# WEKA ANALYSIS PROTOCOL

August 2024

Erasmus+
Enriching lives, opening minds.

# SUMMARY

# 1.- CHOOSING AND UNDERSTANDING THE DATASET

The first step is to search for a dataset on the web or generate your own. In the previous sessions we focused on learning how to generate your own, now we will see how to search for ready-made datasets. In any case, both must be treated with Weka. There are various websites with datasets. This is a brief list.

| |
|---|
| **https://www.kaggle.com/datasets**<br>Dataset Finder |
| **datos.gob.es/es/catalogo** (Copy and paste)<br>Datasets from the Government and Public Institutions of Spain |
| **https://ec.europa.eu/eurostat/data/database**<br>EU datasets |

***Real dataset example: NASA Exoplanets***

*For example, we are going to apply it to the following dataset, extracted from the NASA website.*

***http://exoplanetarchive.ipac.caltech.edu***

- ✓ *The theme is about all the exoplanets currently known, all the data we have about them.*
- ✓ *The different files are in .csv format, the most common and already known to us from previous sessions.*
- ✓ *Pay attention to the comments as they are essential to understand how the dataset was made.*
- ✓ *We will have to learn to decipher the meaning of fields in astronomy. To do this, there is nothing better than using an online AI.*

# 2.- DATASET PROCESSING

Real datasets can have hundreds of fields and thousands of records. Therefore, it is essential to process them beforehand to normalize values, discard fields that are not needed, eliminate duplicate rows, separate data, etc.

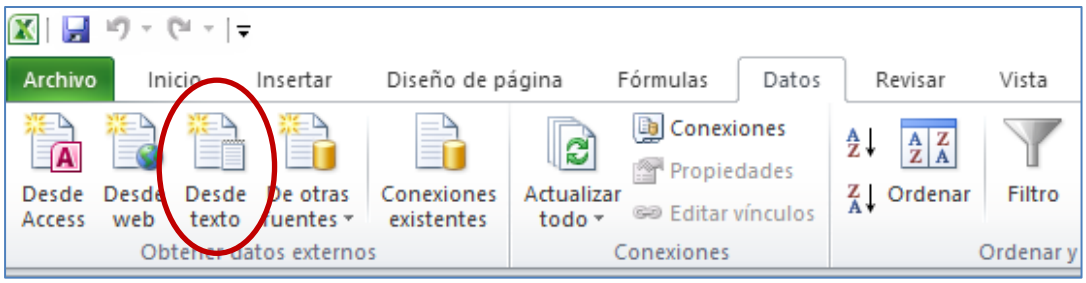
***NASA Exoplanets***

*In our dataset we have 134 fields and 44,799 records or known exoplanets.*

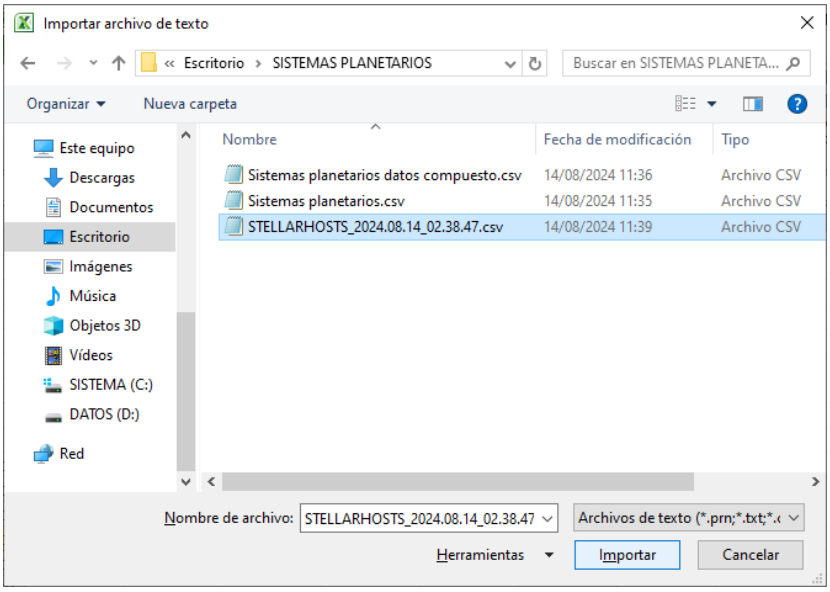
It is always better to process data in a spreadsheet. We will use Microsoft Excel. We have already seen how to convert from Excel to CSV, but now we will do it in reverse order, from CSV to Excel.

