3 .arff FILE

Weka works with a format called ARFF, which stands for *Attribute-Relation File Format*.
This format is composed of a structured three-part layout:

- Header.
- Attribute declaration.
- Data section.

HEADER

It's a label that helps organize and describe data sets in Weka, providing a clear and descriptive name to the relation or dataset being represented in the file.
It is represented like:

```
@relation <relation_name>
```

We can also add filters. Here is an example of the file we will download below.

```
@relation 'people-weka.filters.unsupervised.attribute.StringToNominal
-R1-2,7-weka.filters.unsupervised.attribute.StringToNominal-R1-3,5,7-8'
```

For now, we will only mention the fact that filters can be added; further explanation will be provided later.

ATTRIBUTE DECLARATION

The attribute declaration format depends on the data type, but in general, it will be written as follows:

```
@attribute <attribute_name> <data_type>
```

There are different data types that these files can accept:

- STRING
  They include words, character sequences, or phrases. It can contain any combination of letters, numbers, and other characters. In the case of containing spaces, the phrase will be placed between single quotes (' ') or spaces can be replaced with underscores (_).

```
@attribute 'character string' STRING
@attribute character_string STRING
```

- NOMINAL
  They represent categorical data with a finite and predefined set of values. The values are labels representing categories, and there is no inherent order among them. The values will be written within curly braces { } and inside them, they should be separated by commas. In the case that the data consists of two or more words, it should be enclosed in single quotes (' ').

  ```
  @attribute countries {Spain, Italy, France, 'New Zeland', Turkey}
  @attribute operating_systems {Windows, Linux, 'Mac OS', Unix, Android}
  ```

Nominal attributes are used for categorical variables with predefined values, while string attributes are more flexible and can contain any character sequence, including free-form text. The choice between them depends on the nature of the data you are trying to represent in your dataset.

- NUMERIC
  It represents any type of numeric value (both decimal and integer, interchangeably). Weka will always represent numeric values as "numeric."

  ```
  @attribute number NUMERIC
  ```

- REAL
  They represent real numbers, which means numbers with decimals.

  ```
  @attribute real_number REAL
  ```

- INTEGER
  They represent integers, which means numbers without decimals.

  ```
  @attribute integer_number INTEGER
  ```

Numeric data, whether "real" or "integer," will be represented in Weka as "numeric." This information is used to inform the user about the numeric data type, and therefore, it is optional to include it.

- DATE
  It represents different units of time. We need to indicate the name, the data type (DATE), and the desired format. Combinations are also possible:

  ```
  @attribute full_date DATE 'dd-MM-yyyy HH:mm:ss'
  @attribute hour DATE HH:mm
  @attribute year DATE yyyy
  ```

DATA SECTION (DATASETS)

This section begins with the following declaration:

```
@data
```

Following this declaration are the actual data. Each line represents an instance. Within each instance, attributes are separated by commas. These attributes should appear in the order declared in the header section.

If any value is not defined, it is represented by a "?".

DATA TYPE		
STRING	Character sequences, such as words, phrases, or text strings.	
NOMINAL	They have a finite and predefined set of discrete values without a specific order, and are written within curly braces { }.	
NUMERIC	REAL	Decimal numbers.
	INTEGER	Integer numbers (without decimals).
DATE	Units of time.	

EXAMPLE

This is a shortened example of the .arff file that we will download below.

