

Universidad ORT Uruguay
Facultad de Ingeniería

Data Mining

Obligatorio

Clavijo, Tomás (235426)

Barreto, Sofía (258216)

Docente: Esther Hochsztain

2023

Índice

Parte A	1
Ejercicio 1)	1
1.1) Reglas de Asociación – (Tanagra, asociacionObligatorio2023.txt)	1
1.2) Reglas de Asociación – (Weka, ProyectosObligatorio2023.txt)	15
Ejercicio 2)	38
2.1) Árboles de decisión – (Tanagra, empresasObligatorio2023.txt)	38
2.2) Árboles de decisión – (Weka, ProyectosObligatorio2023.txt)	46
Ejercicio 3)	58
3.1) Clustering – (Tanagra, empresasObligatorio2023.txt)	58
3.2) Clustering – (Knime, text_mining_clustering_1Obligatorio2023.txt)	69
Ejercicio 4)	85
4.1) Redes Neuronales – (Tanagra, empresasObligatorio2023.txt)	85
4.2) Redes Neuronales – (Weka, ProyectosObligatorio2023.txt)	94
Ejercicio 5)	101
5.1) SVM – (Tanagra, empresasObligatorio2023.txt)	101
Ejercicio 6)	110
6.1) Supervisado – (Tanagra, text_mining_clas1Obligatorio2023.txt)	110
6.2) No Supervisado – (Tanagra, text_mining_clustering_1Obligatorio2023.txt)	119
Ejercicio 7)	126
7.1) Datos Atípicos – (Knime, ProyectoObligatorio2023.txt)	126
Parte B	139
Ejercicio 1)	139
Ejercicio 2)	147
Ejercicio 3)	149
Ejercicio 4)	151
Ejercicio 5)	153
Ejercicio 6)	156
Parte C	160
Ejercicio 1)	160
Ejercicio 2)	160
Parte D	162
Ejercicio 1)	162
Ejercicio 2)	163

Parte E	165
Ejercicio 1)	165
Ejercicio 2)	167
Ejercicio 3)	167
Parte F	168
Ejercicio 1)	168
Ejercicio 2)	169
Bibliografía	170

Parte A

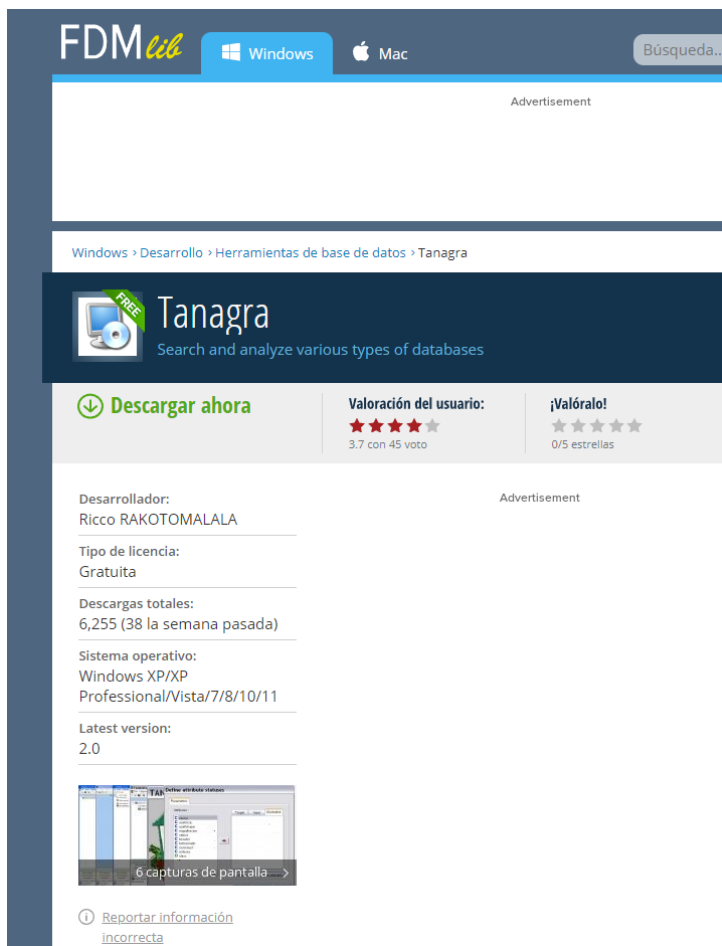
Ejercicio 1)

Construya por lo menos 2 reglas de asociación para cada pareja (herramienta, dataset) utilizando dos herramientas y dos datasets (un dataset para cada herramienta).

1.1) Reglas de Asociación – (Tanagra, asociacionObligatorio2023.txt)

Debemos comenzar por descargar el programa Tanagra desde el link:
<https://es.freedownloadmanager.org/Windows-PC/Tanagra-GRATIS.html>.

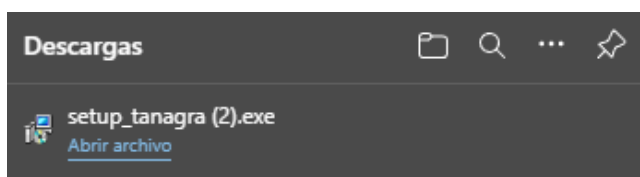
Observándose la siguiente pantalla:



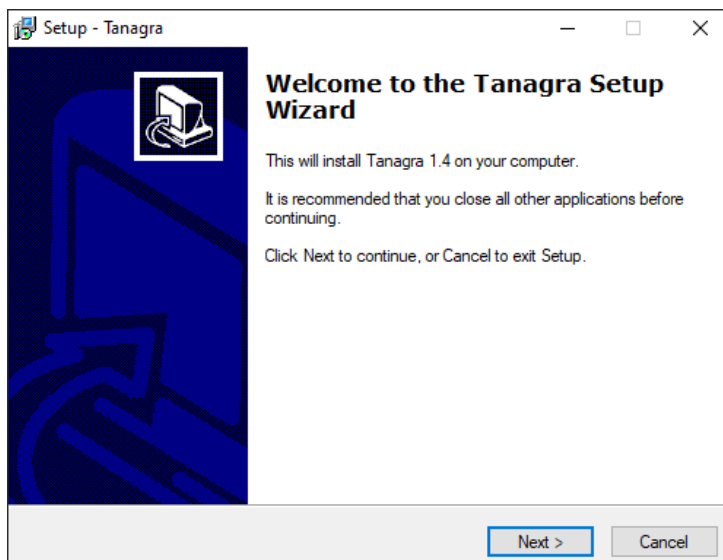
Una vez en esta pantalla, hacemos clic sobre “Descargar ahora”, comenzará a descargarse el archivo correspondiente y se observará la siguiente pantalla:



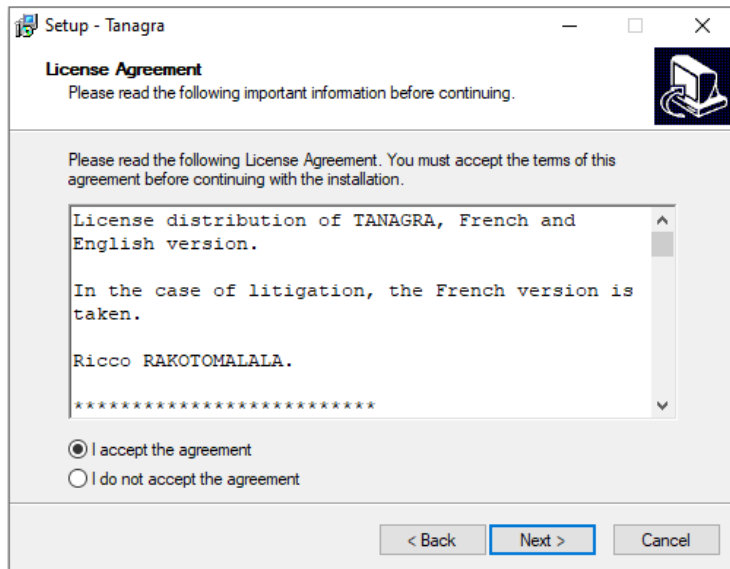
Como se menciona, en caso de no comenzar a descargarse el archivo automáticamente, debemos hacer clic donde dice “Si no se inicia la descarga, haz clic aquí”. El archivo descargado será de extensión .exe. Se puede observar a continuación:



Para proceder con la instalación debemos abrirlo:



Hacemos clic en “Next”, y preguntará si aceptamos el acuerdo de licencia (“Licence Agreement”, presionaremos la opción “I accept the agreement”).




Continuamos haciendo clic en “Next” en las próximas ventanas hasta que nos aparezca “Install”, hacemos clic en la opción y cuando finalice presionamos “Finish”. Ya tendremos instalado el programa para realizar nuestro trabajo.

Por último, debemos descargar el archivo con el que trabajaremos, siendo en este caso: “asociacionObligatorio2023.txt”.

Para esto ingresaremos en el curso de “Data Mining” del sitio de Aulas de la Universidad ORT Uruguay.



Una vez en el sitio nos dirigimos a la carpeta “Obligatorios” (en la imagen se marca con color azul).



ORT

UNIVERSIDAD ORT

TIENEY

Aulas

Apoyo online a cursos

Información General + Carpetas

TAREAS

Acceso a salas Zoom

Grabaciones

Obligatorio

Tema 5

Espacio Docente

Teórico clase a clase. Grupo A. Martes y Miércoles

Teórico clase a clase. Grupo B. Viernes

Tareas Grupo A

Tareas Grupo B

Data mining

Avisos

Foro Data Mining. Primer semestre 2023

FORMA DE EVALUACIÓN

Parcial: 35 puntos

Obligatorio: 50 puntos. El obligatorio se puede hacer en grupo de dos, o individualmente. No se admiten grupos de tres estudiantes

Participación en clase (Tareas): 15 puntos

Las tareas tienen como fecha de entrega la hora de comienzo de la próxima clase, pueden subirse hasta dos semanas después de presentada la tarea (a la hora de comienzo de la clase). Esto puede modificarse por los feriados, dado que en los feriados no vencen tareas. Se eliminan del total las tareas de dos clases. Después del vencimiento (2 semanas) la tarea no se puede subir.

Las tareas de cada semana se pueden hacer en forma individual o por dos estudiantes. Si se realizan por dos estudiantes, debe indicarse en la primera línea del documento los nombres de los autores (si son dos), y cada uno de los integrantes del grupo sube la tarea (aunque sea la misma).

Programa Data Mining

Datos

Materiales

Artículos

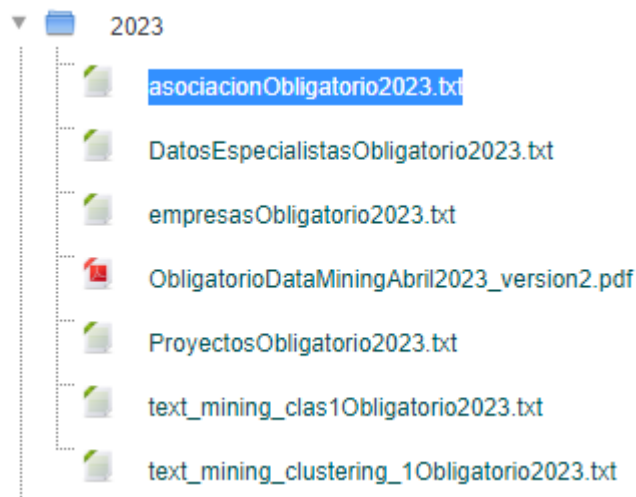
Libros

Parciales

Obligatorios

Manuales-Ayuda Herramientas

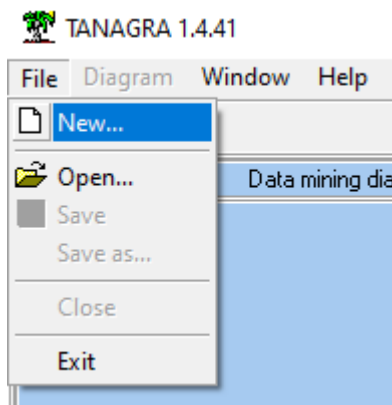
Buscaremos la carpeta correspondiente a “2023” y descargaremos el archivo: “asociacionObligatorio2023.txt” haciendo clic sobre el mismo.



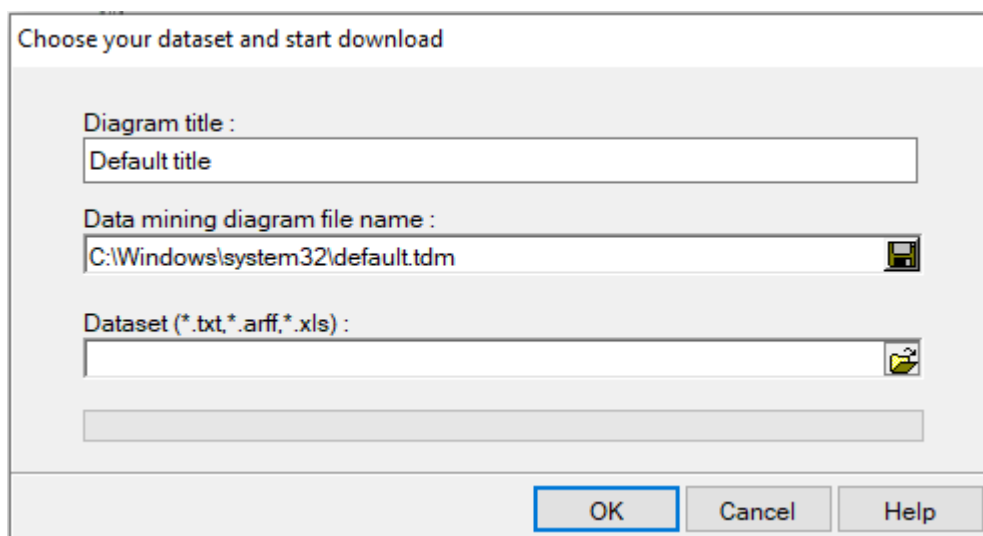
Destacar que este proceso de descarga será igual para todos los archivos que se trabajarán en este obligatorio.

Teniendo Tanagra instalado y el archivo descargado, podemos proceder con el trabajo para hallar reglas de asociación.

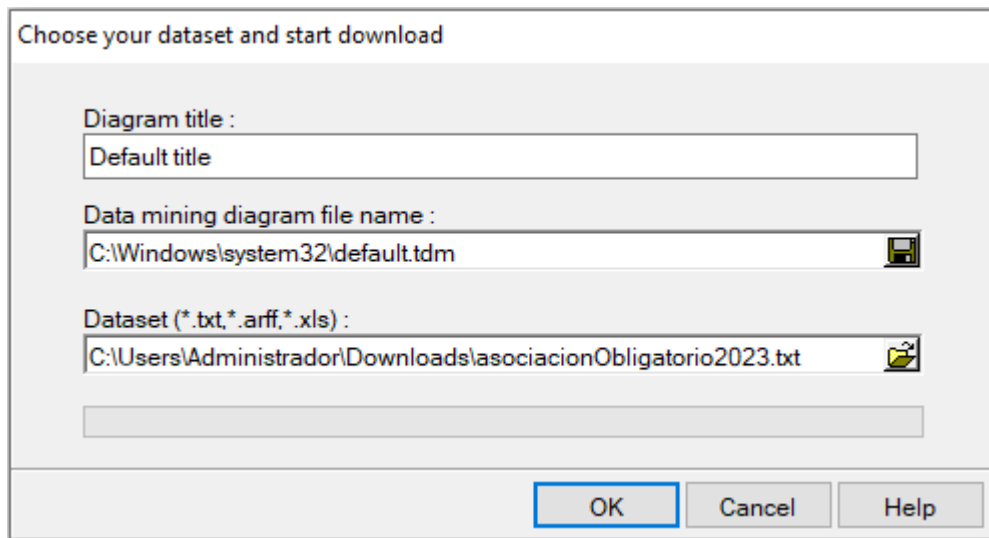
Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



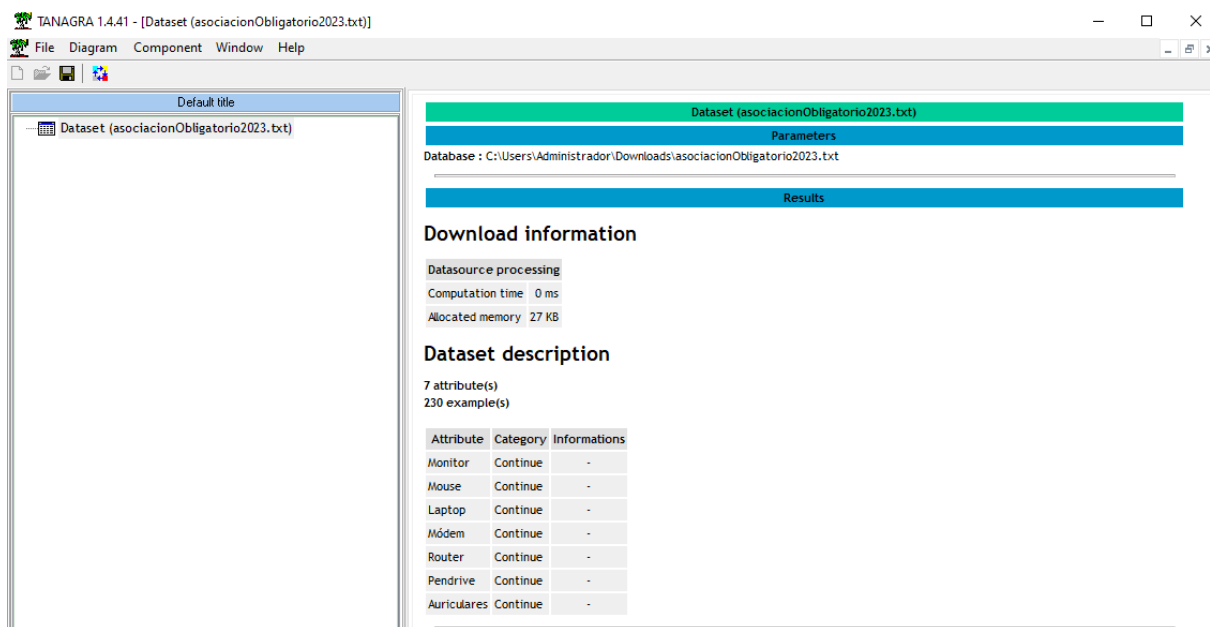
Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo Dataset (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo asociacionObligatorio2023.txt en la ubicación donde fue guardado.



Una vez seleccionado el archivo se nos cargará en el campo Dataset la ruta de este.

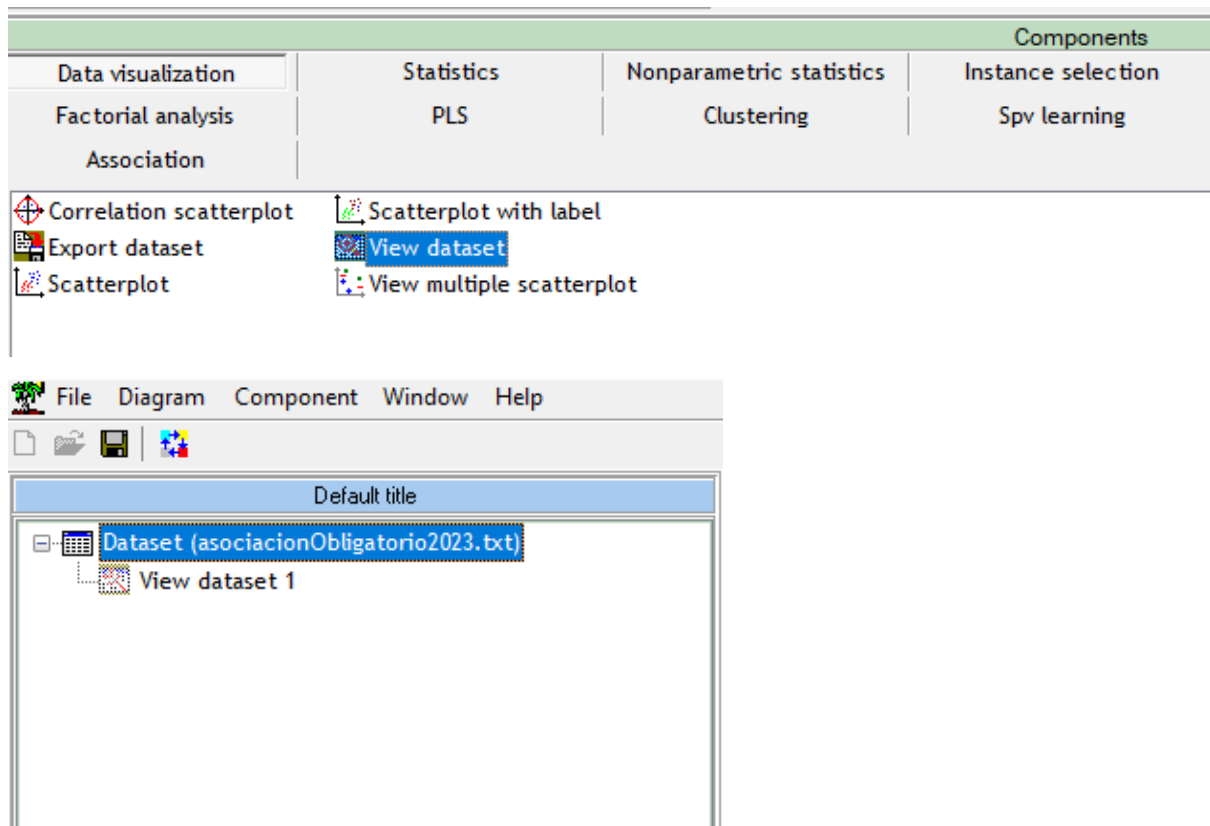


Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:



Si previo a realizar el trabajo sobre el archivo deseamos conocerlo, es decir, observar la información que contiene, podemos crear un *View DataSet*

Nos dirigimos a la sección *Components* que se encuentra en la parte inferior de la ventana, hacemos clic en *Data visualization* y seleccionamos *View dataset*, arrastrándolo hacia la sección de la izquierda, justo debajo del Dataset.



Para proceder con su ejecución hacemos doble clic en el mismo. Observaremos los datos en forma de tabla.

The screenshot shows the same software application, but now the 'View dataset 1' option is selected, and the data is displayed in a table. The table has 33 rows and 8 columns. The columns are labeled: Monitor, Mouse, Laptop, Módem, Router, Pendrive, and Auriculare. The data is as follows:

	Monitor	Mouse	Laptop	Módem	Router	Pendrive	Auriculare
1	1	1	0	1	0	0	0
2	0	0	1	0	0	1	1
3	0	0	0	0	0	1	1
4	1	0	1	1	0	0	1
5	0	0	0	0	0	1	1
6	0	1	1	1	0	1	1
7	1	1	1	0	0	1	0
8	0	0	1	1	0	1	1
9	0	1	1	0	0	1	1
10	0	0	0	0	0	1	1
11	0	0	0	0	0	1	1
12	0	0	0	0	0	1	1
13	1	0	1	1	0	0	1
14	0	0	0	0	0	0	1
15	0	1	0	1	1	0	1
16	1	1	1	1	1	0	0
17	0	1	0	1	1	0	0
18	0	0	0	0	1	0	0
19	1	0	1	1	0	0	0
20	1	0	1	1	0	1	1
21	1	0	1	1	0	0	0
22	0	0	0	0	0	1	1
23	0	0	0	0	0	1	1
24	1	0	1	1	0	1	1
25	0	0	0	0	0	0	1
26	0	1	0	1	1	0	1
27	1	1	1	1	1	0	0
28	0	1	0	1	1	0	0
29	0	0	0	0	1	0	0
30	1	0	1	1	0	0	0
31	1	0	1	1	0	1	1
32	0	0	0	0	0	1	1
33	0	0	0	0	0	1	1

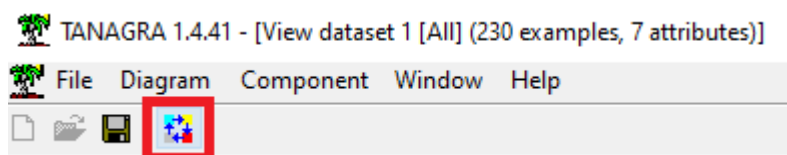
Podemos observar que las columnas tienen los nombres de los atributos, pero los valores de las mismas son 0's y 1's. Es decir, cada variable puede tener 2 valores de carácter booleano; a pesar de que Tanagra las identifique como variables continuas. Cada fila del archivo representa un ticket, y cada variable dentro del ticket, representa cada uno de los productos que pueden o no estar presentes en la compra. Si tiene el valor 1, significa que dicho producto se ha comprado, y en contraparte, si tiene el valor 0 significa que no se ha comprado.

Para generar reglas de asociación trabajaremos con los siguientes parámetros:

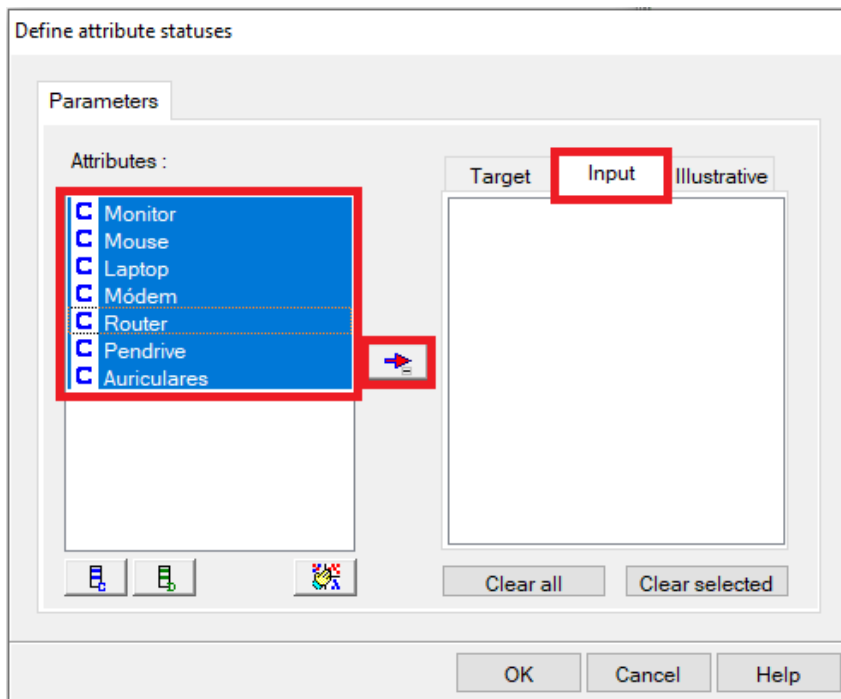
- **Soporte mínimo: 20%**
- **Confianza mínima: 50%**
- **MaxCardItemsets: 4**
- **Lift: 1,1**

Para determinar las reglas de asociación con los parámetros dados, procederemos a seguir los siguientes pasos:

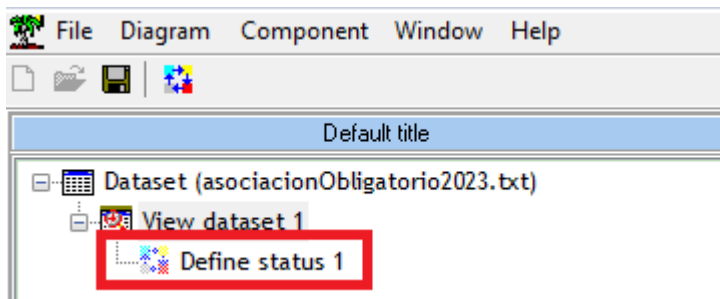
- 1) Haremos un *define status* con el objetivo de indicar con qué atributos trabajaremos. Para esto, debemos hacer clic en el ícono marcado en rojo en la siguiente imagen:



- 2) Se abrirá una ventana denominada *Define attribute statuses*, donde debemos seleccionar todos los atributos y presionar la flecha indicada para colocarlas en el *input*. Observar que en el cuadrante de la derecha esté indicada la opción *input*.



- 3) Una vez los atributos marcados se encuentren del lado derecho, se debe hacer clic sobre el botón **OK**. Luego, deberemos hacer clic sobre “Define status 1”, que se encontrará colgado debajo del dataset, con esto lo ejecutaremos.

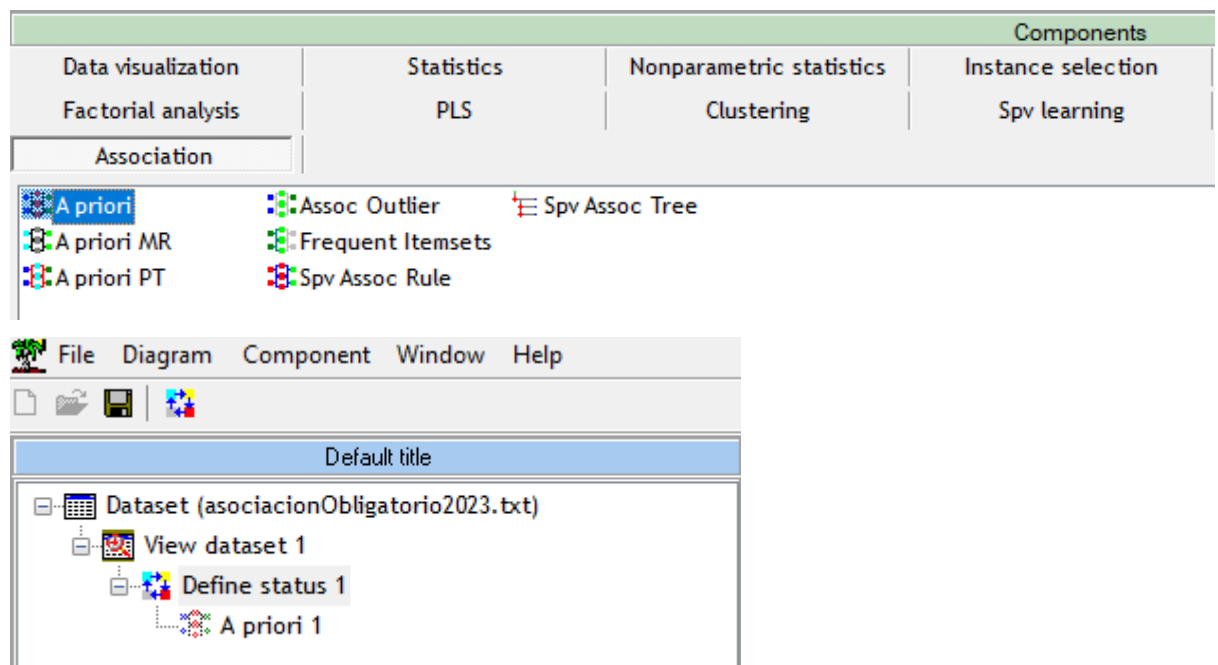


Después de ejecutar, si deseamos verificar lo hecho hasta el momento, en la tabla que se visualiza deberá aparecer el valor **yes** en la columna **input** para todos los atributos marcados en el paso anterior.

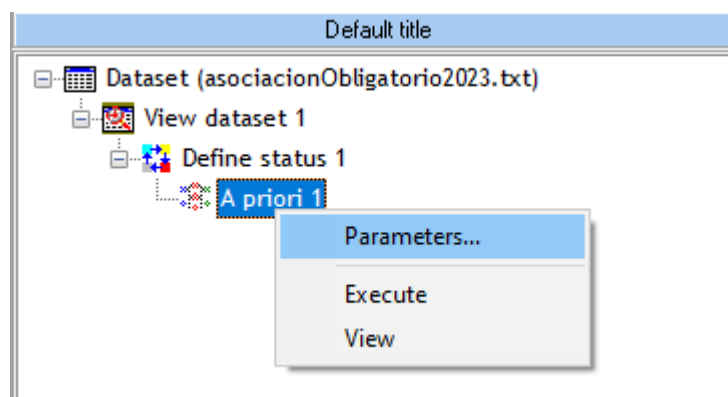
Attribute	Target	Input	Illustrative
Monitor	-	yes	-
Mouse	-	yes	-
Laptop	-	yes	-
Módem	-	yes	-
Router	-	yes	-
Pendrive	-	yes	-
Auriculares	-	yes	-

- 4) Para proceder con las reglas de asociación debemos ir a la sección Components y seleccionar Association. Dentro, seleccionaremos A priori, arrastrándolo debajo del define status creado anteriormente.

El algoritmo A priori se basa en el principio de a priori, que establece que si un conjunto de elementos aparece con frecuencia en un conjunto de transacciones, entonces es probable que exista una asociación significativa entre esos elementos. Es decir, busca patrones frecuentes en los datos y los utiliza para generar reglas de asociación.



- 5) Continuaremos indicando los parámetros definidos anteriormente y que usaremos para generar las reglas de asociación. Para esto, hacemos clic derecho sobre “A priori 1”, y en el menú que se desplegará, seleccionamos la opción *Parameters*



- 6) Después de hacer clic en *Parameters*, se nos abre una ventana en donde podemos indicar los parámetros: *Support* (soporte), *Confidence* (confianza), *Max card itemsets* (número máximo de atributos que se pueden considerar en una regla de asociación) y *Lift*. Allí modificaremos los valores que vienen por defecto, colocando los que se mencionaron anteriormente. Una vez hecho esto, hacemos clic en *OK* para finalizar la configuración.

- 7) Finalmente, hacemos doble clic en A priori 1.

En este caso se pueden observar 30 reglas de asociación:

RULES

Number of rules : 30					
N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"Mouse =true"	"Módem=true" - "Router=true"	2,65725	23,478	65,854
2	"Módem=true" - "Router=true"	"Mouse =true"	2,65725	23,478	94,737
3	"Router=true"	"Módem=true" - "Mouse =true"	2,52439	23,478	65,854
4	"Módem=true" - "Mouse =true"	"Router=true"	2,52439	23,478	90,000
5	"Módem=true" - "Monitor=true"	"Laptop =true"	2,30896	29,130	94,366
6	"Laptop =true"	"Módem=true" - "Monitor=true"	2,30896	29,130	71,277
7	"Monitor=true"	"Módem=true" - "Laptop =true"	2,14355	29,130	73,626
8	"Módem=true" - "Laptop =true"	"Monitor=true"	2,14355	29,130	84,810
9	"Router=true"	"Mouse =true"	1,94973	24,783	69,512
10	"Mouse =true"	"Router=true"	1,94973	24,783	69,512
11	"Laptop =true"	"Monitor=true"	1,88216	30,435	74,468
12	"Monitor=true"	"Laptop =true"	1,88216	30,435	76,923
13	"Auriculares=true" - "Laptop =true"	"Pendrive=true"	1,81669	23,043	96,364
14	"Auriculares=true" - "Monitor=true"	"Pendrive=true"	1,80984	20,870	96,000
15	"Laptop =true" - "Monitor=true"	"Módem=true"	1,80445	29,130	95,714
16	"Módem=true"	"Laptop =true" - "Monitor=true"	1,80445	29,130	54,918
17	"Mouse =true" - "Router=true"	"Módem=true"	1,78602	23,478	94,737
18	"Laptop =true"	"Módem=true"	1,58441	34,348	84,043
19	"Módem=true"	"Laptop =true"	1,58441	34,348	64,754
20	"Módem=true"	"Monitor=true"	1,47091	30,870	58,197
21	"Monitor=true"	"Módem=true"	1,47091	30,870	78,022
22	"Auriculares=true"	"Pendrive=true"	1,44454	51,304	76,623
23	"Pendrive=true"	"Auriculares=true"	1,44454	51,304	96,721
24	"Módem=true" - "Pendrive=true"	"Auriculares=true"	1,43255	20,435	95,918
25	"Pendrive=true" - "Laptop =true"	"Auriculares=true"	1,41350	23,043	94,643
26	"Pendrive=true" - "Monitor=true"	"Auriculares=true"	1,40565	20,870	94,118
27	"Mouse =true"	"Módem=true"	1,37945	26,087	73,171
28	"Auriculares=true" - "Módem=true"	"Pendrive=true"	1,34252	20,435	71,212
29	"Router=true"	"Módem=true"	1,31048	24,783	69,512
30	"Laptop =true"	"Pendrive=true"	1,12313	24,348	59,574

Regla número 2:

El “true” en el atributo significa que compro, por lo que dicha regla se lee de la siguiente manera: “Si compro modem y router (antecedente) → (entonces) compro mouse (consecuente)”.

Modem y Router → Mouse

Observando los demás datos de las columnas podemos determinar que el soporte es de 23,478%, este porcentaje se refiere a la cantidad del total de tickets emitidos donde se compraron los tres productos (modem, router, mouse). En cuanto a la confianza es del 94,737%, este porcentaje significa que de aquellos que compraron un modem y un router, el 94,737% también compró un mouse.

Regla número 23:

Se lee como: “Si compro pendrive (antecedente) → (entonces) compro auriculares (consecuente)”.

Pendrive → Auriculares

Observando los demás datos de las columnas, podemos apreciar que el soporte es de un 51,304%, es decir, que en poco más de la mitad del total de los tickets emitidos se compraron ambos productos. En cuanto a la confianza es del 96,721%, este porcentaje indica que de la totalidad de los que compraron un pendrive, el 96,721% también compró auriculares.

Relación con los negocios

Habiendo observado las dos reglas descritas, se puede determinar que desde el punto de vista de los negocios, estas reglas podrían ser utilizadas en comercios de electrónica, donde a partir de las mismas podrían optar por colocar los elementos juntos, u ofrecer promociones. Por ejemplo, si asiste un cliente a comprar un pendrive, ofrecerle auriculares en descuento.

Preguntas

- 1) ¿Qué significa que una variable sea continua en Tanagra? ¿Existe otro tipo de variable?
- 2) En caso de que las categorizaciones de los atributos estén invertidas, ¿cómo podemos solucionarlo?
- 3) ¿Cómo trabaja el algoritmo A priori utilizado en el ejercicio?

Respuestas

- 1) En Tanagra la clasificación de una variable no está determinada por los valores en sí, sino por la naturaleza de la variable y cómo se utiliza en el análisis. Una variable continua es el equivalente a una variable cuantitativa (numérica). Como se observó en el ejercicio realizado, esta puede ser de carácter discreta (0's y 1's) y sin embargo ser identificada como continua.
Sí, existe otro tipo de variable que es la discreta, siendo el equivalente a una variable cualitativa (alfanumérica).
- 2) La forma más sencilla para cambiarlo es ingresar al archivo .txt y cambiar las “,” (comas) por “.” puntos.
- 3) El proceso del algoritmo A priori en Tanagra generalmente sigue los siguientes pasos: Determinar el umbral de soporte mínimo: El umbral de soporte es un valor mínimo establecido para determinar qué conjuntos de elementos son considerados frecuentes. Los conjuntos de elementos que aparecen con una frecuencia igual o superior al umbral de soporte mínimo se consideran frecuentes. Generar conjuntos de elementos frecuentes: El algoritmo examina el conjunto de datos y busca conjuntos de elementos que cumplan con el umbral de soporte mínimo. Esto se hace a través de un proceso de combinación y poda de conjuntos de elementos candidatos. Generar reglas de asociación: Una vez que se han identificado los conjuntos de elementos frecuentes, el algoritmo A priori genera reglas de asociación a partir de estos conjuntos. Las reglas de asociación indican las relaciones entre diferentes conjuntos de elementos. Evaluar las reglas de asociación: Las reglas de

asociación generadas se evalúan utilizando métricas como el soporte, la confianza y la medida de interés. Estas métricas ayudan a determinar la relevancia y la fuerza de las reglas de asociación

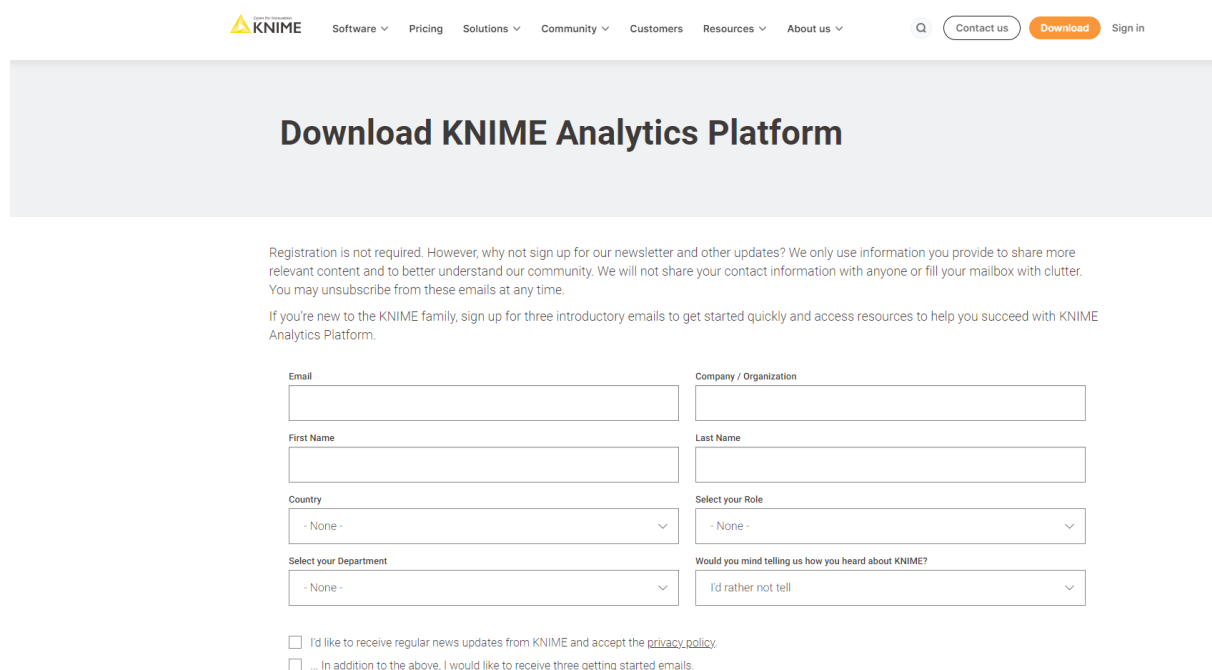
1.2) Reglas de Asociación – (Weka, ProyectosObligatorio2023.txt)

Previo a comenzar a trabajar con Weka, es importante hacer la transformación del archivo .txt al formato .arff. Para esto, utilizaremos otra herramienta llamada **Knime**. Por lo que comenzaremos el tutorial explicando cómo descargarla y cómo generar un archivo .arff a partir de uno .txt.

Destacar que el archivo se descarga de igual manera que se realizó en la parte 1.1), solo que en vez de seleccionar “asociacionObligatorio2023.txt” seleccionamos el que nos compete para el ejercicio.

La descarga de Knime se realizará a través del siguiente enlace:
<https://www.knime.com/downloads>

Se observará la siguiente pantalla:



The screenshot shows the 'Download KNIME Analytics Platform' page. At the top is a navigation bar with links: Software, Pricing, Solutions, Community, Customers, Resources, About us, a search icon, 'Contact us', 'Download', and 'Sign in'. The main heading is 'Download KNIME Analytics Platform'. Below this is a registration form with the following fields and options:

- Email**: Text input field.
- Company / Organization**: Text input field.
- First Name**: Text input field.
- Last Name**: Text input field.
- Country**: Dropdown menu with '- None -' selected.
- Select your Role**: Dropdown menu with '- None -' selected.
- Select your Department**: Dropdown menu with '- None -' selected.
- Would you mind telling us how you heard about KNIME?**: Dropdown menu with 'I'd rather not tell' selected.

Below the form are two checkboxes:

- ☐ I'd like to receive regular news updates from KNIME and accept the [privacy policy](#).
- ☐ ... In addition to the above, I would like to receive three getting started emails.

Una vez completados los datos solicitados y habiendo aceptado los términos y condiciones, se proseguirá seleccionando el archivo correspondiente al sistema operativo de quien realizará el trabajo.

Windows

Microsoft Defender SmartScreen may block download in its attempt to prevent malicious software installations. To solve the problem [click here](#).

KNIME Analytics Platform for Windows (installer) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	Download (464 MB)
KNIME Analytics Platform for Windows (self-extracting archive) <i>The self-extracting archive only creates a folder holding the KNIME installation</i>	Download (466 MB)
KNIME Analytics Platform for Windows (zip archive)	Download (556 MB)

Linux

KNIME Analytics Platform for Linux	Download (567 MB)
------------------------------------	-----------------------------------

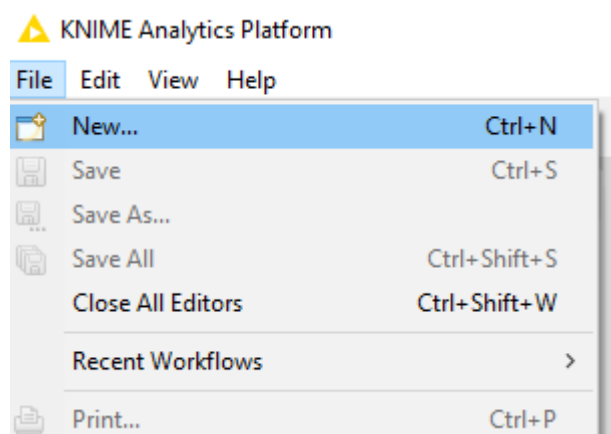
Mac

KNIME Analytics Platform for macOS x86_64 (Intel)	Download (540 MB)
KNIME Analytics Platform for macOS arm64 (Apple silicon)	Download (533 MB)

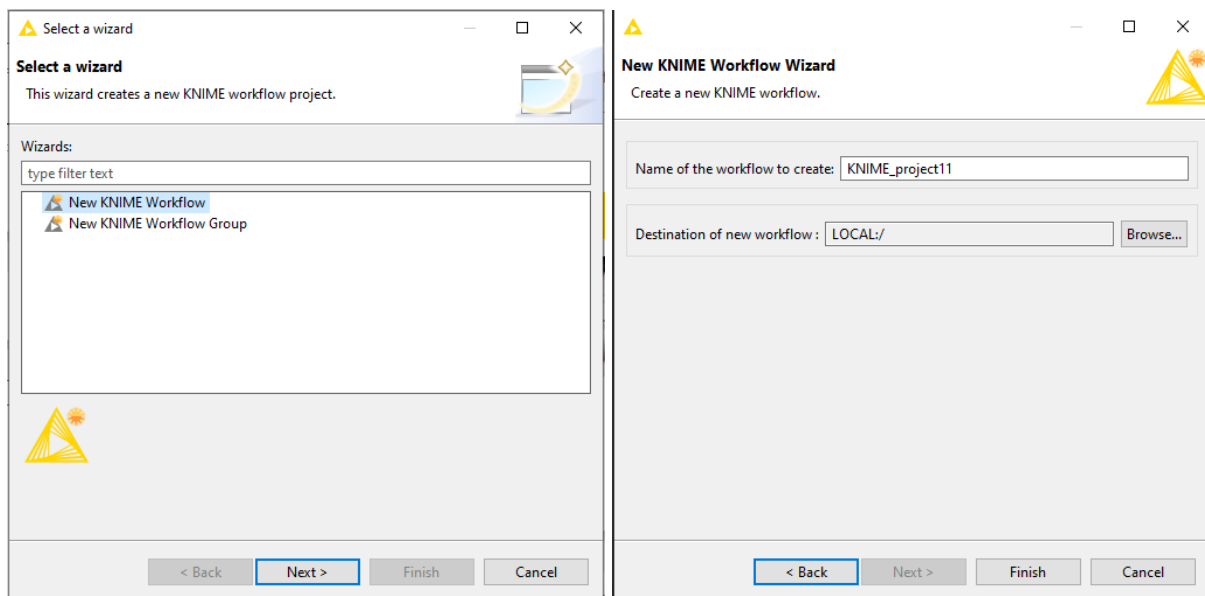
En nuestro caso seleccionaremos la primera opción de Windows y haremos clic en “Download”, descargando el archivo .exe correspondiente. Una vez finalice la descarga lo ejecutaremos.

Para continuar con la instalación simplemente debemos ir presionando siguiente hasta que finalice.

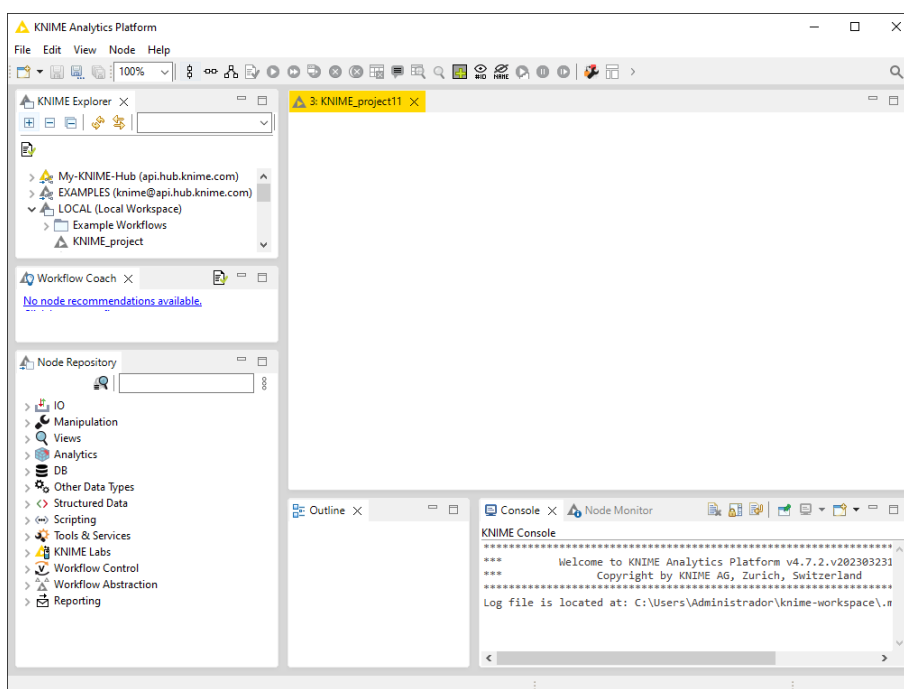
Luego, abriremos Knime. Una vez abierto debemos crear un nuevo flujo de trabajo, por lo que haremos clic en *File* (ubicado en la esquina superior izquierda), y posteriormente hacemos clic en la opción *New...* (indicada en azul).



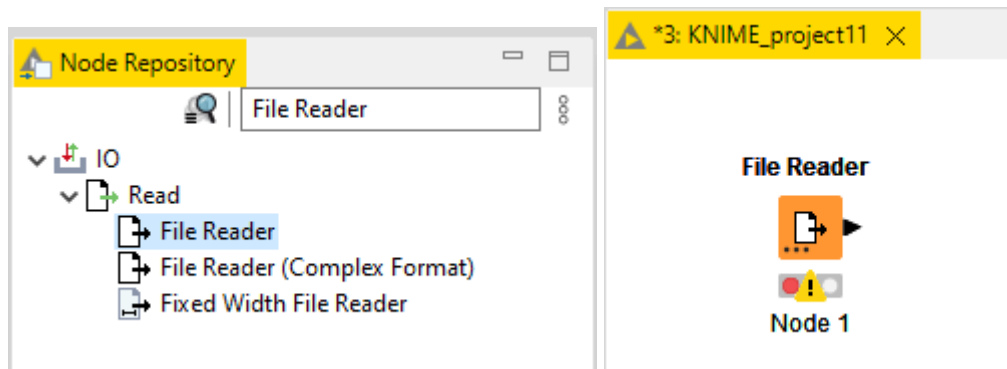
Se desplegará una nueva ventana denominada *Select a wizard*, donde seleccionaremos *New KNIME Workflow*. Una vez seleccionada la opción, marcaremos *Next*. Se abrirá otra ventana donde podemos escribir el nombre del nuevo Workflow a crear o simplemente dejar el por defecto, y cambiar la ubicación donde se guardará el mismo. Una vez realizado lo mencionado, se debe apretar el botón *Finish*.



Obtendremos lo que se observa en la siguiente imagen:



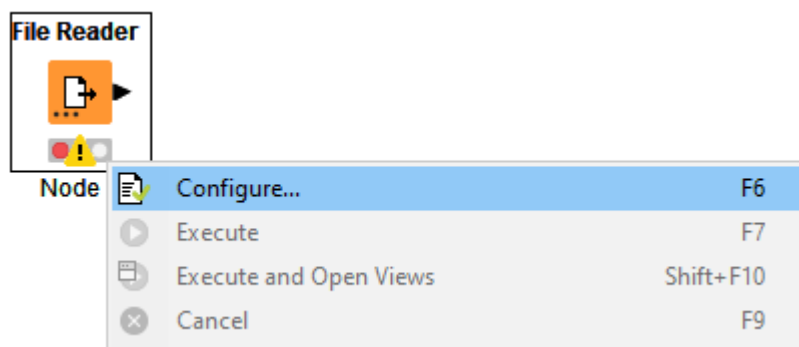
Estando posicionado en dicha pantalla, debemos crear un nodo File Reader. Utilizando el buscador de Node Repository, vamos a buscar File Reader. Una vez encontrado, hacemos doble clic encima de él (indicado en color celeste). Obteniendo como resultado el siguiente nodo:



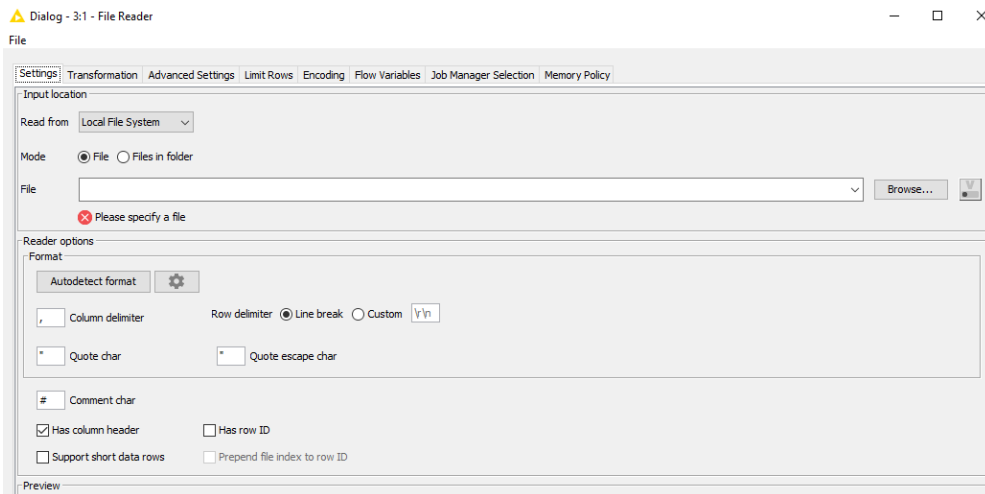
El nodo aparece, en principio, en color rojo. Esto significa que debemos configurarlo para poder ejecutarlo.

