**Aula Prática 3**

**Tema: Ferramenta Weka**

**Murielly Oliveira Nascimento – 11921BSI222**

1. Exercícios a serem resolvidos usando a ferramenta Weka
   1. Faça o download da ferramenta weka e instale-a na sua máquina

<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

* 1. Explore a ferramenta usando os seguintes filtros em pelo menos uma base de dados.
     1. Replace Missing Values: filtro que substitui os valores ausentes da base de dados. Pelo menos uma das bases usadas tem que ter valores ausentes a ser substituídos.
     2. Resample (supervisionado): filtro que cria uma amostra da base de dados. O tamanho da amostra deve ser igual à 50% do tamanho da base original.
     3. Discretize: filtro que discretiza um atributo numérico em nominal.
     4. Nominal to Binary: filtro que converte um atributo nominal para binário.
     5. Normalize: converte um atributo numérico para uma dada escala. Vamos trabalhar com a escala [0,1]
  2. Para testar os filtros escolha bases de dados do site UCI ou de outro repositório. Use bases de dados no formato arff. Use quantas bases forem necessárias para testar cada um dos 5 filtros.

<https://archive.ics.uci.edu/ml/datasets.php>

* 1. É importante que você descreva cada uma das bases usadas. É preciso saber o nome da base, quais são seus atributos e o tipo de cada atributo.

A base escolhida para fazer as análises é a de íris. Os atributos são sepallength, sepalwidth, petallength e petalwidth do tipo numérico e class do tipo nominal com os valores: Iris-setosa, Iris-versicolor, Iris-virginica.

* 1. A cada filtro aplicado em uma base, faça um print screen da tela antes e depois da aplicação do filtro ressaltando a mudança nos dados da base com a aplicação do filtro. Indique o que foi alterado após a aplicação do filtro (ou seja, como era o dado antes e após a aplicação do filtro). Explique em detalhes, comente o que ocorreu.
* ReplaceMissingValues

Na base de dados Iris não tinha valores faltando, por isso, usei o filtro ReplaceWithMissingValue primeiro. Note que 16 valores de sepallength, 11 de sepalwidth, 23 de petallength e 16 de petalwidth sumiram.

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Agora aplicando o filtro ReplaceMissingValues obtemos o seguinte resultado.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Note que antes a maior sepallength da classe Iris Virginica era 32 e agora passou a ser 48. Isso vale para os demais atributos. Em outras palavras, os missing values foram substituídos por instâncias com valores na média das instâncias.

* Resample

Antes

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Depois

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Com esse filtro metade dos dados selecionados. Veja que no gráfico de classes temos 25 instâncias para cada valor: Iris-Virginica, Setosa e Versicolor.

* Discretize

Antes

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Depois

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Os atributos sepallength, sepalwidth, petallength e petalwidth são separados em três rótulos de acordo com intervalos. Para sepallength, por exemplo, temos (-infinito, 5.55], (5.55, 6.15] e (6.15, inifinito).

* Nominal to Binary

Antes.

Nesse caso, estou usando os dados após a aplicação do filtro Discretize.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Depois

Diagrama

Descrição gerada automaticamente

Observa-se que os rótulos sepallength, sepalwidth, petallength e petalwidth que haviam sido divididos em três cada de acordo com intervalos foram convertidos para valores binários. O máximo e o mínimo de cada atributo é 1 e 0, respectivamente.

* Normalize

Antes

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Depois

Gráfico, Histograma

Descrição gerada automaticamente

Com esse filtro o valor máximo e mínimo para cada um dos atributos fica restrito a uma escala 0 a 1.