## CSV TO EXCEL CONVERSION

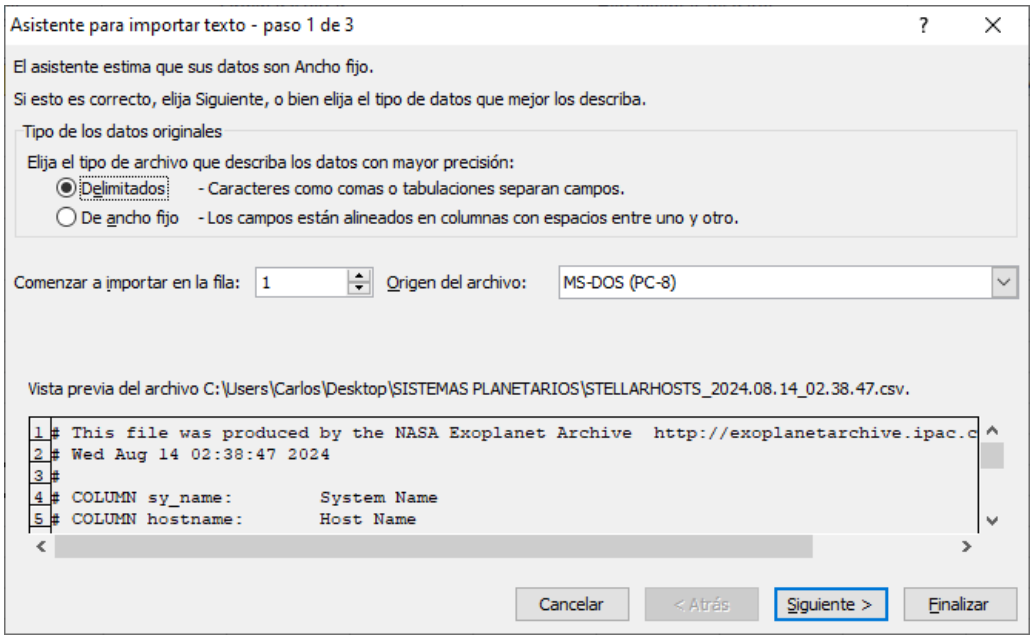
**1st.-** We open Excel with a blank file and go to the ***Data menu*** to select the ***Get External Data From Text option*** .



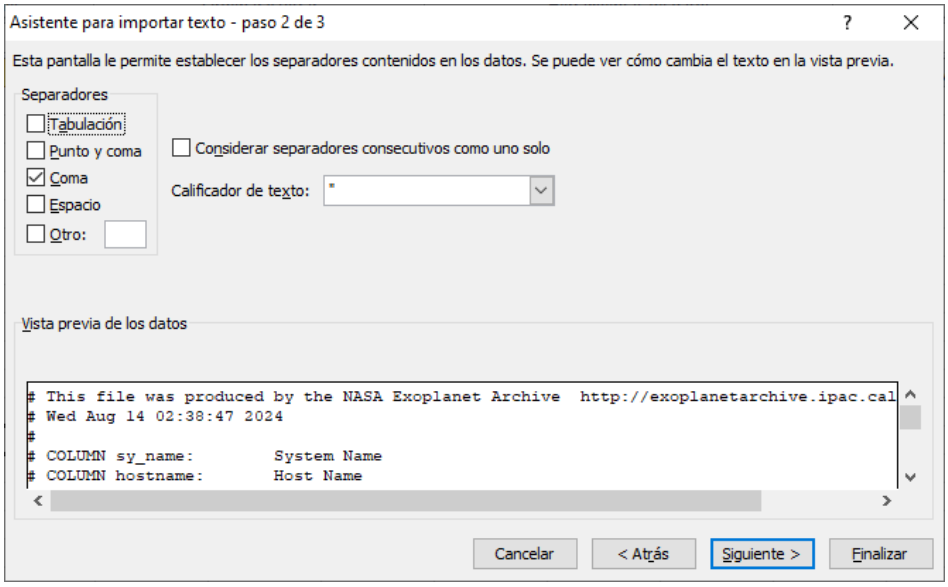

**2nd.-** In the ***Import Text File window*** that opens, we search for and open the desired csv file by clicking on ***Import*** .



