

## **Laboratório**

### **Preprocessamento com WEKA Explorer**

Faça o download do dataset **test\_credit.csv\*** , e execute as seguintes tarefas:

1. Use o **Weka Viewer** para ter uma visão geral do dataset original.
2. Substitua os **missing data** se houver algum.
3. Verifique se há **outliers** ou **valores extremos** no conjunto de dados.
4. Realize **normalização** de duas características.
5. Realize a **discretização** em duas características.
6. Realize a **substituição** em duas características.

\*disponível em: [www.inf.ufpr.br/menotti/am-231/data.zip](http://www.inf.ufpr.br/menotti/am-231/data.zip)

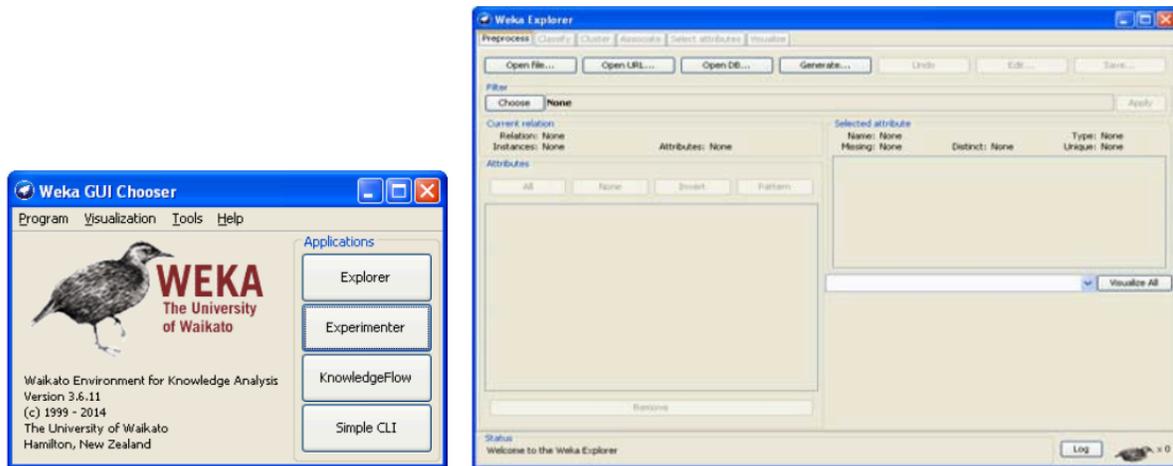
Nas páginas abaixo, encontram-se **tutoriais** explicando / ilustrando cada passo:

## 1. Visualização dos dados bruto (raw data)

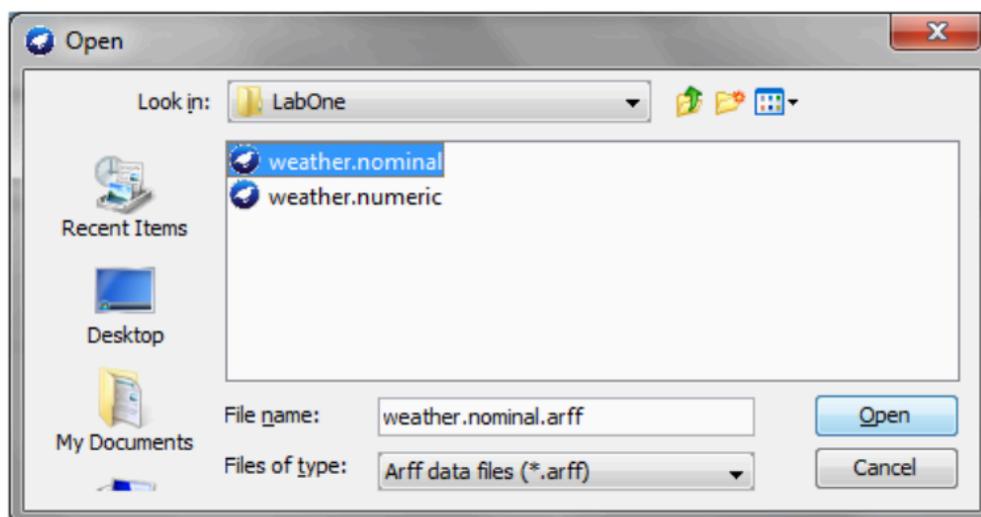
1. Inicie uma sessão do **Weka** ou execute em linha de comando:

```
java -jar weka.jar.
```

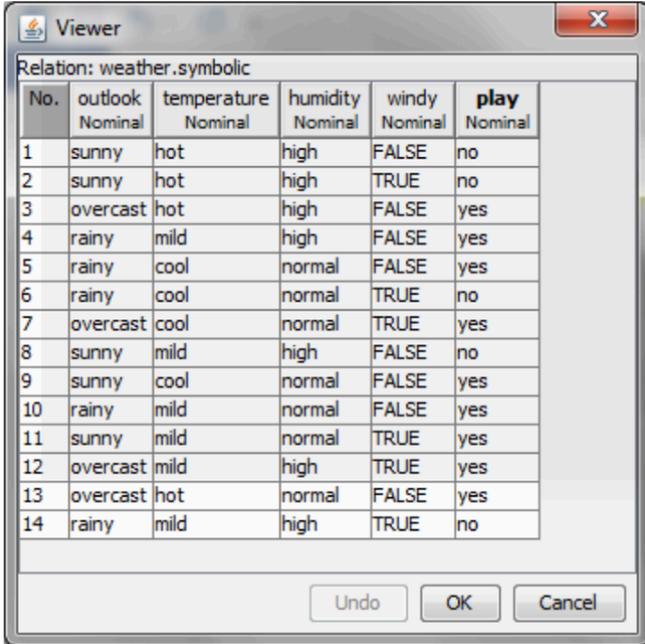
2. Quando a **GUI Chooser** surgir, selecione o **Explorer** a partir das quatro opções do lado direito.



3. A tela acima é a principal do Explorer. Existem 6 guias no topo do aplicativo que representam as operações básicas que o Explorer suporta. Agora, estamos no **Preprocess**. Clique no botão **Open file** para abrir a janela padrão de diálogo através da qual você pode selecionar um arquivo. Escolha o arquivo **weather.nominal.arff**. Se você tem um arquivo no formato **CSV**, modifique de “**ARFF** data files” para “**CSV** data files” em “Files of type”. Quando você especificar um arquivo **.csv** ele é convertido automaticamente para o formato ARFF.

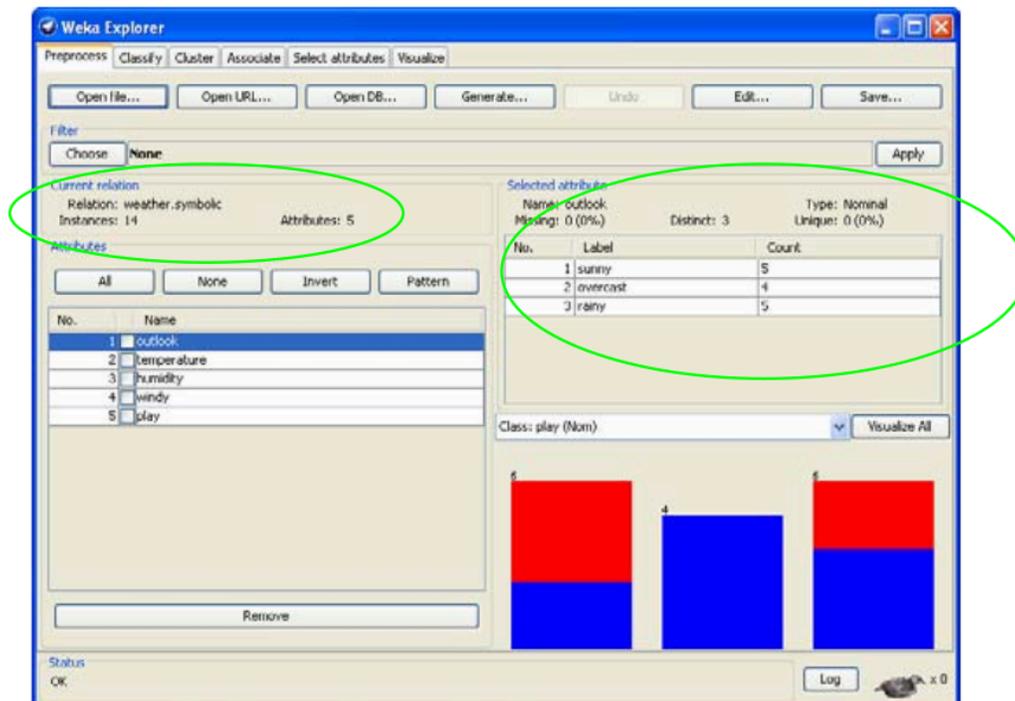


4. Para visualizar todo o dataset, clique no botão **Edit**, e então uma janela de visualização será aberta com o dataset carregado.

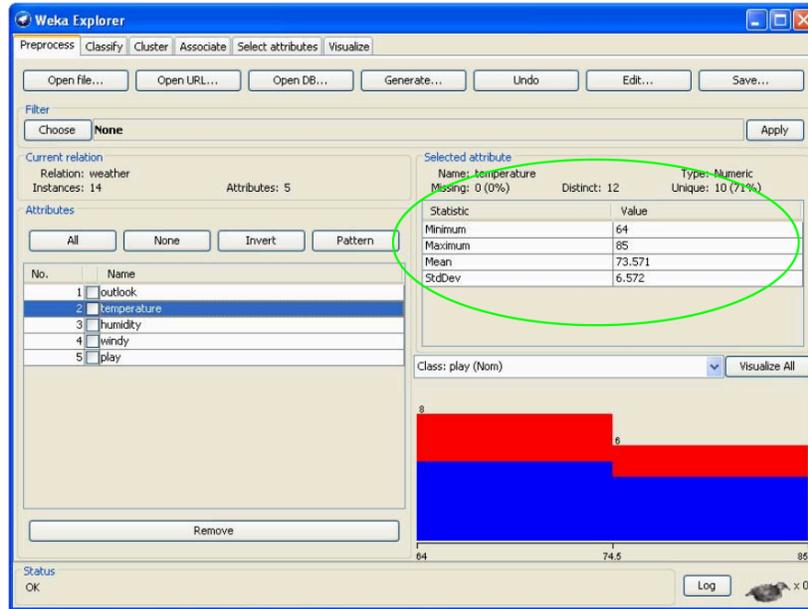


| No. | outlook<br>Nominal | temperature<br>Nominal | humidity<br>Nominal | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | hot                    | high                | FALSE            | no              |
| 2   | sunny              | hot                    | high                | TRUE             | no              |
| 3   | overcast           | hot                    | high                | FALSE            | yes             |
| 4   | rainy              | mild                   | high                | FALSE            | yes             |
| 5   | rainy              | cool                   | normal              | FALSE            | yes             |
| 6   | rainy              | cool                   | normal              | TRUE             | no              |
| 7   | overcast           | cool                   | normal              | TRUE             | yes             |
| 8   | sunny              | mild                   | high                | FALSE            | no              |
| 9   | sunny              | cool                   | normal              | FALSE            | yes             |
| 10  | rainy              | mild                   | normal              | FALSE            | yes             |
| 11  | sunny              | mild                   | normal              | TRUE             | yes             |
| 12  | overcast           | mild                   | high                | TRUE             | yes             |
| 13  | overcast           | hot                    | normal              | FALSE            | yes             |
| 14  | rainy              | mild                   | high                | TRUE             | no              |

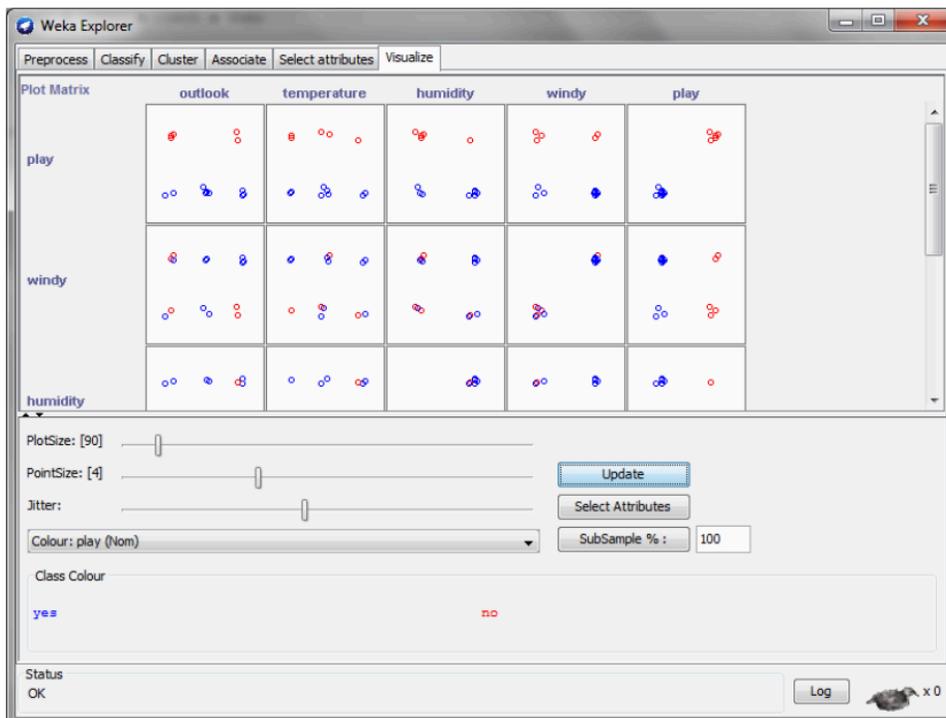
5. O primeiro atributo, **outlook**, é selecionado por default. As características deste atributo são apresentadas. Um histograma no canto inferior direito mostra o quão frequente cada um dos dois valores da classe **play** ocorre para cada valor do atributo **outlook**. Você pode visualizar esta análise para outros atributos bastando realizar a seleção na esquerda.



