

Práctica 1b de Minería de Datos

Preparación de la Vista Minable con



Curso de Postgrado
Minería de Datos

Máster y Postgrado del DSIC
Universitat Politècnica de València

José Hernández Orallo. (jorallo@dsic.upv.es). Diciembre 2006

Índice

1.	Limpeza y transformación de datos	2
1.1	Un problema con datos erróneos y faltantes: De nuevo la “Agrupación de Empleados”.....	2
1.2	Resolución.....	2
2.	Obtención y transformación de datos a partir de bases de datos y almacenes de datos vía ODBC.....	27

En esta práctica se trabaja sobre la limpieza, integración y transformación de datos, así como la conexión con un almacén de datos, con el objetivo de aprender a configurar correctamente la vista minable.

1. Limpieza y transformación de datos

1.1 *Un problema con datos erróneos y faltantes: De nuevo la “Agrupación de Empleados”*

La empresa de software para Internet “Memolum Web” ha aumentado vertiginosamente su plantilla en los últimos años, debido principalmente a una absorción de la compañía “Intelligent Stones” y quiere ver si las tipologías de empleados existentes siguen valiendo. Las variables que se recogen de las fichas de los 40 empleados actuales de la empresa son:

- Sueldo: sueldo anual en euros.
- Casado: si está casado o no.
- Coche: si viene en coche a trabajar (o al menos si lo aparca en el parking de la empresa).
- Hijos: si tiene hijos.
- Alq/Prop: si vive en una casa alquilada o propia.
- Sindic.: si pertenece al sindicato revolucionario de Internet
- Bajas/Año: media del nº de bajas por año
- Antigüedad: antigüedad en la empresa
- Sexo: H: hombre, M: mujer.
- **Estudios: Obl: obligatorios, Bac: bachillerato, FP: formación profesional, Uni: universitarios, Doc: doctorado/master**

Como vemos, existe un nuevo campo “Estudios”, que recoge los estudios realizados por el trabajador. Los datos de los 40 empleados se encuentran en el directorio “..\LabKDD\empleados\empleados4.txt”). El problema de estos datos es que, al haber crecido vertiginosamente la compañía y debido a la conversión de datos en la fusión, pueden tener un número significativo de datos faltantes o erróneos. Se intenta extraer grupos de entre estos quince empleados.

1.2 *Resolución*

El objetivo de este punto es, fundamentalmente, ver que normalmente hay que dedicar bastante tiempo a conseguir que los datos estén de la manera más idónea para la modelización. Incluso en este ejemplo sencillo, veremos que podemos (o debemos) hacer muchas cosas antes de empezar a utilizar nodos de modelado. De hecho, este ejercicio es largo así que cada cierto tiempo **graba la ruta en un fichero “.str”, p.ej. “empleados2.str”**.

En primer lugar vamos a leer los datos de los empleados. Limpiamos la zona de trabajo o empezamos una ruta nueva y volvemos a conectar con el fichero “empleados4.txt” de manera similar al caso anterior:

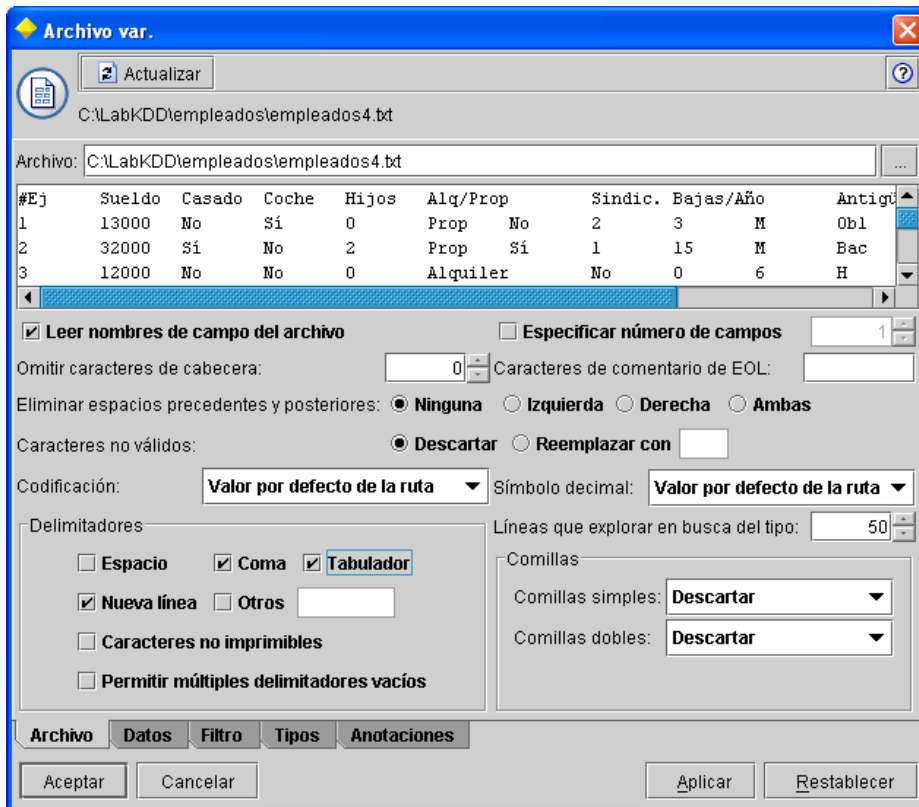


Figura 1. Leyendo del fichero de datos.

Si enganchamos un nodo "Tabla" a la salida de este nodo, podemos ver si los datos se leen correctamente.

Tabla (11 campos, 41 registros) #1

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Estudios
1	13000	No	Si	0	Prop No	No	2	3	M	Obl
2	32000	Si	No	2	Prop Si	No	1	15	M	Bac
3	12000	No	No	0	Alquiler	No	0	6	H	Obl
4	41000	Si	Si	3	Prop No	No	3	13	H	Uni
5	5000	No	No	0	Si	No	0	1	H	
6	65000	No	Si	0	Prop No	No	3	8	M	Doc
7	53000	Si	Si	5	Prop No	No	4	18	M	Bac
8	23000	No	Si	0	Alquiler	Si	7	2	H	
9	31000	Si	No	0	Prop Si	No	0	5	H	Bac
10	30000	Si	Si	2	Prop No	No	1	20	H	Bac
11	20000	No	Si	1	Alquiler	Si	3	3	M	Uni
12	13000	No	No	0	Prop No	No	12	2	H	Bac
13	11000	No	Si	0	Alquiler	No	0	7	H	FP
14	9000	No	Si	1	Prop Si	No	2	3	H	FP
15	60000	Si	Si	4	Prop No	No	0	10	M	Uni
16	380000	No	Si	0	Prop No	No	2	5	H	Uni
17	6000	No	Si	0	No	No	0	1	H	Obl
18	30000	Si	Si	-7	Prop Si	No	10	10	H	Uni
19	23000	No	Si	0	Prop No	No	2	4	M	Bac
20	43000	No	Si	3	Alquiler	Si	20	7	H	Uni
21	13000	No	Si	0	Alquiler	Si	3	3	M	FP
22	21000	Si	Si	1	Prop No	No	1	7	M	Bac
23	15000	Si	Si	2	Prop Si	No	5	10	H	Obl
24	30000	Si	Si	1	Alquiler	No	15	7	M	Uni
25	10000	Si	Si	0	Prop Si	No	1	6	H	
26	40000	No	Si	0	Alquiler	Si	3	16	M	Bac
27	25000	No	No	0	Alquiler	Si	0	8	H	Bac
28	20000	No	Si	0	Si	No	2	6	M	Bac
29	20000	Si	Si	3	Prop No	No	7	5	H	Obl
30	10000	Si	No	0	Alquiler	No	7	4	H	
31	50000	No	No	0	Alquiler	No	2	12	M	Doc
32	8000	Si	Si	2	Prop No	No	3	1	H	Obl
33	20000	No	No	0	Alquiler	No	27	5	M	Bac
34	10000	No	Si	0	Alquiler	Si	0	7	H	Obl
35	8000	No	Si	0	Alquiler	No	3	2	H	FP
36	50000	Si	Si	1	Prop No	No	1	12	H	Doc
37	7000	No	Si	1	Prop Si	No	1	2	M	Obl
38	30000	Si	Si	2	Prop Si	No	10	8	H	FP
39	32000	No	No	0	Prop No	No	2	3	M	Uni
40	33000	No	Si	3	Prop No	No	5	7	H	Uni
41	\$n...	\$null\$		\$null\$			\$null\$	\$null\$		

Tabla Anotaciones

Figura 2. Mostrando los datos tabularmente.

Ya simplemente a primera vista vemos datos faltantes. Pero además, vemos que existe un registro 41 con todos los valores a \$null\$. De hecho, podemos ver que en el Clementine aparece un símbolo de advertencia en la parte de abajo, que si pinchamos nos detalla los mensajes:

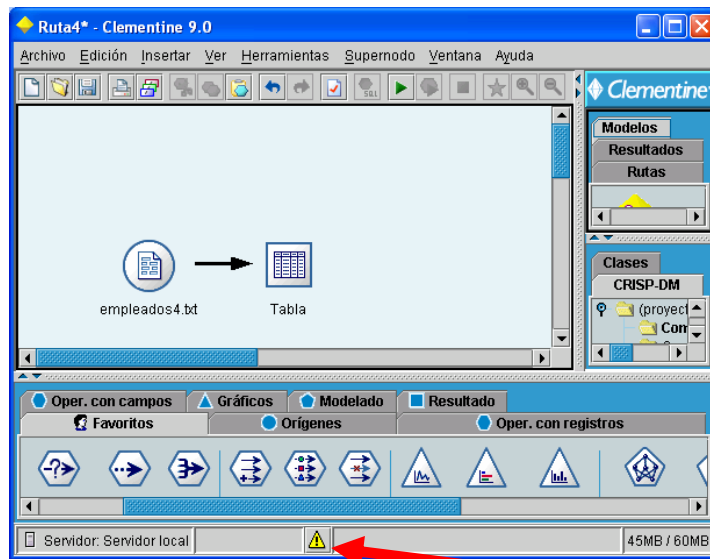


Figura 3. Mensaje de advertencia.

Esta fila “extra” se puede deber a un salto de línea final en el fichero de datos. Podemos arreglarlo en el fichero original pero lo podemos arreglar directamente en el Clementine, mediante un nodo “Seleccionar”. Una vez conectado al nodo de origen se edita de la siguiente forma, poniendo el campo “@NULL(‘#Ej’)” tal cual. Las funciones posibles del Clementine ya las trataremos más adelante (si quieres verlas, puedes pinchar en el icono azul en forma de calculadora y se abrirá el “generador de expresiones” donde verás todas las funciones que hay).

Ojo que hay también que pinchar en “Descartar”:

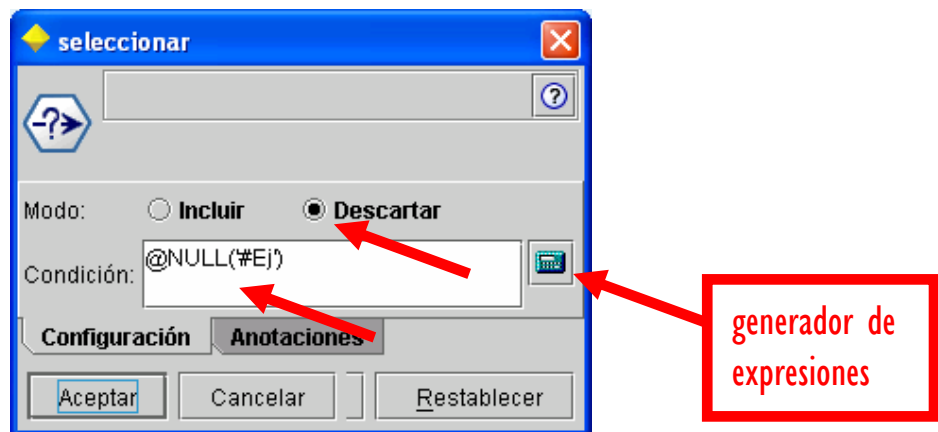


Figura 4. Estableciendo una condición para eliminar la fila que tiene un nulo en #Ej.

Ahora conectamos el nodo tabla a la salida de este nodo seleccionar y volvemos a mirar los datos que tenemos.

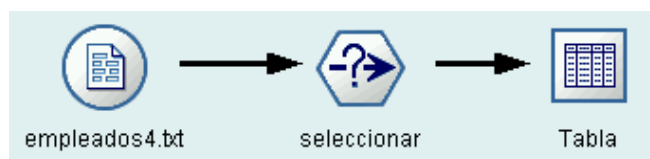


Figura 5. Ruta que importa ya los datos sin el último caso vacío.

Ahora ya vemos sólo 40 registros.

Ahora pasemos a los atributos que pueden ser faltantes o erróneos. Podemos analizar cuántos y para cuántos atributos, de una manera automática, con el nodo "Calidad". Si añadimos uno y lo conectamos con la fuente de datos (después del nodo "seleccionar") tenemos:

Campo	% completado	Registros válidos
#Ej	100	40
Sueldo	100	40
Casado	100	40
Coche	100	40
Hijos	100	40
Alq/Prop	92,5	37
Sindic.	100	40
Bajas/Año	100	40
Antigüedad	100	40
Sexo	100	40
Estudios	90	36

Figura 6. Porcentaje de datos faltantes por campo.

Vemos que faltan valores en el campo Alq/Prop y en el campo Estudios.

Una vez detectados los valores faltantes, podemos pasar a ver los valores anómalos. Para ello es muy útil el nodo "Auditar Datos". Si lo conectamos como hemos hecho con el de Calidad, es decir como se muestra en la siguiente figura:

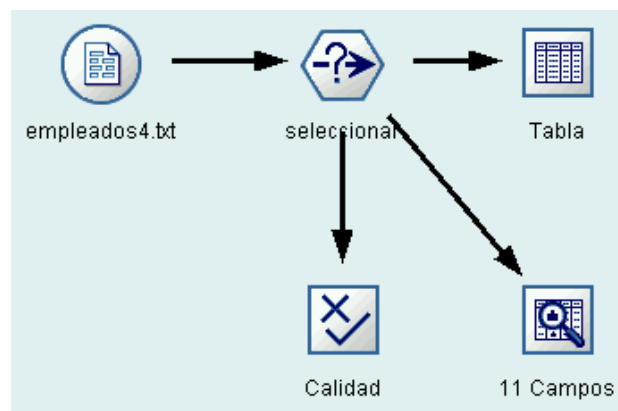


Figura 7. Ruta con nodo de Calidad y con nodo de Auditar.

Y lo ejecutamos, tenemos:

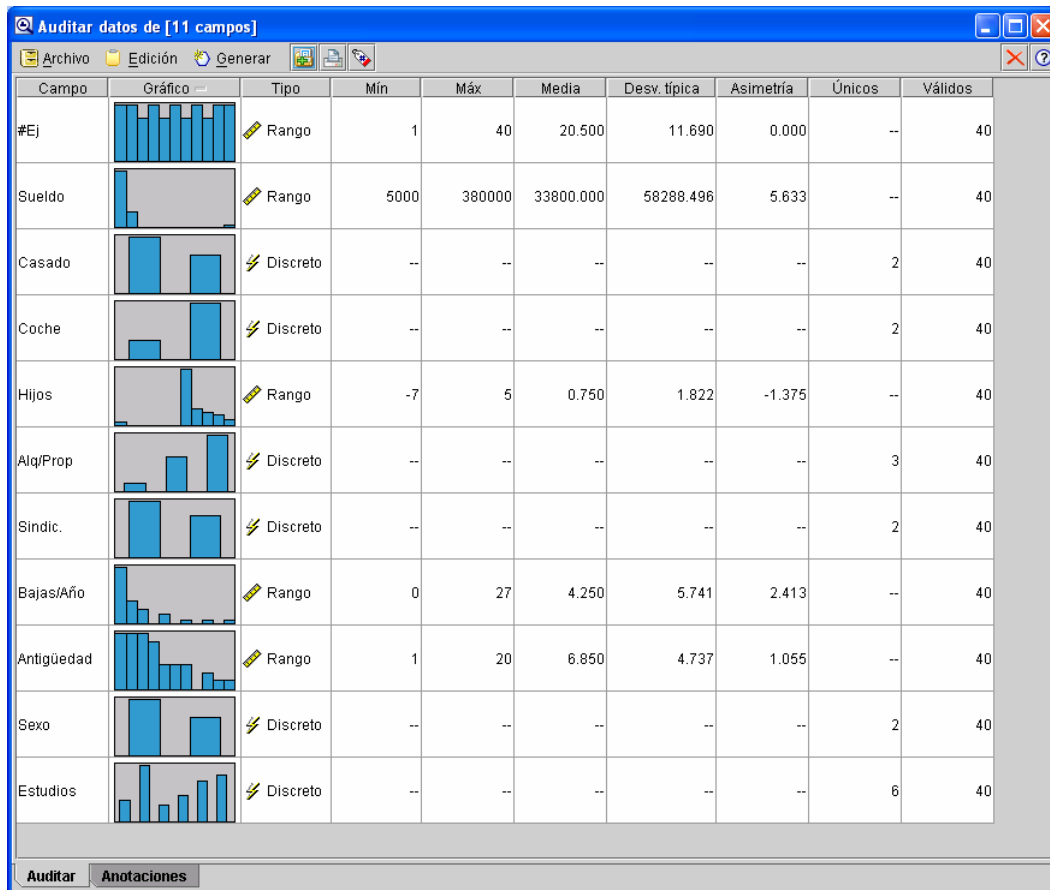


Figura 8. Resultado del nodo “Auditar datos”.

La figura anterior nos muestra una foto bastante detallada de los valores mínimos, máximos, media y desviación de los datos, así como un gráfico con su distribución. Como podemos observar, hay dos gráficos de distribución bastante llamativos, el del sueldo, donde parece haber un valor mucho mayor que el resto y, especialmente, en el campo “Hijos”, donde tenemos un valor mínimo que es negativo. En el nodo “Auditar datos” los valores vacíos (faltantes) se cuentan para los nominales. Por eso, vemos que existen tres categorías para “Alq/Prop”. En “Estudios” también hay una más, aunque eso lo sabemos por el nodo de Calidad anterior.

Podemos investigar los valores anómalos con más detalle, mediante el uso de gráficas por campos. Nos interesa ver los datos numéricos, ya que en estos casos es más fácil detectar *outliers* gráficamente. Para ello podemos añadir ciertos nodos “Gráfico” para comparar diferentes valores. Por ejemplo, podemos querer visualizar la antigüedad respecto al sueldo y además mostrando los estudios. Añadimos un nodo “Gráfico” y lo enganchamos con la fuente de datos (ver Figura 16). Si lo editamos podemos especificar que el campo X será “Sueldo”, el campo Y será “Antigüedad” y el campo de Superponer será “Estudios” (por Color y Forma):

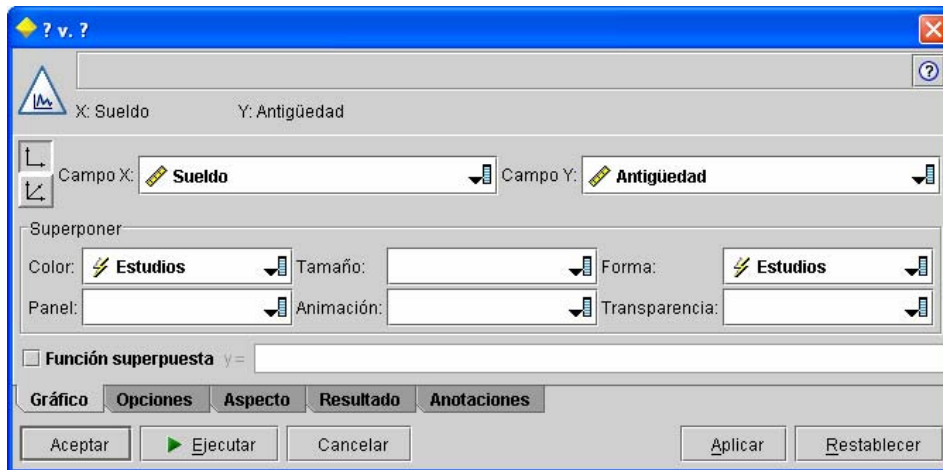


Figura 9. Configurando un nodo “Gráfico”.