Proceso de configuración:

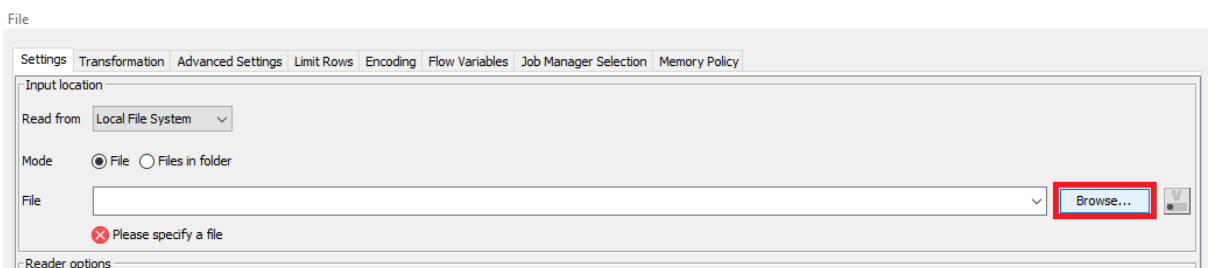
- 1) Hacer clic derecho sobre el nodo *File Reader*, seleccionar la opción *Configure...*



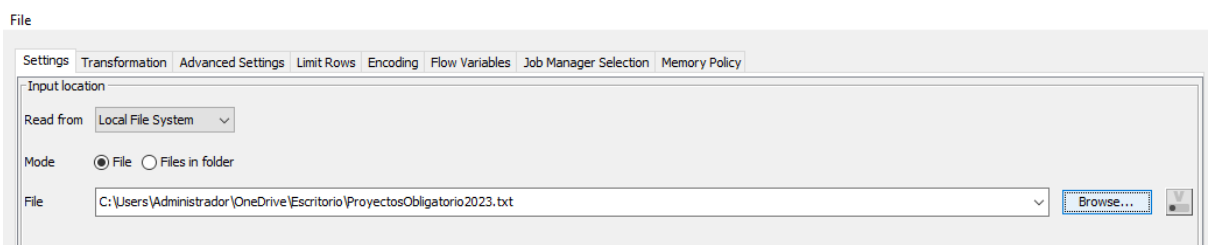
Se abrirá la siguiente ventana:



- 2) Seguidamente, hacemos clic en Browse..., esto se realiza para cargar el archivo deseado, en nuestro caso, ProyectosObligatorio2023.txt. Se nos abrirá un nuevo diálogo donde deberemos dirigirnos a la carpeta donde guardamos el archivo, seleccionarlo y posteriormente presionar *Abrir*.




- 3) Una vez realizado este paso, observaremos que el archivo ya aparece como cargado, como se observa en la imagen adjunta.



Además, nos permitirá ver una vista previa:

Preview

 The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	S	Sexo	Nivel	Estudios	Situacion	Laboral	Otros	Proyectos	P...				
Row0		Masculino	Universitaria	Incompleta	Independiente	S	S	I	N	O	9	7	
Row1		Masculino	Universitaria	Incompleta	Independiente	No	S	I	S	I	2	1	
Row2		Femenino	Universitaria	Incompleta	Independiente	No	S	I	S	I	9	6	
Row3		Masculino	Universitaria	Completa	Independiente	No	S	I	S	I	2	4	
Row4		Femenino	Postgrado	Dependiente	No	S	I	S	I	3	2		
Row5		Femenino	Universitaria	Incompleta	Dependiente	No	S	I	S	I	2	4	
Row6		Femenino	Universitaria	Incompleta	Dependiente	No	S	I	S	I	2	0	
Row7		Femenino	Universitaria	Completa	Independiente	No	N	O	N	O	3	8	
Row8		Masculino	Universitaria	Incompleta	Dependiente	No	N	O	N	O	2	7	
Row9		Masculino	Universitaria	Incompleta	Dependiente	No	N	O	N	O	3	5	
Row10		Masculino	Universitaria	Incompleta	Dependiente	S	I	N	O	S	I	2	8
Row11		Masculino	Universitaria	Completa	Dependiente	No	S	I	N	O	3	2	
Row12		Masculino	Universitaria	Incompleta	Independiente	S	I	S	I	N	O	3	5
Row13		Masculino	Universitaria	Incompleta	Independiente	No	S	I	N	O	3	5	
Row14		Femenino	Universitaria	Incompleta	Independiente	No	S	I	S	I	2	6	
Row15		Masculino	Universitaria	Incompleta	Independiente	No	S	I	S	I	2	6	
Row16		Masculino	Universitaria	Incompleta	Dependiente	No	S	I	S	I	2	7	
Row17		Masculino	Universitaria	Incompleta	Independiente	No	S	I	S	I	3	3	
Row18		Masculino	Universitaria	Completa	Independiente	No	S	I	S	I	3	7	
Row19		Femenino	Postgrado	Independiente	S	I	S	I	3	1			
Row20		Masculino	Universitaria	Incompleta	Dependiente	No	N	O	N	O	2	6	
Row21		Masculino	Universitaria	Incompleta	Dependiente	No	N	O	N	O	2	6	
Row22		Masculino	Universitaria	Incompleta	Independiente	S	I	S	I	N	O	3	2

Al observar la tabla de la siguiente manera, debemos cambiar el delimitador de columnas de manera inmediata. La manera más sencilla de hacerlo es haciendo clic sobre la opción *Autodetect Format* que se presenta en pantalla.

Reader options

Format

Autodetect format

Column delimiter: , Row delimiter: ☒ Line break ☐ Custom

Quote char: " Quote escape char: "

Comment char: #

☒ Has column header ☐ Has row ID

☐ Support short data rows ☐ Prepend file index to row ID

Posteriormente se observarán los datos de la siguiente forma:

Preview

i The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	S Sexo	S NivelEstudios	S Situacio...	S OtrosPr...	S Proyect...	S Financi...	I EdadRe...
Row0	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	97
Row1	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	21
Row2	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	96
Row3	Masculino	Universitaria Completa	Independiente	No	SI	SI	24
Row4	Femenino	Postgrado	Dependiente	No	SI	SI	32
Row5	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	24
Row6	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	20
Row7	Femenino	Universitaria Completa	Independiente	No	NO	NO	38
Row8	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	27
Row9	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	35
Row10	Masculino	Universitaria Incompleta	Dependiente	Si	NO	SI	28
Row11	Masculino	Universitaria Completa	Dependiente	No	SI	NO	32
Row12	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	35
Row13	Masculino	Universitaria Incompleta	Independiente	No	SI	NO	24
Row14	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row15	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row16	Masculino	Universitaria Incompleta	Dependiente	No	SI	SI	27
Row17	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	33
Row18	Masculino	Universitaria Completa	Independiente	No	SI	SI	37
Row19	Femenino	Postgrado	Independiente	Si	SI	SI	31
Row20	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row21	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row22	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	32

Observamos que los datos se identifican como *string* o *integer* y se encuentran correctamente clasificados.

4) Finalmente, hacemos clic en el botón OK para finalizar.

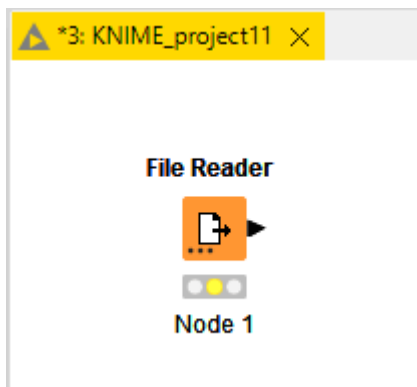
Preview

i The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

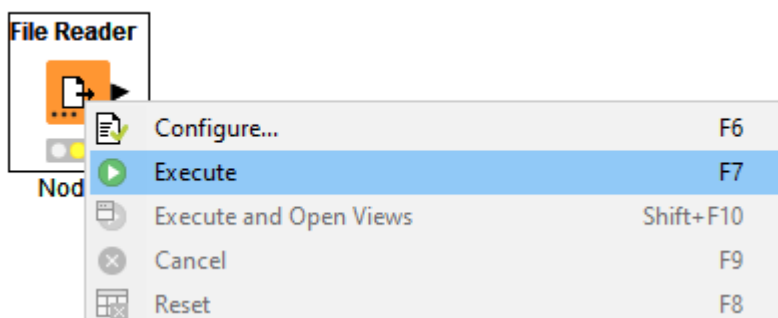
Row ID	S Sexo	S NivelEstudios	S Situacio...	S OtrosPr...	S Proyect...	S Financi...	I EdadRe...
Row0	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	97
Row1	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	21
Row2	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	96
Row3	Masculino	Universitaria Completa	Independiente	No	SI	SI	24
Row4	Femenino	Postgrado	Dependiente	No	SI	SI	32
Row5	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	24
Row6	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	20
Row7	Femenino	Universitaria Completa	Independiente	No	NO	NO	38
Row8	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	27
Row9	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	35
Row10	Masculino	Universitaria Incompleta	Dependiente	Si	NO	SI	28
Row11	Masculino	Universitaria Completa	Dependiente	No	SI	NO	32
Row12	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	35
Row13	Masculino	Universitaria Incompleta	Independiente	No	SI	NO	24
Row14	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row15	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row16	Masculino	Universitaria Incompleta	Dependiente	No	SI	SI	27
Row17	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	33
Row18	Masculino	Universitaria Completa	Independiente	No	SI	SI	37
Row19	Femenino	Postgrado	Independiente	Si	SI	SI	31
Row20	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row21	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row22	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	32

OK Apply Cancel ?

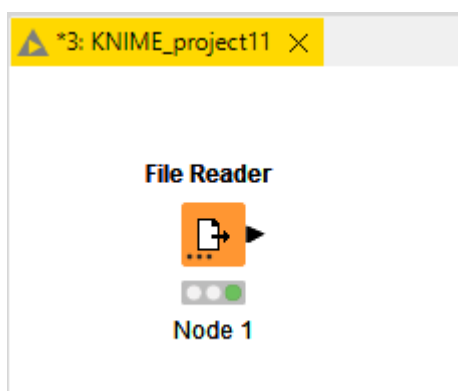
Luego de finalizado, el nodo pasa a ser de esta forma:



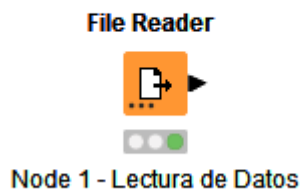
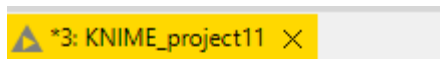
Se observa que ahora se encuentra en color amarillo. Por lo que ahora debemos ejecutar el nodo. Para ejecutar el nodo debemos hacer clic derecho sobre el mismo y luego hacer clic sobre *Execute*.



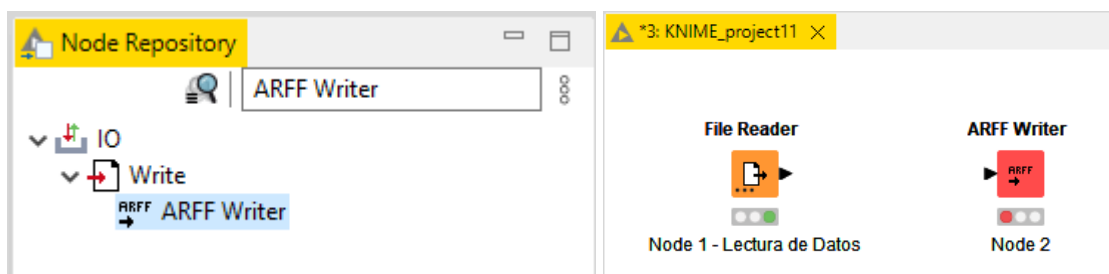
Una vez ejecutado podremos observar que si la ejecución fue exitosa pasará a tener color verde.



Si se desea, con el objetivo de entender de mejor forma la función del nodo, podemos renombrarlo. Para esto, debemos hacer doble clic donde dice “Node 1”, y escribimos la tarea que realiza.

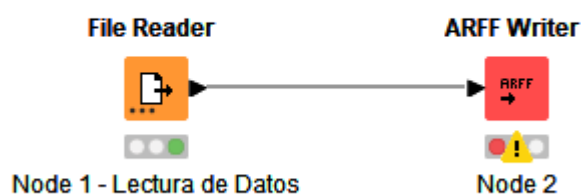


Proseguiremos creando un nodo *ARFF Writer*. Igual que anteriormente, nos dirigimos al buscador de *Node Repository* y escribimos “ARFF Writer”. Una vez hallado, hacemos doble clic sobre el mismo.



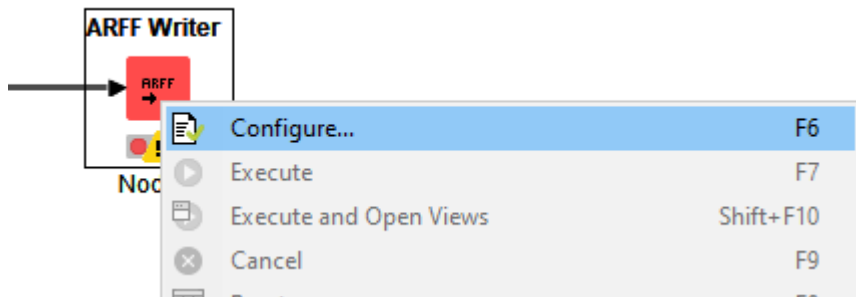
Es importante que ambos nodos estén vinculados. Para vincularlos simplemente debemos hacer clic en el triángulo que aparece en *File Reader* y arrastrar la línea hasta el triángulo que presenta *ARFF Writer*

Observándose de la siguiente manera:

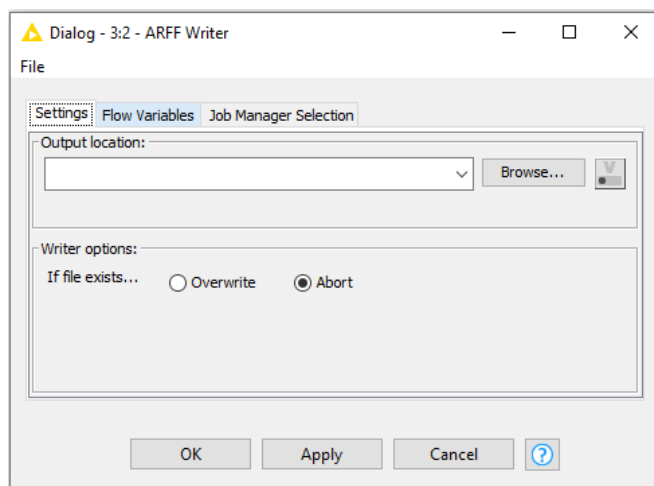


Se puede apreciar que se presenta con un indicador rojo igual que el File Reader cuando comenzamos el trabajo. Esto significa que también debemos configurar el nodo ARFF Writer. Este proceso será de la siguiente manera:

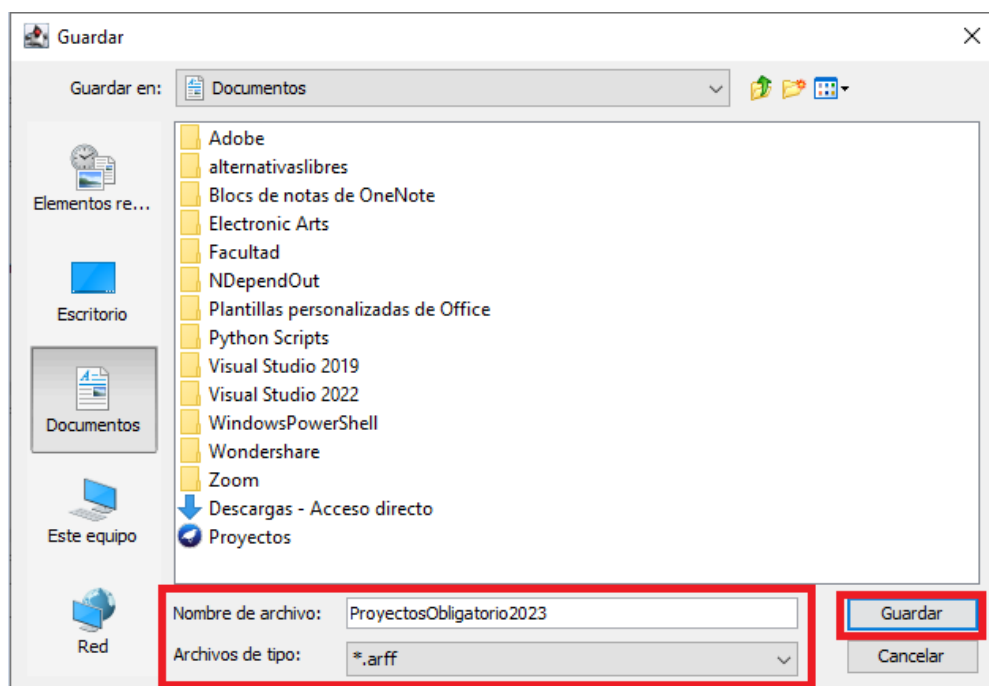
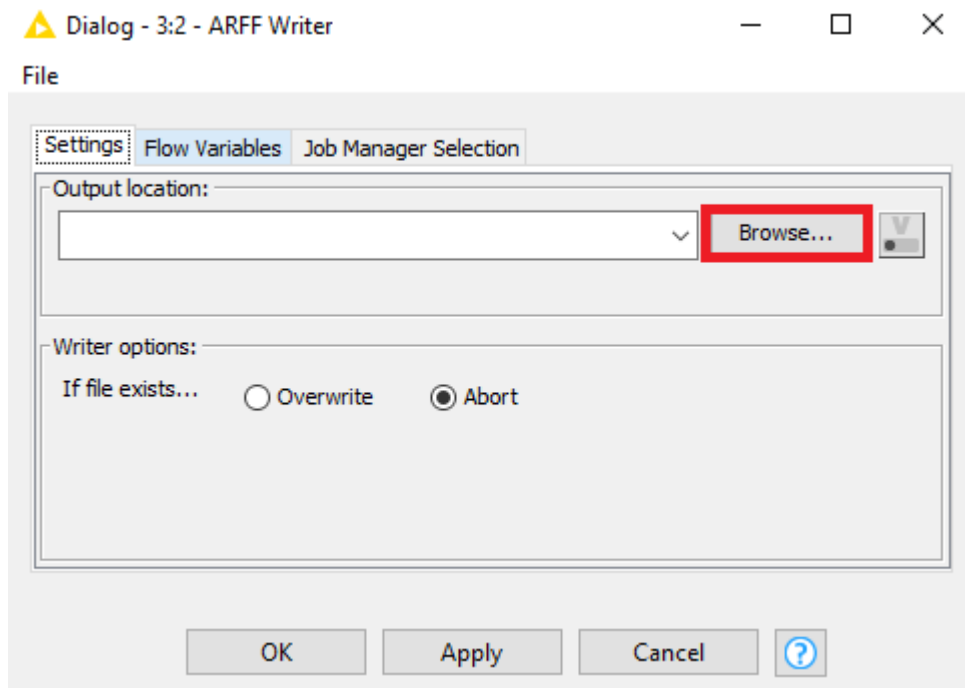
- 1) Hacer clic derecho sobre el nodo *ARFF Writer* y seleccionar la opción *Configure...*



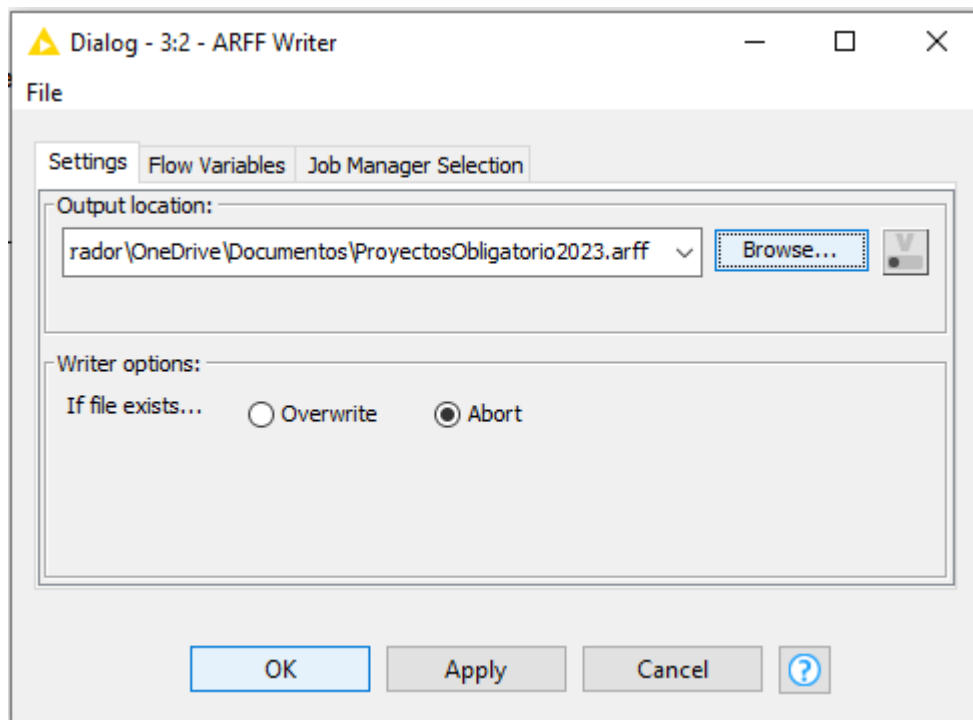
Se abrirá la siguiente ventana:



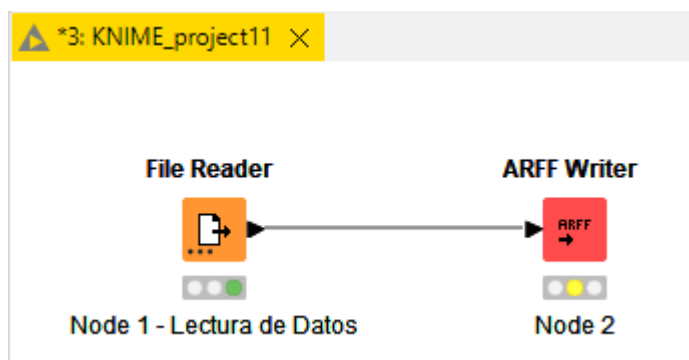
- 2) Hacemos clic en *Browse...* y seleccionaremos la carpeta donde deseamos guardar el archivo, así como también su nombre. Observar que presenta una terminación *.arff*. Una vez lo hayamos ubicado y nombrado, hacemos clic en *Guardar*.



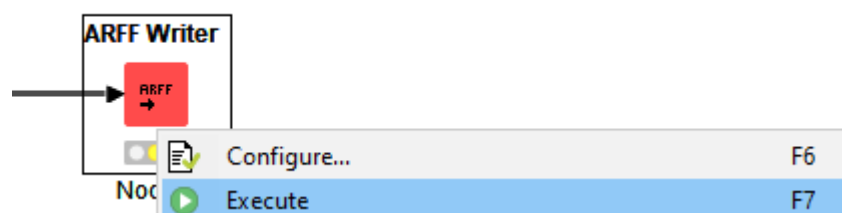
3) Por último, hacemos clic en **OK**



Obteniendo como resultado lo siguiente:



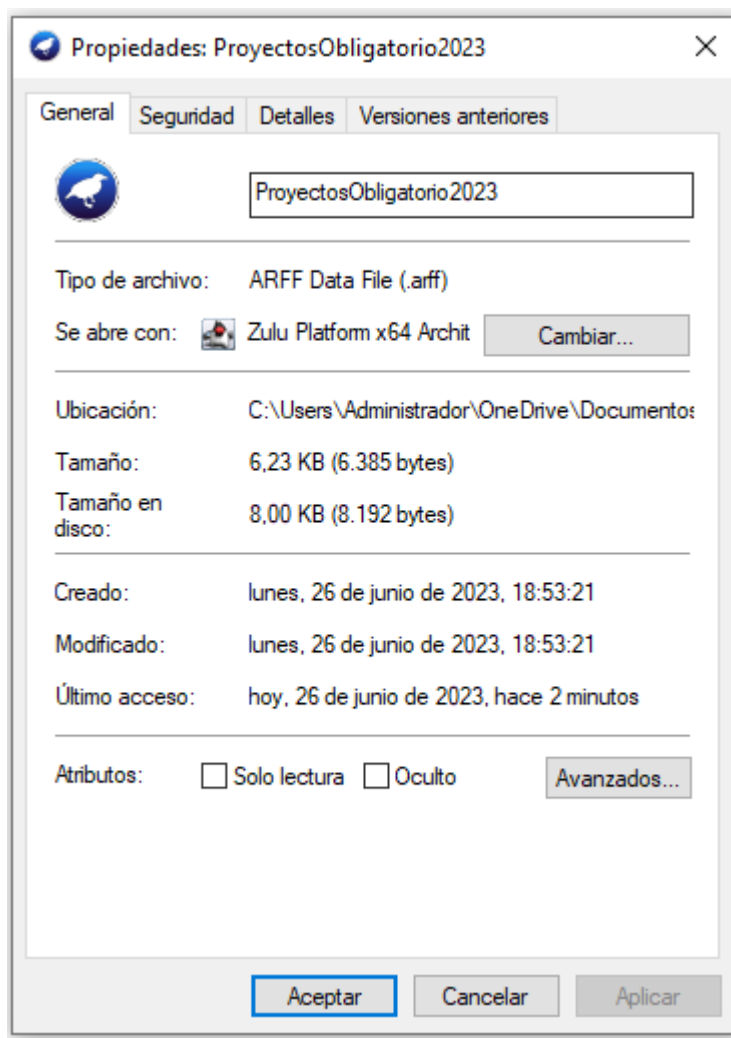
Dado que ya se encuentra el nodo configurado, debemos ejecutarlo. Esto se realiza haciendo clic derecho sobre el nodo *ARFF Writer* y posteriormente clic en *Execute*.



Una vez ejecutado, observaremos el flujo de la siguiente manera:



Solo resta verificar que el archivo fue creado, para esto, nos dirigiremos a la carpeta que seleccionamos como ubicación y lo buscaremos.



Ya tenemos el archivo .arff que necesitaremos para continuar trabajando con Weka, por lo que a continuación describiremos cómo instalar la herramienta que utilizaremos para obtener las reglas de asociación.

Para descargar el programa Weka debemos dirigirnos al siguiente link:

https://waikato.github.io/weka-wiki/downloading_weka/

Se observará la siguiente pantalla:

Downloading and installing Weka

There are two versions of Weka: Weka 3.8 is the latest stable version and Weka 3.9 is the development version. New releases of these two versions are normally made once or twice a year. For the bleeding edge, it is also possible to download nightly snapshots of these two versions.

The stable version receives only bug fixes and feature upgrades that do not break compatibility with its earlier releases, while the development version may receive new features that break compatibility with its earlier releases.

Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

Snapshots

Every night, a snapshot of the Git repository with the Weka source code is taken, compiled, and put together in ZIP files. This happens for both the development branch of the software and the stable branch. Those who want the latest bug fixes before the next official release is made can download these [snapshots](#).

Stable version

Weka 3.8 is the latest stable version of Weka. This branch of Weka only receives bug fixes and upgrades that do not break compatibility with earlier 3.8 releases, although major new features may become available in packages. There are different options for downloading and installing it on your system:

WINDOWS

- Click [here](#) to download a self-extracting executable for 64-bit Windows that includes Azul's 64-bit OpenJDK Java VM 17 (weka-3-8-6-azul-zulu-windows.exe; 133.2 MB)

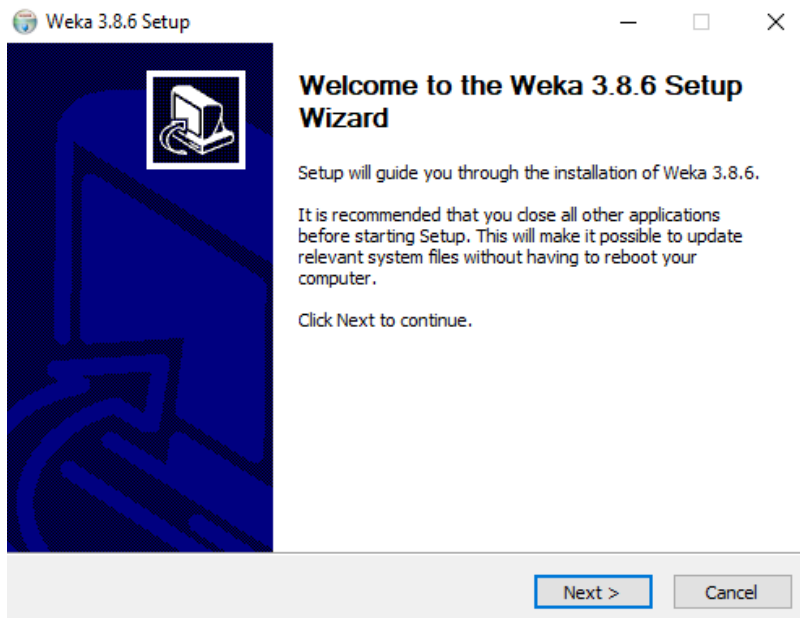
This executable will install Weka in your Program Menu. Launching via the Program Menu or shortcuts will automatically use the included JVM to run Weka.

Table of contents

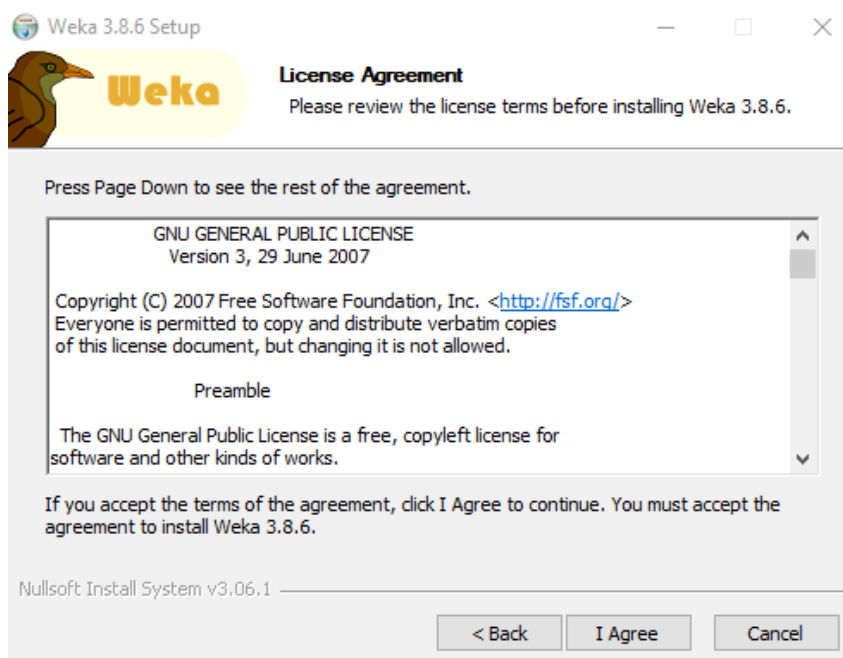
- Snapshots
- Stable version
 - Windows
 - Mac OS - Intel processors
 - Mac OS - ARM processors
 - Linux
 - Other platforms
- Developer version
 - Windows
 - Mac OS - Intel processors
 - Mac OS - ARM processors
 - Linux
 - Other platforms
- Old versions
- Upgrading from Weka 3.7

Una vez nos encontremos en la página, en nuestro caso, nos dirigimos a la sección de Windows (esto dependerá del sistema operativo del usuario). En el recuadro rojo se puede observar que aparece “here” en color azul, haciendo clic sobre dicha opción comenzará la descarga de manera automática.

Continuaremos abriendo el archivo ejecutable, y se observará la siguiente ventana:



Debemos hacer clic en Next, y se presentará la License Agreement correspondiente. Observándose lo siguiente:



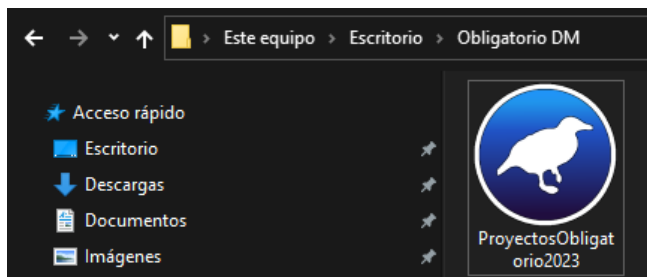
A partir de este punto presionaremos *I Agree*, posteriormente *Next* en las próximas ventanas hasta que aparezca la opción *Install*. Una vez nos encontremos en dicha ventana, seleccionamos instalar y esperamos a que finalice. Al finalizar, hacemos clic en *Next* y luego en *Finish*. Ya tendremos Weka instalado para trabajar.

Luego, debemos abrir el archivo. En este punto tenemos dos opciones:

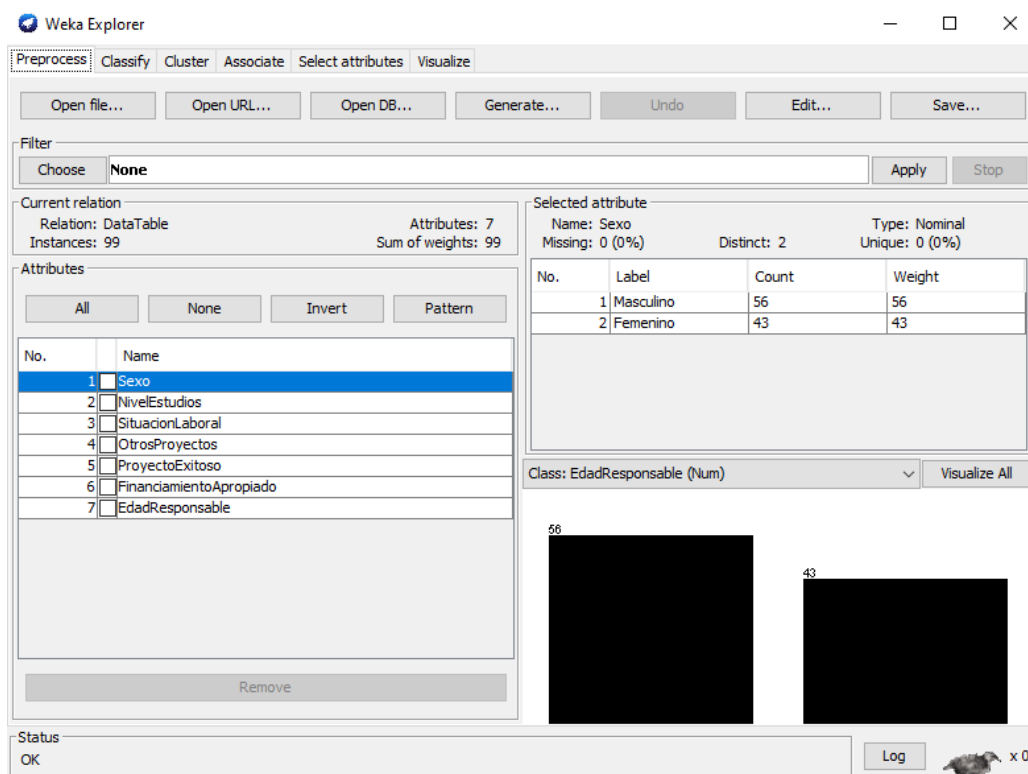
- i) Abrir la aplicación y posteriormente el archivo
- ii) Buscar el archivo en nuestra biblioteca y abrirlo con Weka.

En este tutorial realizaremos la segunda opción ya que facilita el ejercicio, permitiéndole al usuario ahorrar una serie de pasos.

Por lo que nos dirigiremos a donde tenemos guardado el archivo

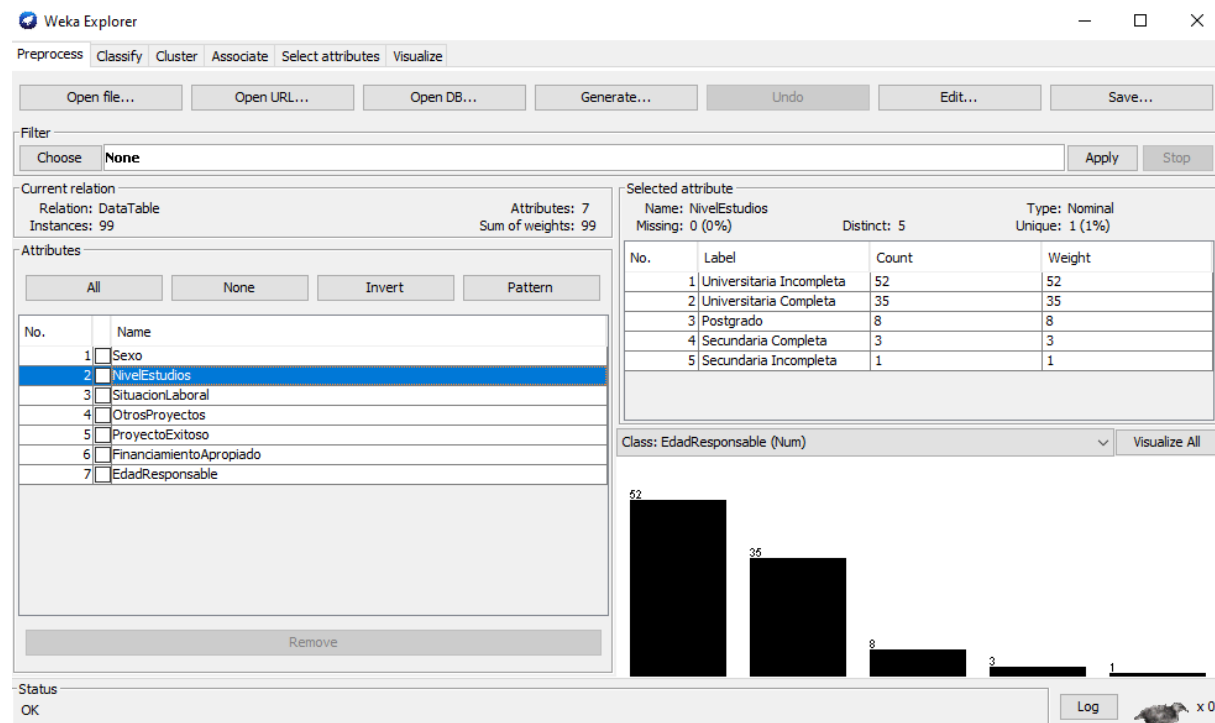


Como se puede observar, ya está identificado con el símbolo de Weka debido a su terminación .arff. Esto nos permite simplemente abrirlo haciendo doble clic sobre el mismo. Se abrirá así la siguiente ventana:



Si deseamos ir conociendo los datos que se tienen, podemos ir marcando de a un atributo y observando en la tabla de la derecha. Por ejemplo, en la imagen anterior está marcado el Sexo, pudiéndose determinar que en los datos hay 56 personas del sexo masculino y 43 del sexo femenino.

Si hacemos clic sobre NivelEstudios, observaremos lo siguiente:

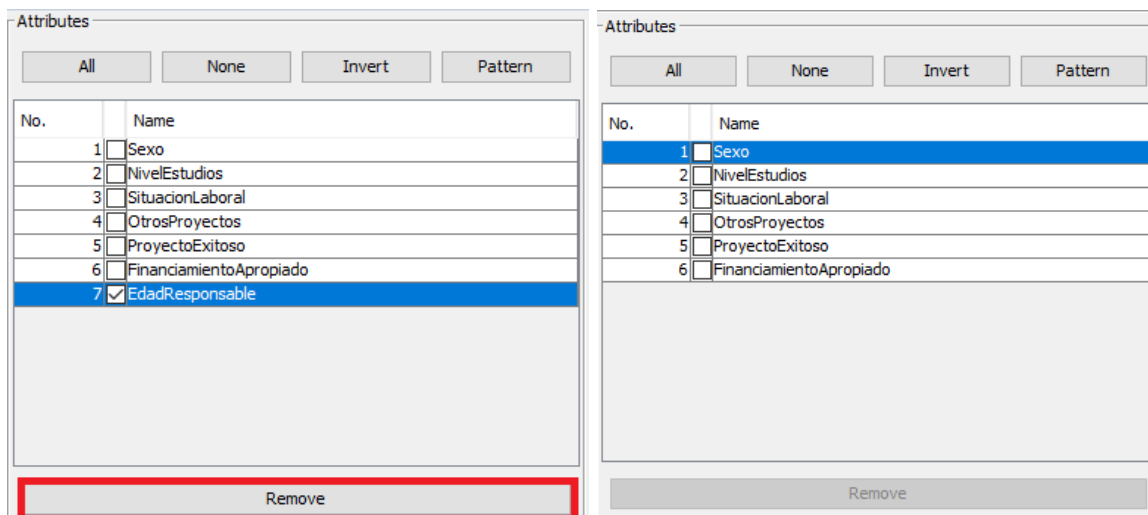


Donde logramos apreciar que 52 personas tienen un nivel universitario incompleto, otras 35 personas universidad completa, 8 tienen postgrado, 3 secundaria completa y 1 secundaria incompleta.

Este proceso se puede realizar para todos los atributos.

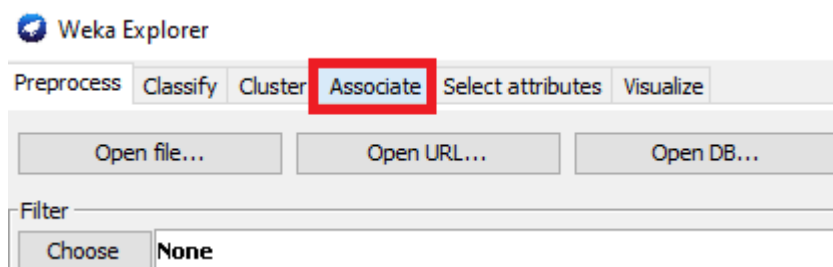
Si observamos detenidamente en la información que nos brinda Weka, podemos determinar que identifica a todos los atributos como tipo Nominal, a excepción del último, *EdadResponsable* que lo identifica como Numeric. Es por esto, que previo a buscar las distintas reglas de asociación, procederemos a eliminar el atributo *EdadResponsable*.

Para eliminar el atributo debemos seleccionarlo haciendo clic sobre el cuadrado que se encuentra a su izquierda, y hacer clic en el botón *Remove*, como se observa en las imágenes a continuación:

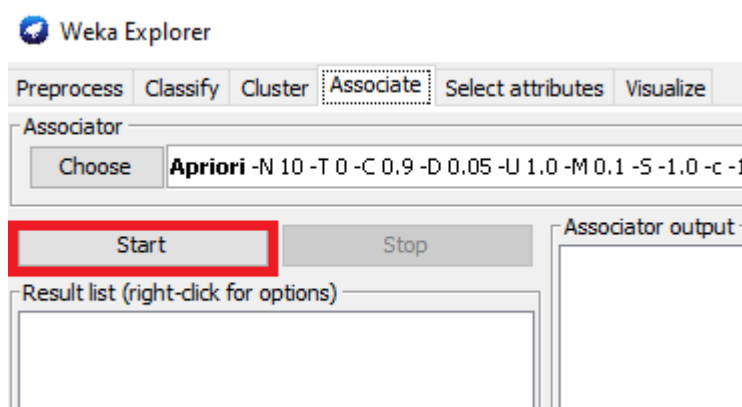


Para generar las reglas de asociación, utilizaremos el algoritmo A priori, siguiendo los pasos que se indican a continuación:

- 1) Dirigirse a la opción ubicada en la parte superior izquierda llamada *Associate* (opción indicada en rojo)



- 2) Luego de hacer clic, se nos abrirá la siguiente pantalla, donde haremos clic en el botón *Start* (el cual se encuentra remarcado en rojo). Esto se debe a que no modificaremos los parámetros originales y por defecto ya se encuentra seleccionado el algoritmo A priori.



Luego de finalizados los pasos de la parte anterior, podemos observar como resultado la generación de reglas de asociación, las cuales se presentan a continuación:



Al final se pueden observar con más detalle las reglas de asociación obtenidas:

```
Best rules found:

1. Sexo=Femenino ProyectoExitoso=SI 22 ==> FinanciamientoApropiado=SI 22 <conf:(1)> lift:(1.55) lev:(0.08) [7] conv:(7.78)
2. SituacionLaboral=Independiente OtrosProyectos=No FinanciamientoApropiado=SI 22 ==> ProyectoExitoso=SI 22 <conf:(1)> lift:(1.74) lev:(0.09) [9] conv:(9.33)
3. Sexo=Femenino NivelEstudios=Universitaria Incompleta 21 ==> FinanciamientoApropiado=SI 21 <conf:(1)> lift:(1.55) lev:(0.07) [7] conv:(7.42)
4. Sexo=Masculino SituacionLaboral=Independiente FinanciamientoApropiado=SI 20 ==> ProyectoExitoso=SI 20 <conf:(1)> lift:(1.74) lev:(0.09) [8] conv:(8.48)
5. NivelEstudios=Universitaria Incompleta SituacionLaboral=Independiente 21 ==> ProyectoExitoso=SI 20 <conf:(0.95)> lift:(1.65) lev:(0.08) [7] conv:(4.45)
6. SituacionLaboral=Dependiente FinanciamientoApropiado=NO 21 ==> OtrosProyectos=No 20 <conf:(0.95)> lift:(1.35) lev:(0.05) [5] conv:(3.08)
7. SituacionLaboral=Independiente FinanciamientoApropiado=SI 32 ==> ProyectoExitoso=SI 30 <conf:(0.94)> lift:(1.63) lev:(0.12) [11] conv:(4.53)
8. Sexo=Masculino SituacionLaboral=Independiente 30 ==> ProyectoExitoso=SI 28 <conf:(0.93)> lift:(1.62) lev:(0.11) [10] conv:(4.24)
9. NivelEstudios=Universitaria Incompleta OtrosProyectos=No ProyectoExitoso=SI 25 ==> FinanciamientoApropiado=SI 23 <conf:(0.92)> lift:(1.42) lev:(0.07) [6] conv:(2.95)
10. SituacionLaboral=Independiente OtrosProyectos=No ProyectoExitoso=SI 24 ==> FinanciamientoApropiado=SI 22 <conf:(0.92)> lift:(1.42) lev:(0.07) [6] conv:(2.83)
```

Regla número 1

Se puede observar que si el sexo es femenino y el proyecto exitoso (antecedente) → (entonces) tuvo el financiamiento apropiado (consecuente).

Sexo = Femenino y ProyectoExitoso = SI → FinanciamientoApropiado = SI

Esto presenta una confianza de 1, ya que se puede observar que son 22 casos que cumplen el antecedente y de ellos, los 22 cumplen el consecuente.

Regla número 8

Se lee como: “Si el sexo es masculino y su situación laboral es independiente (antecedente) → (entonces) el proyecto es exitoso (consecuente).

Sexo = Masculino y SituacionLaboral = Independiente → ProyectoExitoso = SI

Esta regla presenta una confianza de 0,93. Dado que hay 30 casos que cumplen el antecedente, y de ellos, 28 cumplen el consecuente.

Relación con los negocios

Estas reglas de asociación podrían ser utilizadas en negocios para identificar patrones y relaciones entre diferentes variables o características de los clientes, proyectos o situaciones laborales. Algunas posibles podrían ser:

Estrategias de financiamiento: Si una empresa proporciona financiamiento a sus clientes, puede utilizar estas reglas de asociación para determinar qué características están asociadas con un financiamiento exitoso. Por ejemplo, la regla 1 indica que, si el sexo es femenino y el proyecto es exitoso, es más probable que se apruebe un financiamiento apropiado.

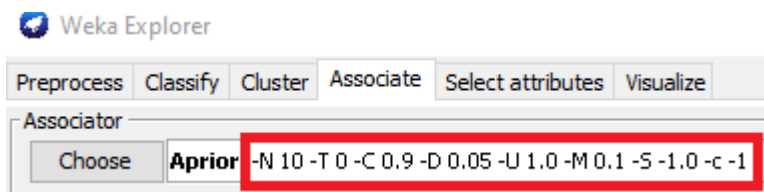
Segmentación de clientes: Se pueden utilizar para segmentar a los clientes en grupos o categorías más específicas en función de sus características. Por ejemplo, la regla 3 sugiere que las mujeres con educación universitaria incompleta tienen una alta probabilidad de obtener un financiamiento apropiado. Con esto se podrían dirigir campañas de marketing específicas para este grupo.

Preguntas

- 1) ¿Cómo podemos modificar los parámetros impuestos para el algoritmo Apriori?
- 2) ¿Por qué utilizamos Association en este ejercicio?
- 3) ¿Cómo podemos guardar los resultados en un archivo de texto?

Respuestas

- 1) Para modificar los parámetros del algoritmo simplemente debemos hacer clic sobre los mismos (ver cuadrante rojo)



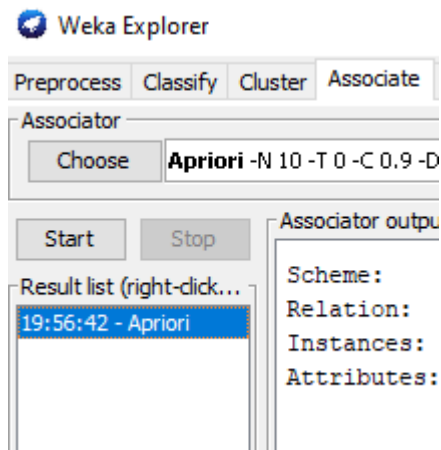
Esto desplegará una nueva ventana de configuración. En ella, podremos modificar algunos parámetros tales como:

numRules: número de reglas a encontrar

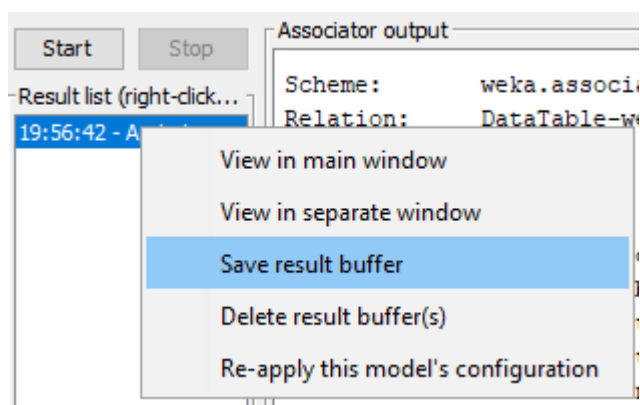
minMetric: nivel mínimo de la métrica seleccionada.

lowerBoundMinSupport: nivel mínimo de soporte.

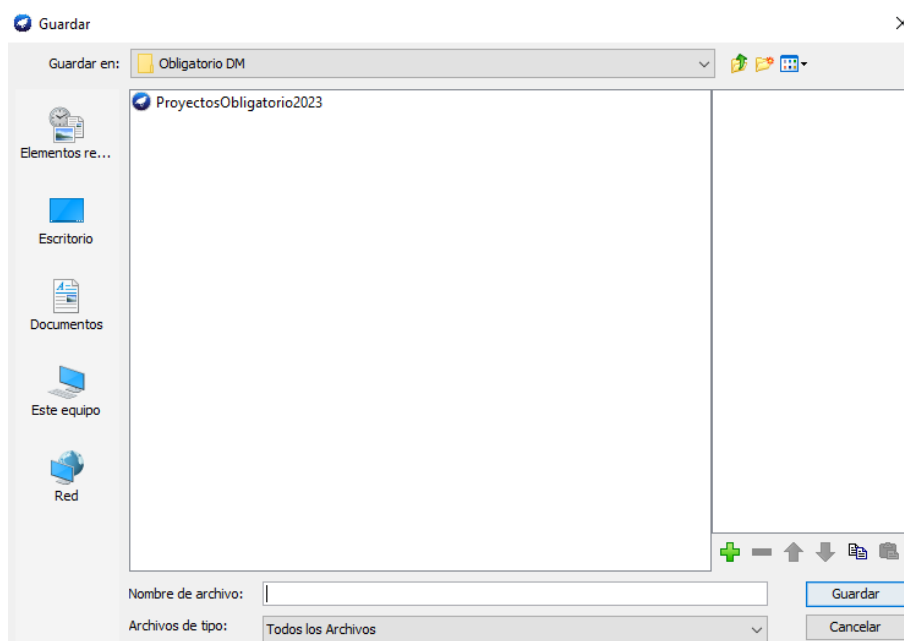
- 2) Utilizamos *Associate* porque esta ventana nos permite aplicar métodos orientados a buscar asociaciones entre datos. Como el objetivo es generar reglas de asociación, estamos buscando un método que nos permita encontrar asociaciones entre datos.
- 3) Si deseamos guardar los datos en un archivo de texto, una vez finalizado el ejercicio, podemos observar que en la parte izquierda se encuentra un cuadrante denominado "*Result list*" que indica la hora en que se realizó la ejecución.



Si hacemos clic derecho sobre la ejecución que deseamos exportar, aparecerán múltiples opciones. Debemos seleccionar aquella que dice “Save result buffer”



En la nueva ventana le indicaremos el nombre deseado y la ubicación, y haremos clic en *Guardar*.



Como resultado obtendremos un archivo con la información obtenida en la ejecución, en este caso, del algoritmo A priori.

```
Proyectos: Bloc de notas
Archivo Edición Formato Ver Ayuda
=== Run information ===

Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:    DataTable-weka.filters.unsupervised.attribute.Remove-R7
Instances:   99
Attributes:  6
             Sexo
             NivelEstudios
             SituacionLaboral
             OtrosProyectos
             ProyectoExitoso
             FinanciamientoApropiado
=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.2 (20 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 36
Size of set of large itemsets L(3): 25
Size of set of large itemsets L(4): 4

Best rules found:

1. Sexo=Femenino ProyectoExitoso=SI 22 ==> FinanciamientoApropiado=SI 22 <conf:(1)> lift:(1.55) lev:(0.08) [7] conv:(7.78)
2. SituacionLaboral=Independiente OtrosProyectos=No FinanciamientoApropiado=SI 22 ==> ProyectoExitoso=SI 22 <conf:(1)> lift:(1.74) lev:(0.09) [9] conv:(9.33)
3. Sexo=Femenino NivelEstudios=Universitaria Incompleta 21 ==> FinanciamientoApropiado=SI 21 <conf:(1)> lift:(1.55) lev:(0.07) [7] conv:(7.42)
4. Sexo=Masculino SituacionLaboral=Independiente FinanciamientoApropiado=SI 20 ==> ProyectoExitoso=SI 20 <conf:(1)> lift:(1.74) lev:(0.09) [8] conv:(8.48)
5. NivelEstudios=Universitaria Incompleta SituacionLaboral=Independiente 21 ==> ProyectoExitoso=SI 20 <conf:(0.95)> lift:(1.65) lev:(0.08) [7] conv:(4.45)
6. SituacionLaboral=Dependiente FinanciamientoApropiado=NO 21 ==> OtrosProyectos=No 20 <conf:(0.95)> lift:(1.35) lev:(0.05) [5] conv:(3.08)
7. SituacionLaboral=Independiente FinanciamientoApropiado=SI 32 ==> ProyectoExitoso=SI 30 <conf:(0.94)> lift:(1.63) lev:(0.12) [11] conv:(4.53)
8. Sexo=Masculino SituacionLaboral=Independiente 30 ==> ProyectoExitoso=SI 28 <conf:(0.93)> lift:(1.62) lev:(0.11) [10] conv:(4.24)
9. NivelEstudios=Universitaria Incompleta OtrosProyectos=No ProyectoExitoso=SI 25 ==> FinanciamientoApropiado=SI 23 <conf:(0.92)> lift:(1.42) lev:(0.07) [6] conv:(2.95)
10. SituacionLaboral=Independiente OtrosProyectos=No ProyectoExitoso=SI 24 ==> FinanciamientoApropiado=SI 22 <conf:(0.92)> lift:(1.42) lev:(0.07) [6] conv:(2.83)
```

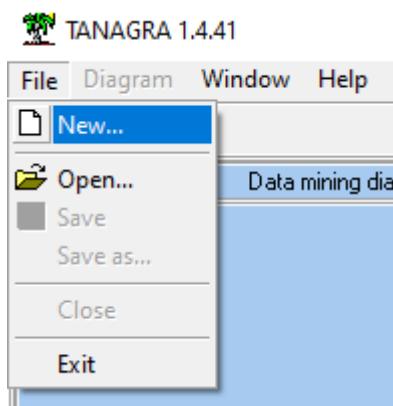

Ejercicio 2)

Construya un árbol de decisión para cada pareja (herramienta, dataset) utilizando dos herramientas y dos datasets. (un dataset para cada herramienta).