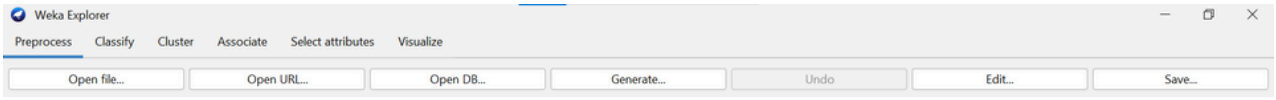
```
@relation 'people-weka.filters.unsupervised.attribute.StringToNominal-R1-2,7-
weka.filters.unsupervised.attribute.StringToNominal-R1-3,5,7-8'

@attribute Name {ESTEFANIA,QUERALT,JOAN,MARC,JOSEP,ESTHER}
@attribute Lastname {'AROCAS PASADAS','VISO GILABERT','AYALA FERRERAS','BAEZ TEJADO','BASTARDES
SOTO','ANGUERA VILAFRANCA','PASCUAL ALOY'}
@attribute City {Zaragoza,Barcelona,Tarragona,Valencia,Girona}
@attribute Sex {Female,Male}
@attribute State {Others,Separated,Single,Married,Divorced}
@attribute Sons numeric
@attribute Profesion {Administrative,Draftsman,Accountant,Dependent,Student}
@attribute Studies {Bachelor,Secundary,Diplomat,Elementant,VT}
@attribute Salary numeric
@attribute Preferred_customer{YES,NO}

@data
ESTEFANIA,'AROCAS PASADAS',   Zaragoza,   Female,  Others,    1,  Administrative, Bachelor,   1500,   NO
QUERALT,  'VISO GILABERT',    Barcelona,  Female,  Others,    0,  Draftsman,      Secundary,  1100,  YES
JOAN,     'AYALA FERRERAS',   Zaragoza,   Male,    Others,    0,  ?,              Diplomat,   1500,   NO
JOAN,     'BAEZ TEJADO',      Zaragoza,   Male,    Separated, 1,  Accountant,     Bachelor,   1900,   NO
MARC,     'BASTARDES SOTO',   Tarragona,  Male,    Single,    3,  Dependent,      Elementant, 1100,   NO
JOSEP,    'ANGUERA VILAFRANCA',Valencia,  Male,    Married,   1,  Student,        VT,         500,    NO
ESTHER,   'PASCUAL ALOY',     Girona,     Female,  Divorced,  0,  ?,              Diplomat,   2500,  YES
```

4 Basic resources. Explorer

In this window, we find the options: *preprocess*, *classify*, *cluster*, *associate*, *select attributes*, and *visualize*.

4.1 Preprocess

The first step is to select the file to work with it. Weka provides the following options for importing a file:



- Open file: Select the file directly from our computer's file explorer.

- **Open URL:** This is used to load datasets directly from an online location via a URL. This feature is useful when the data you want to analyze is available on the web, and you can access it through a URL. You will generally be prompted to enter the URL of the dataset you want to load. Once you provide the URL, Weka will automatically download the data and load it into the application, allowing you to start working with it.

- **Open DB:** Connects Weka to databases to access and analyze data directly stored in a relational database. This involves using JDBC (Java Database Connectivity), which is a standard Java interface for connecting to database management systems.

- **Generate:** Used to create additional attributes, modify existing ones by applying functions or expressions, or remove them. It can be helpful for generating new features that might be more informative or useful for analysis.

The functions "*Undo*," "*Edit*," and "*Save*" are basic operations that allow you to perform specific actions during the data analysis process.

- **Undo**: Allows you to revert the last action you performed. It is useful when you make a mistake or perform an operation that you want to undo.