Si lo ejecutamos es muy posible que nos encontremos con el siguiente problema:

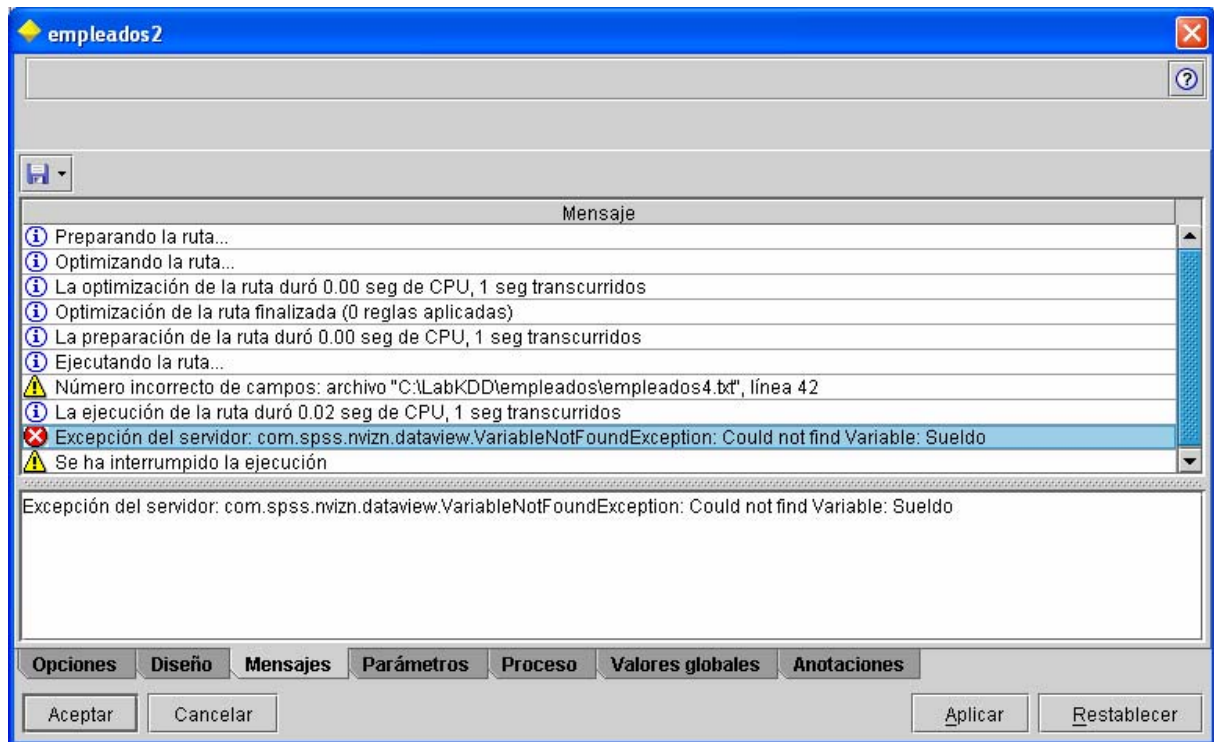


Figura 10. Error aparentemente espeluznante.

Si no caemos presa del pánico y miramos detenidamente el error, vemos que nos dice que no encuentra la variable “Sueldo”, cuando, supuestamente, existe un campo llamado así.

El detalle es difícil de encontrar, así que vamos a dar unas pistas. Desconecta el nodo de “Gráfico” de su antecesor y añade un nodo “Filtro” y únelo con el nodo “seleccionar”, como se muestra en la siguiente figura.

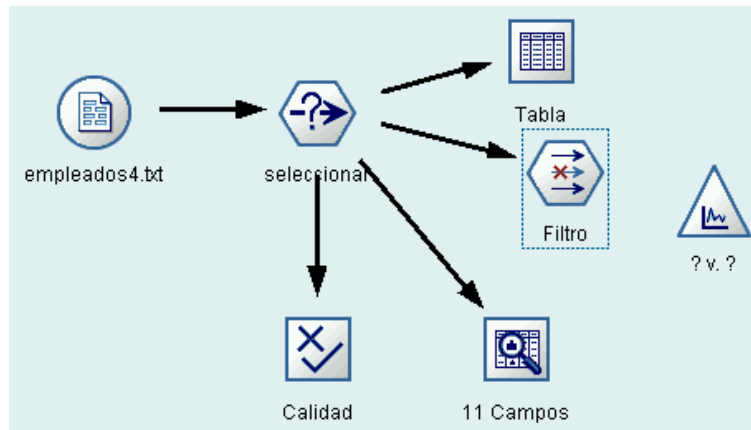


Figura 11. Solventando el problema del campo “Sueldo”.

Ahora editamos el nodo filtro (que sirve para renombrar y filtrar campos) y miramos el nombre “Sueldo”, a ver si tuviera algún misterio.



Figura 12. Solventando el problema del campo “Sueldo”.

Efectivamente; hay un problema. El nombre del campo no es “Sueldo” sino “Sueldo ”, es decir, se nos ha colado un espacio extra en el nombre del campo (proviene del fichero). La solución, una vez detectado el problema, es sencilla. En esa misma pantalla, le quitamos el último espacio al nombre del campo “Sueldo ”, para que sea “Sueldo”.

Ahora conectamos el nodo “Filtro” con el “Gráfico” y editamos éste, dónde volveremos a seleccionar el campo “Sueldo”, como hicimos en la Figura 9. Por fin, ejecutamos y tenemos la siguiente gráfica:

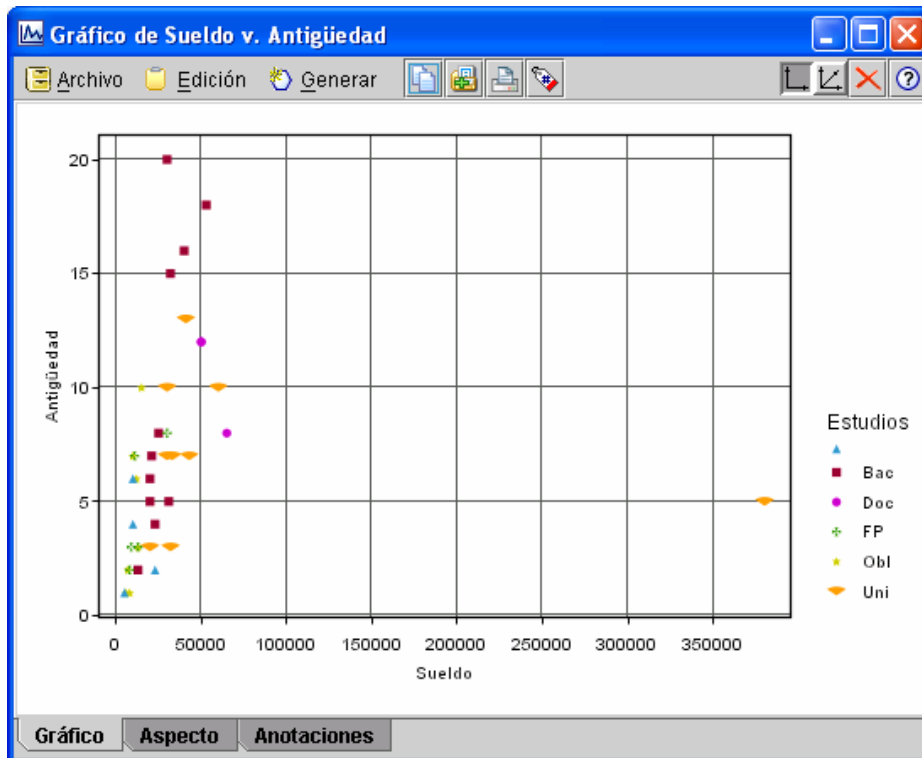


Figura 13. Gráfica: Antigüedad x Sueldo [Estudios].

Parece destacar un dato cerca de los 400000€ de sueldo que parece ser un dato anómalo, dadas las características de la empresa (o puede ser el sueldo del jefe).

De modo similar podemos añadir y conectar otro nodo Gráfico para intentar relacionar el nº de Hijos con las Bajas mostrando el Sexo como "superponer". El resultado es el siguiente.

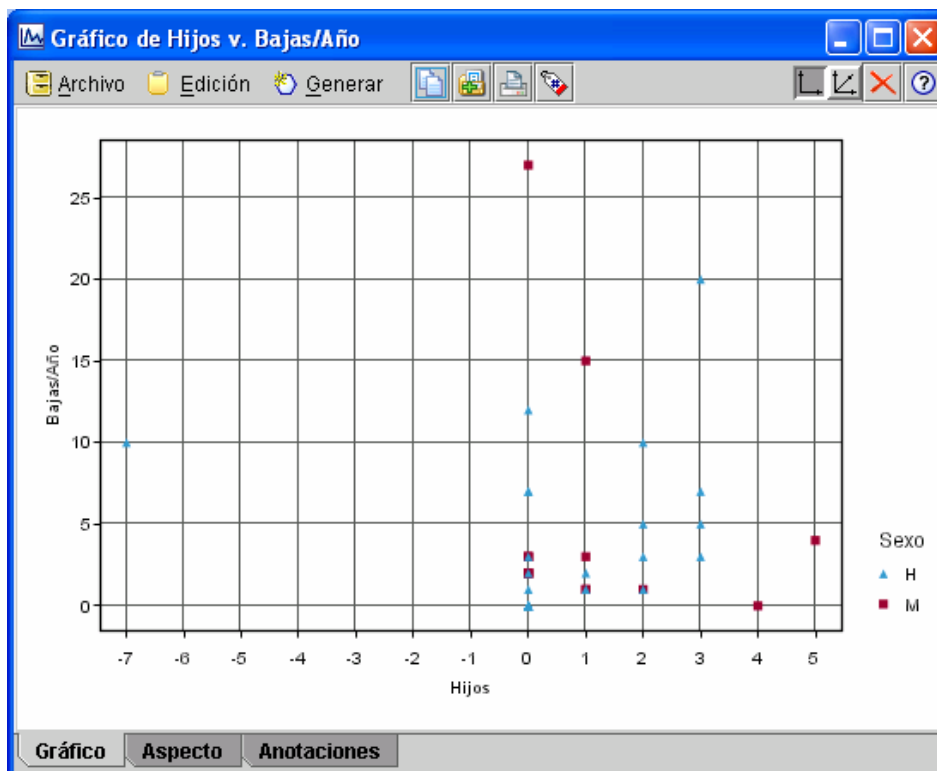


Figura 14. Gráfica "Gráfica": Hijos x Bajas/Año [Sexo].

También destaca claramente el hecho de que existe un registro con el nº de hijos negativo, lo cual es claramente un dato erróneo.

Podemos añadir más Gráficos, por ejemplo, uno que combine “Antigüedad” x “Sueldo” y de campo superpuesto tenga el campo “Casado”

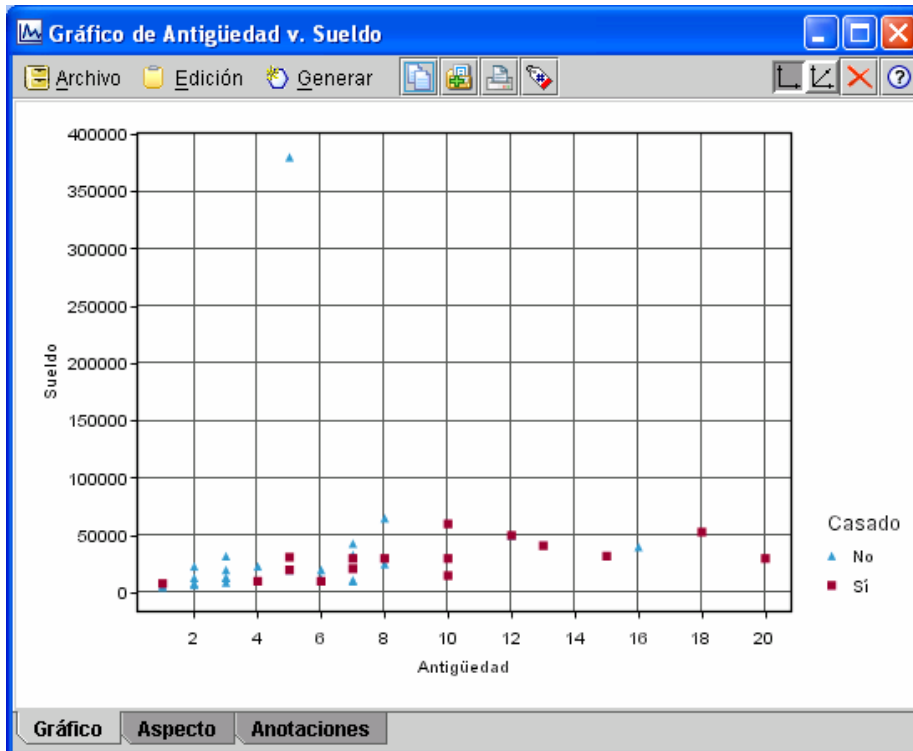


Figura 15. Gráfica: Antigüedad x Sueldo [Casado].

Que, de momento, no nos aporta demasiada información, aparte que la antigüedad parece estar relacionada con estar casado.

La ruta que llevamos hasta el momento es la que se muestra en la siguiente figura:

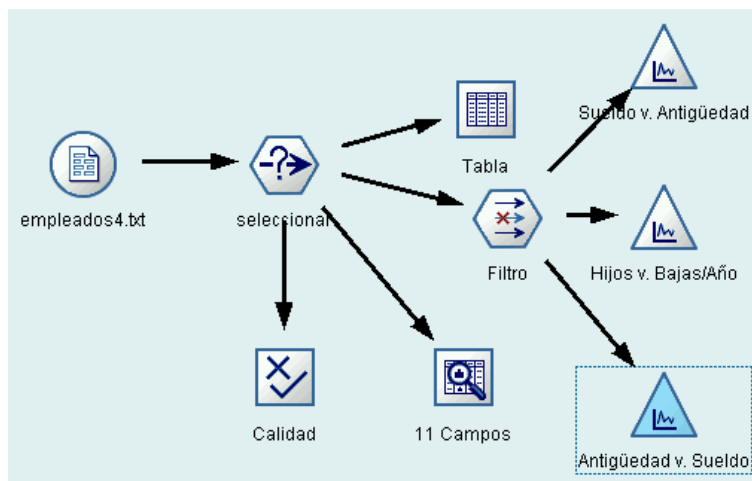


Figura 16. Ruta de los empleados con datos imperfectos.

Siguiendo con nuestro problema, hemos identificado dos campos con valores faltantes y dos registros anómalos (el de los hijos negativos y el del sueldo descomunal). Ya que tenemos 40 empleados, puede resultar conveniente en este caso eliminar los dos registros anómalos, pues los resultados con 38 empleados van a ser similares que con 40 (y más fiables al haber eliminado registros dudosos).

Para eliminarlos, ejecutamos el nodo "Tabla" que ya teníamos en la zona de trabajo y pinchamos sobre los datos anómalos: en este caso el sueldo del registro 16 y los hijos del registro 18, como se muestra en la siguiente figura (usa la tecla CTRL para seleccionar ambos).

	#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Estudios
1	1	13000	No	Sí	0	Prop	No	2	3	M	Obl
2	2	32000	Sí	No	2	Prop	Sí	1	15	M	Bac
3	3	12000	No	No	0	Alquiler	No	0	6	H	Obl
4	4	41000	Sí	Sí	3	Prop	No	3	13	H	Uni
5	5	5000	No	No	0		Sí	0	1	H	
6	6	65000	No	Sí	0	Prop	No	3	8	M	Doc
7	7	53000	Sí	Sí	5	Prop	No	4	18	M	Bac
8	8	23000	No	Sí	0	Alquiler	Sí	7	2	H	
9	9	31000	Sí	No	0	Prop	Sí	0	5	H	Bac
10	10	30000	Sí	Sí	2	Prop	No	1	20	H	Bac
11	11	20000	No	Sí	1	Alquiler	Sí	3	3	M	Uni
12	12	13000	No	No	0	Prop	No	12	2	H	Bac
13	13	11000	No	Sí	0	Alquiler	No	0	7	H	FP
14	14	9000	No	Sí	1	Prop	Sí	2	3	H	FP
15	15	60000	Sí	Sí	4	Prop	No	0	10	M	Uni
16	16	380000	No	Sí	0	Prop	No	2	5	H	Uni
17	17	6000	No	Sí	0		No	0	1	H	Obl
18	18	30000	Sí	Sí	-7	Prop	Sí	10	10	H	Uni
19	19	23000	No	Sí	0	Prop	No	2	4	M	Bac
20	20	43000	No	Sí	3	Alquiler	Sí	20	7	H	Uni
21	21	13000	No	Sí	0	Alquiler	Sí	3	3	M	FP
22	22	21000	Sí	Sí	1	Prop	No	1	7	M	Bac
23	23	15000	Sí	Sí	2	Prop	Sí	5	10	H	Obl
24	24	30000	Sí	Sí	1	Alquiler	No	15	7	M	Uni
25	25	10000	Sí	Sí	0	Prop	Sí	1	6	H	
26	26	40000	No	Sí	0	Alquiler	Sí	3	16	M	Bac
27	27	25000	No	No	0	Alquiler	Sí	0	8	H	Bac

Figura 17. Marcando los datos anómalos.

Pulsando en el menú "Generar", seleccionamos la opción "Nodo Seleccionar ("O")":

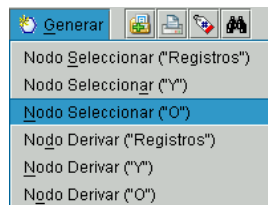


Figura 18. Generando un nodo para eliminar estos registros anómalos.

Parece que nada ha ocurrido, pero si vamos a la zona de trabajo veremos que nos ha generado un nodo "seleccionar" con el nombre "(generado)". Enganchamos el nodo fuente (empleados4.txt) con él y editamos el nodo "(generado)", para comprobar que las condiciones se han generado bien y para modificar el modo a "Descartar" (porque son los que queremos eliminar), como se muestra en la siguiente figura:

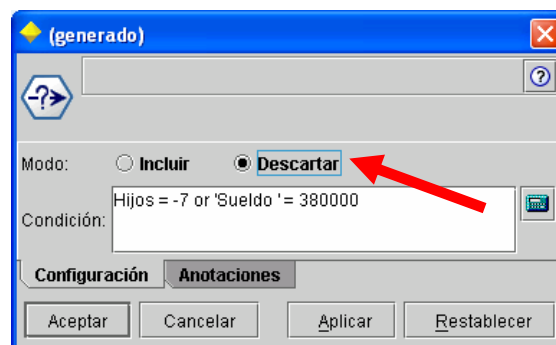


Figura 19. Modificando el nodo select para eliminar estos registros anómalos.

Podemos añadir un nodo “Tabla” y engancharlo con el nodo “(generado)” para comprobar que realmente elimina esos dos registros, resultando en 38 registros de datos.

Una vez resuelto el tema de los datos anómalos vamos a ocuparnos de los datos faltantes (que Clementine denomina “Perdidos”). En primer lugar vamos a abordar el campo “Estudios”. Según se tiene conocimiento de la manera de adquirir este dato (en el momento de contratación en la empresa) es muy posible que la ausencia de valor en este campo pueda significar que el contratado no tenía estudios aparte de los elementales, dejando en blanco este campo. Por tanto, vamos a suponer que aquellos registros sin estudios van a ser realmente “Estudios obligatorios”. Para arreglarlo, y siguiendo este criterio, vamos a añadir un nodo “Relleno”. No obstante, previamente debemos añadir un nodo “Tipo” y conectarlo al nodo “(Generado)” (ver Figura 22).