2.1) Árboles de decisión – (Tanagra, empresasObligatorio2023.txt)

Asumimos ya descargado el programa Tanagra, debido a que se utilizó en el ejercicio anterior. El archivo empresasObligatorio2023.txt se descarga de igual manera que se mencionó al comienzo; encontrándose en la misma carpeta.

Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo DataSet (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo empresasObligatorio2023.txt en la ubicación donde fue guardado.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :

OK Cancel Help

Una vez seleccionado el archivo se nos cargará en el campo DataSet la ruta de este.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :
C:\Users\Administrador\Downloads\empresasObligatorio2023.txt

OK Cancel Help

Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:

TANAGRA 1.4.41 - [Dataset (empresasObligatorio2023.txt)]

File Diagram Component Window Help

Default title

Dataset (empresasObligatorio2023.txt)

Dataset (empresasObligatorio2023.txt)

Parameters

Database : C:\Users\Administrador\Downloads\empresasObligatorio2023.txt

Results

Download information

DataSource processing
 Computation time 0 ms
 Allocated memory 9 KB

Dataset description

6 attribute(s)
 100 example(s)

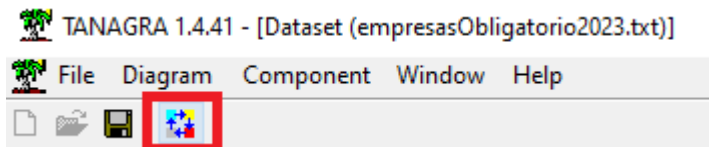
Attribute	Category	Informations
id	Continue	-
FlujodeCaja/DeudaTotal	Continue	-
IngresosIeto/ActivoTotal	Continue	-
ActivoCorriente/PasivoCorriente	Continue	-
ActivoCorriente/VentasNetas	Continue	-
Quiebra	Discrete	2 values

Computation time : 0 ms.
 Created at 26/6/2023 22:26:47

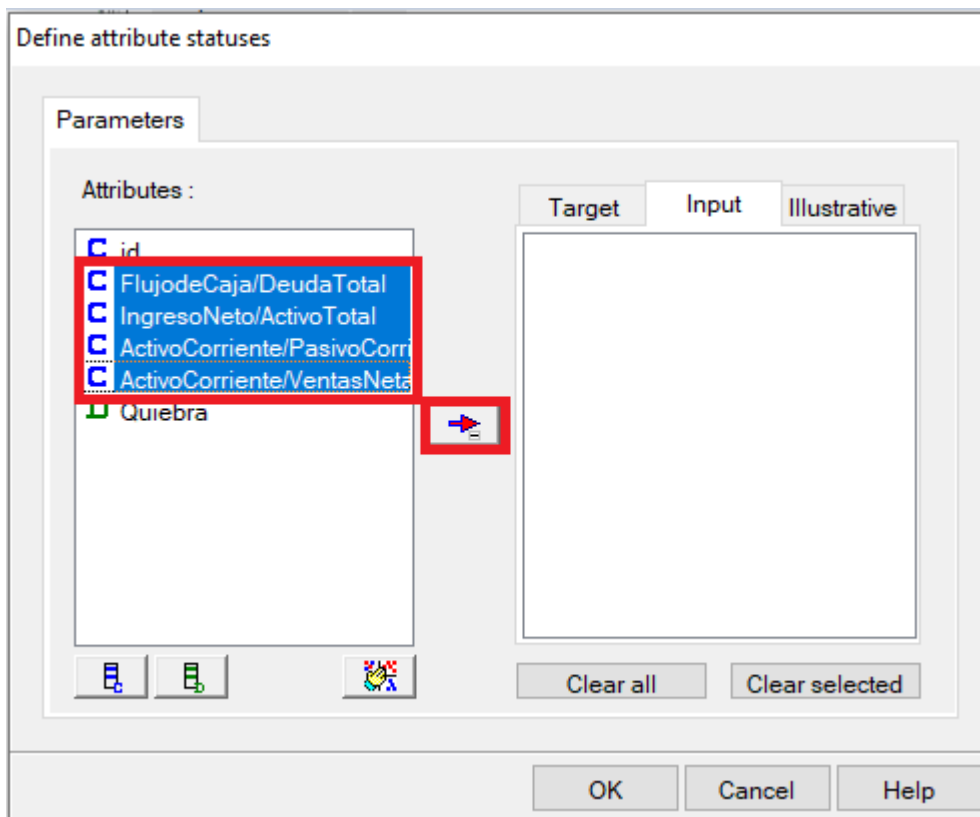
Recordamos que si se desean observar los datos de manera más clara podemos realizar un Data Visualization; en este caso omitiremos dicho paso.

Para determinar el árbol de decisión, procederemos a seguir los siguientes pasos:

- 1) Haremos un *define status* con el objetivo de indicar con qué atributos trabajaremos. Para esto, debemos hacer clic en el ícono marcado en rojo en la siguiente imagen:

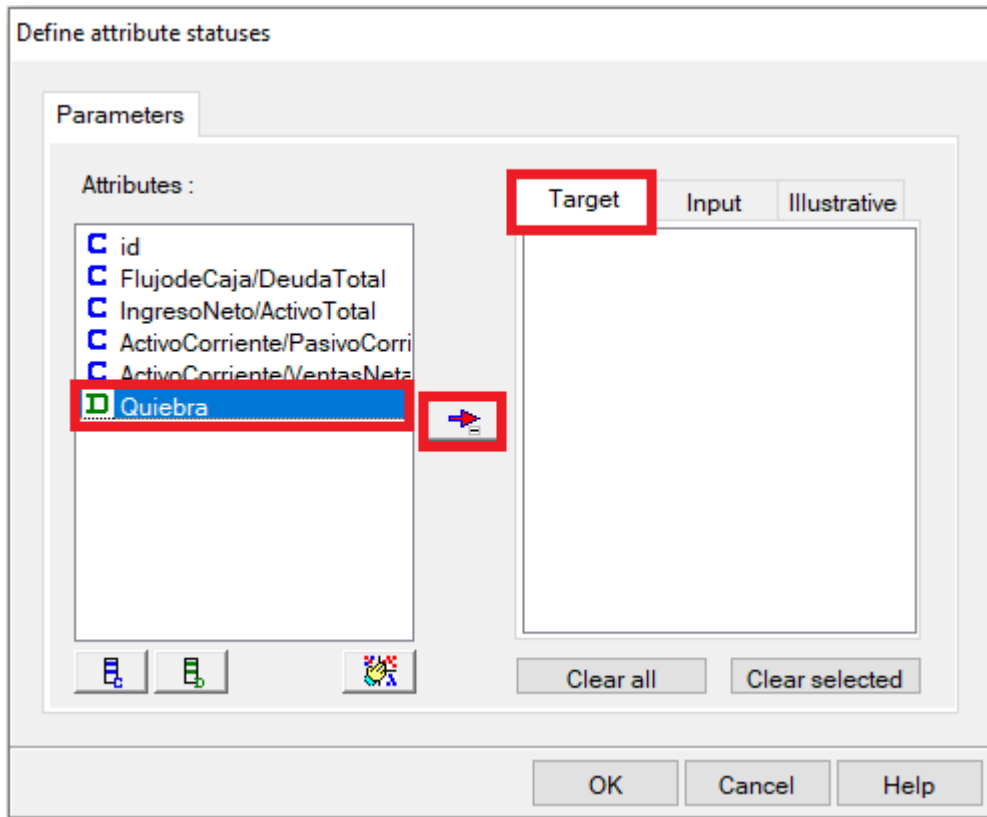


- 2) Se abrirá una ventana denominada *Define attribute statuses*, donde debemos seleccionar todos los atributos que son ratios financieros y presionar la flecha indicada para colocarlas en el *input*. Observar que en el cuadrante de la derecha esté indicada la opción *input*.



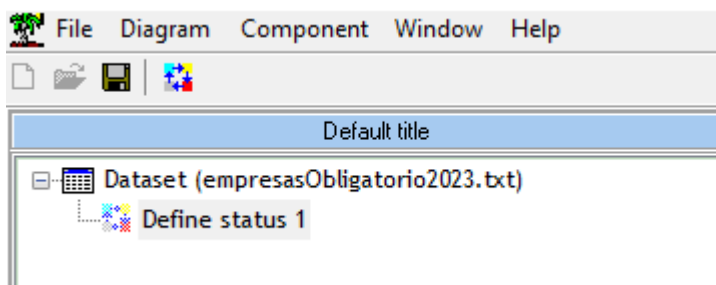
- 3) Una vez los atributos marcados se encuentren del lado derecho, debemos indicar el atributo *Target*, es decir, nuestra variable objetivo. En este caso será *Quiebra*. Por lo que, siguiendo el procedimiento anterior, seleccionaremos

dicho atributo y lo desplazaremos hacia la derecha. Previamente debemos seleccionar Target, ya de que no hacerlo, pondríamos al atributo en Input.



El atributo id queda por fuera ya que es únicamente un identificador y no aporta información.

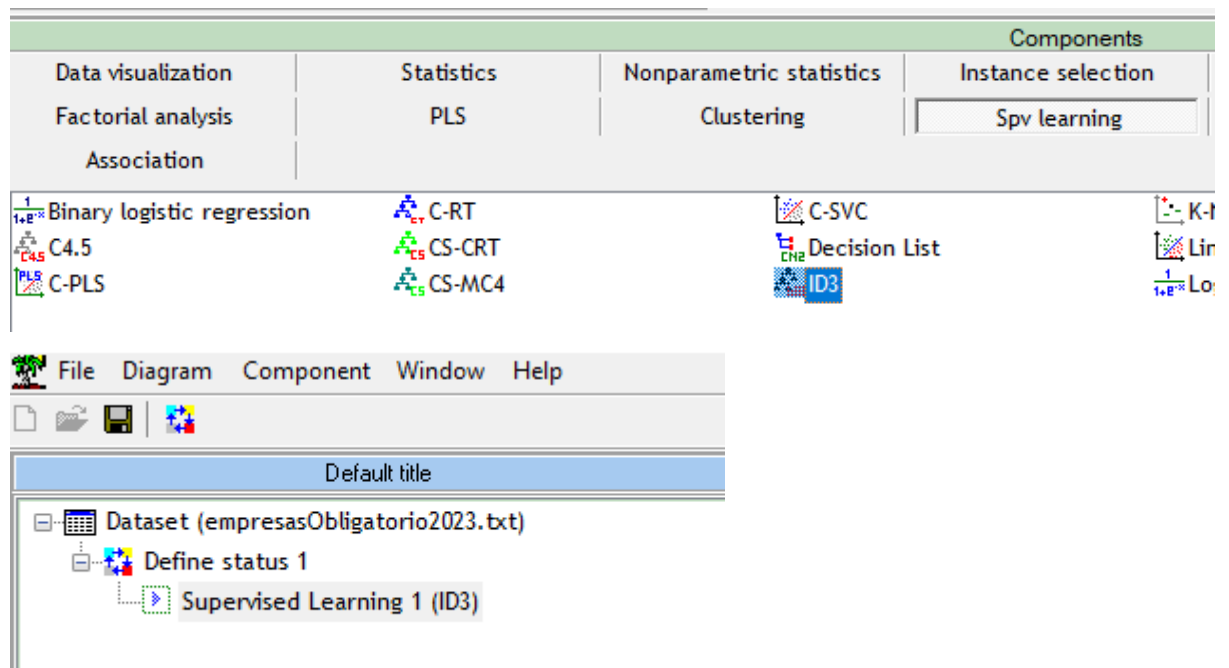
- 4) Una vez realizados los dos pasos anteriores, se debe hacer clic sobre el botón OK. Luego, deberemos hacer clic sobre “Define status 1”, que se encontrará colgado debajo del dataset, con esto lo ejecutaremos.



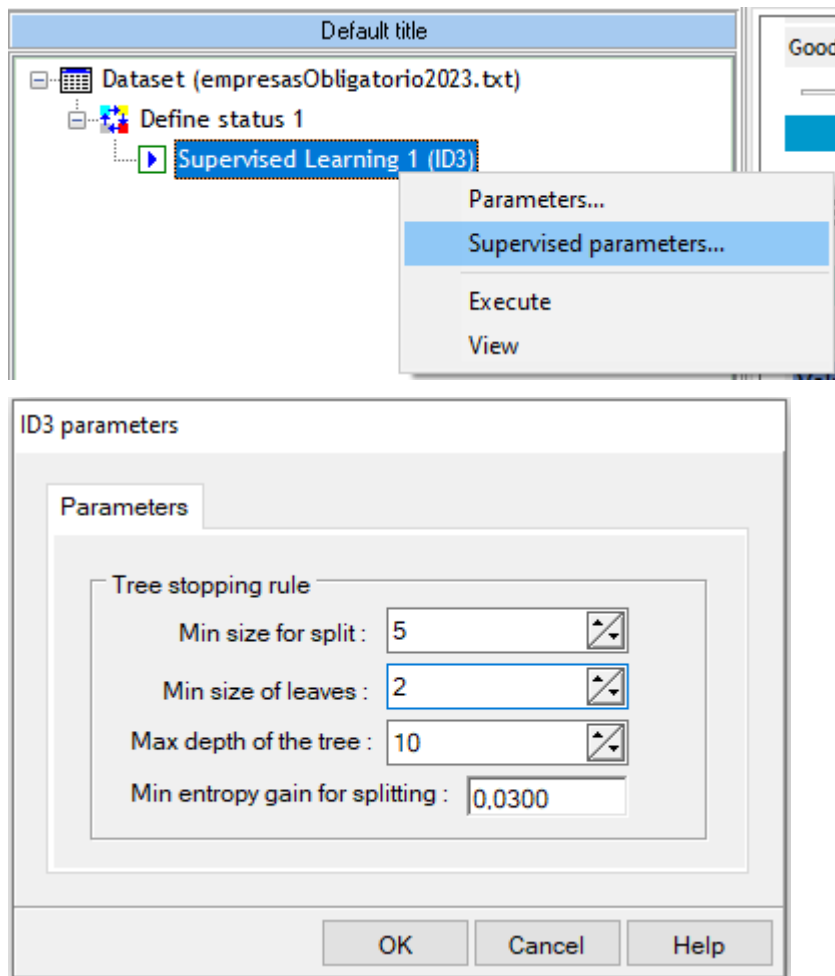
Después de ejecutar, si deseamos verificar lo hecho hasta el momento, en la tabla que se visualiza deberá aparecer el valor yes en la columna input para todos los ratios financieros indicados y en la columna Target para el atributo Quiebra.

Results			
Attribute	Target	Input	Illustrative
id	-	-	-
FlujodeCaja/DeudaTotal	-	yes	-
IngresoNeto/ActivoTotal	-	yes	-
ActivoCorriente/PasivoCorriente	-	yes	-
ActivoCorriente/VentasNetas	-	yes	-
Quiebra	yes	-	-

- 5) Para proceder con el árbol de decisión debemos ir a la sección *Components* y seleccionar *Spv Learning*. Dentro, seleccionaremos *ID3* (indicado de color azul), arrastrándolo debajo del define status creado anteriormente.



- 6) Previo a ejecutar el algoritmo, modificaremos los valores de los dos primeros parámetros por 5 y 2 respectivamente. Para llevar a cabo esta tarea debemos hacer clic derecho sobre "Supervised Learning 1" y al desplegarse el menú seleccionar la opción "Supervised parameters...". Una vez en el mismo, como ya mencionamos, modificaremos los valores mencionados.



- 7) Una vez que finalicemos, le damos clic en OK y ejecutamos el algoritmo haciendo doble clic en el mismo. Obteniendo así un árbol de decisión.

File Diagram Component Window Help

Default title

Dataset (empresasObligatorio2023.txt)

Define status 1

Supervised Learning 1 (ID3)

Values prediction			Confusion matrix			
Value	Recall	1-Precision		NO	SI	Sum
NO	1,0000	0,0000	NO	53	0	53
SI	1,0000	0,0000	SI	0	47	47
			Sum	53	47	100

Classifier characteristics

Data description

Target attribute: Quiebra (2 values)

descriptors: 4

Tree description

Number of nodes: 11

Number of leaves: 6

Decision tree

- ActivoCorriente/PasivoCorriente < 2,1593
 - FlujodeCaja/DeudaTotal < 0,0747 then Quiebra = SI (100,00 % of 41 examples)
 - FlujodeCaja/DeudaTotal >= 0,0747
 - IngresoNeto/ActivoTotal < 0,0983
 - ActivoCorriente/VentasNetas < 0,2871
 - ActivoCorriente/PasivoCorriente < 1,5603 then Quiebra = NO (100,00 % of 3 examples)
 - ActivoCorriente/PasivoCorriente >= 1,5603 then Quiebra = SI (100,00 % of 3 examples)
 - ActivoCorriente/VentasNetas >= 0,2871 then Quiebra = NO (100,00 % of 9 examples)
 - IngresoNeto/ActivoTotal >= 0,0983 then Quiebra = SI (100,00 % of 3 examples)
- ActivoCorriente/PasivoCorriente >= 2,1593 then Quiebra = NO (100,00 % of 41 examples)

Árbol de decisión

Decision tree

- ActivoCorriente/PasivoCorriente < 2,1593
 - FlujodeCaja/DeudaTotal < 0,0747 then Quiebra = SI (100,00 % of 41 examples)
 - FlujodeCaja/DeudaTotal >= 0,0747
 - IngresoNeto/ActivoTotal < 0,0983
 - ActivoCorriente/VentasNetas < 0,2871
 - ActivoCorriente/PasivoCorriente < 1,5603 then Quiebra = NO (100,00 % of 3 examples)
 - ActivoCorriente/PasivoCorriente >= 1,5603 then Quiebra = SI (100,00 % of 3 examples)
 - ActivoCorriente/VentasNetas >= 0,2871 then Quiebra = NO (100,00 % of 9 examples)
 - IngresoNeto/ActivoTotal >= 0,0983 then Quiebra = SI (100,00 % of 3 examples)
- ActivoCorriente/PasivoCorriente >= 2,1593 then Quiebra = NO (100,00 % of 41 examples)

Se observa que la variable que mejor discrimina es ActivoCorriente/PasivoCorriente, ya que se comienzan a abrir ramas a partir de ella. Luego, continúan habiendo nuevos casos (solo si dicha variable es <2,1593), empleando el atributo FlujodeCaja/DeudaTotal para continuar abriendo ramas del árbol.

En la descripción dada por Tanagra se observa que el árbol tiene 11 nodos y 6 hojas.

Relación con los negocios

Un árbol de decisión en este contexto puede ser útil de múltiples maneras, algunas de ellas son:

Planificación financiera y mitigación de riesgos: Puede ser utilizado como una herramienta en la planificación financiera y en la mitigación de riesgos empresariales. Al considerar los diferentes escenarios y resultados que presenta el árbol, las empresas pueden anticipar y prepararse para posibles dificultades financieras.

Toma de decisiones de inversión: Los inversores pueden utilizar el árbol de decisión como una herramienta para evaluar la viabilidad financiera de una empresa antes de tomar decisiones de inversión; evaluando así el riesgo asociado con una inversión potencial y tomando una decisión informada.

Preguntas

- 1) ¿Qué otro algoritmo se podría emplear en Tanagra para hallar un árbol de decisión? ¿Qué diferencia presenta con ID3?
- 2) ¿Por qué en el ejercicio realizado no utilizamos el atributo id?
- 3) ¿Qué significan los parámetros que modificamos en el ejercicio?

Respuestas

- 1) Se podría emplear el algoritmo C4.5, también conocido como el sucesor del algoritmo ID3. Ambos son algoritmos de aprendizaje automático utilizados para construir árboles de decisión a partir de conjuntos de datos de entrenamiento. Si bien comparten similitudes, C4.5 es una mejora del algoritmo ID3, ya que permite trabajar tanto con atributos categóricos como continuos, tiene la capacidad de manejar valores faltantes en los datos e introduce la técnica de poda en la construcción del árbol de decisión.
- 2) No utilizamos el atributo id ya que no aporta información significativa sobre la predicción o clasificación del problema en cuestión. No tienen relación con la variable objetivo que se intenta predecir. Incluso, podría llegar a generar sobreajuste del modelo o dificultar la interpretación.
- 3) Min size for split (tamaño mínimo para dividir): Este parámetro establece el número mínimo de ejemplos que deben estar presentes en un nodo para que se considere elegible para dividirse en subárboles adicionales.
Min size of leaves (tamaño mínimo de las hojas): Este parámetro establece el número mínimo de ejemplos que deben estar presentes en una hoja del árbol.

2.2) Árboles de decisión – (Weka, ProyectosObligatorio2023.txt)

Partimos de la base que ya tenemos instalado el programa Weka y ya convertimos el archivo ProyectosObligatorio2023.txt a tipo arff (ambos procedimientos se mostraron en el ejercicio 1.1)

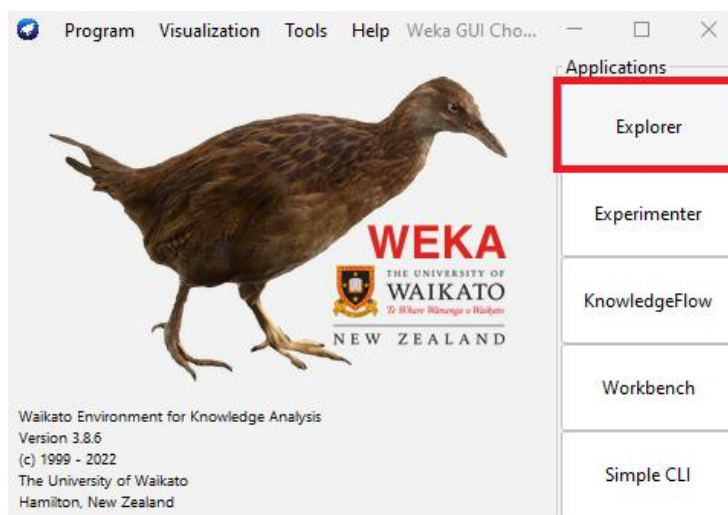
Para comenzar, debemos abrir el archivo, pero en esta ocasión optaremos por hacerlo de otra forma, así en el documento se presentan las dos alternativas.

Proceso para abrir y configurar el archivo ProyectosObligatorio2023.arff en Weka:

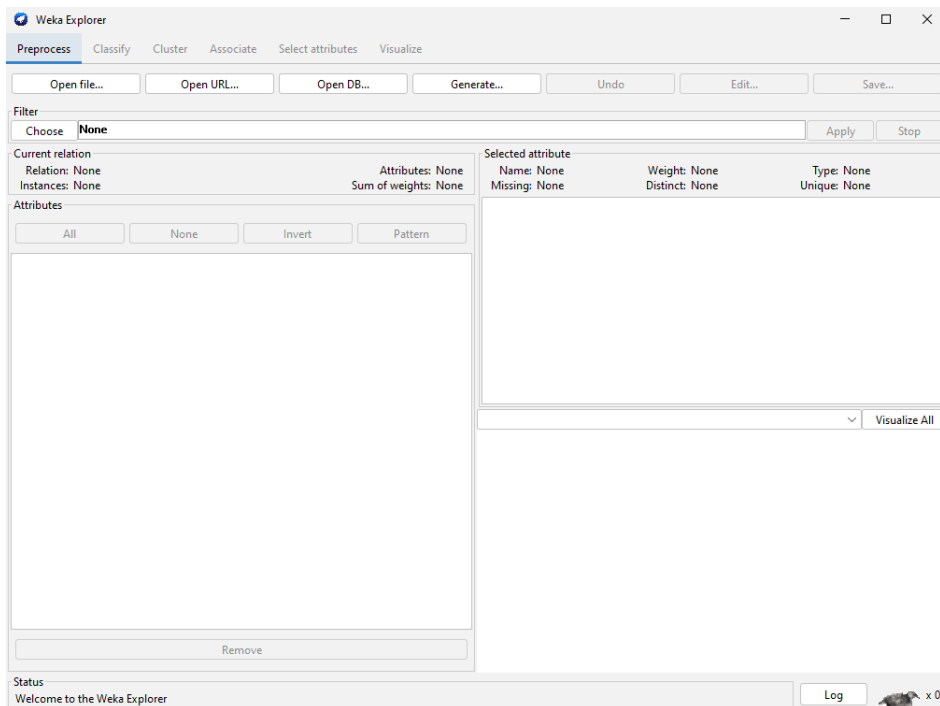
- 1) Abrir la aplicación de Weka, observando la siguiente pantalla:



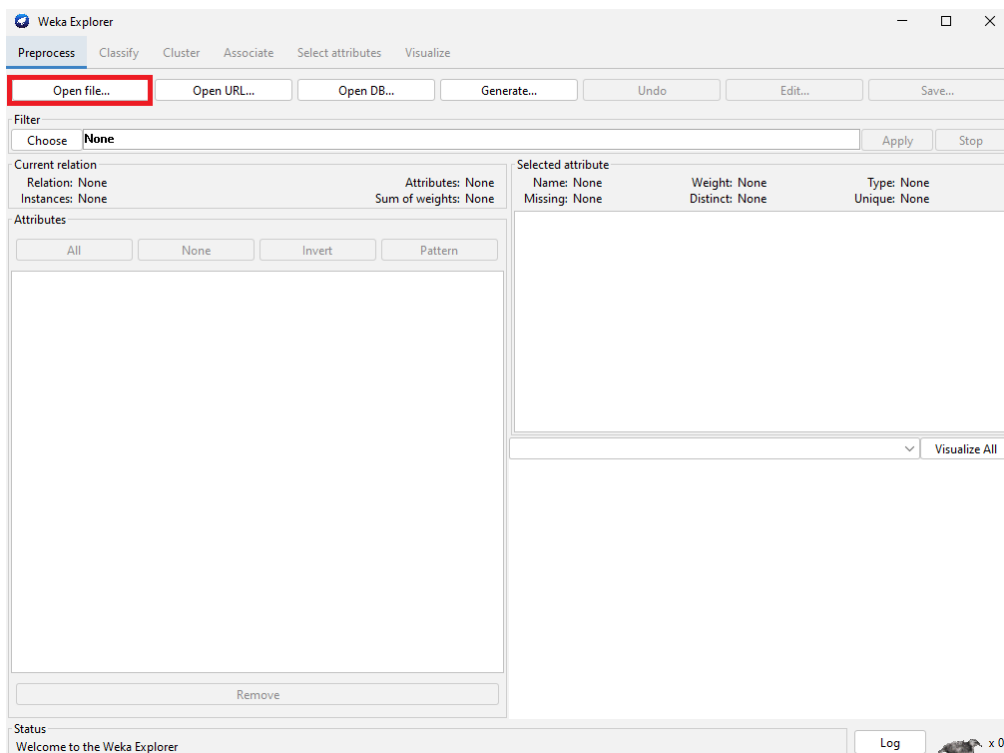
- 2) Luego de abierto debemos hacer clic sobre Explorer (opción que se encuentra marcada en rojo)



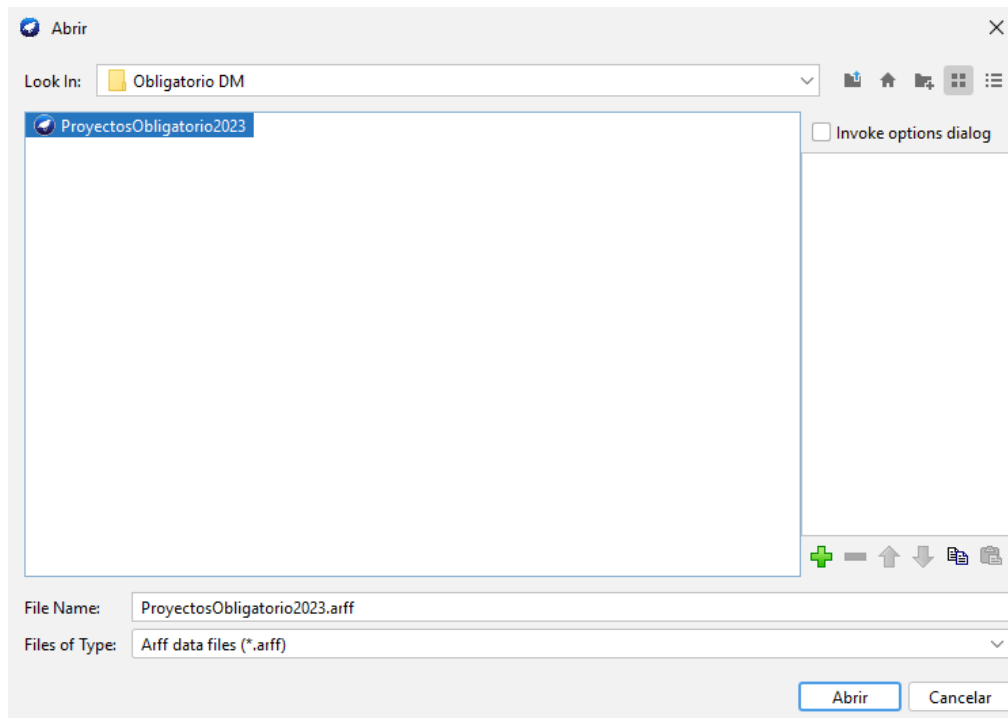
El hacer clic nos abrirá la siguiente pantalla:



- 3) Finalizada la parte anterior, en la pantalla recientemente abierta, deberemos hacer clic sobre la opción *open file*, esto nos permitirá seleccionar el archivo a utilizar.



Una vez que hicimos clic en la opción mencionada, se nos abrirá el siguiente recuadro donde deberemos seleccionar el archivo a cargar. En este caso cargaremos *ProyectosObligatorio2023.arff*. Al cargarlo, hacemos clic en *Abrir*.



- 4) Posterior a hacer clic en Abrir, se obtiene como resultado de la apertura del archivo lo siguiente:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter: Choose **None** Apply Stop

Current relation: Relation: DataTable Instances: 99 Attributes: 7 Sum of weights: 99

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> Sexo
2	<input type="checkbox"/> NivelEstudios
3	<input type="checkbox"/> SituacionLaboral
4	<input type="checkbox"/> OtrosProyectos
5	<input type="checkbox"/> ProyectoExitoso
6	<input type="checkbox"/> FinanciamientoApropiado
7	<input type="checkbox"/> EdadResponsable

Remove

Selected attribute: Name: Sexo Missing: 0 (0%) Distinct: 2 Type: Nominal Unique: 0 (0%)

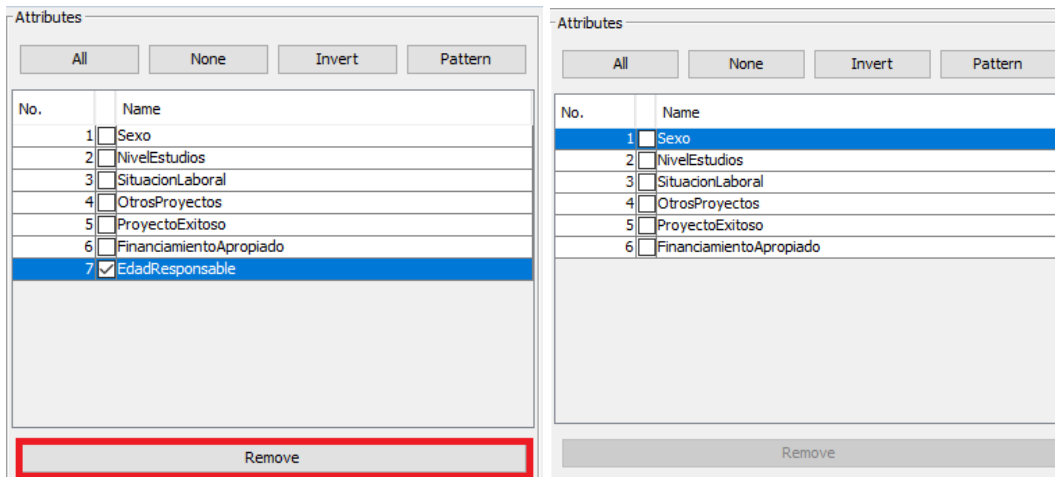
No.	Label	Count	Weight
1	Masculino	56	56
2	Femenino	43	43

Class: EdadResponsable (Num) Visualize All

Status: OK Log x 0

- 5) Dado que Weka no puede manejar directamente variables cualitativas y cuantitativas al mismo tiempo. Debemos realizar una modificación a los datos, por lo que procederemos eliminando el atributo EdadResponsable; realizando el árbol con los demás restantes.

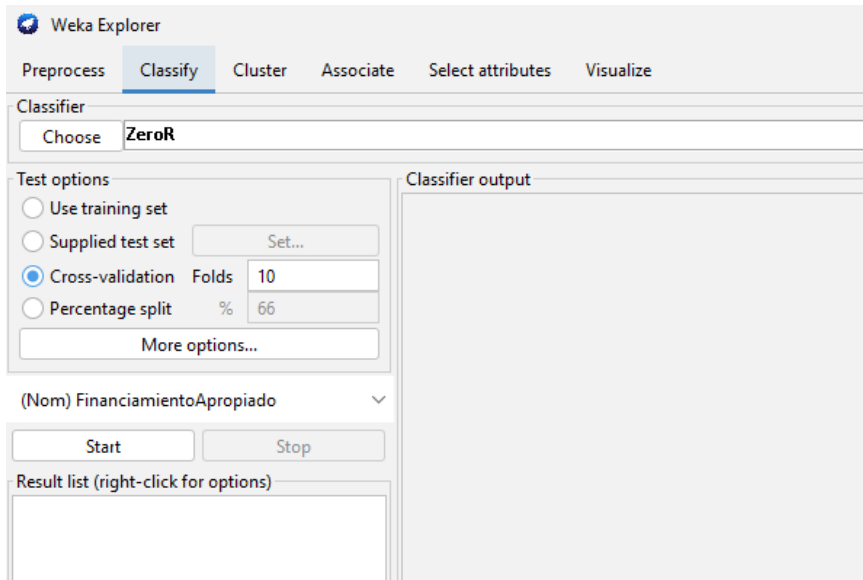
Para esto debemos seleccionarlo haciendo clic sobre el cuadrado que se encuentra a su izquierda, y hacer clic en el botón *Remove*, como se observa en las imágenes a continuación:



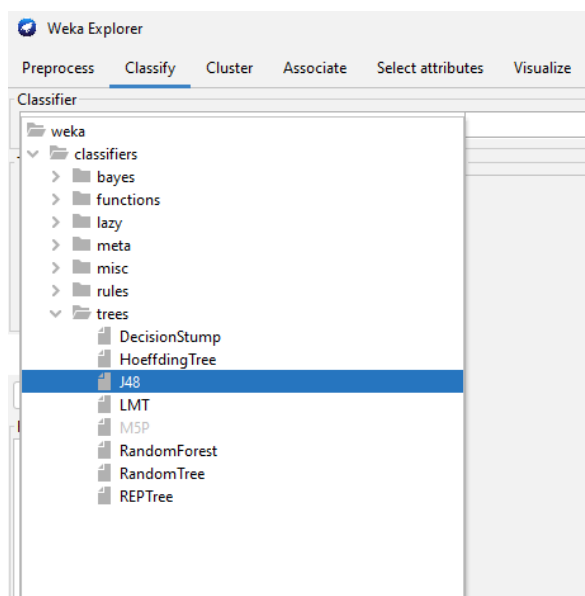
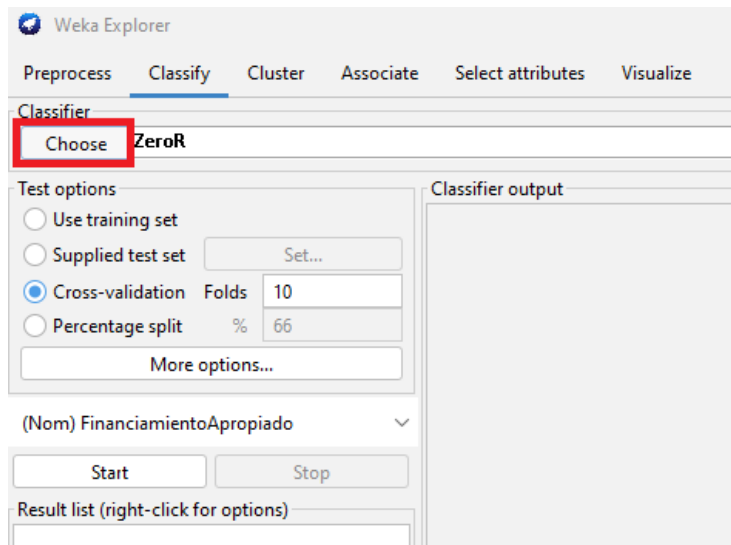
Una vez tengamos el archivo configurado con los atributos a utilizar, podemos proseguir con el proceso de crear el árbol de decisión.

Para generarlo, utilizaremos el algoritmo J48, siguiendo los pasos a continuación:

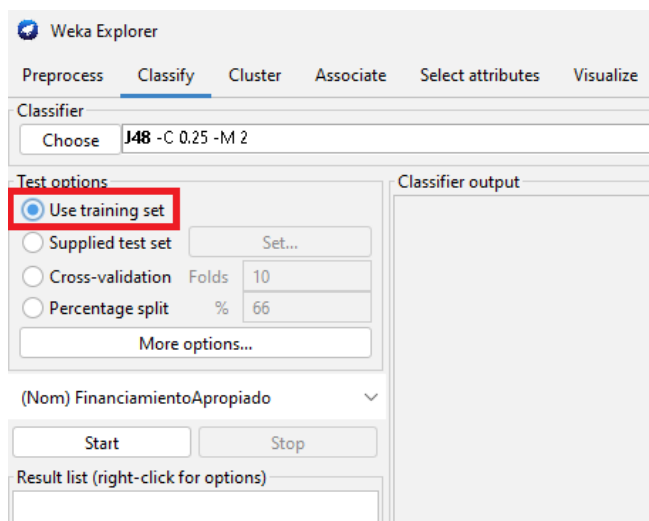
- 1) Clic en *Classify*, en la barra superior. Observaremos la siguiente pantalla:



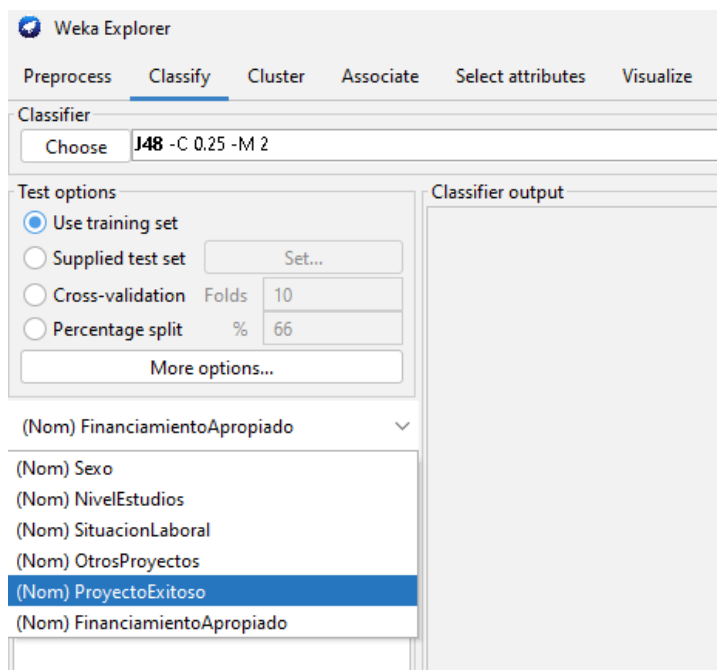
- 2) Posteriormente haremos clic en *Choose*, esto abrirá diversas opciones, nos dirigiremos a la carpeta tres y haremos clic sobre ella. Por último, seleccionaremos el algoritmo J48.



- 3) Continuando con la configuración, en el cuadrante denominado *Test options*, seleccionaremos la opción *Use training set*

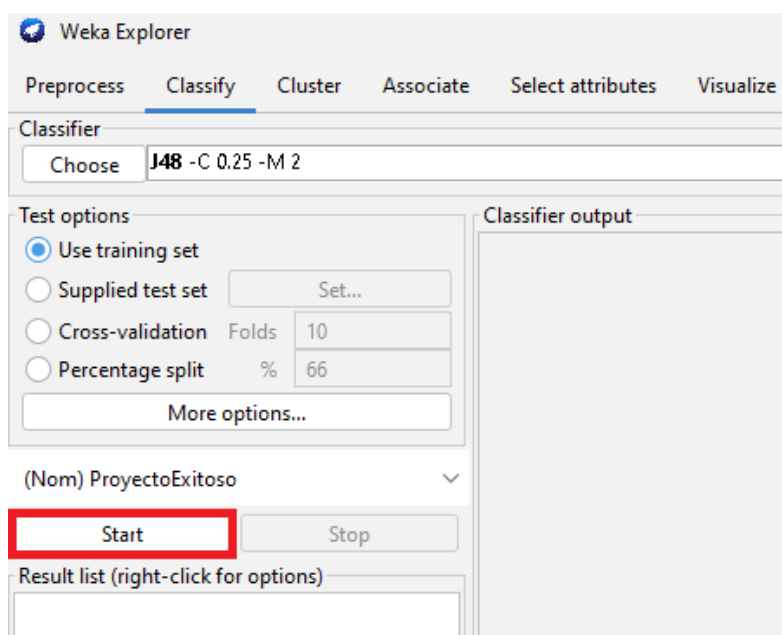


- 4) Verificamos que donde dice (Nom) se encuentre el atributo que deseamos predecir, en este caso, ProyectoExitoso. En caso de no encontrarse, hacemos clic sobre el que se encuentra y seleccionamos el deseado.



En este ejemplo se observa que se encontraba FinanciamientoApropiado, y se modifica seleccionando ProyectoExitoso

- 5) Habiendo finalizado la configuración, hacemos clic en Start y obtendremos el árbol de decisión correspondiente.



Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **J48 -C 0.25 -M 2**

Test options:
☒ Use training set
☐ Supplied test set (Set...)
☐ Cross-validation (Folds: 10)
☐ Percentage split (%: 66)
 More options...

(Nom) ProyectoExitoso

Start Stop

Result list (right-click for options):
 05:08:18 - trees.J48

Classifier output

```

=== Classifier model (full training set) ===

J48 pruned tree
-----

SituacionLaboral = Independiente
| FinanciamientoApropiado = NO
| | Sexo = Masculino: SI (10.0/2.0)
| | Sexo = Femenino: NO (3.0)
| FinanciamientoApropiado = SI: SI (32.0/2.0)
SituacionLaboral = Dependiente
| FinanciamientoApropiado = NO: NO (21.0/2.0)
| FinanciamientoApropiado = SI
| | NivelEstudios = Universitaria Incompleta: SI (21.0/8.0)
| | NivelEstudios = Universitaria Completa: NO (4.0)
| | NivelEstudios = Postgrado: SI (4.0/1.0)
| | NivelEstudios = Secundaria Completa: NO (3.0/1.0)
| | NivelEstudios = Secundaria Incompleta: SI (0.0)
SituacionLaboral = No trabajaba: NO (1.0)

Number of Leaves :    10
Size of the tree :    15

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances      83      83.8384 %
Incorrectly Classified Instances    16      16.1616 %
Kappa statistic                    0.6585
Mean absolute error                 0.2354
Root mean squared error             0.3431
Relative absolute error             48.169 %
Root relative squared error        69.4197 %
Total Number of Instances          99

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,947    0,310    0,806    0,947    0,871    0,674    0,897    0,887    SI
0,690    0,053    0,906    0,690    0,784    0,674    0,897    0,852    NO
Weighted Avg.    0,838    0,201    0,849    0,838    0,834    0,674    0,897    0,872
  
```

Si deseamos tener una mejor visualización del árbol de decisión, podemos dirigirnos al cuadrante *Result list* donde se encuentra la ejecución realizada, hacer clic derecho sobre la misma y seleccionar *Visualize tree*. Esto nos permitirá observar los nodos y hojas de manera más gráfica.

(Nom) ProyectoExitoso

Start Stop

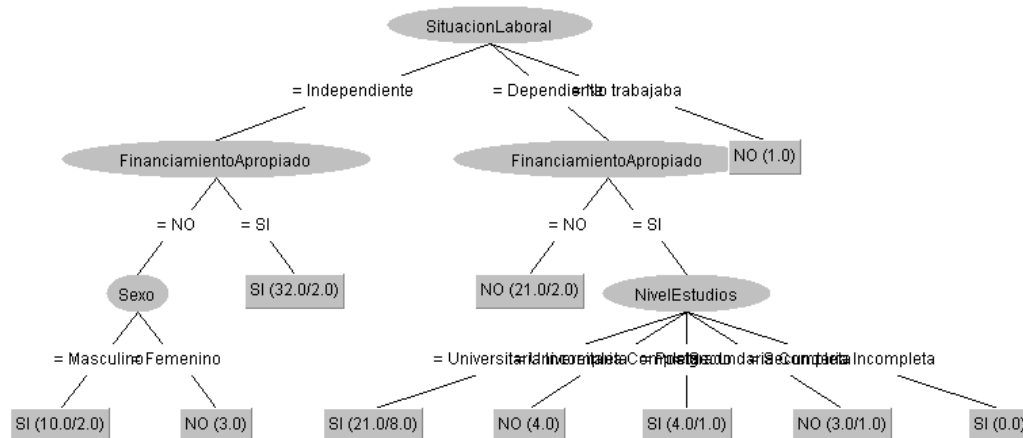
Result list (right-click for options):
 05:08:18 - trees.J48

Test mode

- View in main window
- View in separate window
- Save result buffer
- Delete result buffer(s)
- Load model
- Save model
- Re-evaluate model on current test set
- Re-apply this model's configuration
- Visualize classifier errors
- Visualize tree**
- Visualize margin curve
- Visualize threshold curve >
- Cost/Benefit analysis >
- Visualize cost curve >

El resultado se muestra a continuación:

Árbol de Decisión



Como podemos observar la variable que más discrimina es la situación laboral, abriendo tres posibles ramas: No trabaja, Independiente o Dependiente. Para los casos de Independiente y Dependiente se siguen abriendo ramas a partir de la variable FinanciamientoApropiado.

Relación con los negocios

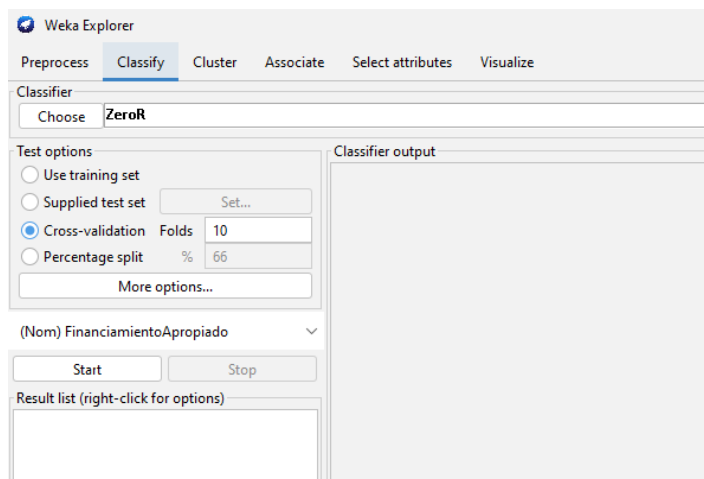
Este árbol de decisión podría ayudarnos a identificar factores clave ya que muestra qué variables son relevantes para predecir el éxito de un proyecto. En este caso, la situación laboral, el financiamiento apropiado y el nivel de estudios son considerados como factores importantes. Ayuda a los negocios a comprender qué aspectos deben tener en cuenta al evaluar un proyecto y tomar decisiones informadas.

Preguntas

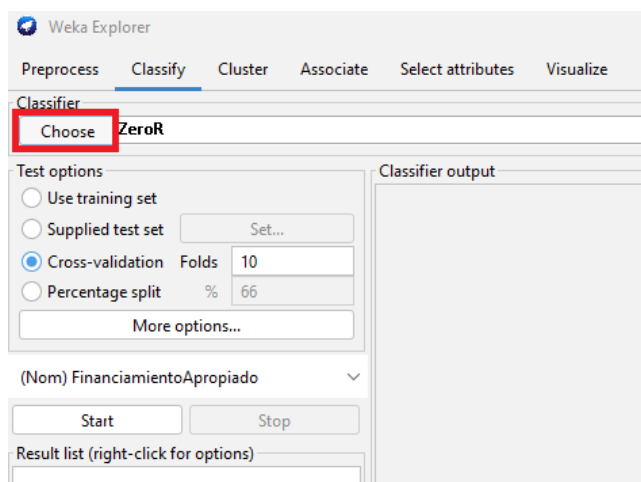
- 1) ¿Cómo podemos hacer para cambiar el algoritmo utilizado en la generación del árbol de decisión?
- 2) ¿Cómo modificamos los parámetros del algoritmo utilizado?
- 3) ¿Cómo podemos cambiar la forma en que se toman los datos para realizar el entrenamiento y testeo del modelo?

Respuestas

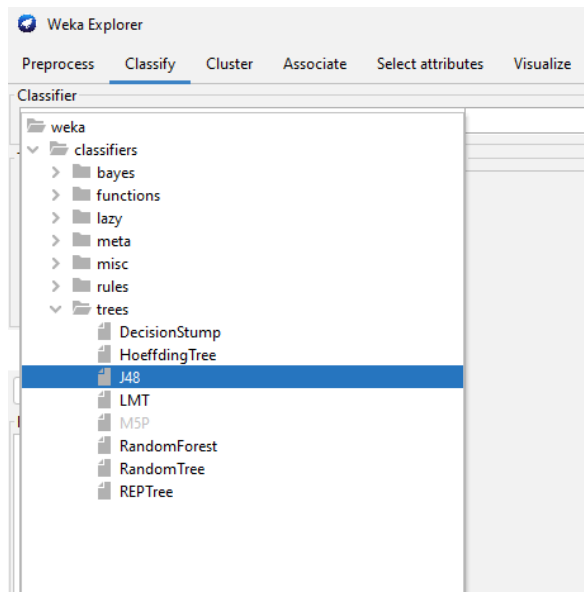
- 1) El proceso es el mismo que al seleccionar el algoritmo J48. Debemos dirigirnos a *Classify*



Posteriormente seleccionamos *Choose*

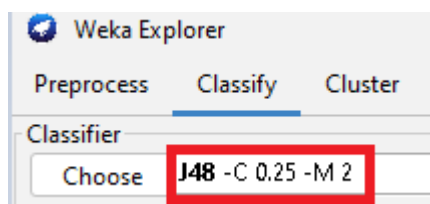


Dentro de la pantalla desplegada, seleccionamos la carpeta denominada Trees. Luego de seleccionarla simplemente elegimos el algoritmo que deseamos utilizar.

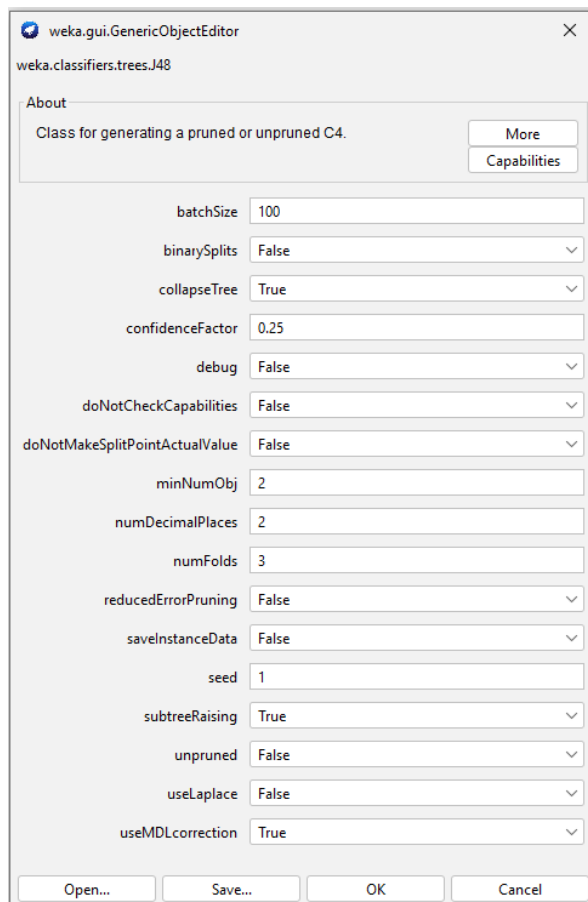


Se observan distintos archivos, entre ellos, DecisionStump: un modelo de aprendizaje automático que se utiliza en problemas de clasificación binaria. Es una forma muy simple de árbol de decisión que consta de una sola capa y una única división en una variable predictora. Otra opción disponible es RandomForest: algoritmo de aprendizaje automático para realizar tareas de clasificación y regresión. Es una técnica basada en ensambles que combina múltiples árboles de decisión individuales para obtener predicciones más precisas y robustas.

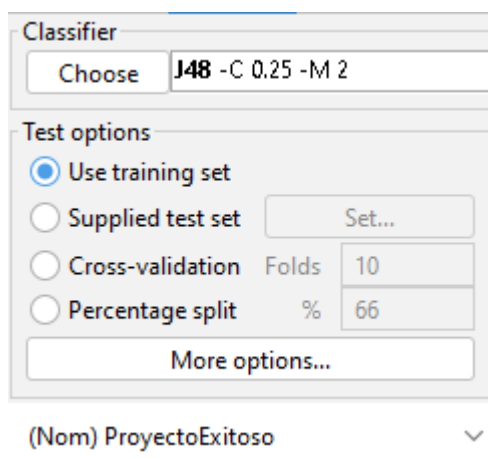
- 2) Para modificar los parámetros del algoritmo utilizado, en este caso J48, lo que debemos es hacer clic sobre el espacio que se muestra a continuación marcado por color rojo.



Una vez hecho el clic se abrirá una nueva ventana donde podremos modificar los parámetros que se desee.



- 3) Si deseamos cambiar la forma en que se toman los datos para testear y entrenar el modelo, como ya lo hicimos anteriormente, debemos dirigirnos a la sección de test options, y dentro de ella marcar la opción deseada.



Use training set: Se utilizan para testear los mismos datos que fueron utilizados para entrenar el modelo.

Supplied test set: Se utiliza un conjunto de datos independientes para probar el modelo.