6. Se você abrir o outro arquivo Weather, *weather.numeric.arff*, a visualização dos atributos é diferente. Selecionando-se o segundo atributo, *temperature*, você visualiza seus valores máximos e mínimos, bem como a média e o desvio padrão. O histograma apresenta a distribuição da classe como uma função deste atributo.



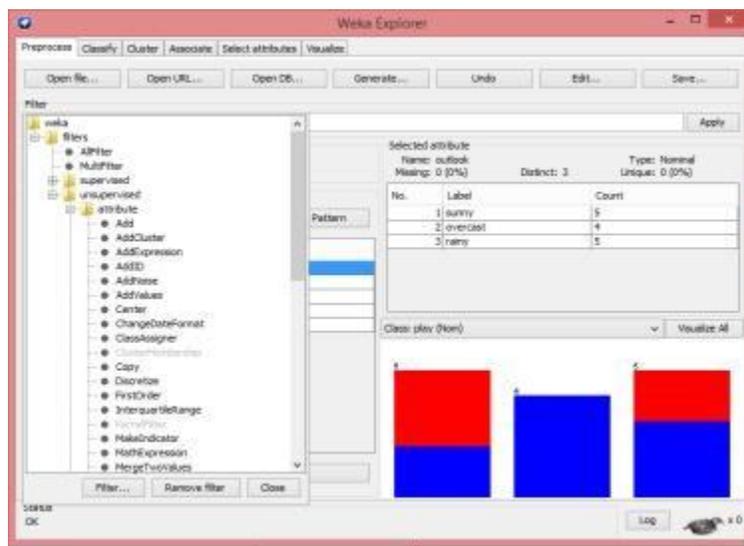
7. Clique na guia **Visualize** para visualizar gráficos 2D do dataset.



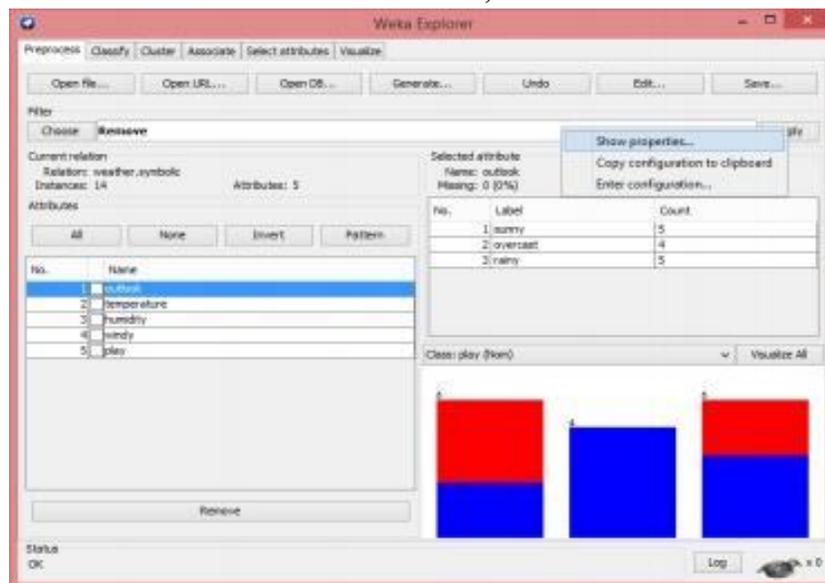
## 1.1 Usando Filtros para Remover Atributos

**Unsupervised Attribute Filter – Remove:** Este filtro remove/deleta atributos específicos de um dataset. O mesmo efeito pode ser obtido mais facilmente selecionando-se os atributos relevantes usando as **tick boxes** e então pressionando-se o botão **Remove**.

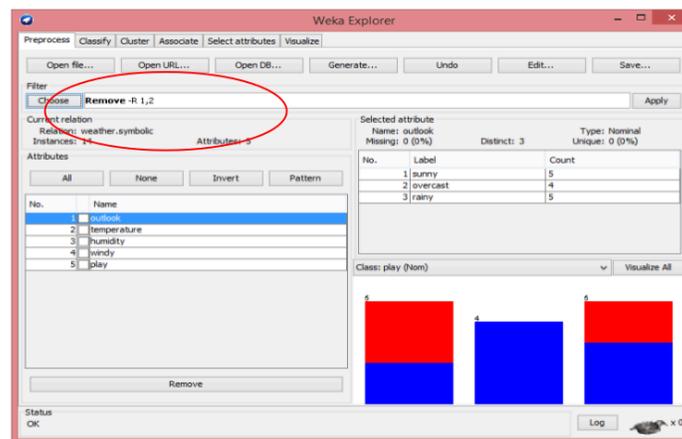
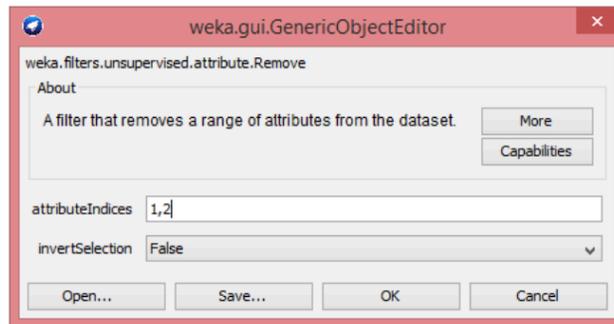
1. Abra um dataset, tal como o *weather.nominal* dataset.
2. Clique no botão **Choose** dentro da caixa **Filter** (acima a esquerda).  
E então clique em: **filters** => **unsupervised** => **attribute** => **Remove**.



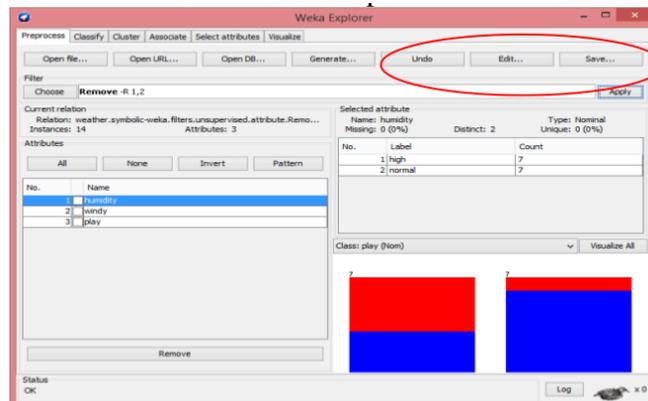
3. Clique com o botão **Direito** sobre a caixa **Remove**, e então escolha **Show Properties**.



4. Existem duas opções para o filtro **Remove**. Uma opção é **attributeIndices** que especifica a faixa de atributos a ser removida. (No exemplo, os índices **1,2** – **outlook** and **temperature** - foram escolhidos). A outra opção é **invertSelection** que determina se o filtro seleciona ou deleta os atributos (Foi selecionado **False** (*default*), que indica para remover ao invés de selecioná-los). E então você clica em **OK**.



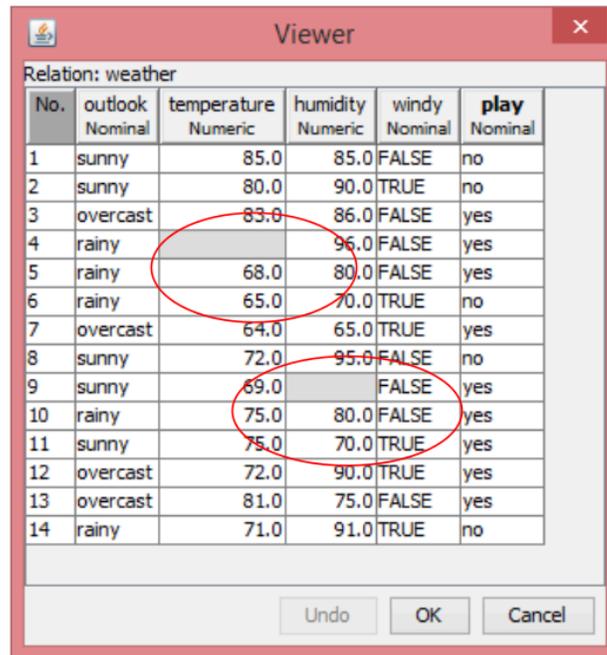
5. Clique no botão **Apply** próximo da caixa do filtro **Remove**, e então os dois primeiros atributos são removidos do dataset, e apenas três sobram. Você pode clicar no botão **Undo** para desfazer a operação de filtragem e restaurar o dataset original. Você pode também clicar no botão **Save** para gravar o dataset processado.



## 2. Manipulando *missing data*

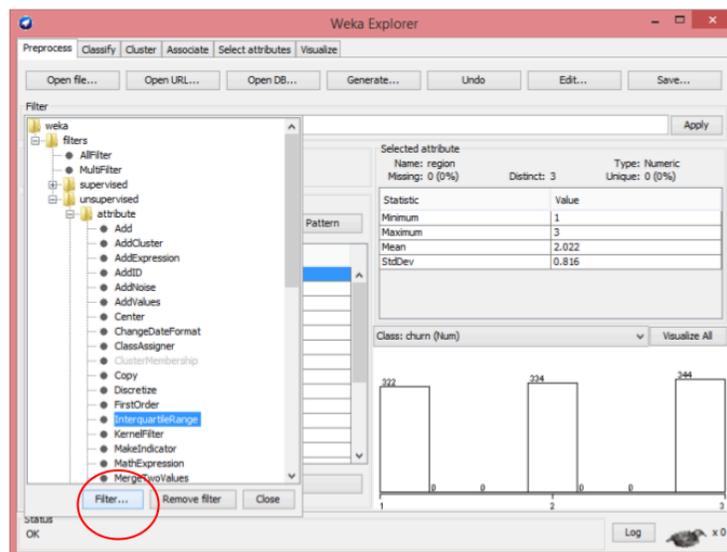
**Unsupervised Attribute Filter – ReplaceMissingValues:** Estes filtro substitui todos os valores faltantes (*missing values*) para atributos nominais e numéricos com a **moda** para atributos nominais e a **média** para atributos numéricos com base nos dados de treinamento.

1. Abra o dataset – **weather.numeric.arff**. Clique no botão **Edit** para visualizar os dados brutos. Você pode verificar que dois dos atributos têm *missing values*.



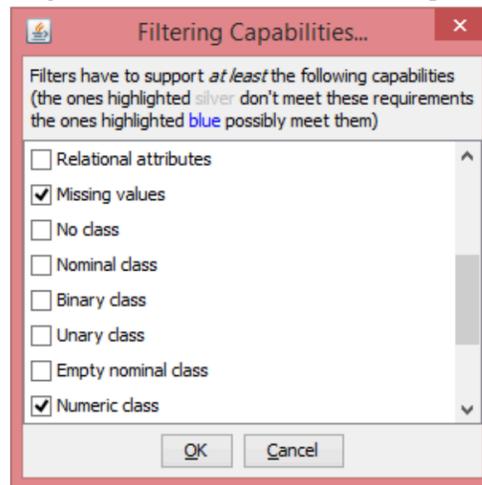
| No. | outlook<br>Nominal | temperature<br>Numeric | humidity<br>Numeric | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | 85.0                   | 85.0                | FALSE            | no              |
| 2   | sunny              | 80.0                   | 90.0                | TRUE             | no              |
| 3   | overcast           | 83.0                   | 86.0                | FALSE            | yes             |
| 4   | rainy              |                        | 96.0                | FALSE            | yes             |
| 5   | rainy              | 68.0                   | 80.0                | FALSE            | yes             |
| 6   | rainy              | 65.0                   | 70.0                | TRUE             | no              |
| 7   | overcast           | 64.0                   | 65.0                | TRUE             | yes             |
| 8   | sunny              | 72.0                   | 95.0                | FALSE            | no              |
| 9   | sunny              | 69.0                   |                     | FALSE            | yes             |
| 10  | rainy              | 75.0                   | 80.0                | FALSE            | yes             |
| 11  | sunny              | 75.0                   | 70.0                | TRUE             | yes             |
| 12  | overcast           | 72.0                   | 90.0                | TRUE             | yes             |
| 13  | overcast           | 81.0                   | 75.0                | FALSE            | yes             |
| 14  | rainy              | 71.0                   | 91.0                | TRUE             | no              |

2. Clique no botão **Choose** dentro da caixa **Filter**. Clique no botão **Filter** na parte de baixo da janela *drop-down*.

