

# WEKA ANALYSIS PROTOCOL

August 2024





- **1.- CHOOSING AND UNDERSTANDING THE DATASET**
- 2.- DATASET PROCESSING
- **3.- DATASET TYPE**
- 4.- STEPS OF THE TREATMENT IN WEKA
- **5.- CREATING A REPORT**



## 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 Dataset Finder
datos.gob.es/es/catalogo (Copy and paste) Datasets from the Government and Public Institutions of Spain
https://ec.europa.eu/eurostat/data/database EU datasets



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.



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*.

🔣 Importar archivo d	de texto				×
← → ~ ↑ 📙	« Escrit	torio > SISTEMAS PLANETARI	ٽ × So	Buscar en SISTEMAS	planeta 🔎
Organizar 👻 Nu	ueva carp	eta			• 🔟 🕐
💻 Este equipo	^	Nombre		Fecha de modificación	Тіро
👆 Descargas		🧾 Sistemas planetarios dato	s compuesto.csv	14/08/2024 11:36	Archivo CSV
Documentos		Sistemas planetarios.csv		14/08/2024 11:35	Archivo CSV
Escritorio		STELLARHOSTS_2024.08.1	4_02.38.47.csv	14/08/2024 11:39	Archivo CSV
Imágenes					
Música					
🗊 Objetos 3D					
Vídeos					
SISTEMA (C:)					
👝 DATOS (D:)					
🔿 Red					
-	~ <				>
	<u>N</u> ombre	e de archivo: STELLARHOSTS_	2024.08.14_02.38.47、	Archivos de texto (*	.prn;*.txt;*.c >
			<u>H</u> erramientas	✓ Importar	Cancelar

**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.

Asistente para importar texto - paso	de 3				?	Х
El asistente estima que sus datos son Ar	cho fijo.					
Si esto es correcto, elija Siguiente, o bie	n elija el tipo de datos que	mejor los desc	iba.			
Tipo de los datos originales						
Elija el tipo de archivo que describa los	datos con mayor precisión					
Delimitados - Caracteres co	mo comas o tabulaciones s	eparan campos	s.			
O De <u>a</u> ncho fijo - Los campos es	tán alineados en columnas	con espacios e	entre uno y otro.			
Comenzar a importar en la fila: 1	Origen del archivel	vo: MS-DO	S (PC-8)			$\sim$
Vista previa del archivo C:\Users\Carlo	s \Desktop \SISTEMAS PLAN	VETARIOS (STE	LARHOSTS_2024.0	8.14_02.38.47.csv	•	
1 # This file was produced	by the NASA Exopl	anet Archiv	/e http://exo	planetarchive	.ipac.c	^
2 # Wed Aug 14 02:38:47 20	24					
4 # COLUMN sy name:	System Name					
5 # COLUMN hostname:	Host Name					v
<					>	
		Cancelar	< Atrás	<u>S</u> iguiente >	<u>F</u> inaliz	ar

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



Asistente para importar texto - paso 2 de 3	?	×
Esta pantalla le permite establecer los separadores contenidos en los datos. Se puede ver cómo cambia el texto en la vist Separadores Tabulación Punto y coma Considerar separadores consecutivos como uno solo	a previa.	
<pre># This file was produced by the NASA Exoplanet Archive http://exoplanetarchive.ip # Wed Aug 14 02:38:47 2024 # # COLUMN sy_name: System Name # COLUMN hostname: Host Name </pre>	ac.cal	<b>^</b>
Cancelar < At <u>r</u> ás <u>Siguiente</u> >	<u>F</u> inaliza	ar

