Instituto Federal de Educação, Ciência e Tecnologia da Bahia
Curso de Tecnologia em Análise e Desenvolvimento de Sistemas
Prof.: Grinaldo Lopes de Oliveira
Disciplina: Banco de Dados 2

Atividade de Laboratório

Membros da Dupla: _____

_____

Objetivo do Laboratório:

Fixar conceitos relativos à **Mineração de Dados** aprendido em sala de aula através de uma abordagem prática utilizando o programa WEKA escrito em linguagem JAVA. Neste laboratório teremos a oportunidade de exercitar a técnica de mineração de dados intitulado regressão e aglomeração (*cluster)*;

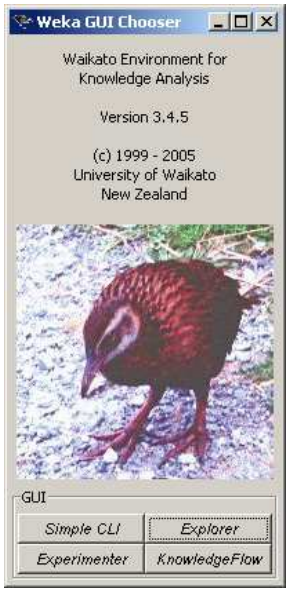
.

INICIANDO OS MOTORES

Iniciando as Atividades

Este laboratório deverá ser feito **OBRIGATORIAMENTE** em dupla. Iniciaremos nossas atividades praticando nossos conhecimentos através de um programa que procura representar o conhecimento acerca de uma árvore genealógica.

Inicialmente, execute o programa WEKA instalado em seu computador. Pressione o botão EXPLORER logo em seguida.



Desenvolvido com o objetivo de gerar uma regressão linear. Esse protótipo cria a estrutura de uma função linear que é interpretada de acordo com os coeficientes encontrados.

A análise deste laboratório segue a seguinte situação hipotética: Um grupo de estudiosos de plantas orquídeas querem identificar uma fórmula matemática que defina o tamanho de uma flor, aqui chamada no banco de dados de *petalwidth***.** Para isto, será utilizada uma técnica de mineração de dados chamado de **regressão** que busca um relacionamento entre atributos a fim de compor uma expressão matemática que explique um determinado atributo-alvo.

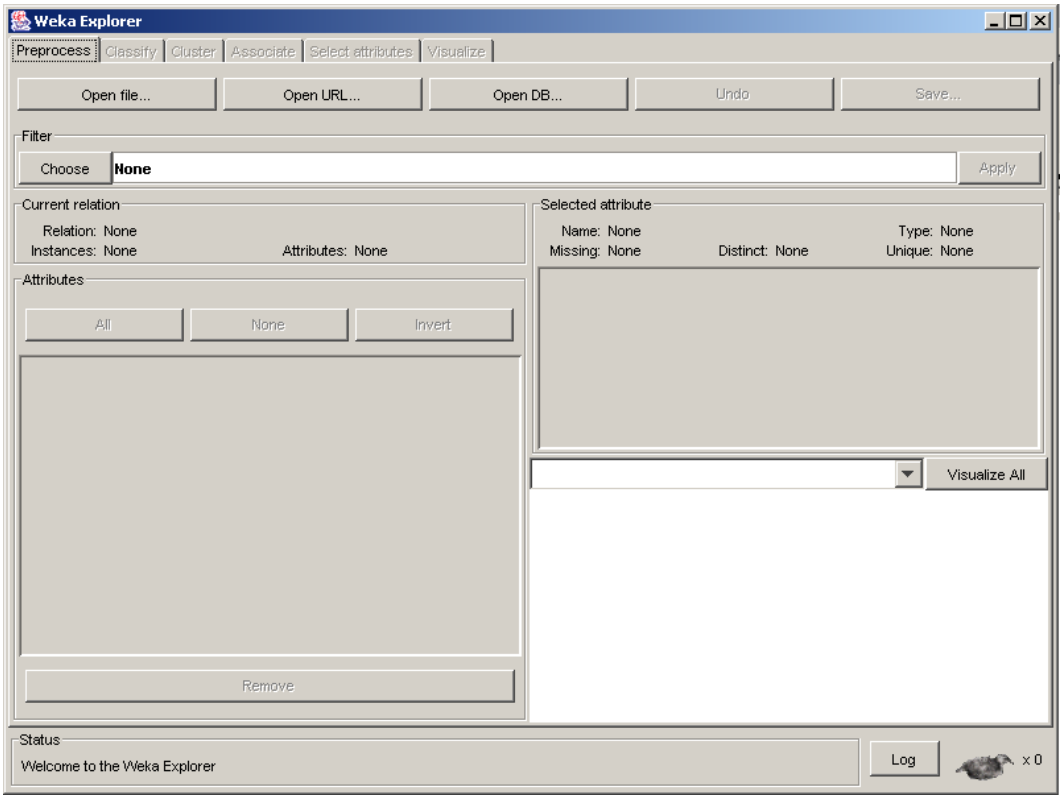
Para isto, foram coletados diversos dados sobre flores de orquídeas e disponibilizadas em um arquivo denominada **íris.arff**.