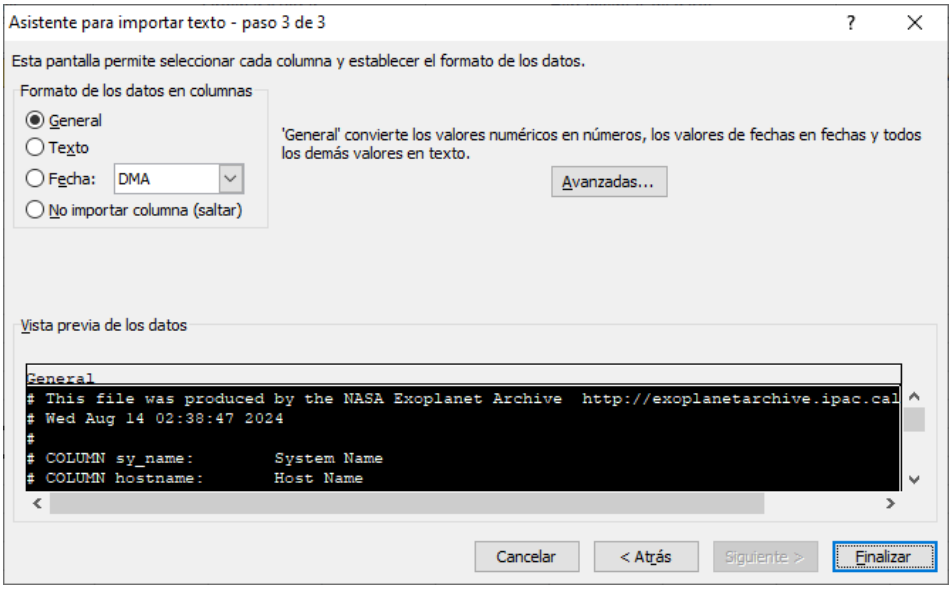

**3rd.-** In step 1 of the wizard that appears, we will mark the ***Delimited option*** to be able to choose in the next step what our separator character is.



