Universidad Don Bosco, El Salvador

Datawarehouse y Minería de Datos

DMD941 G01T

Desafío Práctico 2

Docente:

MG. Karens Medrano

Estudiante

LUIS JOSE	CRUZ MARTINEZ	CM170741

Enlace del repositorio de GitHub

GitHub - Link al Desafío Práctico 1

Porcentaje alcanzado

100%

Fecha de entrega:

Sábado 8 de Noviembre, 2025

INTRODUCCIÓN

El documento presentado a continuación corresponde al Desafío Práctico número 2 de la materia Datawarehouse y Minería de Datos. El objetivo principal del desafío es la aplicación de técnicas de minería de datos utilizando el algoritmo de árboles de decisión para analizar los distintos archivos de datos compartidos y así poder determinar los patrones de comportamiento de los clientes aplicando los filtros requeridos.

En el desarrollo de este desafío se ha utilizado los archivos compartidos como recursos en el aula digital, los cuales son: primer_compra.xlsx, estado_civil.xsl y whisky.xslx, tales archivos han sido procesados en la herramienta RapidMiner que actualmente se llama Altair AI Studio luego del rebranding. Es un programa para minería y análisis de datos, permite crear modelos de datos para el análisis predictivo de grandes volúmenes de datos. Se han aplicado diversos criterios y filtros según la naturaleza de los recursos compartidos.

En el desarrollo del desafío se han llevado a cabo los siguientes ejercicios principales:

1. Generación de árboles de decisión, esto con el objetivo de determinar cuando una persona puede ser considerada cliente potencial, a partir de los campos del archivo.

2. Aplicación de filtro por tipo de pago, en este caso se consideró únicamente las categorías específicas cash y cheque.

3. Aplicación de filtro de edad, con el objetivo de segmentar datos y observar la variación de los resultados del modelo de datos según los rangos definidos.

En el desarrollo de cada tarea he adquirido una comprensión mayor de la utilidad de los árboles de decisión como una de las herramientas que nos permiten clasificar y predecir los comportamientos ya sea de usuarios individuales o de una magnitud más amplia de usuarios/clientes. La aplicación de filtros que son un punto importante para mejorar la segmentación y precisión de los datos para futuros análisis de los procesos.

Análisis de los resultados

Como resultado del desarrollo y análisis de los resultados, aplicando los algoritmos de árboles de decisión y operadores de filtro para identificar patrones sobre los clientes con mayor probabilidad de compra.

1. Tomando como base el primer recurso primer_compra, podemos determinar las variables que influyen para que una persona realice su primera compra. El árbol de decisión generado muestra que las variables de ingreso y edad son las que más influyen en el resultado.

   Las personas con ingresos medios - altos y una edad entre 25 y 40 años tienen mayor tendencia a realizar una compra.También el género masculino tiene mayor probabilidad de compra en comparación con las mujeres.

   Esto lo podemos interpretar así:
   Los hombres jóvenes con estabilidad económica y poder adquisitivo medio o alto son los clientes más propensos a realizar su primera compra.

2. Ahora nos basamos en el estado_civil y con el árbol de decisión construido con el dataset vemos que las variables estado civil y nivel de ingreso están relacionadas.

   Las personas casadas o en pareja y con ingresos medios - altos demuestran un comportamiento más estable y con planes en sus decisiones de compra, y los jóvenes solteros tienden a un consumo no tan recurrente.

   Lo podemos detallar así:
   La estabilidad económica y familiar es un factor importante para mantener los hábitos de consumo recurrentes. Es por eso que los casados con ingresos de medios a altos tienden a comprar productos de precio medio en adelante.

3. Luego analizamos con el tercer recurso whisky aquí tenemos el precio ya incluido lo que nos permite obtener más detalle. Los whiskies con mayor añejamiento y precio superior a 100 tienen calificaciones de mayor calidad, esto nos permite saber que los clientes con un poder adquisitivo mayor prefieren este tipo de productos.

   Podemos detallar que:
   Los clientes que pueden pagar los precios superiores de un producto de calidad y mayor añejamiento pertenecen a un segmento adulto, y tienen ingresos altos.

Como síntesis podemos decir lo siguiente:

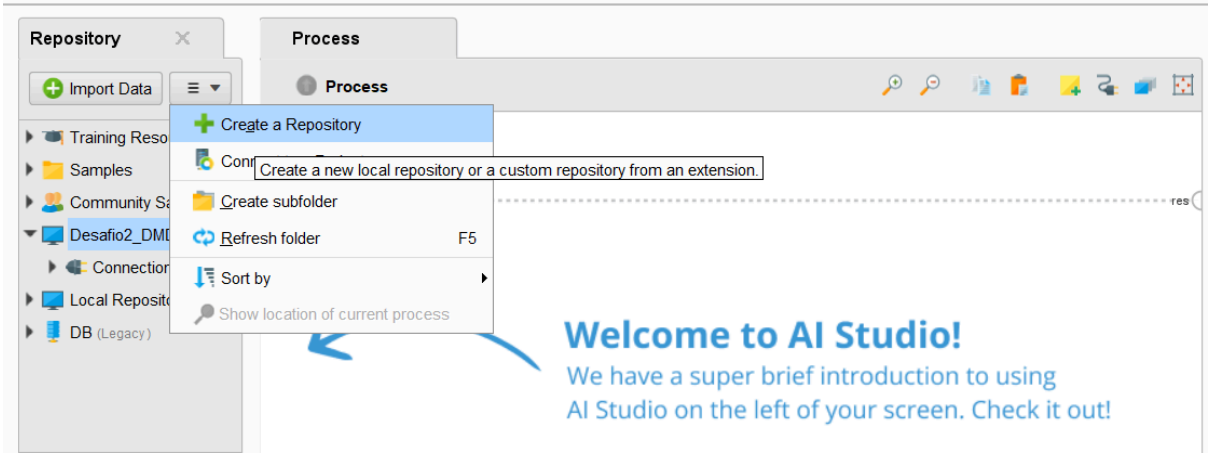
Tomando los resultado de los tres modelos, se puede definir un perfil de cliente con mayor probabilidad de comprar whisky:

Clientes con edad entre 30 y 45 años, con ingresos medio - alto y estado civil casados o en pareja y de género másculino buscan calidad en sus compras.