Primeiro abriremos o arquivo íris, disponível no site do Moodle, no formato ARFF, pois é o tipo de arquivo que o WEKA consegue interpretar.

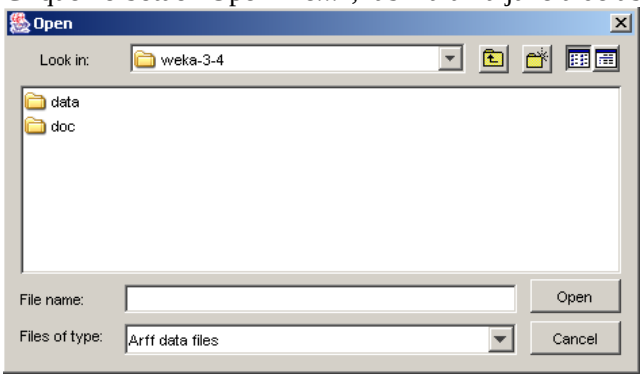
O resultado é o seguinte:

```
@RELATION iris

@ATTRIBUTE sepallength     REAL
@ATTRIBUTE sepalwidth      REAL
@ATTRIBUTE petallength     REAL
@ATTRIBUTE petalwidth      REAL
@ATTRIBUTE class    {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
...
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
...
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
```

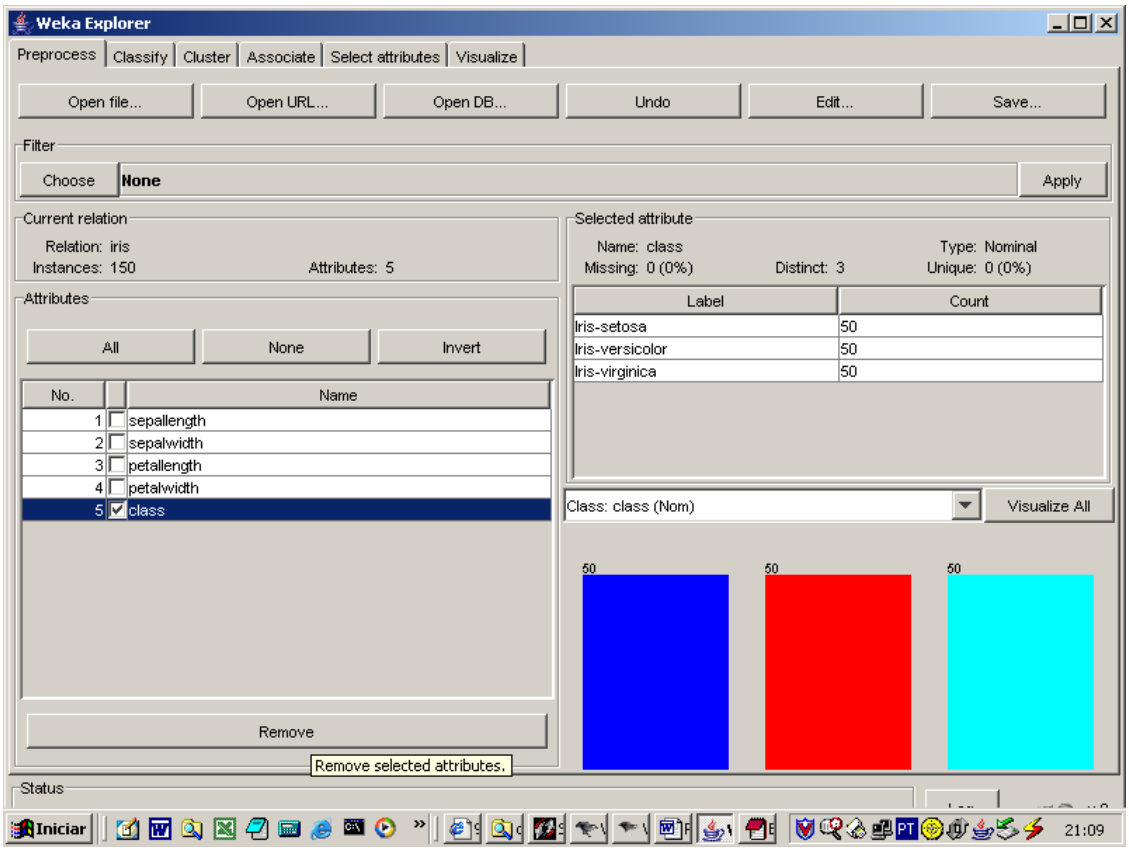
Clicamos no botão “Explorer”, logo será aberto a seguinte janela:



Clique no botão “Open file...”, abrirá uma janela de abertura de arquivo:



Coloque a localização do arquivo iris.arff e clique no botão “open”, a janela WEKA Explorer mostra os campos, marcamos todos os campos para serem minerados:



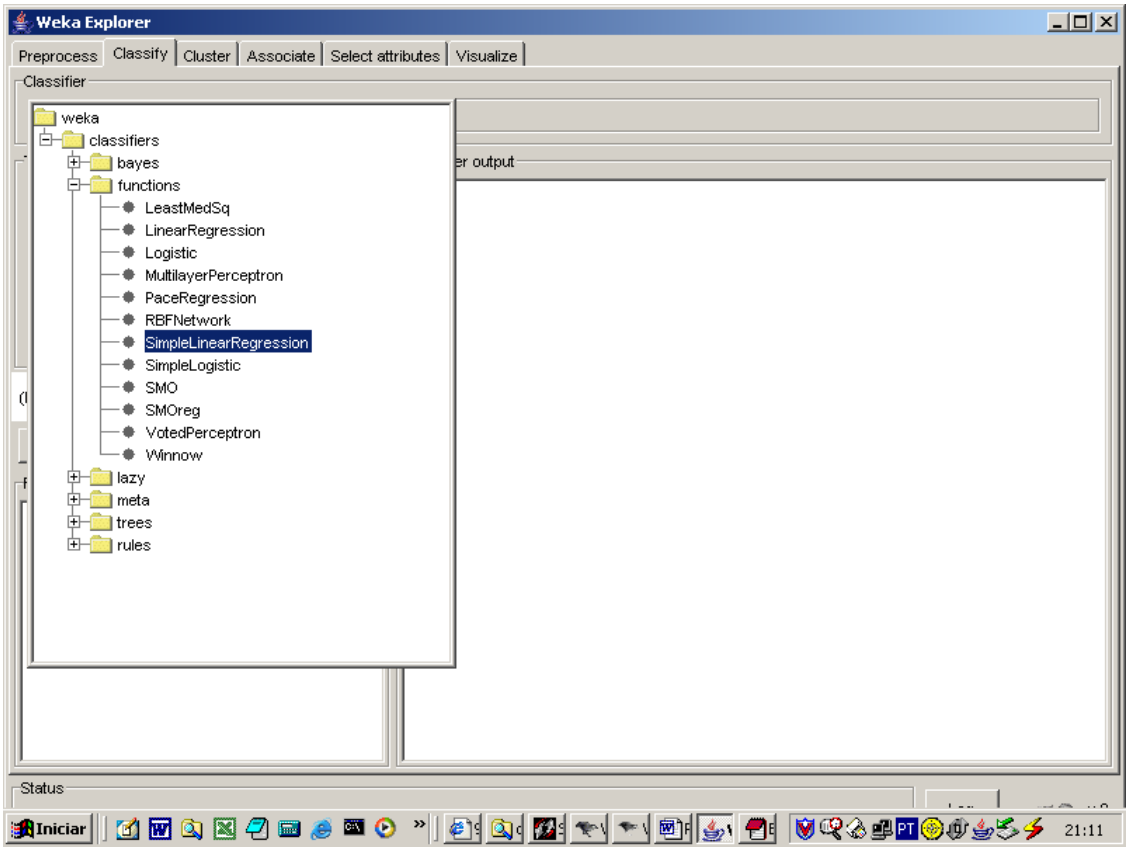
Localize no WEKA a informação de quantos atributos ele possui.