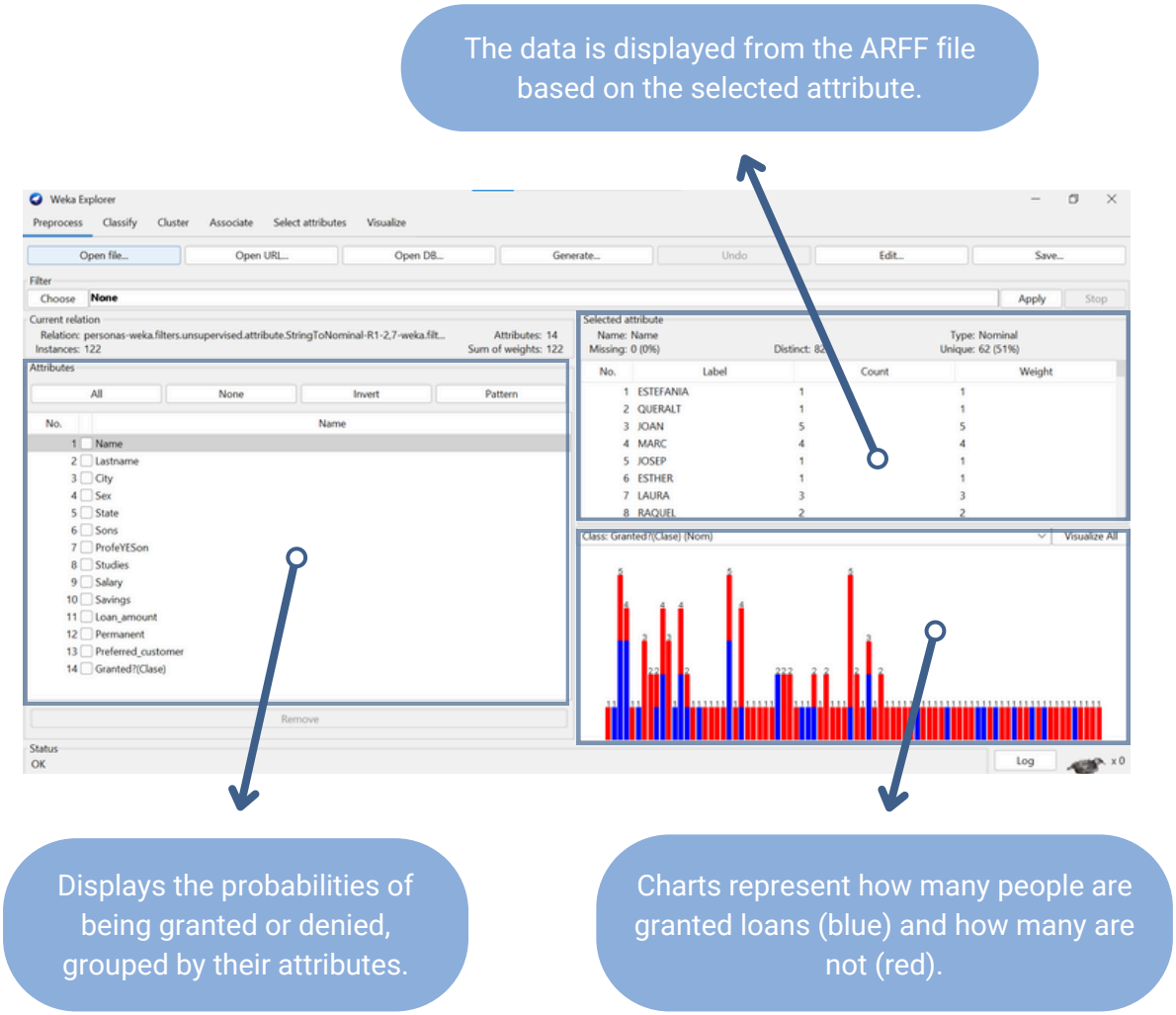
- **Edit**: Enables you to modify specific configurations or properties of a selected element, such as an applied filter or a loaded dataset.

- **Save**: Permits you to save changes made to configurations, datasets, models, etc.

In this case, we will download and import the following file:
people.arff

After downloading, we should select the "*Open file*" option. It will open a window displaying our files. Usually, the downloaded file is located in "Downloads" as "people.arff".
Once the .arff file is imported, we can visualize and analyze the data.

These are the three main windows:



On the attributes screen (bottom left corner), we can select various attribute options and combine them by checking the empty square. The data for these attributes will be displayed on the "Selected attribute" screen (top right corner) individually. In other words, even if we have selected multiple attributes, only the one we click will be displayed. These data can be visually represented just below to facilitate understanding.

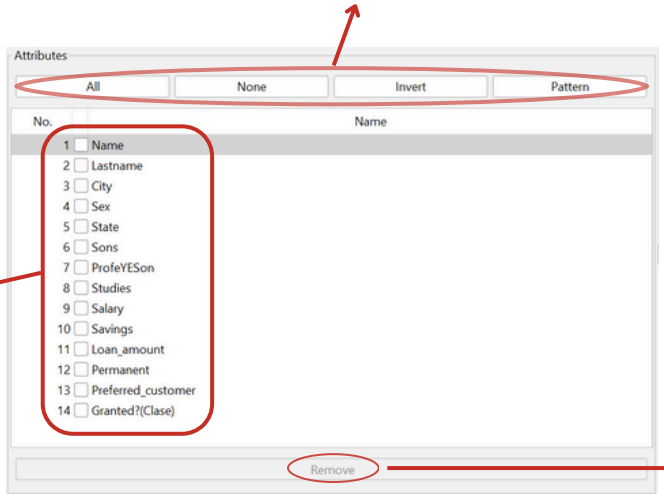
Attribute display, selection and filtering window

It provides detailed information about the attributes present in the loaded dataset.

This window provides us options to select or deselect data, invert the selection, remove attributes or add patterns. It organizes content grouped by attributes.

It also organize the attributes and assign them a number (No.).
On the right, it display the name of each attribute (Name).
Between them is the selection system (empty squares), with this system, we can select multiple attributes simultaneously to operate on them collectively.