Editamos el nodo “Tipo” y en la columna “Perdidos” del campo “Estudios”, pinchamos y seleccionamos “Activado (*)”, como se muestra en la siguiente figura:

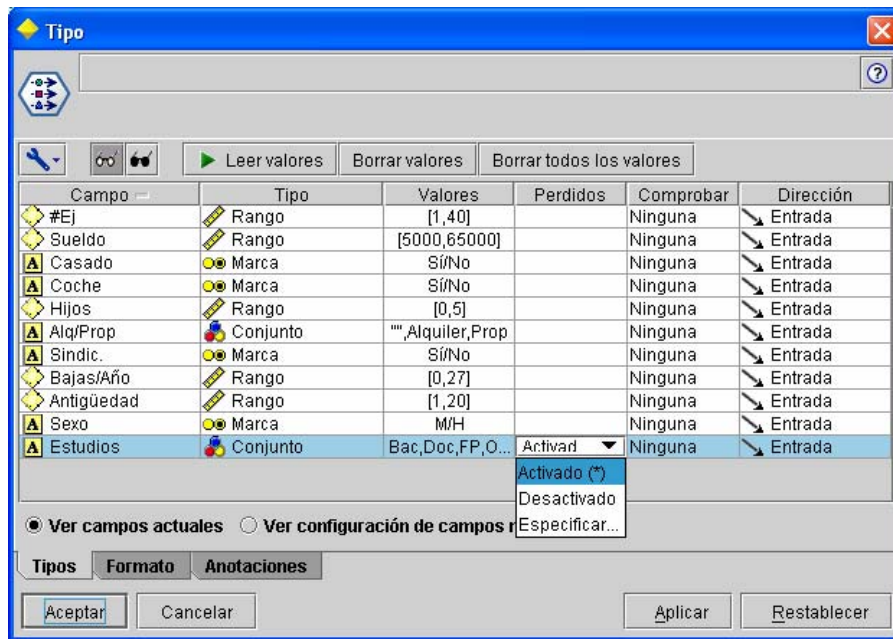


Figura 20. Tipando los atributos blancos del campo “Estudios”.

Esto hace que ahora marque como perdidos los que están vacíos. Y ahora sí que enganchemos el nodo “Relleno” con el nodo “Tipo”. Editamos el nodo “Relleno” como se muestra en la siguiente figura:



Figura 21. Modificando el nodo “relleno” para rellenar blancos con valores.

Añadimos un nodo “Tabla” y tenemos la ruta de la siguiente figura:

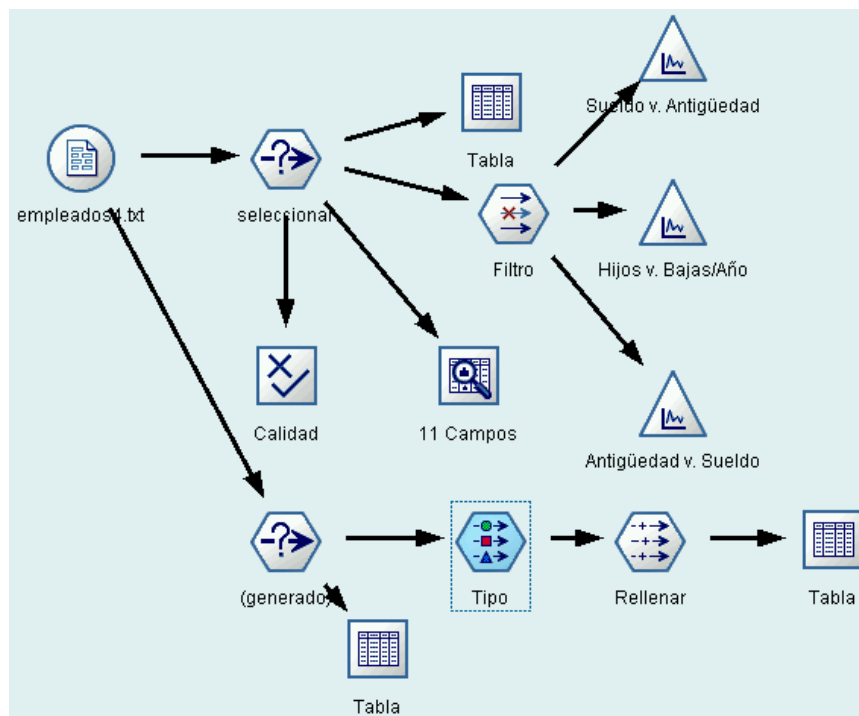


Figura 22. Estado de la ruta.

Si ejecutamos el último nodo “Tabla” podemos ver que ya no hay datos faltantes en la columna “Estudios” y que los faltantes se han rellenado con “Obl”.

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Estudios
1	13000	No	Sí	0	Prop	No	2	3	M	Obl
2	32000	Sí	No	2	Prop	Sí	1	15	M	Bac
3	12000	No	No	0	Alquiler	No	0	6	H	Obl
4	41000	Sí	Sí	3	Prop	No	3	13	H	Uni
5	5000	No	No	0		Sí	0	1	H	Obl
6	65000	No	Sí	0	Prop	No	3	8	M	Doc
7	53000	Sí	Sí	5	Prop	No	4	18	M	Bac
8	23000	No	Sí	0	Alquiler	Sí	7	2	H	Obl
9	31000	Sí	No	0	Prop	Sí	0	5	H	Bac
10	30000	Sí	Sí	2	Prop	No	1	20	H	Bac
11	20000	No	Sí	1	Alquiler	Sí	3	3	M	Uni
12	13000	No	No	0	Prop	No	12	2	H	Bac
13	11000	No	Sí	0	Alquiler	No	0	7	H	FP
14	9000	No	Sí	1	Prop	Sí	2	3	H	FP
15	60000	Sí	Sí	4	Prop	No	0	10	M	Uni
16	6000	No	Sí	0		No	0	1	H	Obl
17	23000	No	Sí	0	Prop	No	2	4	M	Bac
18	43000	No	Sí	3	Alquiler	Sí	20	7	H	Uni
19	13000	No	Sí	0	Alquiler	Sí	3	3	M	FP
20	21000	Sí	Sí	1	Prop	No	1	7	M	Bac
21	15000	Sí	Sí	2	Prop	Sí	5	10	H	Obl
22	30000	Sí	Sí	1	Alquiler	No	15	7	M	Uni
23	10000	Sí	Sí	0	Prop	Sí	1	6	H	Obl
24	40000	No	Sí	0	Alquiler	Sí	3	16	M	Bac
25	25000	No	No	0	Alquiler	Sí	0	8	H	Bac

Figura 23. Datos con el campo “Estudios” rellenado.

Ahora ya sólo nos queda abordar los blancos en “Alq/Prop”. Al haber tres casos con nulo en este atributo, no parece aconsejable eliminar tres registros, ya que su información puede ser valiosa. La idea es intentar rellenarlos con algún valor relativamente *razonable*. Una idea sería ver qué valor es

más frecuente y rellenar con ese valor, pero en este caso prácticamente los dos valores tienen una frecuencia similar. Otra opción sería **predecir ese valor faltante**. Eso es lo que vamos a hacer.

Para ello vamos a crear un clasificador para obtener este valor. En primer lugar, vamos a eliminar (momentáneamente) los tres valores blancos. Para ello, ejecutamos el último “tabla” y seleccionamos uno de los campos que estén en blanco. Pulsamos en el menú “Generar”:

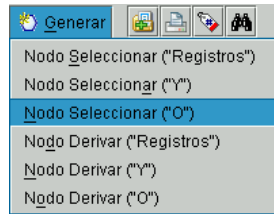


Figura 24. Generando un nodo para eliminar registros en blanco.

Nos ha generado un nodo “(generado)” que enganchamos al nodo “rellenar”. Editamos el nodo “(generado)” para que excluya (“descartar”) estos registros nulos:

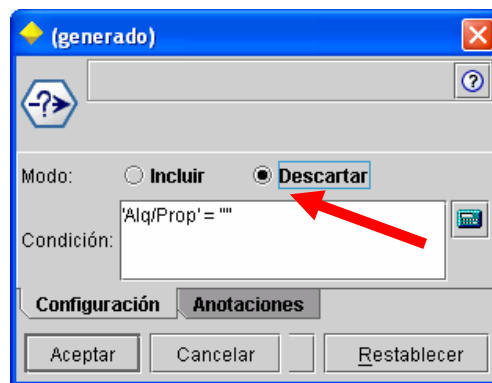


Figura 25. Configurando el nodo para eliminar registros en blanco.

Ahora añadimos un nodo “Tipo” a este nodo “(generado)” y vamos a señalar el campo “Alq/Prop” como campo de “SALIDA” (además pondremos el campo “#Ej” a “ninguna”)

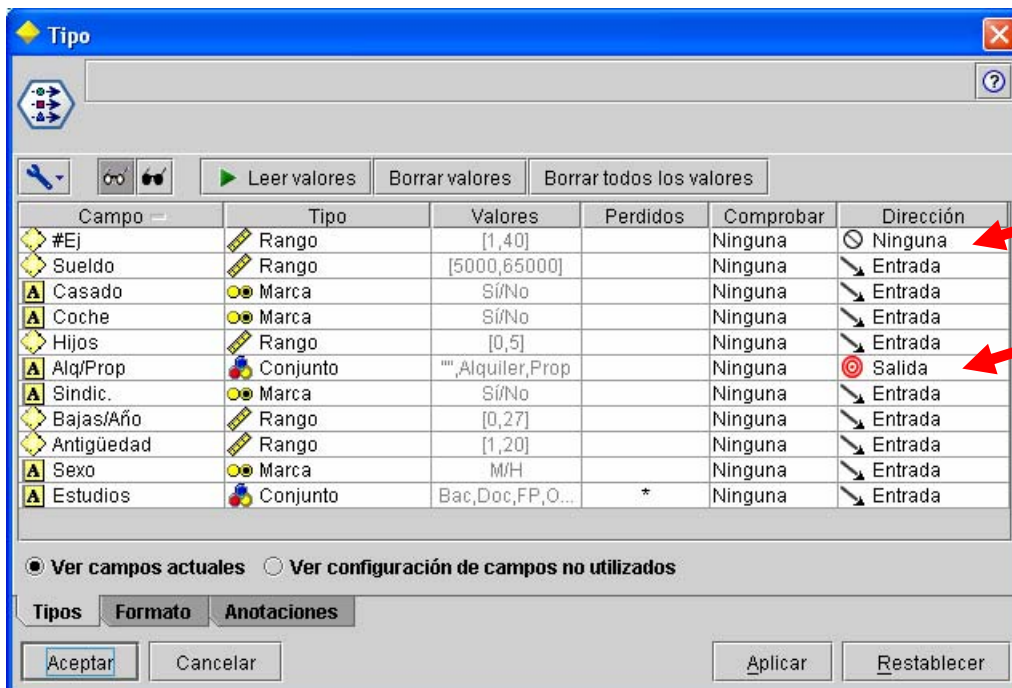


Figura 26. Tipando los datos para aprender un clasificador para el campo “Alq/Prop”.

Ahora añadimos un nodo de clasificación, por ejemplo, un “C&RT” y lo enganchamos al nodo Tipo como se muestra en la siguiente figura:

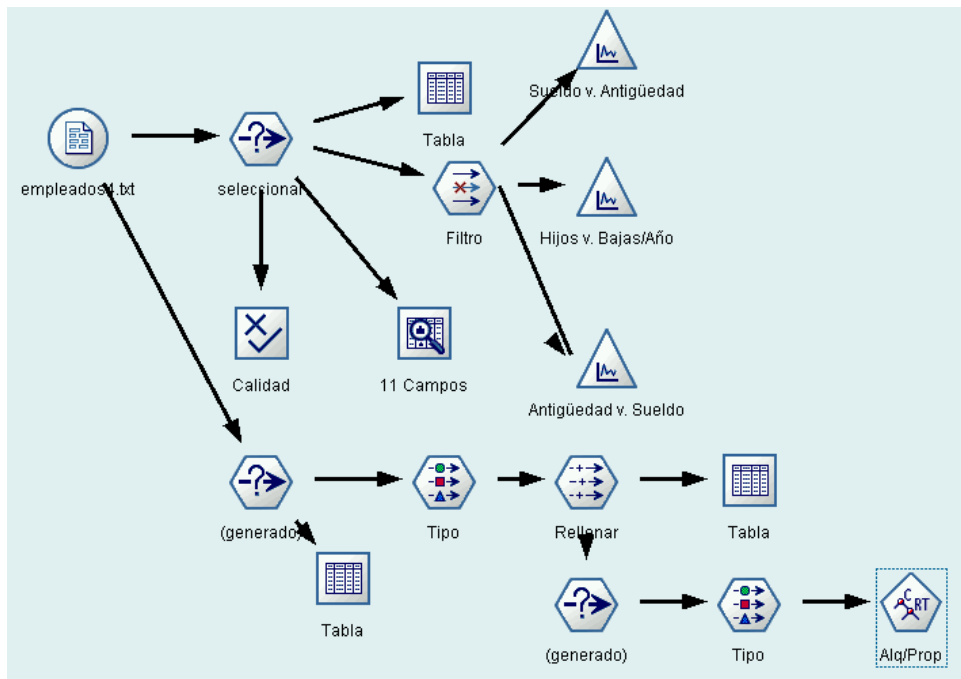


Figura 27. Estado de la ruta.

Ahora ejecutamos la ruta y obtenemos un modelo en el área de trabajo derecha del Clementine. Si editamos el nodo diamante “Alq/Prop” podemos ver cuáles son las reglas:

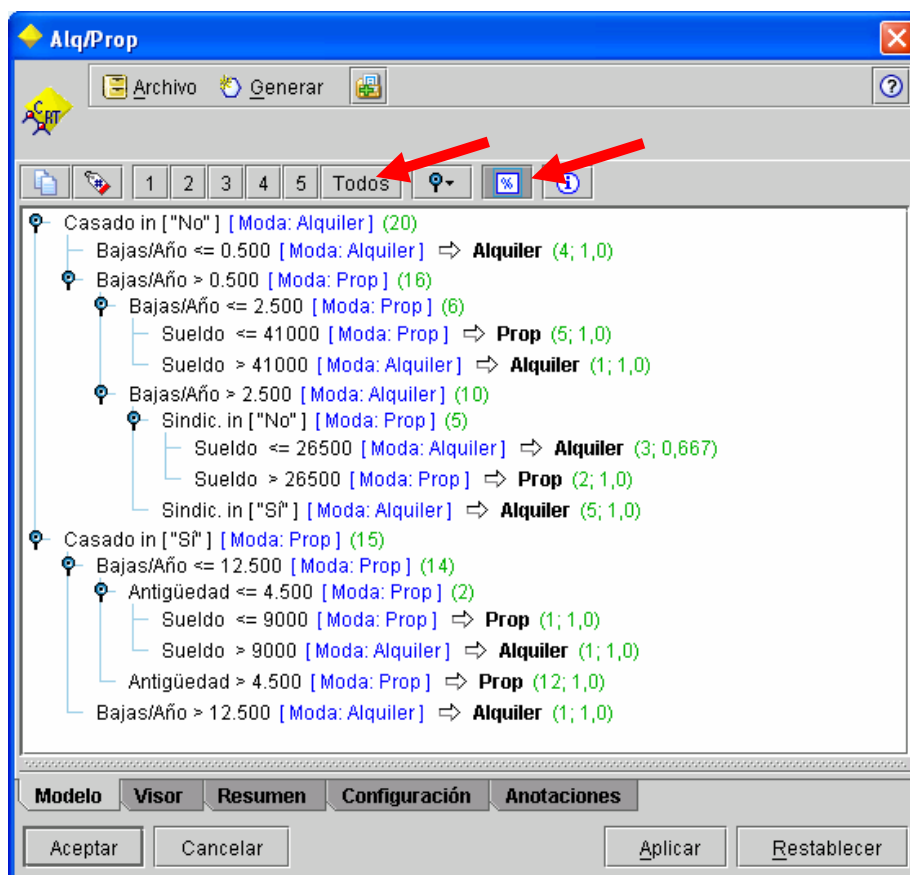


Figura 28. Modelo generado.

Añadimos el modelo a la zona de trabajo de la izquierda y lo enganchamos con el nodo Tipo de abajo, donde añadimos un nodo “Análisis”, como se muestra en la siguiente figura:

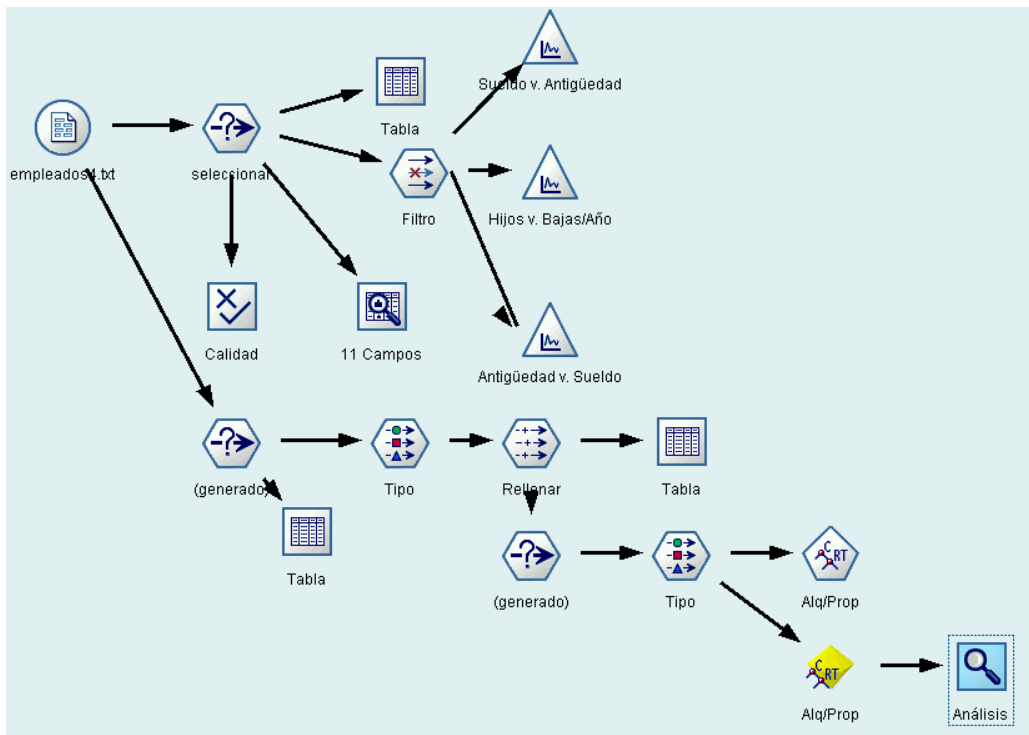


Figura 29. Estado de la ruta.

Podemos evaluar la calidad del modelo ejecutando el nodo “análisis”:

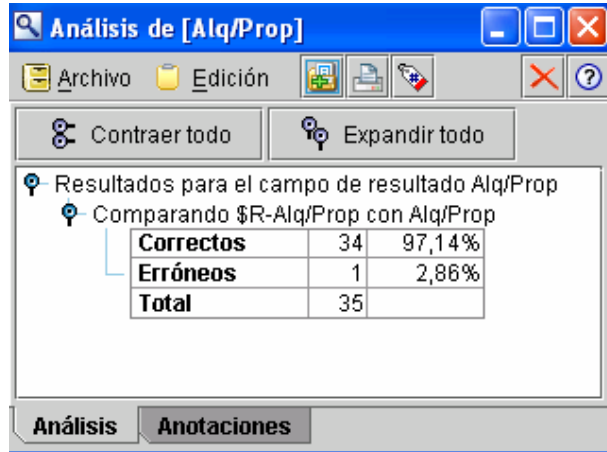


Figura 30. Análisis del modelo.