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

Asistente para importar texto - pa	io 3 de 3	?	×
Esta pantalla permite seleccionar cad	a columna y establecer el formato de los datos.		
Formato de los datos en columnas			
<ul> <li>● <u>G</u>eneral</li> <li>○ Te<u>x</u>to</li> <li>○ F<u>e</u>cha: DMA </li> </ul>	'General' convierte los valores numéricos en números, los valores de fech los demás valores en texto. <u>A</u> vanzadas	has en fechas y t	odos
○ <u>N</u> o importar columna (saltar)			
<u>Vi</u> sta previa de los datos			
Vista previa de los datos General # This file was produced	by the NASA Exoplanet Archive http://exoplanetarch	hive.ipac.ca	,1 ^
Vista previa de los datos <u> General</u> # This file was produced # Wed Aug 14 02:38:47 20 #	by the NASA Exoplanet Archive http://exoplanetarch 24	hive.ipac.ca	1 ^
Vista previa de los datos <u>Feneral</u> # This file was produced # Wed Aug 14 02:38:47 20 # # COLUMN sy_name:	by the NASA Exoplanet Archive http://exoplanetarch 24 System Name	hive.ipac.ca	,1 ^
Vista previa de los datos General # This file was produced # Wed Aug 14 02:38:47 20 # # COLUMN sy_name: # COLUMN hostname:	by the NASA Exoplanet Archive http://exoplanetarch 24 System Name Host Name	hive.ipac.ca	<b>1</b> ^
Vista previa de los datos <u>Peneral</u> # This file was produced # Wed Aug 14 02:38:47 20 # # COLUMN sy_name: # COLUMN hostname: <	by the NASA Exoplanet Archive http://exoplanetarch 24 System Name Host Name	hive.ipac.ca	• • •

**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*.

Importa	ar datos	?	×
Dónde) آ ()	desea situ <u>H</u> oja de cál		
	=Hoja 1!\$/	Ē	<b>6</b>
0	N <u>u</u> eva hoja		
Propie	dades	Can	celar

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



<b>X</b>   <b>→</b> <sup>(1</sup> -   <del>-</del>	Libro1	<ul> <li>Microsoft Excel</li> </ul>					- 0 ×
Archivo Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista							ھ 🕝 🕞 ۵
Derde Desde De otras Access web texto fuerites – existentes	Borrar Volver a aplicar <b>Avanzadas</b>	Texto en Quitar columnas duplicado	Validación Consolidar Análisis se de datos * Y si *	Agrupar Desagrup	ar Subtotal	Mostrar detaile Dcuitar detaile	_
Obtener datos externos Conexiones Ordenar y filtr	ar	Herra	amientas de datos		Esquema	Gi.	
E13 • ( <i>f</i> x							
A		В	С	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 could			hostnama	hd name	hin name	tio id	gaia id
140 FOWIG	sy_name		nostname	nd_name	nip_name	tic_id	gala_id
141	1 11 Com		11 Com P	HD 107565	HIP 00202	TIC 054047652	Gaia DR2 39409
142	2 11 Com		11 Com B	UD 107292	HID 60202	TIC 72427047	Gala DR2 39409
143	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
145	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 394694
148	8 11 Com		11 Com	HD 107383	HIP 60202	TIC 72437047	Gaia DR2 394694
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
H + H Hoja1 Hoja2 Hoja3							▶ []
Listo 🔚						<b>I</b> II 100% (-	)

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.



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) 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*.

Nar Missi	me: play ing: 0 (0%)	Distinct: 2	Type: Nominal Unique: 0 (0%)
No.	Label	Count	Weight
	1 yes	9	9
	2 no	5	5
ss: p	lay (Nom)		Visualize
iss: p	lay (Nom)		• Visualize
ss: p	lay (Nom)		▼ Visualize
ss: p	lay (Nom)		Visualize
ss: p	lay (Nom)		Visualize
ss: p	lay (Nom)		Visualize
9 9	lay (Nom)	í	Visualize

✓ 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.

Missin	e: outlook g: 0 (0%)	Distinct 3	Type: Nominal Unique: 0 (0%)
No.	Label	Count	Weight
	1 sunny	5	5
	2 overcast	4	4
	3 rainy	5	5
ass: pla	ay (Nom)		Visualiz
5			6

## **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.



: humidity	(Num)			Y: windy (No	om)			 _
olour: pla	ıy (Nom)		•	Select Insta	nce			
Reset	Clear	Open	Save		Jitter =		-0	 
ot: weathe	er							
	×	×	××	××		××	X 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
T R U E 65	×		80.5		×*			
iss colou	r							

#### **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.



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).