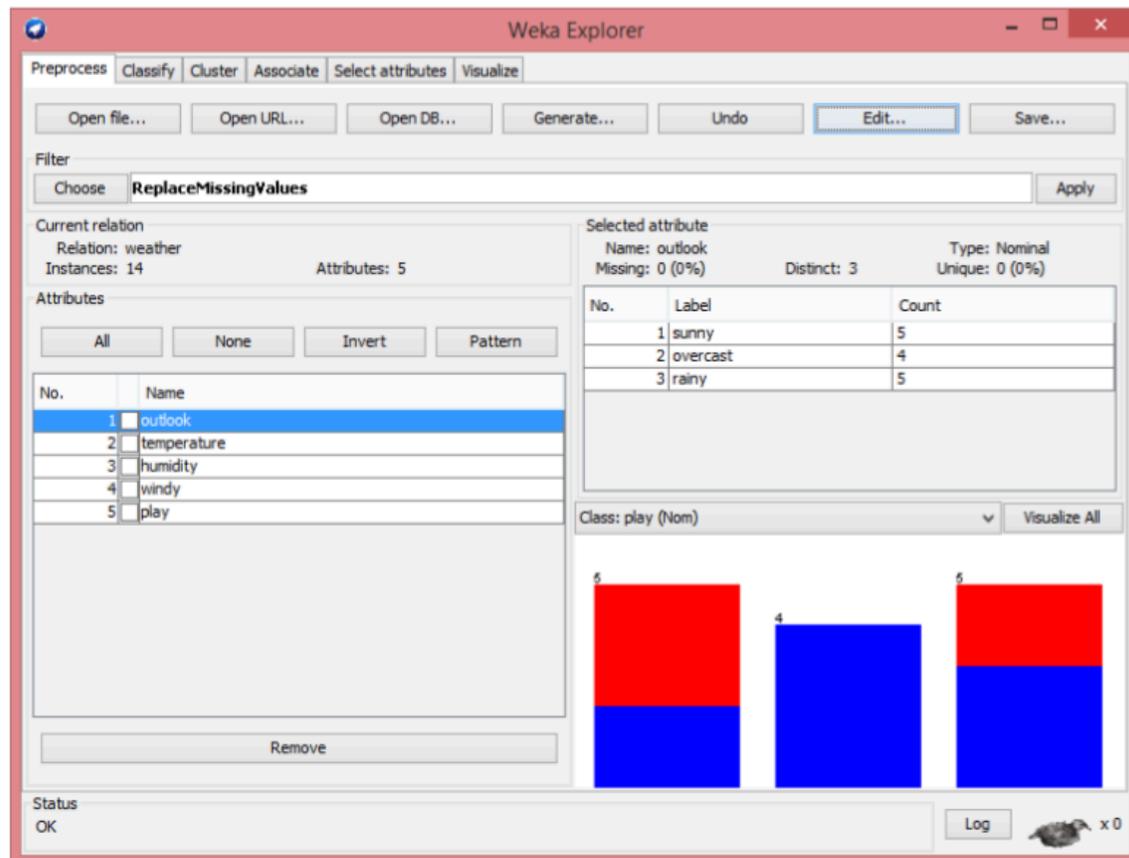


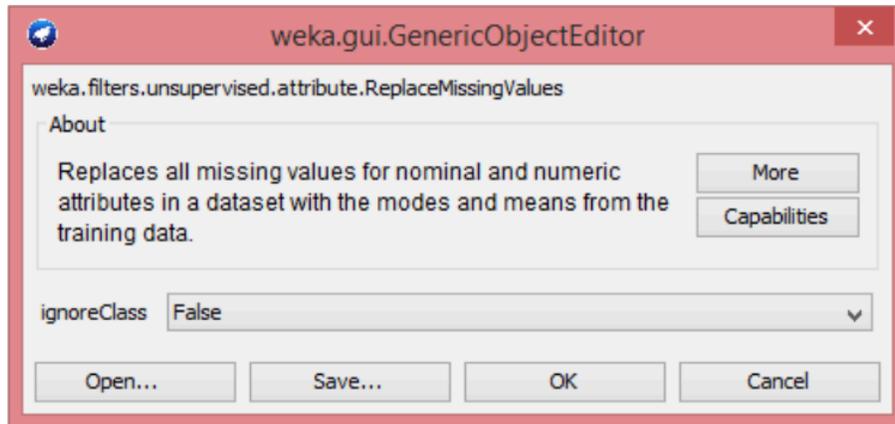
The screenshot shows the Weka Explorer interface. The 'Filter' panel is open, displaying a tree view of filters. The 'ReplaceMissingValues' filter is selected. The 'Filter...' button at the bottom of the panel is circled in red. The main window shows a histogram for the 'churn' attribute, with a bar chart showing the distribution of values (1, 2, 3) and their corresponding counts (322, 324, 344).

3. Uma janela chamada **Filtering Capabilities** se abrirá. Esta janela mostra qual tipo de atributos os filtros suportam. Certifique-se de que apenas os tipos **Numeric Attributes**, **Missing values** e **Numeric Class** estejam selecionados. E então clique em **OK**.



4. Escolha o filtro **ReplaceMissingValues** a partir da lista *drop-down*  
Para isso clique em: **filters** => **unsupervised** => **attribute** => **ReplaceMissingValues**.  
E então clique na caixa **Filter** para mostrar a janela de propriedade do filtro selecionado.





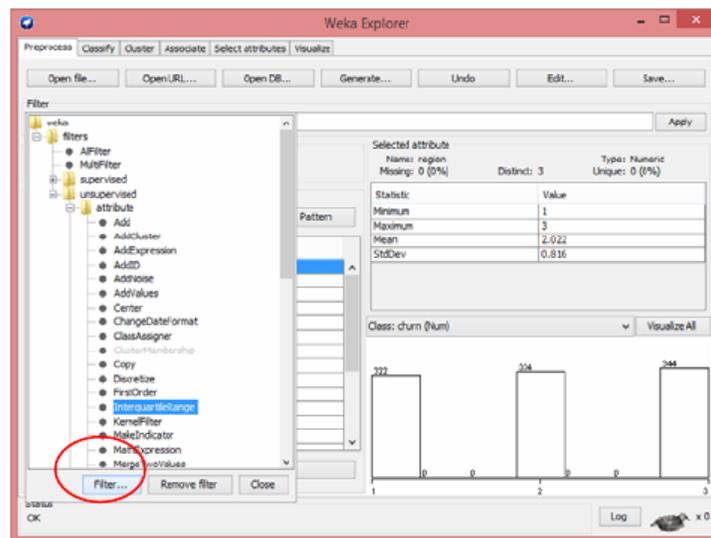
5. Clique no botão **Apply** dentro da caixa **Filter**. Então clique no botão **Edit** para verificar se o dataset foi processado – você verá que os *missing values* foram preenchidos. Grave os dados modificados (clique no botão **Save** na tela principal). Escolha um nome diferente para salvá-lo de forma que o dataset original seja mantido - **weather.numeric.nomissing.arff**

| No. | outlook<br>Nominal | temperature<br>Numeric | humidity<br>Numeric | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | 85.0                   | 85.0                | FALSE            | no              |
| 2   | sunny              | 80.0                   | 90.0                | TRUE             | no              |
| 3   | overcast           | 83.0                   | 86.0                | FALSE            | yes             |
| 4   | rainy              | 73.846153...           | 96.0                | FALSE            | yes             |
| 5   | rainy              | 68.0                   | 80.0                | FALSE            | yes             |
| 6   | rainy              | 65.0                   | 70.0                | TRUE             | no              |
| 7   | overcast           | 64.0                   | 65.0                | TRUE             | yes             |
| 8   | sunny              | 72.0                   | 95.0                | FALSE            | no              |
| 9   | sunny              | 69.0                   | 82.538...           | FALSE            | yes             |
| 10  | rainy              | 75.0                   | 80.0                | FALSE            | yes             |
| 11  | sunny              | 75.0                   | 70.0                | TRUE             | yes             |
| 12  | overcast           | 72.0                   | 90.0                | TRUE             | yes             |
| 13  | overcast           | 81.0                   | 75.0                | FALSE            | yes             |
| 14  | rainy              | 71.0                   | 91.0                | TRUE             | no              |

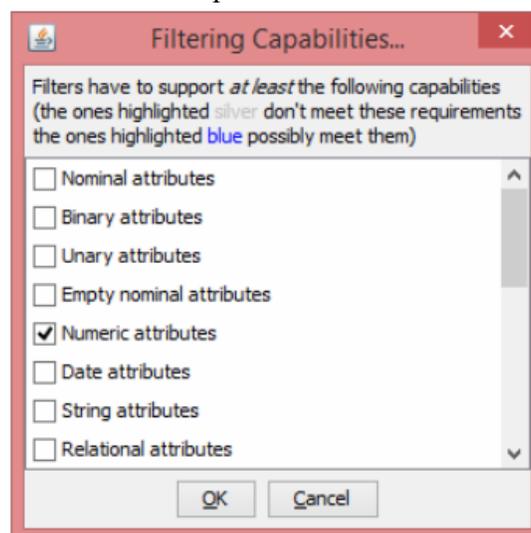
### 3. Usando Filtros para detectar/manipular *outliers* e *extreme values*

**Unsupervised Attribute Filter – InterquartileRange**: Este filtro adiciona novos **atributos** que indicam se valores de **instâncias** podem ser considerados **outliers** ou **extreme Values**.

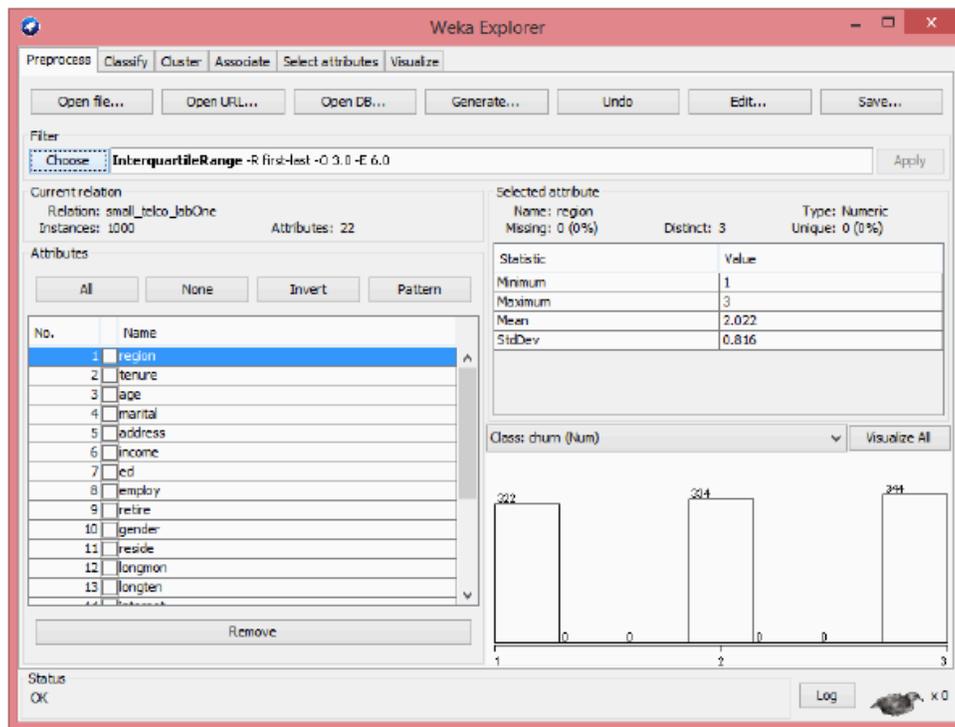
1. Abra o dataset – **small\_telco.csv**. Realize o passo de substituição de *missing values* com o filtro – **ReplaceMissingValues**. Observe que há um total de 22 atributos neste dataset.
2. Então clique no botão **Choose** dentro da caixa **Filter**. Clique no botão **Filter** na parte de baixo da janela *drop-down*.