Este modelo es lo suficientemente aceptable (aunque esté validado sólo con los datos de entrenamiento) para sustituirnos los valores blancos que teníamos en ese campo. Para ello, volvemos a copiar el modelo en la zona de trabajo de la izquierda y lo enganchamos a través de un nuevo nodo Tipo (en el que pondremos el valor de “Salida” el Alq/Prop) con el Rellenar y añadimos un nodo “Tabla” detrás del diamante, como muestra la siguiente figura:

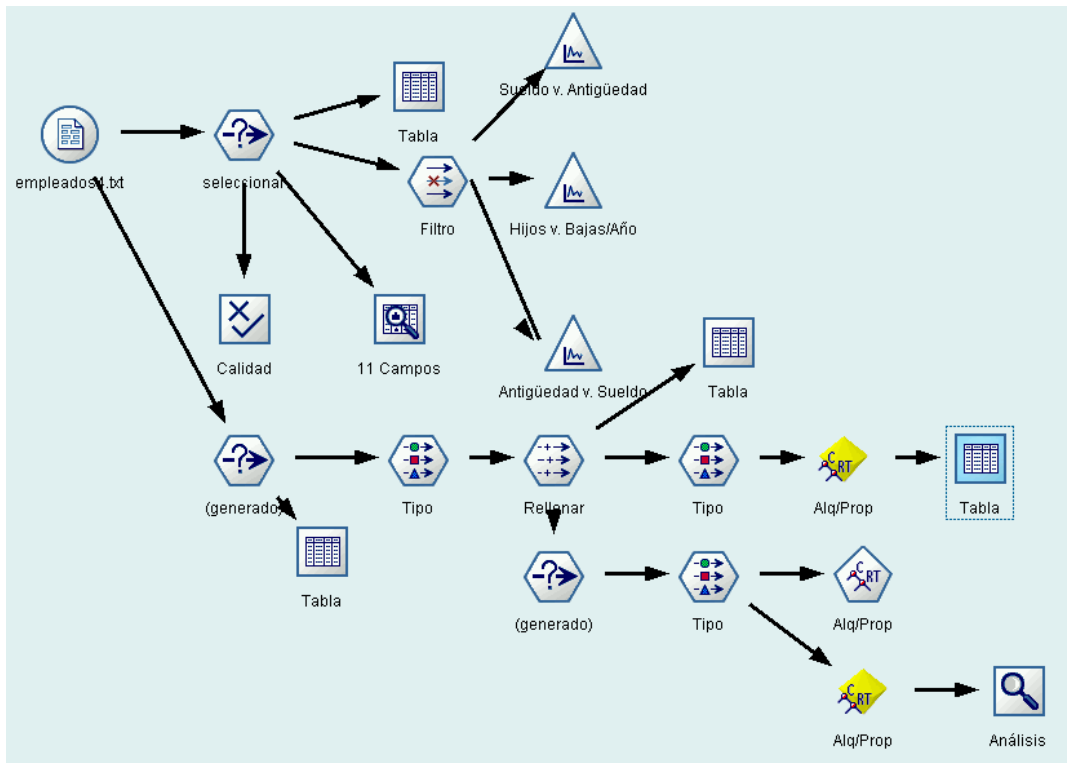


Figura 31. Estado de la ruta.

Si ejecutamos la tabla, vemos que el modelo genera valores para Alq/Prop en un nuevo campo "\$R-Alq/Prop":

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Estudios	\$R-Alq/Prop	\$RC-Alq/Prop	
1	1	13000	No	Sí	0	Prop	No	2	3	M	Obl	Prop	0.750
2	2	32000	Sí	No	2	Prop	Sí	1	15	M	Bac	Prop	0.867
3	3	12000	No	No	0	Alquiler	No	0	6	H	Obl	Alquiler	0.714
4	4	41000	Sí	Sí	3	Prop	No	3	13	H	Uni	Prop	0.867
5	5	5000	No	No	0		Sí	0	1	H	Obl	Alquiler	0.714
6	6	65000	No	Sí	0	Prop	No	3	8	M	Doc	Prop	0.600
7	7	53000	Sí	Sí	5	Prop	No	4	18	M	Bac	Prop	0.867
8	8	23000	No	Sí	0	Alquiler	Sí	7	2	H	Obl	Alquiler	0.750
9	9	31000	Sí	No	0	Prop	Sí	0	5	H	Bac	Prop	0.867
10	10	30000	Sí	Sí	2	Prop	No	1	20	H	Bac	Prop	0.867
11	11	20000	No	Sí	1	Alquiler	Sí	3	3	M	Uni	Alquiler	0.750
12	12	13000	No	No	0	Prop	No	12	2	H	Bac	Alquiler	0.500
13	13	11000	No	Sí	0	Alquiler	No	0	7	H	FP	Alquiler	0.714
14	14	9000	No	Sí	1	Prop	Sí	2	3	H	FP	Prop	0.750
15	15	60000	Sí	Sí	4	Prop	No	0	10	M	Uni	Prop	0.867
16	17	6000	No	Sí	0		No	0	1	H	Obl	Alquiler	0.714
17	19	23000	No	Sí	0	Prop	No	2	4	M	Bac	Prop	0.750
18	20	43000	No	Sí	3	Alquiler	Sí	20	7	H	Uni	Alquiler	0.750
19	21	13000	No	Sí	0	Alquiler	Sí	3	3	M	FP	Alquiler	0.750
20	22	21000	Sí	Sí	1	Prop	No	1	7	M	Bac	Prop	0.867
21	23	15000	Sí	Sí	2	Prop	Sí	5	10	H	Obl	Prop	0.867
22	24	30000	Sí	Sí	1	Alquiler	No	15	7	M	Uni	Alquiler	0.500
23	25	10000	Sí	Sí	0	Prop	Sí	1	6	H	Obl	Prop	0.867
24	26	40000	No	Sí	0	Alquiler	Sí	3	16	M	Bac	Alquiler	0.750
25	27	25000	No	No	0	Alquiler	Sí	0	8	H	Bac	Alquiler	0.714
26	28	20000	No	Sí	0		Sí	2	6	M	Bac	Prop	0.750
27	29	20000	Sí	Sí	3	Prop	No	7	5	H	Obl	Prop	0.867
28	30	10000	Sí	No	0	Alquiler	No	7	4	H	Obl	Alquiler	0.500
29	31	50000	No	No	0	Alquiler	No	2	12	M	Doc	Alquiler	0.500
30	32	8000	Sí	Sí	2	Prop	No	3	1	H	Obl	Prop	0.500
31	33	20000	No	No	0	Alquiler	No	27	5	M	Bac	Alquiler	0.500
32	34	10000	No	Sí	0	Alquiler	Sí	0	7	H	Obl	Alquiler	0.714
33	35	8000	No	Sí	0	Alquiler	No	3	2	H	FP	Alquiler	0.500
34	36	50000	Sí	Sí	1	Prop	No	1	12	H	Doc	Prop	0.867
35	37	7000	No	Sí	1	Prop	Sí	1	2	M	Obl	Prop	0.750
36	38	30000	Sí	Sí	2	Prop	Sí	10	8	H	FP	Prop	0.867
37	39	32000	No	No	0	Prop	No	2	3	M	Uni	Prop	0.750
38	40	33000	No	Sí	3	Prop	No	5	7	H	Uni	Prop	0.600

Figura 32. Campos Generados.

También hay un campo "\$RC-Alq/Prop" que nos da la fiabilidad de la predicción para los campos rellenos. No es excesivamente alta pero es aceptable. Ahora, vamos a copiar los tres campos predichos en el campo original "Alq/Prop". Para ello, en primer lugar vamos a modificar el nodo "Tipo" que tenemos en el camino al nodo tabla que acabamos de ver, y lo modificamos para que nos detecte los campos en blanco del campo "Alq/Prop", como sigue:

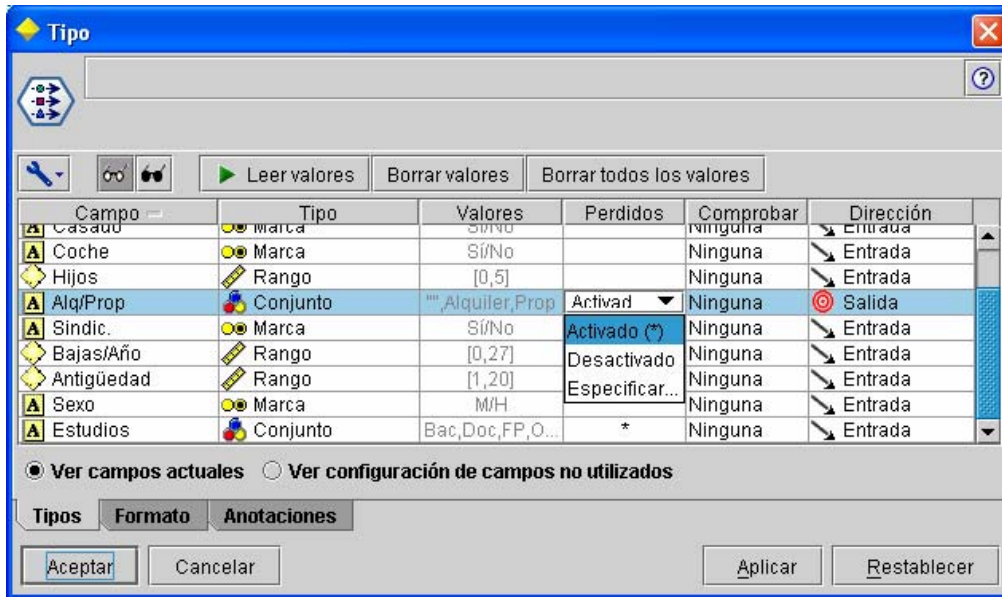


Figura 33. Marcamos "Activado (*)" para que nos reconozca los valores faltantes de Alq/Prop.

Ahora añadimos un nodo "rellenar", donde seleccionamos el campo "Alq/prop" y ante la condición "@(BLANK(@FIELD))" reemplazaremos con el campo predicho. En la ventanita de "Reemplazar con" podemos pinchar en el icono de la calculadora y en el "Generador de expresiones", a la derecha, seleccionar el campo '\$R-Alq/Prop', como se ve en la siguiente figura.

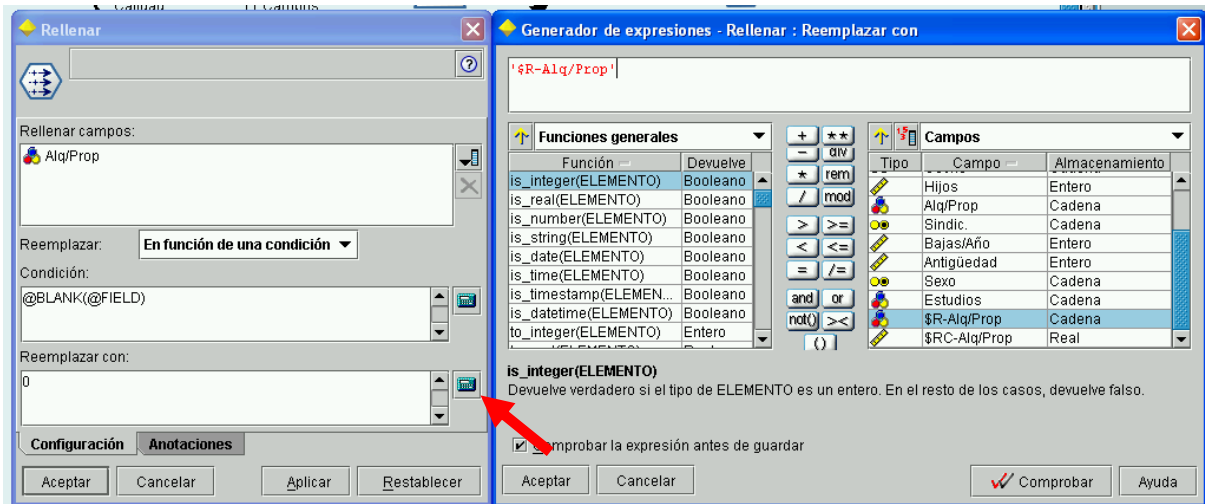


Figura 34. Copiar del campo predicho al campo original mediante un nodo Rellenar.

Le damos a aceptar y añadimos un nodo Tabla que enganchamos con el Rellenar y ejecutamos para ver el resultado.

#E	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Estudios	\$R-Alq/Prop	\$RC-Alq/Prop	
1	13000	No	Sí	0	Prop	No	2	3	M	Obl	Prop	0.750	
2	32000	Sí	No	2	Prop	Sí	1	15	M	Bac	Prop	0.867	
3	12000	No	No	0	Alquiler	No	0	6	H	Obl	Alquiler	0.714	
4	41000	Sí	Sí	3	Prop	No	3	13	H	Uni	Prop	0.867	
5	5000	No	No	0	Alquiler	Sí	0	1	H	Obl	Alquiler	0.714	
6	65000	No	Sí	0	Prop	No	3	8	M	Doc	Prop	0.600	
7	53000	Sí	Sí	5	Prop	No	4	18	M	Bac	Prop	0.867	
8	23000	No	Sí	0	Alquiler	Sí	7	2	H	Obl	Alquiler	0.750	
9	31000	Sí	No	0	Prop	Sí	0	5	H	Bac	Prop	0.867	
10	30000	Sí	Sí	2	Prop	No	1	20	H	Bac	Prop	0.867	
11	20000	No	Sí	1	Alquiler	Sí	3	3	M	Uni	Alquiler	0.750	
12	13000	No	No	0	Prop	No	12	2	H	Bac	Alquiler	0.500	
13	11000	No	Sí	0	Alquiler	No	0	7	H	FP	Alquiler	0.714	
14	9000	No	Sí	1	Prop	Sí	2	3	H	FP	Prop	0.750	
15	60000	Sí	Sí	4	Prop	No	0	10	M	Uni	Prop	0.867	
16	6000	No	Sí	0	Alquiler	No	0	1	H	Obl	Alquiler	0.714	
17	23000	No	Sí	0	Prop	No	2	4	M	Bac	Prop	0.750	
18	43000	No	Sí	3	Alquiler	Sí	20	7	H	Uni	Alquiler	0.750	
19	21	13000	No	Sí	0	Alquiler	Sí	3	3	M	FP	Alquiler	0.750
20	22	21000	Sí	Sí	1	Prop	No	1	7	M	Bac	Prop	0.867
21	23	15000	Sí	Sí	2	Prop	Sí	5	10	H	Obl	Prop	0.867
22	24	30000	Sí	Sí	1	Alquiler	No	15	7	M	Uni	Alquiler	0.500
23	25	10000	Sí	Sí	0	Prop	Sí	1	6	H	Obl	Prop	0.867
24	26	40000	No	Sí	0	Alquiler	Sí	3	16	M	Bac	Alquiler	0.750
25	27	25000	No	No	0	Alquiler	Sí	0	8	H	Bac	Alquiler	0.714
26	28	20000	No	Sí	0	Prop	Sí	2	6	M	Bac	Prop	0.750
27	29	20000	Sí	Sí	3	Prop	No	7	5	H	Obl	Prop	0.867
28	30	10000	Sí	No	0	Alquiler	No	7	4	H	Obl	Alquiler	0.500
29	31	50000	No	No	0	Alquiler	No	2	12	M	Doc	Alquiler	0.500
30	32	8000	Sí	Sí	2	Prop	No	3	1	H	Obl	Prop	0.500
31	33	20000	No	No	0	Alquiler	No	27	5	M	Bac	Alquiler	0.500
32	34	10000	No	Sí	0	Alquiler	Sí	0	7	H	Obl	Alquiler	0.714
33	35	8000	No	Sí	0	Alquiler	No	3	2	H	FP	Alquiler	0.500
34	36	50000	Sí	Sí	1	Prop	No	1	12	H	Doc	Prop	0.867
35	37	7000	No	Sí	1	Prop	Sí	1	2	M	Obl	Prop	0.750
36	38	30000	Sí	Sí	2	Prop	Sí	10	8	H	FP	Prop	0.867
37	39	32000	No	No	0	Prop	No	2	3	M	Uni	Prop	0.750
38	40	33000	No	Sí	3	Prop	No	5	7	H	Uni	Prop	0.600

Figura 35. Datos copiados.

Como se observa, hemos copiado de \$R-Alq/Prop a Alq/Prop los tres datos que nos faltaban. La ruta en este momento la tenemos como sigue:

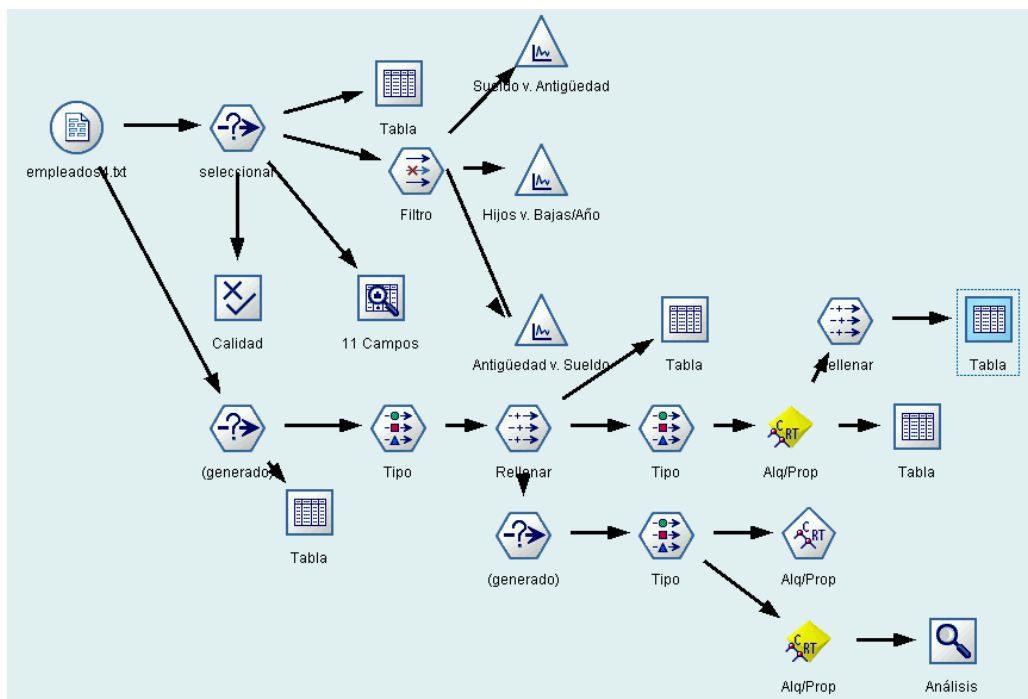


Figura 36. Estado de la ruta.

Por último, el campo “Estudios” sería interesante que se considerara como una escala en vez de como un valor nominal sin orden. El Clementine, si vemos el nodo Tipo, lo considera como un “conjunto”. Eso quiere decir que no le da mayor o menor valor a un doctorado que a los estudios obligatorios. Para solventarlo, una primera idea sería definirlo como un “conjunto ordenado”, pero el Clementine no lo pone fácil para darle el orden que nosotros queramos.