Cross-validation: Consiste en que dado un número n se divide los datos en n partes, y por cada parte, se construye el clasificador con las $n-1$ partes restantes y se prueba con esa. Este proceso es por cada una de las n particiones.

Percentage Split: Se define un porcentaje que se destinará para entrenar el modelo, y el porcentaje restante se utiliza para testear.

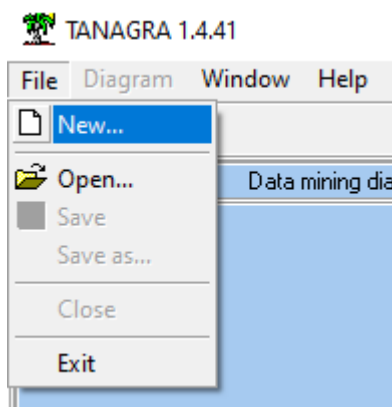
Ejercicio 3)

Identifique grupos de elementos parecidos aplicando clustering utilizando dos herramientas y dos datasets. (un dataset para cada herramienta). Utilice variables cuantitativas por su naturaleza (no cuantifique variables no cuantitativas).

3.1) Clustering – (Tanagra, empresasObligatorio2023.txt)

Asumiendo que el programa Tanagra ya se encuentra instalado, y el archivo empresasObligatorio2023.txt descargado, daremos el instructivo a partir de la carga del archivo en Tanagra.

Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo Dataset (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo empresasObligatorio2023.txt en la ubicación donde fue guardado.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :

OK Cancel Help

Una vez seleccionado el archivo se nos cargará en el campo *Dataset* la ruta de este.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :
C:\Users\Administrador\Downloads\empresasObligatorio2023.txt

OK Cancel Help

Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:

TANAGRA 1.4.41 - [Dataset (empresasObligatorio2023.txt)]

File Diagram Component Window Help

Default title
Dataset (empresasObligatorio2023.txt)

Dataset (empresasObligatorio2023.txt)

Parameters

Database : C:\Users\Administrador\Downloads\empresasObligatorio2023.txt

Results

Download information

DataSource processing
Computation time 0 ms
Allocated memory 9 KB

Dataset description

6 attribute(s)
100 example(s)

Attribute	Category	Informations
id	Continue	-
FlujodeCaja/DeudaTotal	Continue	-
IngresosIeto/ActivoTotal	Continue	-
ActivoCorriente/PasivoCorriente	Continue	-
ActivoCorriente/VentasNetas	Continue	-
Quiebra	Discrete	2 values

Computation time : 0 ms.
Created at 26/6/2023 22:26:47

En esta parte, debemos verificar que todos los atributos excepto el denominado Grupo (que es de tipo Discrete) sean de tipo Continue. Siendo que, como ya hemos mencionado, en Tanagra estas categorizaciones se corresponden de la siguiente manera:

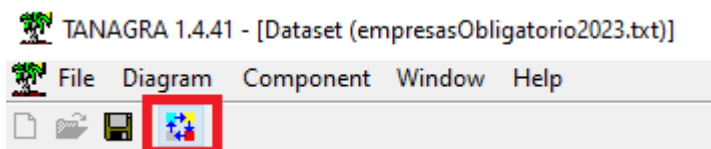
- **Discreta:** Es el equivalente a cualitativa
- **Continua:** Es el equivalente a cuantitativa

En caso de que las categorizaciones de los atributos estén invertidas a como se describió, debemos solucionarlo cambiando en el archivo las “,” (comas) por “.” (puntos).

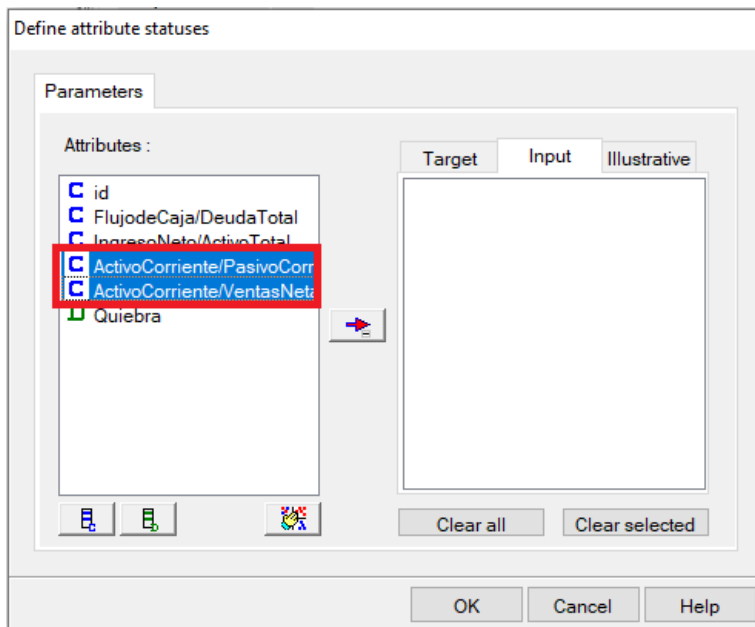
En este ejercicio aplicaremos el algoritmo HAC1 con las variables de input: ActivoCorriente/PasivoCorriente, ActivoCorriente/VentasNetas.

Para continuar, debemos tener en cuenta que antes de aplicar un algoritmo vamos a tener que seleccionar los datos con los que trabajaremos, por lo que, haremos un *Define Status*, haciendo los pasos que se muestran a continuación:

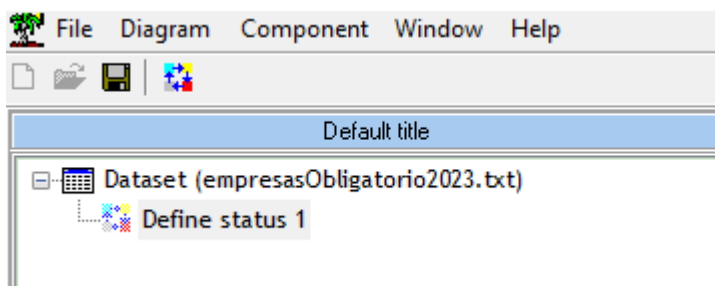
- 1) Debemos hacer clic en el ícono marcado en rojo en la siguiente imagen:



- 2) Se abrirá una ventana denominada *Define attribute statuses*, donde debemos colocar en el input las variables ActivoCorriente/PasivoCorriente y ActivoCorriente/VentasNetas.



3) Finalmente presionamos el botón Ok, quedando de la siguiente forma:



Luego hacemos doble clic en *Define Status 1* para verificar que el input se haya configurado de forma adecuada. Se debería visualizar lo siguiente:

Results			
Attribute	Target	Input	Illustrative
id	-	-	-
FlujodeCaja/DeudaTotal	-	-	-
IngresoNeto/ActivoTotal	-	-	-
ActivoCorriente/PasivoCorriente	-	yes	-
ActivoCorriente/VentasNetas	-	yes	-
Quiebra	-	-	-

Continuando, para aplicar el algoritmo HAC 1 con las variables input: ActivoCorriente/PasivoCorriente y ActivoCorriente/VentasNetas, debemos seguir los siguientes pasos:

- 1) Nos dirigimos a Components, en la sección de Clustering, y dentro de ella vamos a seleccionar HAC (marcado en azul). Lo arrastramos y colgamos debajo del Define Status previamente creado. HAC es un algoritmo de agrupamiento aglomerativo que usa enlace-simple.

The screenshot shows the 'Components' menu with the following structure:

Components			
Data visualization	Statistics	Nonparametric statistics	Instance selection
Factorial analysis	PLS	Clustering	Spv learning
Association			

Below the menu, various clustering algorithms are listed with icons:

- CT, CTP, EM-Clustering
- EM-Selection, **HAC** (highlighted in blue), K-Means
- Kohonen-SOM, LVQ, Neighborhood Graph
- VARCLUS, VARHCA, VARKMeans

The project tree on the right shows the following structure:

- Dataset (empresasObligatorio2023.txt)
 - Define status 1
 - HAC 1** (highlighted in blue)

- 2) Hacemos doble clic sobre HAC 1, ejecutándolo con los parámetros por defecto. Obteniendo así el siguiente resultado:

The screenshot shows the 'HAC 1' parameters and results. The 'Report' tab is selected.

HAC 1 Parameters

# clusters	
Detection	Automatic

Data transformation

Transformation	
Transformation	None

Visualization

Index selection	
Index selection	1
Tree structure	
Tree structure	0
Anova per variable	
Anova per variable	0

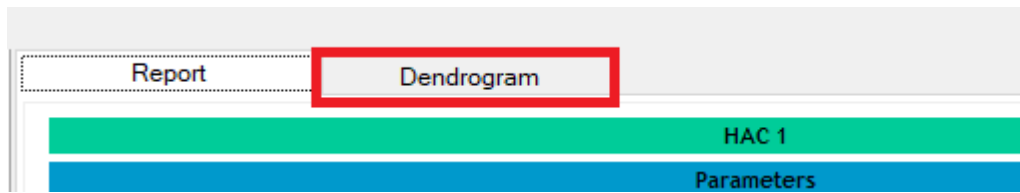
Clustering results

# clusters		
Clusters	3	
Cluster	Description	Size
cluster n°1	c_hac_1	47
cluster n°2	c_hac_2	10
cluster n°3	c_hac_3	43

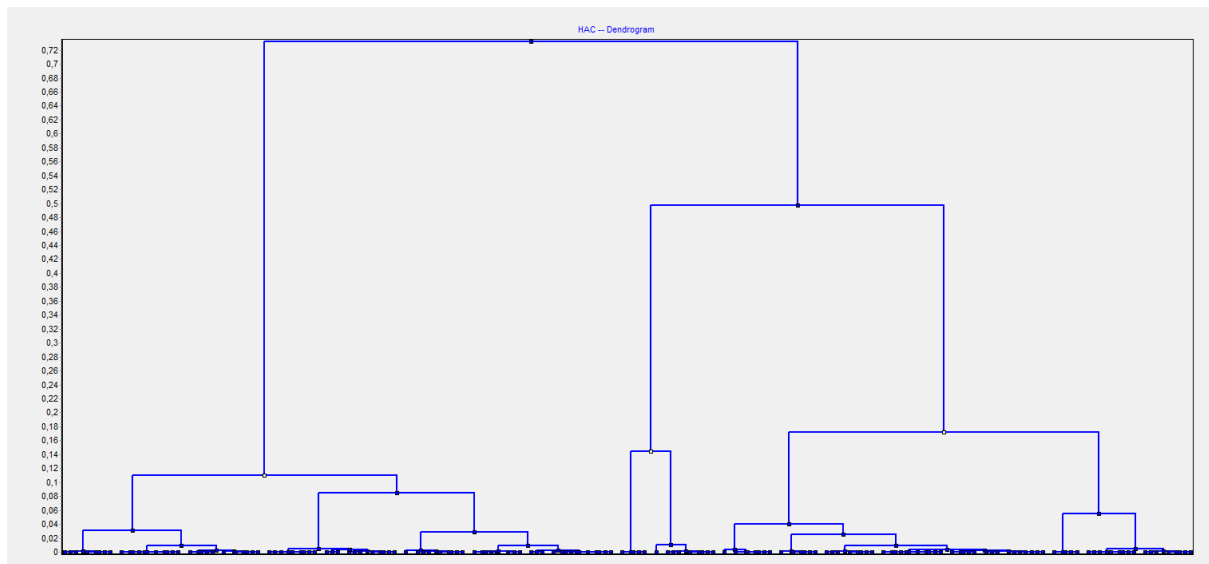
Best cluster selection

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,3661	0,2354
3	0,6144	0,3247
4	0,7004	0,0276
5	0,7727	0,0349
6	0,8274	0,0252

Si deseamos acceder directamente al dendrograma, debemos hacer clic sobre Dendrogram, a la derecha de Report, como se muestra en la siguiente imagen:



Obteniendo el siguiente dendrograma que se aprecia en la figura:

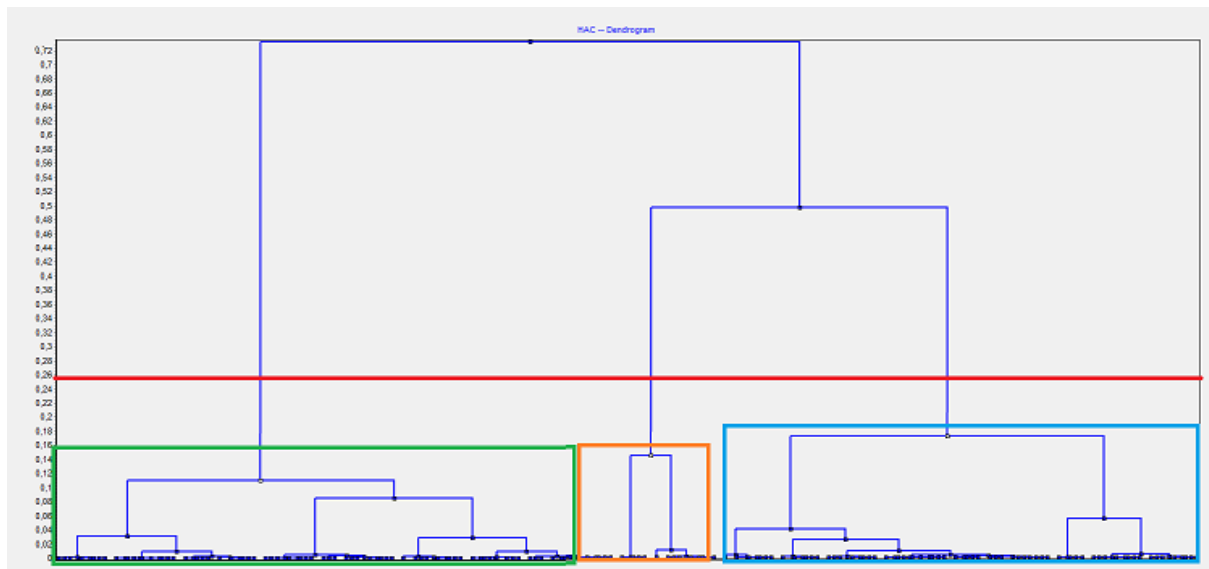


El eje vertical representa una medida de distancia. Nos muestra que los elementos que se juntan primero son los que tienen una distancia más chica, siendo más parecidos. Luego vemos que los elementos se siguen juntando pero a distancias mayores.

Podemos realizar cortes horizontales a diferentes distancias y ver cuántos clusters tenemos a partir del mismo, es decir, plantear un eje imaginario para observar cantidad de clusters.

A medida que aumentamos la distancia estamos juntando cosas muy diferentes.

A continuación, plantearemos un posible corte:



De esta forma podemos observar que se generan 3 clusters.

El algoritmo que utilizamos realiza este trabajo de manera automática, seleccionando los 3 clusters que tienen mejor distancia entre los elementos.

Se puede observar a continuación:

Clustering results

Clusters	3	
Cluster	Description	Size
cluster n° 1	c_hac_1	47
cluster n° 2	c_hac_2	10
cluster n° 3	c_hac_3	43

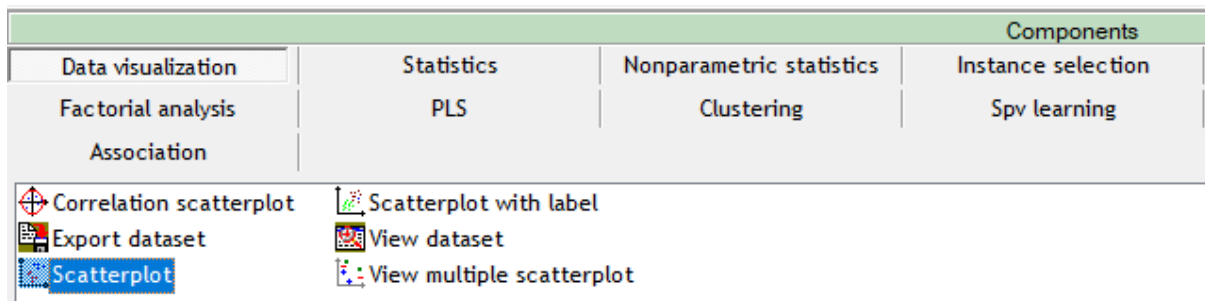
El primer cluster tiene 47 elementos, el segundo tiene 10 elementos y el último tiene 43 elementos.

En la siguiente figura, podemos observar el trabajo que fue realizando el algoritmo para determinar la cantidad de clusters a formar:

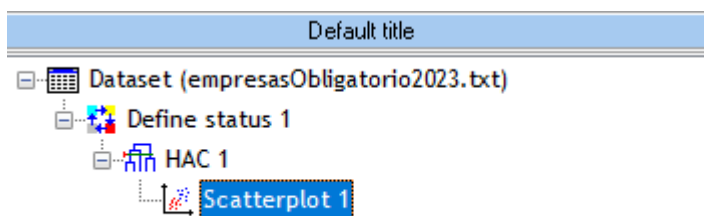
Best cluster selection

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,3661	0,2354
3	0,6144	0,3247
4	0,7004	0,0276
5	0,7727	0,0349

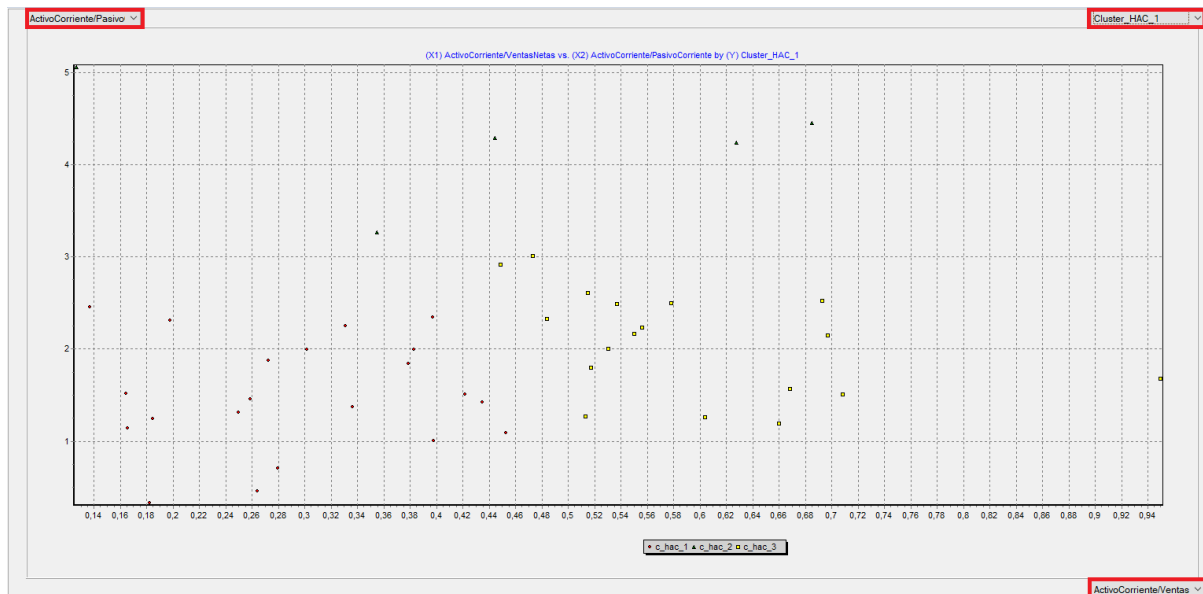
Si deseamos poder visualizar gráficamente los clusters, podemos generar un diagrama de dispersión para las variables trabajadas. Para realizar esto, nos dirigimos a *Components*, posteriormente a *Data Visualization* y seleccionamos *Scatter Plot*, colgándolo de HAC 1



Quedando de la siguiente forma:



Al hacer doble clic sobre Scatterplot 1 y ejecutarlo, podremos observar el diagrama. He de destacar que debemos configurar las variables del mismo. Como se podrá observar en la imagen a continuación, en el cuadro de la esquina superior izquierda debe decir: ActivoCorriente/PasivoCorriente, en el cuadro de la esquina superior derecha deberá decir Cluster_HAC_1 y por último, en el cuadro de la esquina inferior derecha deberá decir ActivoCorriente/VentasNetas. Esto nos permitirá observar los clusters con mayor precisión.



Se determinan que los círculos rojos se encuentran todos juntos, los triángulos verdes también y los cuadrados amarillos. Cada figura representa elementos de un cluster.

Relación con los negocios

El clustering de dos ratios financieros, como es en este caso, ActivoCorriente/PasivoCorriente y ActivoCorriente/VentasNetas, en los negocios puede ser útil para identificar patrones o grupos similares dentro de un conjunto de datos. Puede permitir identificar diferentes perfiles financieros dentro de un conjunto de empresas, segmentar el mercado en base a características financieras similares (esto se puede utilizar en estrategias de marketing y ventas) o puede ser útil para comparar el desempeño de empresas con características financieras similares y así identificar las mejores prácticas o estrategias utilizadas por las más exitosas dentro de cada grupo.

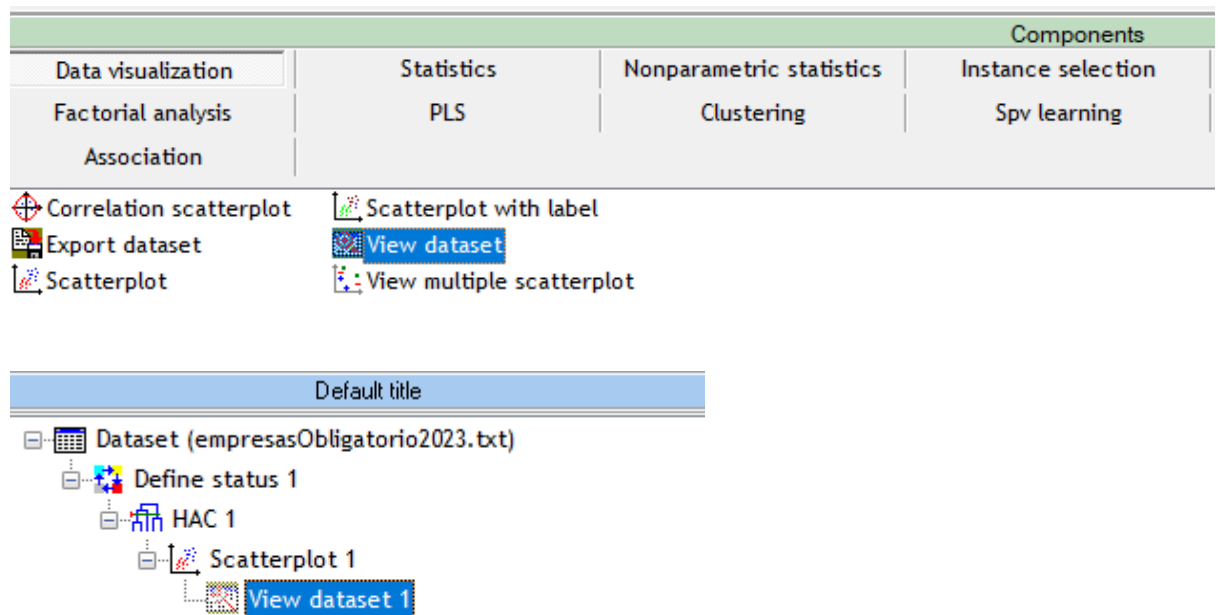
Preguntas

- 1) ¿Cómo podemos determinar a qué cluster pertenece cada uno de los elementos?
- 2) ¿Qué otro algoritmo se podría haber utilizado en este ejercicio?
- 3) ¿Podríamos haber trabajado con las 4 variables cuantitativas?

Respuestas

- 1) Si deseamos saber a qué cluster pertenece cada elemento, podemos utilizar un View Dataset y su última columna denominada Cluster_HAC_1 dirá a cuál pertenece.

Para hacer esto debemos dirigirnos a *Components*, a la sección *Data Visualization* y seleccionar *View dataset*, lo arrastramos y colgamos debajo de HAC 1.



Hacemos doble clic sobre el mismo para ejecutarlo. Una vez ejecutado como se mencionó, podremos observar en la última columna a qué cluster pertenece cada dato.

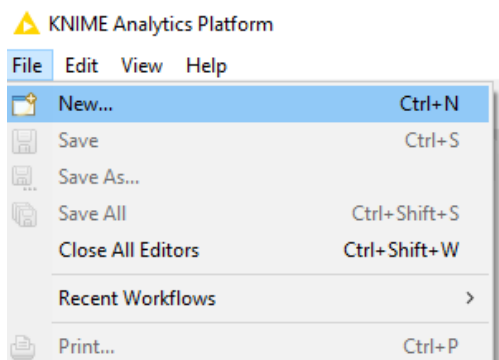
	id	FlujodeCaj	IngresoNet	ActivoCorr	ActivoCorr	Quiebra	Cluster_HA
1	1	0,4746	0,138	2,9166	0,4487	NO	c_hac_3
2	2	0,146	0,0518	2,1692	0,55	NO	c_hac_3
3	3	-0,2777	-0,2316	1,1918	0,6601	SI	c_hac_3
4	4	0,2029	0,0792	1,9936	0,3018	NO	c_hac_1
5	5	0,1184	0,0499	2,521	0,6925	NO	c_hac_3
6	6	-0,1002	-0,0917	1,5644	0,6683	SI	c_hac_3
7	7	0,2169	0,0779	2,3489	0,397	NO	c_hac_1
8	8	0,5808	0,0371	5,0594	0,1268	NO	c_hac_2
9	9	0,1486	0,0564	2,2347	0,5563	NO	c_hac_3
10	10	0,0451	0,0263	1,6756	0,9494	SI	c_hac_3
11	11	0,0713	0,0205	1,3124	0,2497	SI	c_hac_1
12	12	-0,4485	-0,4106	1,0865	0,4526	SI	c_hac_1
13	13	0,1454	0,05	1,8762	0,2723	SI	c_hac_1
14	14	0,3248	0,0718	4,2401	0,6279	NO	c_hac_2
15	15	0,1633	0,0486	2,308	0,1978	NO	c_hac_1
16	16	-0,0757	-0,0821	1,5077	0,4215	SI	c_hac_1
17	17	-0,5633	-0,3114	1,5134	0,1642	SI	c_hac_1
18	18	-0,1421	-0,0651	0,7066	0,2794	SI	c_hac_1
19	19	-0,0721	-0,093	1,4544	0,2589	SI	c_hac_1
20	20	0,5383	0,1064	2,3293	0,4835	NO	c_hac_3
21	21	0,1703	0,0695	1,7973	0,5174	NO	c_hac_3
22	22	-0,0351	-0,0147	1,5046	0,708	SI	c_hac_3
23	23	0,1933	0,0473	2,2506	0,3309	NO	c_hac_1
24	24	-0,333	-0,0854	3,0124	0,473	NO	c_hac_3
25	25	0,5603	0,1112	4,2918	0,4443	NO	c_hac_2
26	26	0,3776	0,1075	3,2651	0,3548	NO	c_hac_2
27	27	0,4785	0,091	1,2444	0,1847	NO	c_hac_1
28	28	0,0109	0,0011	2,1495	0,6969	SI	c_hac_3
29	29	-0,2298	-0,2961	0,331	0,1824	SI	c_hac_1
30	30	0,1227	0,1055	1,1434	0,1655	SI	c_hac_1
31	31	0,2907	0,0597	1,8381	0,3786	NO	c_hac_1
32	32	0,1398	-0,0312	0,4611	0,2643	NO	c_hac_1
33	33	0,0769	0,0195	2,0069	0,5304	NO	c_hac_3

- 2) Se podría haber utilizado el algoritmo K-Means, es un algoritmo de clustering ampliamente utilizado para agrupar datos en K clústers. Es un método de particionamiento, busca dividir el conjunto de datos en K clústers, donde cada observación pertenece a uno de los mismos.
- 3) Sí, podríamos haber trabajado con las 4 variables cuantitativas. Clustering se puede aplicar a conjuntos de datos con cualquier número de variables, siempre y cuando los datos sean cuantitativos y se pueda calcular una medida de similitud o distancia entre los puntos de datos.

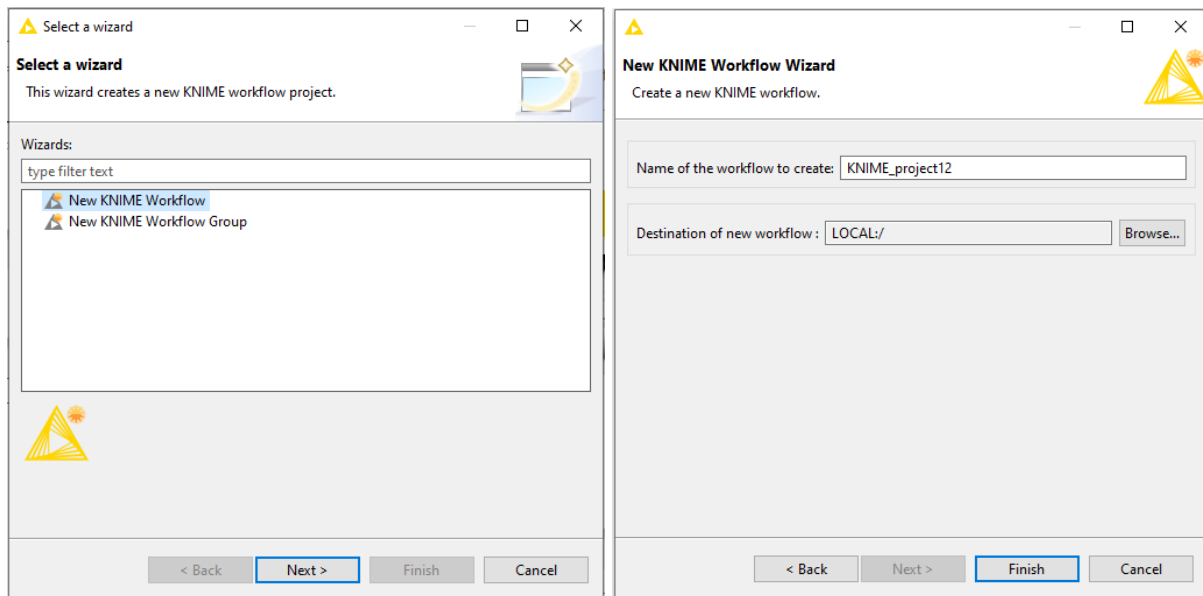
3.2) Clustering – (Knime, text_mining_clustering_1Obligatorio2023.txt)

Asumiendo que ya tenemos instalado el programa Knime dado que lo utilizamos en el primer ejercicio para convertir un archivo .txt en .arff, procederemos a explicar cómo utilizarlo para trabajar con clustering.

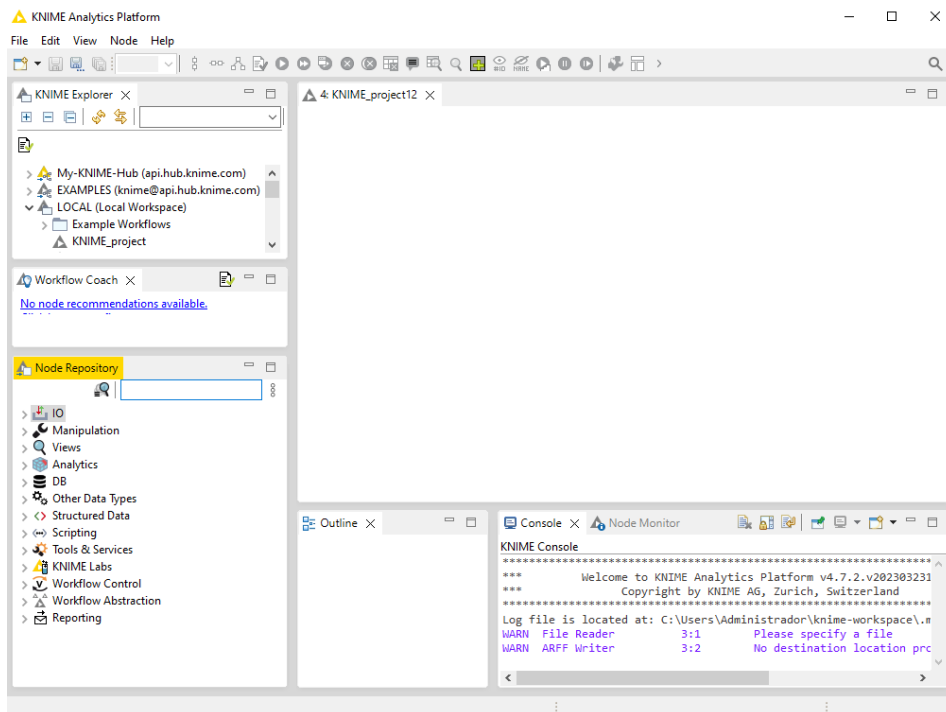
Comenzaremos abriendo Knime. Una vez abierto debemos crear un workflow, por lo que haremos clic en *File* (ubicado en la esquina superior izquierda), y posteriormente hacemos clic en la opción *New...* (indicada en azul).



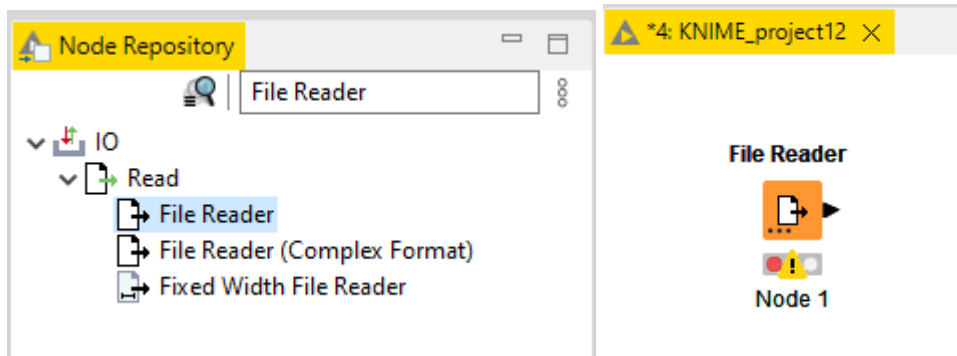
Se desplegará una nueva ventana denominada *Select a wizard*, donde seleccionaremos *New KNIME Workflow*. Una vez seleccionada la opción, marcaremos *Next*. Se abrirá otra ventana donde podemos escribir el nombre del nuevo Workflow a crear o simplemente dejar el por defecto, y cambiar la ubicación donde se guardará el mismo. Una vez realizado lo mencionado, se debe apretar el botón *Finish*.



Obtendremos lo que se observa en la siguiente imagen:



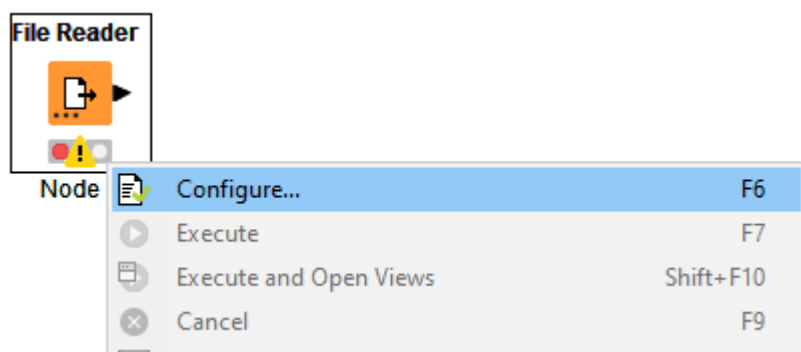
Estando posicionado en dicha pantalla, debemos crear un nodo File Reader. Utilizando el buscador de Node Repository, vamos a buscar File Reader. Una vez encontrado, hacemos doble clic encima de él (indicado en color celeste). Obteniendo como resultado el siguiente nodo:



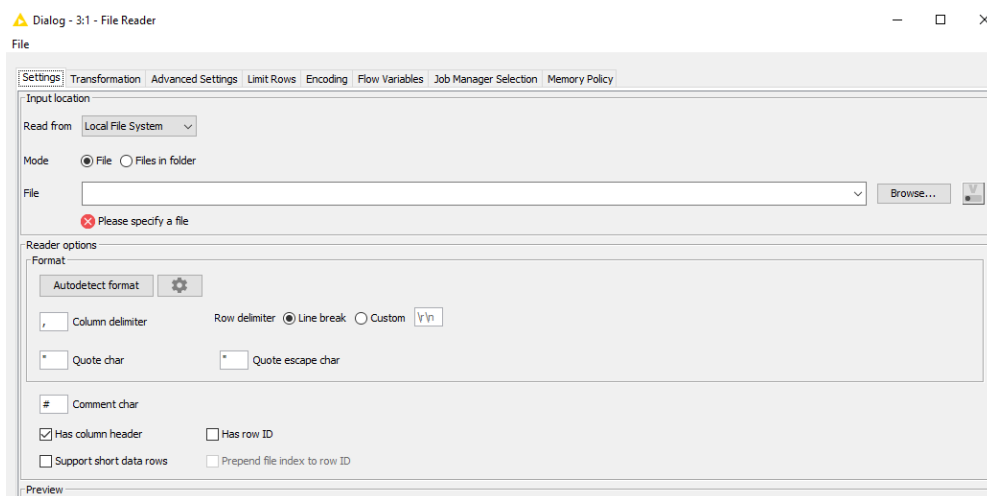
El nodo aparece, en principio, en color rojo. Esto significa que debemos configurarlo para poder ejecutarlo.

Proceso de configuración:

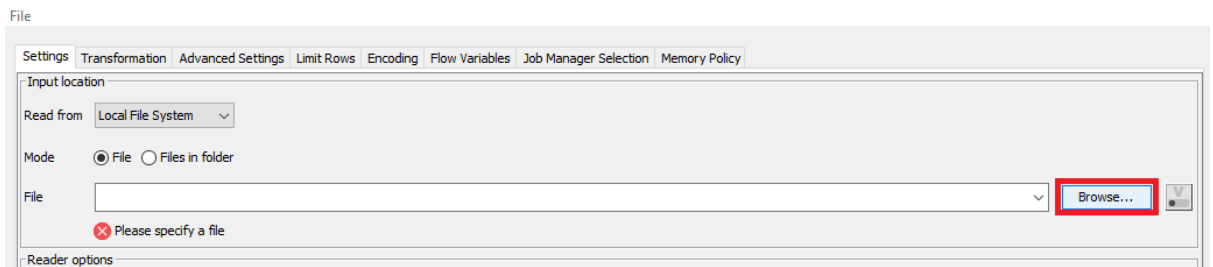
- 1) Hacer clic derecho sobre el nodo *File Reader*, seleccionar la opción *Configure...*



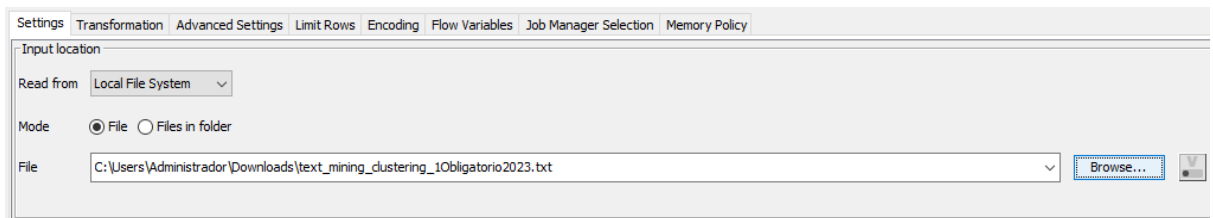
Se abrirá la siguiente ventana:



- 2) Seguidamente, hacemos clic en *Browse...*, esto se realiza para cargar el archivo deseado, *text_mining_clustering_1Obligatorio2023.txt*. Se nos abrirá un nuevo diálogo donde deberemos dirigirnos a la carpeta donde guardamos el archivo, seleccionarlo y posteriormente presionar *Abrir*.



- 3) Una vez realizado este paso, observaremos que el archivo ya aparece como cargado, como se observa en la imagen adjunta.



Además, nos permitirá ver una vista previa:

Preview


i The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	S termino...
Row0	8741344395
Row1	7760552375
Row2	7631350787
Row3	8761337676
Row4	9641338295
Row5	9741431295
Row6	9631540796
Row7	7540544885
Row8	9661432077
Row9	8631555285
Row10	7560534776
Row11	8551534295
Row12	7650534095
Row13	8631540586
Row14	7560537076
Row15	9551433695
Row16	7650435477
Row17	7550534386

Al observar la tabla de la siguiente manera, debemos cambiar el delimitador de columnas de manera inmediata. La manera más sencilla de hacerlo es haciendo clic sobre la opción *Autodetect Format* que se presenta en pantalla.

Reader options

Format

Autodetect format 

Column delimiter: , Row delimiter: ☒ Line break ☐ Custom `\r\n`

Quote char: " Quote escape char: "

Comment char: #

☒ Has column header ☐ Has row ID

☐ Support short data rows ☐ Prepend file index to row ID

Posteriormente se observarán los datos de la siguiente forma:

Preview

i The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	I termino1	I termino2	I termino3	I termino4	I termino5	I termino6	I termino7	I termino8	I termino9	I termino10
Row0	8	7	4	1	3	4	4	3	9	5
Row1	7	7	6	0	5	5	2	3	7	5
Row2	7	6	3	1	3	5	0	7	8	7
Row3	8	7	6	1	3	3	7	6	7	6
Row4	9	6	4	1	3	3	8	2	9	5
Row5	9	7	4	1	4	3	1	2	9	5
Row6	9	6	3	1	5	4	0	7	9	6
Row7	7	5	4	0	5	4	4	8	8	5
Row8	9	6	6	1	4	3	2	0	7	7
Row9	8	6	3	1	5	5	5	2	8	5
Row10	7	5	6	0	5	3	4	7	7	6
Row11	8	5	5	1	5	3	4	2	9	5
Row12	7	6	5	0	5	3	4	0	9	5
Row13	8	6	3	1	5	4	0	5	8	6
Row14	7	5	6	0	5	3	7	0	7	6
Row15	9	5	5	1	4	3	3	6	9	5
Row16	7	6	5	0	4	3	5	4	7	7
Row17	7	5	5	0	5	3	4	3	8	6
Row18	9	6	4	1	5	3	4	5	7	7
Row19	9	6	4	0	4	4	7	0	9	5
Row20	9	6	6	0	3	4	5	7	7	6
Row21	7	6	6	1	3	5	7	2	9	5

Observamos que los datos se identifican como *string* o *integer* y se encuentran correctamente clasificados.

4) Finalmente, hacemos clic en el botón OK para finalizar.

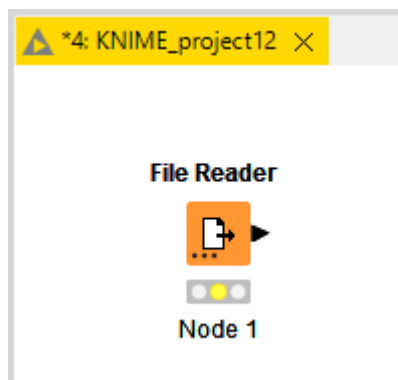
Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

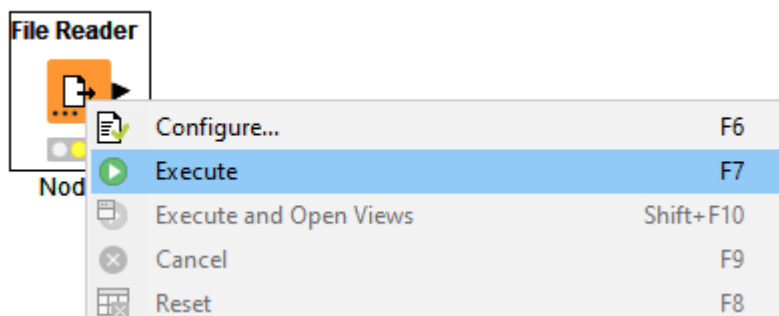
Row ID	termino1	termino2	termino3	termino4	termino5	termino6	termino7	termino8	termino9	termino10
Row0	8	7	4	1	3	4	4	3	9	5
Row1	7	7	6	0	5	5	2	3	7	5
Row2	7	6	3	1	3	5	0	7	8	7
Row3	8	7	6	1	3	3	7	6	7	6
Row4	9	6	4	1	3	3	8	2	9	5
Row5	9	7	4	1	4	3	1	2	9	5
Row6	9	6	3	1	5	4	0	7	9	6
Row7	7	5	4	0	5	4	4	8	8	5
Row8	9	6	6	1	4	3	2	0	7	7
Row9	8	6	3	1	5	5	5	2	8	5
Row10	7	5	6	0	5	3	4	7	7	6
Row11	8	5	5	1	5	3	4	2	9	5
Row12	7	6	5	0	5	3	4	0	9	5
Row13	8	6	3	1	5	4	0	5	8	6
Row14	7	5	6	0	5	3	7	0	7	6
Row15	9	5	5	1	4	3	3	6	9	5
Row16	7	6	5	0	4	3	5	4	7	7
Row17	7	5	5	0	5	3	4	3	8	6
Row18	9	6	4	1	5	3	4	5	7	7
Row19	9	6	4	0	4	4	7	0	9	5
Row20	9	6	6	0	3	4	5	7	7	6
Row21	7	6	6	1	3	5	7	2	9	5
Row22	7	6	6	0	3	3	1	6	8	7

OK Apply Cancel ?

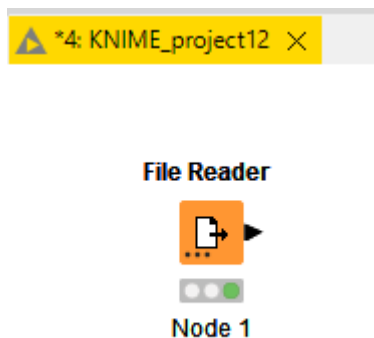
Luego de finalizado, el nodo pasa a ser de esta forma:



Se observa que ahora se encuentra en color amarillo. Por lo que ahora debemos ejecutar el nodo. Para ejecutar el nodo debemos hacer clic derecho sobre el mismo y luego hacer clic sobre *Execute*.

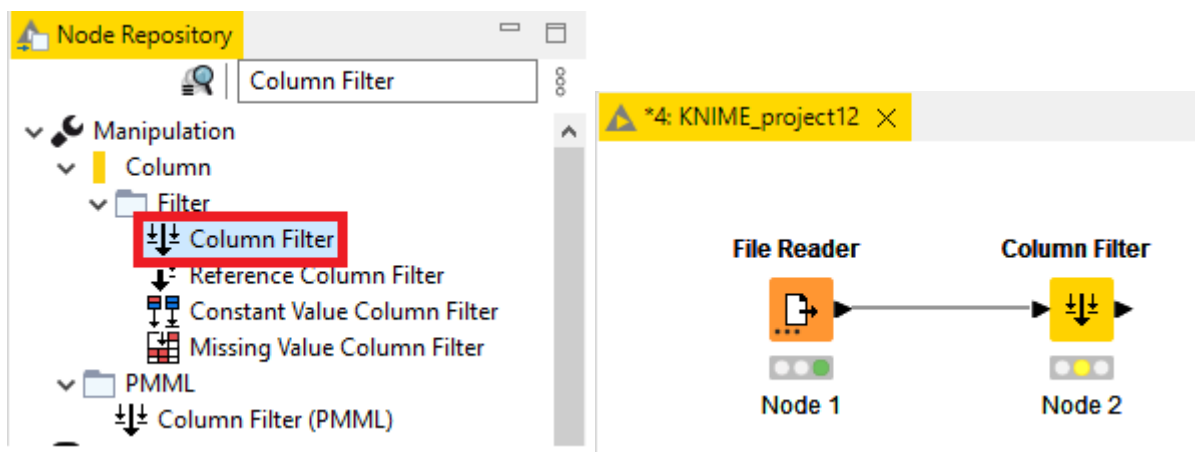


Una vez ejecutado podremos observar que si la ejecución fue exitosa pasará a tener color verde.

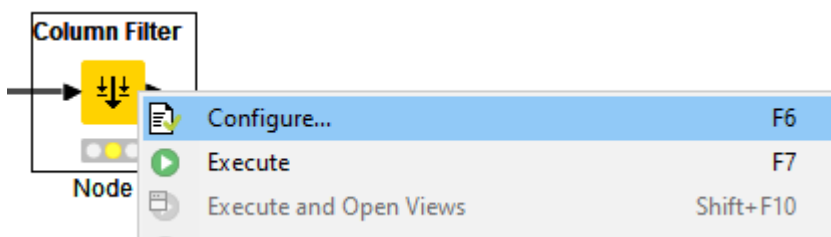


Para continuar seleccionaremos dos atributos con los que trabajar. En nuestro caso utilizaremos únicamente término1 y término3.

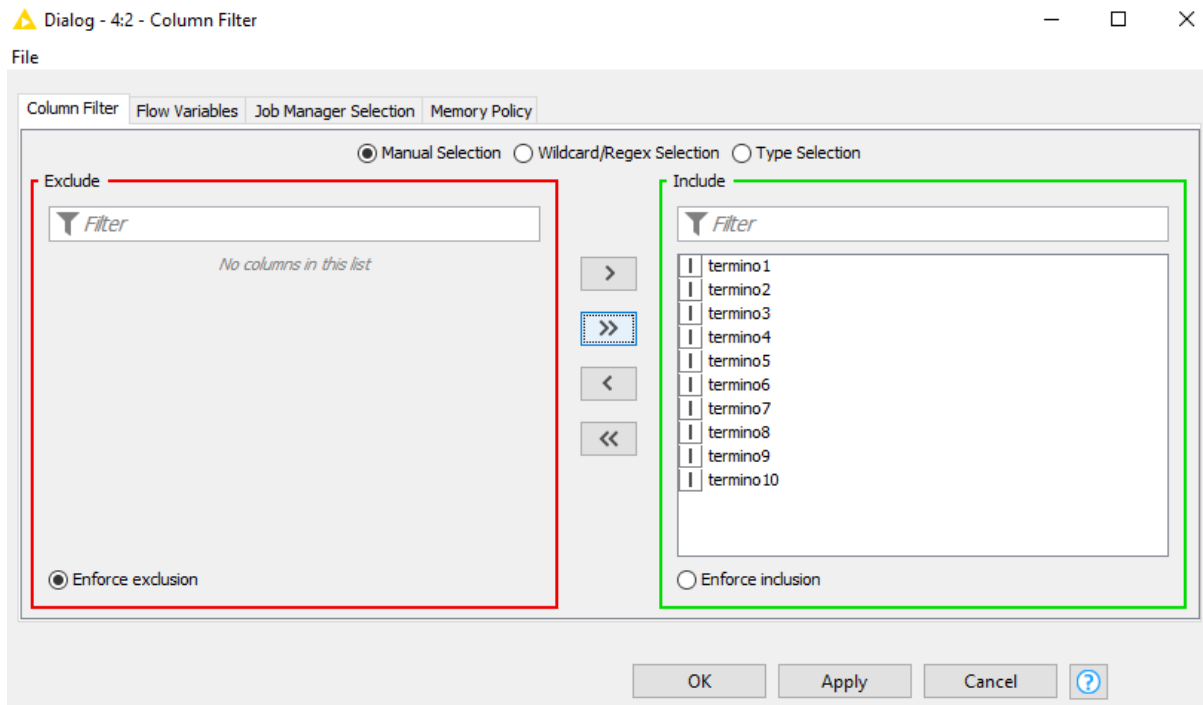
Para esto, crearemos un nodo *Column Filter*. Nos vamos a dirigir nuevamente al buscador de *Node Repository* y escribiremos *Column Filter*. Una vez lo encontremos, lo arrastraremos a la sección de trabajo y lo relacionaremos con el File Reader ya colocado.



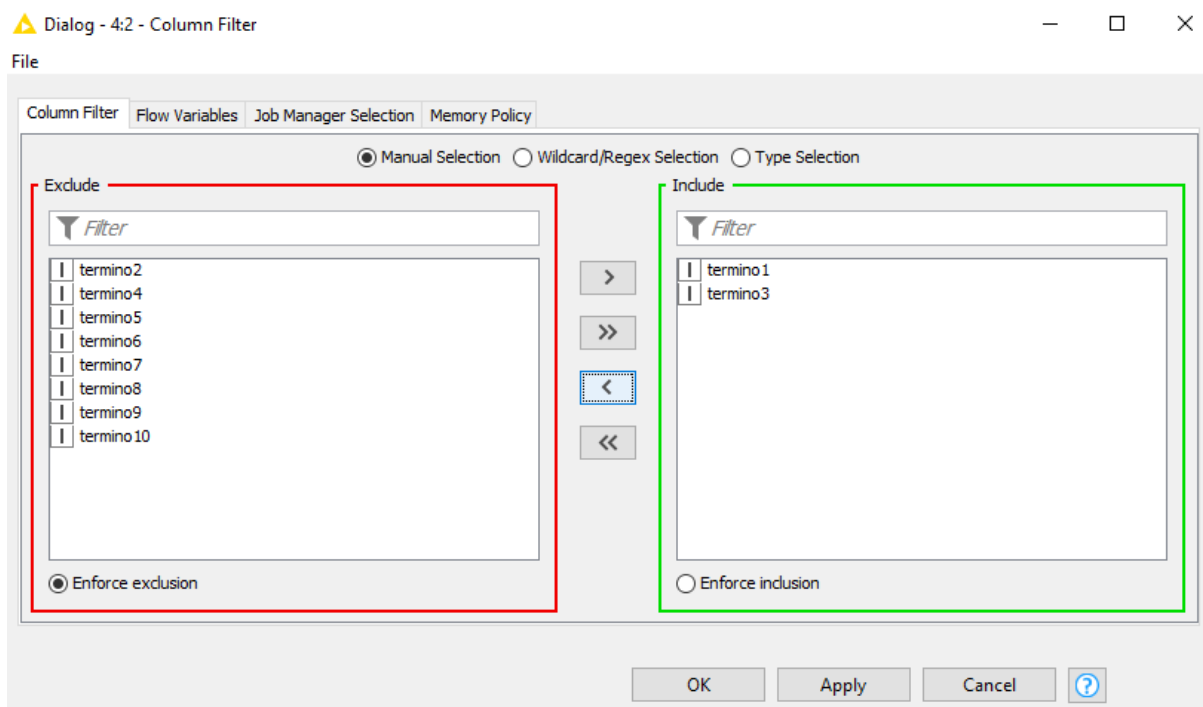
Una vez están relacionados, procederemos a configurar el nuevo nodo. Para esto, haremos clic derecho sobre el mismo y luego haremos clic en *Configure*.



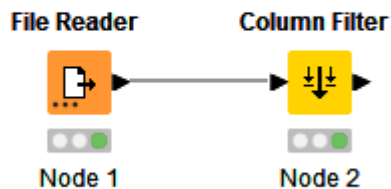
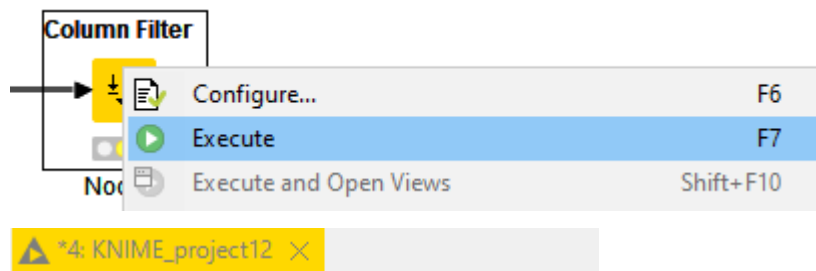
Se observará la siguiente pantalla:



En la misma deberemos seleccionar todos los atributos que no utilizaremos y desplazarlos hacia la izquierda, al cuadrante rojo que los excluye. Al finalizar, quedará de la siguiente manera:



Continuamos haciendo clic en Ok, y posteriormente clic derecho sobre el nodo y clic en Execute.