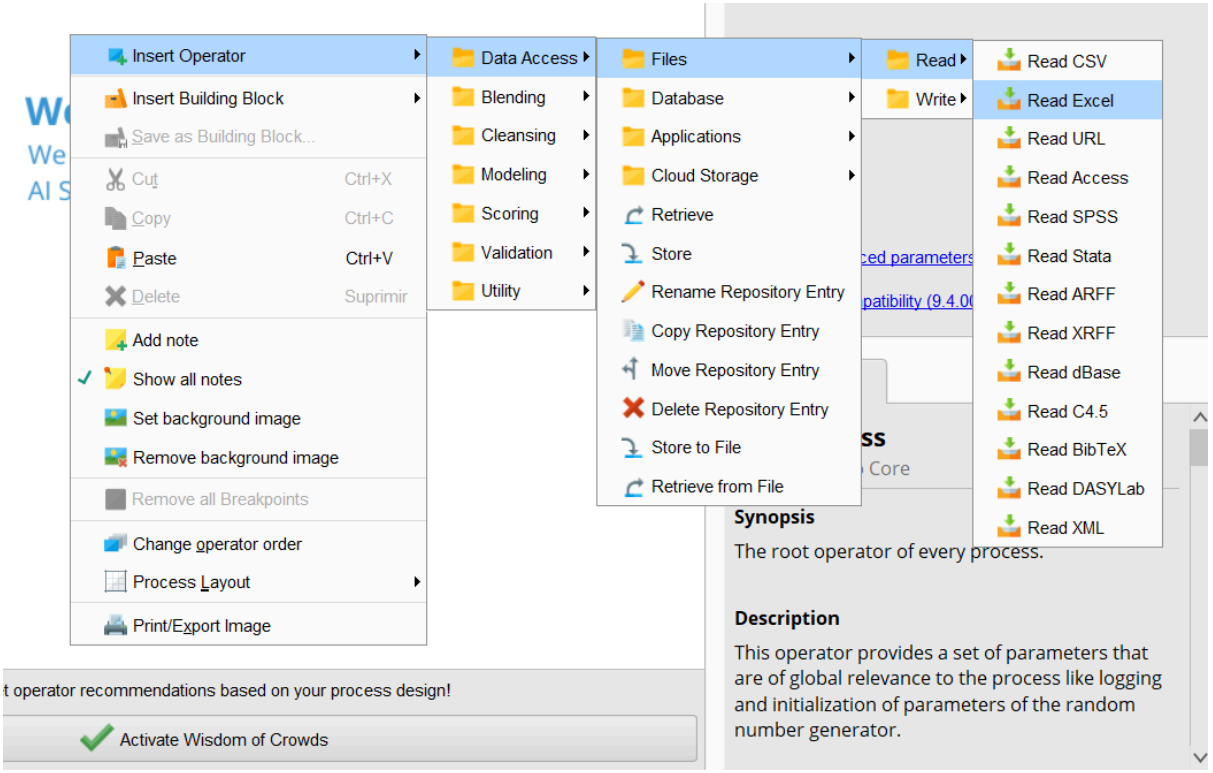
Documentación

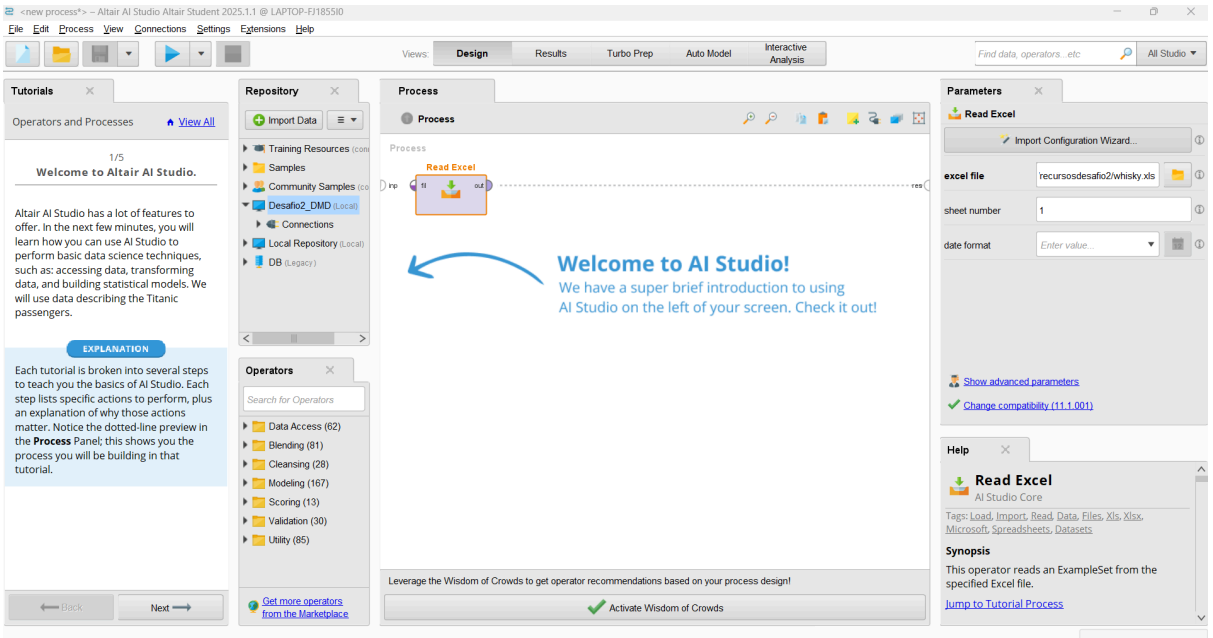
Como primer paso abrimos la herramienta de RapidMiner, ahora llamada Altair AI Studio.

Creamos un nuevo repositorio, en ese caso llamado Desafio2_DMD.

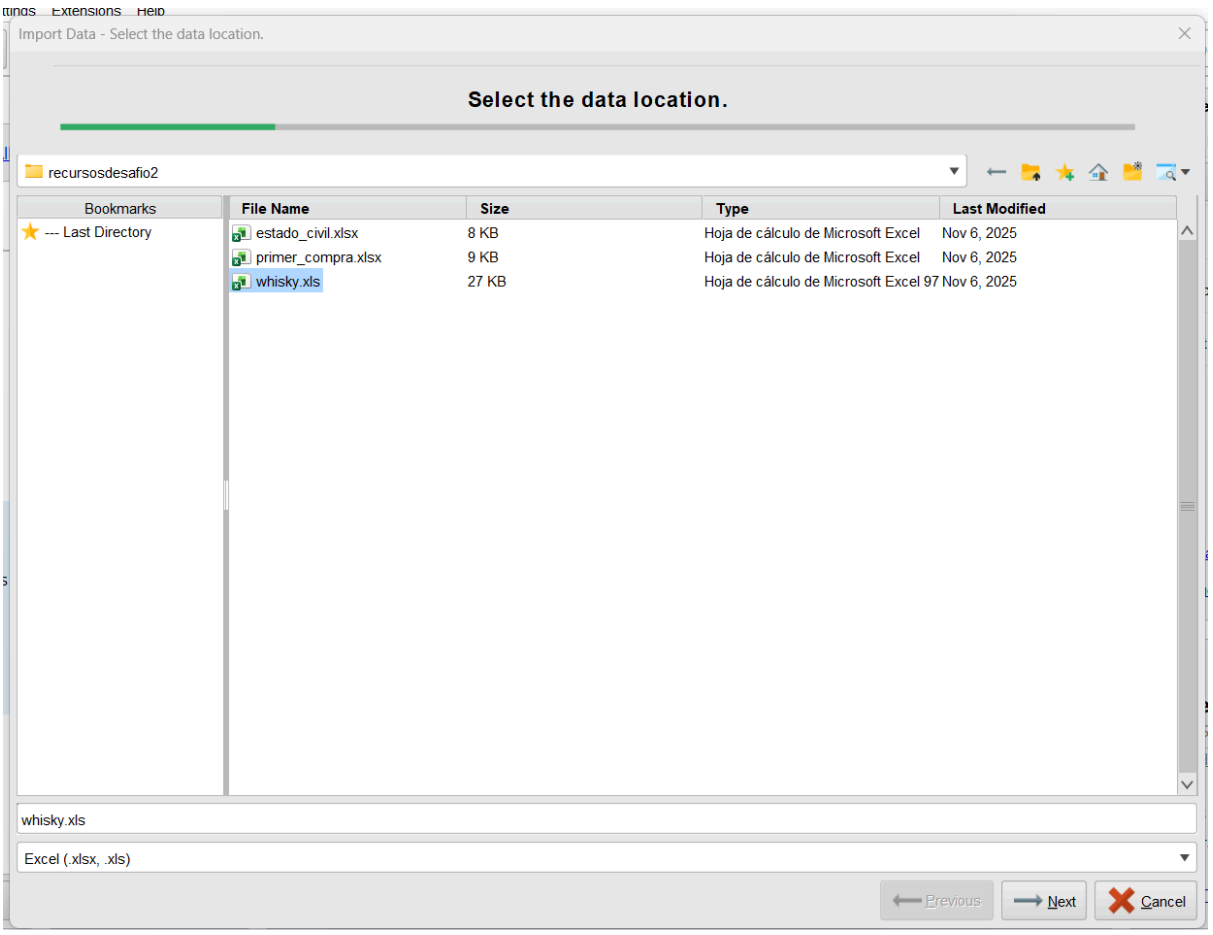


Agregamos un Operador, en Insertar Operador, Data Access, Files, Read, Read Excel.

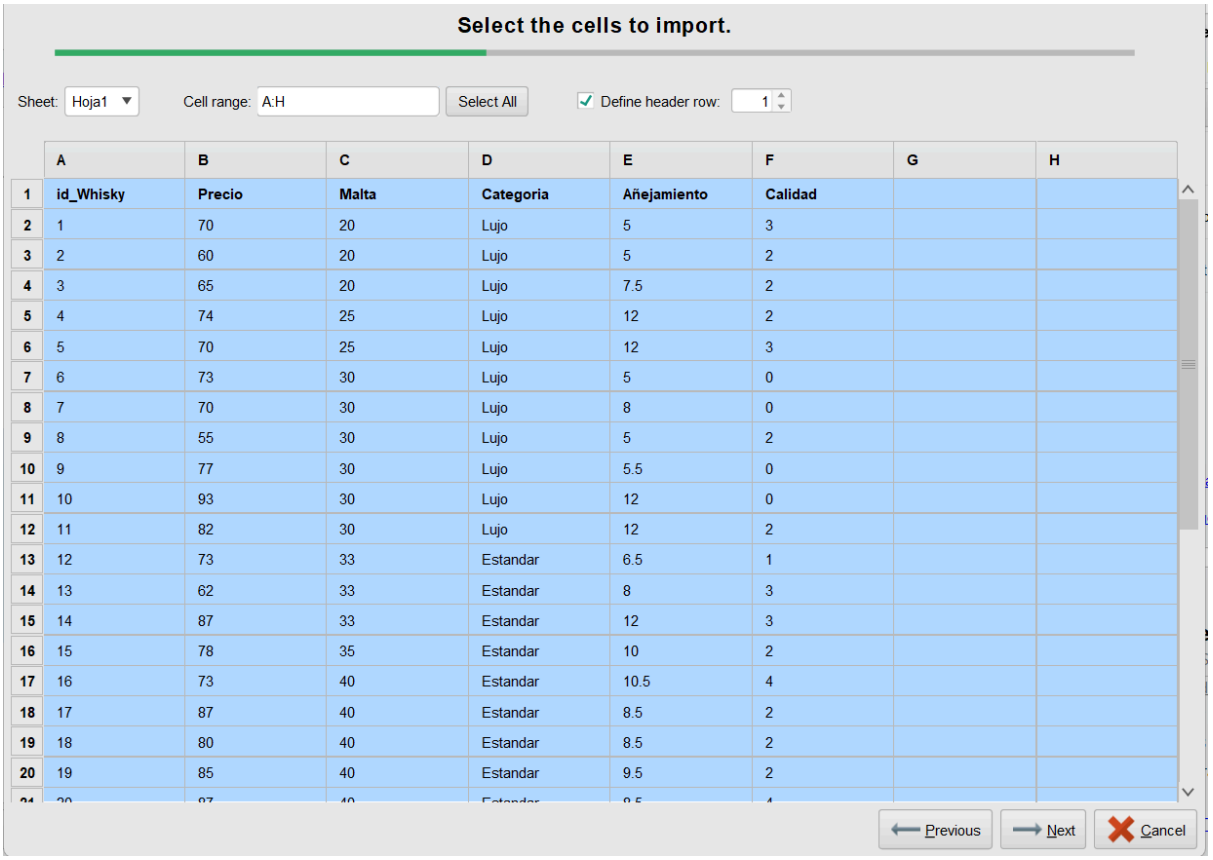




Luego en el input de Excel File, seleccionamos el recurso.



Seleccionamos los datos del archivo que serán importados.