Para evitar problemas, vamos a convertirlo a numérico en vez de simbólico. Para ello añadimos un nodo “Derivar”. Lo conectamos al nodo diamante “Alq/Prop” de arriba y lo editamos, poniendo como nombre “Vestudios” y en la parte de “Derivar como”, pondremos “Conjunto”, y editaremos cinco filas como se muestra en la figura:

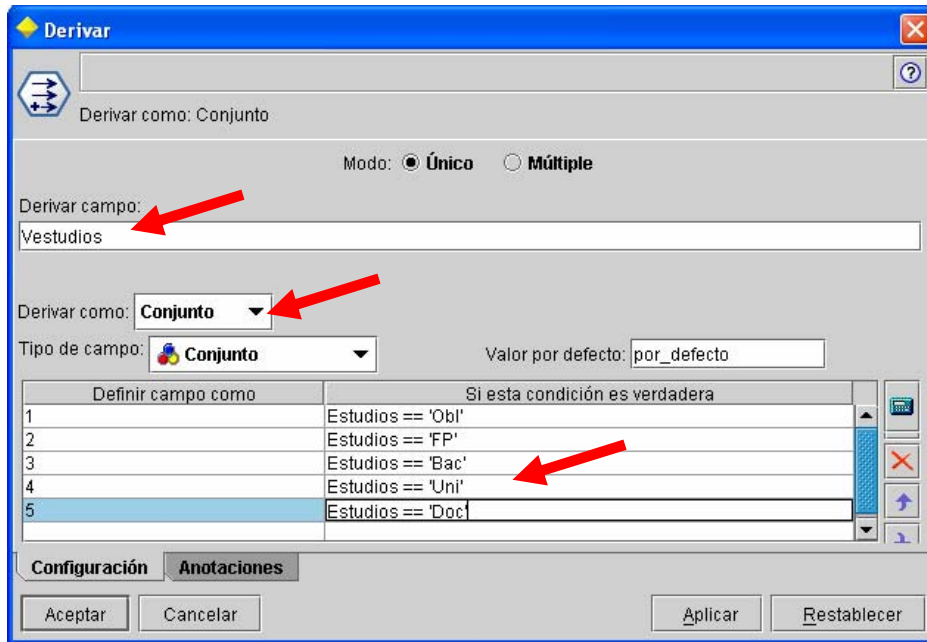


Figura 37. Configurando un nodo “Derivar”.

Podemos observar el resultado añadiendo un nodo “Tabla”:

#Ej	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic.	Bajas/Año	Antigüedad	Sexo	Estudios	\$R-Alq/Prop	\$RC-Alq/P.	Vestudios	
1	13000	No	Sí	0	Prop	No	2	3	M	Obl	Prop	0.500	1	
2	32000	Sí	No	2	Prop	Sí	1	15	M	Bac	Prop	0.867	3	
3	12000	No	No	0	Alquiler	No	0	6	H	Obl	Alquiler	0.625	1	
4	41000	Sí	Sí	3	Prop	No	3	13	H	Uni	Prop	0.867	4	
5	5000	No	No	0	Alquiler	Sí	0	1	H	Obl	Alquiler	0.625	1	
6	65000	No	Sí	0	Prop	No	3	8	M	Doc	Prop	No	0.667	5
7	53000	Sí	Sí	5	Prop	No	4	18	M	Bac	Prop	0.867	3	
8	23000	No	Sí	0	Alquiler	Sí	7	2	H	Obl	Alquiler	0.625	1	
9	31000	Sí	No	0	Prop	Sí	0	5	H	Bac	Prop	0.867	3	
10	30000	Sí	Sí	2	Prop	No	1	20	H	Bac	Prop	0.867	3	
11	20000	No	Sí	1	Alquiler	Sí	3	3	M	Uni	Alquiler	0.625	4	
12	13000	No	No	0	Prop	No	12	2	H	Bac	Prop	0.667	3	
13	11000	No	Sí	0	Alquiler	No	0	7	H	FP	Alquiler	No	0.625	2
14	9000	No	Sí	1	Prop	Sí	2	3	H	FP	Alquiler	0.625	2	
15	60000	Sí	Sí	4	Prop	No	0	10	M	Uni	Prop	0.867	4	
16	6000	No	Sí	0	Alquiler	No	0	1	H	Obl	Alquiler	0.625	1	
17	23000	No	Sí	0	Prop	No	2	4	M	Bac	Prop	0.667	3	
18	43000	No	Sí	3	Alquiler	Sí	20	7	H	Uni	Alquiler	0.818	4	
19	13000	No	Sí	0	Alquiler	Sí	3	3	M	FP	Alquiler	0.818	2	
20	21000	Sí	Sí	1	Prop	No	1	7	M	Bac	Prop	No	0.867	3
21	15000	Sí	Sí	2	Prop	Sí	5	10	H	Obl	Prop	0.867	1	
22	30000	Sí	Sí	1	Alquiler	No	15	7	M	Uni	Alquiler	0.500	4	
23	10000	Sí	Sí	0	Prop	Sí	1	6	H	Obl	Prop	No	0.867	1
24	40000	No	Sí	0	Alquiler	Sí	3	16	M	Bac	Alquiler	0.818	3	
25	25000	No	No	0	Alquiler	Sí	0	8	H	Bac	Alquiler	No	0.818	3

Figura 38. Mostrando el campo derivado.

El problema de este campo es que todavía es un conjunto y no un rango (un numérico). Con el nodo "Tipo" no se puede cambiar el tipo de una cadena a un número, con lo que hemos de hacerlo con otro nodo "Derivar", que conectaremos al nodo "Derivar" anterior y lo editaremos de la siguiente forma.

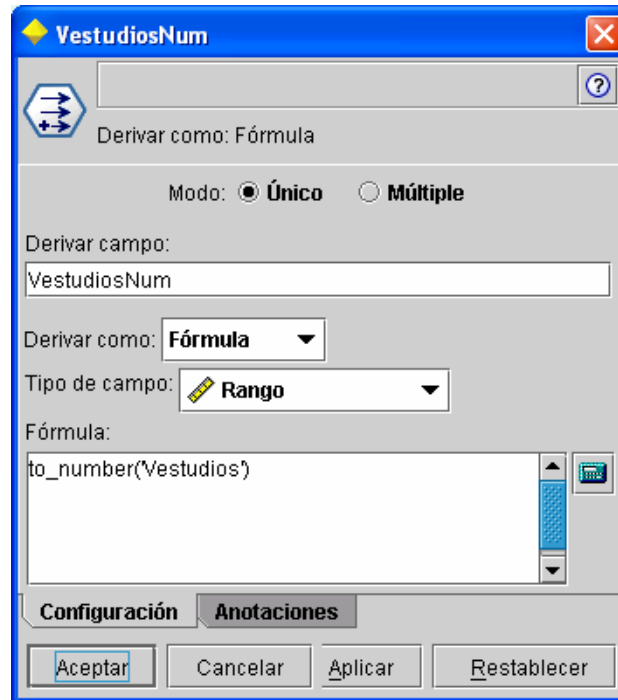


Figura 39. Cambiando el tipo de cadena a rango usando un nodo Derivar.

Ahora vamos a ver si tenemos todos los tipos correctos. Añadimos un nodo Tipo y lo conectamos al último nodo Derivar. Editamos de la siguiente manera: cambiamos el tipo de Alq/Prop de Conjunto a "Por Defecto" (dejará "Discreto" y los valores pondrá "<Leer>", cambiamos la dirección de este campo a Entrada, como se muestra a continuación. También marca el campo "#Ej" como que no es ni de Entrada ni de Salida (Ninguna):

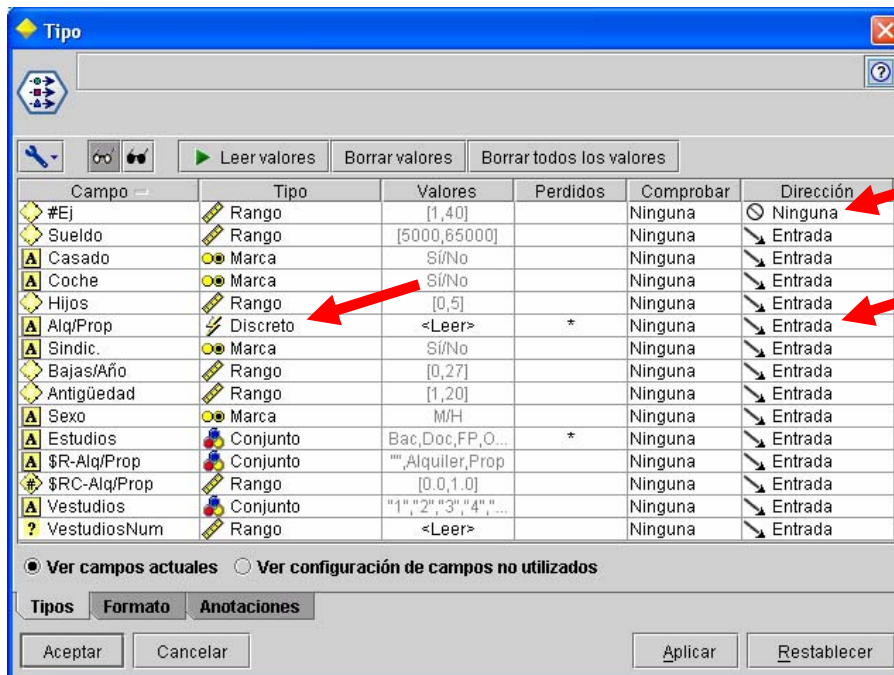


Figura 40. Comprobando los tipos.

Ahora pincharemos en la parte de arriba donde pone "Leer valores" y veremos que Alq/Prop ya es una "Marca" con sólo dos valores Prop y Alquiler.:

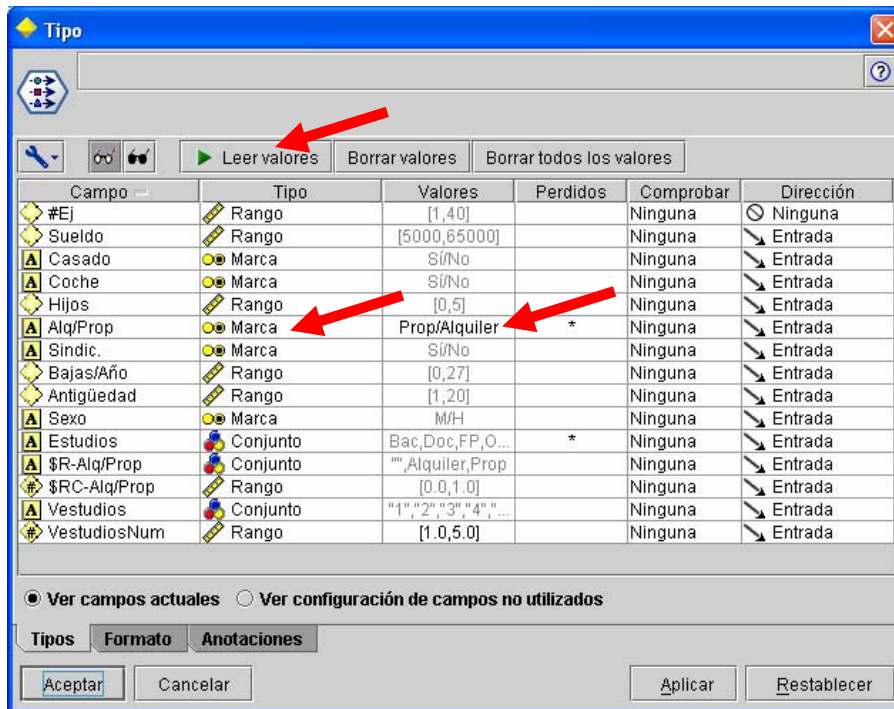


Figura 41. Comprobando los tipos.

Por último, ya sólo nos falta quedarnos con las columnas válidas. En este caso, se trata de quedarnos con la columna VestudiosNum y no la vieja Estudios ni VEstudios y en el caso de Alq/Prop quitar todas las otras que generamos. Para eso simplemente añadimos un nodo Filtro que enganchamos con el nodo "VEstudios" y lo editamos de la siguiente manera:

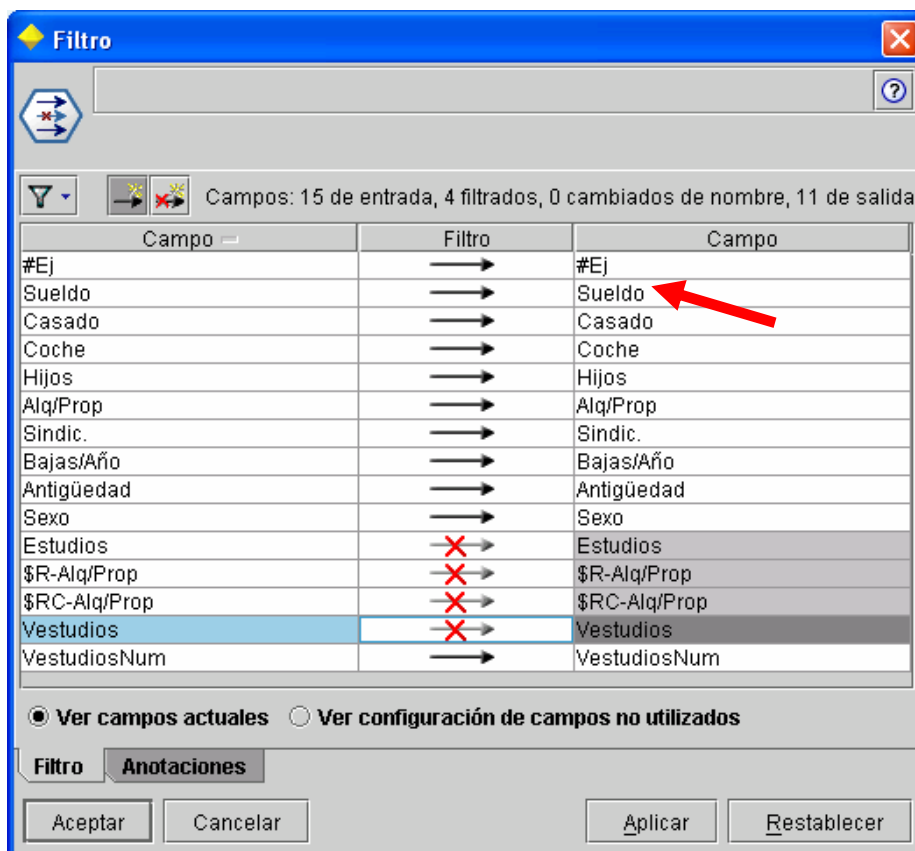


Figura 42. Filtrando los campos que ya no vamos a usar.

También acuérdate de modificar el campo “Sueldo “ para quitarle el espacio del final. Por fin tenemos los datos preparados para trabajar con la ruta que se muestra en la siguiente figura:

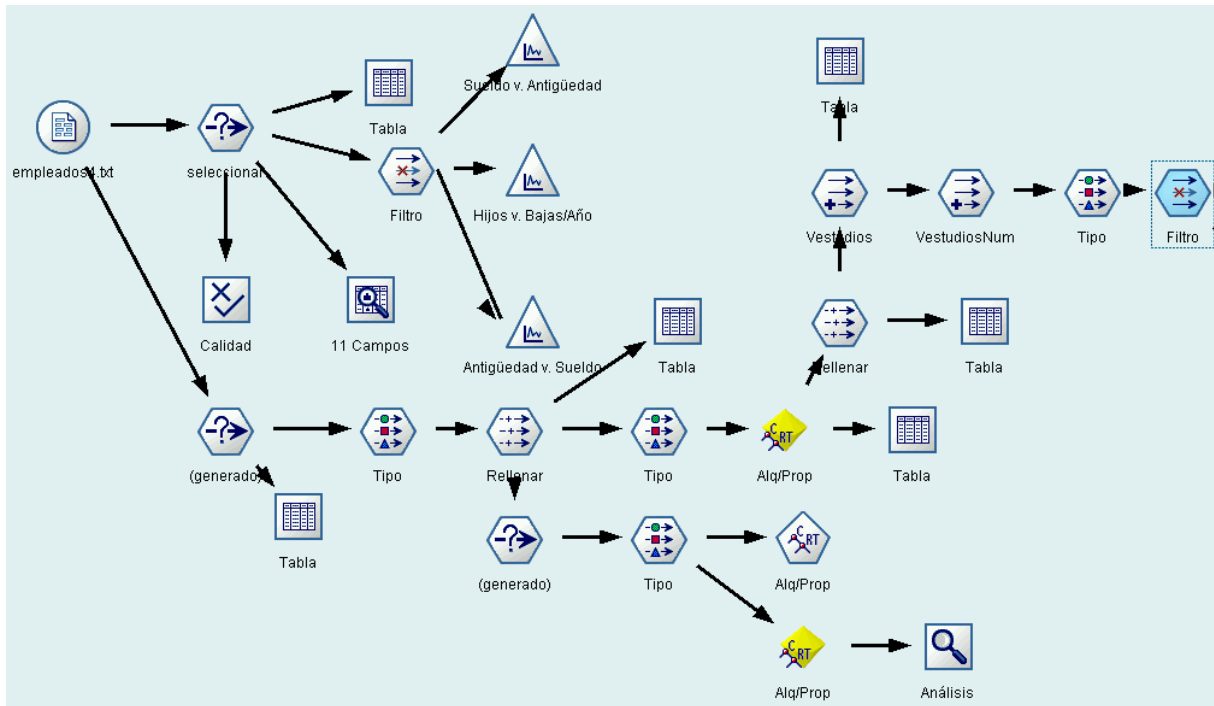


Figura 43. Ruta tras el proceso de limpieza de datos.

Ahora ya podemos examinar mejor los datos y realizar múltiples gráficas y tablas para analizar los datos, p.ej. los Gráficos del estilo de los que ya vimos, histogramas, nodos “Matriz” (están en el apartado de “Resultado”, no en “gráficas”, y vienen bien para comparar campos nominales), etc., como se muestra en la siguiente figura:

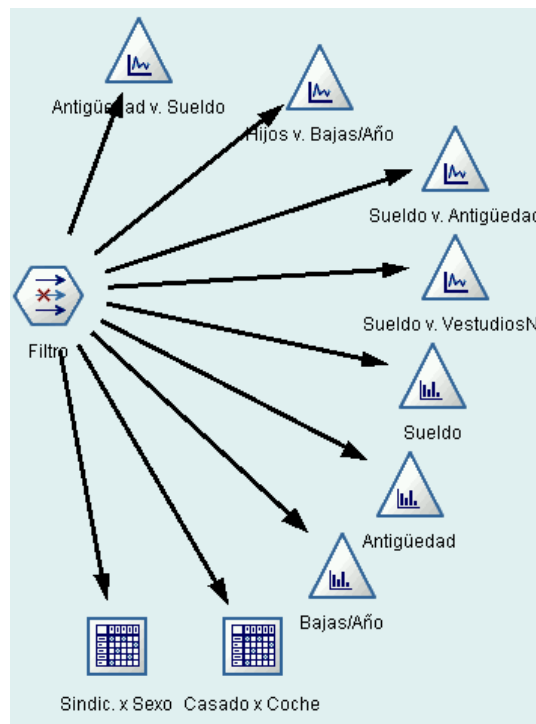


Figura 44. Analizando los datos mediante gráficas (sólo se muestra a partir del nodo “Filtro”).

Por ejemplo, ahora la gráfica entre antigüedad y sueldo se ve mejor sin el dato anómalo de “Sueldo”. También se ven mucho mejor todas las gráficas que tienen en cuenta la edad. También vemos en el

gráfico entre el “Sueldo” y los “Estudios”, que la correlación es alta entre ambos valores (casi lineal), como vemos en la siguiente figura:

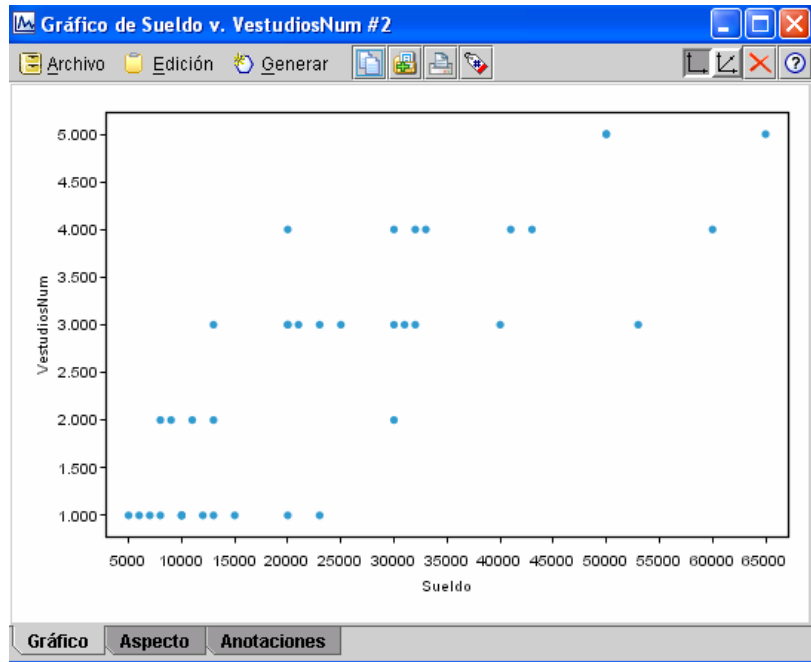


Figura 45. Relación entre Sueldo y Estudios.