Veja no WEKA quantas instâncias existe no conjunto de dados digitado.

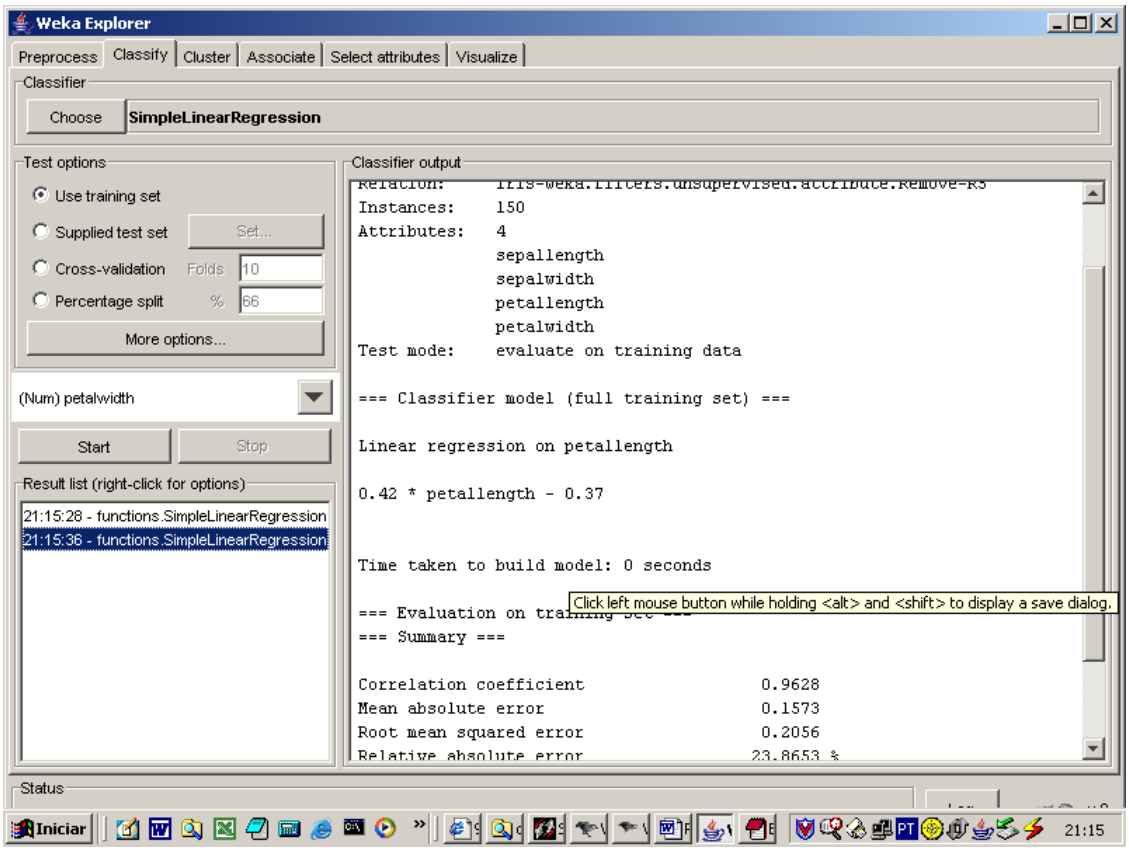
Clique no botão Visualize All. Conforme o laboratório anterior, veja se você já consegue minerar alguma informação sobre as flores.