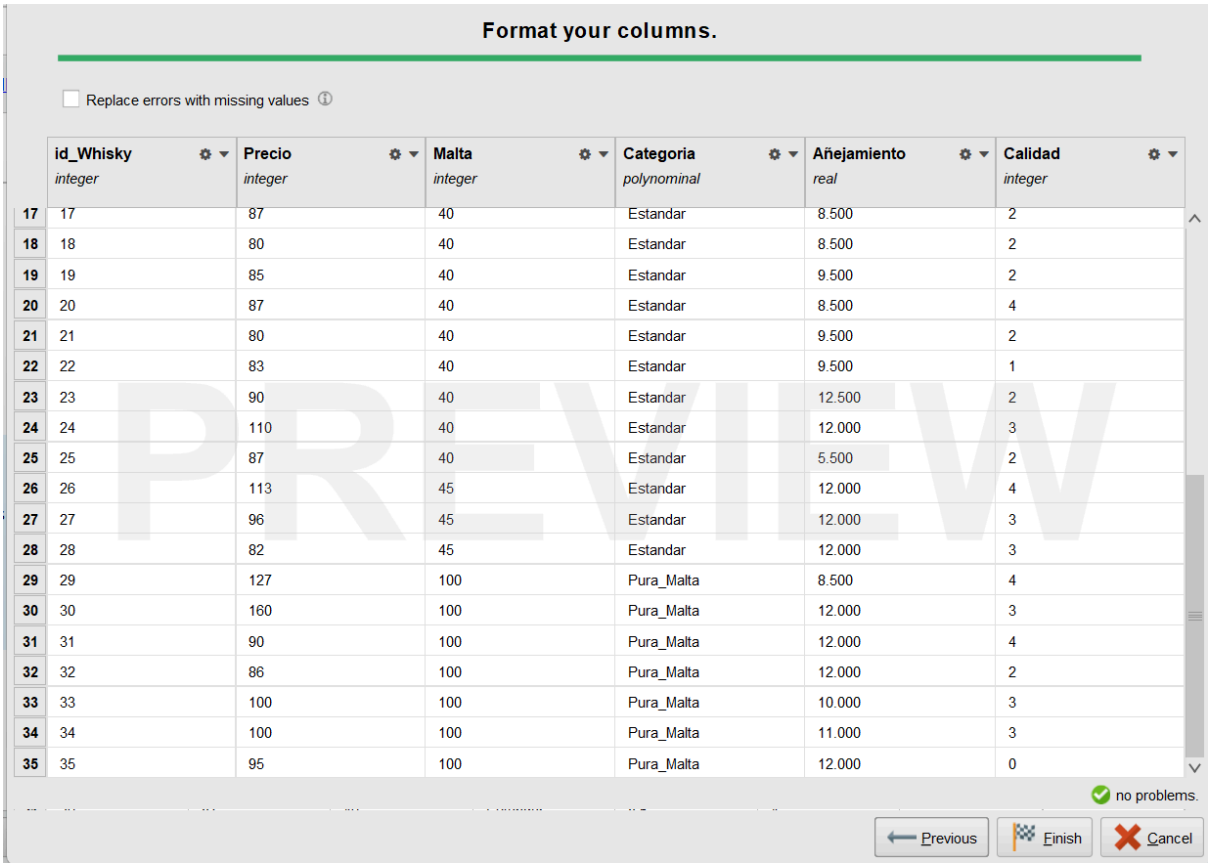
Select the cells to import.

Sheet: Hoja1 | Cell range: A:H | Select All | Define header row: 1

	A	B	C	D	E	F	G	H
1	id_Whisky	Precio	Malta	Categoria	Añejamiento	Calidad		
2	1	70	20	Lujo	5	3		
3	2	60	20	Lujo	5	2		
4	3	65	20	Lujo	7.5	2		
5	4	74	25	Lujo	12	2		
6	5	70	25	Lujo	12	3		
7	6	73	30	Lujo	5	0		
8	7	70	30	Lujo	8	0		
9	8	55	30	Lujo	5	2		
10	9	77	30	Lujo	5.5	0		
11	10	93	30	Lujo	12	0		
12	11	82	30	Lujo	12	2		
13	12	73	33	Estandar	6.5	1		
14	13	62	33	Estandar	8	3		
15	14	87	33	Estandar	12	3		
16	15	78	35	Estandar	10	2		
17	16	73	40	Estandar	10.5	4		
18	17	87	40	Estandar	8.5	2		
19	18	80	40	Estandar	8.5	2		
20	19	85	40	Estandar	9.5	2		

Previous | Next | Cancel

Damos en Next y visualizamos el formato de los datos de la siguiente manera.



Format your columns.

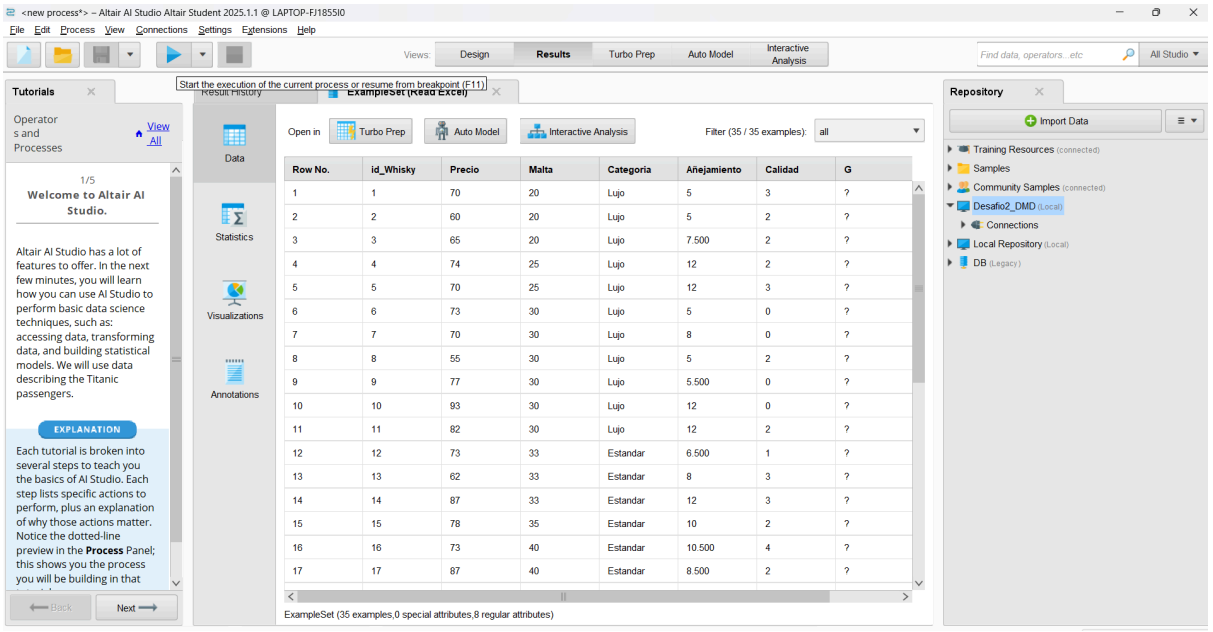
Replace errors with missing values

	id_Whisky *integer*	Precio *integer*	Malta *integer*	Categoria *polynominal*	Añejamiento *real*	Calidad *integer*
17	17	87	40	Estandar	8.500	2
18	18	80	40	Estandar	8.500	2
19	19	85	40	Estandar	9.500	2
20	20	87	40	Estandar	8.500	4
21	21	80	40	Estandar	9.500	2
22	22	83	40	Estandar	9.500	1
23	23	90	40	Estandar	12.500	2
24	24	110	40	Estandar	12.000	3
25	25	87	40	Estandar	5.500	2
26	26	113	45	Estandar	12.000	4
27	27	96	45	Estandar	12.000	3
28	28	82	45	Estandar	12.000	3
29	29	127	100	Pura_Malta	8.500	4
30	30	160	100	Pura_Malta	12.000	3
31	31	90	100	Pura_Malta	12.000	4
32	32	86	100	Pura_Malta	12.000	2
33	33	100	100	Pura_Malta	10.000	3
34	34	100	100	Pura_Malta	11.000	3
35	35	95	100	Pura_Malta	12.000	0

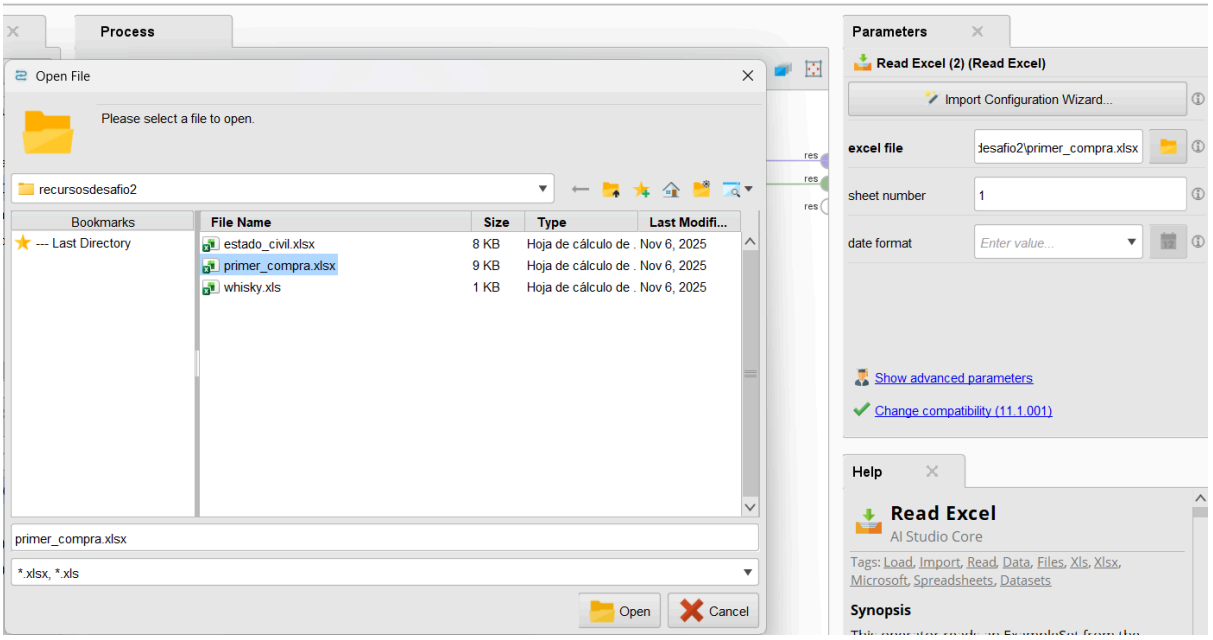
no problems.

