

MAESTRÍA EN CIENCIA DE DATOS

**Asignatura
Inteligencia Artificial**

Trabajo final

**Robinson Gualotuña
Wilson Arroyo**

epoch

**Orellana, Ecuador
2024**

CONTENIDO

1. Descripción de la actividad.	3
2. Determine el objetivo de negocio.	3
3. Realice la gestión de datos de la data de Monterey	3
4. Responder las preguntas de gestión de datos inicial.	3
1) ¿Cuántos atributos (features) hay y cuál es su significado y tipo de datos?	4
Tenemos 10 atributos:	4
2) ¿Cuántas clases (class) hay y tipo de datos?	5
3) ¿Cuántas variedades hay en esta clase?	5
4) ¿Cuál es el número total de instancias en el dataset?	5
5) ¿Existen valores desconocidos o faltantes en algún atributo?	6
5. Analítica	6
1) Análisis unidimensional de las variables	6
2) Análisis multidimensional de las variables	7
6. Modelación	9

1. Descripción de la actividad.

Tarea en grupo: Análisis de Datos del Municipio de Monterrey en RapidMiner

Para esta actividad, se trabajará de forma grupal utilizando la fuente de datos generada por el municipio de Monterrey. Para el grupo 8, se deberá usar el año 2020. Cada grupo deberá completar las tareas indicadas a continuación:

2. Determine el objetivo de negocio.

El objetivo del negocio es entender la relación de pagos realizados a contratistas, proveedores, prestadores de servicio por honorarios pagados a profesionistas; gastos en comunicación social, representaciones, asesorías y en general todo desembolso realizado por el Municipio de Monterrey.

3. Realice la gestión de datos de la data de Monterrey

Se descargan los datos desde la ubicación:

https://portal.monterrey.gob.mx/transparencia/Oficial/Index_Proveedores_Contratistas.asp

Y se cargan los datos en la herramienta RapidMiner.

4. Responder las preguntas de gestión de datos inicial.

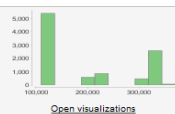


The screenshot displays the RapidMiner software interface. The main workspace shows a process diagram with a 'Read Excel' operator connected to an input port. The 'Parameters' panel on the right is open, showing the configuration for the 'Read Excel' operator. The 'excel file' parameter is set to 'I\\Monterrey_trabajo.xlsx', the 'sheet number' is '1', and the 'date format' is set to 'Enter value...'. Below the parameters, there are links for 'Show advanced parameters' and 'Change compatibility (10.5.000)'. The 'Help' panel at the bottom right provides information about the 'Read Excel' operator, including its tags and a synopsis: 'This operator reads an ExampleSet from the specified Excel file.' At the bottom of the interface, there is a message: 'Leverage the Wisdom of Crowds to get operator recommendations based on your process design!' with a button to 'Activate Wisdom of Crowds'.