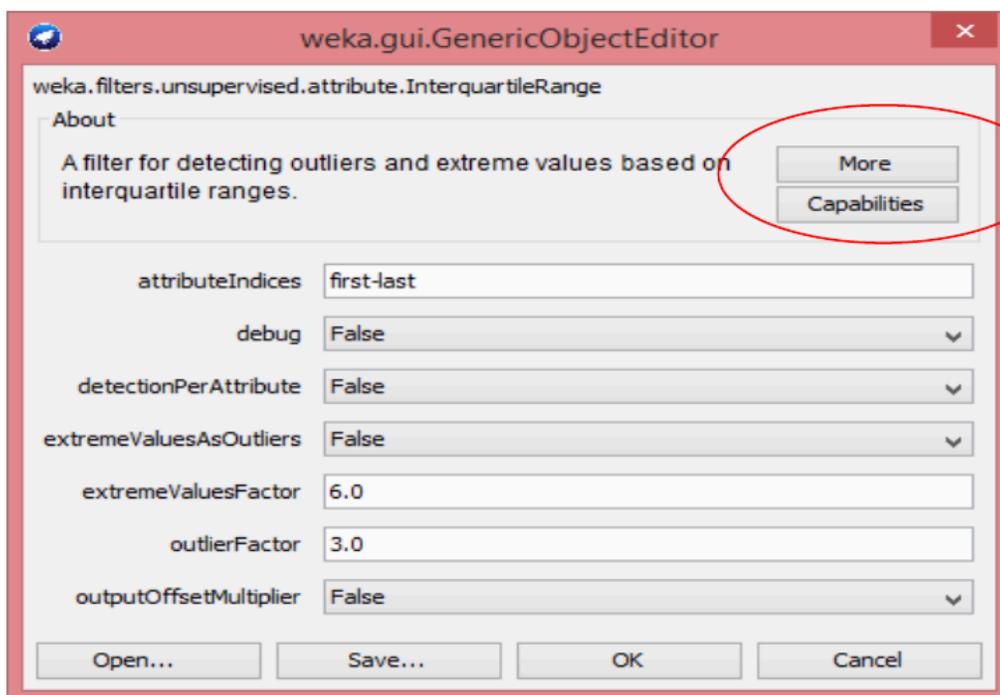
3. Uma janela chamada **Filtering Capabilities** se abrirá. Esta janela mostra qual tipo de atributos os filtros suportam. Certifique-se de que apenas os tipos **Numeric Attributes** e **Numeric Class** estejam selecionados. E então clique em **OK**.



4. Escolha o filtro **InterquartileRange** a partir da lista drop-down list de filtros não supervisionado para atributos. Para isso clique em: **filters** => **unsupervised** => **attribute** => **InterquartileRange**.



5. Clique (com o botão **Esquerdo**) dentro da caixa **Filter**, e então a janela de propriedades é apresentada. Clique no botão **More** para mostrar mais informações sobre este filtro.



Information

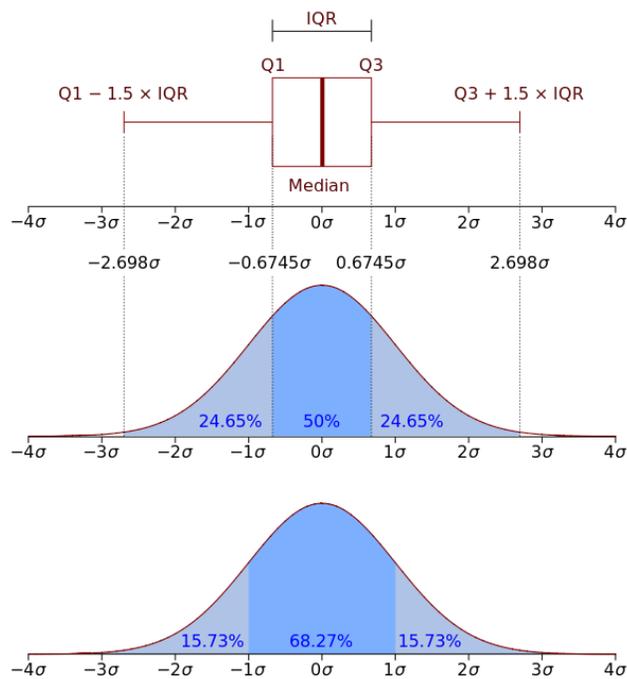
**NAME**  
weka.filters.unsupervised.attribute.InterquartileRange

**SYNOPSIS**  
A filter for detecting outliers and extreme values based on interquartile ranges. The filter skips the class attribute.

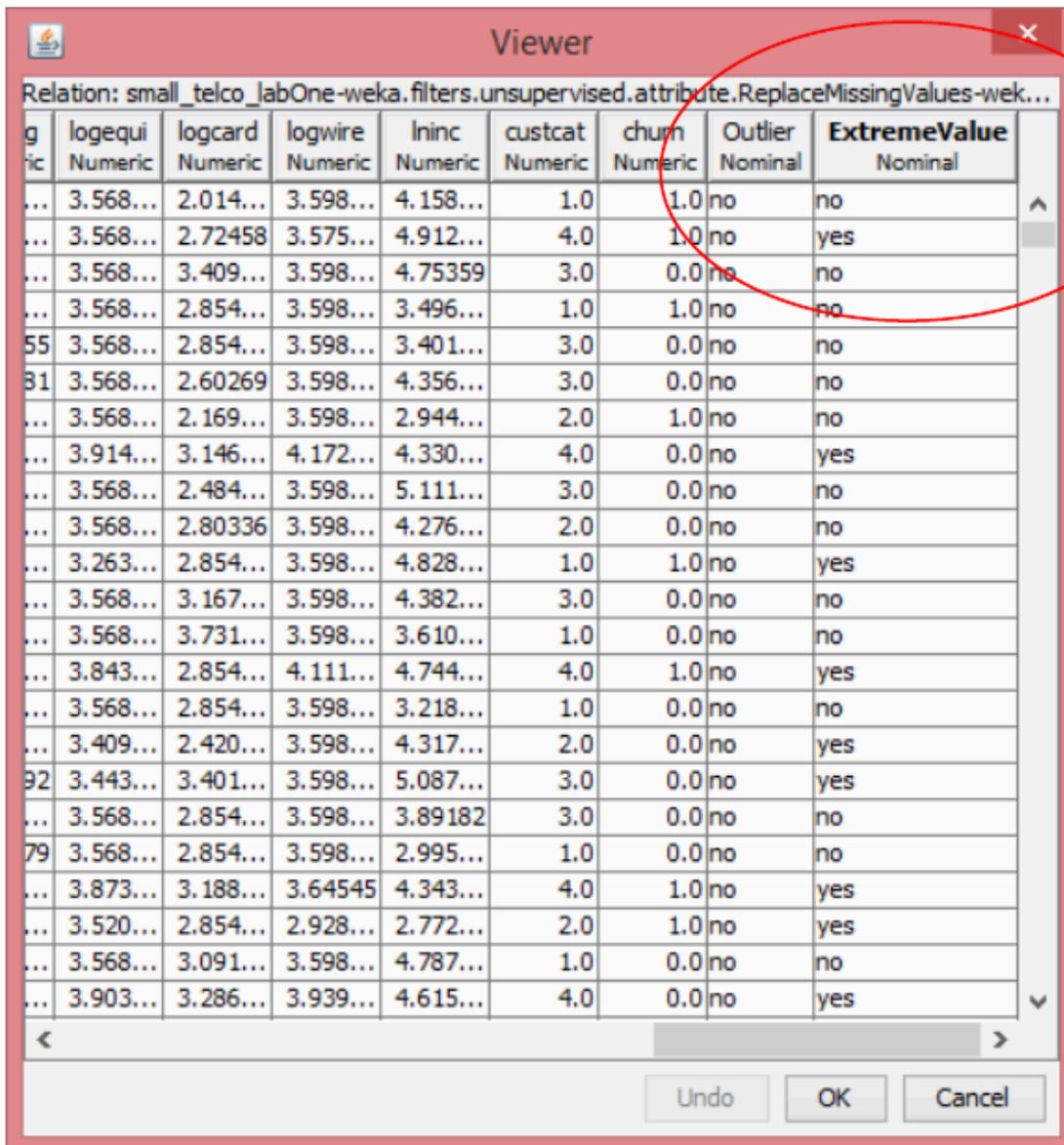
**Outliers:**  
 $Q3 + OF \cdot IQR < x \leq Q3 + EVF \cdot IQR$   
 or  
 $Q1 - EVF \cdot IQR \leq x < Q1 - OF \cdot IQR$

**Extreme values:**  
 $x > Q3 + EVF \cdot IQR$   
 or  
 $x < Q1 - EVF \cdot IQR$

Os fatores são usados para definir os *extreme values* e *outliers* de acordo com a definição de Q1, Q3 e IQR (veja ilustração abaixo).



6. Clique no botão **Apply** dentro da caixa **Filter**. Você encontrará dois atributos extras/novos que foram gerados. Estes dois atributos marcam uma instância como um **outlier** ou um **extreme value** se qualquer um dos seus valores de seus atributos são tidos como **outlier** ou **extreme value**.

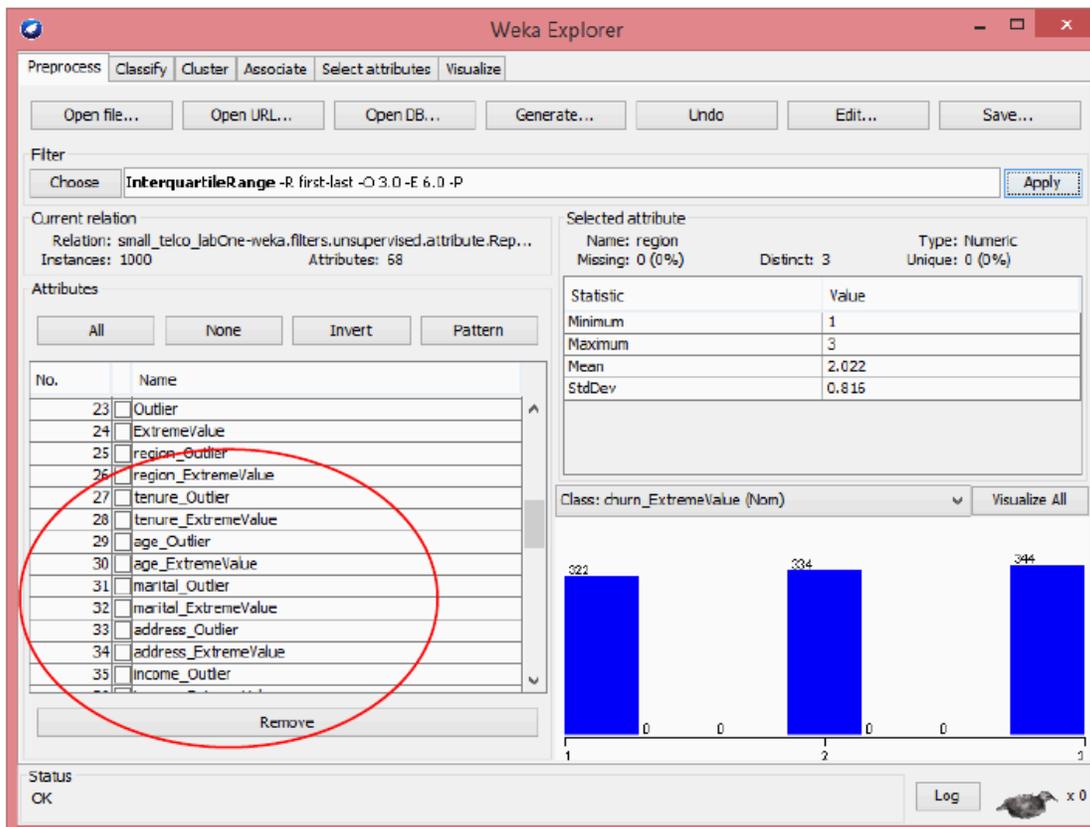
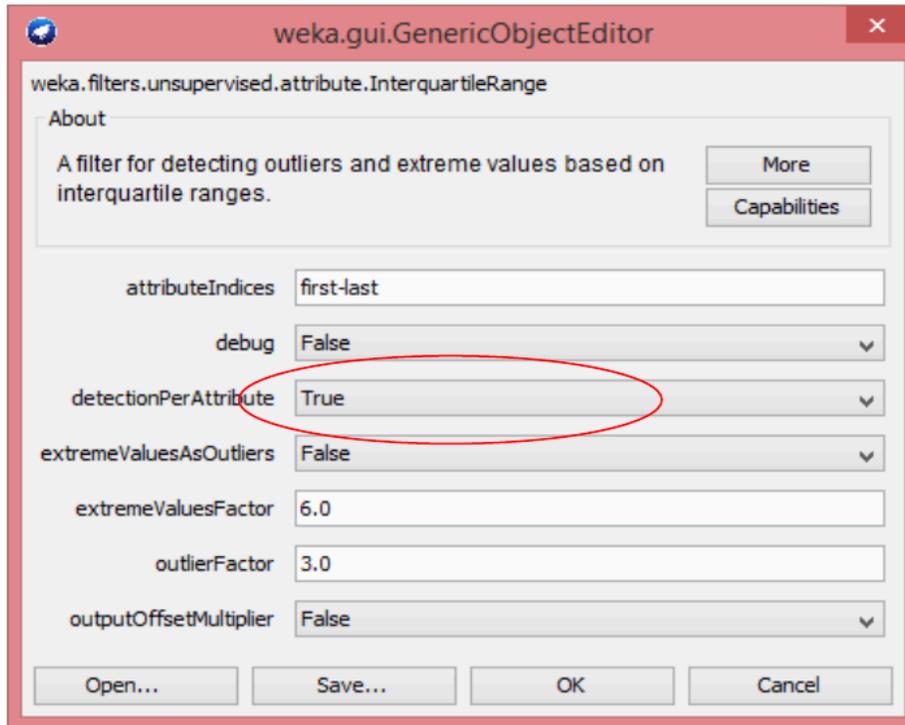