Previous | Finish | Cancel

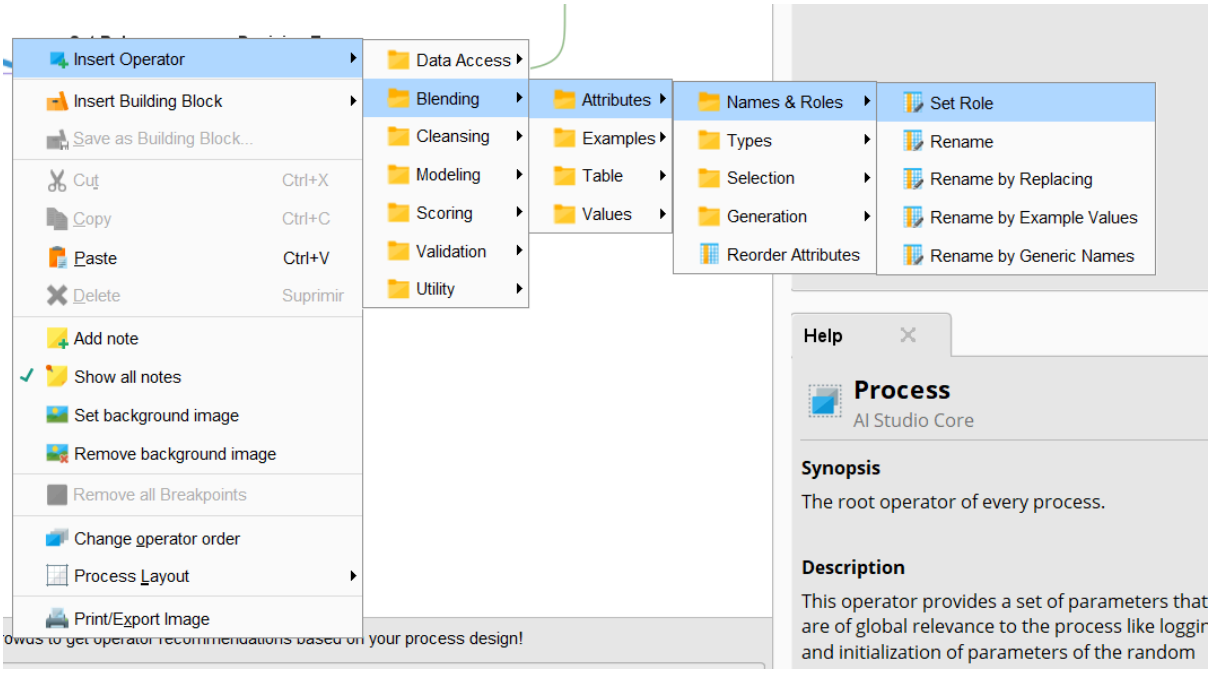
Si corremos el proceso haciendo clic en el botón azul, veremos los datos importados en la siguiente tabla.



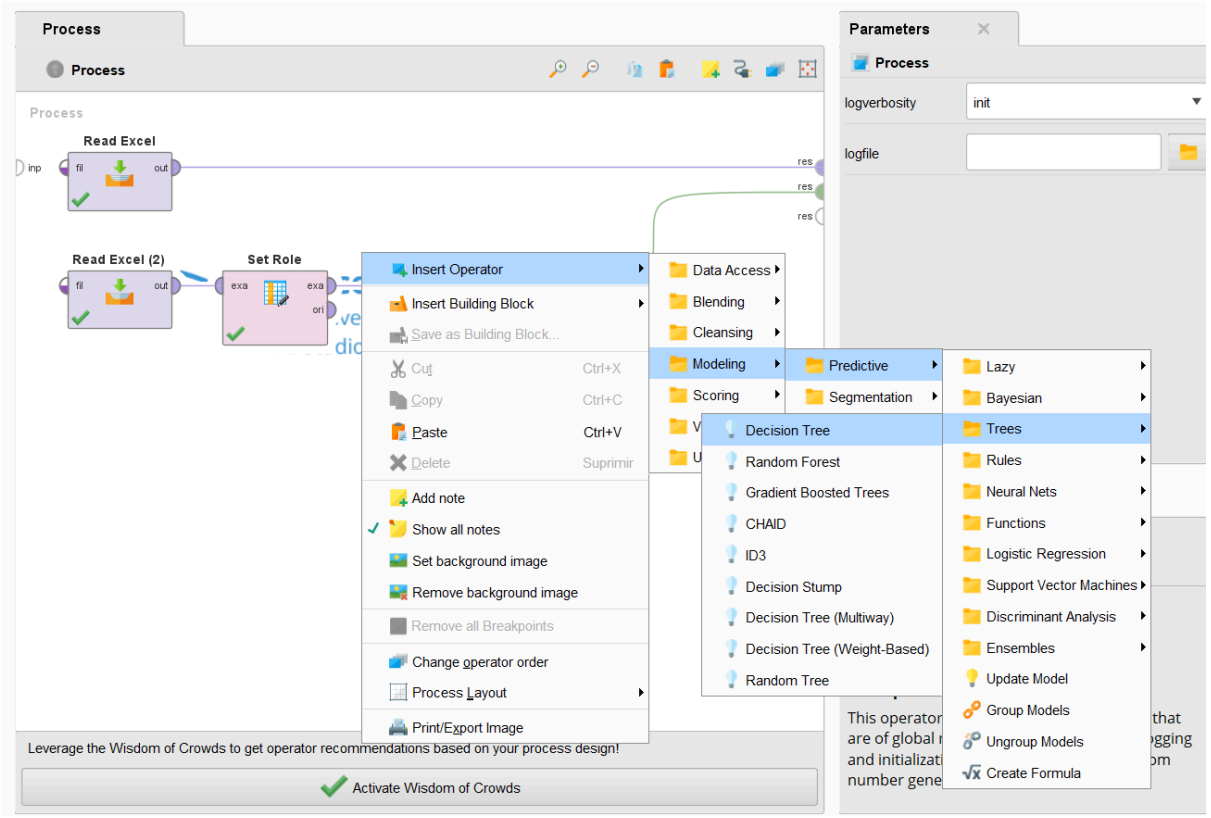
Ahora procedemos a cargar el siguiente recurso de la misma manera, Read Excel.



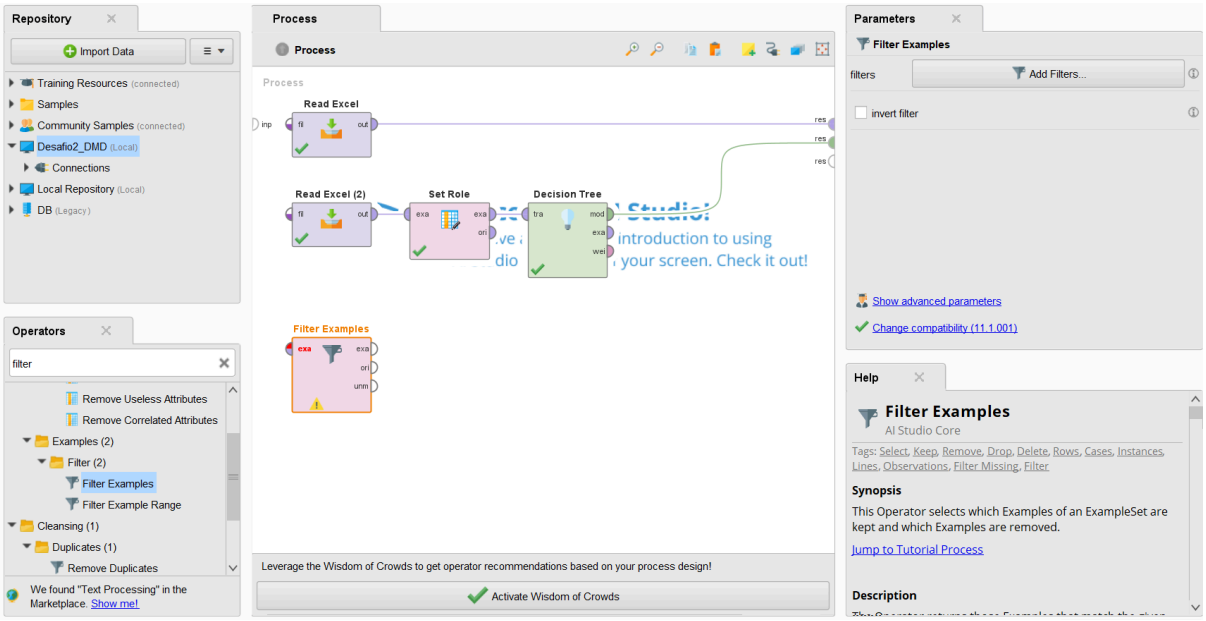
Luego agregaremos el operador Set Role y colocamos Edad como label



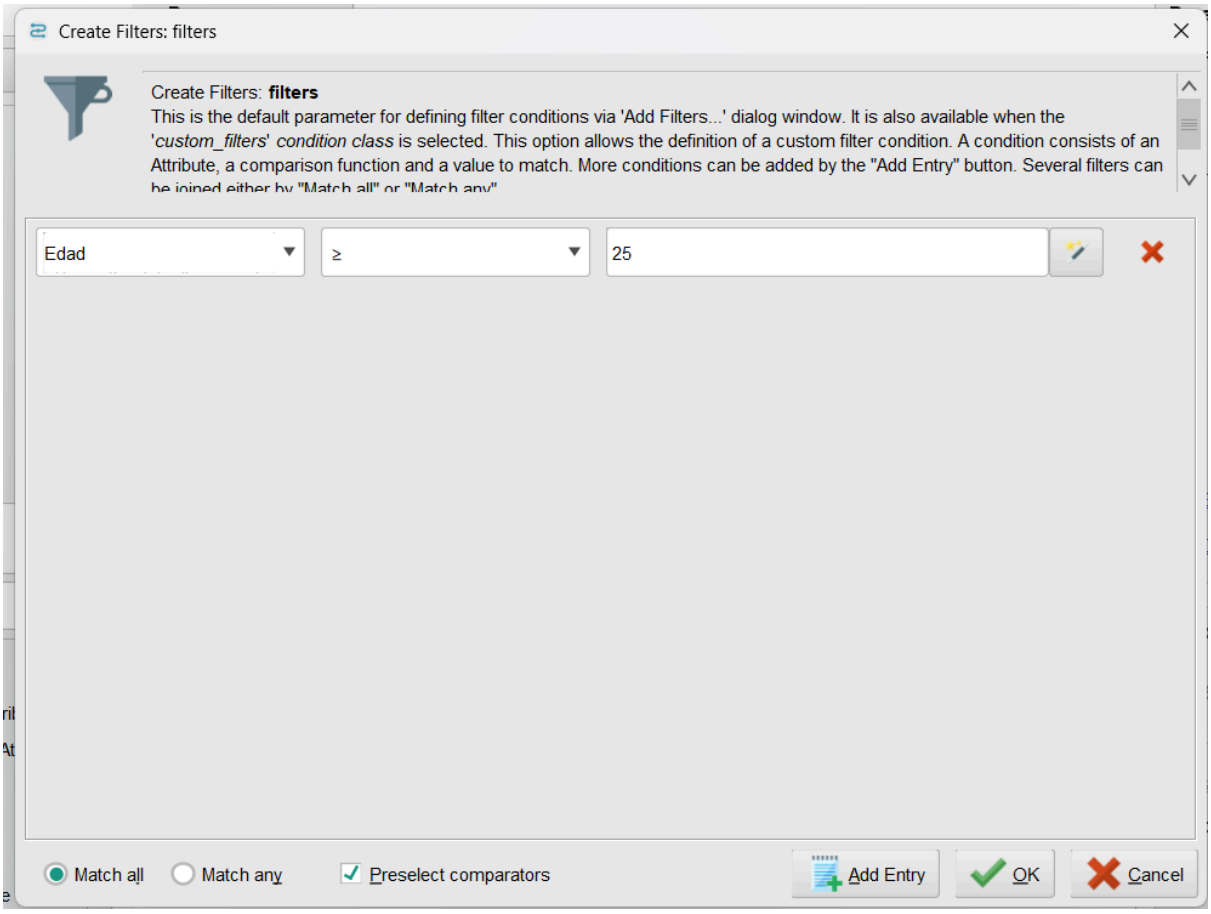
Agregaremos el operador árbol de decisión.