Ya tenemos dos nodos configurados correctamente. Si deseamos ver cómo quedó la tabla después de utilizar el nodo Column Filter, podemos hacer clic derecho sobre el mismo y seleccionar la última opción: *Filtered table*.

Filtered table - 4:2 - Column Filter

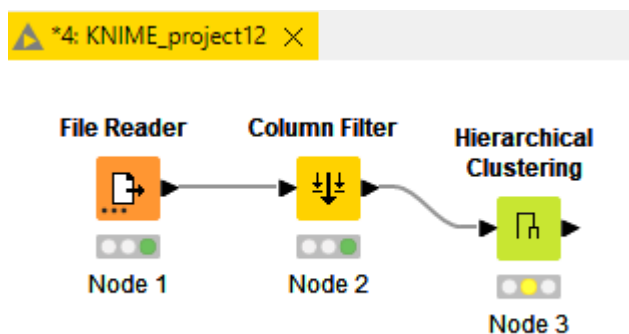
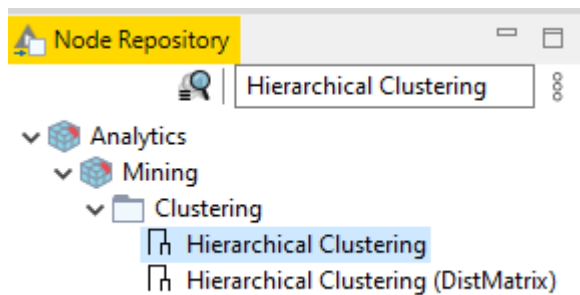
File Edit Hilite Navigation View

Table "default" - Rows: 150 Spec - Columns: 2 Properties Flow Variables

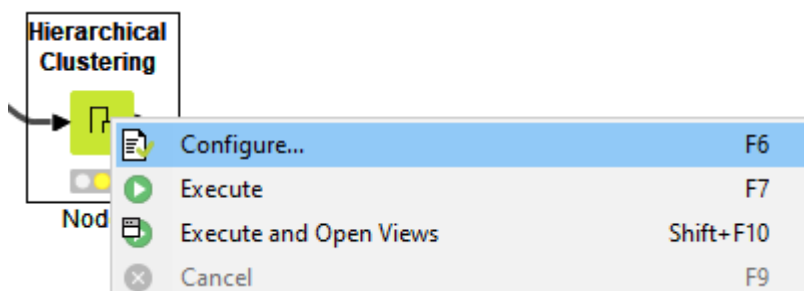
Row ID	termino1	termino3
Row0	8	4
Row1	7	6
Row2	7	3
Row3	8	6
Row4	9	4
Row5	9	4
Row6	9	3
Row7	7	4
Row8	9	6
Row9	8	3
Row10	7	6
Row11	8	5
Row12	7	5
Row13	8	3
Row14	7	6
Row15	9	5
Row16	7	5
Row17	7	5
Row18	9	4
Row19	9	4
Row20	9	6
Row21	7	6
Row22	7	6
Row23	8	6
Row24	7	5

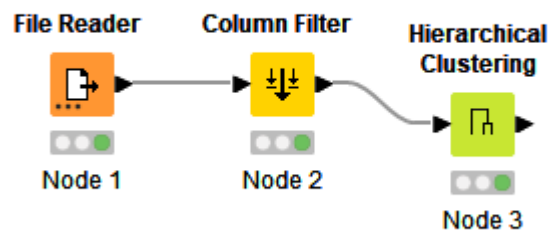
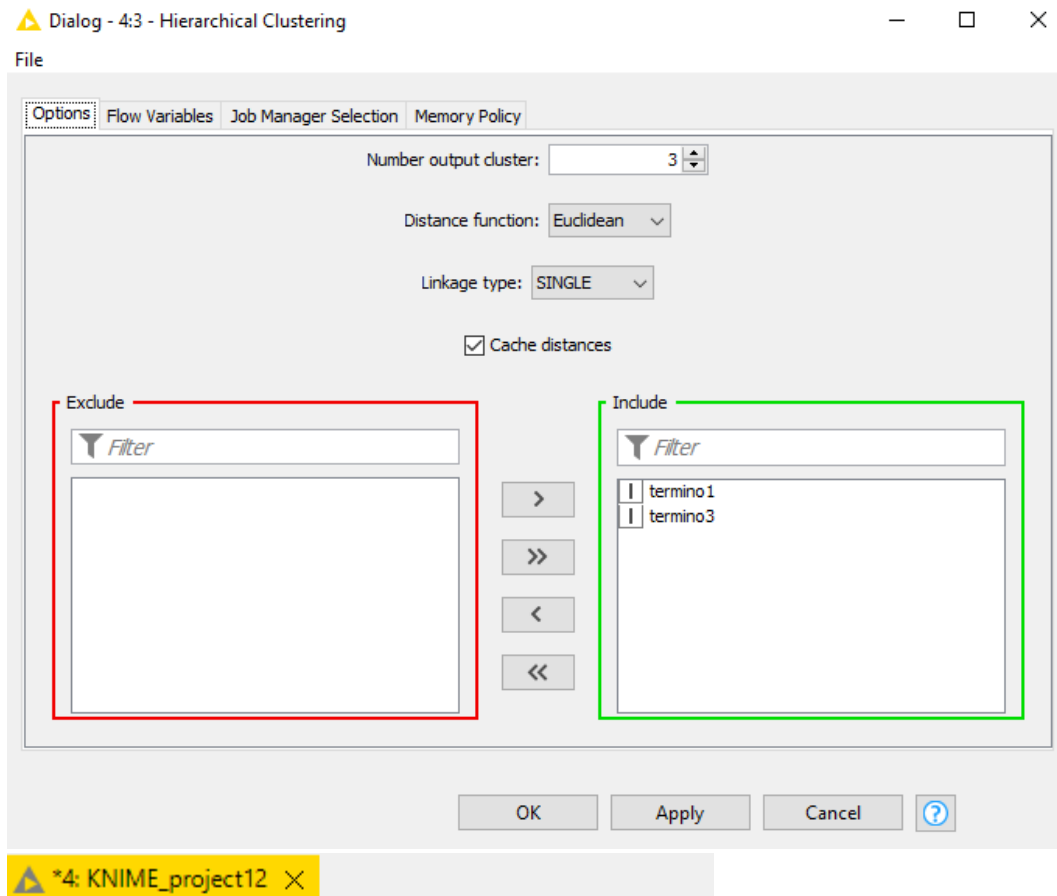
Como se observa, solo se encuentran las columnas que seleccionamos para trabajar. Ahora, estamos listos para continuar con la función de clustering.

Nos dirigimos nuevamente al buscador de *Node Repository*, y en este caso buscaremos *Hierarchical Clustering*, lo seleccionamos y arrastramos al workspace, relacioándolo con *Column Filter*.

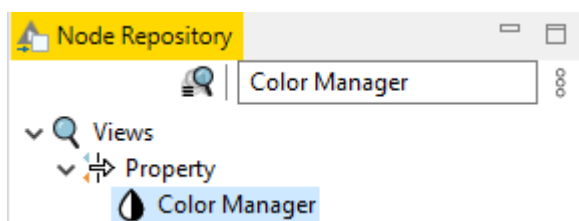


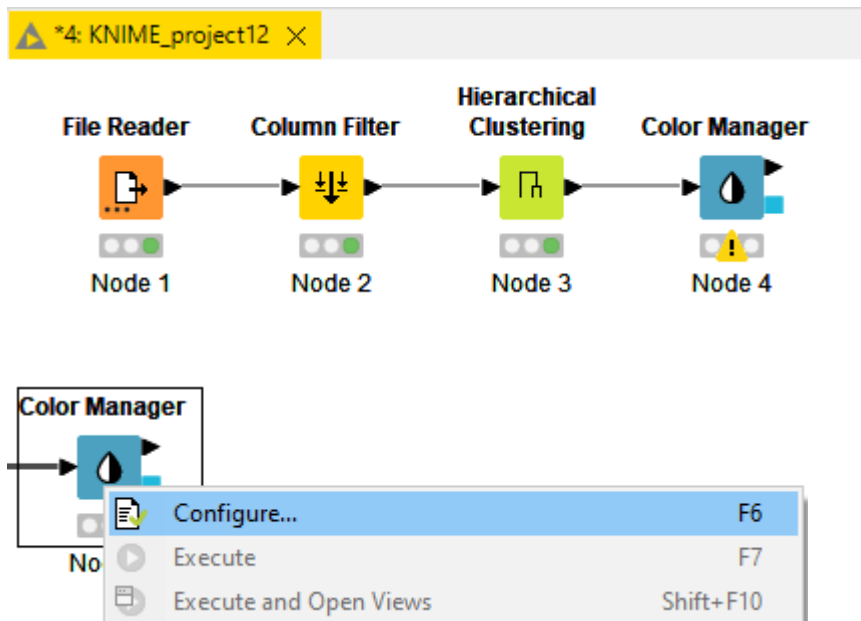
Para continuar debemos configurar el nodo Hierarchical Clustering, para esto haremos clic derecho sobre el mismo y seleccionaremos la opción *Configure*.



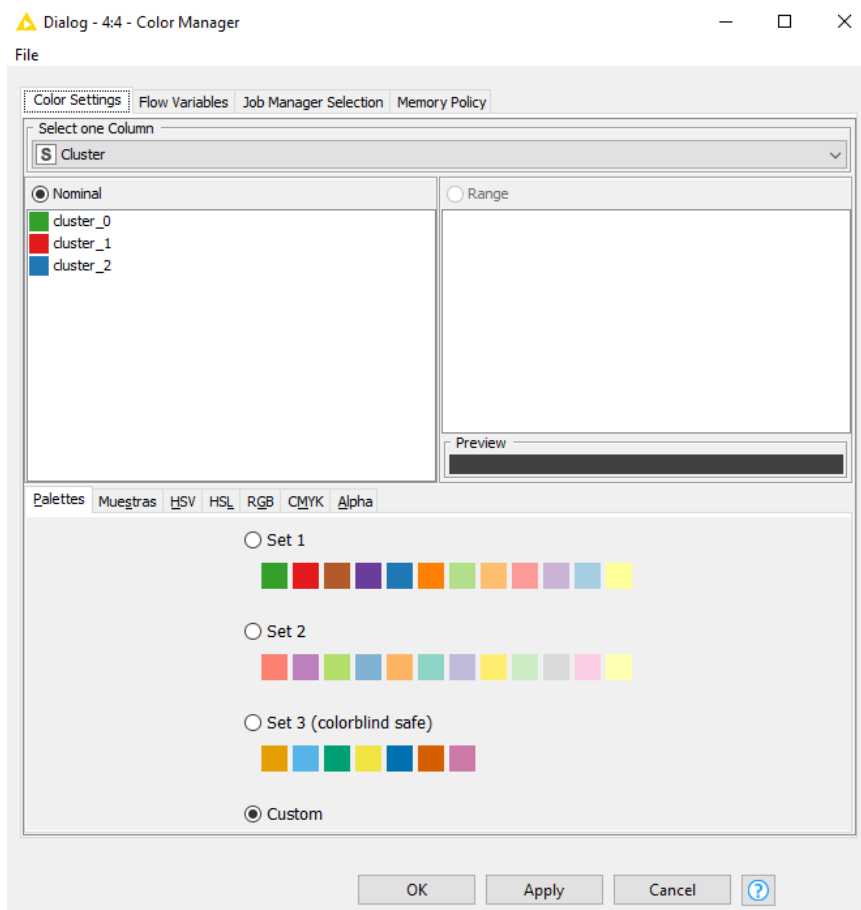


Para identificar los clusters con colores podemos utilizar el nodo Color Manager. Al igual que con los anteriores, procedemos a buscarlos en el buscador de Node Repository. Lo seleccionamos y arrastramos al workspace relacionándolo con Hierarchical Clustering.

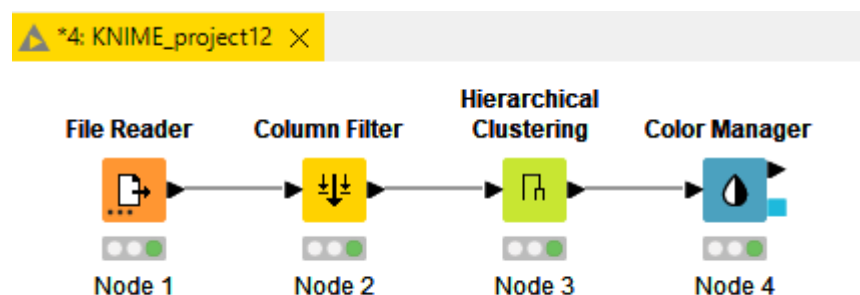
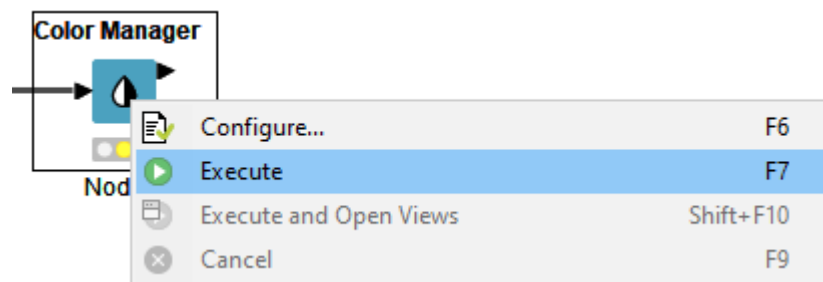




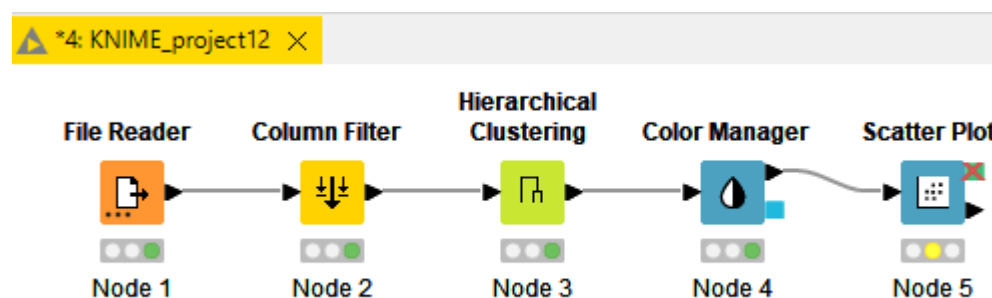
Una vez estén relacionados los nodos, procedemos a configurar este último. Para realizar la configuración correspondiente, hacemos clic derecho sobre *Color Manager* y clic en *Configure...*



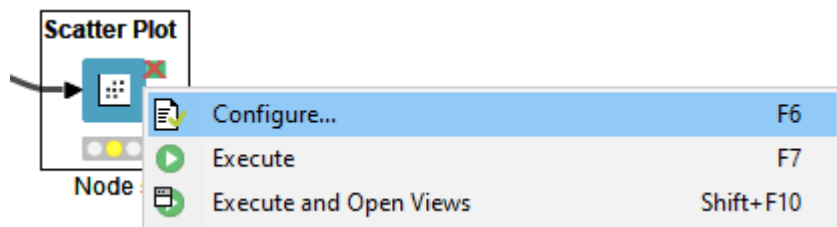
Asignamos el color deseado a cada cluster y hacemos clic en *Ok*. Posteriormente, volvemos a hacer clic derecho en el nodo y clic en *Execute*.



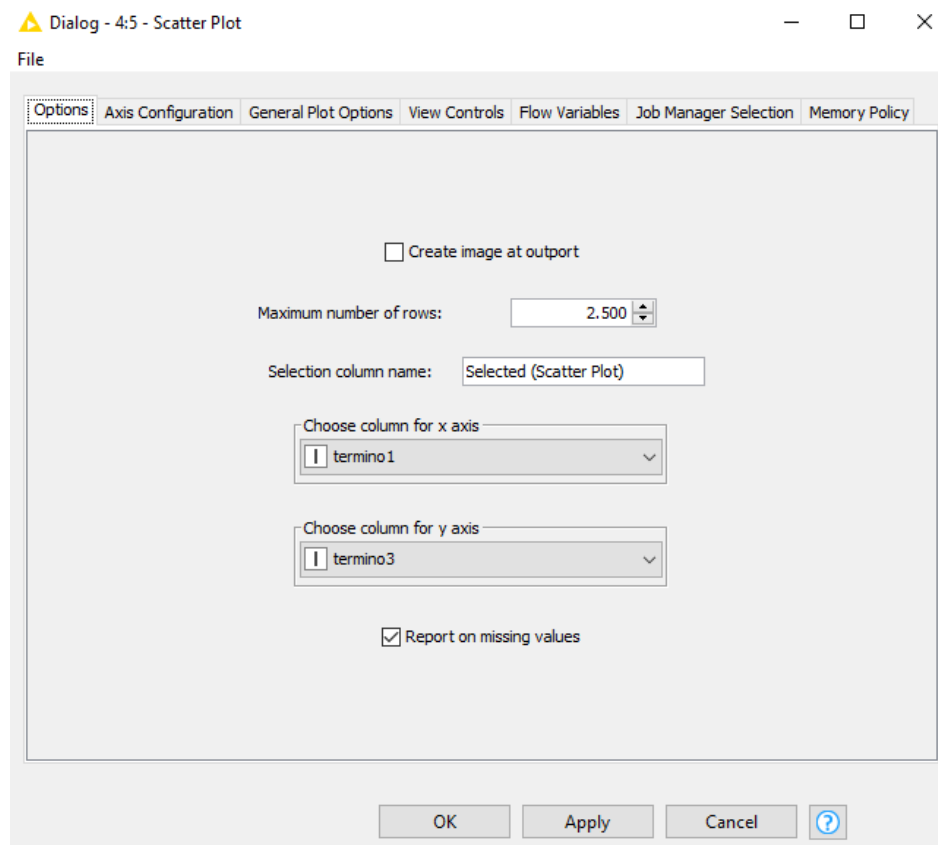
Por último, para poder visualizar los clusters, haremos un nodo *Scatter Plot*. Lo buscamos en el buscador de Node Repository, y una vez lo encontremos, lo relacionamos con Color Manager (triángulo con triángulo).



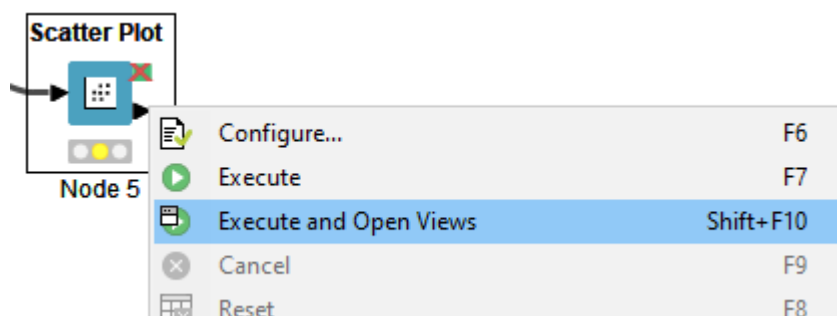
Para configurar el nodo hacemos clic derecho sobre el mismo y clic en *Configure...*

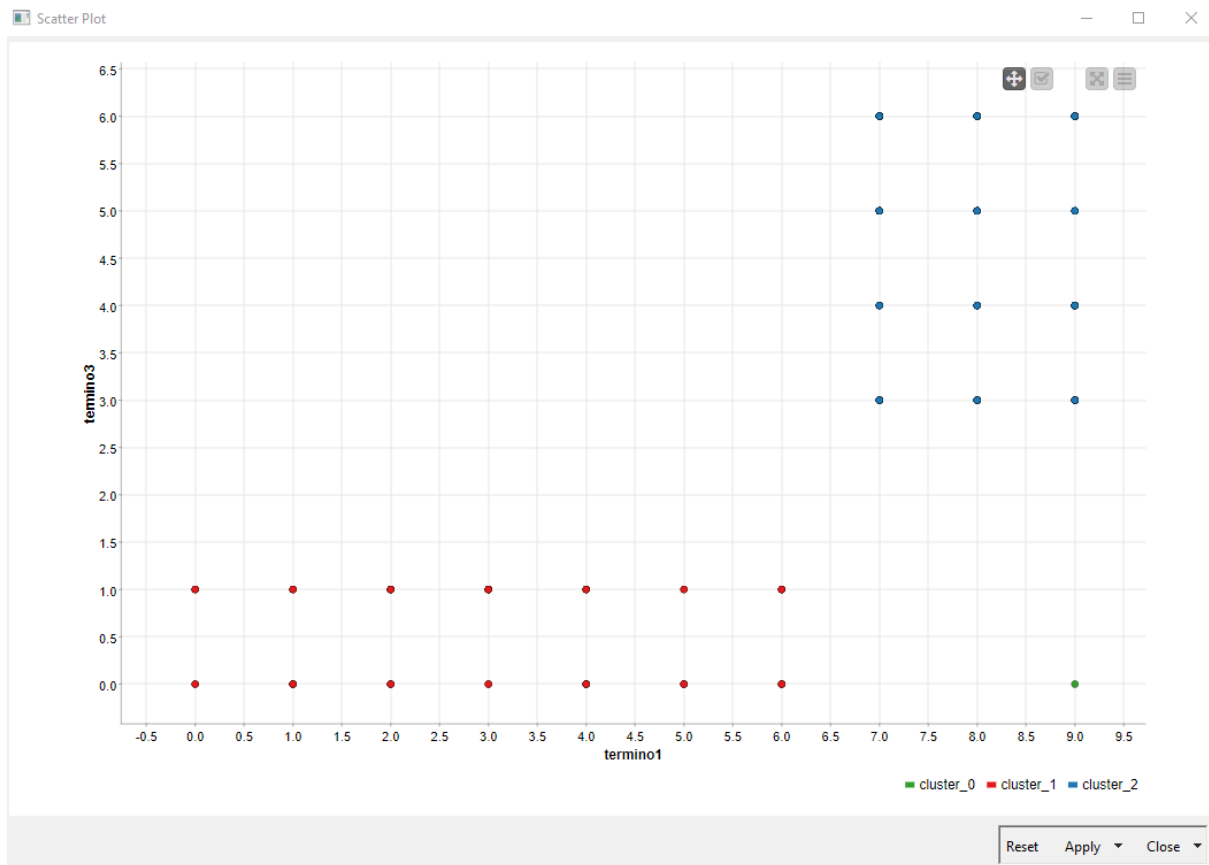


En la configuración verificamos que se encuentren termino1 en **x** y termino3 en **y**. De ser así, hacemos clic en *Ok*. En caso de no ser así, los intercambiamos.



Por último, al igual que con todos los nodos, hacemos clic derecho sobre él y lo ejecutamos haciendo clic en *Execute and Open Views*.





Relación con los negocios

Entendiendo que el archivo lleva la frecuencia con la que aparecen algunos términos en un texto dado, puede utilizarse como medida de similitud entre documentos o para identificar patrones o temas comunes en un conjunto de datos. Por ejemplo, si dos documentos tienen una alta frecuencia de co-ocurrencia de términos relacionados con “ventas” y “marketing”, es posible que pertenezcan a un mismo grupo relacionado con estrategias de ventas y marketing.

Preguntas:

- 1) ¿Qué es un diagrama de dispersión?
- 2) ¿Cuál es la principal diferencia entre un diagrama de dispersión y un dendrograma?
- 3) ¿Por qué en este caso utilizamos variables cuantitativas?

Respuestas:

- 1) Un diagrama de dispersión en el contexto de clustering es una representación gráfica que muestra la distribución de los puntos de datos en un espacio bidimensional o tridimensional. Cada punto representa una instancia o un elemento de datos y su posición está determinada por sus características o atributos. Puede ser útil para visualizar y comprender la estructura de los datos y la separación de los grupos o clusters identificados por el algoritmo.
- 2) La principal diferencia radica en la forma en que representan la estructura de los clusters y la relación entre ellos. Un dendrograma es un tipo de diagrama de árbol que muestra la jerarquía de agrupamiento en un algoritmo de clustering jerárquico; agrupando los datos en forma de una estructura de árbol. En cambio, el diagrama de dispersión no representa explícitamente la jerarquía de los clusters, se centra en la proximidad o separación de los puntos de datos.
- 3) Porque trabajamos con algoritmos que se basan en cálculos de distancias, y para eso necesitan que las variables sean numéricas.

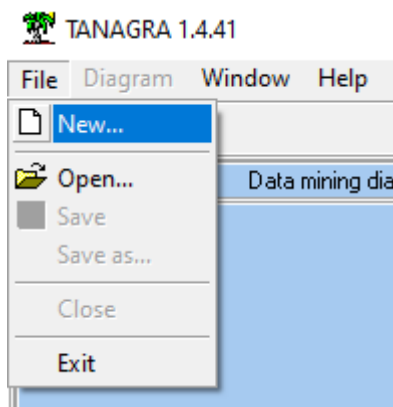
Ejercicio 4)

Presente aplicaciones de redes neuronales utilizando dos herramientas.

4.1) Redes Neuronales – (Tanagra, empresasObligatorio2023.txt)

Asumiendo que ya tenemos el programa Tanagra instalado y el archivo descargado por ejercicios que fuimos realizando anteriormente. Procederemos a detallar el tutorial correspondiente al proceso de creación de la red neuronal.

Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo Dataset (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo empresasObligatorio2023.txt en la ubicación donde fue guardado.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :

OK Cancel Help

Una vez seleccionado el archivo se nos cargará en el campo Dataset la ruta de este.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :
C:\Users\Administrador\Downloads\empresasObligatorio2023.txt

OK Cancel Help

Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:

TANAGRA 1.4.41 - [Dataset (empresasObligatorio2023.txt)]

File Diagram Component Window Help

Default title

Dataset (empresasObligatorio2023.txt)

Dataset (empresasObligatorio2023.txt)

Parameters

Database : C:\Users\Administrador\Downloads\empresasObligatorio2023.txt

Results

Download information

Datasource processing

Computation time 0 ms

Allocated memory 9 KB

Dataset description

6 attribute(s)

100 example(s)

Attribute	Category	Informations
id	Continue	-
FlujodeCaja/DeudaTotal	Continue	-
IngresoNeto/ActivoTotal	Continue	-
ActivoCorriente/PasivoCorriente	Continue	-
ActivoCorriente/VentasNetas	Continue	-
Quiebra	Discrete	2 values

Computation time : 0 ms.

Created at 28/6/2023 2:45:10

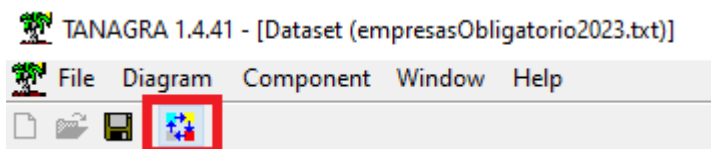
En esta parte, debemos verificar que todos los atributos excepto el denominado Grupo (que es de tipo Discrete) sean de tipo Continue. Siendo que, como ya hemos mencionado, en Tanagra estas categorizaciones se corresponden de la siguiente manera:

- **Discreta:** Es el equivalente a cualitativa
- **Continua:** Es el equivalente a cuantitativa

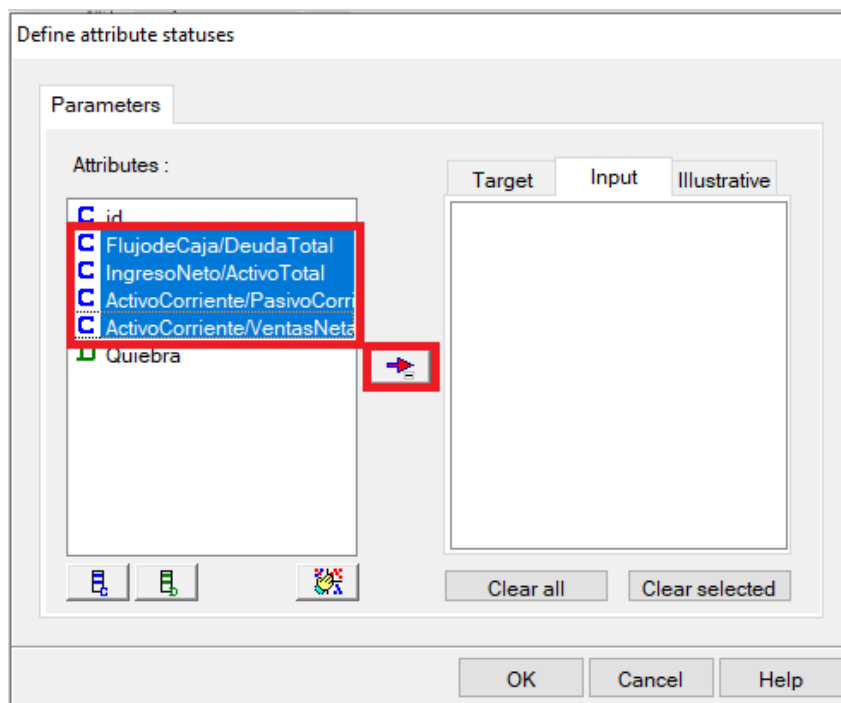
En caso de que las categorizaciones de los atributos estén invertidas a como se describió, debemos solucionarlo cambiando en el archivo las “,” (comas) por “.” (puntos).

Para continuar, debemos tener en cuenta que antes de aplicar un algoritmo vamos a tener que seleccionar los datos con los que trabajaremos, por lo que, haremos un *Define Status*, haciendo los pasos que se muestran a continuación:

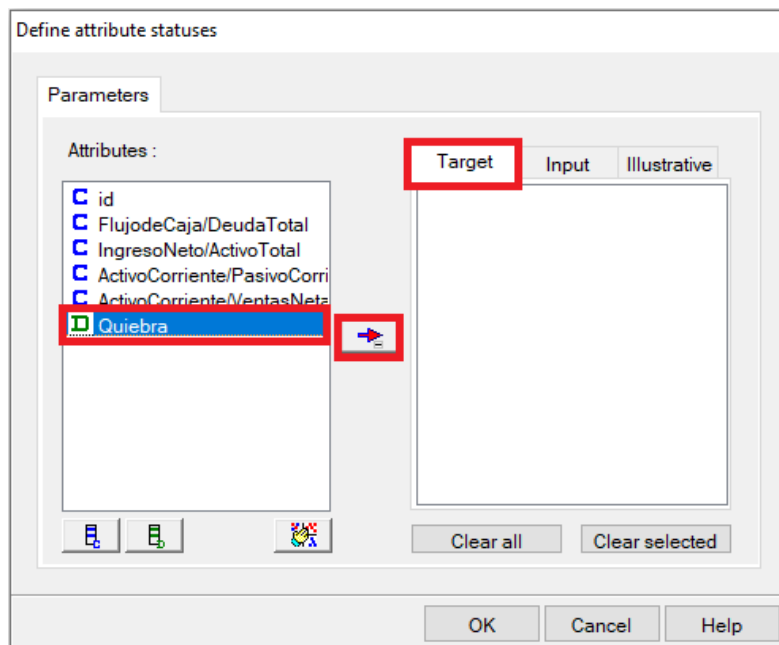
- 1) Debemos hacer clic en el ícono marcado en rojo en la siguiente imagen:



- 2) Se abrirá una ventana denominada *Define attribute statuses*, donde debemos seleccionar todos los atributos que son ratios financieros y presionar la flecha indicada para colocarlas en el *input*. Observar que en el cuadrante de la derecha esté indicada la opción input.

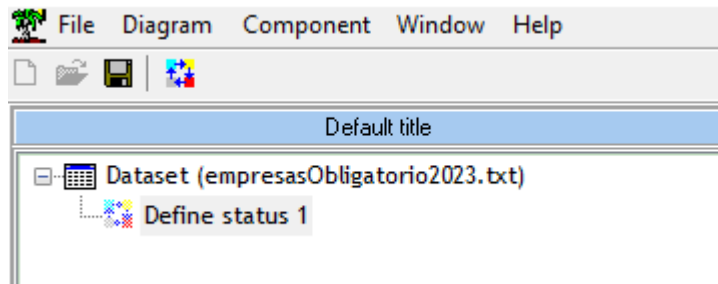


- 3) Una vez los atributos marcados se encuentren del lado derecho, debemos indicar el atributo Target, es decir, nuestra variable objetivo. En este caso será Quiebra. Por lo que, siguiendo el procedimiento anterior, seleccionaremos dicho atributo y lo desplazaremos hacia la derecha. Previamente debemos seleccionar Target, ya de que no hacerlo, pondríamos al atributo en Input.



El atributo id queda por fuera ya que es únicamente un identificador y no aporta información.

- 4) Una vez realizados los dos pasos anteriores, se debe hacer clic sobre el botón OK. Luego, deberemos hacer clic sobre “Define status 1”, que se encontrará colgado debajo del dataset, con esto lo ejecutaremos.



Después de ejecutar, si deseamos verificar lo hecho hasta el momento, en la tabla que se visualiza deberá aparecer el valor yes en la columna input para todos los ratios financieros indicados y en la columna Target para el atributo Quiebra.

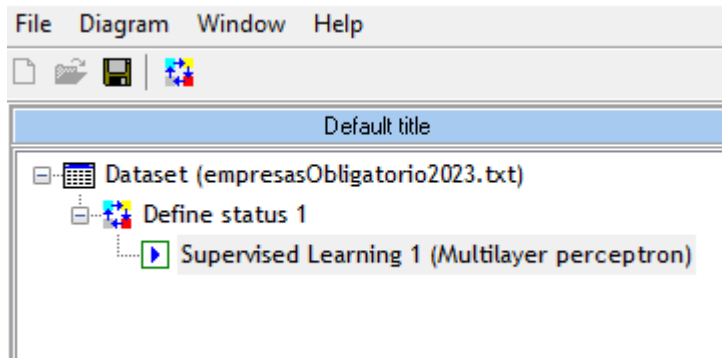
Results			
Attribute	Target	Input	Illustrative
id	-	-	-
FlujodeCaja/DeudaTotal	-	yes	-
IngresoNeto/ActivoTotal	-	yes	-
ActivoCorriente/PasivoCorriente	-	yes	-
ActivoCorriente/VentasNetas	-	yes	-
Quiebra	yes	-	-

Habiendo finalizado con la selección de los datos correspondientes (mismo proceso que se ha realizado en ejercicios anteriores), podemos proceder con la explicación de cómo llevar a cabo la red neuronal.

Para esto, nos vamos a dirigir a *Components*, en la sección *Spv learning*, y seleccionaremos *Multilayer perceptron*. Una vez esté seleccionado, lo arrastramos debajo del define status realizado anteriormente.

Components			
Instance selection	Feature construction	Feature selection	Regression
Spv learning	Meta-spv learning	Spv learning assessment	Scoring

List	K-NN	Multilayer perceptron	Naive bayes con
	Linear discriminant analysis	Multinomial Logistic Regression	PLS-DA
	Log-Reg TRIRLS	Naive bayes	PLS-LDA



Una vez esté colgado, con los parámetros por defecto, procedemos a ejecutarlo haciendo doble clic sobre *Supervised Learning 1 (Multilayer perceptron)*

Se observará la siguiente pantalla:

Supervised Learning 1 (Multilayer perceptron)

Parameters

MLP architecture	
Use hidden layer	yes
Neurons in the hidden layer	10

Learning parameters	
Validation set proportion	0,20
Learning rate	0,15
Attribute transformation	standardized

Stopping rule	
Max iteration	100
Error rate threshold	0,0100
Verify error stagnation	no

Results

Classifier performances

Error rate		0,0800	
Values prediction		Confusion matrix	
Value	Recall	1-Precision	
NO	0,9623	0,1053	
SI	0,8723	0,0465	

	NO	SI	Sum
NO	51	2	53
SI	6	41	47
Sum	57	43	100

Matriz de confusión

Results

Classifier performances

Error rate			0,0800			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		NO	SI	Sum
NO	0,9623	0,1053	NO	51	2	53
SI	0,8723	0,0465	SI	6	41	47
			Sum	57	43	100

Se puede observar que de un total de 100 casos totales, 92 de ellos fueron clasificados correctamente (se encuentran en la diagonal principal), los restantes 8 casos se clasificaron incorrectamente.

Relación con los negocios

Una red neuronal que predice si una empresa quiebra o no puede ser una herramienta valiosa para los negocios en varios aspectos.

Toma de decisiones estratégicas: La predicción de quiebra puede ayudar a los líderes empresariales a tomar decisiones estratégicas informadas. Si la red neuronal indica que una empresa tiene un alto riesgo de quiebra, los directivos pueden tomar medidas preventivas, como ajustar las operaciones, buscar fuentes de financiamiento adicionales o explorar nuevas oportunidades.

Gestión de riesgos: Las empresas pueden utilizar esta herramienta para evaluar el riesgo asociado con sus clientes, proveedores o socios comerciales.

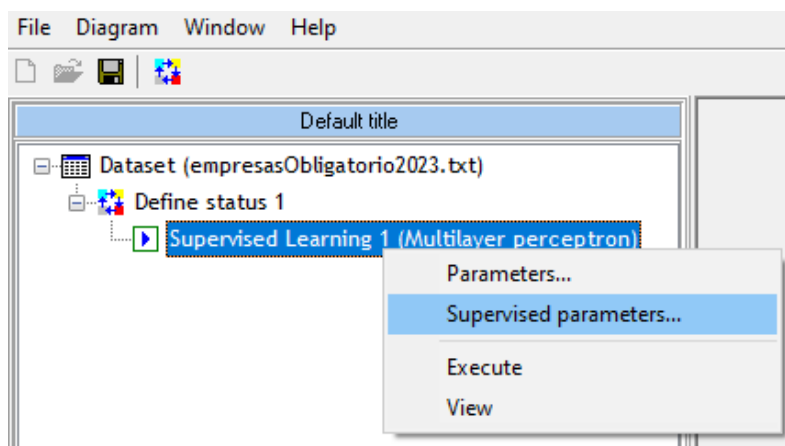
Evaluación de inversiones: Los inversores pueden utilizar la predicción de quiebra para evaluar la viabilidad financiera de una empresa en la que estén considerando invertir.

Preguntas

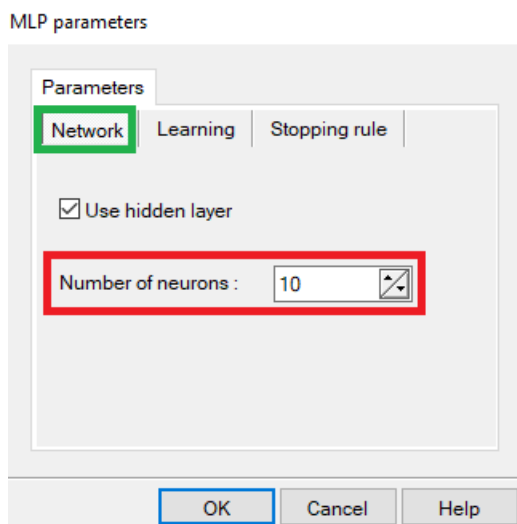
- 1) ¿Qué cantidad de neuronas tiene la capa oculta?
- 2) ¿Cuántas de las empresas que quebraron quedaron bien clasificadas? ¿Y de las que no quebraron?
- 3) ¿Para qué tipo de problemas se utiliza el algoritmo Multilayer Perceptron?

Respuestas

- 1) Para conocer la cantidad de neuronas que tiene la capa oculta debemos dirigirnos a *Supervised Learning 1 (Multilayer perceptron)* y hacer clic derecho sobre el mismo. Al desplegarse las opciones seleccionamos *Supervised parameters...*



Si nos dirigimos a la sección Network (indicada en color verde), podremos ver que se presentan 10 neuronas (se indica en color rojo).



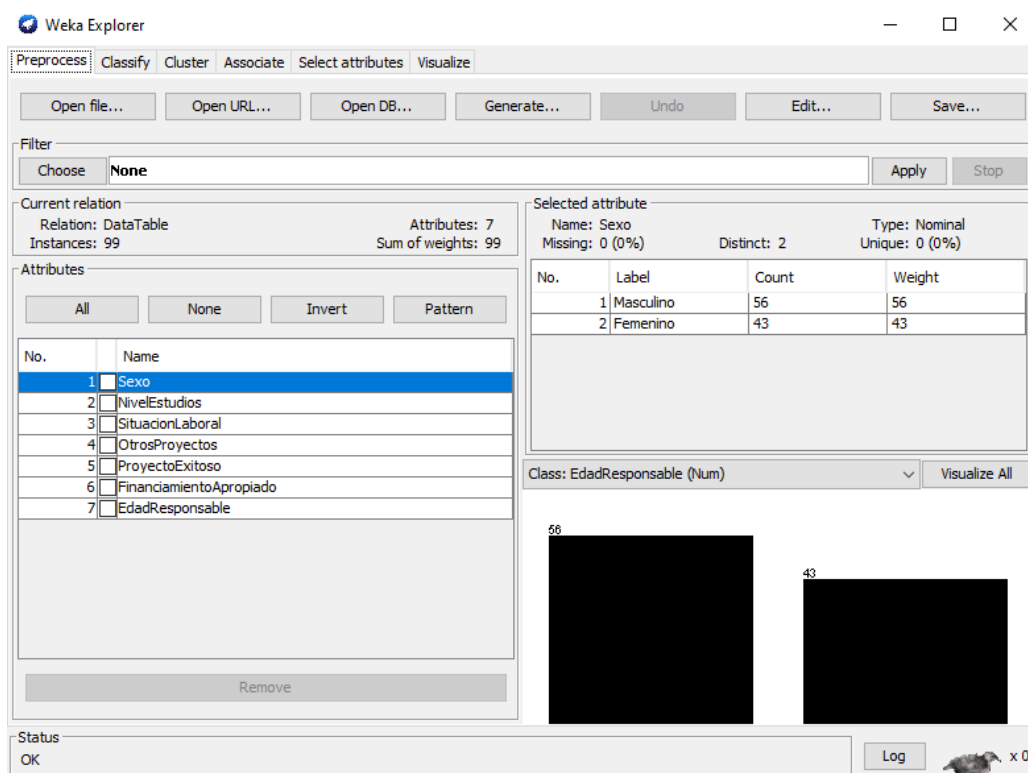
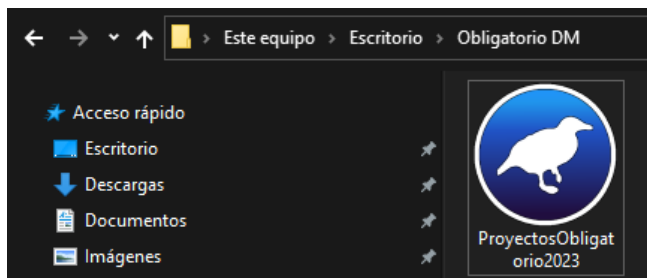
- 2) Para responder a esta pregunta debemos visualizar la matriz de confusión. Recordando que los bien clasificados se encuentran en la diagonal principal, podemos observar que de las 47 empresas que quebraron, 41 de ellas fueron bien clasificadas. Con respecto a las que no quebraron, del total de 53 empresas, 51 fueron bien clasificadas.

- 3) El Multilayer Perceptron (MLP) es un algoritmo de aprendizaje automático utilizado en la biblioteca Tanagra. En dicho contexto se utiliza para problemas de clasificación y regresión. Es un tipo de red neuronal artificial con múltiples capas ocultas. Cada capa contiene nodos (neuronas) que están conectados con pesos ajustables. Estos pesos se aprenden durante el entrenamiento del modelo para lograr una buena capacidad de generalización.

4.2) Redes Neuronales – (Weka, ProyectosObligatorio2023.txt)

Asumiendo que ya tenemos instalado el programa Weka y hemos transformado el archivo ProyectosObligatorio2023.txt a .arff utilizando la herramienta Knime (se explica en el primer ejercicio), continuaremos con el paso a paso.

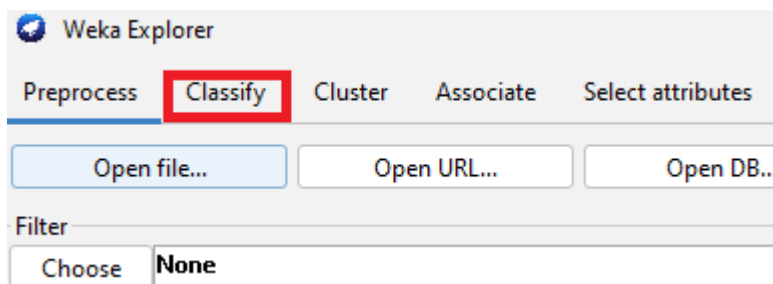
En esta ocasión, dado que ya hemos explicado de dos maneras distintas cómo abrir un archivo, queda a elección del lector el cómo hacerlo. En nuestro caso, optaremos por la manera que consideramos más sencilla, siendo buscar el archivo en nuestra biblioteca y hacer doble clic sobre el mismo.



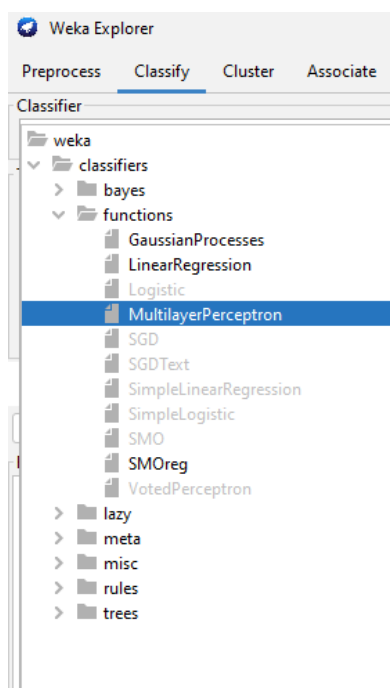
Si recorremos atributo por atributo podremos observar la cantidad de datos que hay de cada uno. Por ejemplo, en la imagen adjunta se observa que hay 56 personas del sexo masculino y 43 del sexo femenino, siendo el atributo “Sexo” el que está indicado. Si observamos detenidamente en la información que nos brinda Weka, podemos determinar que identifica a todos los atributos como tipo Nominal, a excepción del último, *EdadResponsable* que lo identifica como Numeric.

Para continuar, utilizaremos el algoritmo Multilayer Perceptron, siguiendo los pasos que se indican a continuación:

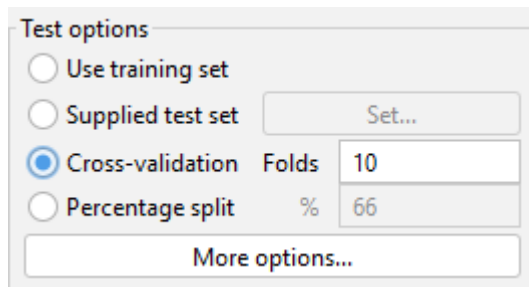
- 1) Dirigirse a la opción ubicada en la parte superior izquierda llamada *Classify* (opción indicada en rojo)



- 2) Luego haremos clic en *Choose*, para seleccionar el algoritmo con el que trabajaremos. Al desplegarse las distintas carpetas, seleccionaremos *functions* y posteriormente *Multilayer Perceptron*.



- 3) Por último, debemos verificar que en la opción que se encuentra marcada por defecto (“cross-validation”) en Test options tenga valor “Folds” en 10.



Test options

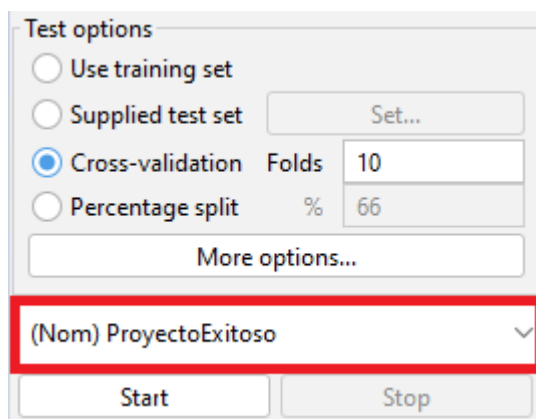
☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

Seguidamente, debemos confirmar que la variable objetivo con la que estamos trabajando es ProyectoExitoso, esto lo podemos ver en la siguiente imagen.



Test options

☐ Use training set

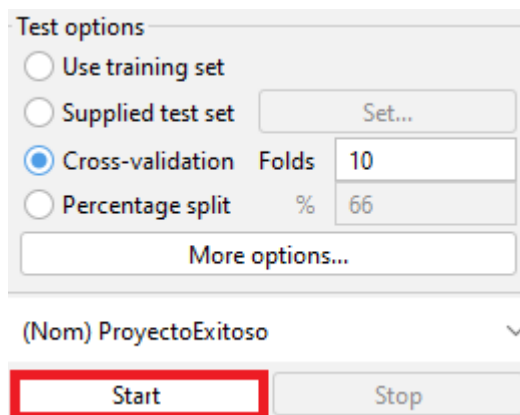
☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Nom) ProyectoExitoso

Finalmente, en la pantalla principal, presionamos el botón Start para realizar la ejecución.



Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Nom) ProyectoExitoso

Luego de finalizados los pasos de la parte anterior, podemos observar la matriz de confusión y otros datos de la clasificación.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a**

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds **10**

☐ Percentage split % **66**

More options...

(Nom) ProyectoExitoso

Start Stop

Result list (right-click for options)

23:05:00 - functions.MultilayerPerceptron

Classifier output

Input

Node 1

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	85	85.8586 %
Incorrectly Classified Instances	14	14.1414 %
Kappa statistic	0.7087	
Mean absolute error	0.202	
Root mean squared error	0.3665	
Relative absolute error	41.3045 %	
Root relative squared error	74.1077 %	
Total Number of Instances	99	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,895	0,190	0,864	0,895	0,879	0,709	0,886	0,910	SI
	0,810	0,105	0,850	0,810	0,829	0,709	0,886	0,859	NO
Weighted Avg.	0,859	0,154	0,858	0,859	0,858	0,709	0,886	0,888	

=== Confusion Matrix ===

a b <-- classified as

51 6 | a = SI

8 34 | b = NO

Matriz de confusión

```
=== Confusion Matrix ===  
  a  b  <-- classified as  
51  6 | a = SI  
 8 34 | b = NO
```

Analizando la matriz de confusión podemos observar como en la diagonal principal se encuentran aquellos datos que fueron bien clasificados, siendo un total de 85. Dentro de ellos vemos como 51 en realidad fueron exitosos y otros 34 no lo fueron. Luego en la restante diagonal podemos apreciar que 14 casos fueron mal clasificados.

Relación con los negocios

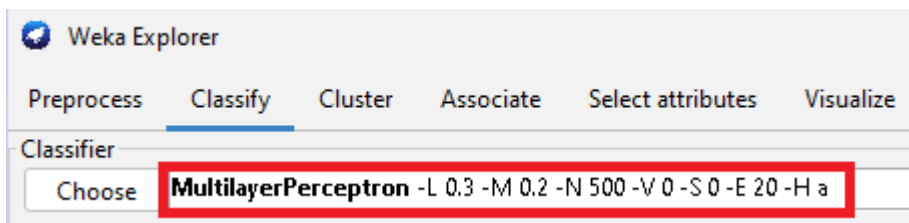
La utilización de una matriz de confusión obtenida con una red neuronal para clasificar si un proyecto será exitoso o no, basándose en ciertas características, tiene una relevancia significativa en el ámbito empresarial. Esta herramienta puede proporcionar información valiosa para la toma de decisiones estratégicas y la gestión de proyectos. En este caso, la red neuronal estaría entrenada para clasificar proyectos como exitosos o no exitosos, utilizando características específicas como entrada.

Preguntas

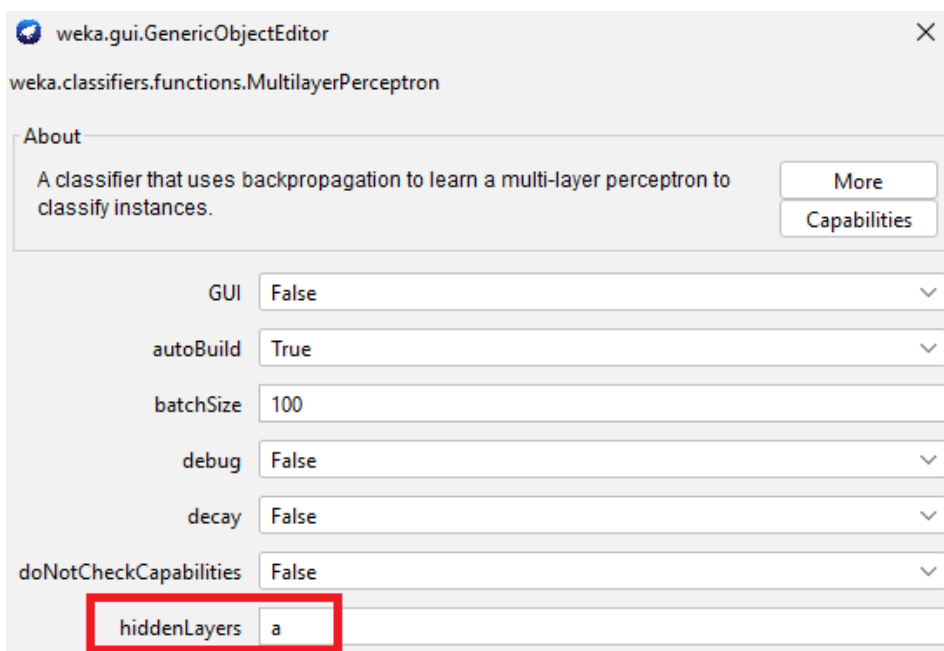
- 1) ¿Cómo podemos especificar cuántos nodos deseamos colocar en cada capa oculta en Weka?
- 2) ¿Cuál es el significado de los 10 folds que indicamos en el campo Cross-Validation?
- 3) ¿Qué es el momento (momentum)? ¿Y en qué puede ayudar ajustarlo?

Respuestas

- 1) Como todo parámetro de una función en Weka, debemos hacer clic sobre la función y sus respectivos parámetros a la derecha de Choose, como se muestra a continuación:



Esto finalmente abrirá una nueva ventana de diálogo, donde observaremos que uno de los campos se denomina “HiddenLayers”. En el mismo podemos especificar el número de nodos de cada oculta que deseamos de ser necesario.



- 2) El número de folds en el campo de Cross-Validation se refiere a la técnica utilizada para evaluar y validar el rendimiento de un modelo de aprendizaje automático. En la validación cruzada, el conjunto de datos se divide en múltiples subconjuntos llamados folds. Cada fold se utiliza alternativamente como conjunto de prueba y como conjunto de entrenamiento, ya que se itera sobre los k pliegues, utilizando cada uno de ellos como conjunto de prueba y el resto de los pliegues como conjunto de entrenamiento. Durante cada iteración, se entrena la red neuronal en el conjunto de entrenamiento y se evalúa su rendimiento en el conjunto de prueba.
- 3) El momento es un factor que afecta la velocidad de convergencia y la estabilidad del modelo. Ajustarlo puede ayudar a superar mínimos locales y mejorar el rendimiento del Multilayer Perceptron.