Relation: small\_telco\_labOne-weka.filters.unsupervised.attribute.ReplaceMissingValues-wek...

| g   | logequi  | logcard  | logwire  | lninc    | custcat | churn   | Outlier | ExtremeValue |
|-----|----------|----------|----------|----------|---------|---------|---------|--------------|
| ic  | Numeric  | Numeric  | Numeric  | Numeric  | Numeric | Numeric | Nominal | Nominal      |
| ... | 3.568... | 2.014... | 3.598... | 4.158... | 1.0     | 1.0     | no      | no           |
| ... | 3.568... | 2.72458  | 3.575... | 4.912... | 4.0     | 1.0     | no      | yes          |
| ... | 3.568... | 3.409... | 3.598... | 4.75359  | 3.0     | 0.0     | no      | no           |
| ... | 3.568... | 2.854... | 3.598... | 3.496... | 1.0     | 1.0     | no      | no           |
| 55  | 3.568... | 2.854... | 3.598... | 3.401... | 3.0     | 0.0     | no      | no           |
| 81  | 3.568... | 2.60269  | 3.598... | 4.356... | 3.0     | 0.0     | no      | no           |
| ... | 3.568... | 2.169... | 3.598... | 2.944... | 2.0     | 1.0     | no      | no           |
| ... | 3.914... | 3.146... | 4.172... | 4.330... | 4.0     | 0.0     | no      | yes          |
| ... | 3.568... | 2.484... | 3.598... | 5.111... | 3.0     | 0.0     | no      | no           |
| ... | 3.568... | 2.80336  | 3.598... | 4.276... | 2.0     | 0.0     | no      | no           |
| ... | 3.263... | 2.854... | 3.598... | 4.828... | 1.0     | 1.0     | no      | yes          |
| ... | 3.568... | 3.167... | 3.598... | 4.382... | 3.0     | 0.0     | no      | no           |
| ... | 3.568... | 3.731... | 3.598... | 3.610... | 1.0     | 0.0     | no      | no           |
| ... | 3.843... | 2.854... | 4.111... | 4.744... | 4.0     | 1.0     | no      | yes          |
| ... | 3.568... | 2.854... | 3.598... | 3.218... | 1.0     | 0.0     | no      | no           |
| ... | 3.409... | 2.420... | 3.598... | 4.317... | 2.0     | 0.0     | no      | yes          |
| 92  | 3.443... | 3.401... | 3.598... | 5.087... | 3.0     | 0.0     | no      | yes          |
| ... | 3.568... | 2.854... | 3.598... | 3.89182  | 3.0     | 0.0     | no      | no           |
| 79  | 3.568... | 2.854... | 3.598... | 2.995... | 1.0     | 0.0     | no      | no           |
| ... | 3.873... | 3.188... | 3.64545  | 4.343... | 4.0     | 1.0     | no      | yes          |
| ... | 3.520... | 2.854... | 2.928... | 2.772... | 2.0     | 1.0     | no      | yes          |
| ... | 3.568... | 3.091... | 3.598... | 4.787... | 1.0     | 0.0     | no      | no           |
| ... | 3.903... | 3.286... | 3.939... | 4.615... | 4.0     | 0.0     | no      | yes          |

Undo OK Cancel

7. Se nós mudamos a opção **detectionPerAttribute** do filtro **InterquartileRange**, de **False** para **True**, um par indicador **outlier-extreme** para cada atributo é gerado.



8. Você pode clicar em cada atributo gerado para verificar se existem valores *outlier* ou *extreme value* para atributos originais. Você pode remover aqueles atributos indicadores que não tenham nenhum *outlier* ou *extreme value* com o botão **Remove**. Salve os dados resultantes como **small\_telco.processed.arff**.

The screenshot shows the Weka Explorer interface. The 'Filter' section has 'NumericCleaner' selected with the following parameters: `-min -1.7976931348623157E308 -min-default -1.7976931348623157E308 -max 1.7976931348623157E308 -max`. The 'Current relation' is 'small\_telco\_labOne-weka.filters.unsupervised.attribute.Rep...' with 1000 instances and 62 attributes. The 'Attributes' list on the left shows several attributes selected with checkboxes, including 'internet\_Outlier' (row 51), 'internet\_ExtremeValue' (row 52), 'ebill\_Outlier' (row 53), 'ebill\_ExtremeValue' (row 54), 'loglong\_Outlier' (row 55), 'loglong\_ExtremeValue' (row 56), 'logequi\_Outlier' (row 57), and 'logequi\_ExtremeValue' (row 58). A red circle highlights the 'Remove' button at the bottom of the attributes list. The 'Selected attribute' section shows 'internet\_Outlier' with a 'Type: Nominal', 'Missing: 0 (0%)', and 'Distinct: 1'. A table below shows the distribution: 'no' (1000) and 'yes' (0). A bar chart below the table shows a red bar for 'no' (height 1000) and a blue bar for 'yes' (height 0). The status bar at the bottom shows 'Status OK' and a 'Log' button.

| No. | Label | Count |
|-----|-------|-------|
| 1   | no    | 1000  |
| 2   | yes   | 0     |

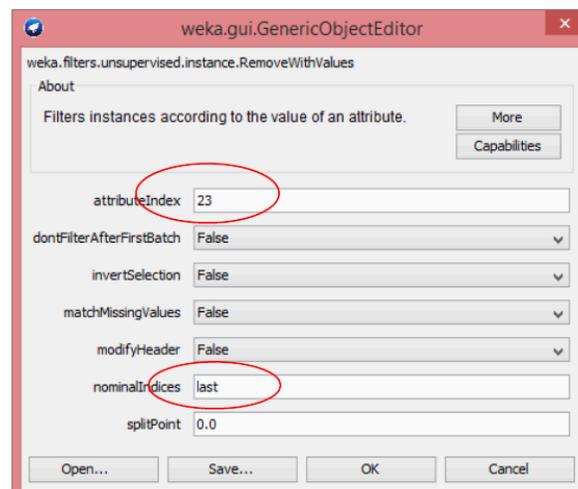
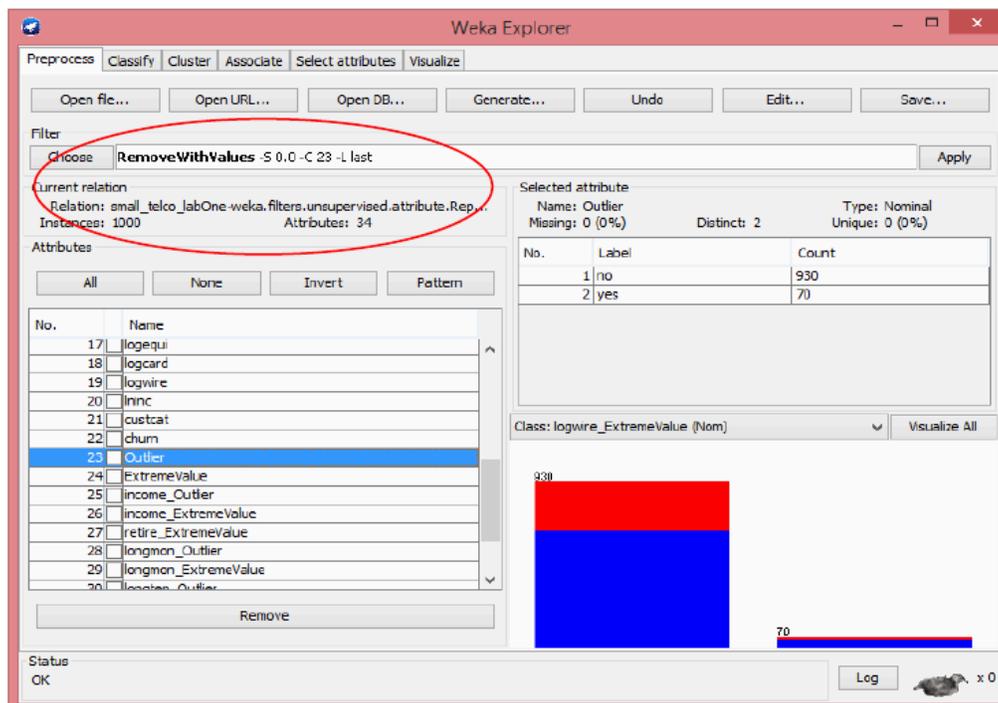
### 3.1. Removendo instâncias com *outliers* e *extreme values*

**Unsupervised Instance Filter – RemoveWithValues:** Este filtro remove instâncias de acordo com valores de um atributo.

1. Após encontrarmos instâncias com valores *outlier* ou *extreme value*, nós podemos removê-las completamente do dataset. Escolha o filtro **RemoveWithValues** na lista drop-down de filtros não-supervisionados de instâncias.

Para isso clique em: **filters** => **unsupervised** => **instance** => **InterquartileRange**.

Como o atributo *outlier* está indexado com 23 e o último valor é “yes”, modifique as opções *attributeIndex* para **23** e o *nominalIndices* para “**last**”



- Então clique no botão **Apply** depois de confirmar as mudanças. 70 instâncias serão removidas do dataset e o atributo *outlier* não terá instâncias com valor “yes”.

The screenshot shows the Weka Explorer interface with the 'RemoveWithValues' filter applied to the 'Outlier' attribute. The 'Current relation' section shows 930 instances. The 'Selected attribute' table shows 930 instances for 'no' and 0 for 'yes'. A bar chart visualizes this distribution.

| No. | Label | Count |
|-----|-------|-------|
| 1   | no    | 930   |
| 2   | yes   | 0     |

Class: logwire\_ExtremeValue (Nom) Visualize All

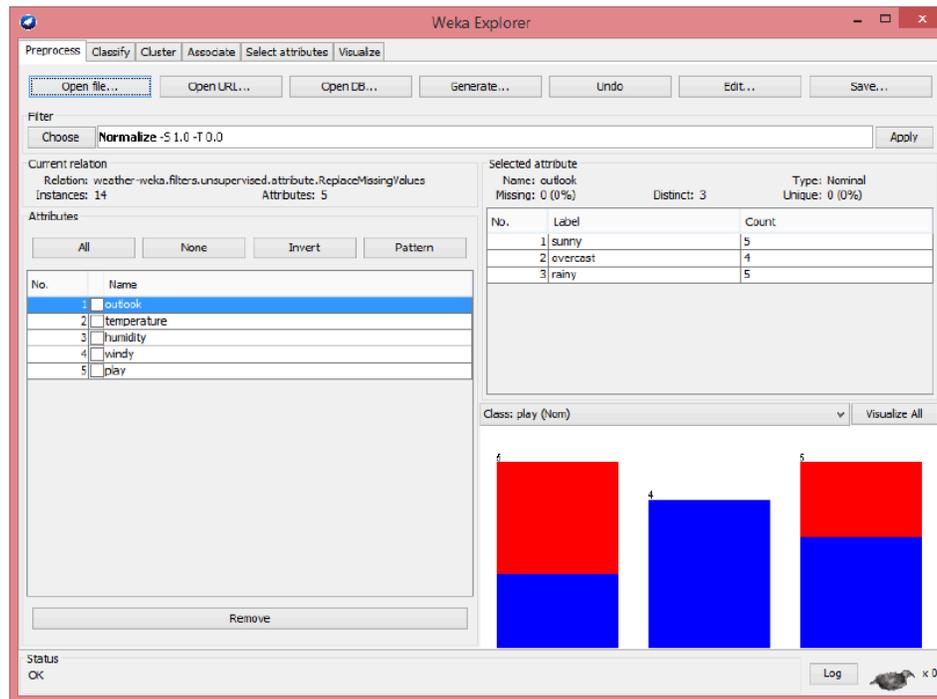
Status OK Log x 0

- Você poderá também remover instâncias por atributo de acordo com pares *outlier-atributo* da mesma forma.