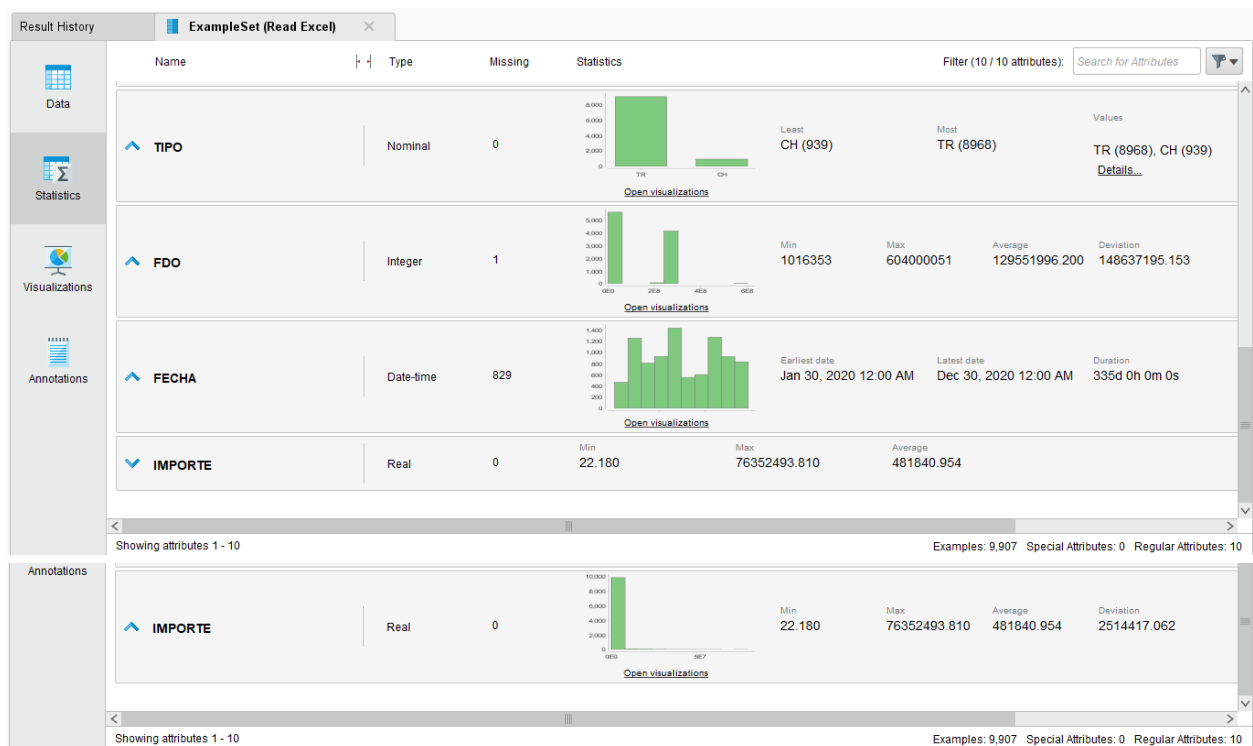
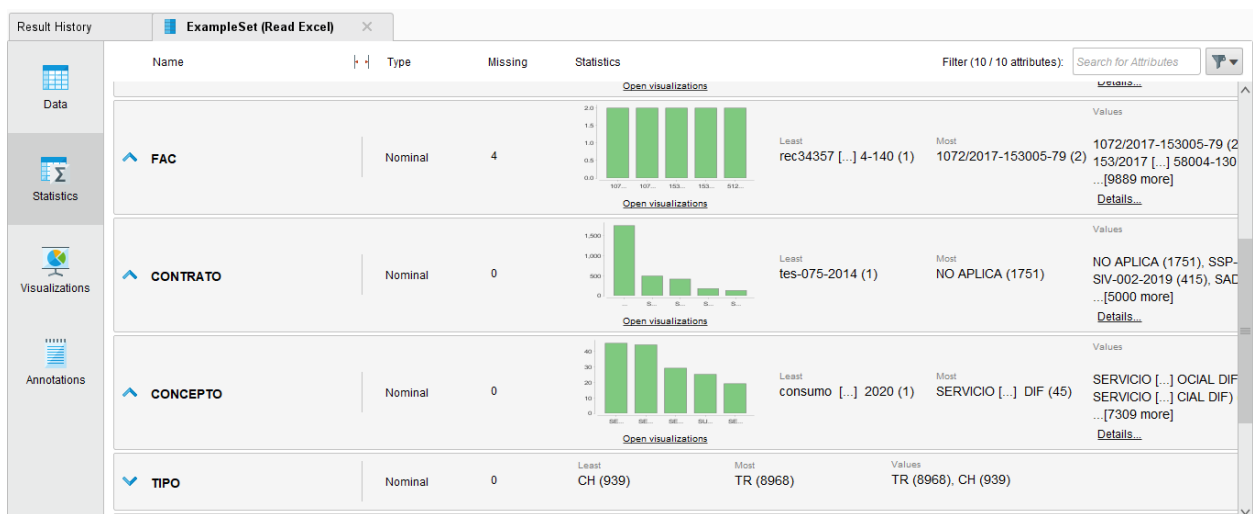
ExampleSet (Read Excel)									
Open in Turbo Prep Auto Model Interactive Analysis					Filter (9,907 / 9,907 examples): all				
Row No.	NUMBENEF	BENEFICIARIO	RFC	FAC	CONTRATO	CONCEPTO	TIPO	FDO	FEC
1	116136	SERVICIOS ...	SGM950714...	500/2016-15...	NO APLICA	CUMPLIMIEN...	CH	1102629	Jan
2	116684	INDUSTRIAS ...	ISB840628IB2	100054946-1...	SSP-189-2017	MANTENIMIE...	TR	604000040	Jan
3	116684	INDUSTRIAS ...	ISB840628IB2	100054945-1...	SSP-189-2017	MANTENIMIE...	TR	604000040	Jan
4	126020	INSTITUTO D...	IJR070509Q24	A66-153001-3	NO APLICA	PRIMERA MI...	TR	1016366	Jan
5	126224	BANCO MUL...	BMI061005N...	FS-1-2020-1...	NO APLICA	FONDO SAP...	TR	1016353	Jan
6	126224	BANCO MUL...	BMI061005N...	FS-2-2020-1...	NO APLICA	FONDO SAP...	TR	1016387	Jan
7	126287	INSTITUTO M...	IMM100301H...	A2652-15300...	NO APLICA	PRIMERA MI...	TR	1016367	Jan
8	126371	INSTITUTO M...	IMP130214DJ0	134-153001-1	NO APLICA	PRIMERA MI...	TR	1016368	Jan
9	126693	SISTEMA PA...	SDI770226674	158002-19	NO APLICA	REINTEGRO ...	TR	292000016	Jan
10	206019	S.U.T.S.M.M.	XAXX010101...	158004-4	NO APLICA	APORTACIO...	TR	1016354	Jan
11	206019	S.U.T.S.M.M.	XAXX010101...	DS010120-1...	NO APLICA	DESCUENT...	TR	1016384	Jan
12	206894	MUNICIPIO D...	MCM610101...	2020-1-A-158...	NO APLICA	PAGO DE NÓ...	TR	11000161	Jan
13	206894	MUNICIPIO D...	MCM610101...	2020-2-A-158...	NO APLICA	PAGO DE NÓ...	TR	11000162	Jan
14	206894	MUNICIPIO D...	MCM610101...	2020-2-A-158...	NO APLICA	PAGO DE NÓ...	TR	11000163	Jan