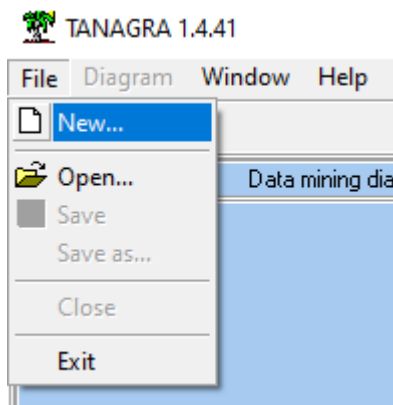
Ejercicio 5)

Presente una aplicación de Support Vector Machines.

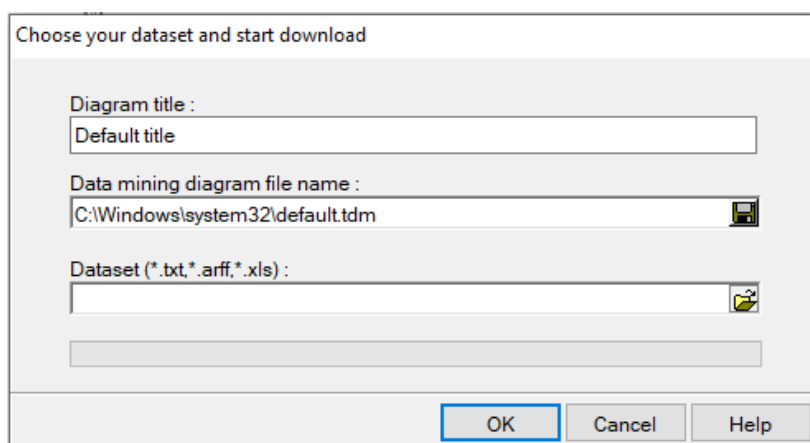
5.1) SVM – (Tanagra, empresasObligatorio2023.txt)

Asumiendo que el programa Tanagra ya se encuentra instalado, se procederá con el instructivo para cargar el archivo y realizar el ejercicio correspondiente.

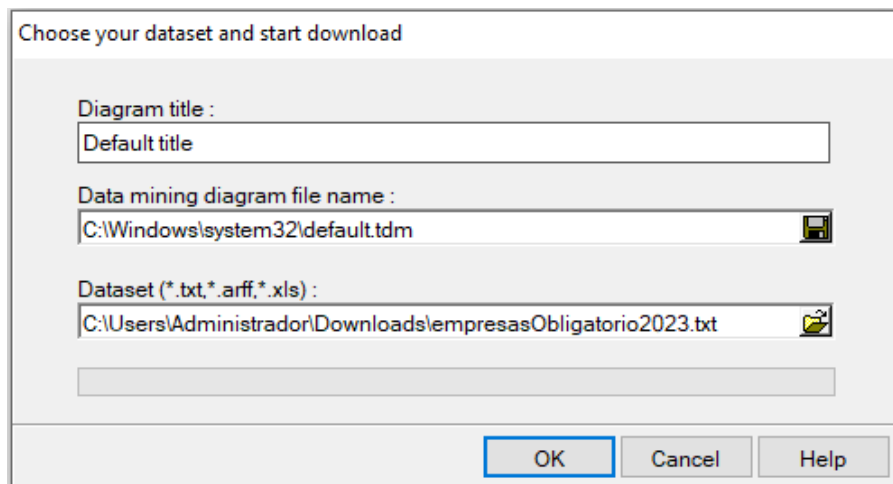
Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



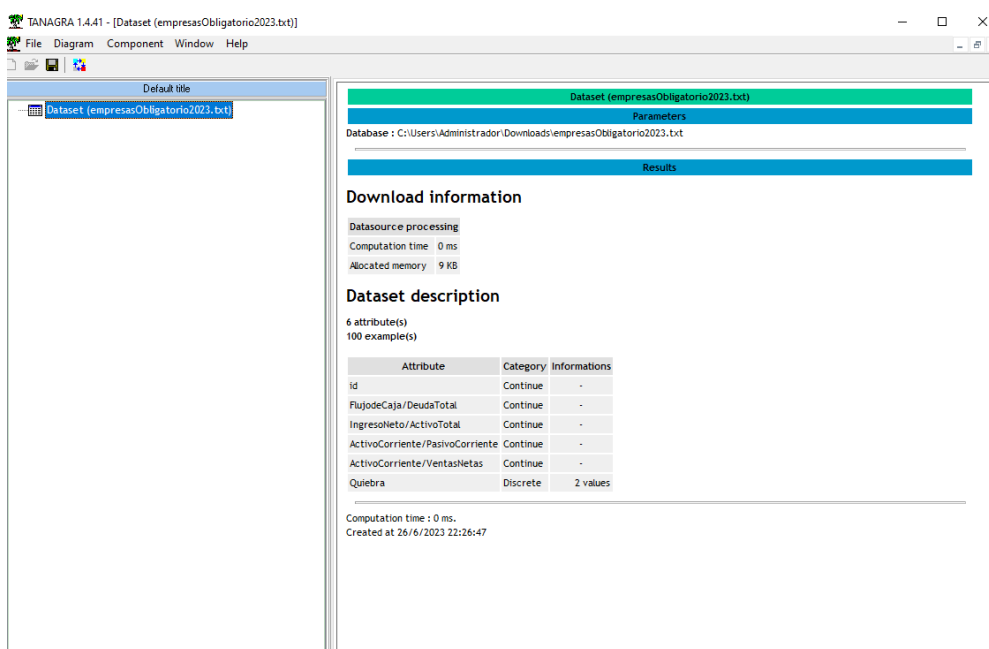
Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo Dataset (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo empresasObligatorio2023.txt en la ubicación donde fue guardado.



Una vez seleccionado el archivo se nos cargará en el campo DataSet la ruta de este.



Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:



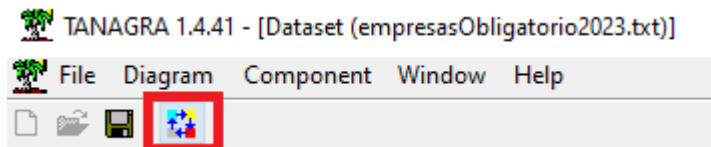
En esta parte, debemos verificar que todos los atributos excepto el denominado Grupo (que es de tipo Discrete) sean de tipo Continue. Siendo que, como ya hemos mencionado, en Tanagra estas categorizaciones se corresponden de la siguiente manera:

- **Discreta:** Es el equivalente a cualitativa
- **Continua:** Es el equivalente a cuantitativa

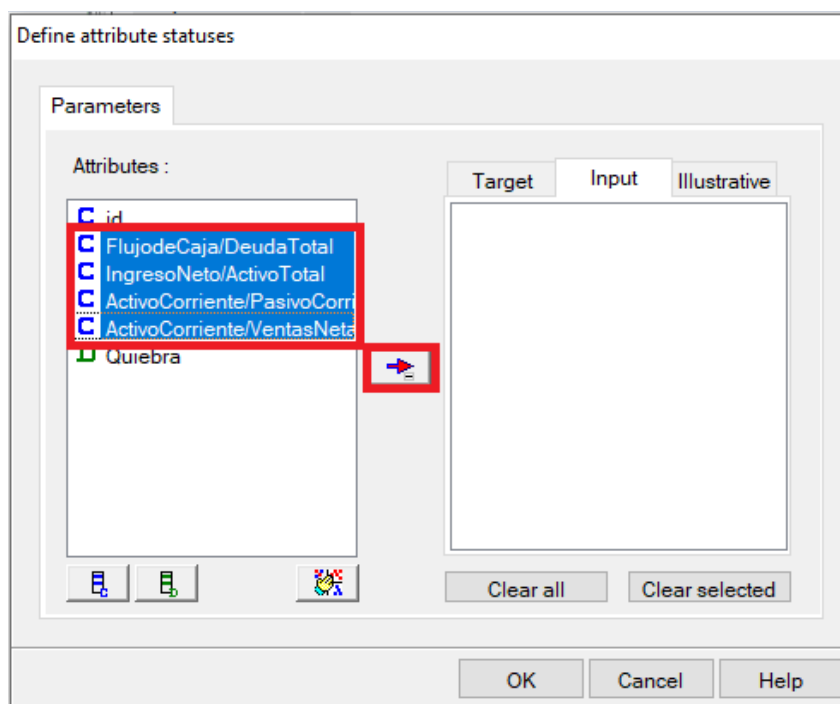
En caso de que las categorizaciones de los atributos estén invertidas a como se describió, debemos solucionarlo cambiando en el archivo las “,” (comas) por “.” (puntos).

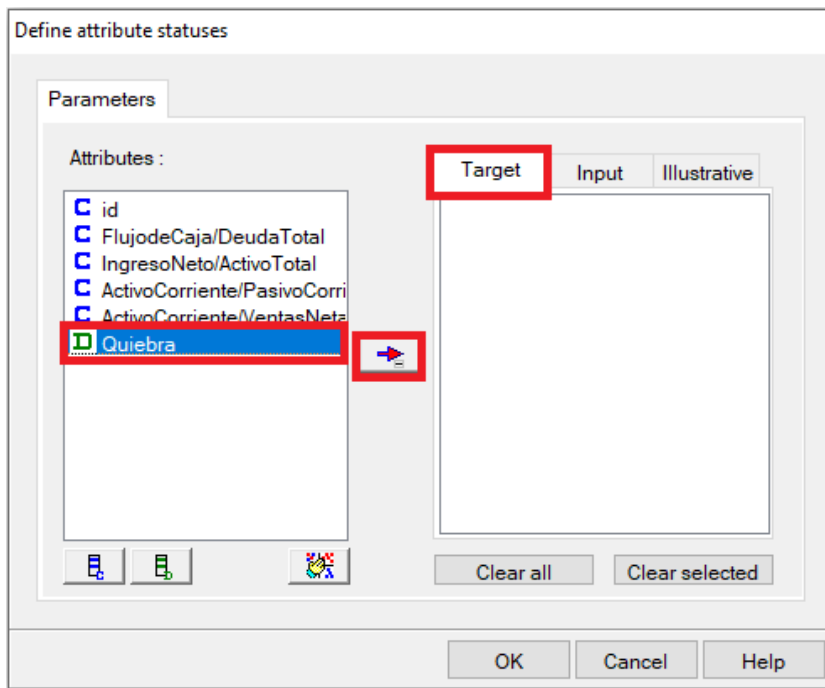
Una vez cargado el archivo para poder aplicar SVM, se deben realizar los siguientes pasos:

- 1) Antes de aplicar SVM, tenemos que hacer un define status. Debemos hacer clic en el ícono marcado en rojo en la siguiente imagen:

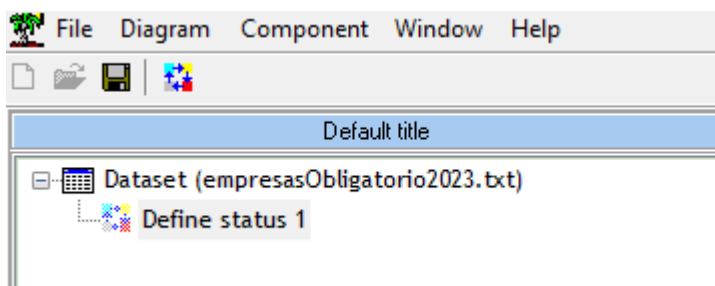


- 2) Se nos abrirá una ventana denominada Define attribute statuses, en donde debemos colocar en el target la variable Quiebra, ya que esa es la variable objetivo (generamos SVM para clasificar si una empresa va a quebrar o no). Luego, en el input vamos a colocar el resto de las variables, excepto el identificador id que no brinda información relevante.





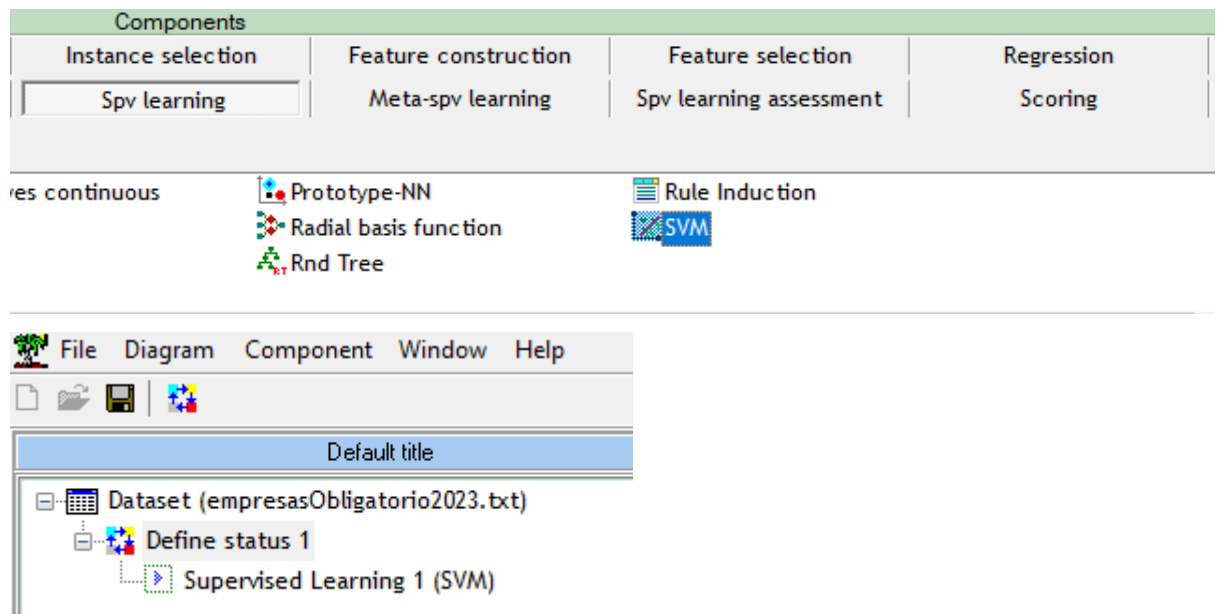
Finalmente presionamos el botón Ok, quedando de la siguiente forma:



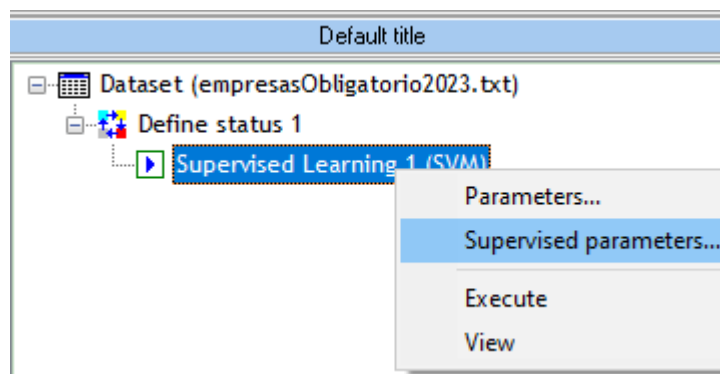
Luego hacemos doble clic en *Define Status 1* para verificar que el input se haya configurado de forma adecuada. Se debería visualizar lo siguiente:

Results			
Attribute	Target	Input	Illustrative
id	-	-	-
FlujodeCaja/DeudaTotal	-	yes	-
IngresoNeto/ActivoTotal	-	yes	-
ActivoCorriente/PasivoCorriente	-	yes	-
ActivoCorriente/VentasNetas	-	yes	-
Quiebra	yes	-	-

- Habiendo hecho el Define Status, pasamos a generar la SVM (Support Vector Machines). El mismo, corresponde a aprendizaje supervisado, ya que esta es una técnica de clasificación que utiliza una variable objetivo (en este caso Quiebra). Se encuentra dentro de la sección *Spv learning* de *Components*.

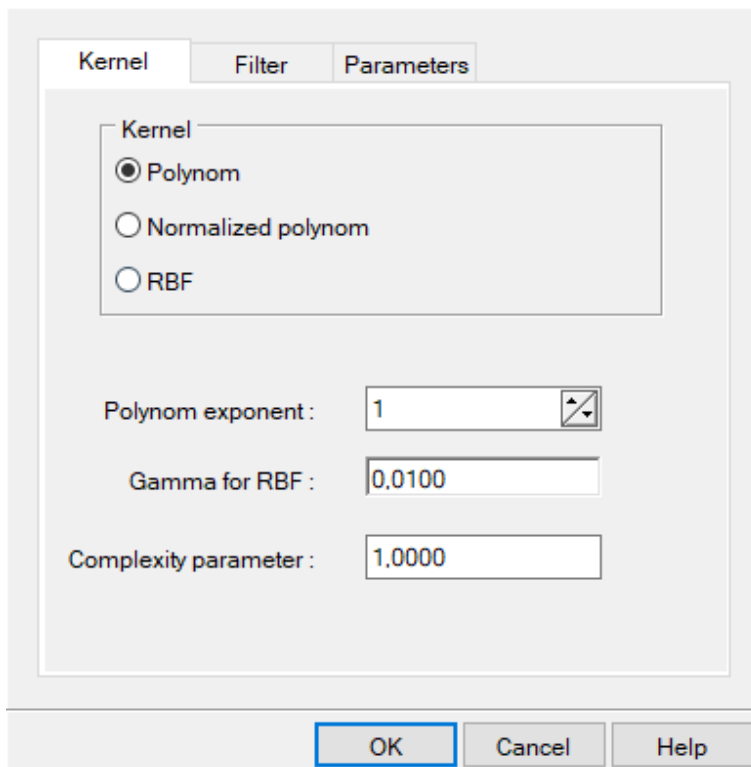


- 4) Para ejecutarlo, hacemos doble clic en SVM (sin cambiar los parámetros). Igualmente, es conveniente hacer clic derecho sobre la SVM y clic en *Supervised parameters...* para verificar el exponente del polinomio sea 1.



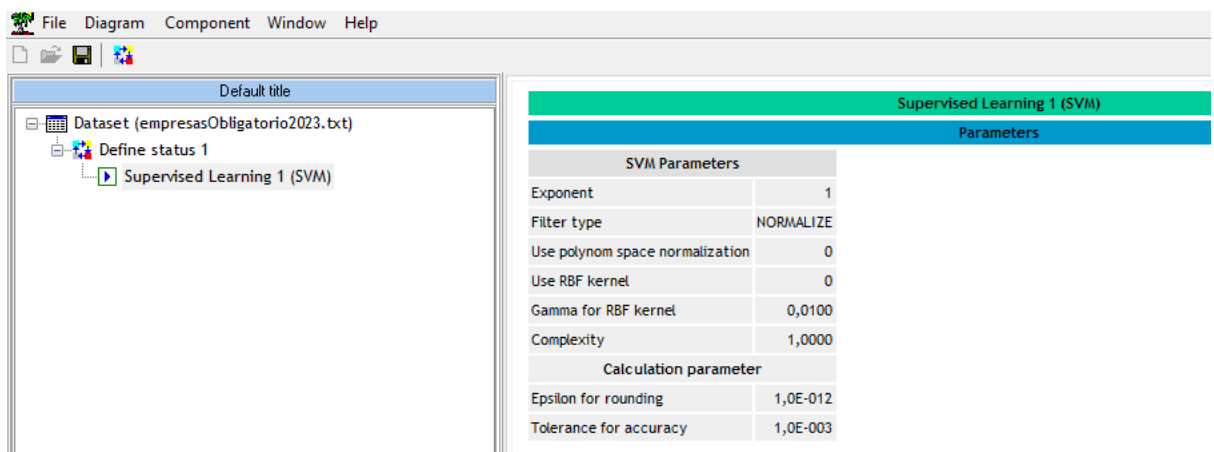
Abriéndose la siguiente ventana de configuración de parámetros:

SVM



Podemos observar dentro de la opción Kernel, tenemos un polinomio con exponente 1, es decir, una función lineal. Por lo que ahora sí, podemos hacer clic en *Ok* para cerrar la nueva ventana y luego doble clic sobre *Supervised Learning 1* para ejecutarlo.

Obteniendo así el siguiente resultado:



Matriz de Confusión

Results

Classifier performances

Error rate			0,1000			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		NO	SI	Sum
NO	0,9434	0,1228	NO	50	3	53
SI	0,8511	0,0698	SI	7	40	47
			Sum	57	43	100

Análisis: Dado que trabajamos con una técnica de clasificación, observando la matriz de confusión podemos determinar qué tan bien hizo su trabajo.

En la diagonal principal hay 90 empresas que quedaron bien clasificadas de un total de 100. Donde 40 de ellas sí quebraron y otras 50 de ellas no quebraron.

En la otra diagonal tenemos las 10 empresas restantes que no se clasificaron correctamente: 7 de ellas sí quebraron pero fueron clasificadas como que no quebraron, y otras 3 no quebraron y fueron clasificadas como que quebraron.

Relación con los negocios

Las Support Vector Machines (SVM) son un tipo de algoritmo de aprendizaje automático supervisado que se utiliza en diversos ámbitos empresariales debido a su capacidad para abordar problemas de clasificación y regresión. Pueden ser útiles para la clasificación de clientes en categorías específicas en función de sus características y comportamiento, por ejemplo, predecir si un cliente es propenso a cancelar un servicio. Tiene gran utilidad en la detección de fraudes en transacciones financieras. Por otra parte, tiene la capacidad de clasificar comentarios y opiniones en redes sociales como positivos, negativos o neutros, e incluso puede llegar a diagnosticar una enfermedad, colaborando así en la medicina.

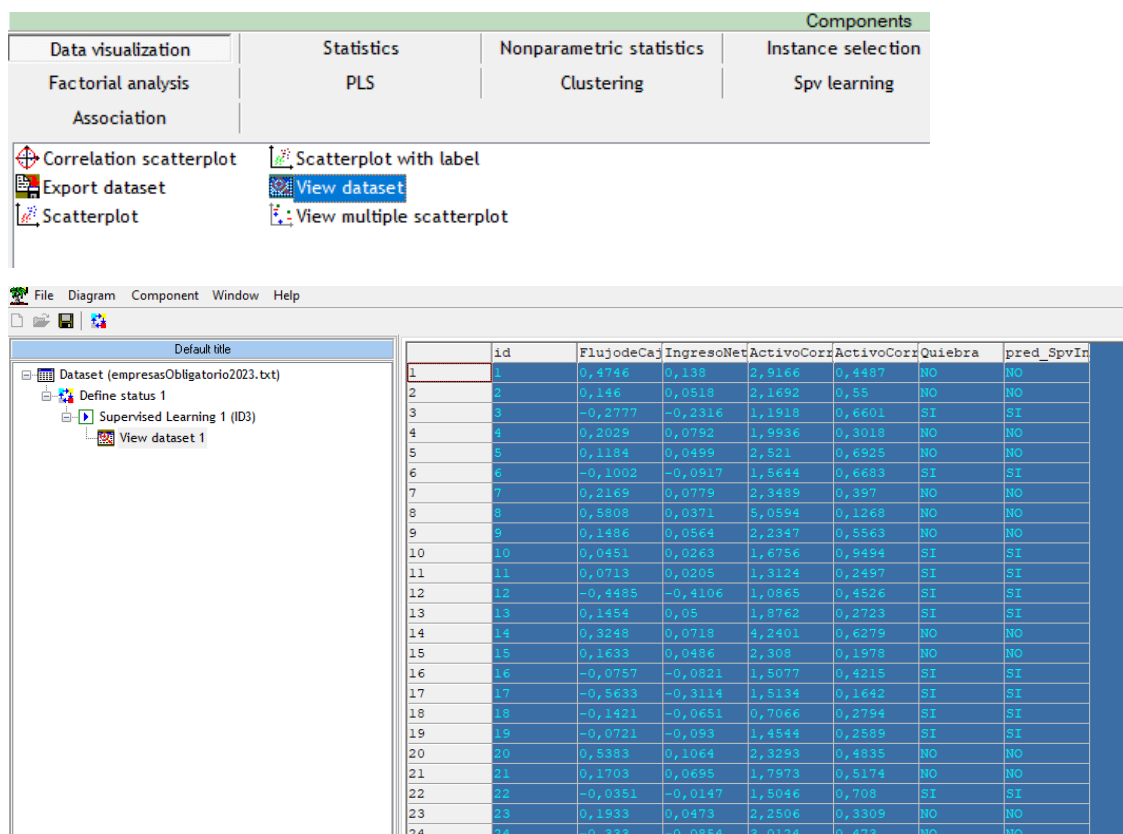
En este ejercicio nos puede ayudar a tomar acciones en caso de que una empresa se encuentre en riesgo de quebrar.

Preguntas

- 1) ¿Qué tan útil sería para un negocio predecir la variable id?
- 2) ¿Cómo puedo observar la clasificación que se le otorgó a un registro en particular?
- 3) ¿Qué indica el valor de Recall?

Respuestas

- 1) Predecir una variable de identificación no sería apropiado ni útil en la mayoría de los casos. Estas variables generalmente no tienen un patrón discernible o una relación predecible con otras características. Suelen ser únicas y no aportan información significativa para predecir una variable objetivo.
- 2) Si deseo saber cómo se clasificó un registro en particular, debajo de SVM debemos colgar un *View Dataset*, esto nos permite visualizar los datos que se están utilizando actualmente. Por lo que nos dirigimos a *Components*, en la sección Data visualization y seleccionamos *View Dataset* colgándolo debajo de *Supervised Learning 1*, posteriormente lo ejecutamos.



Components							
Data visualization		Statistics	Nonparametric statistics	Instance selection			
Factorial analysis		PLS	Clustering	Spv learning			
Association							
Correlation scatterplot	Scatterplot with label						
Export dataset	View dataset						
Scatterplot	View multiple scatterplot						

	id	FlujodeCa	IngresoNet	ActivoCorr	ActivoCorr	Quiebra	pred_SpvIn
1	1	0,4746	0,138	2,9166	0,4487	NO	NO
2	2	0,146	0,0518	2,1692	0,55	NO	NO
3	3	-0,2777	-0,2316	1,1918	0,6601	SI	SI
4	4	0,2029	0,0792	1,9936	0,3018	NO	NO
5	5	0,1184	0,0499	2,521	0,6925	NO	NO
6	6	-0,1002	-0,0917	1,5644	0,6683	SI	SI
7	7	0,2169	0,0779	2,3489	0,397	NO	NO
8	8	0,5808	0,0371	5,0594	0,1268	NO	NO
9	9	0,1486	0,0564	2,2347	0,5563	NO	NO
10	10	0,0451	0,0263	1,6756	0,9494	SI	SI
11	11	0,0713	0,0205	1,3124	0,2497	SI	SI
12	12	-0,4485	-0,4106	1,0865	0,4526	SI	SI
13	13	0,1454	0,05	1,8762	0,2723	SI	SI
14	14	0,3248	0,0718	4,2401	0,6279	NO	NO
15	15	0,1633	0,0486	2,308	0,1978	NO	NO
16	16	-0,0757	-0,0821	1,5077	0,4215	SI	SI
17	17	-0,5633	-0,3114	1,5134	0,1642	SI	SI
18	18	-0,1421	-0,0651	0,7066	0,2794	SI	SI
19	19	-0,0721	-0,093	1,4544	0,2589	SI	SI
20	20	0,5383	0,1064	2,3293	0,4835	NO	NO
21	21	0,1703	0,0695	1,7973	0,5174	NO	NO
22	22	-0,0351	-0,0147	1,5046	0,708	SI	SI
23	23	0,1933	0,0473	2,2506	0,3309	NO	NO
24	24	-0,333	-0,0854	3,0124	0,473	NO	NO

Como se puede visualizar, tenemos las dos columnas. La columna Quiebra que tiene el verdadero resultado, y la columna pred_SpvInstance_1 que tiene la predicción realizada por el modelo. En este caso, podemos ir analizando registro por registro.

- 3) El Recall, también conocido como tasa de verdaderos positivos, es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación, incluyendo los modelos de SVM. Se calcula como la proporción de instancias positivas que son correctamente identificadas por el modelo en relación con todas las instancias positivas presentes en el conjunto de datos.

Ejercicio 6)

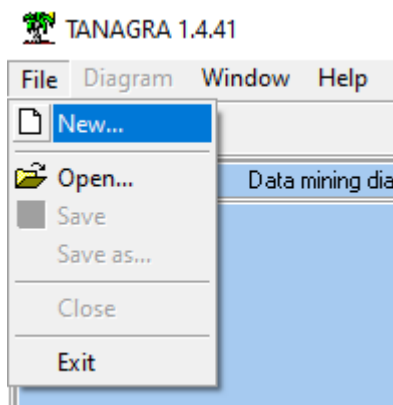
Presente dos aplicaciones de Text Mining: una de aprendizaje supervisado y otra de aprendizaje no supervisado.

6.1) Supervisado – (Tanagra, text_mining_clas1Obligatorio2023.txt)

En esta primera aplicación utilizaremos Tanagra y realizaremos un trabajo de clasificación haciendo uso del aprendizaje supervisado.

Asumiendo que ya se encuentra instalado el programa y el archivo descargado como se explicó al principio, procedemos a describir el paso a paso del trabajo.

Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo Dataset (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo text_mining_clas1Obligatorio2023.txt en la ubicación donde fue guardado.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :

OK Cancel Help

Una vez seleccionado el archivo se nos cargará en el campo Dataset la ruta de este.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :
C:\Users\Administrador\Downloads\text_mining_clas1Obligatorio2023.txt

OK Cancel Help

Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:

File Diagram Component Window Help

Default title

Dataset (text_mining_clas1Obligatorio2023.txt)

Dataset (text_mining_clas1Obligatorio2023.txt)

Parameters

Database : C:\Users\Administrador\Downloads\text_mining_clas1Obligatorio2023.txt

Results

Download information

Datasource processing

Computation time : 0 ms

Allocated memory : 6 KB

Dataset description

4 attribute(s)

160 example(s)

Attribute	Category	Informations
termino1	Continue	-
termino2	Continue	-
termino3	Continue	-
Tipo	Discrete	3 values

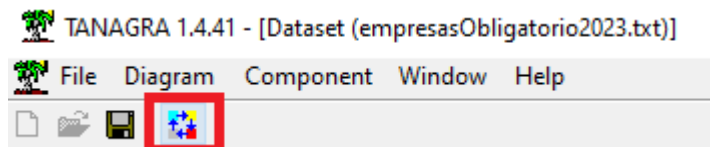
Computation time : 0 ms.

Created at 29/6/2023 6:03:27

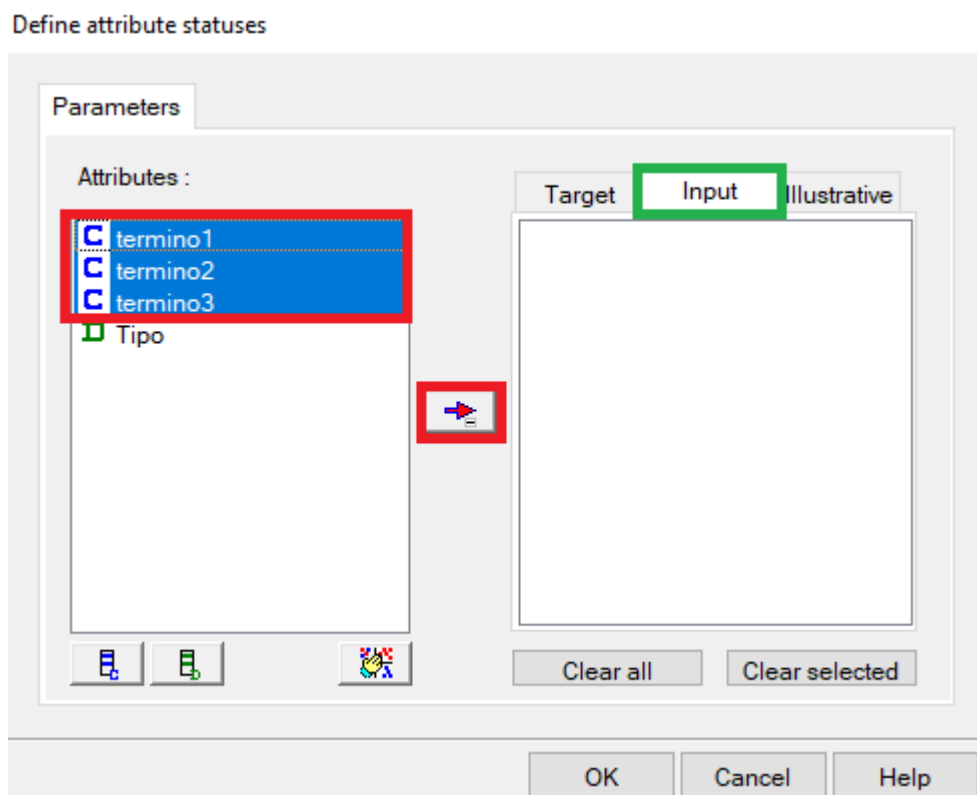
Como mencionamos, nos encontramos realizando una clasificación, por lo que previo a continuar con el trabajo debemos realizar un define status donde en input irán los atributos termino1, termino2 y termino3, y en target estará tipo.

El proceso se describe a continuación:

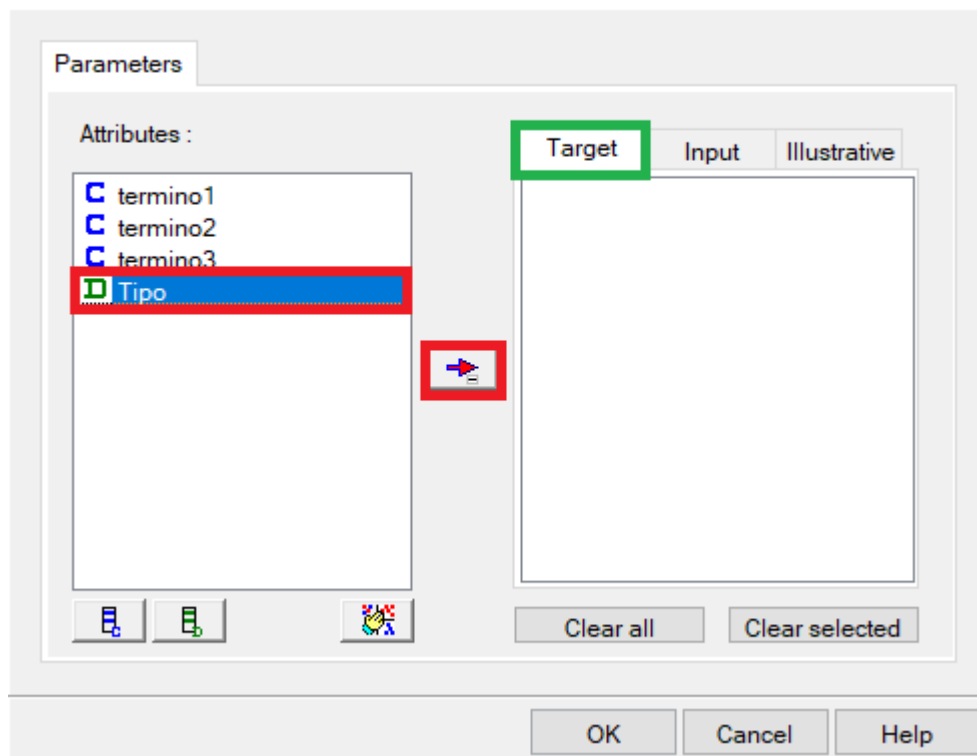
1. Para hacer el define status vamos a hacer clic en el botón que se encuentra remarcado a rojo en la siguiente imagen:



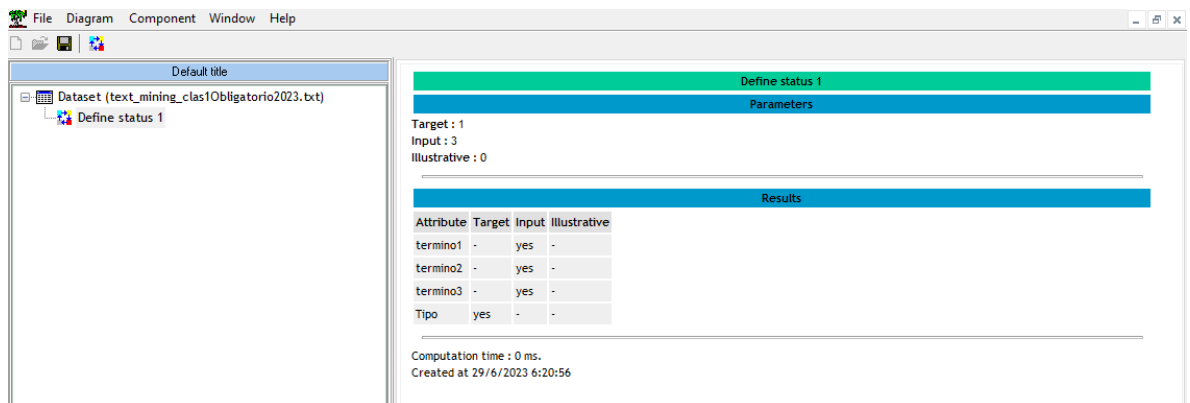
2. Se nos abrirá una ventana denominada Define attribute statuses, en donde debemos colocar en el target el atributo tipo, ya que esa es la variable objetivo. Luego, en el input vamos a colocar el resto de los atributos.



Define attribute statuses

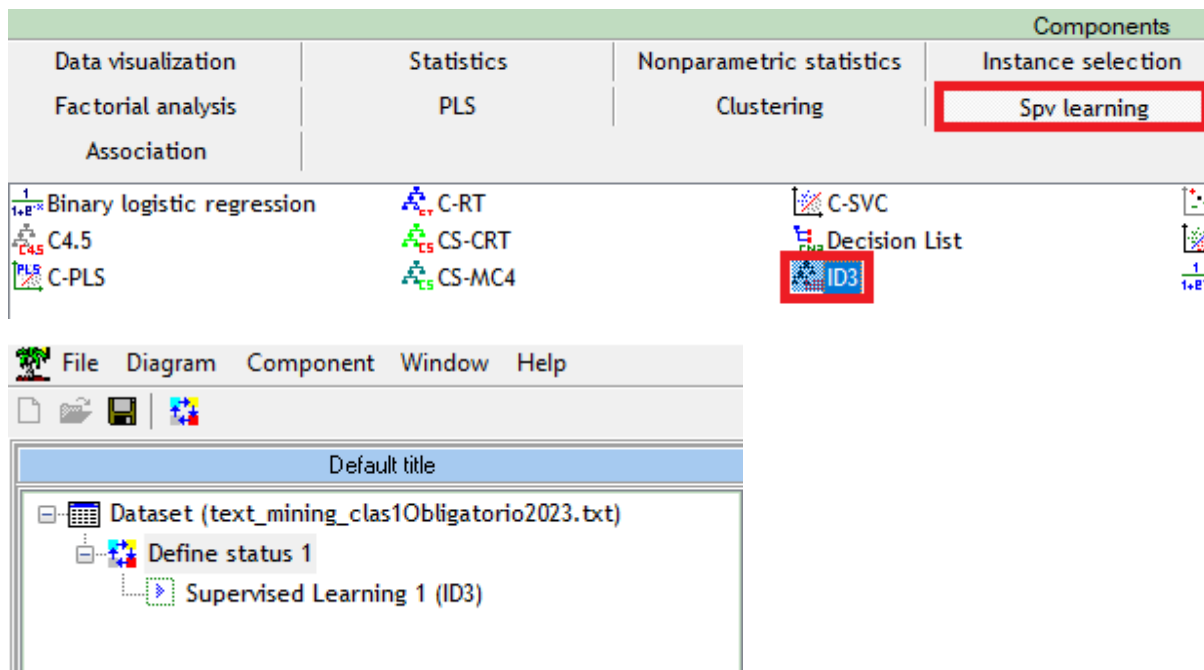


- Al finalizar, simplemente haremos clic en *OK* para cerrar la ventana que se había desplegado anteriormente. Por último, haremos doble clic sobre el *Define Status 1* para ejecutarlo, debiendo observar la siguiente pantalla:

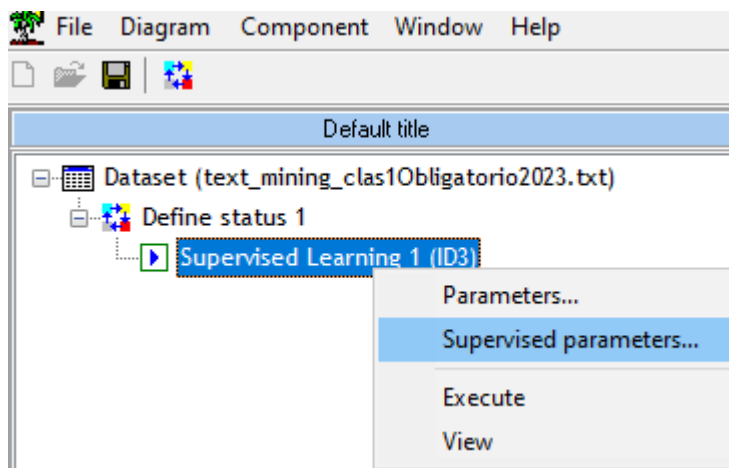


En este punto podemos continuar trabajando con el algoritmo indicado para el aprendizaje supervisado de Text Mining. En este caso utilizaremos el algoritmo ID3 en Tanagra.

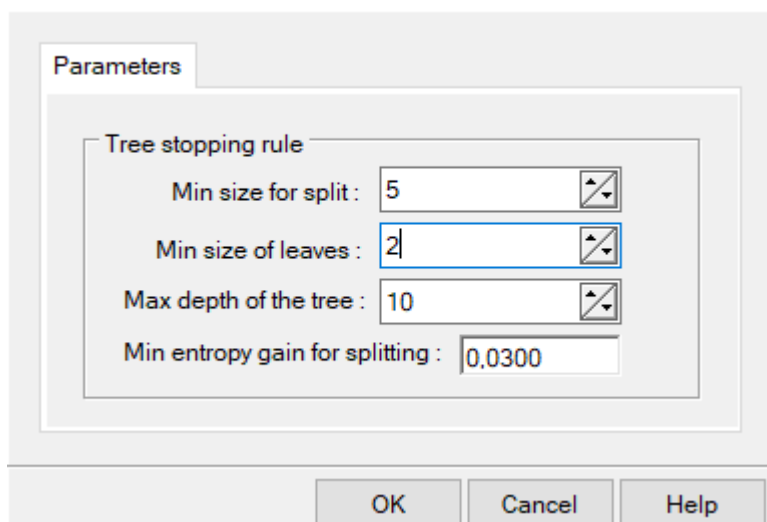
Nos vamos a dirigir a Components, luego a la sección Spv Learning y buscamos el algoritmo mencionado. Una vez encontrado lo arrastramos debajo del Define Status realizado anteriormente.



Previo a ejecutar, haremos clic derecho sobre Supervised Learning 1 con el objetivo de asignarle los valores deseados para los parámetros supervisados. Por lo que haremos clic derecho sobre *Supervised Learning 1* y seleccionaremos la opción *Supervised parameters*. Se desplegará un nuevo cuadro de diálogo, donde modificaremos el valor de *Min size for split* siendo igual a 5, y *Min size for leaves* siendo igual a 2.



ID3 parameters



The dialog box titled 'Parameters' contains a section 'Tree stopping rule' with four input fields:

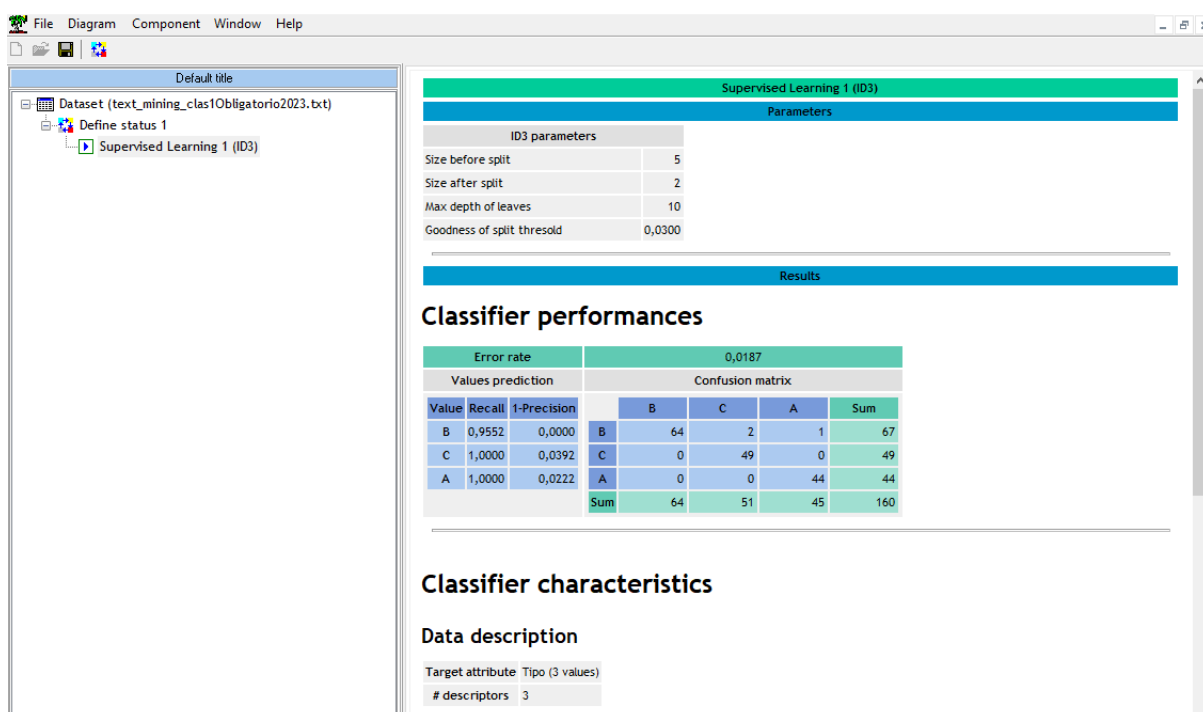
- Min size for split : 5
- Min size of leaves : 2
- Max depth of the tree : 10
- Min entropy gain for splitting : 0,0300

At the bottom are buttons for 'OK', 'Cancel', and 'Help'.

Una vez estén los valores como se observa en la imagen hacemos clic en OK.

Por último, hacemos doble clic sobre *Supervised Learning 1 (ID3)* para ejecutarlo.

Se observará la siguiente pantalla:



The window displays the results of the ID3 algorithm. It includes a tree stopping rule section, classifier performances, and classifier characteristics.

Tree stopping rule

- Size before split : 5
- Size after split : 2
- Max depth of leaves : 10
- Goodness of split threshold : 0,0300

Classifier performances

Error rate : 0,0187

Values prediction			Confusion matrix				
Value	Recall	1-Precision		B	C	A	Sum
B	0,9552	0,0000	B	64	2	1	67
C	1,0000	0,0392	C	0	49	0	49
A	1,0000	0,0222	A	0	0	44	44
			Sum	64	51	45	160

Classifier characteristics

Data description

- Target attribute : Tipo (3 values)
- # descriptors : 3

Matriz de confusión

Classifier performances

Error rate			0,0187				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		B	C	A	Sum
B	0,9552	0,0000	B	64	2	1	67
C	1,0000	0,0392	C	0	49	0	49
A	1,0000	0,0222	A	0	0	44	44
			Sum	64	51	45	160

La matriz de confusión presenta 157 datos en la diagonal principal que fueron bien clasificados y solamente 3 que quedaron mal clasificados.

Árbol de decisión

Tree description

Number of nodes	17
Number of leaves	9

Decision tree

- termino1 < 9,5000
 - termino1 < 4,0000 then Tipo = A (100,00 % of 42 examples)
 - termino1 >= 4,0000
 - termino3 < 10,5000
 - termino3 < 7,5000 then Tipo = B (100,00 % of 33 examples)
 - termino3 >= 7,5000
 - termino3 < 8,5000
 - termino1 < 6,5000 then Tipo = B (100,00 % of 2 examples)
 - termino1 >= 6,5000 then Tipo = A (66,67 % of 3 examples)
 - termino3 >= 8,5000 then Tipo = B (100,00 % of 8 examples)
 - termino3 >= 10,5000
 - termino3 < 11,5000
 - termino2 < 5,5000 then Tipo = C (60,00 % of 5 examples)
 - termino2 >= 5,5000 then Tipo = B (100,00 % of 2 examples)
 - termino3 >= 11,5000 then Tipo = B (100,00 % of 19 examples)
 - termino1 >= 9,5000 then Tipo = C (100,00 % of 46 examples)

El árbol no presenta raíz, pero esto se debe a Tanagra, se encontraría del lado izquierdo, siendo los puntos azules $\text{termino1} < 9,5000$ y $\text{termino1} \geq 9,5000$ sus respectivos hijos.

Parte desde termino1 siendo el atributo que más discrimina, incluso si este es mayor o igual a 9,5 ya afirma a qué tipo pertenece, en caso de ser menor, comienza a abrir nuevos casos sobre termino1 .

Relación con los negocios

El text mining, también conocido como minería de texto o análisis de texto, tiene diversas aplicaciones y utilidades en el ámbito de los negocios. En este caso estamos categorizando, por lo que, al permitir clasificar automáticamente grandes cantidades de textos en diferentes categorías, esto resulta útil en áreas como la gestión de contenido, la clasificación de correos electrónicos, la detección de spam y la organización de documentos. Permite ahorrar tiempo y recursos, y facilita la búsqueda y recuperación de información relevante.

Preguntas

- 1) ¿Cuál es el objetivo principal del Text Mining en el contexto de Tanagra?
- 2) ¿Cuál es el algoritmo utilizado en Tanagra para la categorización de texto utilizando Text Mining?
- 3) ¿Qué información se puede obtener a través de la matriz de confusión generada en el proceso de categorización de texto con ID3 en Tanagra?

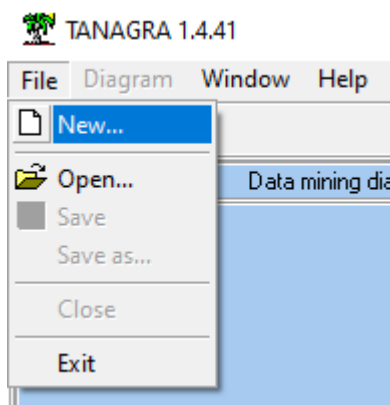
Respuestas

- 1) El objetivo principal del Text Mining en Tanagra es extraer información y conocimiento útil a partir de grandes volúmenes de texto, con el fin de categorizar y clasificar los documentos de manera automatizada.
- 2) En Tanagra, el algoritmo utilizado para la categorización de texto es el ID3 (Iterative Dichotomiser 3). El ID3 es un algoritmo de aprendizaje automático supervisado que utiliza la ganancia de información para realizar las divisiones en los nodos del árbol de decisión.
- 3) La matriz de confusión generada en el proceso de categorización de texto con ID3 en Tanagra proporciona información sobre el rendimiento del modelo de clasificación. Las filas de la matriz representan las categorías reales de los documentos, mientras que las columnas representan las categorías predichas por el modelo. La matriz muestra la cantidad de documentos clasificados correctamente (verdaderos positivos y verdaderos negativos) y los errores de clasificación (falsos positivos y falsos negativos), lo que permite evaluar la precisión y el rendimiento general del modelo.

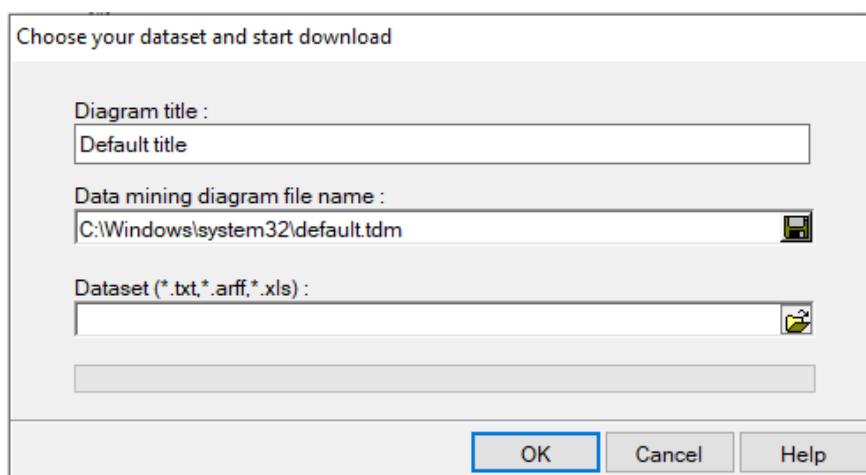
6.2) No Supervisado – (Tanagra, text_mining_clustering_1Obligatorio2023.txt)

Asumiendo que el programa ya se encuentra instalado y el archivo descargado, procedemos a realizar la explicación paso a paso de la aplicación de Text Mining no supervisado.

Abrimos el programa Tanagra, y vamos a *File* (en la parte superior izquierda), y dentro de ese menú seleccionamos New...



Una vez que seleccionamos New..., se nos abre una nueva ventana, la cual tiene diferentes campos, donde haremos énfasis en el campo Dataset (allí se seleccionará el archivo que contiene los datos a analizar). Para esto hacemos clic sobre la carpeta que se visualiza en la imagen, y luego seleccionamos el archivo text_mining_clas1Obligatorio2023.txt en la ubicación donde fue guardado.



Una vez seleccionado el archivo se nos cargará en el campo Dataset la ruta de este.

Choose your dataset and start download

Diagram title :
Default title

Data mining diagram file name :
C:\Windows\system32\default.tdm

Dataset (*.txt,*.arff,*.xls) :
C:\Users\Administrador\Downloads\text_mining_clustering_1Obligatorio2

OK Cancel Help

Posteriormente presionamos el botón “Ok”. Se abrirá la siguiente pantalla:

File Diagram Component Window Help

Default title

Dataset (text_mining_clustering_1Obligatorio2023.txt)

Dataset (text_mining_clustering_1Obligatorio2023.txt)

Parameters

Database : C:\Users\Administrador\Downloads\text_mining_clustering_1Obligatorio2023.txt

Results

Download information

Datasource processing

Computation time 0 ms

Allocated memory 17 KB

Dataset description

10 attribute(s)

150 example(s)

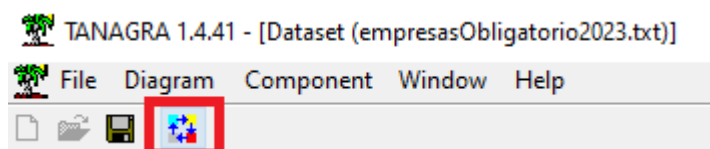
Attribute	Category	Informations
termino1	Continue	-
termino2	Continue	-
termino3	Continue	-
termino4	Continue	-
termino5	Continue	-
termino6	Continue	-
termino7	Continue	-
termino8	Continue	-
termino9	Continue	-
termino10	Continue	-

Computation time : 0 ms.