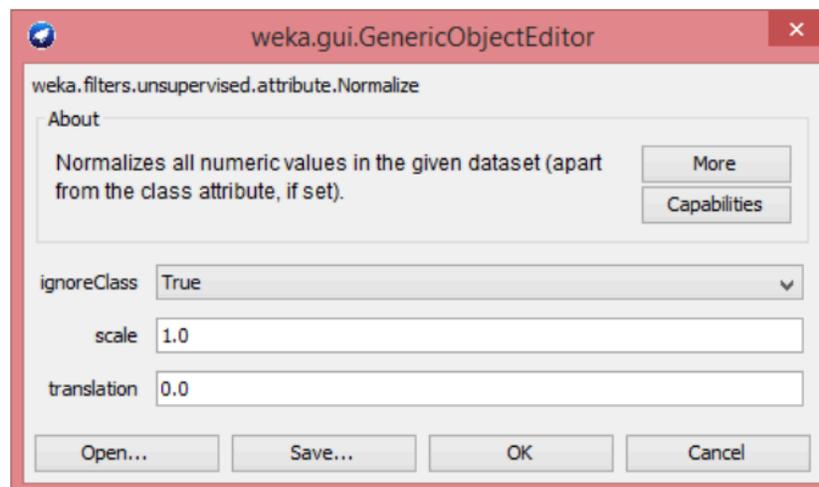
## 4. Usando filtros para executar a normalização

**Unsupervised Attribute Filter – Normalize:** Este filtro normaliza todos os valores numéricos no dataset fornecido para o intervalo padrão de  $[0,0, 1,0]$ .

1. Abra o dataset **weather.numeric.nomissing.arff** (os valores ausentes já foram substituídos).



2. Escolha filtro Normalize na lista drop-down de filtros de atributos não supervisionados. Para isso clique em: **filters** => **unsupervised** => **attribute** => **Normalize**. E, em seguida, clique com o botão **Esquerdo** para abrir sua janela de propriedades. Queremos fazer normalização em todos os atributos numéricos. Clique em **OK** e **Apply**.



Viewer

Relation: weather-weka.filters.unsupervised.attribute.Replace...

| No. | outlook<br>Nominal | temperature<br>Numeric | humidity<br>Numeric | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | 1.0                    | 0.6451...           | FALSE            | no              |
| 2   | sunny              | 0.7619047...           | 0.8064...           | TRUE             | no              |
| 3   | overcast           | 0.9047619...           | 0.6774...           | FALSE            | yes             |
| 4   | rainy              | 0.4688644...           | 1.0                 | FALSE            | yes             |
| 5   | rainy              | 0.1904761...           | 0.4838...           | FALSE            | yes             |
| 6   | rainy              | 0.0476190...           | 0.1612...           | TRUE             | no              |
| 7   | overcast           | 0.0                    | 0.0                 | TRUE             | yes             |
| 8   | sunny              | 0.3809523...           | 0.9677...           | FALSE            | no              |
| 9   | sunny              | 0.2380952...           | 0.5657...           | FALSE            | yes             |
| 10  | rainy              | 0.5238095...           | 0.4838...           | FALSE            | yes             |
| 11  | sunny              | 0.5238095...           | 0.1612...           | TRUE             | yes             |
| 12  | overcast           | 0.3809523...           | 0.8064...           | TRUE             | yes             |
| 13  | overcast           | 0.8095238...           | 0.3225...           | FALSE            | yes             |
| 14  | rainy              | 0.3333333...           | 0.8387...           | TRUE             | no              |

Undo OK Cancel

3. Você pode escolher um intervalo diferente definindo fatores de **scale** e **translation**. A **scale** é a diferença entre os valores mínimo e máximo. Se a **scale** ficar em 2 e a **translation** for mantida em 0, o intervalo será [0.0, 2.0].

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Normalize

About

Normalizes all numeric values in the given dataset (apart from the class attribute, if set).

More

Capabilities

ignoreClass False

scale 2.0

translation 0.0

Open... Save... OK Cancel

Viewer

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMis...

| No. | outlook<br>Nominal | temperature<br>Numeric | humidity<br>Numeric | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | 2.0                    | 1.2903...           | FALSE            | no              |
| 2   | sunny              | 1.5238095...           | 1.6129...           | TRUE             | no              |
| 3   | overcast           | 1.8095238...           | 1.3548...           | FALSE            | yes             |
| 4   | rainy              | 0.9377289...           | 2.0                 | FALSE            | yes             |
| 5   | rainy              | 0.3809523...           | 0.9677...           | FALSE            | yes             |
| 6   | rainy              | 0.0952380...           | 0.3225...           | TRUE             | no              |
| 7   | overcast           | 0.0                    | 0.0                 | TRUE             | yes             |
| 8   | sunny              | 0.7619047...           | 1.9354...           | FALSE            | no              |
| 9   | sunny              | 0.4761904...           | 1.1315...           | FALSE            | yes             |
| 10  | rainy              | 1.0476190...           | 0.9677...           | FALSE            | yes             |
| 11  | sunny              | 1.0476190...           | 0.3225...           | TRUE             | yes             |
| 12  | overcast           | 0.7619047...           | 1.6129...           | TRUE             | yes             |
| 13  | overcast           | 1.6190476...           | 0.6451...           | FALSE            | yes             |
| 14  | rainy              | 0.6666666...           | 1.6774...           | TRUE             | no              |

Undo OK Cancel

4. A **translation** é a distância entre o mínimo e 0.0. Quando a **scale** é deixada em 2 e a **translation** mantida em -1, o intervalo é [-1.0, 1.0].

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Normalize

About

Normalizes all numeric values in the given dataset (apart from the class attribute, if set).

More

Capabilities

ignoreClass False

scale 2.0

translation -1

Open... Save... OK Cancel

Viewer

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingVal...

| No. | outlook<br>Nominal | temperature<br>Numeric | humidity<br>Numeric | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | 1.0                    | 0.2903...           | FALSE            | no              |
| 2   | sunny              | 0.5238095...           | 0.6129...           | TRUE             | no              |
| 3   | overcast           | 0.8095238...           | 0.3548...           | FALSE            | yes             |
| 4   | rainy              | -0.0622710...          | 1.0                 | FALSE            | yes             |
| 5   | rainy              | -0.6190476...          | -0.032...           | FALSE            | yes             |
| 6   | rainy              | -0.9047619...          | -0.677...           | TRUE             | no              |
| 7   | overcast           | -1.0                   | -1.0                | TRUE             | yes             |
| 8   | sunny              | -0.2380952...          | 0.9354...           | FALSE            | no              |
| 9   | sunny              | -0.5238095...          | 0.1315...           | FALSE            | yes             |
| 10  | rainy              | 0.0476190...           | -0.032...           | FALSE            | yes             |
| 11  | sunny              | 0.0476190...           | -0.677...           | TRUE             | yes             |
| 12  | overcast           | -0.2380952...          | 0.6129...           | TRUE             | yes             |
| 13  | overcast           | 0.6190476...           | -0.354...           | FALSE            | yes             |
| 14  | rainy              | -0.3333333...          | 0.6774...           | TRUE             | no              |

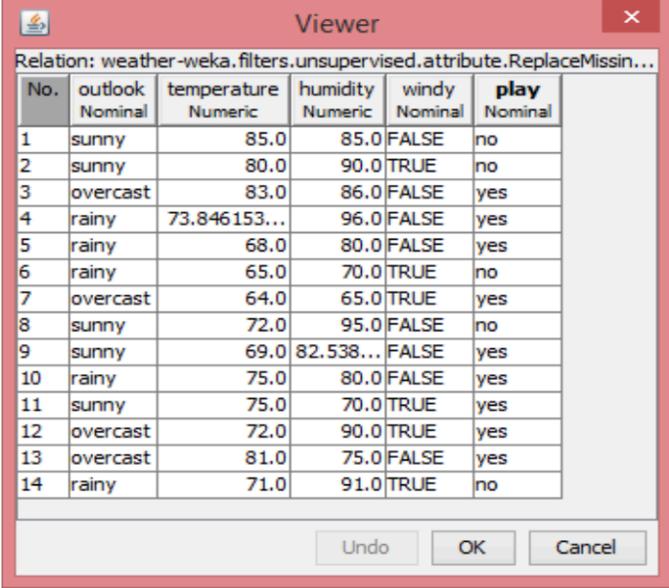
Undo OK Cancel

5. Você pode salvar o dataset se estiver satisfeito com os resultados  
**(weather.numeric.processed.arff)**

## 5. Discretização com Filtros

**Unsupervised Attribute Filter – Discretize:** Este filtro converte atributos numéricos em atributos nominais usando discretização *equal-width* (largura - default) or *equal-depth* (frequência)

1. Abra o dataset **weather.numeric.nomissing.arff**. (sem *missing values*)



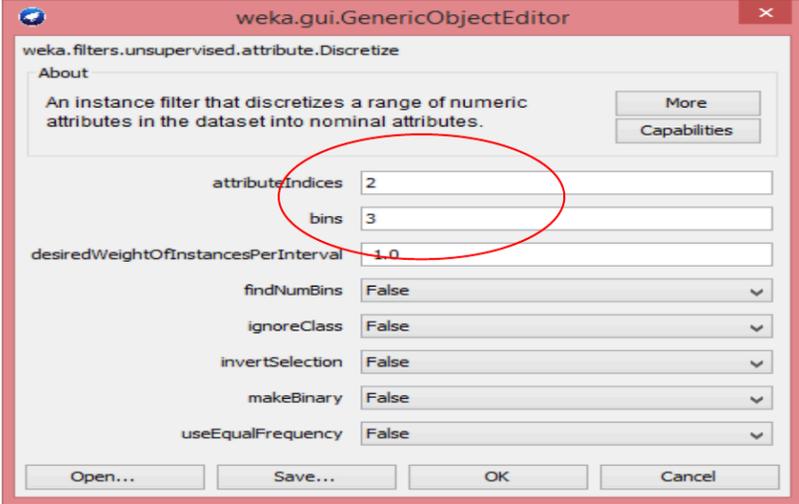
Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissin...

| No. | outlook  | temperature  | humidity  | windy   | play    |
|-----|----------|--------------|-----------|---------|---------|
|     | Nominal  | Numeric      | Numeric   | Nominal | Nominal |
| 1   | sunny    | 85.0         | 85.0      | FALSE   | no      |
| 2   | sunny    | 80.0         | 90.0      | TRUE    | no      |
| 3   | overcast | 83.0         | 86.0      | FALSE   | yes     |
| 4   | rainy    | 73.846153... | 96.0      | FALSE   | yes     |
| 5   | rainy    | 68.0         | 80.0      | FALSE   | yes     |
| 6   | rainy    | 65.0         | 70.0      | TRUE    | no      |
| 7   | overcast | 64.0         | 65.0      | TRUE    | yes     |
| 8   | sunny    | 72.0         | 95.0      | FALSE   | no      |
| 9   | sunny    | 69.0         | 82.538... | FALSE   | yes     |
| 10  | rainy    | 75.0         | 80.0      | FALSE   | yes     |
| 11  | sunny    | 75.0         | 70.0      | TRUE    | yes     |
| 12  | overcast | 72.0         | 90.0      | TRUE    | yes     |
| 13  | overcast | 81.0         | 75.0      | FALSE   | yes     |
| 14  | rainy    | 71.0         | 91.0      | TRUE    | no      |

2. Escolha o filtro **Discretize** a partir da lista drop-down de filtros de atributos não supervisionados.

Para isso clique em: **filters** => **unsupervised** => **attribute** => **Discretize**.

E então clique (com o botão **Esquerdo**) para abrir a janela de propriedades. Vamos realizar uma discretização *equal-width* (largura) no atributo 2 – **temperature** com 3 *bins*.