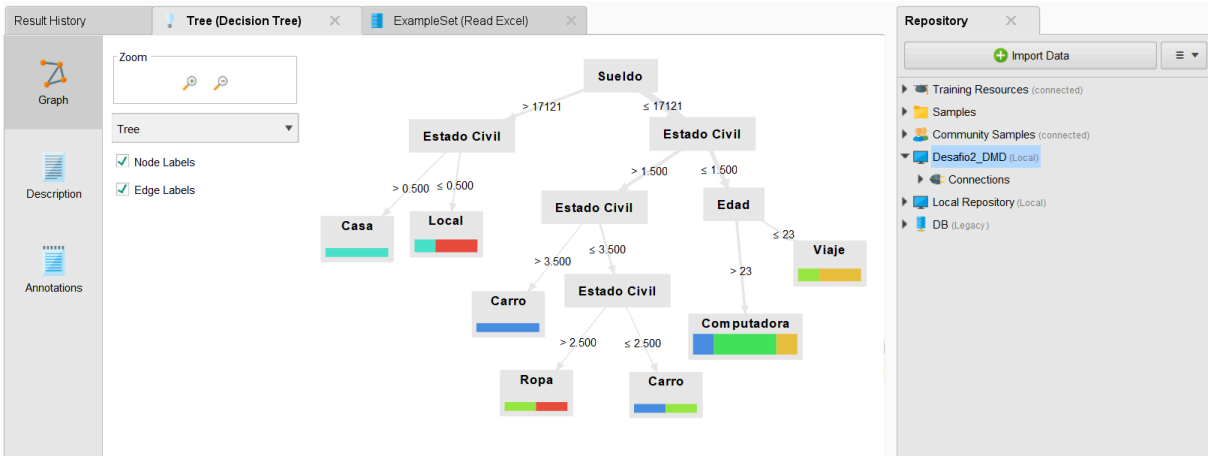
Agregamos el operador Filter Examples



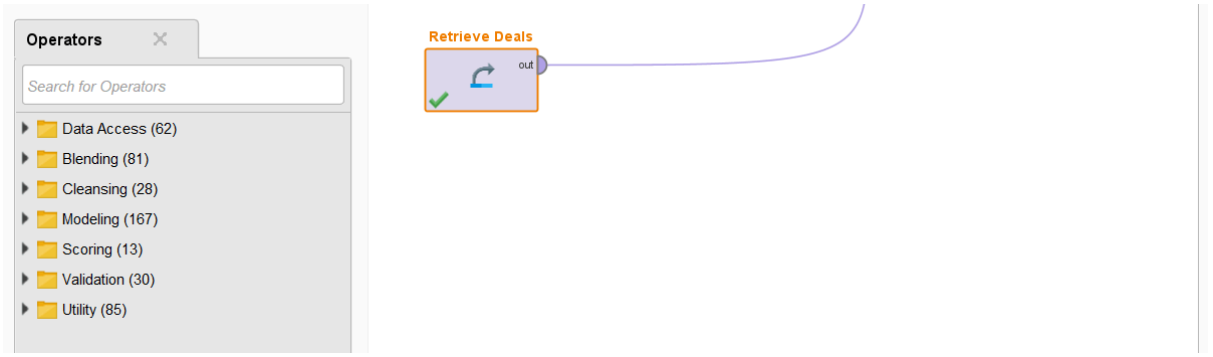
Aplicamos un filtro de ejemplo de edad.



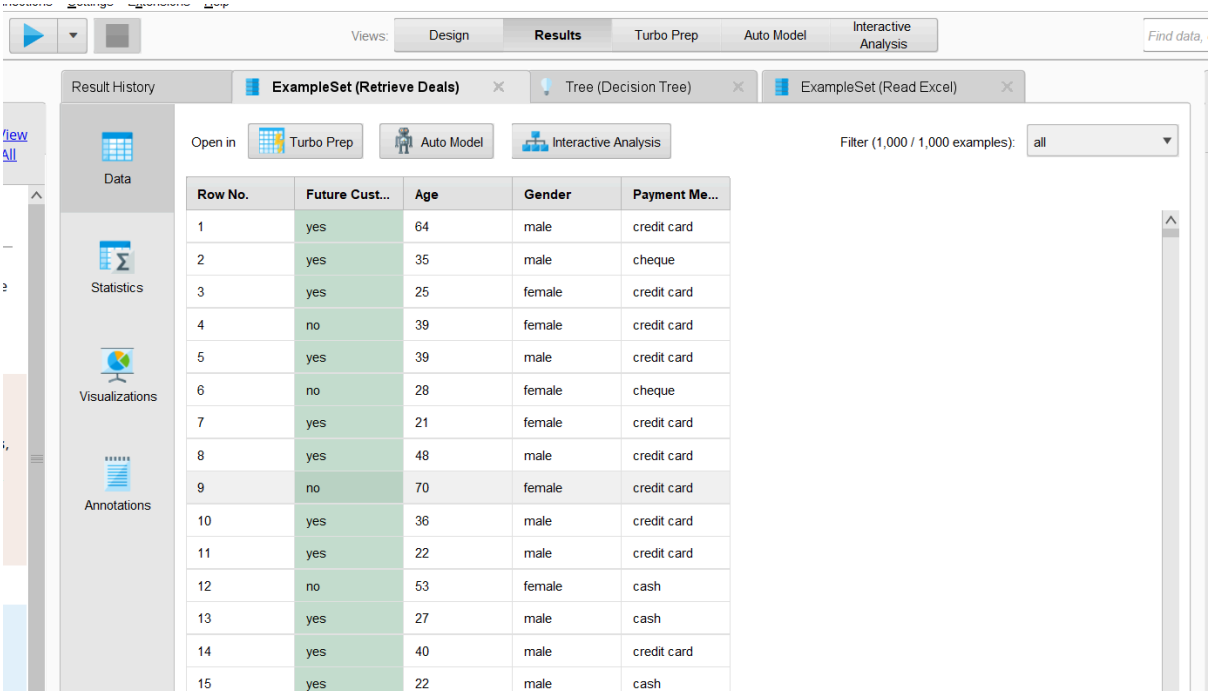
Corremos el proceso y visualizamos el correspondiente árbol de datos.



Ahora procedemos a agregar uno de los samples de datos de RapidMiner.