Options to select or filter instances in datasets.



Shows attributes and provides a selection system.

It's used to eliminate specific attributes from a dataset.

The *selection system* is useful when we're working with large and complex datasets, as it allows you to perform operations on multiple attributes simultaneously.

- **All**: selects all attributes in the dataset. You can use it to include all attributes in preprocessing or analysis operations.

- **None**: deselects all attributes in the dataset. You can use it when you want to perform specific operations only on a subset of attributes and not on all.

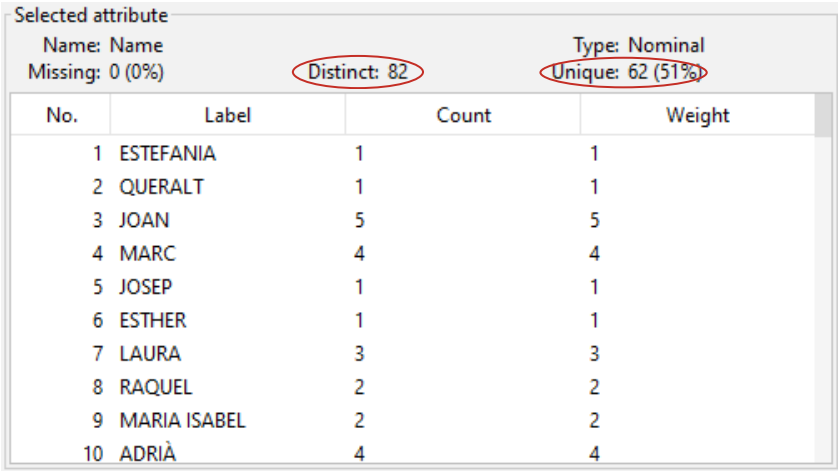
- **Invert**: changes the current attribute selection, selecting those that were deselected and deselecting those that were selected. Useful when you want to invert the current attribute selection without manually selecting them one by one.

- **Pattern**: they are the set of tools and algorithms designed for pattern mining. They are useful when working with transactional or sequential data and aim to discover relationships and patterns among different elements.

Attribute data visualization window

This window will display the data for the selected attribute. It will show the number of records for each data point, the name of each data point, how many distinct values there are, its data type (nominal, string, integer, etc.), the attribute name, and other relevant details that we will explain below.

Example 1: "Name"



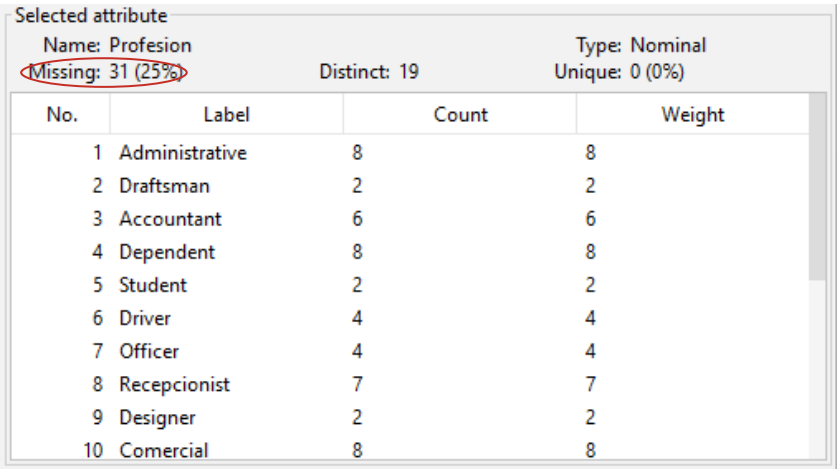
Selected attribute
Name: Name
Missing: 0 (0%)
Distinct: 82
Type: Nominal
Unique: 62 (51%)

No.	Label	Count	Weight
1	ESTEFANIA	1	1
2	QUERALT	1	1
3	JOAN	5	5
4	MARC	4	4
5	JOSEP	1	1
6	ESTHER	1	1
7	LAURA	3	3
8	RAQUEL	2	2
9	MARIA ISABEL	2	2
10	ADRIÀ	4	4

- **Distinct**: indicates how many different values there are in a specific attribute of the dataset. For example, if you are working with a dataset that has a "color" attribute and this attribute has values like "red," "green," and "blue," then the number of distinct values for that attribute will be 3, as there are three different colors.