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

More

Capabilities

attributeIndices: 2

bins: 3

desiredWeightOfInstancesPerInterval: 1.0

findNumBins: False

ignoreClass: False

invertSelection: False

makeBinary: False

useEqualFrequency: False

Open... Save... OK Cancel

3. Clique no botão **Apply**. E então selecione o atributo **temperature** para verificar os resultados.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **Discretize -B 3 -M -1.0 -R 2** Apply

Current relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValues...  
Instances: 14 Attributes: 5

Attributes: All None Invert Pattern

| No. | Name        |
|-----|-------------|
| 1   | outlook     |
| 2   | temperature |
| 3   | humidity    |
| 4   | windy       |
| 5   | play        |

Selected attribute: Name: temperature, Missing: 0 (0%), Distinct: 3, Type: Nominal, Unique: 0 (0%)

| No. | Label       | Count |
|-----|-------------|-------|
| 1   | '(-inf-71]' | 5     |
| 2   | '(71-78]'   | 5     |
| 3   | '(78-inf]'  | 4     |

Class: play (Nom) Visualize All

Bar chart showing distribution of 'play' class across temperature bins:

- Bin 1: 5 instances (2 'no', 3 'yes')
- Bin 2: 5 instances (3 'no', 2 'yes')
- Bin 3: 4 instances (2 'no', 2 'yes')

Status: OK Log x 0

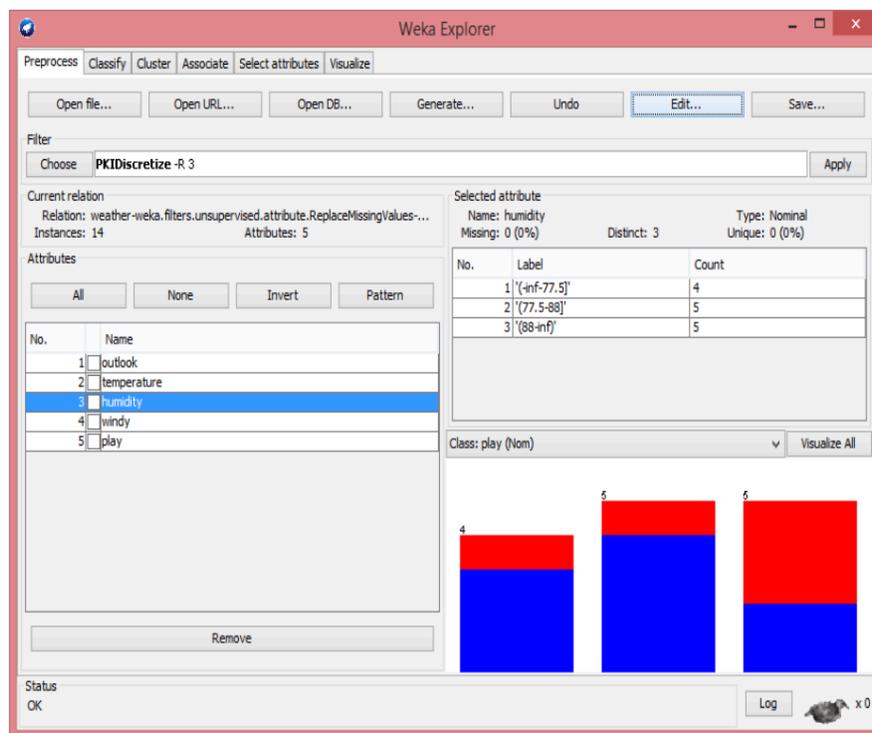
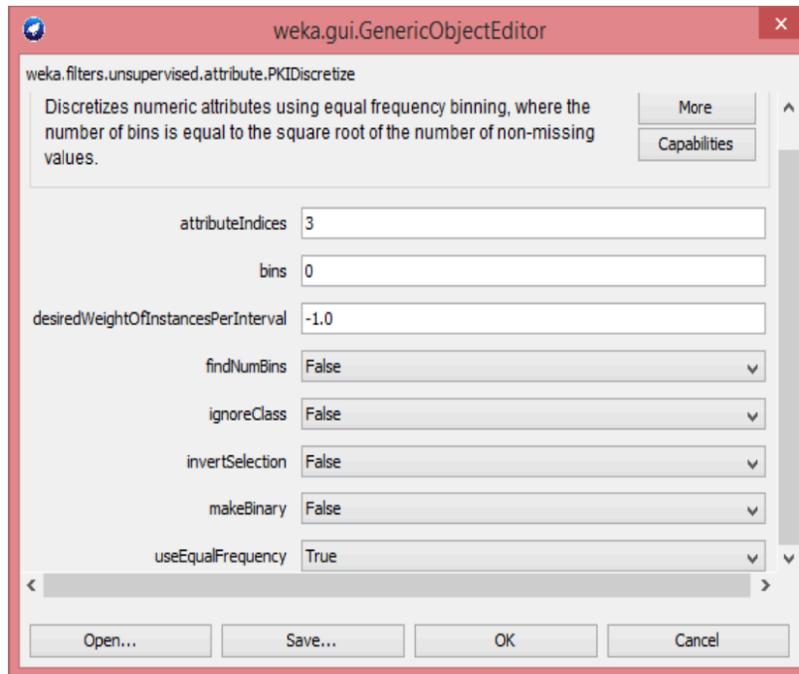
Viewer

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMis...

| No. | outlook<br>Nominal | temperature<br>Nominal | humidity<br>Numeric | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | '(78-inf]'             | 85.0                | FALSE            | no              |
| 2   | sunny              | '(78-inf]'             | 90.0                | TRUE             | no              |
| 3   | overcast           | '(78-inf]'             | 86.0                | FALSE            | yes             |
| 4   | rainy              | '(71-78]'              | 96.0                | FALSE            | yes             |
| 5   | rainy              | '(-inf-71]'            | 80.0                | FALSE            | yes             |
| 6   | rainy              | '(-inf-71]'            | 70.0                | TRUE             | no              |
| 7   | overcast           | '(-inf-71]'            | 65.0                | TRUE             | yes             |
| 8   | sunny              | '(71-78]'              | 95.0                | FALSE            | no              |
| 9   | sunny              | '(-inf-71]'            | 82.538...           | FALSE            | yes             |
| 10  | rainy              | '(71-78]'              | 80.0                | FALSE            | yes             |
| 11  | sunny              | '(71-78]'              | 70.0                | TRUE             | yes             |
| 12  | overcast           | '(71-78]'              | 90.0                | TRUE             | yes             |
| 13  | overcast           | '(78-inf]'             | 75.0                | FALSE            | yes             |
| 14  | rainy              | '(-inf-71]'            | 91.0                | TRUE             | no              |

Undo OK Cancel

4. Para realizar a discretização *equal-depth* (frequência) no atributo 3 - *humidity*, escolhemos o filtro *PKIDiscretize* a partir da lista drop-down de filtros de atributos não supervisionados. Para isso clique em: **filters** => **unsupervised** => **attribute** => **PKIDiscretize**. Este filtro usa a raiz quadrada do número de valores como o número de *bins*.



Viewer

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissing...

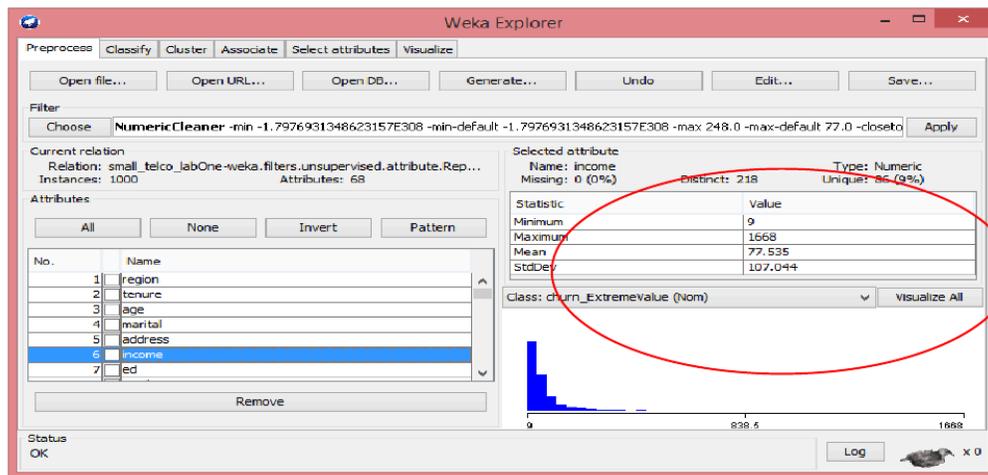
| No. | outlook<br>Nominal | temperature<br>Nominal | humidity<br>Nominal | windy<br>Nominal | play<br>Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1   | sunny              | '(78-inf)'             | '(77.5-...          | FALSE            | no              |
| 2   | sunny              | '(78-inf)'             | '(88-inf)'          | TRUE             | no              |
| 3   | overcast           | '(78-inf)'             | '(77.5-...          | FALSE            | yes             |
| 4   | rainy              | '(71-78]'              | '(88-inf)'          | FALSE            | yes             |
| 5   | rainy              | '(-inf-71]'            | '(77.5-...          | FALSE            | yes             |
| 6   | rainy              | '(-inf-71]'            | '(-inf-7...         | TRUE             | no              |
| 7   | overcast           | '(-inf-71]'            | '(-inf-7...         | TRUE             | yes             |
| 8   | sunny              | '(71-78]'              | '(88-inf)'          | FALSE            | no              |
| 9   | sunny              | '(-inf-71]'            | '(77.5-...          | FALSE            | yes             |
| 10  | rainy              | '(71-78]'              | '(77.5-...          | FALSE            | yes             |
| 11  | sunny              | '(71-78]'              | '(-inf-7...         | TRUE             | yes             |
| 12  | overcast           | '(71-78]'              | '(88-inf)'          | TRUE             | yes             |
| 13  | overcast           | '(78-inf)'             | '(-inf-7...         | FALSE            | yes             |
| 14  | rainy              | '(-inf-71]'            | '(88-inf)'          | TRUE             | no              |

Undo    OK    Cancel

## 6. Usando Filtros para Substituir valores

**Unsupervised Attribute Filter – NumericCleaner:** Este filtro substitui os valores dos atributos numéricos que são muito pequenos ou muito grandes ou muito próximos a um valor particular por valores *default*.

1. Ao invés de remover instâncias com *outlier* e *extreme value*, nós poderíamos substituir os valores dos atributos para valores *default*. Abra o dataset **small\_telco.processed.arff**. Vamos usar o atributo **income** como um exemplo. Clique no atributo **income**, e então suas estatísticas são mostradas na parte direita da janela: o mínimo é 9, o máximo é 1668 e a média é 77.535.

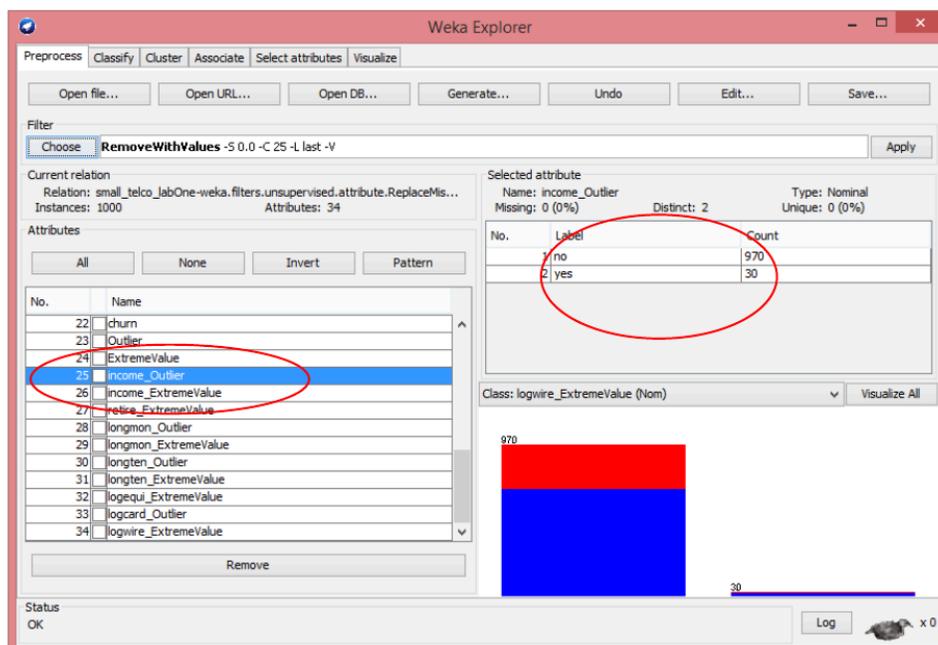


The screenshot shows the Weka Explorer interface with the **NumericCleaner** filter selected. The **Selected attribute** table is highlighted with a red circle:

| Statistic | Value   |
|-----------|---------|
| Minimum   | 9       |
| Maximum   | 1668    |
| Mean      | 77.535  |
| StdDev    | 107.044 |

The **Attributes** list on the left shows 'income' selected. The **Class** dropdown is set to 'churn\_ExtremeValue (Nom)'. A histogram of the 'income' attribute is shown at the bottom right.