0	Viewer				×
Rela	tion: weather				
No.	1: outlook 2: Nominal	temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	rainy	20.0	92.0	FALSE	no
	Add ir	nstance	Undo	ок	Cancel

**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** 

🥥 Guardar	×
Buscar en: 📋 Datasets	🗎 🝙 👔 🝺
weather.numeric.arff	Invoke options dialog
•	
Nombre de archivo: weather.prediction.arff	
Archivos de tipo: Arff data files (*.arff)	
	Guardar Cancelar

**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.

Weka Explo	Orer	fe Cluster	Arrorit	da Č Salact attributa	Viewaliza	otoractivo Pa	vallel Coordin	ates Plot	Vicustine 2D	Enrocaet	Projection	Plot DI4i In	ference		- 0	×
lassifier	01838	in Consten	- ABBOCH		Visualize	instauter e	statet Coordin	alearioi	viauance 30	rorecast	riojecaom		lerence	 		
Choose	J48 · C	0.25 -M 2														
est options				Classifier output												
<ul> <li>Use tra</li> </ul>	ining set			Everage	ION ON CEALUR	my are	_									
O Supplie	ed test set	Se	t	Time taken	to test model	on train:	ing data: 0	seconds								1
O Cross-	validation	Folds 1		=== Summary												- 1
Percent	tage split	% 6	6	Correctly C	lessified In-	tances	14		100							- 1
	More o	ntions		Incorrectly	Classified 1	nstances			0	÷ .						- 1
				Kappa stati	stic		1									- 1
				Mean absolu	te error		0									- 1
Nom) play				Root mean s	guared error		0									- 1
			_	Relative ab	solute error		0									- 1
Start			top	Root relati	ve squared er	ror	0	8								- 1
0.010				Total Numbe	r of Instance	8	14									- 1
esult list (rig	ght-click I	for options)														- 1
42-45-40			_	Detaile	d Accuracy By	Class ===	-									
13, 15, 48 -	trees.J46					-				1000		-	01			
					IF RACE	e pr sace	Precision	Recall	2-neasure	NUC 1. OOO	RUC Area	PHL Area	CIASS			
					1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Yes			
				Valabrad Au	- 1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	110			
				weighted av	9. 1,000	0,000	1,000	2,000	1,000	1,000	2,000	1,000				
				Confinat	on Marris											
				00011481	OIL PROLAG											
				ab c	classified as											
				9014-	vea											
				051b=	00											
																T.
tatus														 		
OK															Log	18
																-

**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.



Choose       J48 - C 0.25 - M 2         est options       Classifier output         Use training set       Image: Classifier output         Supplied test set       Set         Cross-validation       Folds         Percentage split       % 66	Preprocess Classify Cluster Associate	e Select attributes Visualize In	teractive Parallel Coordinates Plot				
Choose       J48 - C 0.25 - M 2         est options         Classifier outnut         O Use training set       Image: Colspan="2">Classifier outnut         Image: Colspan="2">Supplied test set       Set         Relation: None       Attributes: Non         Instances: None       Sum of weights: Non         Open file       Open URL	assifier						
Classifier output         O Use training set         Supplied test set         Supplied test set         Cross-validation         Folds         Percentage split         % 66	Choose J48 -C 0.25 -M 2						
Classifier output       Use training set       Supplied test set       Supplied test set       Cross-validation       Folds       Percentage split       % 66							
O Use training set <ul> <li>Test Instances</li> <li>Test Instances</li> <li>Cross-validation</li> <li>Folds</li> <li>Percentage split</li> <li>%</li> <li>66</li> </ul> <ul> <li>Open file</li> <li>Open URL</li> </ul> <ul> <li>Open URL</li> </ul> <ul> <li>Open URL</li> </ul> <ul> <li>Open file</li> <li>Open URL</li> </ul> <ul> <li>Open file</li> <li>Open URL</li> </ul> <ul> <li>Open file</li> <li>Open URL</li> <li>Open URL</li> <li>Open URL</li> <li>Open URL</li> </ul>	est options	Classifier output					
Supplied test set     Set     Relation: None     Attributes: Non     Instances: None     Sum of weights: Non     Open file     Open URL	<ul> <li>Use training set</li> </ul>	G Test Instances	- 🗆 X				
O Cross-validation     Folds     10       O Percentage split     %     66   Open URL	Supplied test set     Set	Relation: None Instances: None	Attributes: None Sum of weights: None				
O Percentage split % 66 Open lile	Cross-validation Folds 10						
	O Percentage split % 66						
More options Class No class	More options	Class No class					

**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,

- $\checkmark$  The theme of our dataset, type of data model, its origin and the reasons for its choice.
- $\checkmark$  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.
- $\checkmark$  Generation of graphs on the results.



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