Ahora retomemos la gráfica entre antigüedad y sueldo y le añadimos en el campo superponer si están casados o no, que ahora se ve mejor sin el dato anómalo de “Sueldo”. Si la ejecutamos y observamos, con un poco de atención se ven dos zonas diferenciadas, una de bajo sueldo y antigüedad, donde resulta que la mayoría no están casados, y una zona de alto sueldo y antigüedad, donde resulta que la mayoría están casados. Con el ratón seleccionamos ambas regiones, como vemos en la siguiente figura:

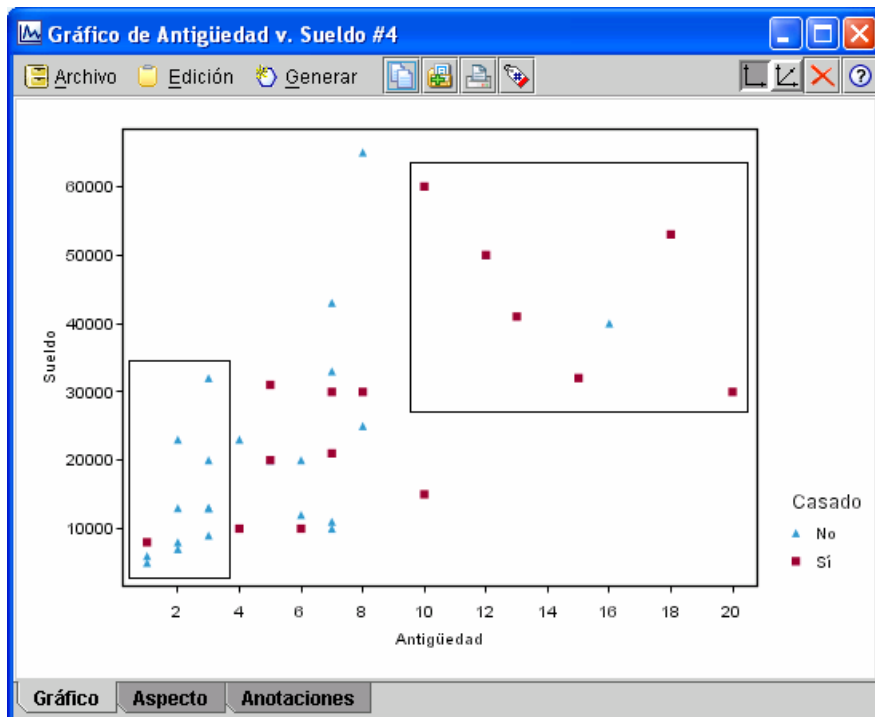


Figura 46. Marcando regiones con el ratón.

Ahora podemos generar campos derivados yendo al menú "Generar" y seleccionando "Nodo Derivar (Conjunto)".

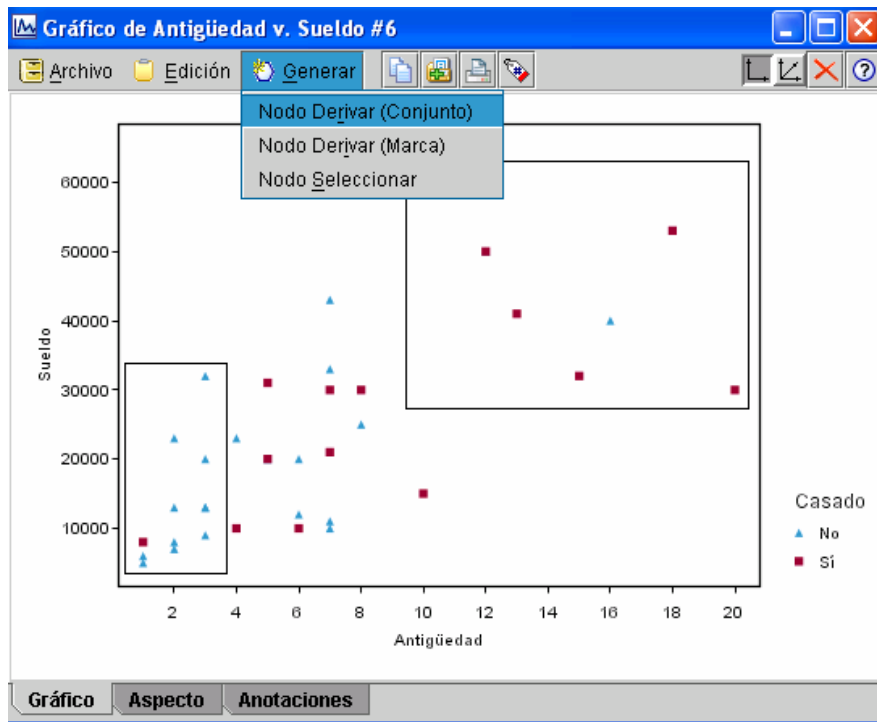


Figura 47. Generando atributos derivados utilizando regiones espaciales marcadas con el ratón.

Podemos ver que nos aparece un nuevo nodo "región" en la zona de trabajo. Si lo conectamos con el último nodo Filtro y le añadimos un nodo "Tabla" podemos ver el siguiente resultado:

#E	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindic	Bajas/Año	Antigüedad	Sexo	Vestudios	Num	región
1	13000	No	Sí	0	Prop	No	2	3	M	1.000	región2	
2	32000	Sí	No	2	Prop	Sí	1	15	M	3.000	región1	
3	12000	No	No	0	Alquiler	No	0	6	H	1.000	por defecto	
4	41000	Sí	Sí	3	Prop	No	3	13	H	4.000	región1	
5	5000	No	No	0	Alquiler	Sí	0	1	H	1.000	región2	
6	65000	No	Sí	0	Prop	No	3	8	M	5.000	por defecto	
7	53000	Sí	Sí	5	Prop	No	4	18	M	3.000	región1	
8	23000	No	Sí	0	Alquiler	Sí	7	2	H	1.000	región2	
9	31000	Sí	No	0	Prop	Sí	0	5	H	3.000	por defecto	
10	30000	Sí	Sí	2	Prop	No	1	20	H	3.000	región1	
11	20000	No	Sí	1	Alquiler	Sí	3	3	M	4.000	región2	
12	13000	No	No	0	Prop	No	12	2	H	3.000	región2	
13	11000	No	Sí	0	Alquiler	No	0	7	H	2.000	por defecto	
14	9000	No	Sí	1	Prop	Sí	2	3	H	2.000	región2	
15	60000	Sí	Sí	4	Prop	No	0	10	M	4.000	región1	
16	6000	No	Sí	0	Alquiler	No	0	1	H	1.000	región2	
17	23000	No	Sí	0	Prop	No	2	4	M	3.000	por defecto	
18	43000	No	Sí	3	Alquiler	Sí	20	7	H	4.000	por defecto	
19	13000	No	Sí	0	Alquiler	Sí	3	3	M	2.000	región2	
20	21000	Sí	Sí	1	Prop	No	1	7	M	3.000	por defecto	
21	15000	Sí	Sí	2	Prop	Sí	5	10	H	1.000	por defecto	
22	30000	Sí	Sí	1	Alquiler	No	15	7	M	4.000	por defecto	
23	10000	Sí	Sí	0	Prop	Sí	1	6	H	1.000	por defecto	
24	40000	No	Sí	0	Alquiler	Sí	3	16	M	3.000	región1	
25	25000	No	No	0	Alquiler	Sí	0	8	H	3.000	por defecto	
26	20000	No	Sí	0	Prop	Sí	2	6	M	3.000	por defecto	
27	20000	Sí	Sí	3	Prop	No	7	5	H	1.000	por defecto	
28	10000	Sí	No	0	Alquiler	No	7	4	H	1.000	por defecto	
29	50000	No	No	0	Alquiler	No	2	12	M	5.000	región1	
30	8000	Sí	Sí	2	Prop	No	3	1	H	1.000	región2	
31	20000	No	No	0	Alquiler	No	27	5	M	3.000	por defecto	
32	10000	No	Sí	0	Alquiler	Sí	0	7	H	1.000	por defecto	
33	8000	No	Sí	0	Alquiler	No	3	2	H	2.000	región2	
34	50000	Sí	Sí	1	Prop	No	1	12	H	5.000	región1	
35	7000	No	Sí	1	Prop	Sí	1	2	M	1.000	región2	
36	30000	Sí	Sí	2	Prop	Sí	10	8	H	2.000	por defecto	
37	32000	No	No	0	Prop	No	2	3	M	4.000	región2	
38	33000	No	Sí	3	Prop	No	5	7	H	4.000	por defecto	

Figura 48. Se ha generado un nuevo campo según la "región".

Según en el orden que hayas hecho las regiones se llamarán 1 o 2. En el caso que se muestra en la figura, la “región2” es la de bajo sueldo y antigüedad mientras que la “región1” es la de alto sueldo y antigüedad.

Ahora lo que vamos a hacer es realizar un agrupamiento en tres clusters (mediante Kmedias) utilizando este nuevo campo y sin utilizarlo, comparando los resultados. La ruta que hay que hacer es la que se muestra en la siguiente figura:

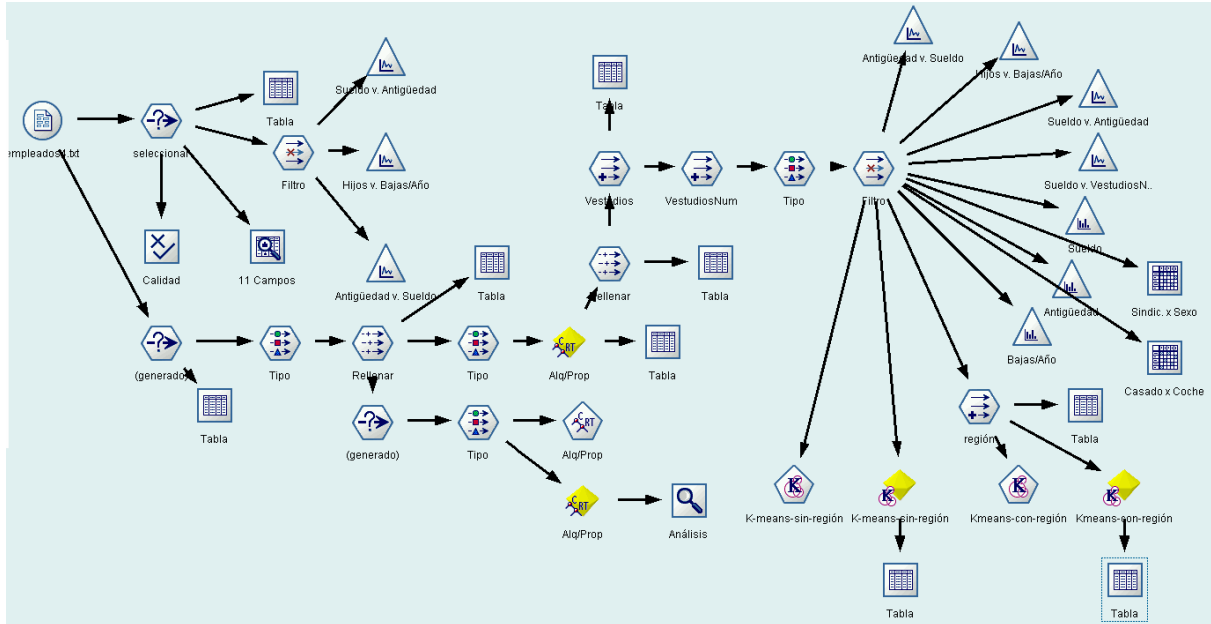


Figura 49. Ruta definitiva.

EJERCICIOS PROPUESTOS:

- Compara los clusters que salen por cada uno de las dos vías (sin las regiones o con ellas) y comenta si los resultados son razonables. Interpreta los grupos obtenidos en ambos casos.
- Varía el número de clusters (2, 4, 5, ...) y vuelve a ejecutar la ruta para cada uno de estos valores. ¿qué se puede observar?
- Analiza las distancias entre grupos y las distancias de los elementos al centro

Ahora graba la ruta en un fichero “.str”, p.ej. “empleados2.str”,

Por último, si te quedan ganas, intenta hacer un agrupamiento de tres clusters a partir de los datos iniciales, sin la limpieza realizada (con los datos originales) y compara los resultados.

2. Obtención y transformación de datos a partir de bases de datos y almacenes de datos vía ODBC

Aunque en la mayoría de ejemplos vamos a trabajar por comodidad a partir de datos que están en ficheros, en la mayoría de aplicaciones reales los datos originales se encuentran en una base de datos o un almacén de datos. Clementine permite enlazar a través de ODBC con cualquier motor de un Sistema de Gestión de Bases de Datos (SGBD).

Pasemos a trabajar con una base de datos clásica, el ejemplo “Neptuno” en MsAccess que versa sobre una compañía de pedidos. La base de datos se encuentra en el directorio “..\LabKDD\Neptuno”. Si abres la base de datos “Neptuno.mdb” (si tienes una versión más moderna de Access, le diremos que la abra, que no la convierta), la estructura de la base de datos es la siguiente:

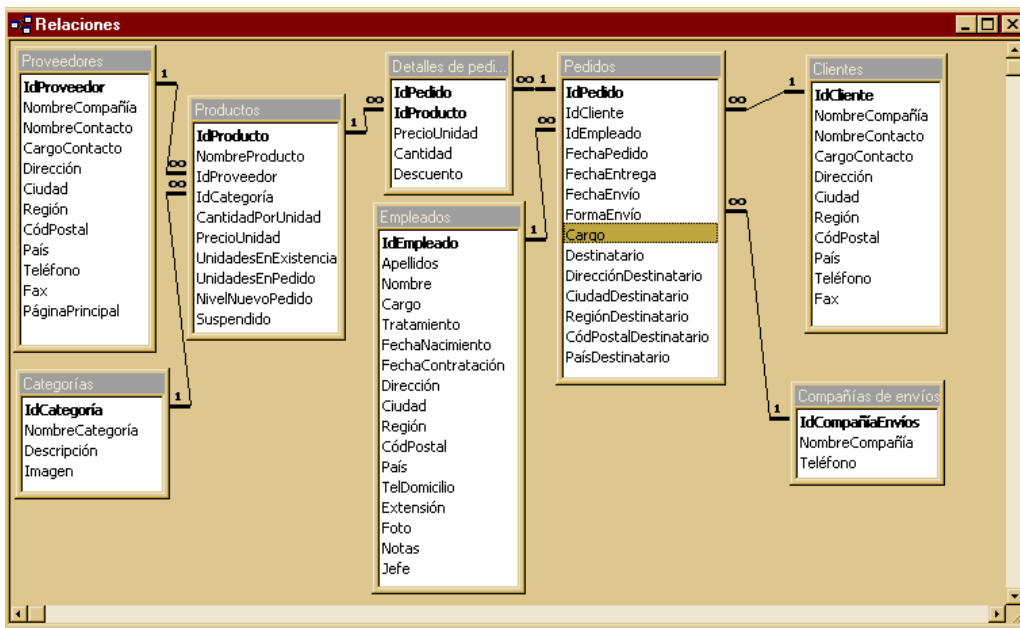


Figura 50. Base de Datos Neptuno

Sin entrar en más detalles en el esquema, el problema que vamos a tratar en este caso es el de predecir el volumen de pedidos para el próximo cuatrimestre (es un ejemplo muy simple, pero de nuevo nos servirá para ir conociendo mejor el sistema y tomar conciencia de que incluso en un problema muy simple, con muy pocos datos, hay que invertir tiempo procesándolos para conseguir una vista minable de cierta calidad).

Para ello, se ha realizado una consulta específica para esta cuestión que obtiene el total de los cargos de los pedidos por cuatrimestre. Dicha consulta (en realidad es una vista) se llama “_VentasPorCuatrimestre”:



Figura 51. Vistas de la Base de Datos

Si la ejecutamos, tenemos el siguiente resultado:

ORD	ANYO	TOTAL
1	1994A	
2	1994B	
3	1994C	2.413 pta
4	1994D	4.404 pta
5	1995A	6.356 pta
6	1995B	7.547 pta
7	1995C	8.458 pta
8	1995D	9.282 pta
9	1996A	12.999 pta
10	1996B	12.496 pta
11	1996C	
12	1996D	

Figura 52. Ventas por Cuatrimestre

Nos interesa realizar una estimación para los cuatrimestres 1996C y 1996D. Para ello, vamos a utilizar el Clementine.

Antes debemos crear una fuente ODBC en el sistema. Para ello (en Windows) vamos al Panel de Control y elegimos en “Herramientas Administrativas”, los “Orígenes de Datos ODBC”:

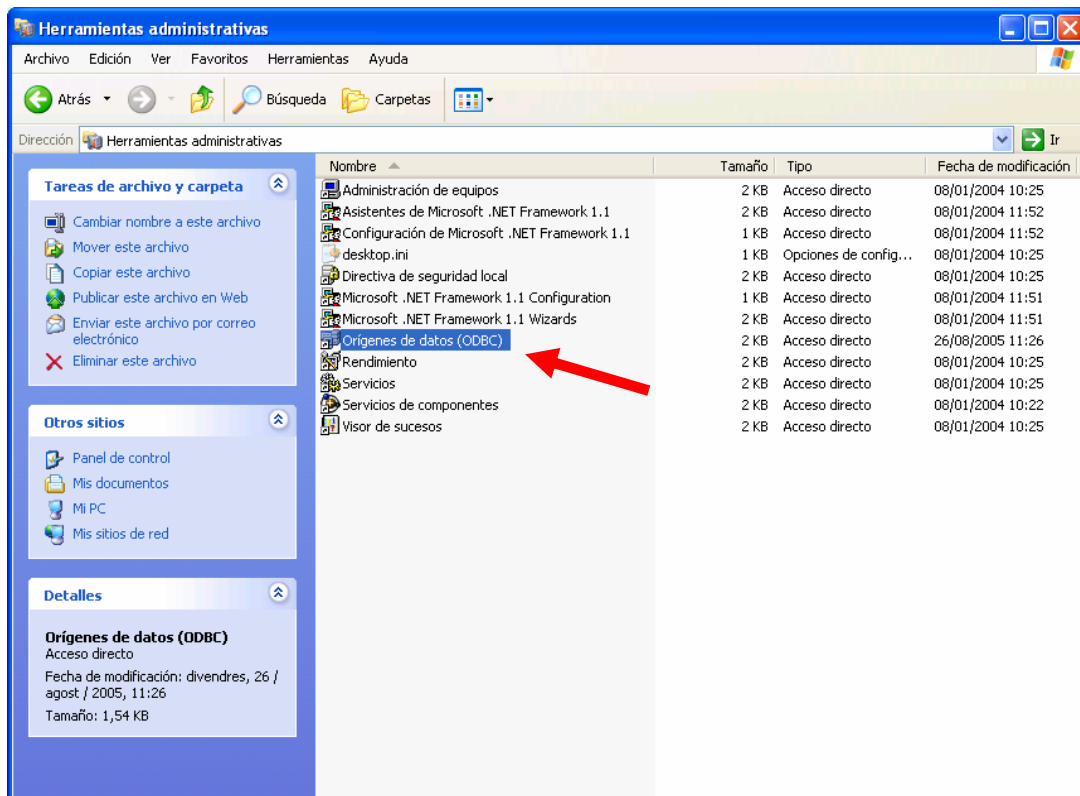


Figura 53. Panel de Control

Pinchamos y, en la pestaña “Orígenes de datos de Usuario” agregamos un origen “Ms Access Database” (hemos de seleccionar ese origen y pinchar en “Agregar”):

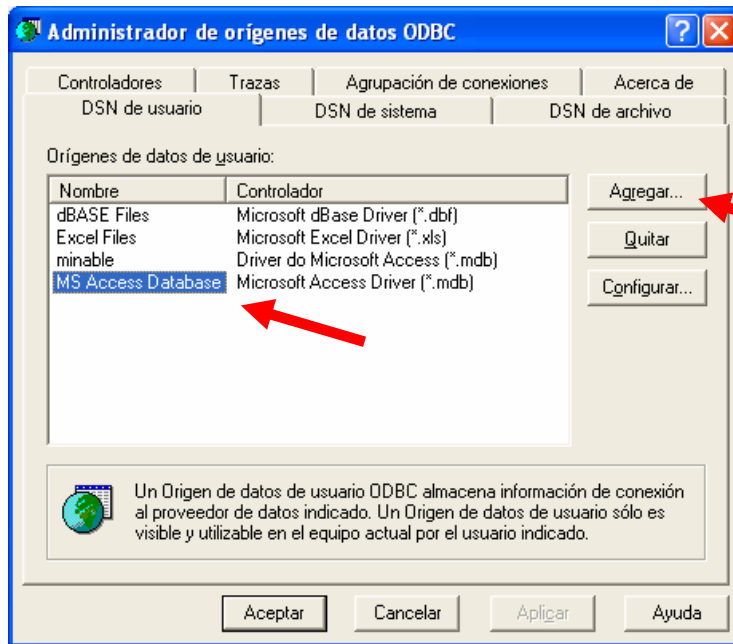


Figura 54. Administrador de orígenes de datos ODBC

En la siguiente pantalla “Crear nuevo origen de datos”, elige “Controlador para Microsoft Access (*.mdb)” y pincha “Finalizar”.