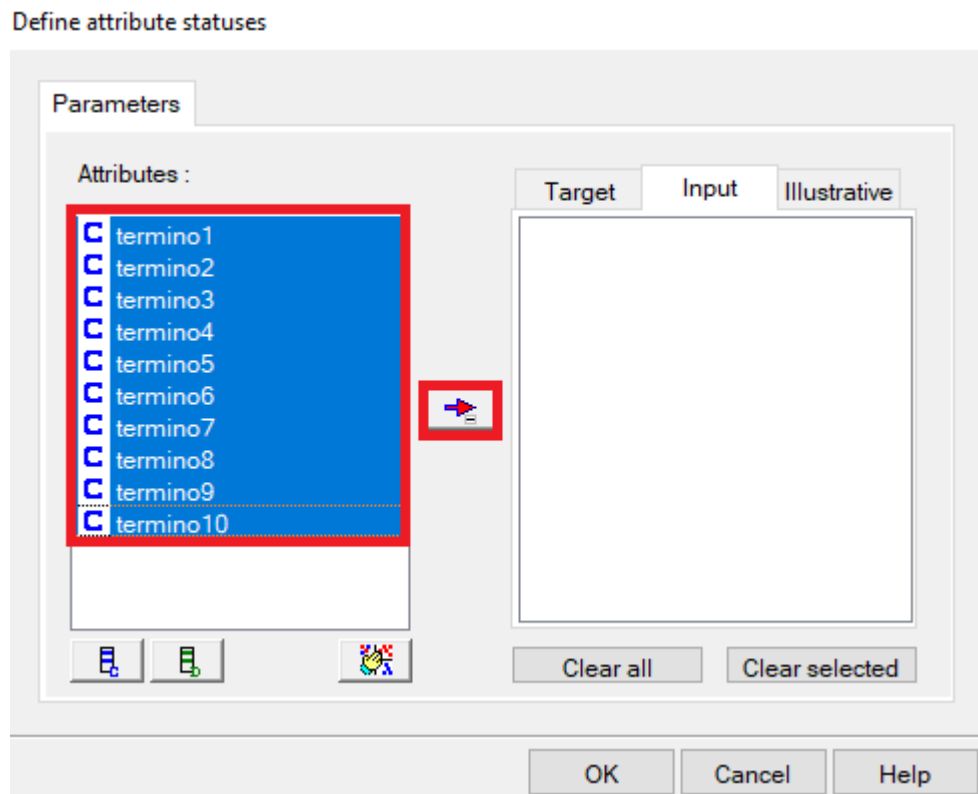
Como mencionamos, nos encontramos realizando una clasificación, por lo que previo a continuar con el trabajo debemos realizar un define status donde en input irán los atributos termino1, termino2 y termino3, y en target estará tipo.

El proceso se describe a continuación:

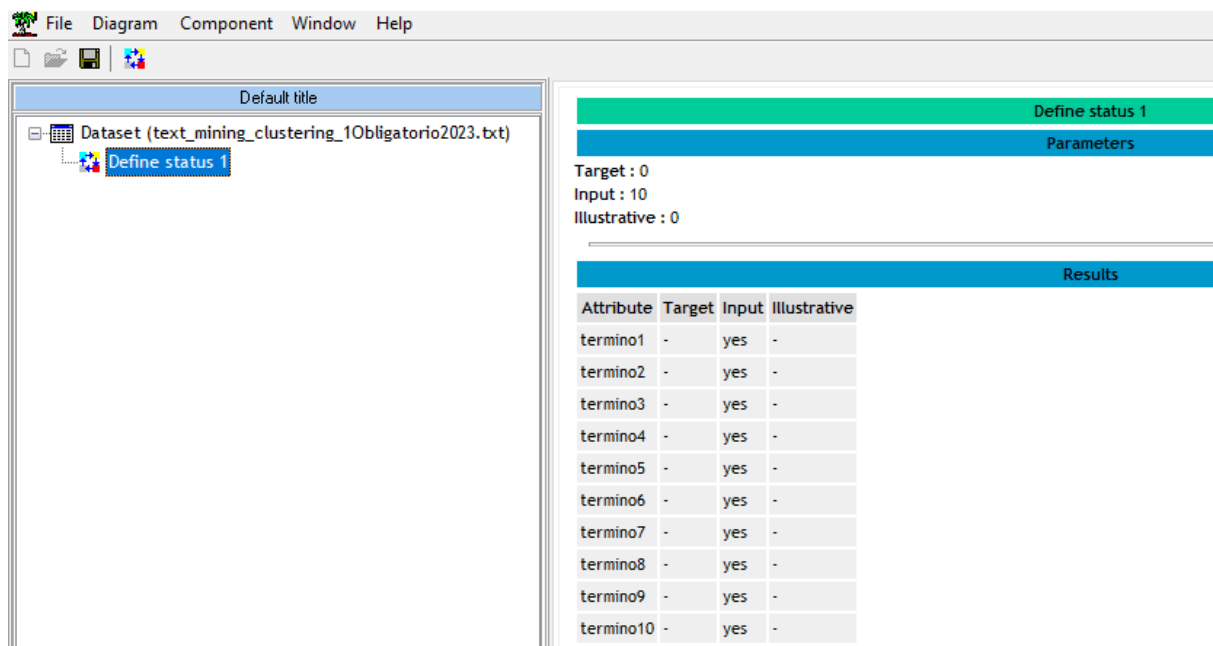
- 1) Para hacer el define status vamos a hacer clic en el botón que se encuentra remarcado a rojo en la siguiente imagen:



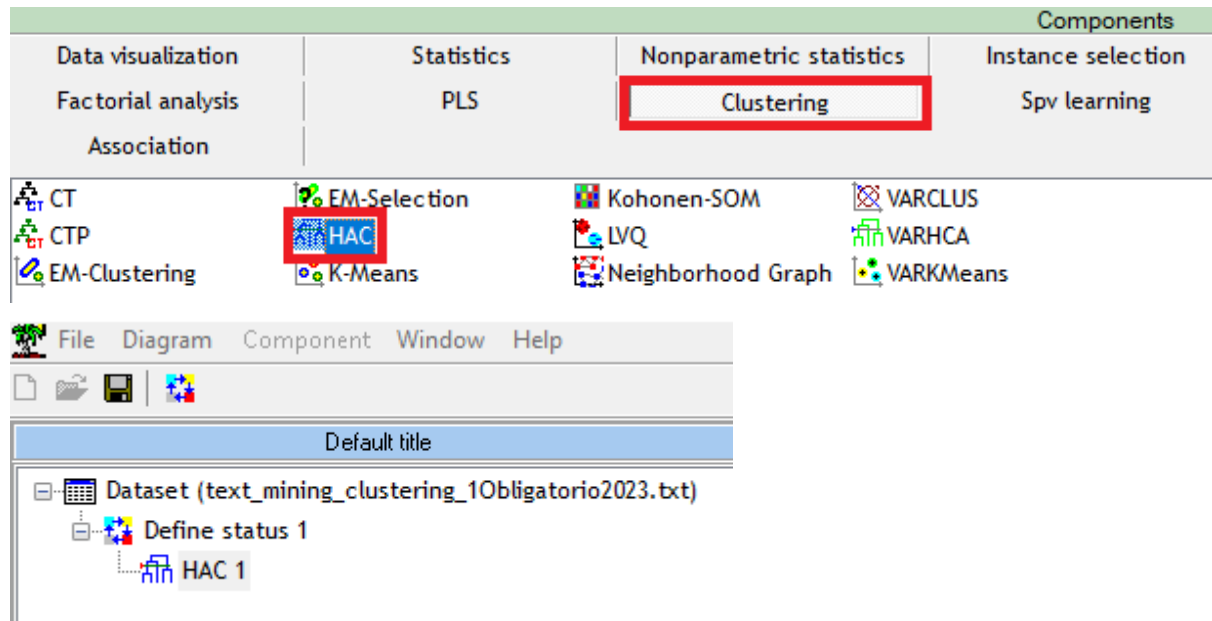
- 2) Se nos abrirá una ventana denominada Define attribute statuses, en donde debemos colocar en el input vamos a colocar todos los atributos.



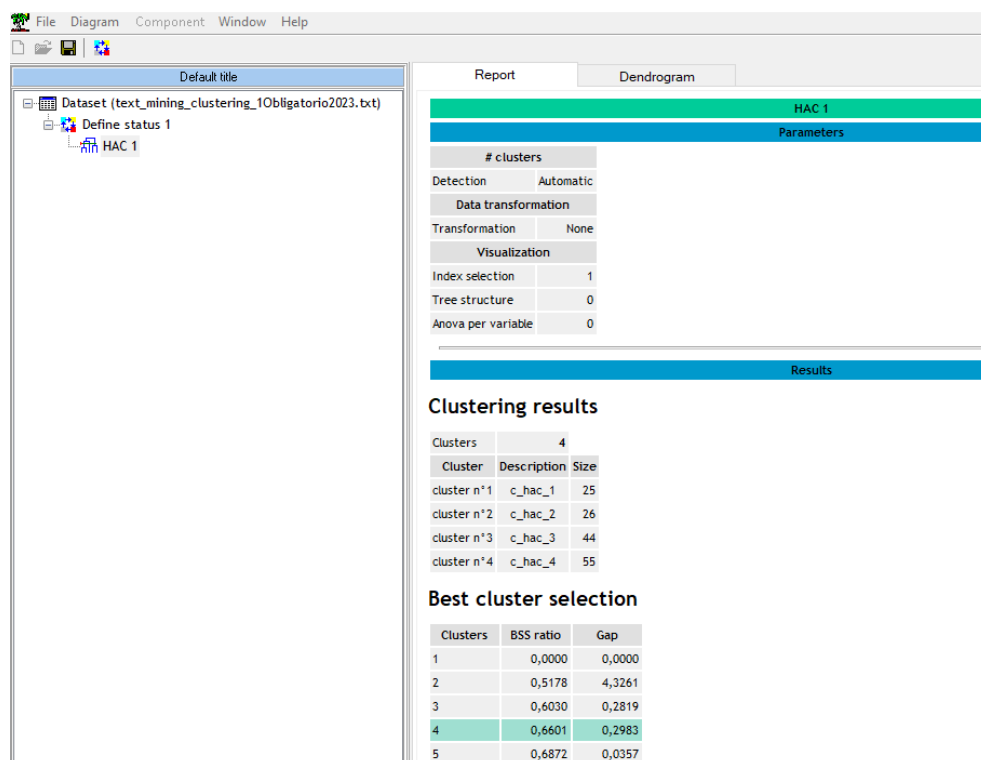
- 3) Al finalizar, simplemente haremos clic en *OK* para cerrar la ventana que se había desplegado anteriormente. Por último, haremos doble clic sobre el *Define Status 1* para ejecutarlo, debiendo observar la siguiente pantalla:



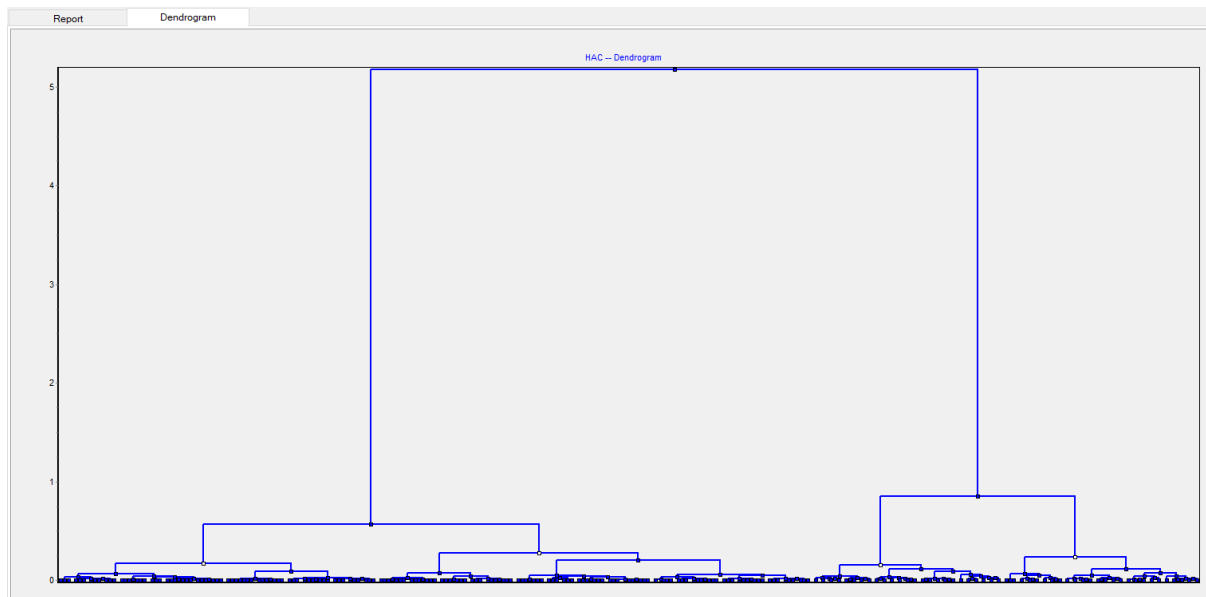
En este punto podemos continuar trabajando con el algoritmo indicado: HAC 1
 Nos vamos a dirigir a Components, luego a la sección Clustering y buscamos el algoritmo mencionado. Una vez encontrado lo arrastramos debajo del Define Status realizado anteriormente.



Hacemos doble clic sobre el algoritmo y lo ejecutamos, obteniendo la siguiente imagen:



Vemos que se formaron 4 clusters, siendo considerada la mejor cantidad. Si hacemos clic donde dice Dendrogram, en la barra superior, podremos visualizarlo de manera gráfica, de la siguiente forma:



Relación con los negocios

En el caso de trabajar con la cantidad de veces que aparece un término (también conocido como análisis de frecuencia de términos), la relación con los negocios puede ser la siguiente:

Monitoreo de redes sociales: puede utilizarse para analizar y agrupar publicaciones, comentarios o menciones en redes sociales. Esto es útil para identificar temas de conversación populares, tendencias emergentes o sentimientos comunes entre los usuarios. Las empresas pueden utilizar esta información para adaptar sus estrategias de marketing, realizar análisis de competencia o gestionar su reputación en línea

Análisis de feedback de clientes: Al analizar los comentarios, reseñas o encuestas de los clientes, puede ayudar a identificar grupos de clientes con opiniones similares o problemas comunes. Esto permite a las empresas comprender las preferencias y necesidades de sus clientes de manera más detallada, realizar mejoras en los productos o servicios, responder de manera más eficiente a los problemas y fortalecer la relación con los clientes.

Preguntas

- 1) ¿Qué significa que sea una aplicación de Text Mining no supervisado?
- 2) ¿Cuándo se trabaja con supervisado?
- 3) ¿Cuándo se trabaja con no supervisado?

Respuestas

- 1) En el contexto de Text Mining, cuando se dice que una aplicación es "no supervisada", significa que el proceso de análisis no requiere la presencia de datos de entrenamiento etiquetados o categorizados. En otras palabras, no se proporciona ninguna información sobre las categorías o clases a las que pertenecen los documentos o textos. En lugar de utilizar datos etiquetados, las técnicas de Text Mining no supervisadas se centran en descubrir patrones, estructuras ocultas o relaciones intrínsecas dentro de los datos sin ninguna guía previa. Estas técnicas exploran automáticamente el contenido textual y agrupan o categorizan los documentos en base a la similitud de su contenido

2) Utilización de Text Mining supervisado

Etiquetado de documentos: Cuando se dispone de un conjunto de documentos previamente etiquetados con categorías o clases, se puede aplicar Text Mining supervisado para desarrollar un modelo predictivo capaz de clasificar automáticamente nuevos documentos en esas categorías. Esto es útil en tareas como la clasificación de noticias, análisis de sentimientos, detección de spam, entre otros.

Construcción de modelos: Cuando se necesita entrenar un modelo predictivo utilizando características extraídas del texto, se puede emplear Text Mining supervisado. Se utilizan técnicas de aprendizaje automático para entrenar modelos que puedan predecir, por ejemplo, la relevancia de un documento o la probabilidad de ocurrencia de ciertos eventos en función del texto.

3) Utilización de Text Mining no supervisado

Exploración de datos: Si no se dispone de etiquetas o categorías previas en los datos, el Text Mining no supervisado puede ser utilizado para explorar y descubrir patrones, temas o agrupaciones latentes en el texto. Esto es útil para realizar

análisis exploratorios y descubrir información oculta en grandes conjuntos de datos no estructurados.

Segmentación de documentos: Cuando se desea agrupar documentos similares sin tener información previa sobre las categorías, el Text Mining no supervisado se utiliza para aplicar técnicas de clustering y agrupar documentos en función de la similitud de su contenido. Esto ayuda a identificar grupos temáticos, detectar tendencias emergentes o segmentar audiencias.

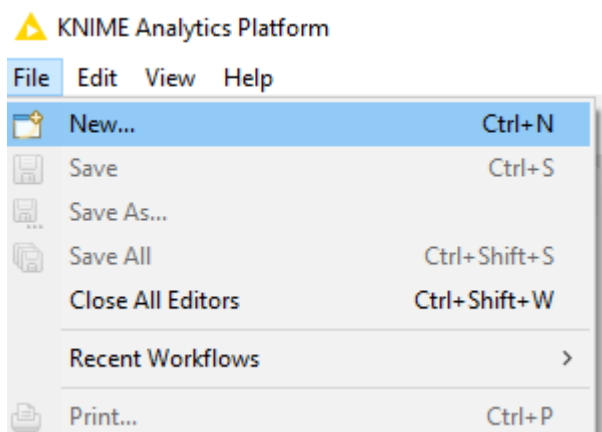
Ejercicio 7)

Presente la identificación de datos atípicos en base al archivo *Proyectos*.

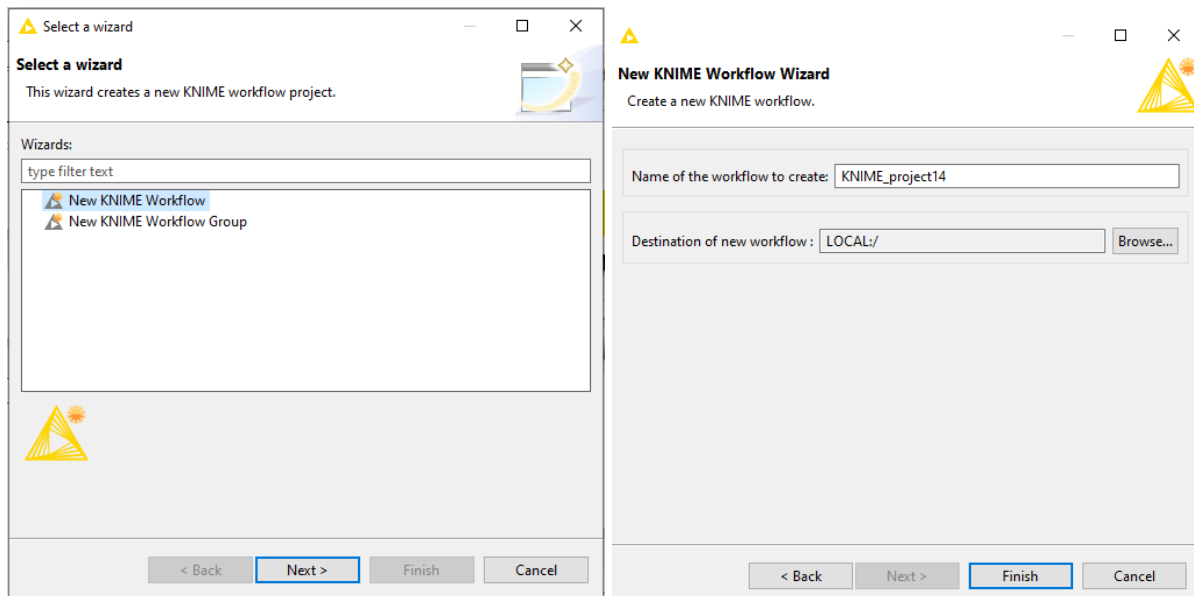
7.1) Datos Atípicos – (Knime, ProyectoObligatorio2023.txt)

Asumiendo que ya se encuentra el programa Knime instalado y el archivo correspondiente descargado porque se han utilizado en ejercicios anteriores, procederemos a explicar el proceso para la identificación de datos atípicos.

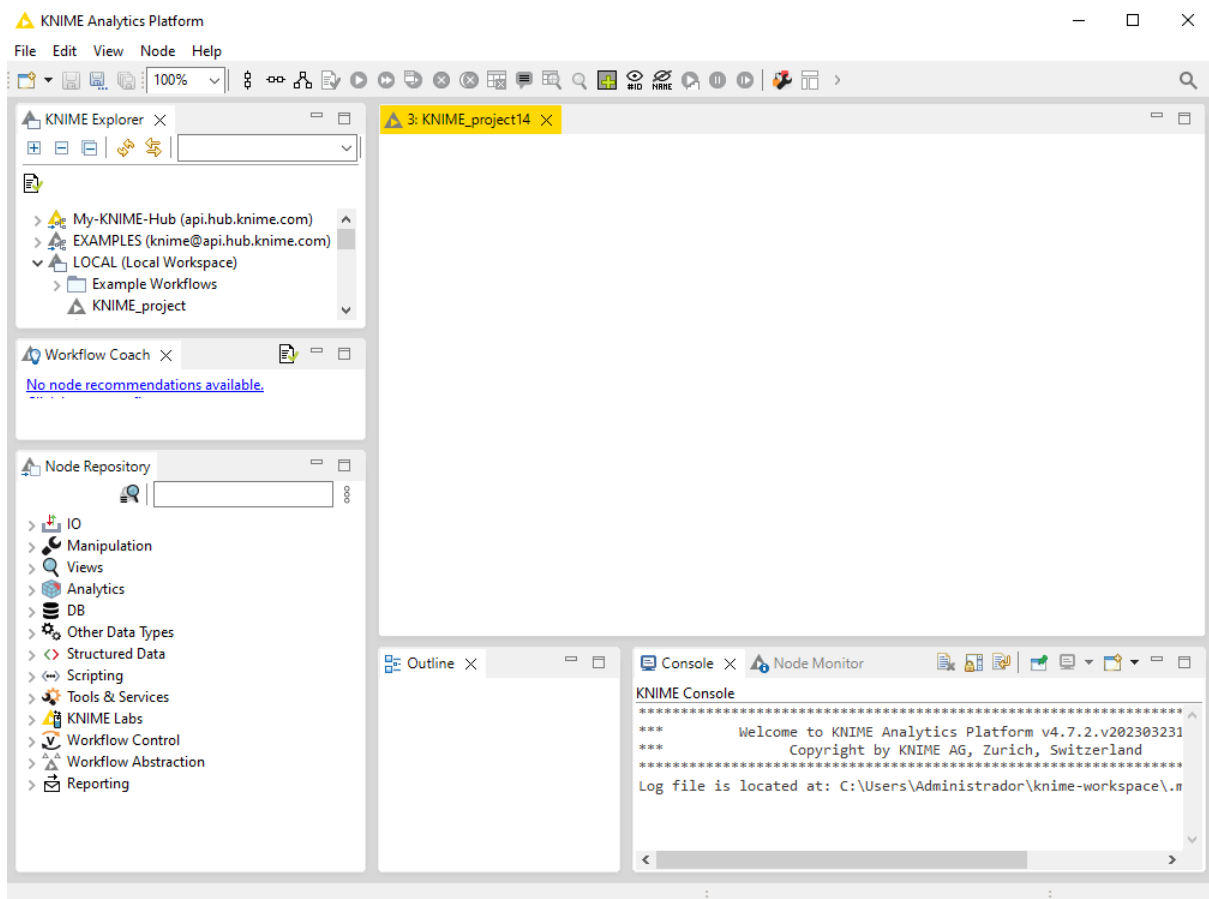
Comenzaremos abriendo Knime. Una vez abierto debemos crear un nuevo flujo de trabajo, por lo que haremos clic en *File* (ubicado en la esquina superior izquierda), y posteriormente hacemos clic en la opción *New...* (indicada en azul).



Se desplegará una nueva ventana denominada *Select a wizard*, donde seleccionaremos *New KNIME Workflow*. Una vez seleccionada la opción, marcaremos *Next*. Se abrirá otra ventana donde podemos escribir el nombre del nuevo Workflow a crear o simplemente dejar el por defecto, y cambiar la ubicación donde se guardará el mismo. Una vez realizado lo mencionado, se debe apretar el botón *Finish*.

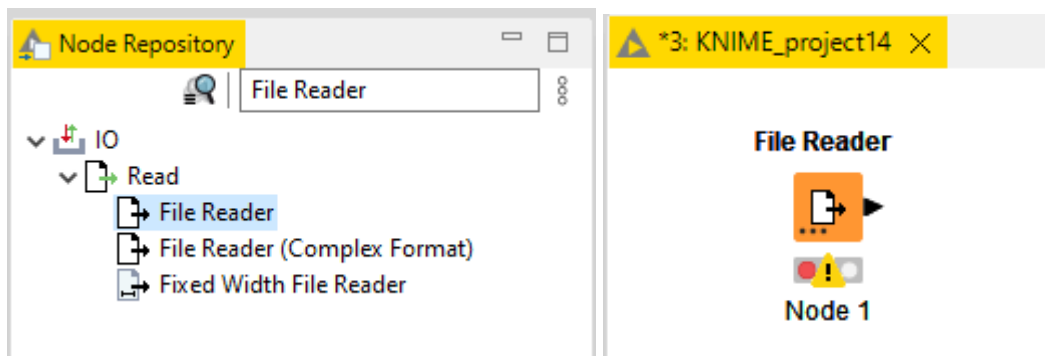


Obtendremos lo que se observa en la siguiente imagen:



Estando posicionado en dicha pantalla, debemos crear un nodo File Reader. Utilizando el buscador de Node Repository, vamos a buscar File Reader. Una vez encontrado, hacemos doble clic encima de él (indicado en color celeste).

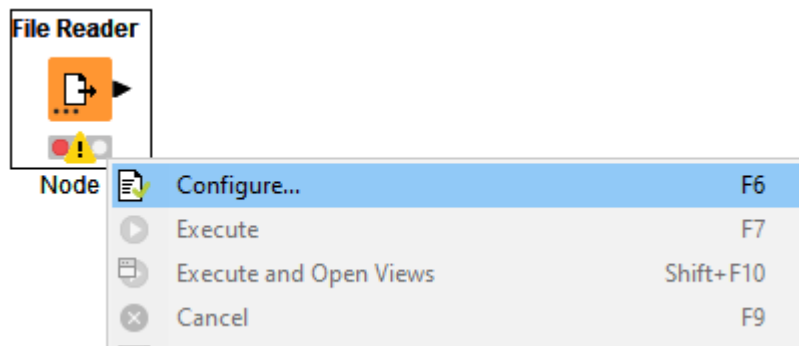
Obteniendo como resultado el siguiente nodo:



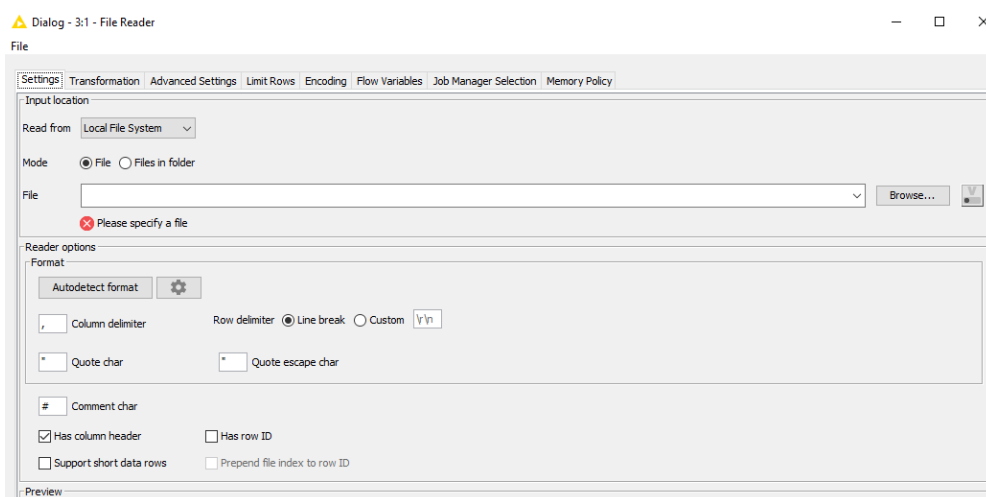
El nodo aparece, en principio, en color rojo. Esto significa que debemos configurarlo para poder ejecutarlo.

Proceso de configuración:

- 1) Hacer clic derecho sobre el nodo *File Reader*, seleccionar la opción *Configure...*

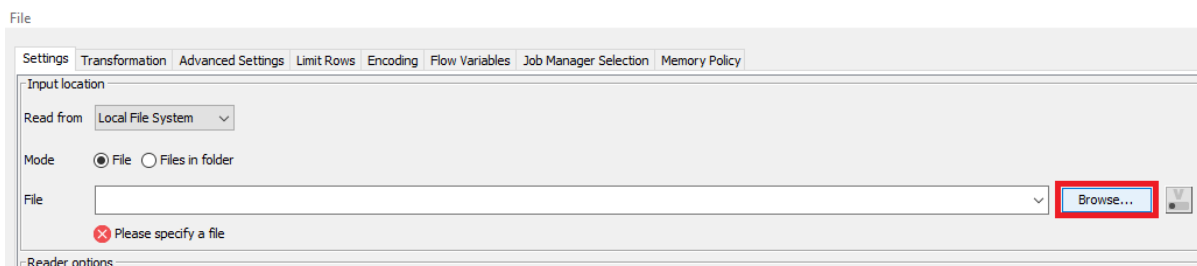


Se abrirá la siguiente ventana:

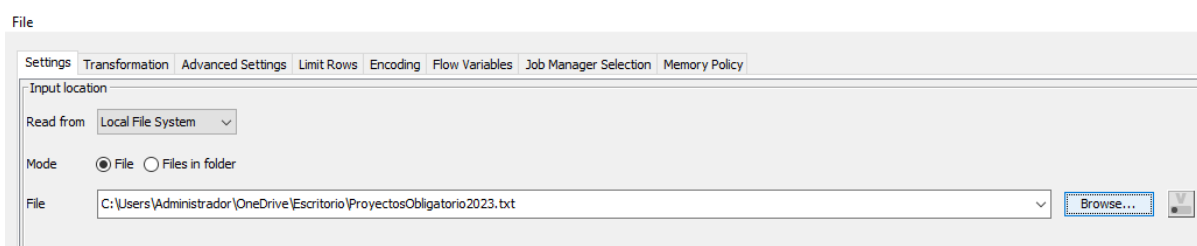


- 2) Seguidamente, hacemos clic en Browse..., esto se realiza para cargar el archivo deseado, en nuestro caso, ProyectosObligatorio2023.txt. Se nos abrirá

un nuevo diálogo donde deberemos dirigirnos a la carpeta donde guardamos el archivo, seleccionarlo y posteriormente presionar *Abrir*.



- 3) Una vez realizado este paso, observaremos que el archivo ya aparece como cargado, como se observa en la imagen adjunta.



Además, nos permitirá ver una vista previa:

Preview


i The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	S	Sexo	NivelEstudios	SituacionLaboral	OtrosProyectosP...
Row0		Masculino	Universitaria	Incompleta	IndependienteSiSINO97
Row1		Masculino	Universitaria	Incompleta	IndependienteNoSISI21
Row2		Femenino	Universitaria	Incompleta	IndependienteNoSISI96
Row3		Masculino	Universitaria	Completa	IndependienteNoSISI24
Row4		Femenino	Postgrado	Dependiente	NoSISI32
Row5		Femenino	Universitaria	Incompleta	DependienteNoSISI24
Row6		Femenino	Universitaria	Incompleta	DependienteNoSISI20
Row7		Femenino	Universitaria	Completa	IndependienteNoNONO38
Row8		Masculino	Universitaria	Incompleta	DependienteNoNONO27
Row9		Masculino	Universitaria	Incompleta	DependienteNoNONO35
Row10		Masculino	Universitaria	Incompleta	DependienteSiNOSI28
Row11		Masculino	Universitaria	Completa	DependienteNoSINO32
Row12		Masculino	Universitaria	Incompleta	IndependienteSiSINO35
Row13		Masculino	Universitaria	Incompleta	IndependienteNoSIN...
Row14		Femenino	Universitaria	Incompleta	IndependienteNoSISI26
Row15		Masculino	Universitaria	Incompleta	IndependienteNoSISI26
Row16		Masculino	Universitaria	Incompleta	DependienteNoSISI27
Row17		Masculino	Universitaria	Incompleta	IndependienteNoSISI33
Row18		Masculino	Universitaria	Completa	IndependienteNoSISI37
Row19		Femenino	Postgrado	Independiente	SiSISI31
Row20		Masculino	Universitaria	Incompleta	DependienteNoNONO26
Row21		Masculino	Universitaria	Incompleta	DependienteNoNONO26
Row22		Masculino	Universitaria	Incompleta	IndependienteSiSINO37

Al observar la tabla de la siguiente manera, debemos cambiar el delimitador de columnas de manera inmediata. La manera más sencilla de hacerlo es haciendo clic sobre la opción *Autodetect Format* que se presenta en pantalla.

Reader options

Format

Autodetect format 

Column delimiter: , Row delimiter: ☒ Line break ☐ Custom `\r\n`

Quote char: " Quote escape char: \"


Comment char: #

☒ Has column header ☐ Has row ID

☐ Support short data rows ☐ Prepend file index to row ID

Posteriormente se observarán los datos de la siguiente forma:

Preview

 The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	S Sexo	S NivelEstudios	S Situacio...	S OtrosPr...	S Proyect...	S Financi...	I EdadRe...
Row0	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	97
Row1	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	21
Row2	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	96
Row3	Masculino	Universitaria Completa	Independiente	No	SI	SI	24
Row4	Femenino	Postgrado	Dependiente	No	SI	SI	32
Row5	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	24
Row6	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	20
Row7	Femenino	Universitaria Completa	Independiente	No	NO	NO	38
Row8	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	27
Row9	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	35
Row10	Masculino	Universitaria Incompleta	Dependiente	Si	NO	SI	28
Row11	Masculino	Universitaria Completa	Dependiente	No	SI	NO	32
Row12	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	35
Row13	Masculino	Universitaria Incompleta	Independiente	No	SI	NO	24
Row14	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row15	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row16	Masculino	Universitaria Incompleta	Dependiente	No	SI	SI	27
Row17	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	33
Row18	Masculino	Universitaria Completa	Independiente	No	SI	SI	37
Row19	Femenino	Postgrado	Independiente	Si	SI	SI	31
Row20	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row21	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row22	Masculino	Universitaria Incompleta	Independiente	Si	SI	NO	32

Observamos que los datos se identifican como *string* o *integer* y se encuentran correctamente clasificados.

4) Finalmente, hacemos clic en el botón OK para finalizar.

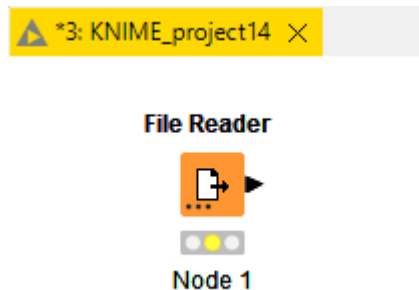
Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

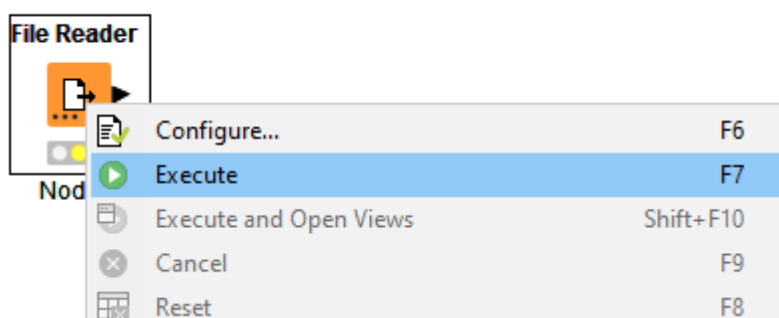
Row ID	S Sexo	S NivelEstudios	S Situacio...	S OtrosPr...	S Proyect...	S Financi...	I EdadRe...
Row0	Masculino	Universitaria Incompleta	Independiente	SI	SI	NO	97
Row1	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	21
Row2	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	96
Row3	Masculino	Universitaria Completa	Independiente	No	SI	SI	24
Row4	Femenino	Postgrado	Dependiente	No	SI	SI	32
Row5	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	24
Row6	Femenino	Universitaria Incompleta	Dependiente	No	SI	SI	20
Row7	Femenino	Universitaria Completa	Independiente	No	NO	NO	38
Row8	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	27
Row9	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	35
Row10	Masculino	Universitaria Incompleta	Dependiente	SI	NO	SI	28
Row11	Masculino	Universitaria Completa	Dependiente	No	SI	NO	32
Row12	Masculino	Universitaria Incompleta	Independiente	SI	SI	NO	35
Row13	Masculino	Universitaria Incompleta	Independiente	No	SI	NO	24
Row14	Femenino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row15	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	26
Row16	Masculino	Universitaria Incompleta	Dependiente	No	SI	SI	27
Row17	Masculino	Universitaria Incompleta	Independiente	No	SI	SI	33
Row18	Masculino	Universitaria Completa	Independiente	No	SI	SI	37
Row19	Femenino	Postgrado	Independiente	SI	SI	SI	31
Row20	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row21	Masculino	Universitaria Incompleta	Dependiente	No	NO	NO	26
Row22	Masculino	Universitaria Incompleta	Independiente	SI	SI	NO	32

OK Apply Cancel ?

Luego de finalizado, el nodo pasa a ser de esta forma:



Se observa que ahora se encuentra en color amarillo. Por lo que ahora debemos ejecutar el nodo. Para ejecutar el nodo debemos hacer clic derecho sobre el mismo y luego hacer clic sobre *Execute*.



Una vez ejecutado podremos observar que si la ejecución fue exitosa pasará a tener color verde.

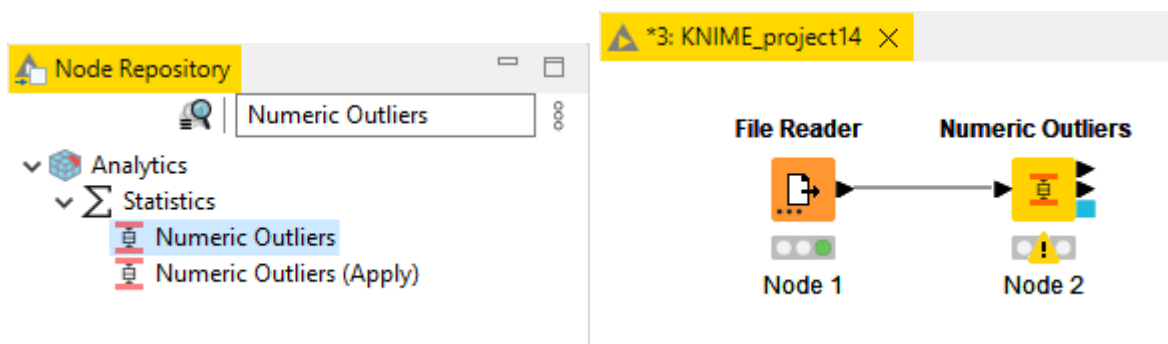
File Reader



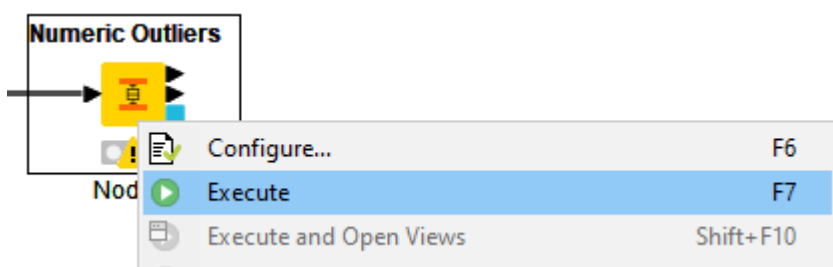
Node 1

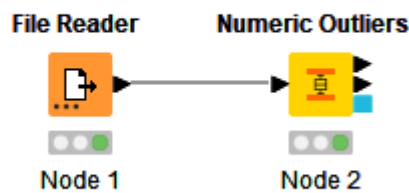
En este punto ya tendremos el archivo cargado correctamente, por lo que debemos proseguir con los nodos que se utilizarán para la detección de datos atípicos. Para esta tarea utilizaremos un único nodo más, llamado “*Numeric Outliers*”.

Por lo que nos vamos a dirigir al buscador de *Node Repository* y escribiremos *Numeric Outliers*. Una vez lo identifiquemos (se observa con celeste en la imagen a continuación), lo arrastramos al área de trabajo y relacionamos con *File Reader*.

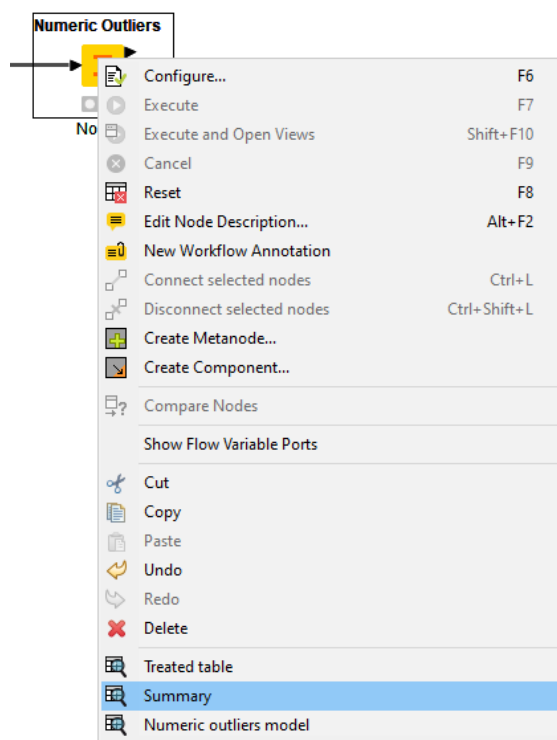


Una vez ya estén relacionados, sin modificar alguna característica, hacemos clic derecho sobre el nuevo nodo y posteriormente clic en *Execute*.





En este punto ya se habrá generado la tabla correspondiente con la información sobre los datos atípicos. Si deseamos observarla debemos hacer clic derecho sobre el nodo *Numeric Outliers* y posteriormente hacer clic en la opción *Summary*.



Summary - 3:2 - Numeric Outliers

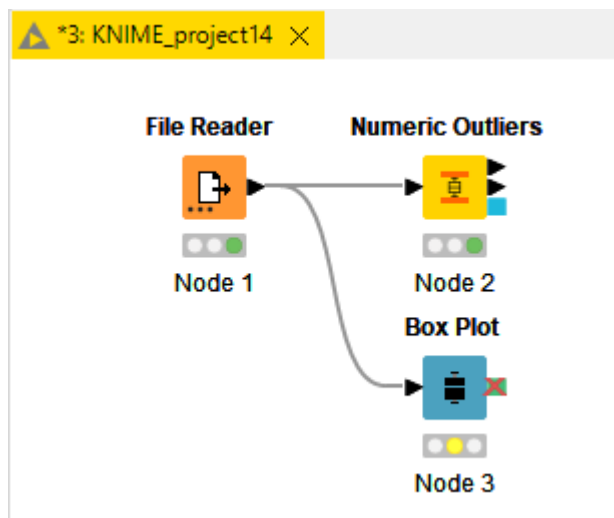
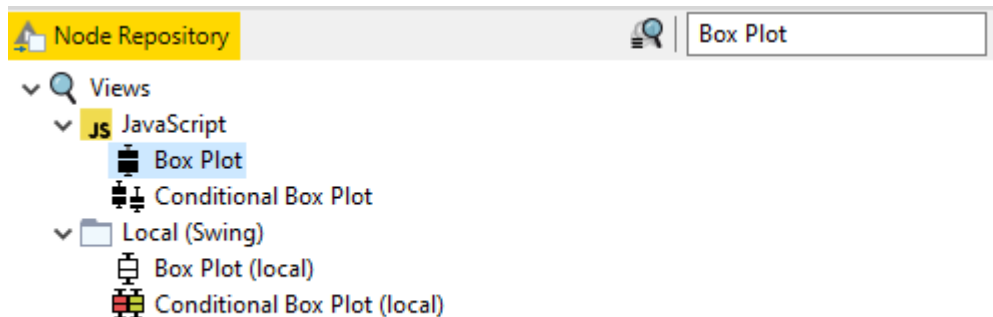
File Edit Hilite Navigation View

Table "default" - Rows: 1 Spec - Columns: 5 Properties Flow Variables					
Row ID	S Outlier column	I Member count	I Outlier count	D Lower bound	D Upper bound
Row0	EdadResponsable	99	5	7.5	51.5

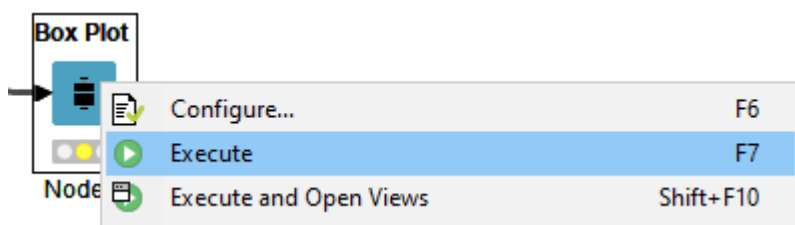
Se puede observar que los datos atípicos son aquellos $<7,5$ y $>51,5$. Habiendo en este dataset un total de 5 datos atípicos.

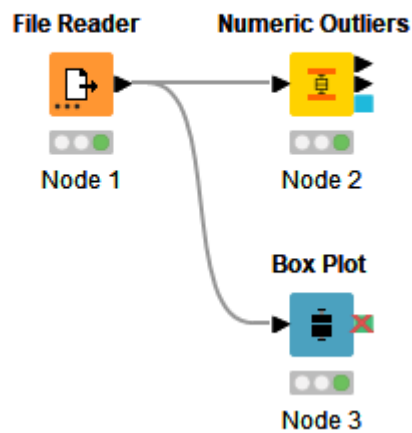
Si deseamos visualizar esto de manera gráfica podemos agregar un tercer nodo a la solución, denominado: Box Plot.

Nos vamos a dirigir nuevamente al buscador de Node Repository y escribiremos Box Plot, cabe destacar que debemos seleccionar el que se encuentra dentro de la pestaña JavaScript (indicado en color celeste). Una vez lo encontremos, lo arrastramos al área de trabajo y relacionamos con File Reader.

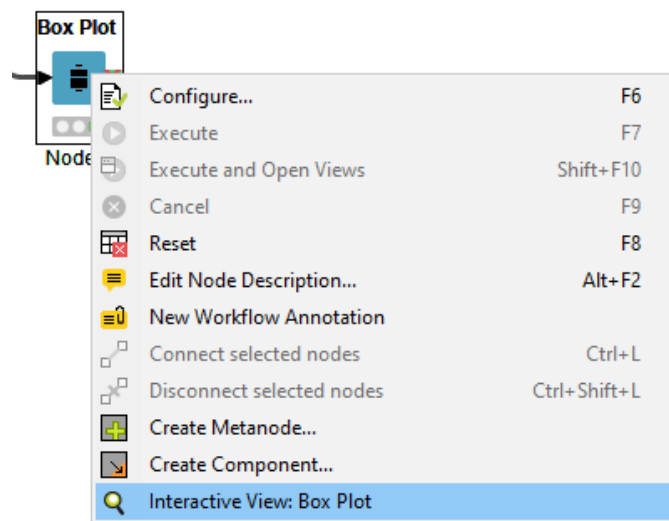


Sin modificar alguna característica, procederemos a ejecutarlo. Para esto haremos clic derecho sobre el nodo y clic en *Execute*.

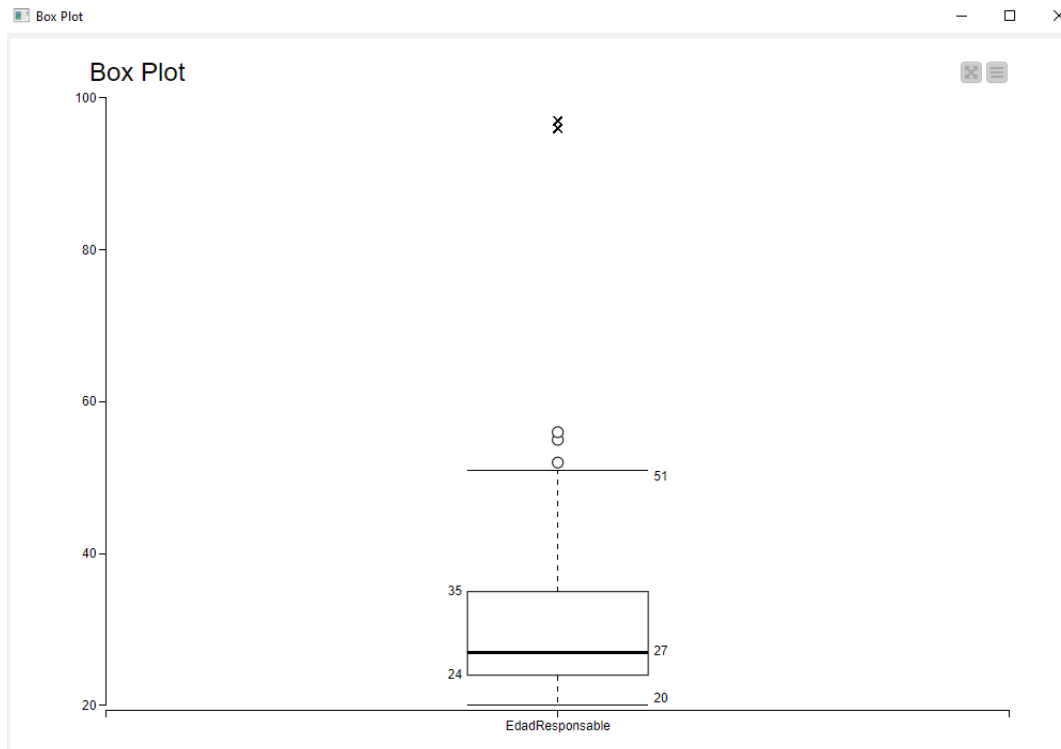




Por último, para visualizar su resultado haremos clic derecho sobre el nodo nuevamente, pero seleccionaremos la opción Interactive View: Box Plot.



Box Plot



Como ya se había observado en la tabla realizada previamente, los puntos que se encuentran por encima de la línea que está en el eje $y = 51$ son los datos atípicos.

Relación con los negocios

La detección de datos atípicos (outliers) puede proporcionar múltiples beneficios en el área de los negocios:

- Identificación de anomalías en transacciones financieras: En el mundo empresarial, es fundamental detectar transacciones fraudulentas o sospechosas que puedan tener un impacto negativo en las finanzas de la empresa. La detección de datos atípicos puede ayudar a identificar patrones o comportamientos anómalos en los datos financieros, permitiendo tomar medidas preventivas y reducir el riesgo financiero.
- Mejora de la calidad de los productos y servicios: Puede ayudar a identificar productos defectuosos o servicios que no cumplen con los estándares de calidad establecidos. Al analizar los datos y encontrar anomalías, las empresas pueden identificar rápidamente problemas en la producción o en la prestación

de servicios, y tomar medidas correctivas para mejorar la calidad y satisfacción del cliente.

- Optimización de la cadena de suministro: En el ámbito de la cadena de suministro, la detección de datos atípicos puede ayudar a identificar problemas en la gestión de inventarios, retrasos en la entrega o fluctuaciones en los precios de los materiales. Al identificar y abordar rápidamente estas anomalías, las empresas pueden optimizar su cadena de suministro, reducir costos y mejorar la eficiencia operativa.

Preguntas

- 1) ¿En qué otros ámbitos puede ser útil la detección de anomalías?
- 2) ¿Cuál es la definición de Box Plot?

Respuestas

- 1) La detección de outliers en Data Mining tiene aplicaciones en varios ámbitos y disciplinas más allá de los negocios, a continuación, mencionaremos dos como ejemplos:
 - Salud y medicina: En el ámbito de la salud, la detección de outliers es valiosa para identificar datos anómalos en registros médicos, resultados de pruebas, datos genómicos, señales biométricas y otros datos relacionados con la salud. Esto puede ayudar a identificar enfermedades raras, diagnosticar condiciones médicas inusuales y detectar errores en los datos.
 - Seguridad y ciberseguridad: En el ámbito de la seguridad, la detección de outliers se utiliza para identificar patrones de comportamiento anómalos en los datos de seguridad, como actividades maliciosas, intrusiones en redes, ataques de hackers y comportamientos sospechosos en sistemas informáticos. Esto permite tomar medidas para prevenir y mitigar amenazas de seguridad.
- 2) Un box plot es una representación gráfica que muestra la distribución de un conjunto de datos numéricos. El diagrama de caja se basa en una serie de estadísticas descriptivas y proporciona una representación visual de la dispersión y los valores atípicos de los datos.

Parte B

A una asociación empresarial le preocupa la escasez de especialistas que parece estarse previendo para el futuro. Se está estudiando el grado actual de satisfacción con la profesión. Como parte de la investigación se pidió a un conjunto de especialistas que indicaran su grado de satisfacción con el trabajo, sueldo y oportunidades de ascenso en su actual trabajo. Cada uno de los aspectos de satisfacción anteriormente mencionados fue medido en una escala de 0 a 100 puntos, y los mayores valores representan mayores niveles de satisfacción. Los especialistas se clasificaron según el tipo de especialidad (A, B, C). Se dispone además de datos del género, estado conyugal y la empresa en que trabajan. Los datos se utilizarán para determinar aspectos que permitan identificar posibilidades de mejora para la captación y conservación de especialistas en las empresas.

Ejercicio 1)

Realizar un análisis exploratorio de los datos.

Con el objetivo de conocer más acerca del archivo de datos, utilizaremos el programa Tanagra. Dado que en este punto no se solicita un tutorial del paso a paso, únicamente iremos mostrando la información que resulte relevante, omitiendo algunos detalles.