- **Unique**: number of unique values present in an attribute. It counts values that don't repeat. It is useful for understanding the variability and cardinality of an attribute. If an attribute has a large number of unique values, it may indicate high variability in that data.

Example 2: "Profesion"



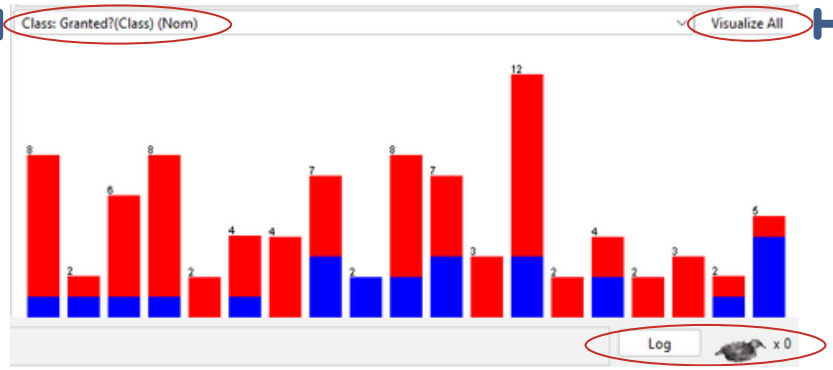
Selected attribute
Name: Profesion
Missing: 31 (25%)
Distinct: 19
Type: Nominal
Unique: 0 (0%)

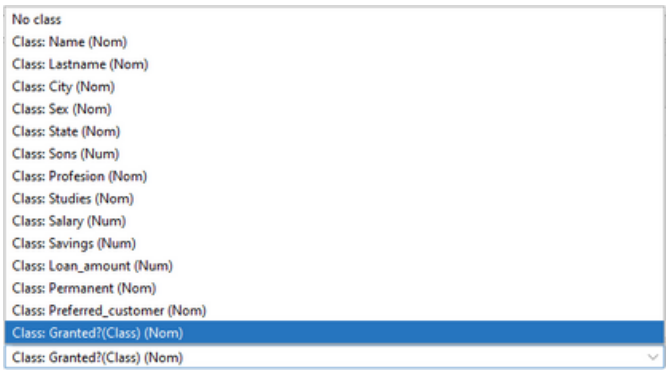
No.	Label	Count	Weight
1	Administrative	8	8
2	Draftsman	2	2
3	Accountant	6	6
4	Dependent	8	8
5	Student	2	2
6	Driver	4	4
7	Officer	4	4
8	Recepcionist	7	7
9	Designer	2	2
10	Comercial	8	8

- **Missing**: (encoded in the ARFF file with "?") indicates how many values are absent or not present for that attribute in the dataset. Identifying the amount of missing values in an attribute can be important for deciding how to address those missing values, whether through imputation, instance deletion, or some other method.

Attribute data visualization window

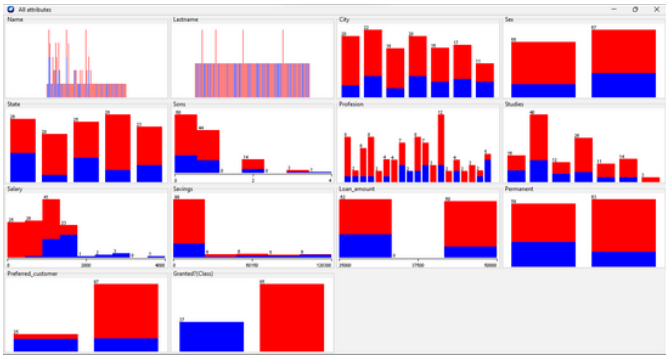
This window allows you to visualize the results graphically.
For example, you can see how many loans are granted and how many are not:





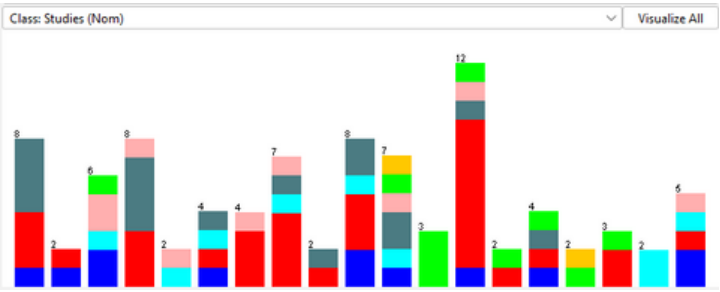
We can also make different combinations based on what we want to visualize.
To do this, we will select the menu at the top and choose a compatible option, that is, of the same data type.