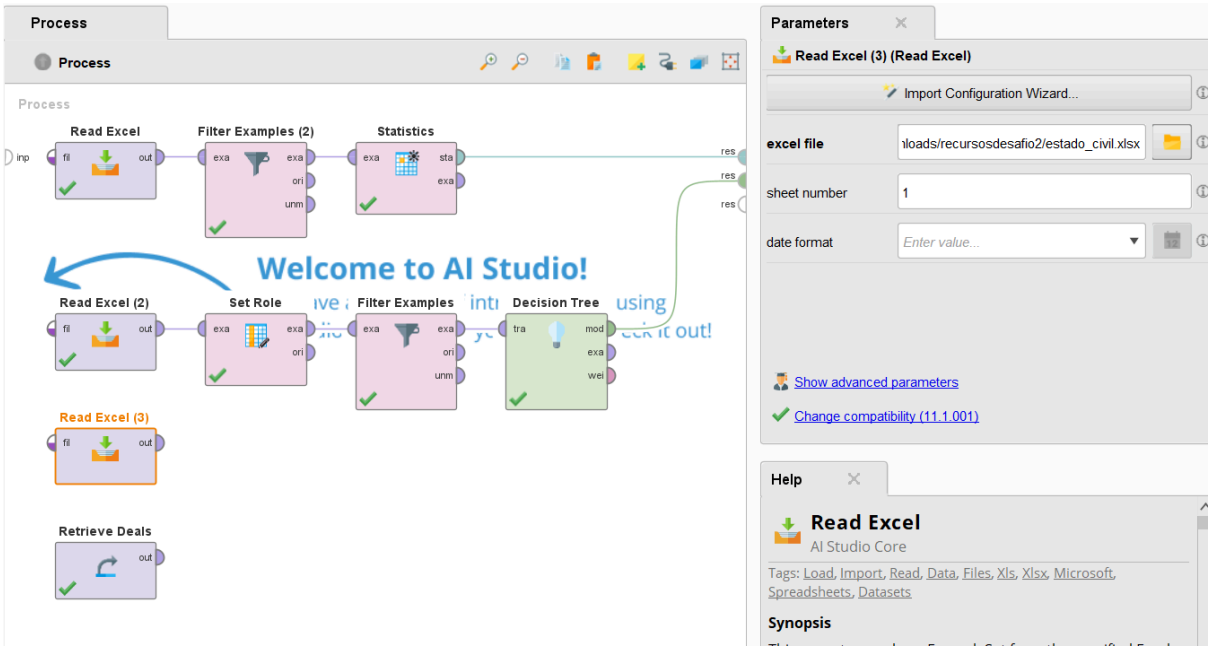


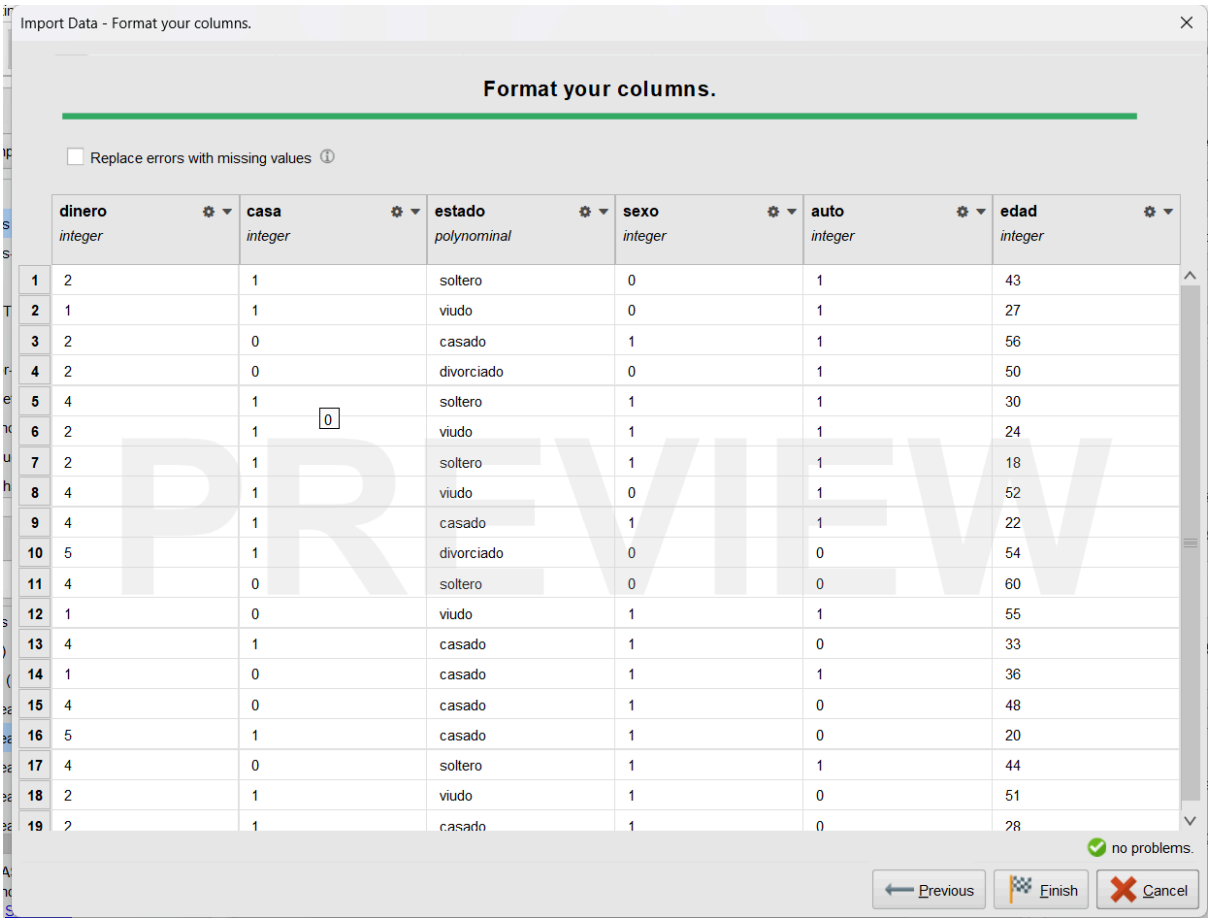
Lo cual nos permitirá obtener el siguiente tipo de datos.



Row No.	Future Cust...	Age	Gender	Payment Me...
1	yes	64	male	credit card
2	yes	35	male	cheque
3	yes	25	female	credit card
4	no	39	female	credit card
5	yes	39	male	credit card
6	no	28	female	cheque
7	yes	21	female	credit card
8	yes	48	male	credit card
9	no	70	female	credit card
10	yes	36	male	credit card
11	yes	22	male	credit card
12	no	53	female	cash
13	yes	27	male	cash
14	yes	40	male	credit card
15	yes	22	male	cash

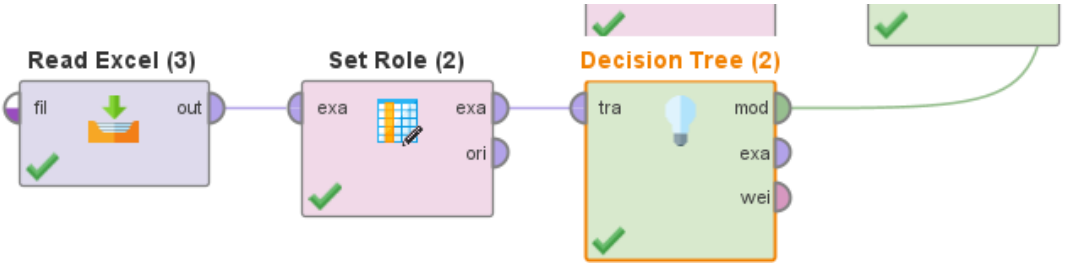
Agregamos el siguiente recurso.

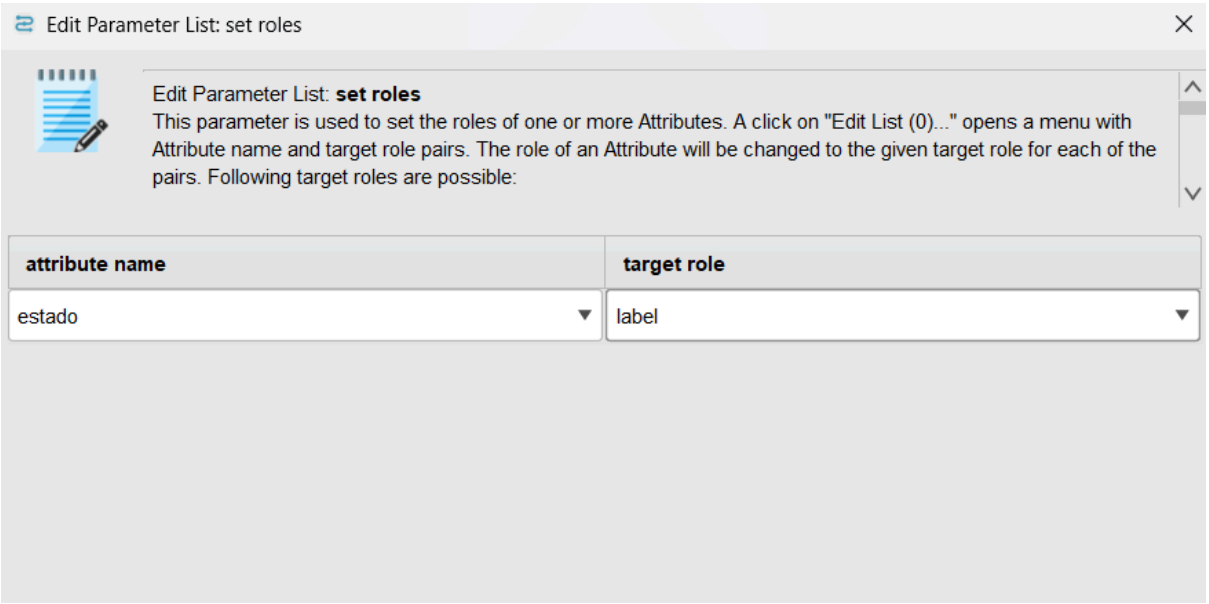




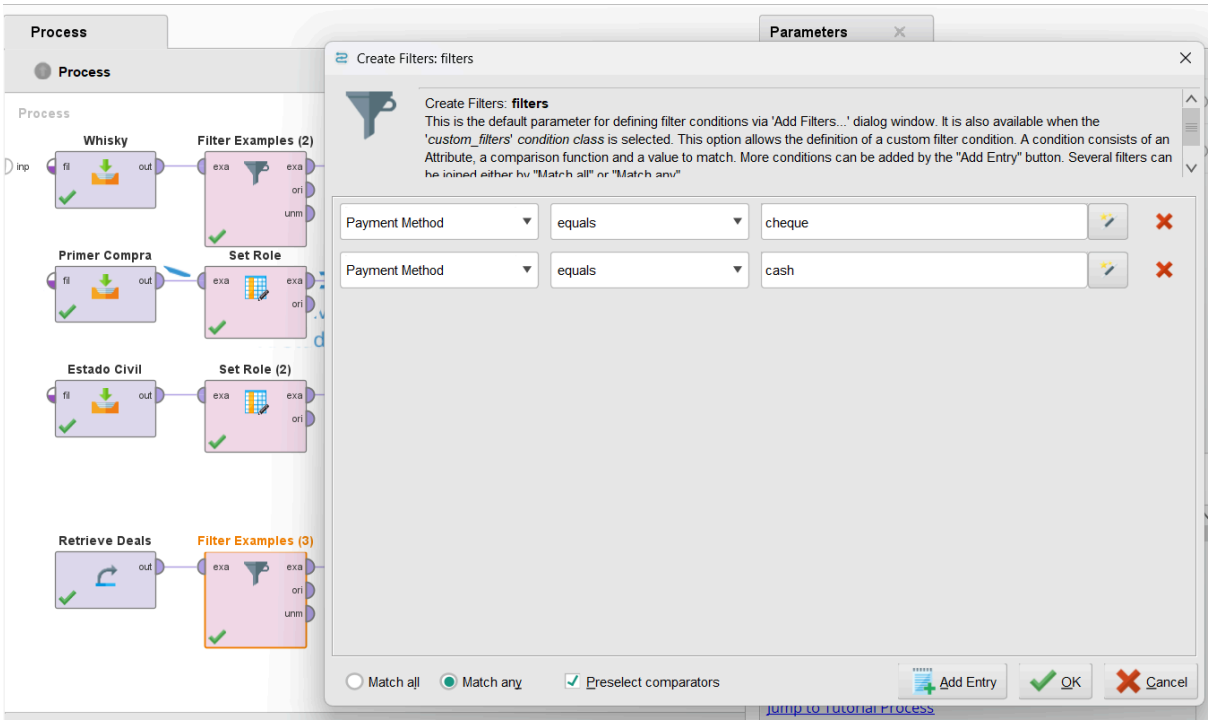
	dinero integer	casa integer	estado polynominal	sexo integer	auto integer	edad integer
1	2	1	soltero	0	1	43
2	1	1	viudo	0	1	27
3	2	0	casado	1	1	56
4	2	0	divorciado	0	1	50
5	4	1	soltero	1	1	30
6	2	1	viudo	1	1	24
7	2	1	soltero	1	1	18
8	4	1	viudo	0	1	52
9	4	1	casado	1	1	22
10	5	1	divorciado	0	0	54
11	4	0	soltero	0	0	60
12	1	0	viudo	1	1	55
13	4	1	casado	1	0	33
14	1	0	casado	1	1	36
15	4	0	casado	1	0	48
16	5	1	casado	1	0	20
17	4	0	soltero	1	1	44
18	2	1	viudo	1	0	51
19	2	1	casado	1	0	28

Usamos nuevamente el operador Set Role para colocar el estado como label.