Como o algoritmo só aceita dados numéricos, vamos remover o campo ***class***. Selecione-o e clique o botão **REMOVE**.

Para selecionar o tipo do algoritmo que foi implementado clicamos no botão “choose”, para isso selecione dentro da pasta weka, a pasta classifiers, dentro dela functions e logo após SimpleLinearRegression. Como mostra a figura abaixo:



Agora que selecionamos o tipo do algoritmo, vamos clicar em “use training test”, para dizer que estamos na fase de treinamento do método de classificação, logo após devemos escolher o campo que será utilizado para explicar a regressão linear, no caso escolheremos tamanho da pétala(Petalwidth), lembrando que a interpretação matemática será y= ax + b, como mostra a figura a seguir:



Logo acima vemos a equação da regressão que explica o tamanho da pétala:

0.42 * petallength - 0.37

Analisamos o coeficiente de correlação se está próximo de 1, é um indicativo de que os dados estão bem ajustados.

Para testar, pegue uma linha qualquer do banco de dados e calcule o tamanho da pétala conforme a fórmula apresentada.

Selecione um outro atributo para encontrar a fórmula e compare seus resultados com o banco de dados original.

Selecione outros algoritmos de regressão, tais como o **PaceRegression** e o **LinearRegression**. As fórmulas encontradas foram melhores ou piores? Justifique sua resposta sob o ponto de vista de um consultor que deseja verificar todas as possibilidades possíveis de análise.

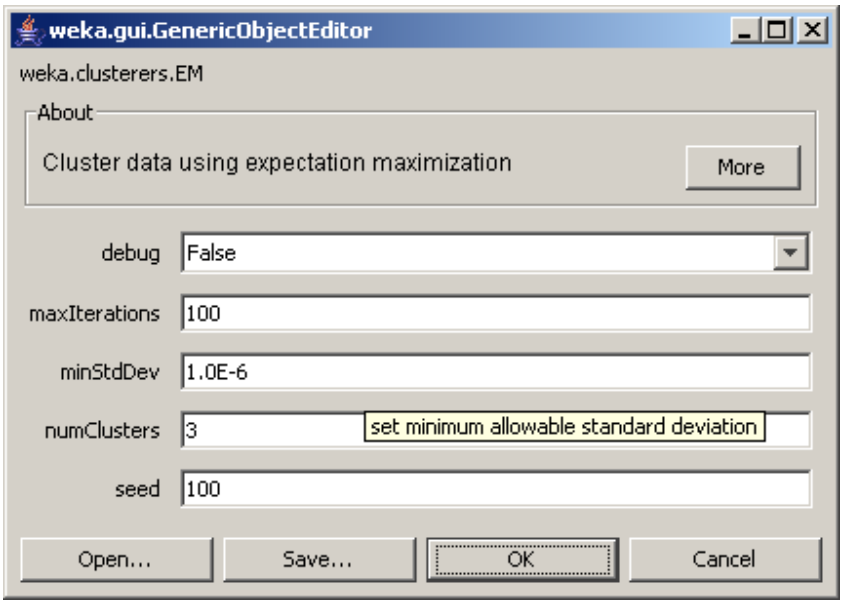
Escreva abaixo o que você descobriu minerando os dados com esta técnica de mineração. Que outros usos pode ela ter?