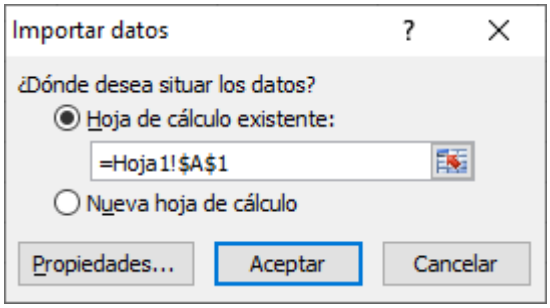

**4th.-** In step 2 of the wizard we select the separator with which our dataset is made, generally the ***comma*** .



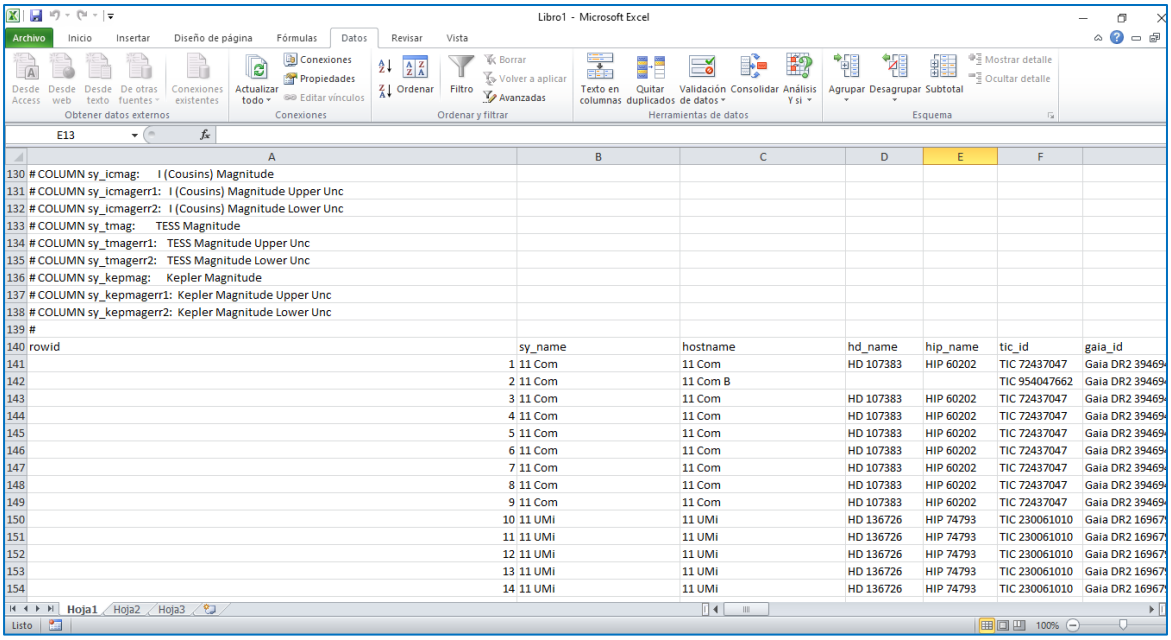

**5th.-** In step 3 of the wizard we leave everything as is and click ***Finish*** .




**6th.-** A small ***Import Data*** window will appear. that asks us where we want the data to be copied. We select cell ***A1*** on the first sheet, if it is not already, and click ***OK*** .




**7th.-** All our data will appear already separated into columns and ready to be processed according to our decisions.



| | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| 130 | # COLUMN sy_icmag: I (Cousins) Magnitude | | | | | | |
| 131 | # COLUMN sy_icmagerr1: I (Cousins) Magnitude Upper Unc | | | | | | |
| 132 | # COLUMN sy_icmagerr2: I (Cousins) Magnitude Lower Unc | | | | | | |
| 133 | # COLUMN sy_tmag: TESS Magnitude | | | | | | |
| 134 | # COLUMN sy_tmagerr1: TESS Magnitude Upper Unc | | | | | | |
| 135 | # COLUMN sy_tmagerr2: TESS Magnitude Lower Unc | | | | | | |
| 136 | # COLUMN sy_kepmag: Kepler Magnitude | | | | | | |
| 137 | # COLUMN sy_kepmagerr1: Kepler Magnitude Upper Unc | | | | | | |
| 138 | # COLUMN sy_kepmagerr2: Kepler Magnitude Lower Unc | | | | | | |
| 139 | # | | | | | | |
| 140 | rowid | sy_name | hostname | hd_name | hip_name | tic_id | gaia_id |
| 141 | 1 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 142 | 2 | 11 Com | 11 Com B | | | TIC 954047662 | Gaia DR2 39469 |
| 143 | 3 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 144 | 4 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 145 | 5 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 146 | 6 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 147 | 7 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 148 | 8 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 149 | 9 | 11 Com | 11 Com | HD 107383 | HIP 60202 | TIC 72437047 | Gaia DR2 39469 |
| 150 | 10 | 11 UMi | 11 UMi | HD 136726 | HIP 74793 | TIC 230061010 | Gaia DR2 16967 |
| 151 | 11 | 11 UMi | 11 UMi | HD 136726 | HIP 74793 | TIC 230061010 | Gaia DR2 16967 |
| 152 | 12 | 11 UMi | 11 UMi | HD 136726 | HIP 74793 | TIC 230061010 | Gaia DR2 16967 |
| 153 | 13 | 11 UMi | 11 UMi | HD 136726 | HIP 74793 | TIC 230061010 | Gaia DR2 16967 |
| 154 | 14 | 11 UMi | 11 UMi | HD 136726 | HIP 74793 | TIC 230061010 | Gaia DR2 16967 |

Once our treatment is finished, we will only have to convert from Excel to ARFF again, going through csv, which we already know and have done previously.

# 3.- DATASET TYPE

When we consider what type of dataset we have (Linear, Non-Linear or with time series) and what we want to predict, we must look for a field to predict it.