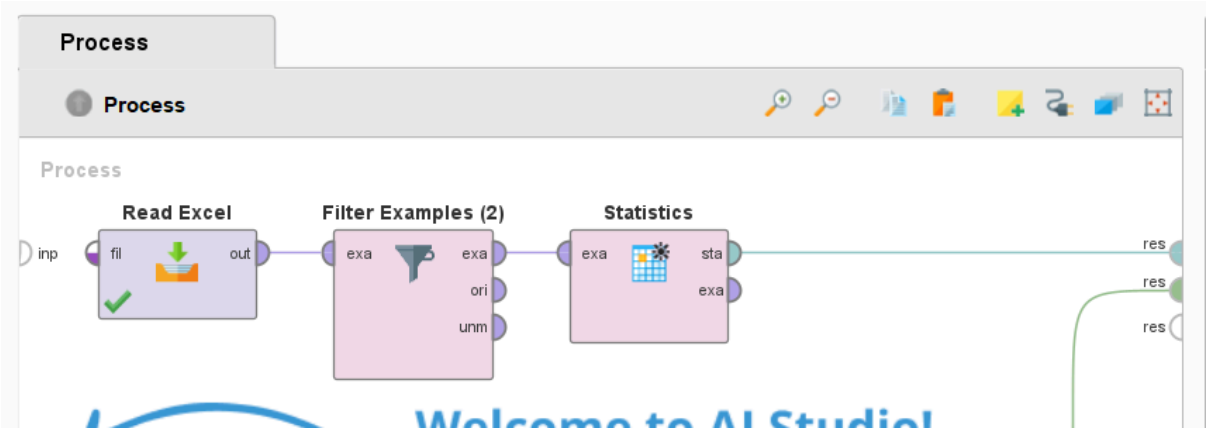


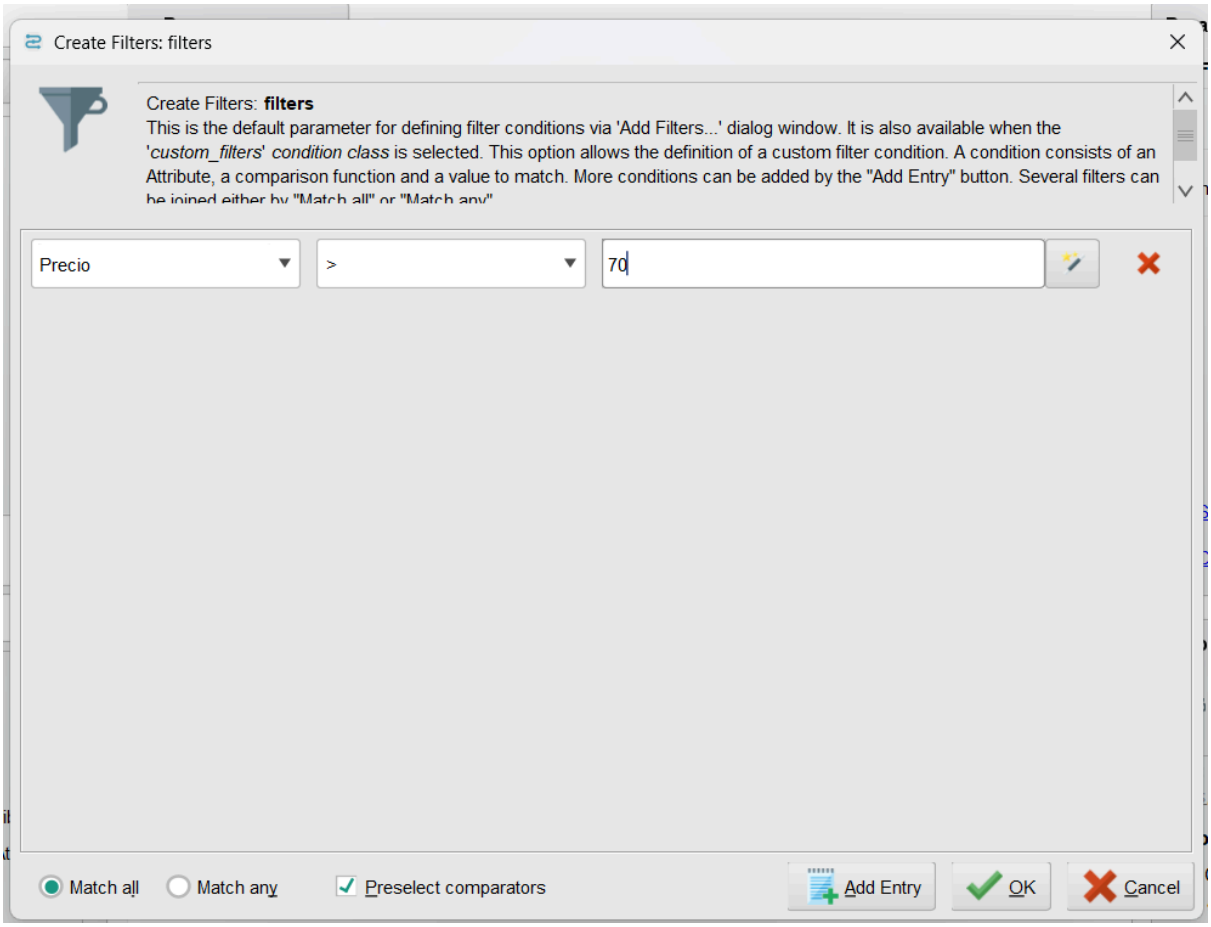


Agregamos un filtro para Deals.



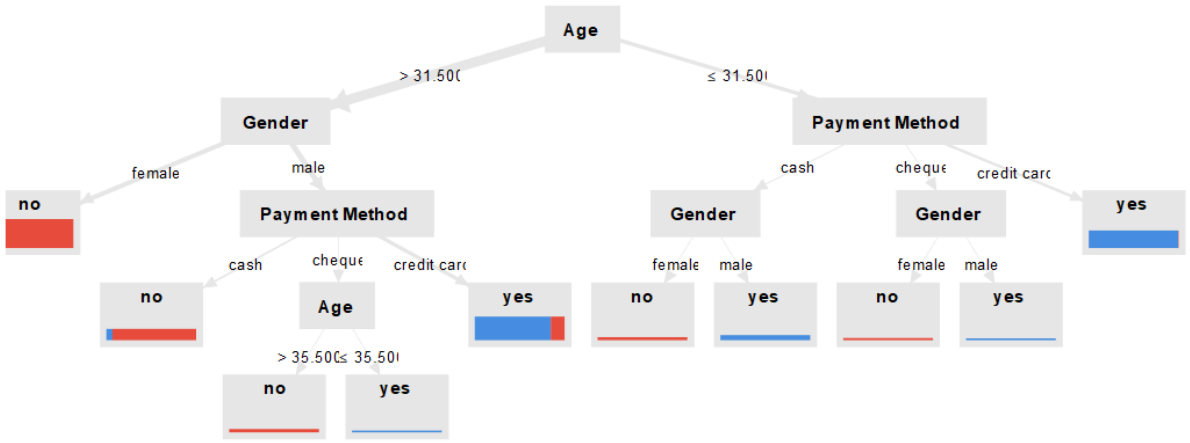
Agregamos un filtro y el operador de estadísticas en el primer recurso.

