**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 6**

**K-MEANS**

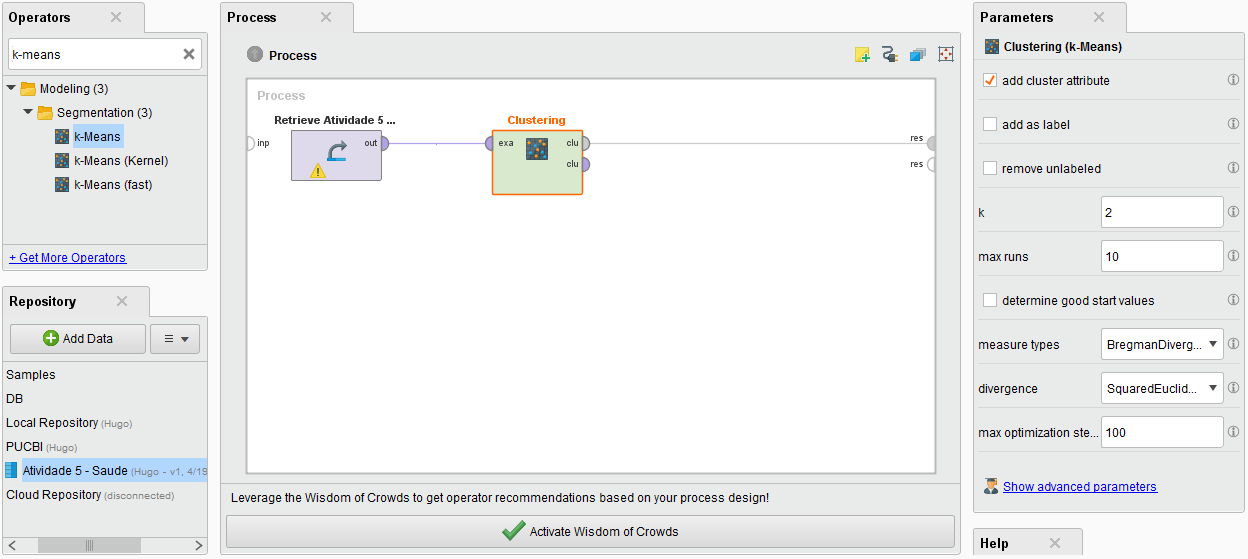
*Adaptado de North, Matthew A. Data Mining for the Masses. 2012.*

# OBJETIVOS

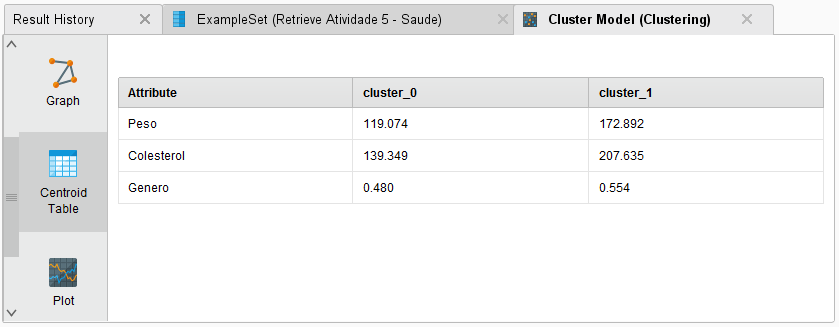
1. Uma operadora de seguro de saúde sabe da forte influência do peso, gênero e nível de colesterol no desenvolvimento de doenças cardíacas. O objetivo dessa atividade é identificar perfis de indivíduos com maior risco para sobrepeso e colesterol alto.
2. Para alcançar este objetivo, iremos utilizar o modelo agrupamento pelo algoritmo das K-Means. Então nosso objetivo se torna encontrar grupos de indivíduos (*clusters*) e definir os limites entre o grupo saudável e o grupo de risco.
3. A base de dados se chama ***Atividade 6 - saúde.xlsx***, e possui os atributos ***peso, colesterol e gênero***. ***Peso e colesterol*** são atributos numéricos. Na variável gênero, ***1*** corresponde ao sexo masculino e ***0*** ao feminino.