***Exoplanets NASA***

*Our dataset is non-linear and we can predict the number of moons, the spectral type, whether it is circumbinary, etc. .*

# 4.- STEPS OF THE TREATMENT IN WEKA

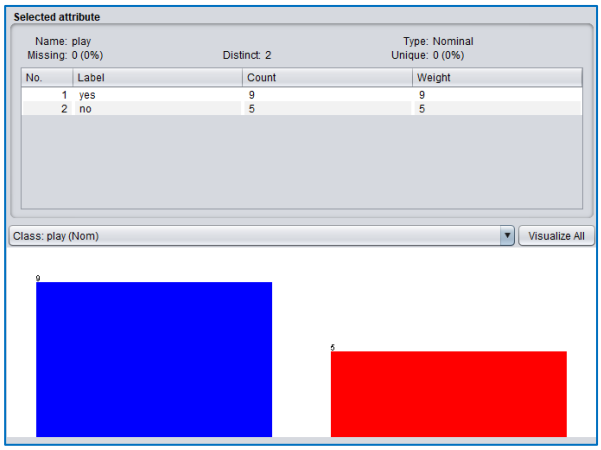
We recall the most used algorithms in each type of data set model and in color those explained in the previous session.

| ***DATASET TYPE*** | ***ALGORITHMS*** |
|---|---|
| ***LINEAR*** | ***ZeroR, OneR, DecisionTable, J48, Random Forest, Random Tree, KNN, Bayes, Linear regression, CostSensitiveClassifer*** |
| ***NON-LINEAR*** | ***KNN, Bayes, CostSensitiveClassifer, Artificial Neural Networks (MultilayerPerceptron)***<br>*SVM, SMO, Voted Perceptron, SGD, SGD Text, Gaussian Processes, Recurrent neural networks, Recurrent neural networks* |
| ***TIME SERIES*** | ***Forecast*** |

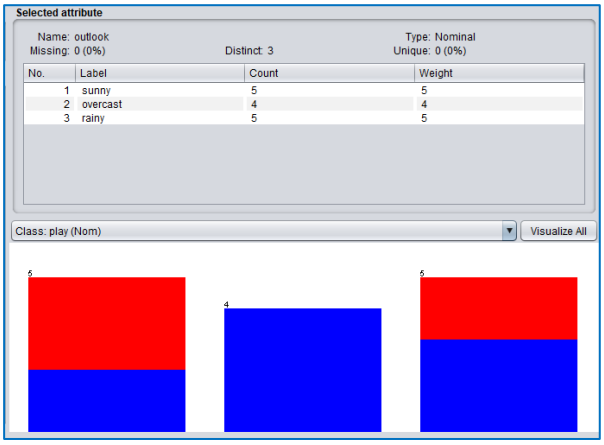
*We will apply the steps to this simple dataset on the prediction of whether one can do sports.*

## STEP 1 - PREPROCESS

- ✓ Place the field to be predicted as the last one and see the distribution of the values and their colors. ***The blue color means that you can do sports and the red color means that you cannot*** .