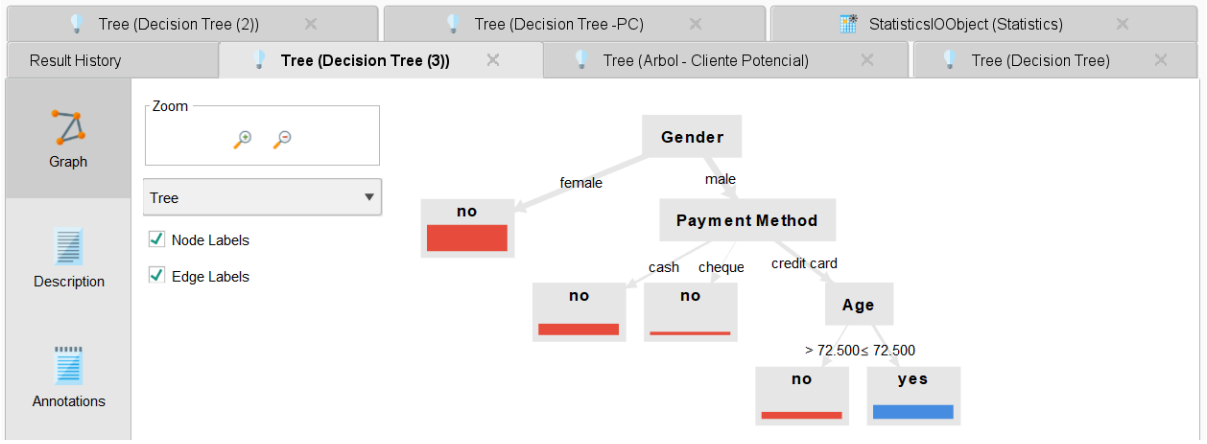


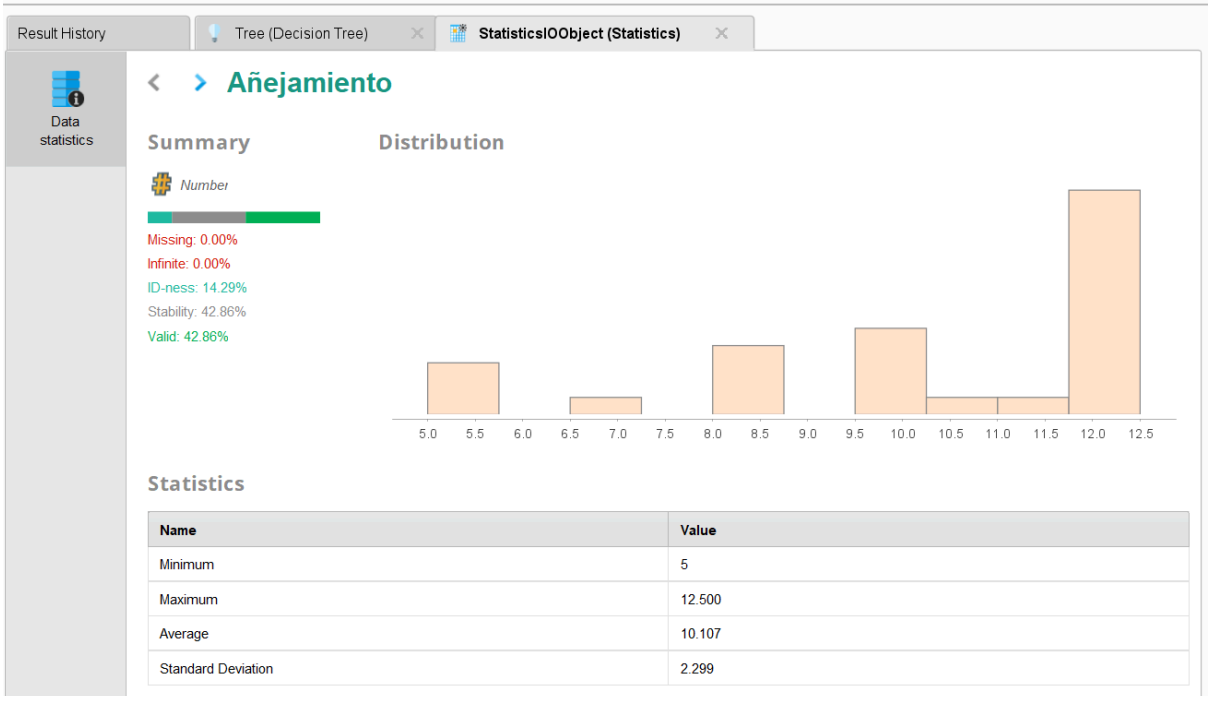


Correr el proceso y validar los resultados

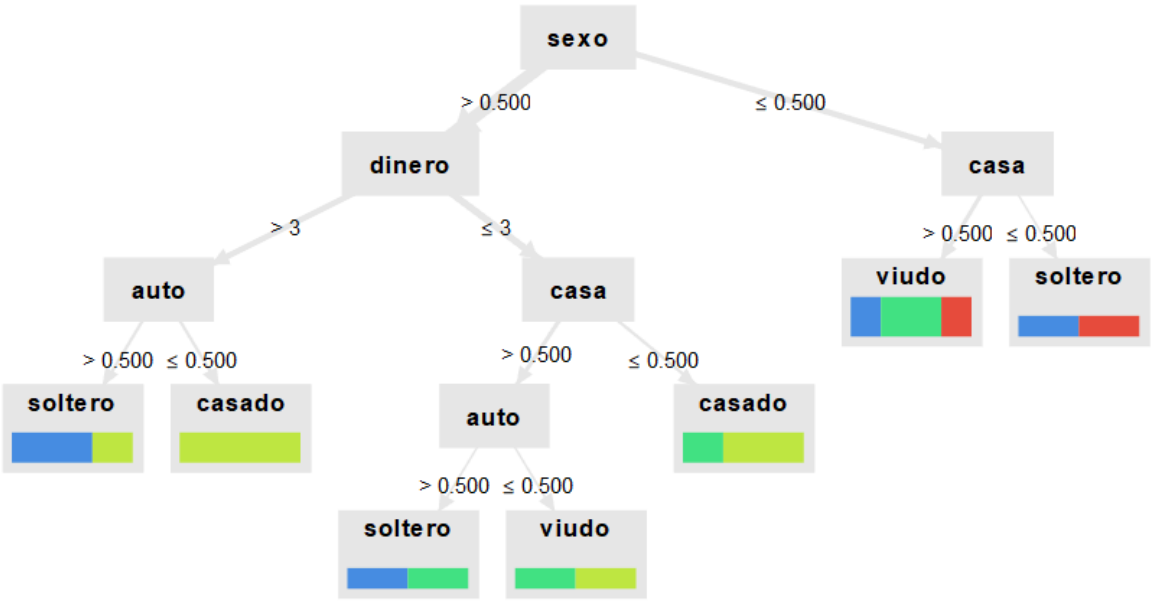
Corremos los procesos configurados para visualizar los resultados.

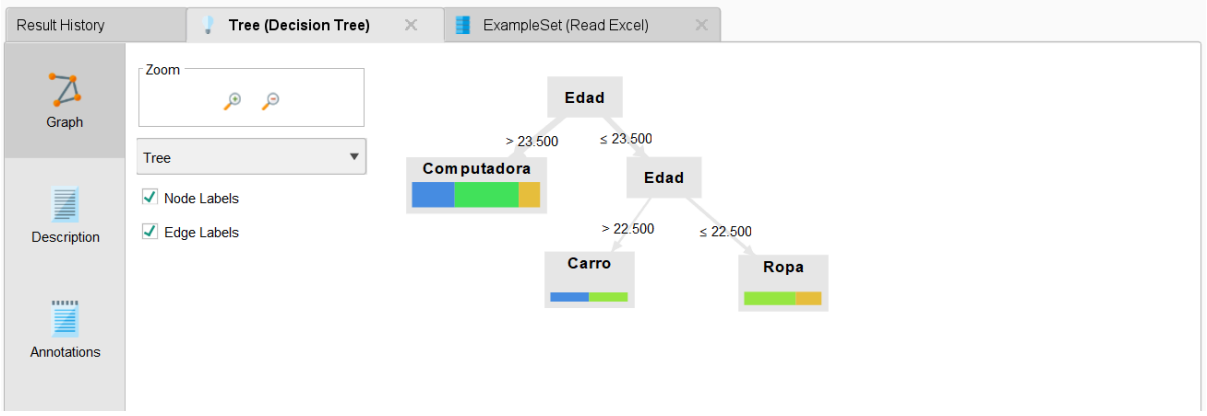


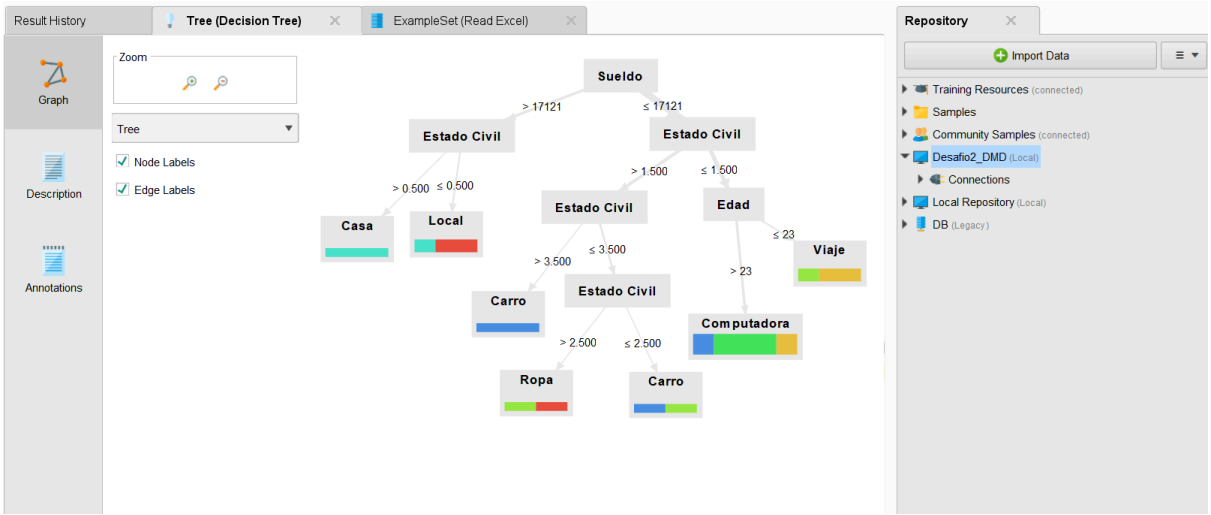




Name	Value
Minimum	5
Maximum	12.500
Average	10.107
Standard Deviation	2.299


sexo
> 0.500
≤ 0.500
dinero
casa
> 3
≤ 3
> 0.500
≤ 0.500
auto
casa
viudo
soltero
> 0.500
≤ 0.500
> 0.500
≤ 0.500
soltero
casado
auto
casado
> 0.500
≤ 0.500
soltero
viudo


Result History
Tree (Decision Tree)
ExampleSet (Read Excel)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
Edad
> 23.500
≤ 23.500
Computadora
Edad
> 22.500
≤ 22.500
Carro
Ropa


Result History
Tree (Decision Tree)
ExampleSet (Read Excel)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
Sueldo
> 17121
≤ 17121
Estado Civil
Estado Civil
> 0.500
≤ 0.500
> 1.500
≤ 1.500
Casa
Local
Estado Civil
Edad
≤ 23
> 23
Viaje
> 3.500
≤ 3.500
Carro
Estado Civil
Computadora
> 2.500
≤ 2.500
Ropa
Carro
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Desafio2_DMD (Local)
Connections
Local Repository (Local)
DB (Legacy)

Control de versiones con Git

He utilizado **Git** en GitHub para versionar el proyecto como se solicita en el Desafío.

- Recursos compartidos.
- Carpeta del repositorio en RapidMiner.
- Archivo Desafio2Completado.rmp

Enlace: GitHub

Conclusion

La aplicación de los árboles de decisión en los distintos conjuntos de datos permitió identificar patrones claros sobre el comportamiento de los clientes. A partir del análisis realizado, determinamos que el cliente con mayor probabilidad de compra corresponde a personas adultas de entre 30 y 45 años, con ingresos medios o altos y tendencia a valorar la calidad de los productos. El uso de esta técnica demuestra cómo la minería de datos facilita la toma de decisiones estratégicas para mejorar las ventas de una empresa.