ExampleSet (9,907 examples,0 special attributes,10 regular attributes)

1) ¿Cuántos atributos (features) hay y cuál es su significado y tipo de datos?

Tenemos 10 atributos:

Result History ExampleSet (Read Excel)									
	Name	Type	Missing	Statistics					
Data				Filter (10 / 10 attributes): <input type="text" value="Search for Attributes"/>					
Statistics	NUMBENEF	Integer	0	 Min: 110036, Max: 370012, Average: 193184.928, Deviation: 91112.882					
Visualizations	BENEFICIARIO	Nominal	0	 Least: ZAPATA V [...], Most: S.I.M.E.P.R.O.D.E. (616), Values: S.I.M.E.P.R.O.D.E. (616), GARCIA Z [...], JOBERAG... [657 more]					
Annotations	RFC	Nominal	0	 Least: ZAVD830923MZ0 (1), Most: SIM870529CA0 (579), Values: SIM870529CA0 (579), IN... GAZE940509117 (280), C... [675 more]					



2) ¿Cuántas clases (class) hay y tipo de datos?

Tenemos 1, es la variable TIPO

Esta variable es Nominal

3) ¿Cuántas variedades hay en esta clase?

Tenemos 2 variedades:

TR con 8698 observaciones

CH con 939 observaciones

4) ¿Cuál es el número total de instancias en el dataset?

Tenemos 9907 registros o instancias

5) ¿Existen valores desconocidos o faltantes en algún atributo?