Al abrir el archivo ya se pueden observar algunas características:

Dataset description

7 attribute(s)
180 example(s)

Attribute	Category	Informations
satisfaccion_trabajo	Continue	-
satisfaccion_sueldo	Continue	-
satisfaccion_ascenso	Continue	-
tipo_de_especialista	Discrete	3 values
genero	Discrete	6 values
estadoconyugal	Discrete	5 values
empresa	Discrete	11 values

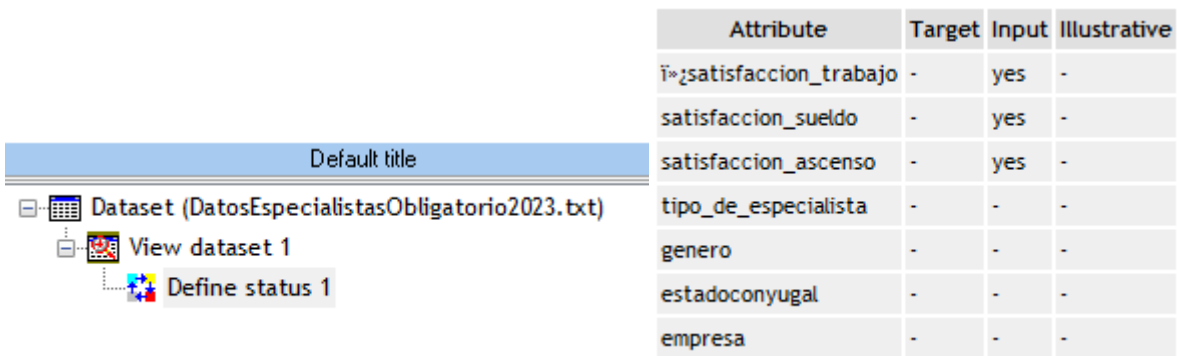
Se logra observar que el archivo tiene 7 atributos/variables. 3 de ellas son continuas (cuantitativas) y las otras 4 son discretas (alfanumérico).

Por otra parte, se logra determinar que el archivo presenta 180 registros.
Con el objetivo de visualizarlos de manera más clara, procederemos a hacer un *View dataset*, colgándolo justo debajo del dataset.

FileDiagramComponentWindowHelp

<

Continuando con el análisis, veremos algunas estadísticas para las variables continuas, tales como los valores mínimo, máximo, promedio y el coeficiente de variación de cada una. Previo a esto debemos seleccionar los datos (define status).



Attribute	Target	Input	Illustrative
i	-	yes	-
s	-	yes	-
satisf	-	yes	-
satisfacci	-	-	-
satisfacci	-	-	-
tipo_de	-	-	-
genero	-	-	-
estadoconyugal	-	-	-
empresa	-	-	-

Ahora sí, nos vamos a dirigir a *Components*, *Statistics* y seleccionar *Univariate continuous stat* para conocer los datos que mencionamos anteriormente.

Default title					
Dataset (DatosEspecialistasObligatorio2023.txt)					
View dataset 1					
Define status 1					
Univariate continuous stat 1					

Attribute	Min	Max	Average	Std-dev	Std-dev/avg
satisfaccion_trabajo	9	100	54,9111	26,4856	0,4823
satisfaccion_sueldo	5	99	57,3333	26,4524	0,4614
satisfaccion_ascenso	8	99	55,5778	26,1508	0,4705

En este caso los valores fueron todos positivos, pero de haber quedado uno negativo podríamos haber aplicado valor absoluto ya que Tanagra no lo hace.

La variable más homogénea por definición es la menos dispersa, es decir, la de menor coeficiente de variación. En este caso es *satisfacción_sueldo*, cuyo coeficiente es 0,4614.

En cuanto a la variable menos homogénea podemos afirmar que es *satisfacción_trabajo*, cuyo coeficiente es 0,4823.

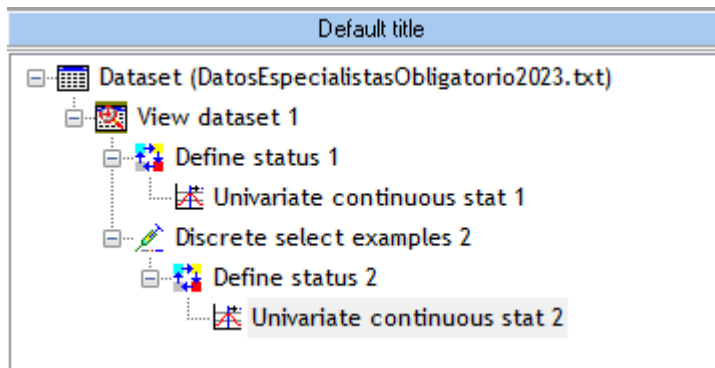
Siguiendo nuestro análisis, nos vamos a dirigir a *Components*, *Instance selection*, y en esta ocasión seleccionar *Discrete select examples*, colgándolo debajo del dataset. En los parámetros pondremos como atributo *tipo_de_especialista* e iremos iterando sobre los valores:

Default title	
Dataset (DatosEspecialistasObligatorio2023.txt)	
View dataset 1	
Define status 1	
Univariate continuous stat 1	
Discrete select examples 2	

Discrete select examples 2	
Parameters	
Attribute selection : tipo_de_especialista	
Value selection : A	
Results	
60 selected examples from 180	

Se logra conocer que hay 60 especialistas que son de tipo A.

Nuevamente, realizaremos un define status con los valores continuos con el objetivo de conocer los mismos datos estadísticos anteriores, pero específicamente para el grupo de especialistas de tipo A. Luego, volveremos a utilizar *Univariate continuous stat*.



Obteniendo los siguientes resultados:

Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
¿satisfaccion_trabajo	9	91	27,5333	15,6004	0,5666
satisfaccion_sueldo	5	97	60,3333	16,6710	0,2763
satisfaccion_ascenso	20	99	78,8333	14,9146	0,1892

Repetimos el proceso, pero para B, esta vez haciendo clic derecho sobre “*Discrete select examples*” y modificando el parámetro a tipo B.

Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
¿satisfaccion_trabajo	10	95	60,9000	14,5354	0,2387
satisfaccion_sueldo	15	99	80,4667	15,8526	0,1970
satisfaccion_ascenso	8	97	29,4333	15,1281	0,5140

Por último, repetimos el proceso para C, obteniendo los siguientes datos:

Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
¿satisfaccion_trabajo	10	100	76,3000	20,2395	0,2653
satisfaccion_sueldo	8	88	31,2000	18,5954	0,5960
satisfaccion_ascenso	12	84	58,4667	19,2323	0,3289

De estos datos, observando simplemente el promedio, ya podemos pensar que los especialistas de tipo A son los más insatisfechos con el trabajo, los de tipo C con el sueldo y los de tipo B con el ascenso. Teniendo cada uno un punto que cree debe mejorar.

Si deseamos explorar vínculos lineales entre las variables, podríamos emplear un Scatterplot e ir probando manualmente, pero buscando optimizar el tiempo, podemos hallar el coeficiente de determinación.

Para esto, realizaremos un nuevo define status con los atributos que son continuos, y luego iremos a *Components*, *Statistics*, y seleccionaremos *Linear correlation*, arrastrándolo debajo del define status previamente realizado.

Default title

- Dataset (DatosEspecialistasObligatorio2023.txt)
 - View dataset 1
 - Define status 1
 - Univariate continuous stat 1
 - Discrete select examples 2
 - Define status 2
 - Univariate continuous stat 2
 - Scatterplot 1
 - Define status 4
 - Linear correlation 2

Linear correlation 1

Parameters

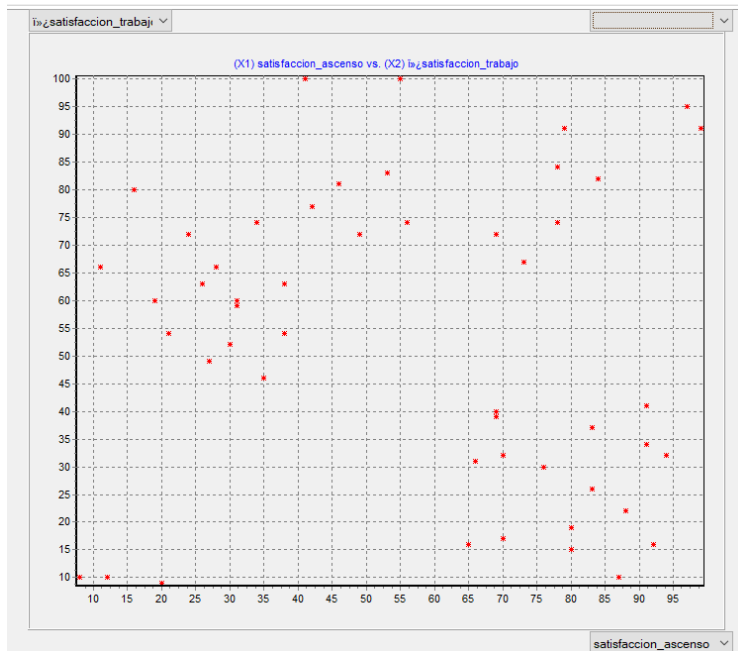
Cross-tab parameters

Sort results	non
Input list	Cross-input (Y x X)

Results

Y	X	r	r ²	t	Pr(> t)
satisfaccion_trabajo	satisfaccion_sueldo	-0,0675	0,0046	-0,9021	0,3682
satisfaccion_trabajo	satisfaccion_ascenso	-0,1870	0,0350	-2,5402	0,0119
satisfaccion_sueldo	satisfaccion_ascenso	-0,1242	0,0154	-1,6702	0,0966

Se puede observar que el mayor valor de r² es 0,0350 siendo la relación satisfacción_trabajo y satisfacción_ascenso. Teniendo este dato, procederemos a utilizar el Scatterplot mencionado anteriormente para poder visualizar cómo quedan graficadas.



Con el objetivo de seguir obteniendo información sobre los datos, realizaremos 5 intervalos de igual frecuencia para trabajar con ellos.

Para esto iremos a *Components, Feature construction* y seleccionaremos *EqFreq Disc* (intervalos con igual frecuencia), colgándolo debajo de un nuevo define status que tendrá las tres variables continuas. El mismo, por defecto, ya separa en 5 intervalos.

Data description

Attributes discretized	3
Examples	180

Generated attributes

Source	New att	Intervals	Cut points
í»¿satisfaccion_trabajo	d_eqF_í»¿satisfaccion_trabajo_1	5	(26,0000 ; 49,0000 ; 66,0000 ; 80,0000)
satisfaccion_sueldo	d_eqF_satisfaccion_sueldo_1	5	(26,0000 ; 52,0000 ; 68,0000 ; 80,0000)
satisfaccion_ascenso	d_eqF_satisfaccion_ascenso_1	5	(28,0000 ; 42,0000 ; 69,0000 ; 80,0000)

Ahora, haremos un nuevo *define status* para estos nuevos atributos que recién creamos, colgándolo debajo del *EqFreq Disc 1*. Luego, utilizaremos *Univariate discrete stat* para obtener el histograma correspondiente.

Results					
Attribute	Gini	Distribution			
d_eqF_i=satisfaccion_trabajo_1	0,7980	Values	Count	Percent	Histogram
		m_<_26,00000000	34	18,89 %	
		26,00000000=<_m_<_49,00000000	34	18,89 %	
		49,00000000=<_m_<_66,00000000	32	17,78 %	
		66,00000000=<_m_<_80,00000000	42	23,33 %	
		m_>=_80,00000000	38	21,11 %	
d_eqF_satisfaccion_sueldo_1	0,7941	Values	Count	Percent	Histogram
		m_<_26,00000000	30	16,67 %	
		26,00000000=<_m_<_52,00000000	40	22,22 %	
		52,00000000=<_m_<_68,00000000	34	18,89 %	
		68,00000000=<_m_<_80,00000000	30	16,67 %	
		m_>=_80,00000000	46	25,56 %	
d_eqF_satisfaccion_ascenso_1	0,7973	Values	Count	Percent	Histogram
		m_<_28,00000000	32	17,78 %	
		28,00000000=<_m_<_42,00000000	36	20,00 %	
		42,00000000=<_m_<_69,00000000	34	18,89 %	
		69,00000000=<_m_<_80,00000000	34	18,89 %	
		m_>=_80,00000000	44	24,44 %	

En la imagen se observa para cada variable, cómo se definieron los intervalos y la cantidad de ocurrencias que ellos alojan, con su respectivo porcentaje.

Repetimos el proceso, pero utilizando EqWidth Disc, obteniendo el siguiente diagrama de barras:

Results					
Attribute	Gini	Distribution			
d_eqW_i=satisfaccion_trabajo_1	0,7869	Values	Count	Percent	Histogram
		m_<_27,20000076	38	21,11 %	
		27,20000076=<_m_<_45,40000153	26	14,44 %	
		45,40000153=<_m_<_63,60000229	36	20,00 %	
		63,60000229=<_m_<_81,80000305	52	28,89 %	
		m_>=_81,80000305	28	15,56 %	
d_eqW_satisfaccion_sueldo_1	0,7869	Values	Count	Percent	Histogram
		m_<_23,79999924	26	14,44 %	
		23,79999924=<_m_<_42,59999847	28	15,56 %	
		42,59999847=<_m_<_61,39999771	38	21,11 %	
		61,39999771=<_m_<_80,19999695	52	28,89 %	
		m_>=_80,19999695	36	20,00 %	
d_eqW_satisfaccion_ascenso_1	0,7840	Values	Count	Percent	Histogram
		m_<_26,20000076	28	15,56 %	
		26,20000076=<_m_<_44,40000153	44	24,44 %	
		44,40000153=<_m_<_62,60000229	22	12,22 %	
		62,60000229=<_m_<_80,80000305	50	27,78 %	
		m_>=_80,80000305	36	20,00 %	

Con todo esto, ya hemos podido conocer los valores mínimo, máximo, promedio y coeficiente de variación para el dataset genérico, así como también específico para cada tipo de especialista. Hemos buscado si existe una relación lineal entre los atributos y por último hemos obtenido los distintos histogramas (Width, Freq) ampliando la información conocida sobre los valores que se adjudicaron en cada uno de los atributos, es decir, las notas asignadas por los 180 especialistas.

Ejercicio 2)

Desarrolle una aplicación de reglas de asociación.

Siguiendo con lo que se planteó en el ejercicio anterior, utilizaremos Tanagra para aplicar reglas de asociación. Para esto utilizaremos el algoritmo A priori.

Si utilizamos el algoritmo con los atributos como están se nos indicará que no hay reglas de asociación:

The screenshot shows the Tanagra interface. On the left, the project tree includes 'Dataset (DatosEspecialistasObligatorio2023.txt)', 'Define status 1', and 'A priori 1'. The main panel displays the 'A priori 1' configuration. Under 'Parameters', the settings are: Support min 0,15, Confidence min 0,50, Max rule length 4, and Lift filtering 1,10. The 'Results' section shows 'ITEMS' with 180 transactions, 3 all items, and 3 filtered items. 'Counting itemsets' shows 3 itemsets of size 2 and 1 itemset of size 3. 'Rules' shows 0 rules. The 'RULES' table is empty, with a header row: N°, Antecedent, Consequent, Lift, Support (%), Confidence (%).

Es por esto que hemos decidido utilizar las variables que creamos al dividir en 5 intervalos con EqWidth. Siendo así, hemos obtenido 6 reglas de asociación:

The screenshot shows the Tanagra interface with the 'A priori 1' configuration. The 'Parameters' are the same as in the previous screenshot. The 'Results' section shows 'ITEMS' with 180 transactions, 15 all items, and 15 filtered items. 'Counting itemsets' shows 4 itemsets of size 2 and 0 itemsets of size 3. 'Rules' shows 6 rules. The 'RULES' table contains 6 rules, each with an N° and columns for Antecedent, Consequent, Lift, Support (%), and Confidence (%).

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"d_eqW_i=satisfaccion_trabajo_1=45,40000153_<_m_c_63,60000229"	"d_eqW_satisfaccion_ascenso_1=26,20000076_<_m_c_44,40000153"	3,18182	15,556	77,778
2	"d_eqW_satisfaccion_ascenso_1=26,20000076_<_m_c_44,40000153"	"d_eqW_i=satisfaccion_trabajo_1=45,40000153_<_m_c_63,60000229"	3,18182	15,556	63,636
3	"d_eqW_i=satisfaccion_trabajo_1=45,40000153_<_m_c_63,60000229"	"d_eqW_satisfaccion_sueldo_1=80,19999695"	2,77778	11,111	55,556
4	"d_eqW_satisfaccion_sueldo_1=80,19999695"	"d_eqW_i=satisfaccion_trabajo_1=45,40000153_<_m_c_63,60000229"	2,77778	11,111	55,556
5	"d_eqW_satisfaccion_ascenso_1=62,60000229_<_m_c_80,80000305"	"d_eqW_satisfaccion_sueldo_1=42,59999847_<_m_c_61,39999771"	2,46316	14,444	52,000
6	"d_eqW_satisfaccion_sueldo_1=42,59999847_<_m_c_61,39999771"	"d_eqW_satisfaccion_ascenso_1=62,60000229_<_m_c_80,80000305"	2,46316	14,444	68,421

Reglas de asociación:

RULES

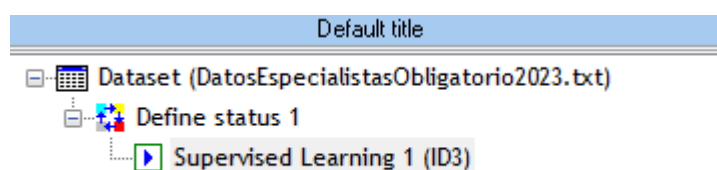
Number of rules : 6					
N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"d_eqW_i>:satisfaccion_trabajo_1=45,40000153_=<_m_<_63,60000229"	"d_eqW_satisfaccion_ascenso_1=26,20000076_=<_m_<_44,40000153"	3,18182	15,556	77,778
2	"d_eqW_satisfaccion_ascenso_1=26,20000076_=<_m_<_44,40000153"	"d_eqW_i>:satisfaccion_trabajo_1=45,40000153_=<_m_<_63,60000229"	3,18182	15,556	63,636
3	"d_eqW_i>:satisfaccion_trabajo_1=45,40000153_=<_m_<_63,60000229"	"d_eqW_satisfaccion_sueldo_1=m_>=_80,19999695"	2,77778	11,111	55,556
4	"d_eqW_satisfaccion_sueldo_1=m_>=_80,19999695"	"d_eqW_i>:satisfaccion_trabajo_1=45,40000153_=<_m_<_63,60000229"	2,77778	11,111	55,556
5	"d_eqW_satisfaccion_ascenso_1=62,60000229_=<_m_<_80,80000305"	"d_eqW_satisfaccion_sueldo_1=42,59999847_=<_m_<_61,39999771"	2,46316	14,444	52,000
6	"d_eqW_satisfaccion_sueldo_1=42,59999847_=<_m_<_61,39999771"	"d_eqW_satisfaccion_ascenso_1=62,60000229_=<_m_<_80,80000305"	2,46316	14,444	68,421

Ejercicio 3)

Desarrolle una aplicación de clasificación.

Continuaremos trabajando con Tanagra, y en esta ocasión optaremos por un árbol de decisión. Destacar que no es la única técnica que se podría utilizar, ya que se podría optar, por ejemplo, por una red neuronal.

Como ya hemos hecho anteriormente, haremos un define status donde pondremos en input los tres atributos cuantitativos y en target el tipo de especialista; buscando determinar si a través de las calificaciones que brindaron se puede determinar qué tipo de especialista es.



Results

Classifier performances

Error rate			0,0111				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		A	B	C	Sum
A	1,0000	0,0323	A	60	0	0	60
B	0,9667	0,0000	B	2	58	0	60
C	1,0000	0,0000	C	0	0	60	60
			Sum	62	58	60	180

En la matriz de confusión se puede observar que, de 180 casos, se lograron clasificar 178 de manera correcta, teniendo un desempeño casi excelente.

En cuanto al árbol de decisión queda determinado de la siguiente forma:

Tree description

Number of nodes	11
Number of leaves	6

Decision tree

- $\text{satisfaccion_trabajo} < 43,5000$
 - $\text{satisfaccion_ascenso} < 16,0000$
 - $\text{satisfaccion_sueldo} < 11,5000$ then $\text{tipo_de_especialista} = C$ (100,00 % of 4 examples)
 - $\text{satisfaccion_sueldo} \geq 11,5000$ then $\text{tipo_de_especialista} = B$ (100,00 % of 2 examples)
 - $\text{satisfaccion_ascenso} \geq 16,0000$ then $\text{tipo_de_especialista} = A$ (100,00 % of 58 examples)
- $\text{satisfaccion_trabajo} \geq 43,5000$
 - $\text{satisfaccion_ascenso} < 39,5000$ then $\text{tipo_de_especialista} = B$ (100,00 % of 56 examples)
 - $\text{satisfaccion_ascenso} \geq 39,5000$
 - $\text{satisfaccion_sueldo} < 92,5000$ then $\text{tipo_de_especialista} = C$ (100,00 % of 56 examples)
 - $\text{satisfaccion_sueldo} \geq 92,5000$ then $\text{tipo_de_especialista} = A$ (50,00 % of 4 examples)

La variable que más discrimina es `satisfacción_trabajo`. El árbol no solo nos permite clasificar, también nos da información valiosa acerca de los valores que son de suma importancia para determinar qué se debe mejorar dependiendo del especialista.

Ejercicio 4)

Desarrolle una aplicación de clustering utilizando los tres atributos de grado de satisfacción en su actual trabajo.

Para la tarea de Clustering, utilizaremos Tanagra y el algoritmo HAC 1. Destacar que no asignaremos previamente la cantidad de clusters, sino que le solicitaremos defina la cantidad que mejor se adecúe a los datos.

Obtenemos como resultado:

Default title

Dataset (DatosEspecialistasObligatorio2023.txt)

Define status 1

HAC 1

Results

Clustering results

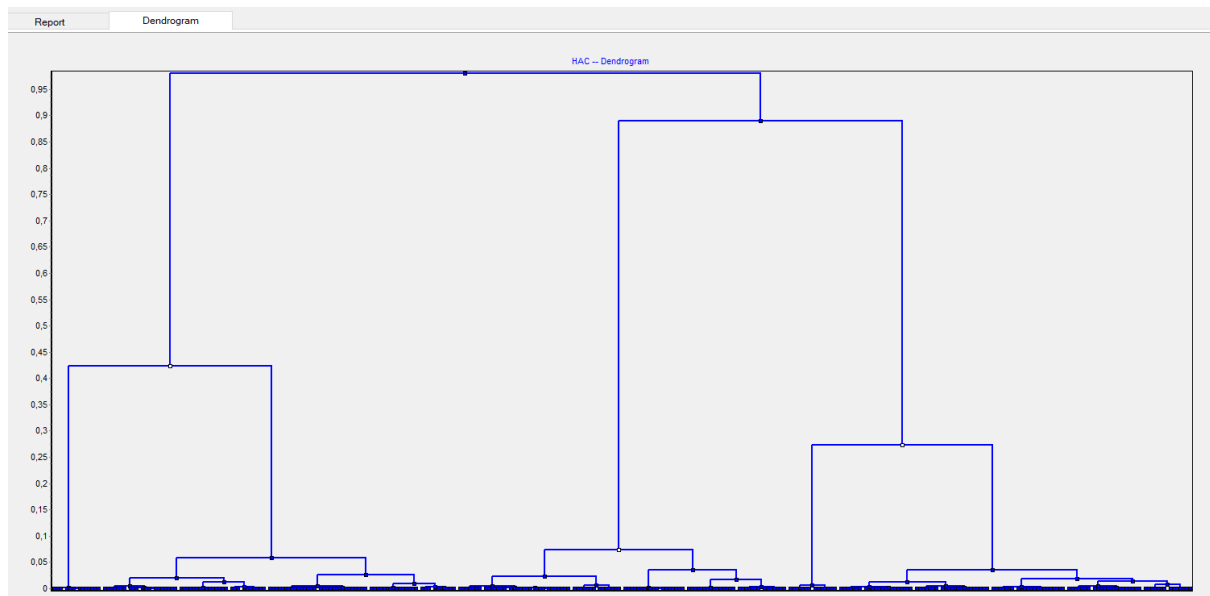
Clusters	3	
Cluster	Description	Size
cluster n°1	c_hac_1	64
cluster n°2	c_hac_2	52
cluster n°3	c_hac_3	64

Best cluster selection

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,3268	0,0910
3	0,6233	0,4654
4	0,7647	0,1516
5	0,8555	0,1982

Indicando que 3 clusters es la mejor opción, y teniendo tanto el primero como el tercero 64 elementos y el segundo con 52 elementos.

Dendrograma:

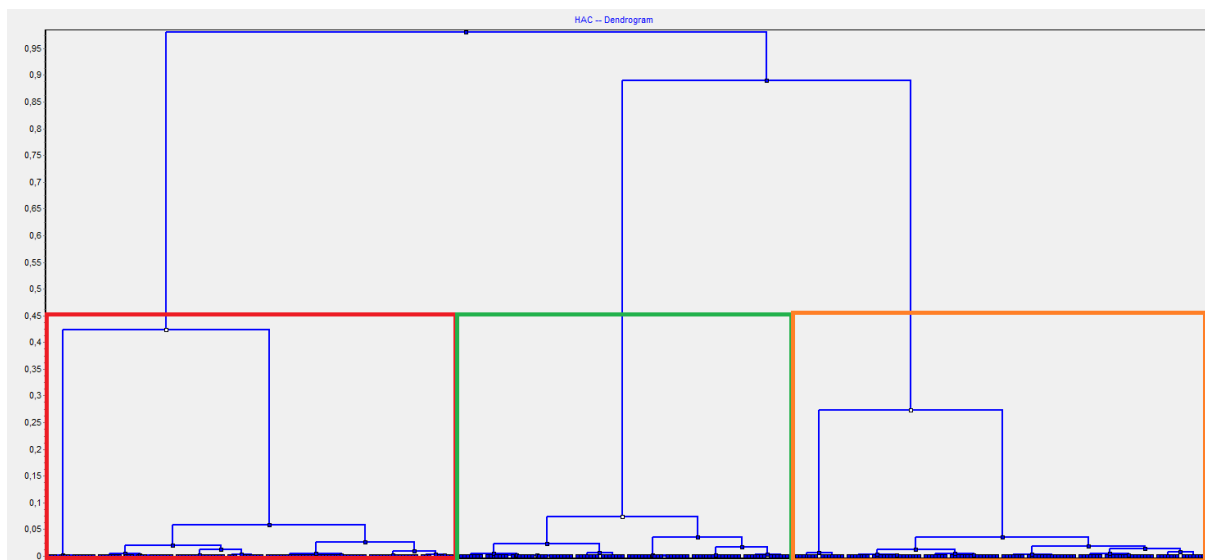


Ejercicio 5)

Presente un ejemplo de validación externa del clustering a partir del resultado del ítem anterior.

La validación externa en clustering es un proceso en el que se evalúa la calidad de los resultados del algoritmo de clustering utilizando información externa o conocimiento previo sobre las clases o categorías a las que pertenecen los datos. Es decir, se compara la estructura de clusters generada por el algoritmo de clustering con las etiquetas o clases reales de los datos, que se consideran como la verdad o la referencia.

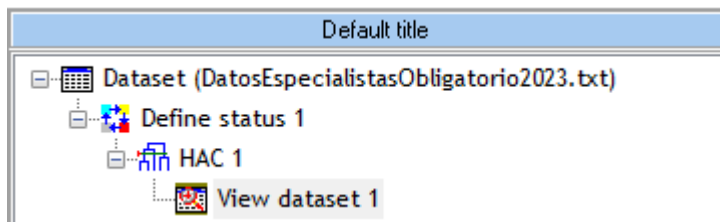
En este caso, como mencionamos en el ejercicio anterior se identificaron 3 clusters, coincidiendo con la cantidad de tipos de especialistas que existen. Los mismos se pueden apreciar en la siguiente imagen:



Por lo que podríamos identificar a cada cluster, por el momento, un número o el color que le asignamos. Lo que deberíamos poder determinar es si se puede etiquetar a cada uno con un tipo de especialista.

Por el momento, con el color rojo está identificado el cluster 1 con un total de 64 casos, el verde es el cluster 2 con 52 casos y por último, el anaranjado es el cluster 3 con 64 casos.

Con el objetivo de determinar más, podemos hacer un View dataset, y así observar a qué grupo pertenece cada dato y en qué cluster quedó asignado.



Podemos observar que la gran mayoría de los especialistas de tipo A fueron asignados al Cluster 1.

	i_satisf	satisfacci	satisfacci	tipo_de_es	genero	estadoconyugal	empresa	Cluster_HAC_1
1	91	97	99	A	Mujer	divorciado/a	FaltaTornillo	c_hac_3
2	32	42	94	A	MujerTrans	divorciado/a	AgroMisProgramas	c_hac_1
3	15	49	80	A	Varon	union libre	SegurosNoEstan	c_hac_1
4	39	74	69	A	VaronTrans	union libre	GrandesAlmacenes	c_hac_1
5	26	72	83	A	NoSabe/NoR	viudo/a	InteligenciaLaboral	c_hac_1
6	22	80	88	A	Otro	viudo/a	ComercioMuyElectronico	c_hac_1
7	10	57	87	A	VaronTrans	soltero/a	InformaticaDivertida	c_hac_1
8	31	59	66	A	NoSabe/NoR	soltero/a	AsocioTodo	c_hac_1
9	16	80	65	A	Otro	casado/a	BanQuito	c_hac_1
10	16	53	92	A	Mujer	casado/a	ClasificoTodo	c_hac_1
11	41	56	91	A	MujerTrans	casado/a	MuchosDatos	c_hac_1
12	19	52	80	A	Varon	casado/a	MuchosDatos	c_hac_1
13	34	63	91	A	VaronTrans	casado/a	MuchosDatos	c_hac_1
14	30	43	76	A	NoSabe/NoR	casado/a	AgroMisProgramas	c_hac_1
15	32	66	70	A	Otro	divorciado/a	SegurosNoEstan	c_hac_1
16	40	61	69	A	VaronTrans	divorciado/a	GrandesAlmacenes	c_hac_1
17	17	64	70	A	NoSabe/NoR	union libre	InteligenciaLaboral	c_hac_1
18	37	42	83	A	Otro	union libre	ComercioMuyElectronico	c_hac_1
19	5	5	20	A	Mujer	viudo/a	InformaticaDivertida	c_hac_1
20	15	49	80	A	MujerTrans	viudo/a	AsocioTodo	c_hac_1
21	39	74	69	A	Varon	soltero/a	BanQuito	c_hac_1
22	26	72	83	A	VaronTrans	soltero/a	ClasificoTodo	c_hac_1
23	22	80	88	A	NoSabe/NoR	casado/a	MuchosDatos	c_hac_1
24	10	57	87	A	Mujer	casado/a	FaltaTornillo	c_hac_1
25	31	59	66	A	MujerTrans	casado/a	AgroMisProgramas	c_hac_1
26	16	80	65	A	Varon	casado/a	SegurosNoEstan	c_hac_1
27	16	53	92	A	VaronTrans	casado/a	GrandesAlmacenes	c_hac_1
28	41	56	91	A	NoSabe/NoR	casado/a	InteligenciaLaboral	c_hac_1
29	19	52	80	A	Mujer	union libre	ComercioMuyElectronico	c_hac_1
30	34	63	91	A	MujerTrans	union libre	InformaticaDivertida	c_hac_1

Lo mismo sucedió con los especialistas de tipo B, fueron asignados al Cluster 3.

	i	satisf	satisfacci	satisfacci	tipo_de_es	genero	estadoconyugal	empresa	Cluster_HAC_1
38	72	84	24	B	NoSabe/NoR	soltero/a		SegurosNoEstan	c_hac_3
39	80	94	16	B	Otro	casado/a		GrandesAlmacenes	c_hac_3
40	74	79	34	B	VaronTrans	casado/a		InteligenciaLaboral	c_hac_3
41	59	69	31	B	NoSabe/NoR	casado/a		ComercioMuyElectronico	c_hac_3
42	46	86	35	B	Otro	casado/a		InformaticaDivertida	c_hac_3
43	52	99	30	B	Mujer	casado/a		AsocioTodo	c_hac_3
44	49	79	27	B	MujerTrans	casado/a		BanQuito	c_hac_3
45	54	80	21	B	Varon	union libre		ClasificoTodo	c_hac_3
46	60	75	19	B	VaronTrans	union libre		MuchosDatos	c_hac_3
47	95	98	97	B	NoSabe/NoR	union libre		MuchosDatos	c_hac_3
48	54	67	38	B	Otro	union libre		MuchosDatos	c_hac_3
49	63	83	38	B	VaronTrans	union libre		AgroMisProgramas	c_hac_3
50	63	93	26	B	NoSabe/NoR	union libre		SegurosNoEstan	c_hac_3
51	66	68	28	B	Otro	viudo/a		GrandesAlmacenes	c_hac_3
52	60	94	31	B	Mujer	viudo/a		InteligenciaLaboral	c_hac_3
53	66	78	11	B	MujerTrans	soltero/a		ComercioMuyElectronico	c_hac_3
54	72	84	24	B	Varon	soltero/a		InformaticaDivertida	c_hac_3
55	80	94	16	B	VaronTrans	casado/a		AsocioTodo	c_hac_3
56	74	79	34	B	NoSabe/NoR	casado/a		BanQuito	c_hac_3
57	59	69	31	B	Mujer	casado/a		ClasificoTodo	c_hac_3
58	46	86	35	B	MujerTrans	casado/a		MuchosDatos	c_hac_3
59	52	99	30	B	Varon	casado/a		FaltaTornillo	c_hac_3
60	49	79	27	B	VaronTrans	casado/a		AgroMisProgramas	c_hac_3

Por último, los especialistas de tipo C no fueron la excepción y la gran mayoría se halla en el Cluster 2.

54	72	84	24	B	Varon	soltero/a	InformaticaDivertida	c_hac_3
55	80	94	16	B	VaronTranscasado/a		AsocioTodo	c_hac_3
56	74	79	34	B	NoSabe/NoRcasado/a		BanQuito	c_hac_3
57	59	69	31	B	Mujer	casado/a	ClasificoTodo	c_hac_3
58	46	86	35	B	MujerTranscasado/a		MuchosDatos	c_hac_3
59	52	99	30	B	Varon	casado/a	FaltaTornillo	c_hac_3
60	49	79	27	B	VaronTranscasado/a		AgroMisProgramas	c_hac_3
61	10	8	12	C	NoSabe/NoRdivorciado/a		SegurosNoEstan	c_hac_1
62	74	14	56	C	Mujer	divorciado/a	GrandesAlmacenes	c_hac_2
63	83	26	53	C	MujerTransunion libre		InteligenciaLaboral	c_hac_2
64	72	26	49	C	Varon	union libre	ComercioMuyElectronico	c_hac_2
65	74	43	78	C	VaronTransviudo/a		InformaticaDivertida	c_hac_2
66	81	23	46	C	NoSabe/NoRviudo/a		AsocioTodo	c_hac_2
67	77	36	42	C	Mujer	soltero/a	BanQuito	c_hac_2
68	100	40	55	C	MujerTranssoltero/a		ClasificoTodo	c_hac_2
69	67	43	73	C	Varon	casado/a	MuchosDatos	c_hac_2
70	100	28	41	C	VaronTranscasado/a		MuchosDatos	c_hac_2
71	91	25	79	C	NoSabe/NoRcasado/a		MuchosDatos	c_hac_2
72	72	15	69	C	Otro	casado/a	AgroMisProgramas	c_hac_2

Esto significa que, a través de la comparación de etiquetas previas con los clusters generados, podemos afirmar lo siguiente:

- Cluster 1 = Especialistas de tipo A
- Cluster 2 = Especialistas de tipo C
- Cluster 3 = Especialistas de tipo B

Ejercicio 6)

En base a los análisis realizados, presente un informe en que presente recomendaciones relativas a la captación y conservación de especialistas.

La presente investigación se centra en el análisis de datos relacionados con el grado de satisfacción de especialistas en su trabajo, sueldo y oportunidades de ascenso. Se han recopilado datos de especialistas clasificados en tres tipos (A, B, C), y se cuenta con información adicional como género, estado conyugal, y empresa en la que trabajan. El objetivo del informe es presentar recomendaciones para la captación y conservación de especialistas, basadas en el análisis de los datos disponibles.

Análisis de las variables cuantitativas:

Las variables cuantitativas, como la satisfacción en el trabajo, sueldo y oportunidades de ascenso han sido analizadas para el conjunto de datos completo y para cada tipo de especialidad por separado. Los resultados obtenidos son los siguientes:

- Satisfacción en el trabajo: El rango de valores va desde 9 hasta 100, con un promedio de 54.9. El coeficiente de variación es de 0.48, lo que indica una moderada variabilidad en los niveles de satisfacción en el trabajo.
- Satisfacción en el sueldo: El rango de valores va desde 5 hasta 100, con un promedio de 57.33. El coeficiente de variación es de 0.46, lo que indica una moderada variabilidad en los niveles de satisfacción salarial.
- Satisfacción en las oportunidades de ascenso: El rango de valores va desde 8 hasta 99, con un promedio de 55.57. El coeficiente de variación es de 0.47, lo que indica una moderada variabilidad en los niveles de satisfacción con las oportunidades de ascenso.
- Análisis por tipo de especialidad:
 - Tipo de especialidad A: Se observa una menor satisfacción en el trabajo en comparación con los otros tipos de especialidad, con un promedio de 27.53. Sin embargo, la satisfacción salarial es mayor, con un promedio de 60.33. Las oportunidades de ascenso también son altas, con un promedio de 78.83.
 - Tipo de especialidad B: Se observa una baja satisfacción en el trabajo, con un promedio de 14.53. Sin embargo, la satisfacción salarial es alta,

con un promedio de 80.46. Las oportunidades de ascenso son bajas, con un promedio de 29.4.

- Tipo de especialidad C: Se observa una alta satisfacción en el trabajo, con un promedio de 76.3. Sin embargo, la satisfacción salarial es baja, con un promedio de 31.2. Las oportunidades de ascenso son también bajas, con un promedio de 19.2.

Reglas de asociación encontradas:

Se han identificado varias reglas de asociación que relacionan las diferentes variables de satisfacción. Estas reglas proporcionan información sobre las posibles dependencias entre las variables y pueden ayudar a comprender mejor los factores que influyen en la satisfacción de los especialistas.

Por ejemplo, se encontró una regla que indica que cuando la satisfacción en el trabajo está en el rango de 49 a 66, es probable que la satisfacción con las oportunidades de ascenso esté en el rango de 28 a 42. Esta regla tiene un soporte del 11.11% y una confianza del 62.5%.

Árbol de decisión:

Se ha construido un árbol de decisión para clasificar a los especialistas en función de las variables de satisfacción. El árbol de decisión ha logrado clasificar correctamente la mayoría de los casos para cada tipo de especialidad.

Para el tipo de especialidad A, se clasificaron correctamente 60 de 60 casos.

Para el tipo de especialidad B, se clasificaron correctamente 58 de 60 casos.

Para el tipo de especialidad C, se clasificaron correctamente 60 de 60 casos.

El árbol de decisión proporciona una estructura clara para clasificar a los especialistas en función de sus niveles de satisfacción en el trabajo, sueldo y oportunidades de ascenso.

Clustering:

Se aplicó un algoritmo de clustering y se identificaron 3 clusters. A grandes rasgos, se observa una correspondencia entre los clusters y los tipos de especialidad A, B y C. Sin embargo, se identificaron algunos errores de clasificación.

Cluster 1: Contiene 64 elementos.

Cluster 2: Contiene 52 elementos.

Cluster 3: Contiene 64 elementos.

Recomendaciones:

Basándose en los resultados obtenidos, se realizan las siguientes recomendaciones para la captación y conservación de especialistas:

Tipo de especialidad A: Los especialistas de este tipo muestran una baja satisfacción en el trabajo, pero altos niveles de satisfacción salarial y oportunidades de ascenso. Es importante brindarles un entorno de trabajo estimulante y proporcionarles oportunidades claras de crecimiento y desarrollo profesional.

Tipo de especialidad B: Los especialistas de este tipo presentan una baja satisfacción en el trabajo y en las oportunidades de ascenso, pero tienen altos niveles de satisfacción salarial. Es fundamental evaluar los factores que están afectando su satisfacción en el trabajo y tomar medidas para mejorar el ambiente laboral y las oportunidades de crecimiento.

Tipo de especialidad C: Los especialistas de este tipo muestran una alta satisfacción en el trabajo, pero bajos niveles de satisfacción salarial y oportunidades de ascenso. Se recomienda revisar las políticas salariales y las posibilidades de promoción para asegurar una mayor satisfacción y retención de estos especialistas.

Para todos los tipos de especialidad: Es importante realizar un seguimiento regular de la satisfacción de los especialistas en el trabajo, sueldo y oportunidades de ascenso. Esto permitirá identificar áreas de mejora y tomar medidas preventivas para evitar la pérdida de especialistas.

Conclusiones:

El análisis de los datos proporciona información valiosa sobre la satisfacción de los especialistas y su relación con las variables estudiadas. Las recomendaciones presentadas permiten orientar las acciones de captación y conservación de especialistas, teniendo en cuenta las particularidades de cada tipo de especialidad. Se recomienda realizar un seguimiento continuo de la satisfacción de los especialistas

y ajustar las estrategias según sea necesario para garantizar un entorno laboral satisfactorio y atractivo para los especialistas.

Parte C

Ejercicio 1)

Realice una presentación con diapositivas del Capítulo 2 del libro Andres Fortino - Data Mining and Predictive Analytics_ A Case Study Approach-Mercury Learning and Information (2023) (disponible en la carpeta Libros)

La presentación se adjunta en el archivo zip.

Ejercicio 2)

Presente el proceso de Data Mining para el caso presentado en la Parte B, describiendo cada una de sus etapas.

Proceso de Data Mining en el contexto de Parte B:

- Fase de exploración: En esta etapa, se busca comprender a fondo el problema planteado y los objetivos del análisis. Se define lo que se espera obtener del análisis de los datos y cómo puede ayudar a abordar la escasez de especialistas en el futuro. También se identifican las variables relevantes y las fuentes de datos disponibles. En este caso, la fuente de datos es el archivo "DatosEspecialistasObligatorio2023.txt", que contiene información sobre el grado de satisfacción de los especialistas en su trabajo, sueldo y oportunidades de ascenso, clasificados por tipo de especialidad, género, estado conyugal y empresa en la que trabajan.
- Fase de análisis: En esta etapa, se realiza un análisis inicial de los datos para comprender su estructura, calidad y distribución. Se utilizan técnicas estadísticas y visualizaciones para identificar patrones, tendencias o valores atípicos en los datos. Este análisis exploratorio proporciona una visión general de los datos y ayuda a resaltar posibles problemas o áreas de interés. Dentro de esta fase, se trabajaría en reglas de asociación para descubrir patrones interesantes y relaciones entre los atributos de los especialistas. Además, se construye un modelo de clasificación utilizando algoritmos como árboles de decisión. Por último, se realiza el clustering, que tiene como objetivo agrupar a los especialistas en segmentos homogéneos en función de su satisfacción

laboral. Una vez que se tiene una comprensión más profunda de los datos, se puede pasar a la siguiente fase.

- Fase de interpretación: Con un conocimiento más profundo de los datos, incluyendo las reglas de asociación descubiertas, los clusters identificados y el modelo de clasificación utilizado para predecir la satisfacción laboral de los especialistas, se busca obtener conclusiones significativas basadas en los hechos. Esta etapa implica identificar conclusiones derivadas de los hechos obtenidos durante el análisis.
- Fase de explotación: En esta fase, se presenta un informe que resume los hallazgos del análisis de Data Mining y se formulan recomendaciones relacionadas con la captación y conservación de especialistas en las empresas. El informe puede incluir los resultados de todas las etapas del proceso, junto con análisis adicionales y visualizaciones que respalden las recomendaciones propuestas.

En resumen, el proceso de Data Mining para el caso presentado en la parte B consta de las siguientes etapas: exploración, análisis, interpretación y explotación. Cada etapa cumple un papel importante en la comprensión del negocio, el análisis de los datos, la interpretación de los resultados y la toma de decisiones basadas en datos.

Parte D

En Weka al hacer árbol de decisión J48 se presentan indicadores de precisión en la clasificación en Detailed Accuracy By Class

Ejercicio 1)

Explique el significado de dos de los indicadores de precisión en la clasificación.

En Weka, al realizar un árbol de decisión J48, se proporcionan varios indicadores de precisión en la clasificación bajo la sección "Detailed Accuracy By Class". Estos indicadores brindan información sobre la precisión de la clasificación en diferentes clases o categorías. A continuación, mencionaremos dos de los más comunes:

Precisión: La precisión es una medida de la exactitud de la clasificación en una clase específica. Se calcula dividiendo el número de instancias correctamente clasificadas en una clase por el número total de instancias clasificadas como pertenecientes a esa clase. Es decir, indica la proporción de instancias clasificadas correctamente dentro de una clase en relación con todas las instancias clasificadas en esa clase. Una precisión alta indica una baja tasa de falsos positivos, lo que significa que la mayoría de las instancias clasificadas como pertenecientes a esa clase son correctas.

Recall: El recall es una medida de qué tan bien se identifican todas las instancias pertenecientes a una clase determinada. Se calcula dividiendo el número de instancias correctamente clasificadas en una clase por el número total de instancias reales que pertenecen a esa clase. Es decir, indica la proporción de instancias clasificadas correctamente dentro de una clase en relación con todas las instancias reales de esa clase. Un recall alto indica una baja tasa de falsos negativos, lo que significa que la mayoría de las instancias reales de esa clase se identifican correctamente.

En resumen, estos identificadores son útiles para evaluar el rendimiento del modelo de árbol de decisión en la clasificación de diferentes clases. Un alto valor de precisión indica que el modelo es capaz de clasificar correctamente las instancias en una clase específica, mientras que un alto valor de recall indica que el modelo puede identificar

correctamente la mayoría de las instancias reales de esa clase. Es importante tener un equilibrio entre la precisión y el recall.

Ejercicio 2)

Para los dos indicadores anteriores presente un ejemplo de su significado usando un ejemplo presentado en la Parte A, para que pueda ser comprendido por un usuario 'de negocio'.

Para este ejercicio tomaremos como ejemplo el archivo de Proyectos, que incluso fue el que se utilizó en Weka para hacer un árbol de decisión en la parte A.

Precisión: La precisión mide la proporción de proyectos clasificados correctamente como exitosos en relación con todos los proyectos clasificados como exitosos por el modelo. En la parte A el valor de esta fue de 0,849; llevado a porcentaje sería un 85% aproximadamente. Esto significa que, si el modelo predice que 100 proyectos son exitosos, de esos 100 proyectos, 85 son clasificados correctamente como exitosos. El modelo tiene una alta capacidad para identificar correctamente los proyectos que realmente son exitosos.

Recall: Siguiendo la definición, mide la proporción de proyectos exitosos correctamente identificados por el modelo en relación con todos los proyectos exitosos presentes en los datos. En la parte A el valor fue de 0,838; es decir, aproximadamente un 84%, Esto significa que, si consideramos que hay un total de 150 proyectos exitosos en el conjunto de datos, el modelo logra identificar 125 de ellos. Por lo que tiene una capacidad bastante alta para detectar los proyectos exitosos presentes en los datos.

En resumen, en el contexto de los proyectos, la precisión nos indica qué tan confiable es el modelo al predecir si un proyecto será exitoso o no. Un alto valor de precisión significa que la mayoría de los proyectos clasificados como exitosos por el modelo realmente lo son. Por otro lado, el recall nos indica qué tan completo es el modelo al identificar los proyectos exitosos. Un alto valor de recall indica que el modelo logra capturar la mayoría de los proyectos exitosos presentes en los datos.

Ambos indicadores son importantes en el contexto de los negocios, ya que permiten evaluar la efectividad del modelo en la clasificación de proyectos exitosos. Dependiendo de los objetivos y las necesidades del negocio, se puede dar más importancia a la precisión (evitar falsos positivos) o al recall (evitar falsos negativos). Por ejemplo, si el costo de un proyecto fallido es alto, se puede priorizar una alta precisión para evitar asignar recursos a proyectos que tienen una alta probabilidad de fracaso.

Parte E

Realice un estudio comparativo de las herramientas utilizadas para desarrollar este obligatorio en base su experiencia.

Ejercicio 1)

Elabore un cuadro comparativo en que en función de diferentes ítems asigne puntajes de 0 a 10. Puede ponderar los ítems si lo considera pertinente

Herramienta	A	B	C	D	E
Weka	8	7	9	8	8
Tanagra	7	6	7	7	6
Knime	9	9	8	9	9

A continuación, describiremos cada una de las columnas:

- Diversidad de datos aceptados (A): Se refiere a la capacidad de la herramienta para trabajar con una amplia variedad de tipos de datos. Weka tiene una puntuación alta en este aspecto, ya que puede manejar datos estructurados, no estructurados y semi-estructurados. Tanagra también puede manejar diferentes tipos de datos, aunque su enfoque principal es en datos estructurados. Knime tiene una excelente capacidad para trabajar con diversos tipos de datos, incluyendo datos tabulares, imágenes, textos y más.
- Usabilidad (B): Este ítem se refiere a la facilidad de uso de la herramienta. Weka obtiene una puntuación alta en usabilidad, ya que ofrece una interfaz gráfica intuitiva y una amplia documentación. Tanagra es un poco menos intuitivo en comparación con Weka, pero sigue siendo relativamente fácil de usar. Knime se destaca en usabilidad, ya que ofrece una interfaz visual amigable para arrastrar y soltar componentes, lo que facilita la construcción de flujos de trabajo.

- Herramientas y Algoritmos (C): Se refiere a la disponibilidad y variedad de herramientas y algoritmos ofrecidos por cada uno. Weka obtiene una alta puntuación, ya que ofrece una amplia gama de algoritmos de aprendizaje automático y técnicas de preprocesamiento de datos. Tanagra también ofrece una buena selección de herramientas y algoritmos, aunque puede tener menos opciones que Weka. Knime tiene una amplia biblioteca de nodos y extensiones que permiten la integración de diferentes algoritmos y herramientas.
- Nivel de soporte oficial (D): Este ítem se refiere al nivel de soporte proporcionado oficialmente por los desarrolladores de la herramienta. Weka tiene un buen nivel de soporte oficial, con una comunidad activa y una documentación sólida. Tanagra ofrece soporte oficial a través de la Universidad de Lyon, aunque puede tener menos recursos y actualizaciones frecuentes en comparación con otras herramientas. Knime tiene un buen nivel de soporte oficial, con una comunidad activa y una amplia documentación.
- Nivel de soporte por la comunidad (E): Este ítem se refiere al nivel de soporte proporcionado por la comunidad de usuarios de la herramienta. Weka tiene una comunidad activa y numerosos foros de discusión donde los usuarios pueden obtener ayuda y compartir conocimientos. Tanagra tiene una comunidad más pequeña en comparación con Weka, pero sigue siendo posible encontrar recursos y ayuda en línea. Knime cuenta con una comunidad activa y ofrece soporte a través de foros y grupos de discusión en línea.

Ejercicio 2)

Si tuviera que seleccionar dos herramientas, ¿cuáles seleccionaría? Fundamente su elección

Si tuviera que seleccionar dos herramientas, elegiría Weka y Knime. Ambas herramientas obtuvieron puntajes altos en el cuadro comparativo en varios aspectos clave. Weka tiene una amplia diversidad de datos aceptados y una gran variedad de algoritmos y herramientas disponibles. También cuenta con un buen nivel de soporte oficial y por la comunidad. Knime también se destaca en diversidad de datos aceptados y ofrece una interfaz visual intuitiva, lo que la hace fácil de usar. Además, tiene una amplia biblioteca de nodos y extensiones, lo que le brinda una buena selección de herramientas y algoritmos. Ambas herramientas ofrecen soporte oficial y por la comunidad, lo que significa que puedes encontrar recursos y ayuda en línea cuando lo necesites.

Ejercicio 3)

Si tuviera que descartar una herramienta, ¿cuál descartaría? Fundamente su elección

Si tuviera que descartar una herramienta, descartaría Tanagra. Aunque Tanagra tiene algunas características destacables, como la capacidad de manejar diferentes tipos de datos y ofrecer soporte oficial a través de la Universidad de Lyon, obtuvo puntajes más bajos en comparación con Weka y Knime en varios aspectos clave. Su usabilidad y variedad de algoritmos pueden ser más limitados en comparación con las otras herramientas. Además, su nivel de soporte por la comunidad puede ser menos activo en comparación con Weka y Knime.

Parte F

De los materiales estudiados para el primer parcial y para este obligatorio: cada uno de los integrantes del grupo indique:

Ejercicio 1)

Cuál le resultó más interesante indicando el motivo

Tomás Clavijo: Considero que el tema más interesante del curso fue Clustering. Siendo lo que más me impresionó la capacidad de interpretar y comprender el significado detrás de un dendrograma. El hecho de poder visualizar y analizar las estructuras jerárquicas de los datos agrupados me permitió obtener una comprensión más profunda de los patrones y relaciones existentes. Además, las técnicas de validación utilizadas para evaluar la calidad de los clusters me parecieron fundamentales para garantizar que los resultados sean confiables y efectivos. Destacar su aplicabilidad práctica en el entorno empresarial. Poder categorizar y agrupar a los clientes en función de características comunes ofrece un enorme potencial para la personalización y optimización de las estrategias de negocio.

Sofía Barreto: El tema que más me gusto fueron las redes neuronales. Las encuentro interesantes por su lado misterioso, especialmente por las partes llamadas "capas ocultas". Aunque sabemos mucho sobre ellas, hay cosas que todavía no entendemos del todo. Esa es la parte que más me gusta, el hecho de que siempre hay algo nuevo para descubrir.

Ejercicio 2)

Cuál le resultó menos interesante indicando el motivo

Tomás Clavijo: Considero que redes neuronales fue el tema que despertó un menor nivel de interés en mí. Sin embargo, esto no significa que no reconozca su importancia y relevancia en el campo del aprendizaje automático. La principal razón por la que las redes neuronales resultaron ser menos interesantes para mí fue que ya tenía cierto conocimiento previo sobre el tema. Había investigado y explorado en asignaturas anteriores, adquiriendo una comprensión básica de su funcionamiento. A pesar de lo dicho, considero que su capacidad para simular el funcionamiento del cerebro humano y su flexibilidad para modelar relaciones complejas entre variables son características impresionantes.

Sofía Barreto: El tema que menos me gusto son los árboles de decisión. Mientras que son una herramienta útil y fácil de entender, a veces encuentro que no son tan eficientes como otros métodos.

Bibliografía

- Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (s. f.). Weka. Recuperado de <https://www.cs.waikato.ac.nz/ml/weka/> (26 de junio de 2023)
- Ag, K. Z. (s. f.). KNIME Documentation. KNIME. Recuperado de <https://docs.knime.com> (26 de junio de 2023)
- Ricco, S. (2005). Le logiciel Tanagra (EGC'2005). Recuperado de https://eric.univlyon2.fr/ricco/tanagra/fichiers/le_logiciel_tanagra_egc_2005.pdf . (27 de junio de 2023)
- Universidad Europea. (2022). Aprendizaje supervisado y no supervisado. Recuperado de <https://universidadeuropea.com/blog/aprendizaje-supervisado-no-supervisado/> (28 de junio de 2023)
- Fortino, A. (2023). Capítulo 2. Data Mining and Predictive Analytics: A Case Study Approach. Mercury Learning and Information.
- Kirkos, E., Spathis, C., Nanopoulos, A., & Manolopoulos, Y. (2007). Identifying Qualified Auditor's Opinions: A Data Mining Approach. Journal of Emerging Technologies in Accounting
- Kantardzic, M. (2020). Capítulo 1: [Data-Mining Concepts]. Data mining: Concepts, models, methods, and algorithms (3rd ed., pp. 01-31). John Wiley & Sons.
- Bhatia, P. (2019). Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press.