# MODELAGEM

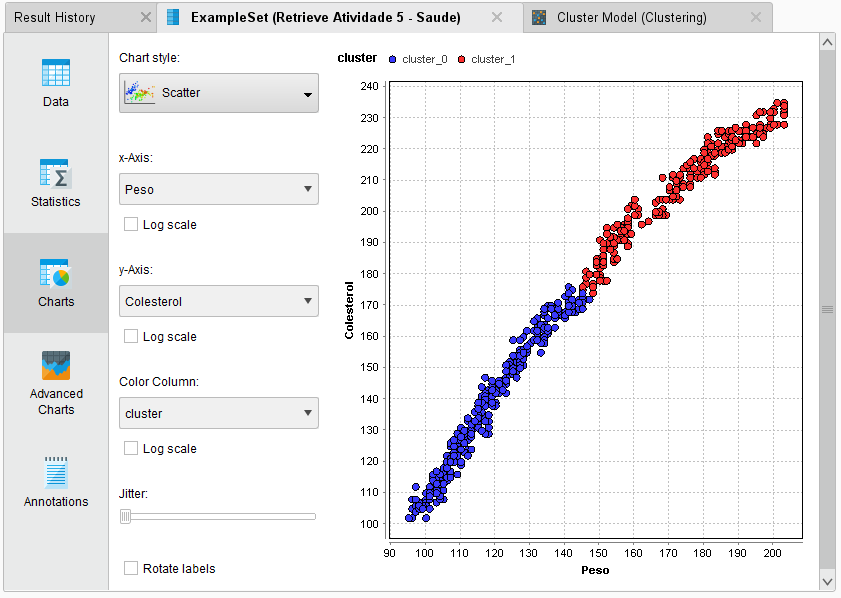
1. Carregue a base de dados ***Atividade 6 - saúde.xlsx***.
2. Localize o operador ***Modelling → Segmentation → K-means*** e adicione ao processo.



1. Observe nos parâmetros o número de clusters igual a 2. A medida de distância é definida no parâmetro ***divergence*** que apresenta valores baseado no tipo de medida selecionado em ***measure types***. No nosso exemplo, a distância default selecionada foi ***SquaredEuclideanDistance***, ou distância euclidiana quadrática.
2. Execute o processo e analise os resultados. O operador ***k-means*** possui duas saídas:
   1. ***Cluster Model***: exibe informações sobre o cluster. É possível se visualizar, por exemplo, os valores do centroide do cluster.



* 1. ***Cluster set***: exibe a base de dados rotulada com os clusters. É possível se visualizar características como média e variância de cada cluster. Bem como plotar os clusters de dados.



# EXERCÍCIO

1. Tende descrever os dois grupos gerados. Utilize as informações dos centroides dos clusteres para tentar diferenciar os dois grupos de risco.
2. Vamos tentar encontrar 4 grupos distintos: alto risco possui colesterol e peso altos; baixo risco possui colesterol e peso baixos; e risco moderado possui apenas o peso ou apenas o colesterol altos. Altere o número de clusters para 4.
3. Verifique os resultados. Investigue a tabela de centroides e a visão dos dados, além dos gráficos. Defina quais são os limites de peso e de colesterol para que um indivíduo seja considerado paciente de risco.
4. Observe que o atributo colesterol varia muito mais que o atributo peso, o que pode causar uma tendência do modelo de valorizar mais o colesterol do que os demais parâmetros.
5. Altere a medida de divergência para distância de Mahalanobis. Veja os resultados. Pesquise na internet a distância de Mahalanobis e tente explicar porque os resultados foram tão ruins.
6. Escreva um documento mostrando o processo realizado, explicando os operadores e seus parâmetros, e mostrando os clusteres obtidos. Interprete também os resultados.
   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 6**, e conter o nome do autor do trabalho.