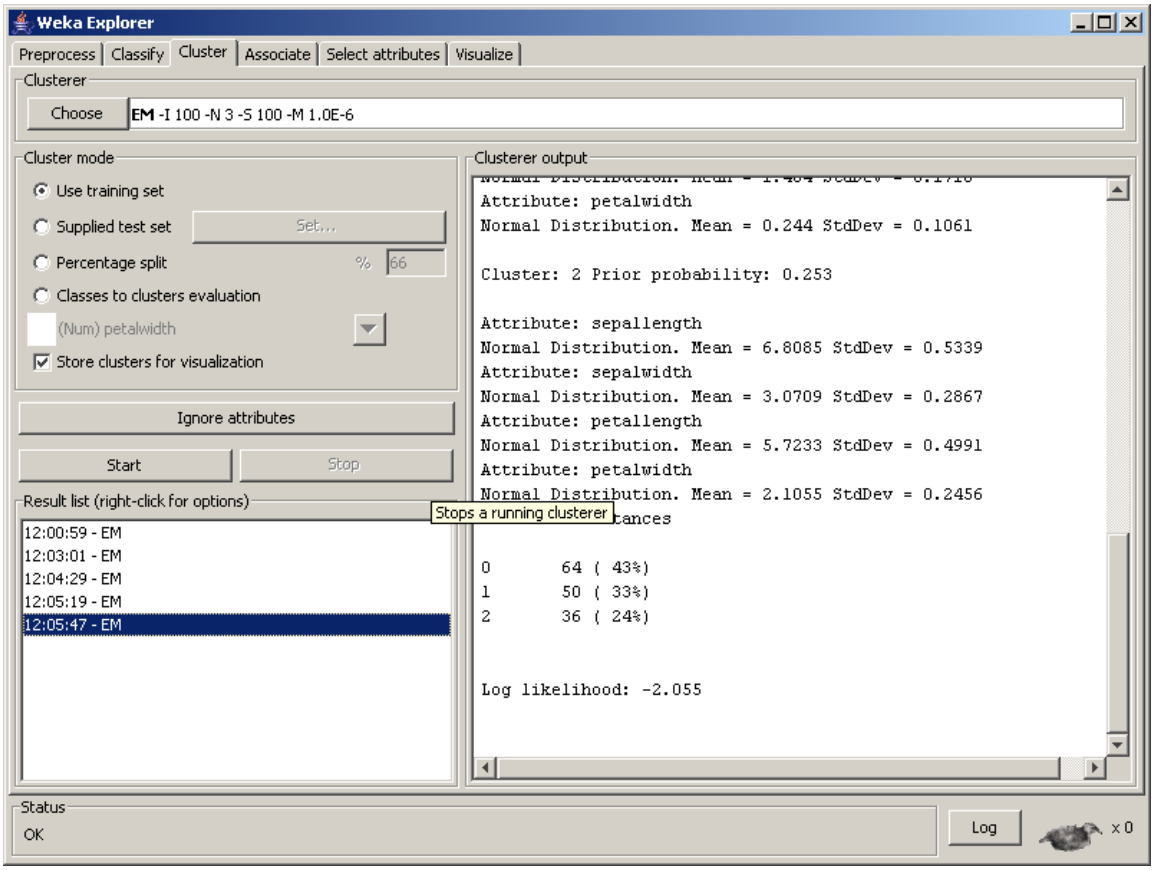
Examinando Aglomerações.

Carregue o arquivo **iris.arff** que representa dados sobre orquídeas.

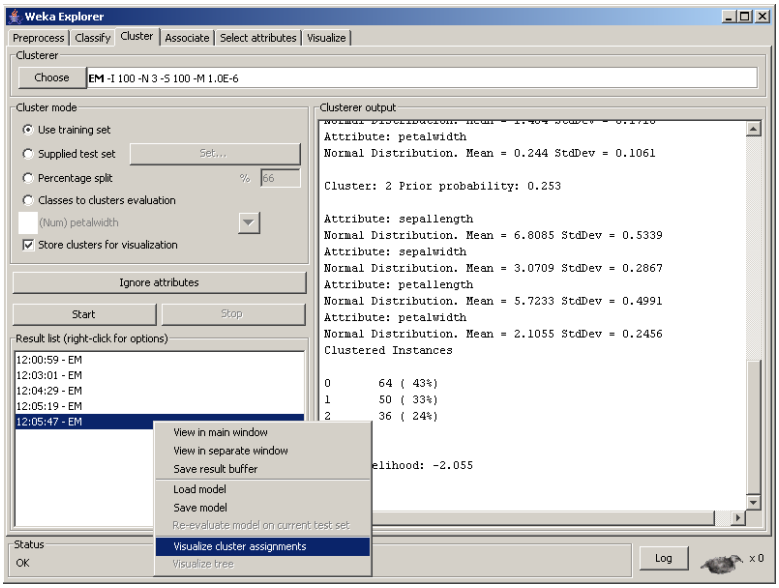
Clique na aba **cluster** e selecione o algoritmo EM.