Or, alternatively, we can directly visualize all the possibilities of the chosen option by marking "visualize all".
A separate window will then open with all the charts.

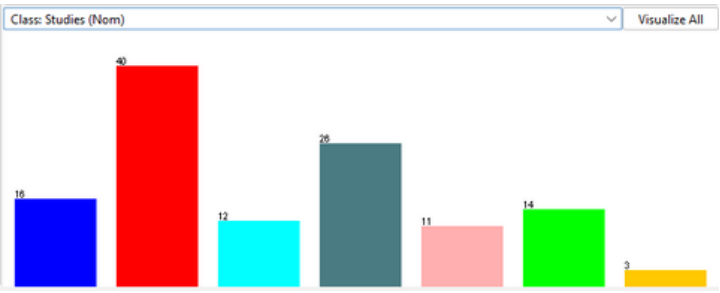


In this case, we have selected the option "Granted?" which tells us whether the loan is granted or not, with red color indicating "not granted" and blue indicating "granted". This way, we can quickly visualize the results without the need to go one by one.

If there are more than two options (as in the case of "Studies," which has multiple options), the different choices will be represented with different colors.



In this case, we will select the attribute "Profession" and in the graphical visualization window, we will choose to classify it by "Studies." This way, we can understand what type of studies individuals have based on their profession.



To know which study corresponds to each color, we should select the "Studies" attribute and classify it in turn by "Studies." This way, by hovering the cursor over the chart, we can determine which study each color represents.

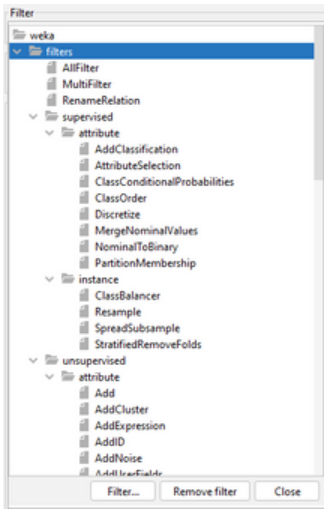
The "Log" section refers to the log of events or messages generated during the execution of operations in the graphical interface or through commands in the command line. This log provides detailed information about what is happening in the system, which can be useful for tracking, debugging, and understanding Weka's internal processes.

When the program is processing, the Weka bird will start moving. This indicates that the algorithm is actively working.

Filter application

The filters section (located at the top) allows the application of data filtering techniques for preprocessing, transformation, or attribute selection tasks in datasets.
Weka has built-in filters that allow manipulations on data at two levels: attributes and instances. Filtering operations can be applied "in cascade," meaning that each filter takes as input the dataset resulting from applying a previous filter.

If you click the "Choose" button, a menu will open, displaying the various filters arranged in a tree structure. They are grouped into:



SupervisedAttributeFilter: these filters help select or transform features (attributes) in your data, taking into account the variable you are trying to predict or classify (the label or class variable).

UnsupervisedAttributeFilter: these filters make changes to your data without considering the variable you are trying to predict. They are applied independently of the label or class variable.

The main difference lies in whether the filter takes into account the variable you are trying to predict. Supervised filters consider that variable, while unsupervised filters do not.

Within these, they are further grouped into two types:

Attribute: applied to the columns (attributes) of your dataset. They modify or select specific features. You can use an attribute filter to select the most relevant features, transform numeric features, discretize attributes, or eliminate attributes that do not provide much information.

Instance: applied to the rows (instances) of your dataset. They modify or select specific instances. You can use an instance filter to select a subset of instances, balance unequal classes, or eliminate instances with certain characteristics.

Once a specific filter is selected using the "Choose" button, its name appears within the filter area (where the word "None" appeared before). You can configure its parameters by clicking on this area, at which point the configuration window for that particular filter appears. If this operation is not performed, the default values of the selected filter would be used.
Once a filter has been applied, the relationship changes for the rest of the operations carried out in the Experimenter, with the option to undo the last applied filtering operation using the Undo button. Additionally, the results of applying filters can be saved in new files, which will also be in ARFF format, for further manipulations.