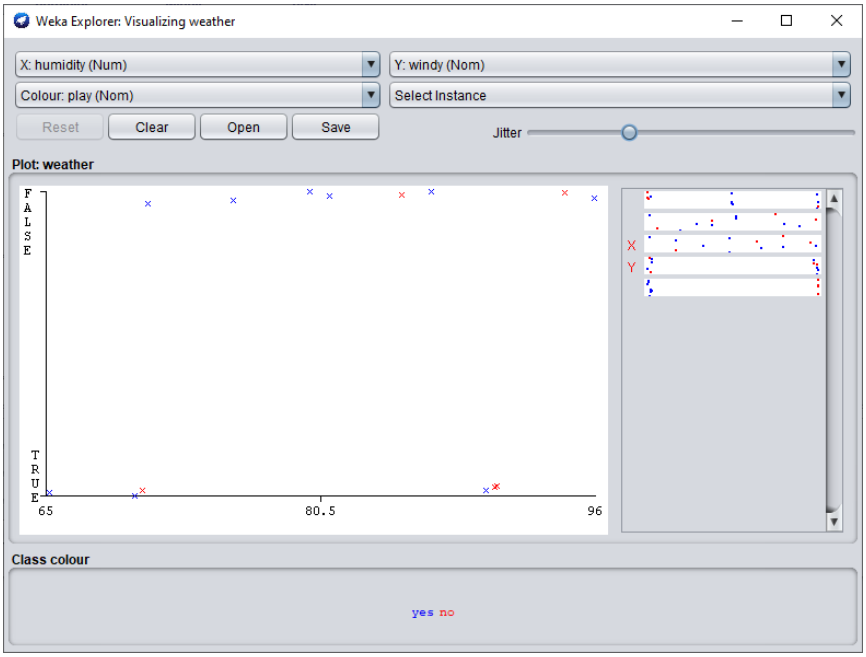



- ✓ Investigate the histograms of each field one by one to determine which one best separates the values of the field to be predicted. The outlook field clearly separates when it is possible to do sports if the weather is cloudy.




## STEP 2 - VISUALIZE

- ✓ Investigate whether there is a pair of fields that clearly separates the data. If so, generate a new field derived from the pair according to mathematical theories.




## STEP 3 - CLASSIFY

Once we know what our dataset is like and what the most important fields are based on the histograms and field pairs, it is time for the algorithms to confirm this for us.

We must run the various algorithms to extract the conclusions and the highest possible precision with each of them. Concepts such as the confusion matrix, the cost matrix, etc. are fundamental.

Once done, we will have to choose the one we believe is the best algorithm for our data model, which will be the one we apply to predict in the next step.

## STEP 4 – CLASSIFY & PREDICTION

Once we have learned about the type of our dataset, the best algorithms applicable to it, and how to configure them correctly, it is time to start making our predictions about future values. This applies to both linear and non-linear models.

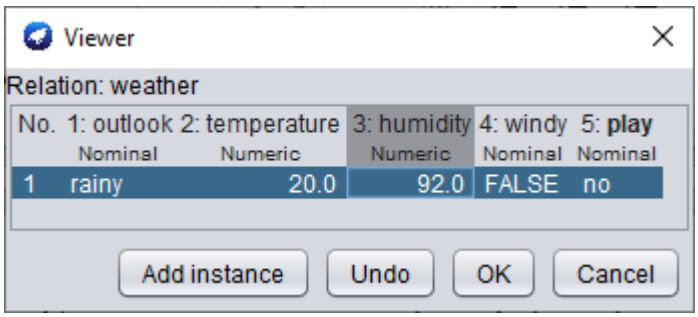
### *Weather.numeric dataset prediction*

*Now we are going to predict whether we can do sports on a particular day. We are going to make a mixture of the values that occur on a rainy day, for example the following,*

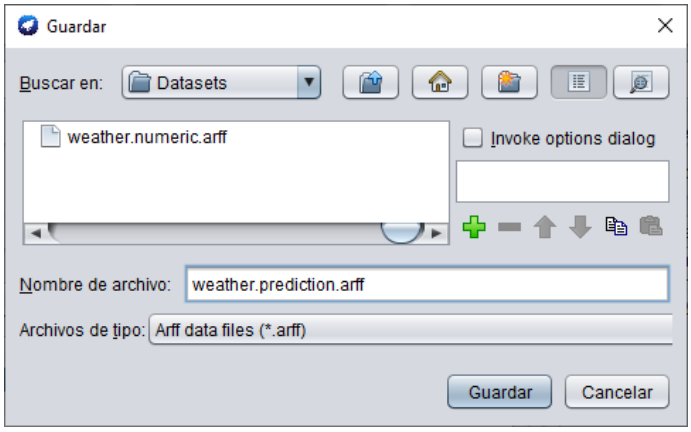
***rainy, 20, 92, FALSE, no***

*The value of the class to be predicted is irrelevant, since what matters is the algorithm's prediction, not what we put in. If we get it right, it will classify it as correct, if not, as incorrect.*