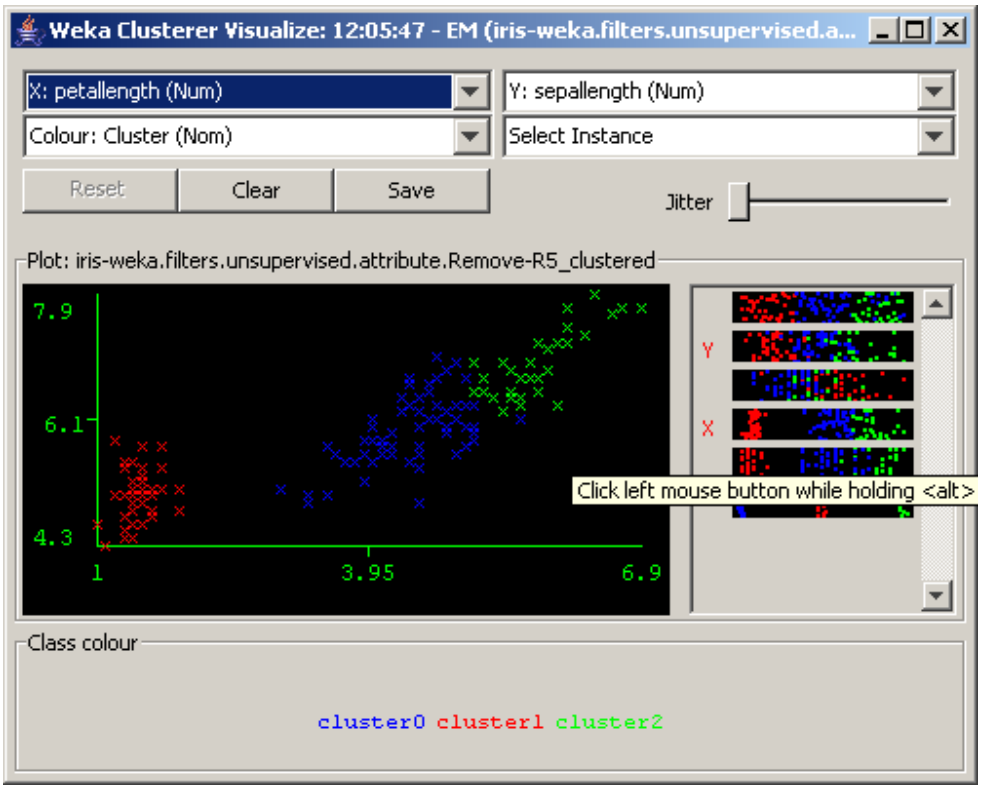
Clique sobre a palavra EM na tela. Configure o número de clusters para 3.





Analisamos o Log likelihood, quando maior for esse valor, melhor será a definição dos grupos, logo após clicamos no botão “direito”, sobre o “Result List” e solicitamos para exibir o gráfico dos cluster.





Clique nas figuras ao lado do gráfico. Ela mostra os cluster de acordo com a orientação dos atributos no eixo X e Y.

Em sua opinião, quais atributos melhor separam os clusters ?

Clique em um dos elementos do cluster no gráfico exibido. Será aberta uma tela com seus atributos. Tente identificar no banco de dados se os cluster conseguiram separar os três tipos de flores (íris-setosa, íris-versicolor e íris-virginica). Dê exemplos abaixo de casos do cluster corretamente separados.

Escreva abaixo o que você descobriu minerando os dados com esta técnica de mineração. Que outros usos pode ela ter?

QUESTÃO DESAFIO

Consiga uma base de dados qualquer e aplique as técnicas de regressão e algomeração. Escreva o que você achou durante a mineração de dados. Esta resposta deve contemplar uma explicação sobre a base de dados e a explicação dos conhecimentos aprendidos a partir da mineração de dados.