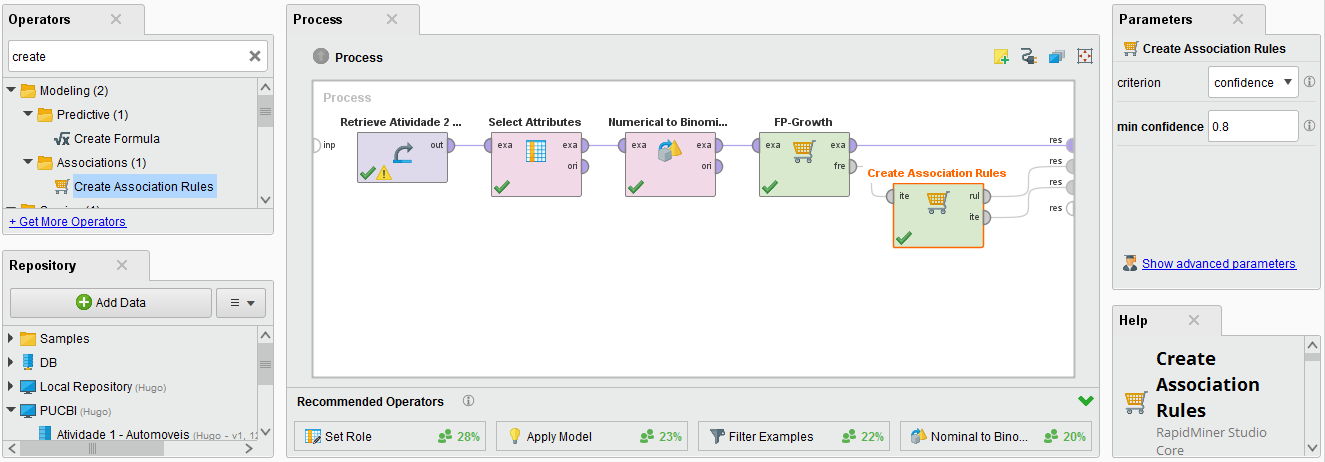
1. Carregue a base de dados ***Atividade 2 - comunidade.xlsx***.
2. Na visão de resultados, analise a ***Statistics***. Note que não há atributos omissos, nem dados inconsistentes. Note que as colunas binárias sobre o respondente pertencerem a alguma comunidade foram codificadas como 0 (Não) e 1 (Sim), e o ***RapidMiner*** as importou como atributos inteiros.
3. Como o algoritmo de regra de associação só suporta dados binominais, faça a preparação dos dados para adequá-los ao algoritmo de regras de associação.
   1. Reduza o número de atributos da sua base de dados eliminando dados desinteressantes. Utilize o operador ***Select Attributes*** e selecione apenas ***Familiar, Hobbies, Clubes, Política, Profissional, Religiosa***, ***Grupo de suporte***.
   2. Mude o tipo de dados de numérico para binominal, através do operador: ***Blending → Attibutes → Types → Numerical to Binominal***.

# MODELAGEM

1. Vamos identificar os conjuntos de itens frequentes. Utilize o operador ***Modelling → Associations → FP-growth***. Verifique o ***min support = 0.2*** na aba ***Parameters***. Desabilite a opção ***find min number of itemsets***.
2. Conecte a saída ***fre***, que representa os ***frequent patterns***, na saída do processo. Ao executarmos o processo identificamos alguns padrões frequentes na base.



1. Vamos agora criar as regras de associação. Adicione o operador ***Modelling → Associations → Create Association Rules***. O processo deve ficar semelhante à figura a seguir. Observe o valor do parâmetro ***min confidence=0.8***.



1. As saídas do operador de regras de associação são as regras e os conjuntos de itens frequentes. **Com esses parâmetros, o processo não irá encontrar nenhuma regra.**

# EXERCÍCIO

1. Lembrando-se que o processo de *Data Science* sugere que **o aprendizado de máquina é um processo iterativo**, volte ao modelo e **altere os valores de suporte e confiança para que eles produzam padrões interessantes**.
   1. Escreva quais os valores encontrados que produziram regras interessantes.
   2. Baseado nesses valores, tente interpretar os resultados. Por exemplo, tente entender qual o potencial de alcance daquela regra na população. Lembre-se de analisar o suporte em termos do tamanho da base de dados. Por exemplo, um suporte pequeno em uma base grande pode representar uma quantidade não desprezível de indivíduos, enquanto um suporte pequeno em uma base pequena pode representar apenas ruído.
2. Em Aprendizado de Máquina, é importante dividirmos a base de dados em uma base de **TREINAMENTO** e outra de **TESTE**. Aprenda a utilizar o operador ***Split Data*** para criar duas partições do seu dado, com 80% dos dados para **TREINAMENTO** e 20% para **TESTE**. Conecte o operador após o processamento dos dados.
3. Em seguida, utilize o operador ***Apply Association Rules*** para aplicar as regras obtidas na base de testes.
4. Utilize o *aggregated lift* como valor de confiança.
5. Escreva um documento mostrando o processo realizado, explicando os operadores e seus parâmetros, e mostrando as regras resultantes.
   1. Tente se limitar a uma ou duas páginas.
   2. Utilize imagens da tela do ***RapidMiner*** para mostrar os operadores, os dados antes e após o processamento.
   3. Este documento não precisa de capa e nem ser formatado em nenhum formato específico de trabalho acadêmico (estilo ABNT).
   4. Basta ter o título **ATIVIDADE 2**, e conter o nome do autor do trabalho.