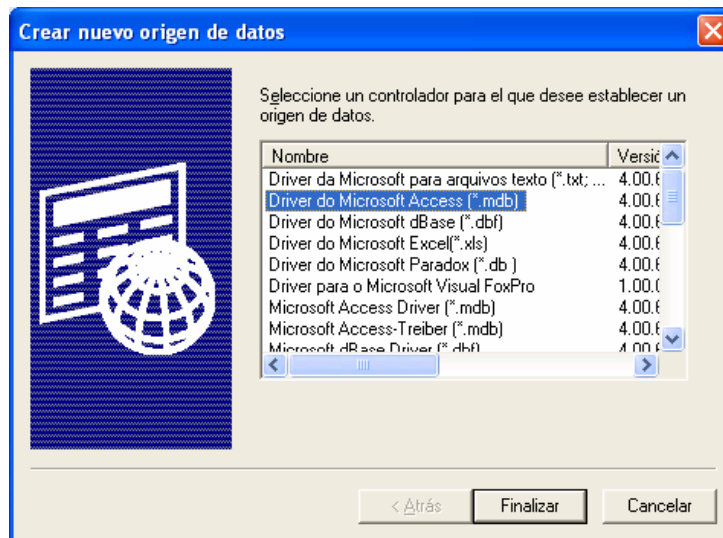


Figura 55. Administrador de orígenes de datos ODBC. Creando un nuevo origen.

En la siguiente pantalla pincha en “Seleccionar” y selecciona el fichero “..\LabKDD\Neptuno\Neptuno.mdb”. Le daremos el nombre del origen de datos “Fuente_Neptuno” y la descripción que quieras y aceptar.

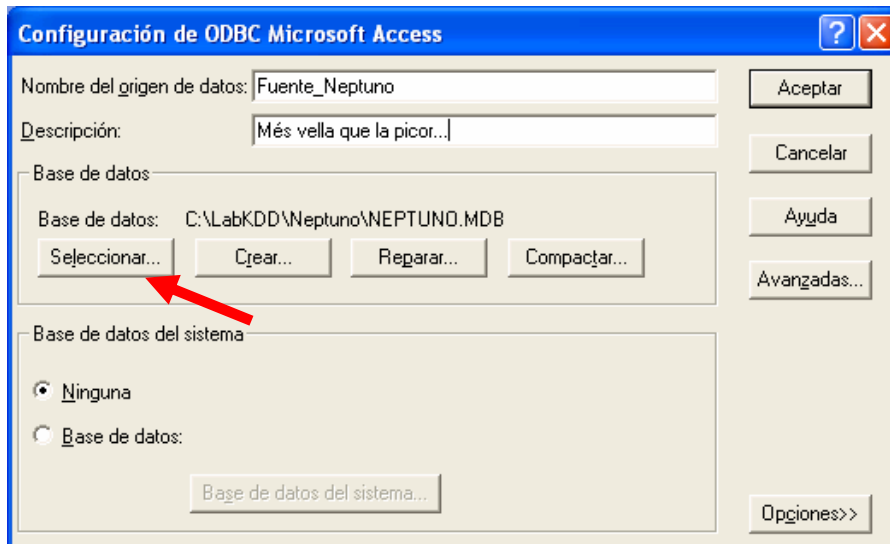


Figura 56. Configurando el nuevo origen.

Ahora volvemos al Clementine. Limpiamos el área de trabajo y añadimos un nodo Base de Datos (SQL) de la categoría Orígenes. Si lo editamos, en el “Nombre de Tabla” pincharemos y seleccionaremos “<Añadir nueva conexión de base de datos..>”, como se muestra en la siguiente figura:

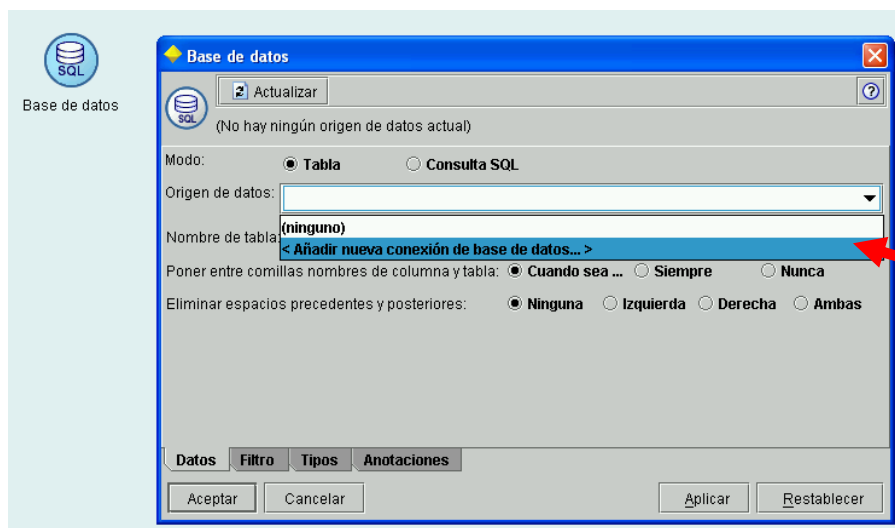


Figura 57. Editando el nodo “Base de datos” de Clementine.

Nos aparecerá la siguiente pantalla, en la que, seleccionando sobre “Fuente_Neptuno”, pulsaremos en “Conectar” (no hace falta ni nombre de usuario ni contraseña) y si todo va bien nos dirá que ha conectado, como se muestra en la siguiente figura:

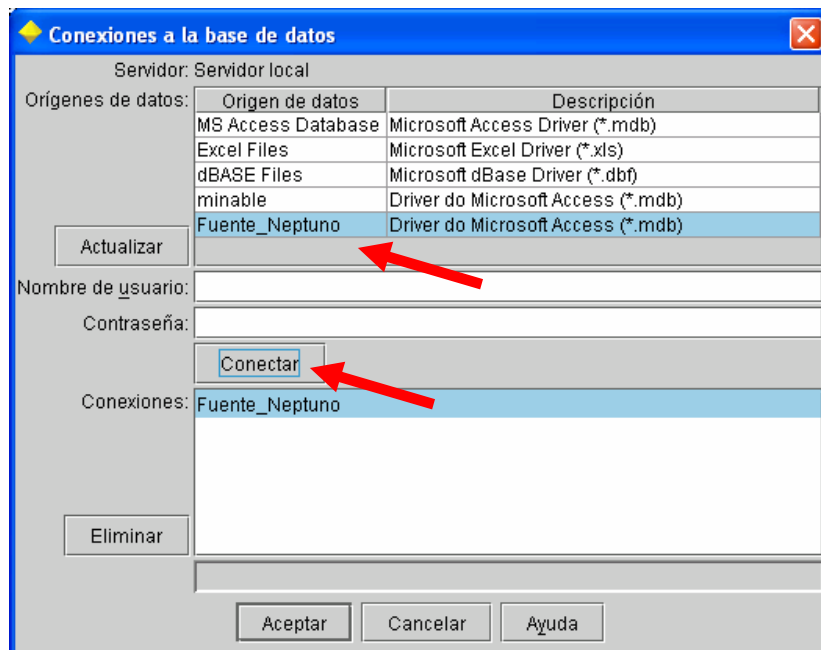


Figura 58. Conectando el nodo Base de Datos con la Fuente ODBC

Pulsa “Aceptar” y pincha en “Seleccionar” el “Nombre de Tabla”:

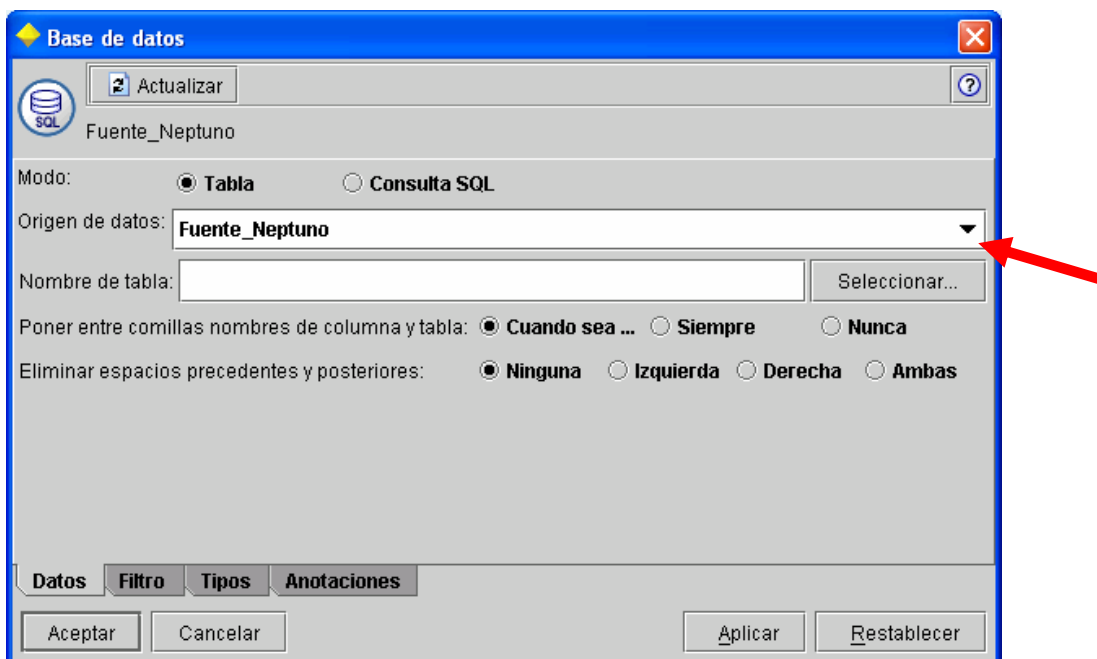


Figura 59. Conectando el nodo Base de Datos con la Fuente ODBC

Ahora, en la siguiente pantalla, selecciona la vista con nombre “_VentasPorCuatrimestre”:

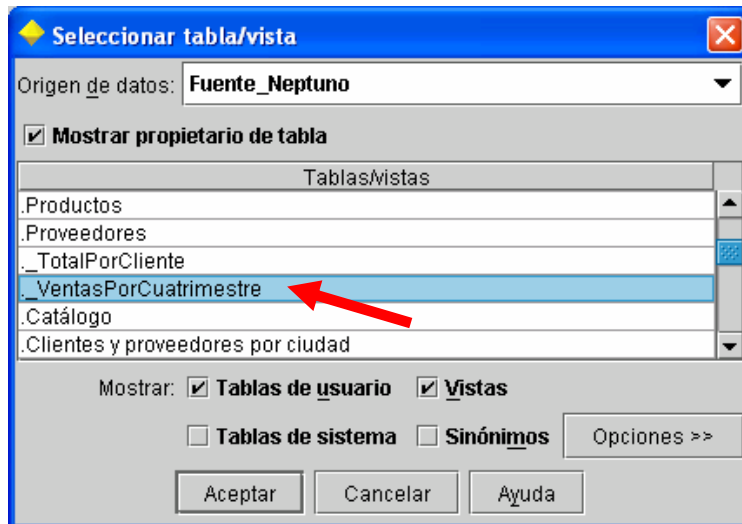


Figura 60. Conectando el nodo Base de Datos con la Fuente ODBC

Pulsa "Aceptar" y tendrás lo siguiente:

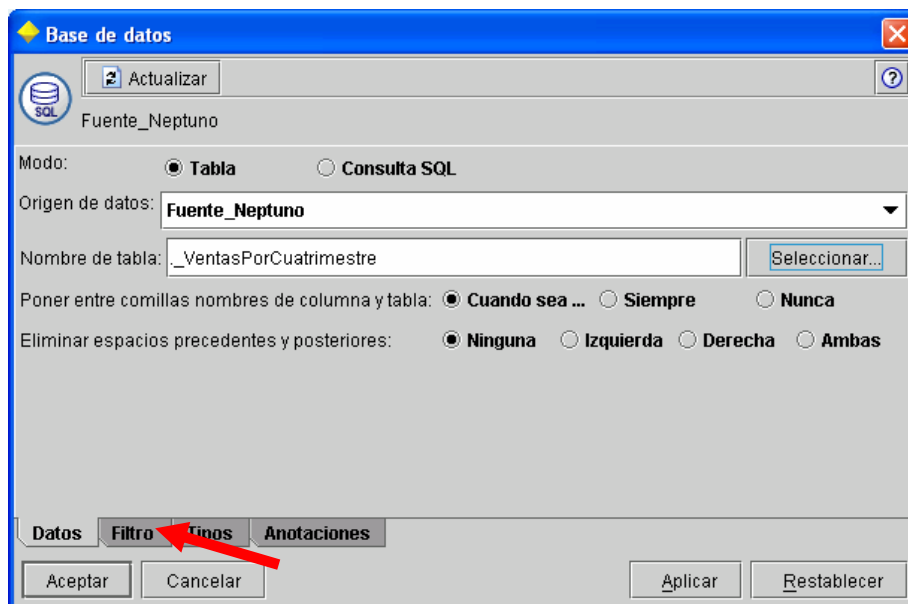


Figura 61. Conectando el nodo Base de Datos con la Fuente ODBC

Ahora pincha en la pestaña "Filtro" y pasarás a la siguiente pantalla:

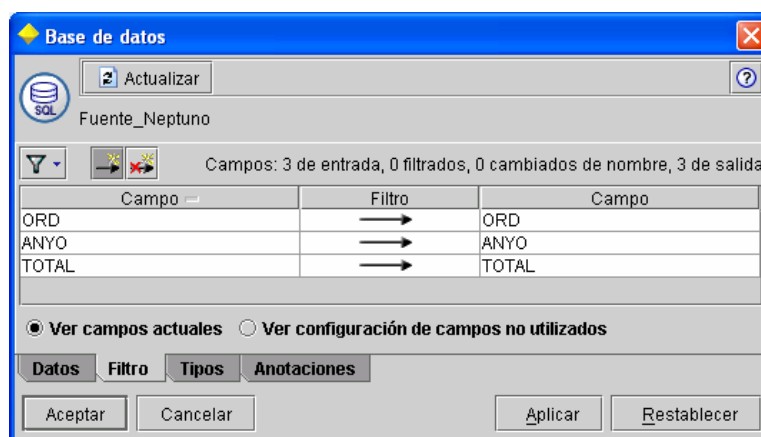


Figura 62. Conectando el nodo ODBC con la Fuente ODBC

Los tres campos nos interesan así que dale a “Aceptar”. Ahora añade un nodo “Tabla” al área de trabajo y conéctalo con la fuente de datos. Ejecuta la ruta. El resultado será que los datos se cargan desde la bases de datos, como se ve en la siguiente figura:

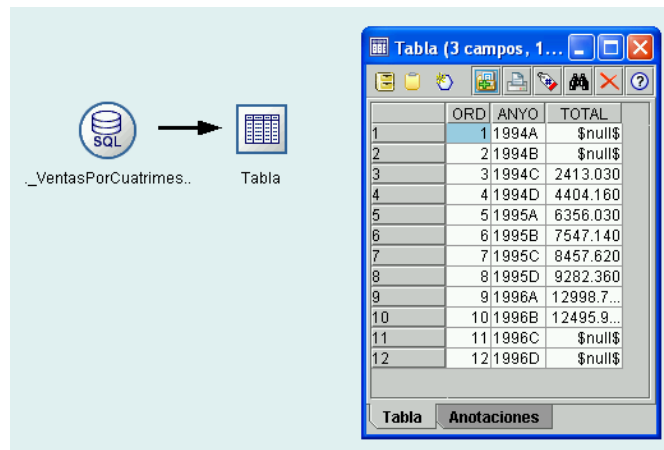


Figura 63. Datos conectados a través de ODBC

Ahora, en primer lugar debemos descartar los valores nulos, ya que no nos van a servir para predecir. Esto se puede hacer con un nodo “Tipo”. Si lo enganchamos al nodo “Fuente_Neptuno”, al editarlo podemos indicar lo siguiente:

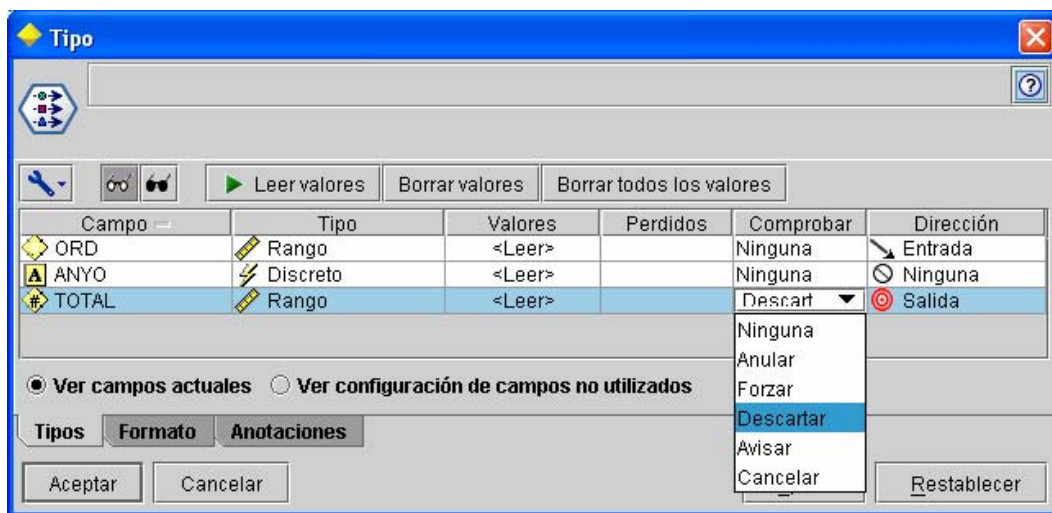


Figura 64. Tipando e indicando entrada y salida

Con la opción “Descartar” en el campo TOTAL descartamos aquellos que no sean del tipo especificado, en este caso reales. Además hemos eliminado ANYO (dirección = NINGUNA) porque utilizaremos ORD como valor numérico para los años y hemos puesto el campo “TOTAL” como el campo “Salida”.