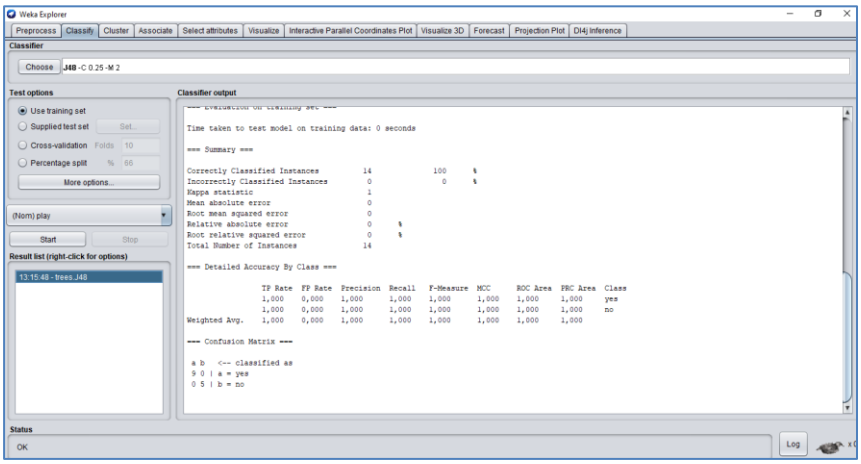
***STEP A:*** *we will load the* ***weather.numeric.arff dataset*** *if it is not already loaded, and go to the* ***Edit… button*** *on the* ***Preprocess tab*** *. In the pop-up window we will delete all the records except one, using the context menu. Then we will change the values it had to the ones above that we want to predict by double-clicking on them. We will click on* ***OK*** *. (It can also be done directly in the notes box manually).*



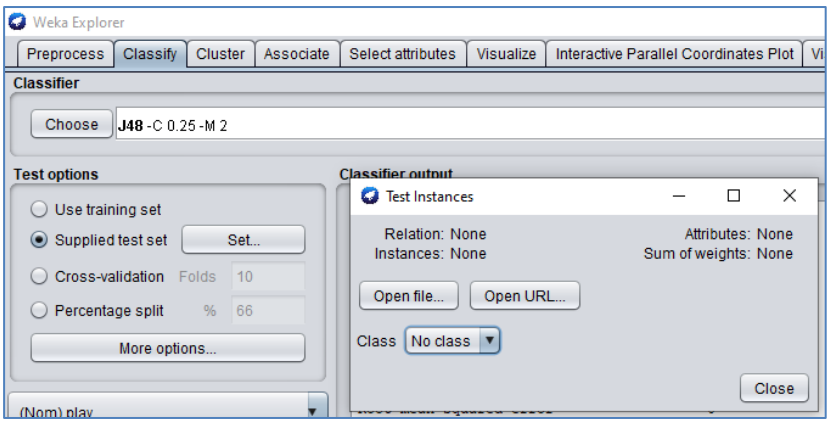

***STEP B:*** *Returning to the* ***Preprocess tab*** *We will see that now the dataset only has a single record. We will use the* ***Save…*** *button to save it with another name, for example* ***weather.prediction.arff***




***STEP C:*** *We will reload the* ***weather.numeric.arff training dataset*** *, choose the desired algorithm, configure it and pass it with* ***Use training set*** *. For example, we will do it with the* ***J48*** *according to the image below to obtain 100% accuracy.*




***STEP D:*** *in the* ***Test options box*** *We will choose* ***Supplied Test Set*** *, press the* ***Set button*** *and in the pop-up window we will press the* ***Open File button…*** *to select the file to predict,* ***weather.prediction.arff*** *, which has our record to predict, and click on the* ***Close button*** *. It can actually contain as many as we want.*




***STEP E:*** *we will press the Start button and observe the result of the prediction,* ***not correct*** *, but above all the confusion matrix. It has classified as yes when we actually put no. Which means that* ***the prediction of the J48 algorithm is yes*** *. (If we had many records we could go to Visualize Classifier Errors to see it better).*

*to b <-- classified as*
*0 0 | a = yes*
***1*** *0 | b = no*

***Weka does not allow us to put a ? in the value of the class to predict because then in the summary it shows us that it has ignored a record.***

# 5.- CREATING A REPORT

The time has come to prepare a final report explaining,

- ✓ The theme of our dataset, type of data model, its origin and the reasons for its choice.
- ✓ The treatment we have carried out and the steps we have followed to generate our arff file.
- ✓ How we have treated it in Weka: preprocessing and analysis. Choice and configuration of the classifier.
- ✓ Examples of predictions made and results obtained.
- ✓ Generation of graphs on the results.

***See you in Madrid in June 2025 – We're done!***

*In Alcorcón, each country will present and defend its data report.*