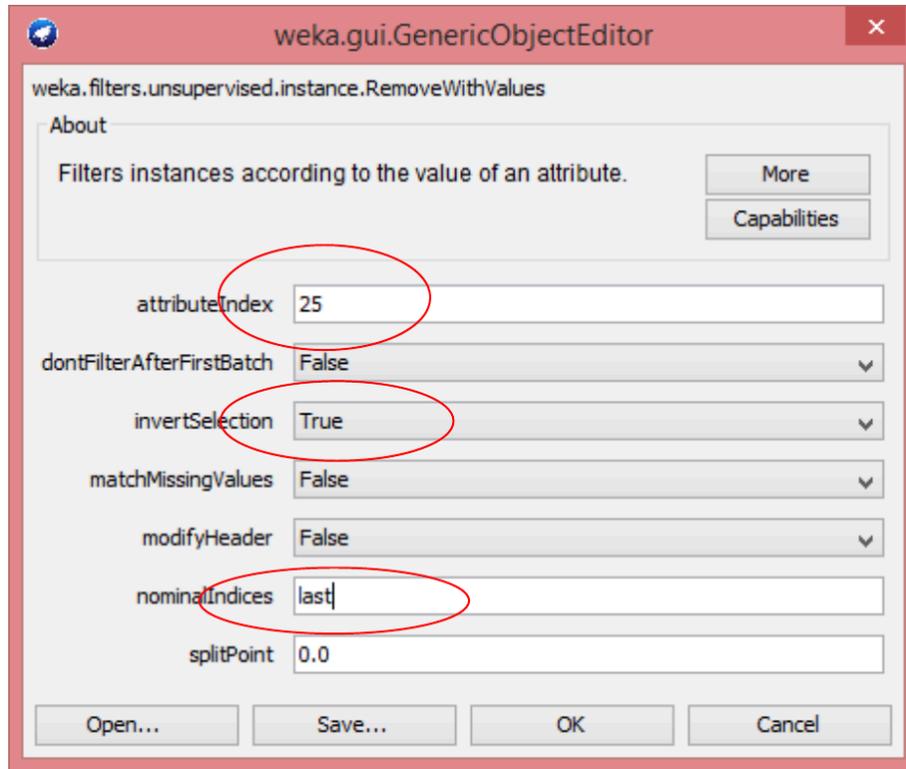
2. Realize uma filtragem **RemoveWithValues** invertida com o atributo **income\_Outlier**.



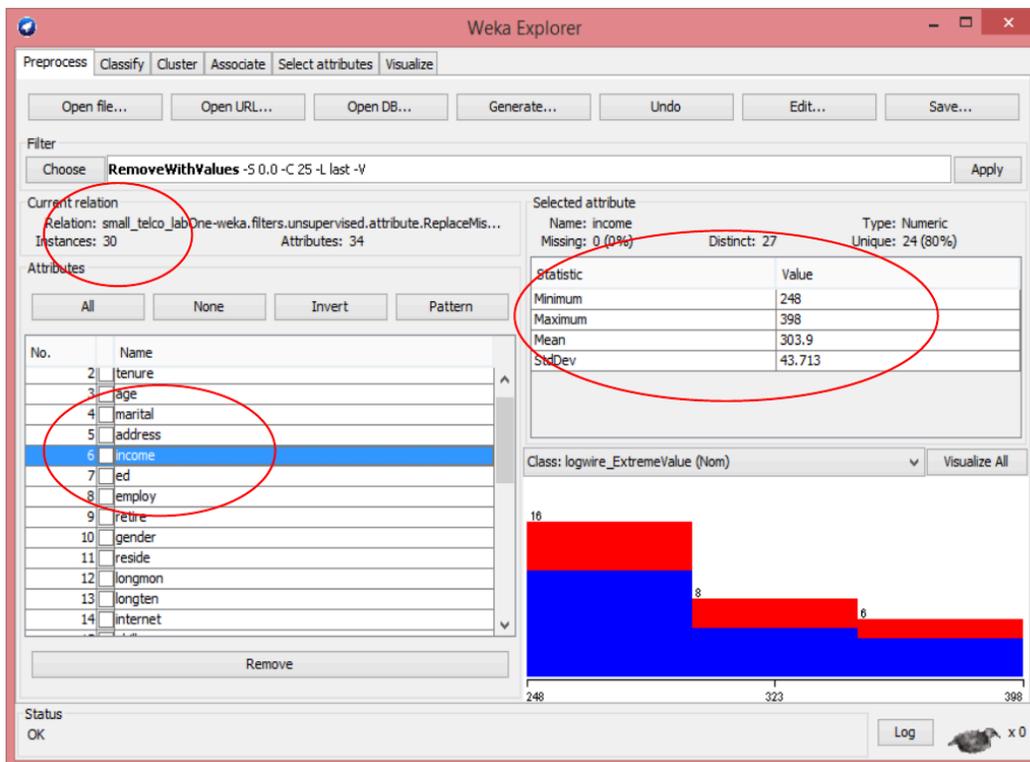
The screenshot shows the Weka Explorer interface with the **RemoveWithValues** filter selected. The **Selected attribute** table is highlighted with a red circle:

| No. | Label | Count |
|-----|-------|-------|
| 1   | no    | 970   |
| 2   | yes   | 30    |

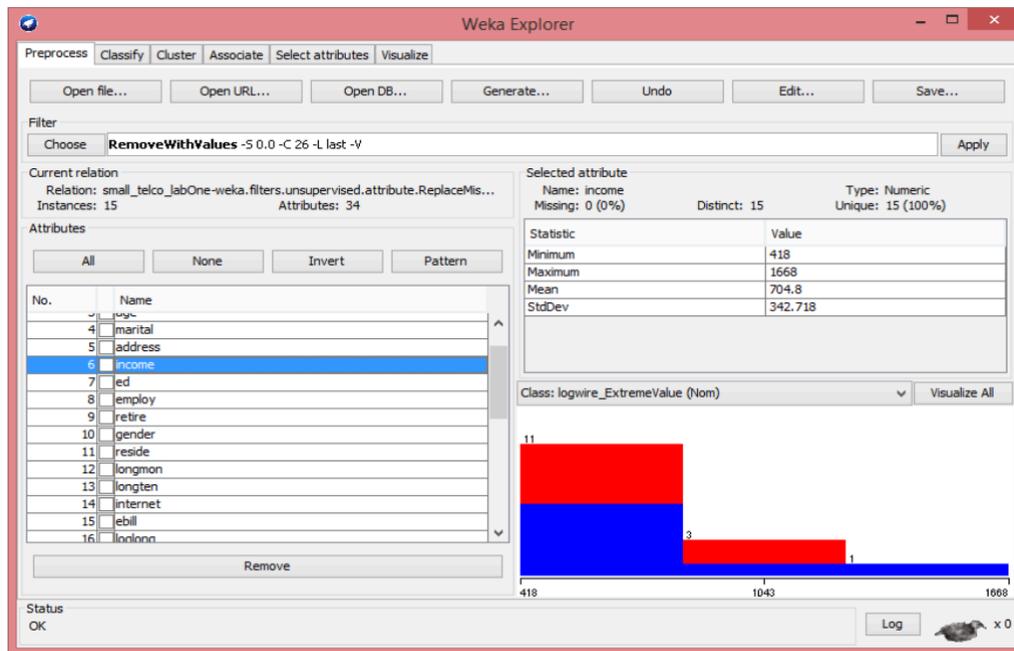
The **Attributes** list on the left shows 'income\_Outlier' selected. The **Class** dropdown is set to 'logwire\_ExtremeValue (Nom)'. A bar chart of the 'income\_Outlier' attribute is shown at the bottom right.



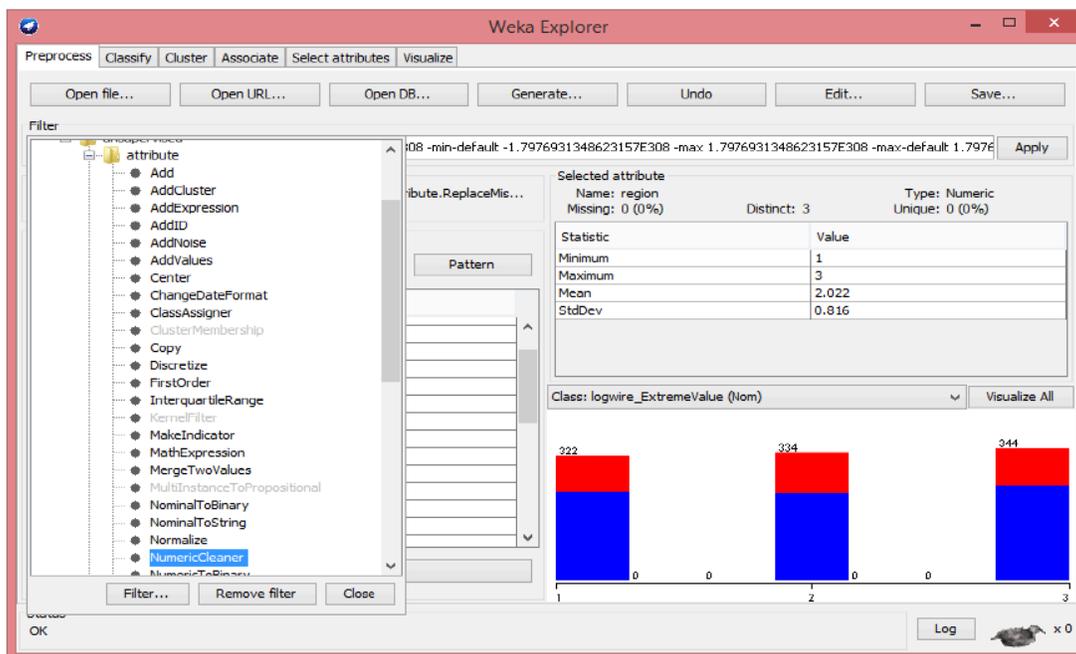
3. Verifique o status do atributo **income** nas 30 instâncias remanescentes. O min. Agora é 248, e o max. é 398.

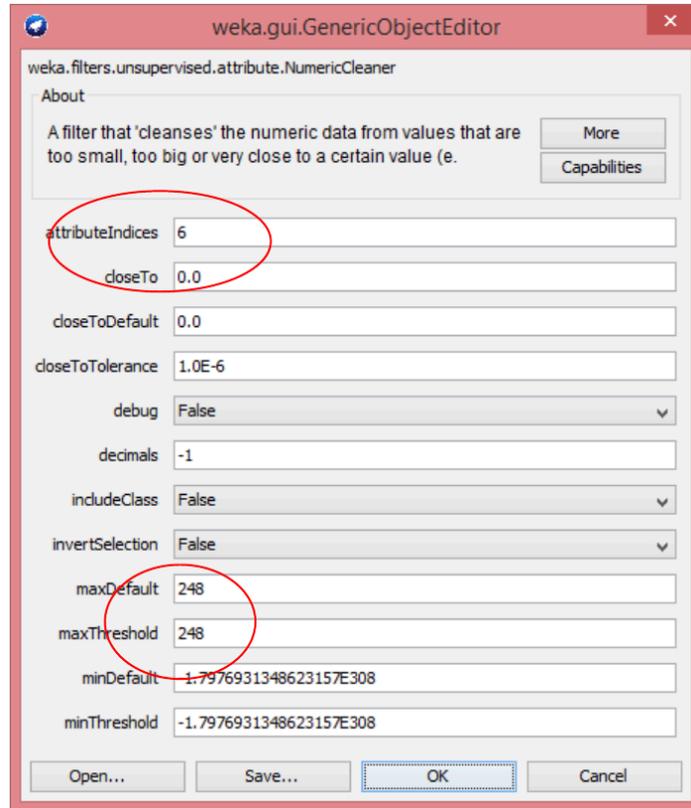


- Clique no botão **Undo** e repita o passo e repita os passo 2 com o atributo **income\_Extremevalue**. Verifique o atributo **income** nas 15 instâncias remanescentes. O min. Ficou em 418, e o max. é 1668.



- Clique no botão **Undo**.
- Então estamos prontos para realizar o filtro não supervisionado no atributo – **NumericCleaner** em todas as instâncias. Escolha o filtro NumericCleaner na lista drop-down, e então clique com o botão **Esquerdo** na caixa do filtro para mostrar a janela de propriedades.





7. Clique no botão **Apply** para realizar a filtragem, então selecionar o atributo **income** para visualizar as estatísticas do atributo modificado. Se você gostou do resultado, salve o dataset.