En segundo lugar, parece observarse en la tabla (ver Figura 63) que los datos del trimestre 1996B pueden no estar acabados y los datos del trimestre 1994C pueden ser incompletos. Lo mejor es ignorarlos. Para ello, necesitamos añadir nodos “seleccionar” para quitarlos. Existe una manera muy cómoda de hacerlo. Sobre los propios datos de la tabla que genera el nodo tabla cuando se ejecuta señalamos aquellas filas que no nos interesan, como se ve en la siguiente figura:

	ORD	ANYO	TOTAL
1	1	1994A	\$null\$
2	2	1994B	\$null\$
3	3	1994C	2413.030
4	4	1994D	4404.160
5	5	1995A	6356.030
6	6	1995B	7547.140
7	7	1995C	8457.620
8	8	1995D	9282.360
9	9	1996A	12998.7...
10	10	1996B	12495.9...
11	11	1996C	\$null\$
12	12	1996D	\$null\$

Figura 65. Generando una condición automáticamente

Acuérdate de pulsar la tecla “Ctrl” para poder seleccionar las dos filas. Seleccionamos el menú “Generar” → “Nodo Seleccionar (“Registros”)”.

Figura 66. Generando una condición automáticamente

Automáticamente nos genera un nodo “Seleccionar” con nombre “(generado)”. Lo conectamos al nodo “Tipo”. Si editamos el nodo “Generar” vemos que nos aparecen justamente las condiciones para incluir esas dos filas. Modificamos el nodo “(generado)” de tal manera que nos excluya los cuatrimestres que no queremos, pinchando en “Descartar”, quedando de la siguiente manera:

Modo: Incluir Descartar

Condición: ORD = 3 and ANYO = "1994C" and TOTAL = 2413.03 or
ORD = 10 and ANYO = "1996B" and TOTAL = 12495.9

Figura 67. Modificando una condición

Si conectamos un nuevo nodo “Tabla” a la salida del nodo “(generado)” podremos ver que ya sólo tenemos los datos que nos interesan (si todavía salieran los valores \$null\$, cierra la tabla y vuelve a ejecutarla, o a veces hace falta editar el nodo “tipo” de nuevo y volver a ejecutar):

	ORD	ANYO	TOTAL
1	4	1994D	4404.160
2	5	1995A	6356.030
3	6	1995B	7547.140
4	7	1995C	8457.620
5	8	1995D	9282.360
6	9	1996A	12998.780

Figura 68. Resultado del filtro

Ahora es el momento de obtener un modelo. En primer lugar vamos a añadir un nodo “Gráfico” para ver la curva de crecimiento. Lo conectamos al nodo “Generado”. Editamos el Nodo “Gráfico” y marcamos los campos X e Y como ORD y TOTAL, respectivamente.

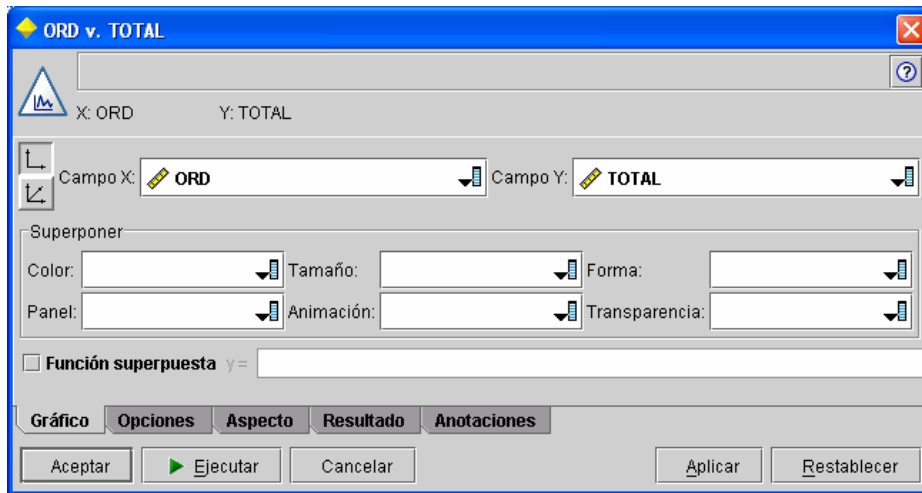


Figura 69. Editando un nodo “Gráfico”

También pincharemos en “Opciones” y allí marcaremos en “Línea”, como se ve en la siguiente figura.

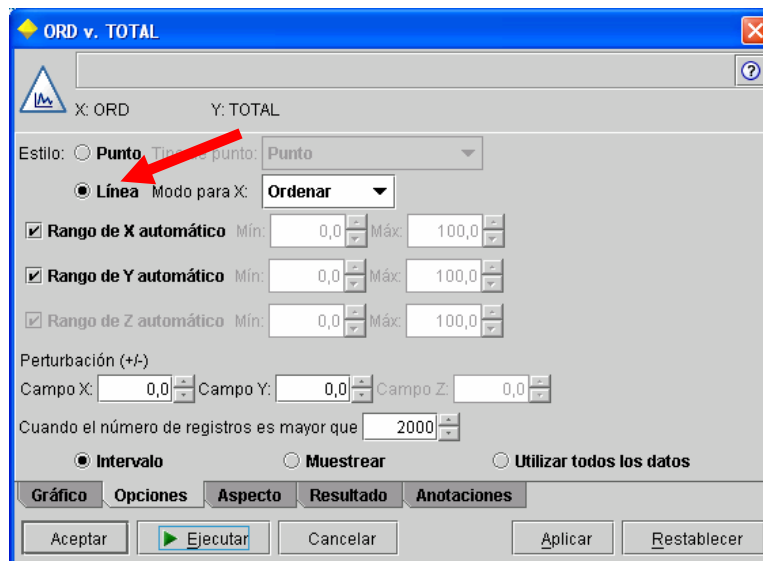


Figura 70. Editando un nodo “Gráfico”

En aspecto seleccionaremos que no nos muestre la “línea cuadriculada”. El resultado del Gráfico se ve en la siguiente figura:

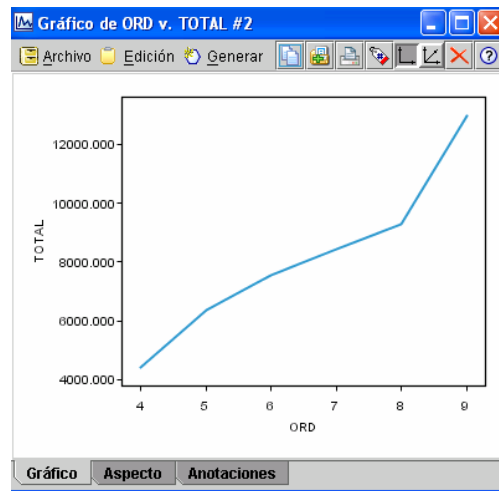


Figura 71. Resultado un gráfico

Vemos que, aunque al final aparece un repunte importante, la evolución se podría aproximar bastante a la linealidad. Por ello, vamos a aplicar un nodo “Regresión” de la categoría de “Modelado” y lo conectamos al nodo “Generado”.

Ya podemos darle a ejecutar. Nos aparece un nuevo modelo en la zona derecha del Clementine que lo pinchamos para que aparezca en la zona de trabajo. Tendremos la siguiente situación:

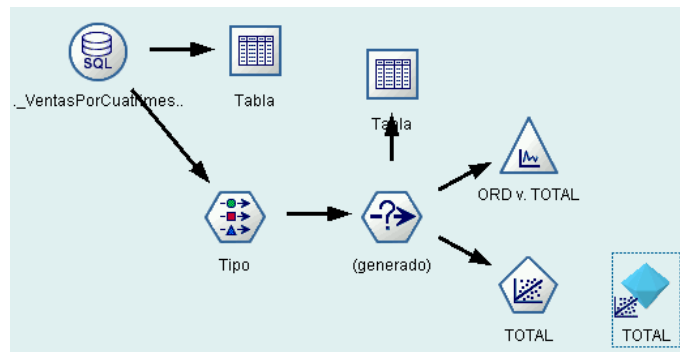


Figura 72. Ruta Resultante

Por último, el modelo resultante se puede ver si editamos el nodo diamante:

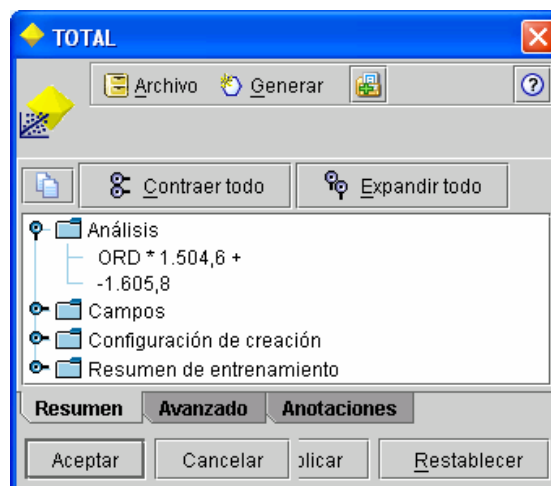


Figura 73. Modelo Resultante (Ecuación lineal)

Para evaluarlo vamos a utilizar el nodo diamante y lo enganchamos al nodo tipo. A continuación del nodo diamante “TOTAL” añadimos un nodo “gráfico múltiple”. En este, ponemos como campo X el campo ORD y como campos Y vamos a poner TOTAL que es el valor real y \$E-TOTAL que es el valor generado.

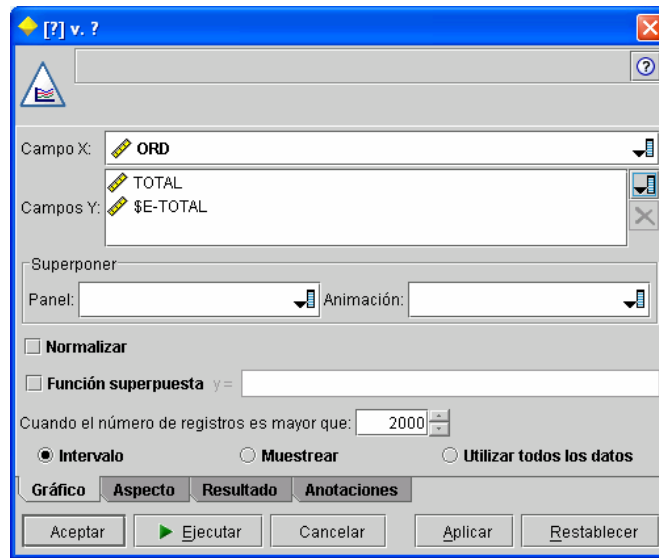


Figura 74. Editando el gráfico múltiple.

Ahora tenemos la siguiente ruta:

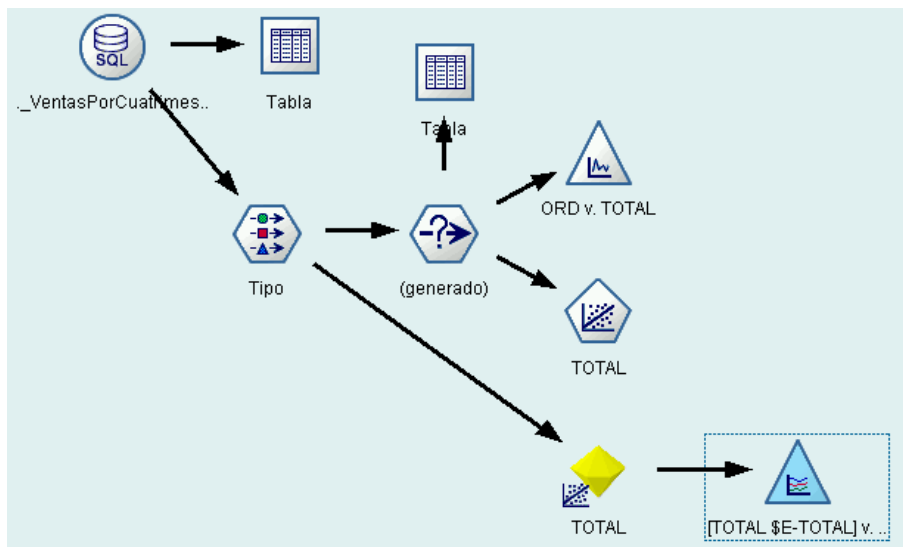


Figura 75. Evaluando el modelo gráficamente

Si lo ejecutamos tenemos el siguiente resultado:

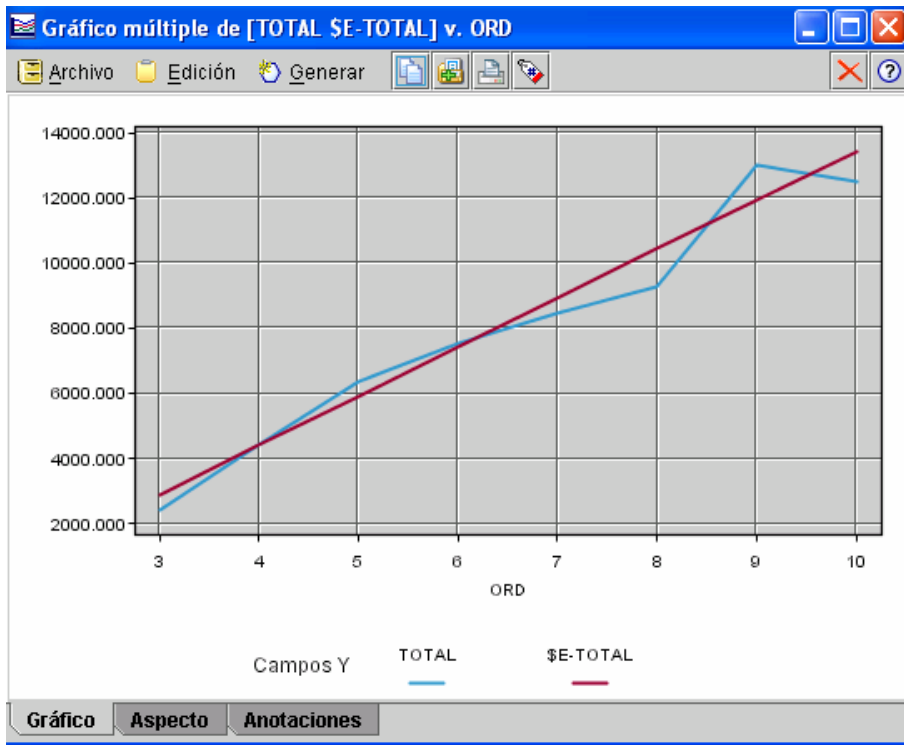


Figura 76. Comparando la curva real con el valor predicho

Donde podemos ver que el modelo se ajusta bastante bien a la curva real.

Ahora si queremos aplicarlo para cualquier valor, utilizamos un nodo “datos de usuario”. Para ello cogemos otro nodo diamante y lo enganchamos con el nodo “(generado)”. Pinchamos con el botón derecho en el nodo diamante y allí pinchamos en la opción “Generar nodo de datos de usuario”, como se ve en la siguiente figura:

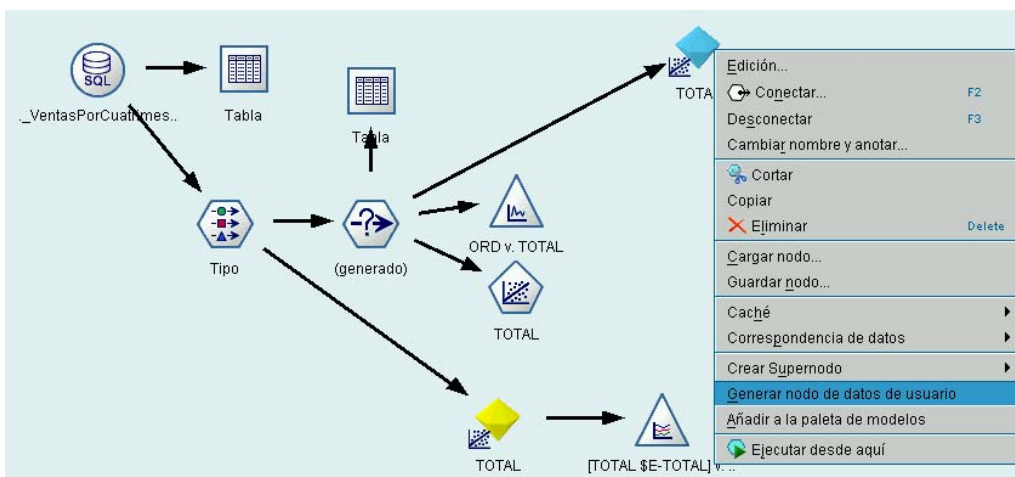


Figura 77. Creando un nodo de datos de usuario.

Nos ha aparecido un nuevo nodo “Datos de usuario”. Lo editamos e insertamos los valores que queremos predecir (trimestres 11 y 12). Por ejemplo, lo que aparece en la siguiente figura (el 10 también lo ponemos para comparar con lo que teníamos):

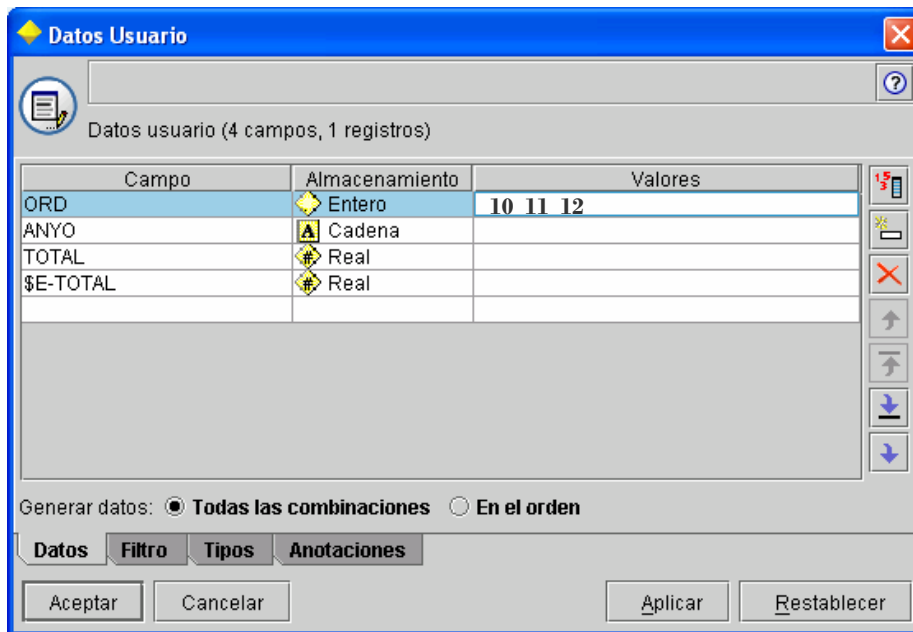


Figura 78. Entrada de datos manual

Le damos a aceptar y ahora desconectamos el nodo diamante "Total" del nodo "(generado)" (Recuerda, pinchando en el enlace con el botón derecho: "Eliminar Conexión"). Conectamos el nodo "Datos de usuario" con el nodo diamante "Total" y éste a su vez con un nuevo nodo Tabla. Ejecutamos el nodo Tabla y vemos lo que tenemos en la figura siguiente:



Figura 79. Utilizando el modelo para predecir los datos

Como vemos, predecimos un valor de 13440,6 para el trimestre 10 (1996B), que supusimos no terminado (y que es ligeramente superior al existente, 12495, lo que parece lógico), y unas predicciones de 14945,3 para el trimestre 11 (1996C) y una predicción de 16449.9 para el trimestre 12.

Ahora podemos grabar la ruta.

EJERCICIOS PROPUESTOS:

- ¿Has notado alguna diferencia por el hecho de estar conectado a la base de datos respecto a los ejercicios anteriores que leías los datos de un fichero?
- Intenta conectarte a alguna tabla más grande de la base de datos y observa si le cuesta cargar los datos mucho más que si estuviera en un fichero de texto.