Si, en los siguientes atributos:

FAC: 4

FDO: 1

FEHA: 829

5. Analítica

1) Análisis unidimensional de las variables

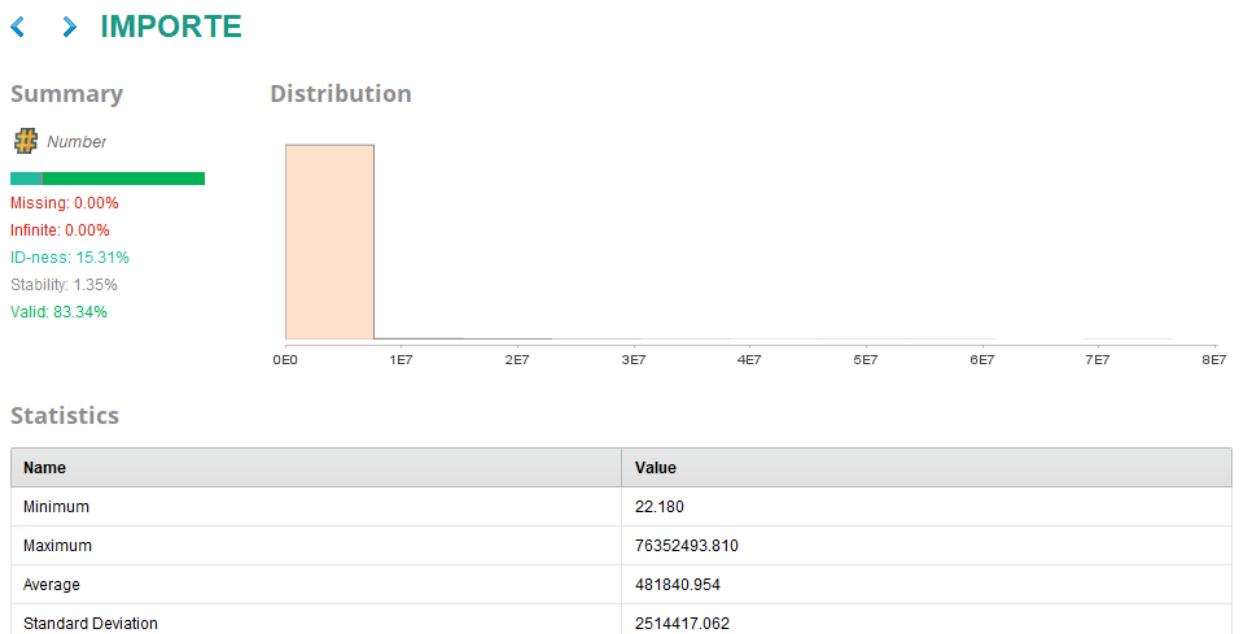
a. Determinar el conteo de la ocurrencia de una variable

Por ejemplo la variable TIPO:



b. Contabilizar el número de pagos por proveedor

c. Análisis numérico de la variable IMPORTE

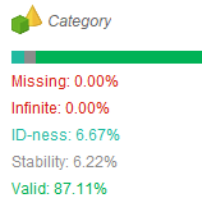


d. Realice además dos análisis adicionales de dos variables seleccionadas

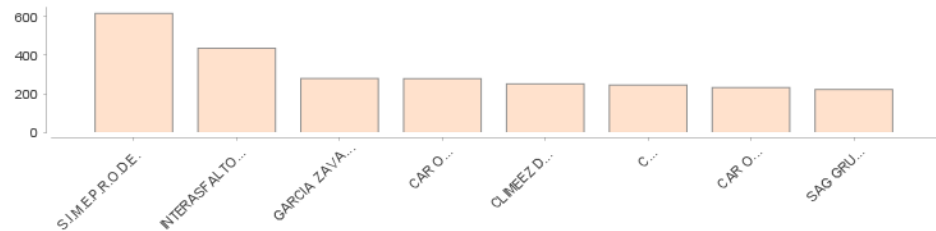
Variable BENEFICIARIO:

< > BENEFICIARIO

Summary



Top Values



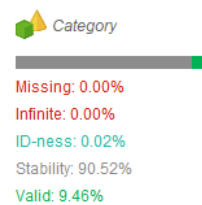
661 Distinct Values:

Value	Count	Percentage
S.I.M.E.P.R.O.D.E.	616	6.22%
INTERASFALTOS, S.A. DE C.V.	436	4.40%
GARCIA ZAVALA EDGAR ALEJANDRO JOBERAGAN	280	2.83%
CAR ONE MONTERREY, S.A. DE C.V.	279	2.82%
CLIMEEZ DEL NORTE, S.A. DE C.V.	252	2.54%
CFE SUMINISTRADOR DE SERVICIOS BASICOS	246	2.48%

Variable TIPO:

< > TIPO

Summary



Top Values

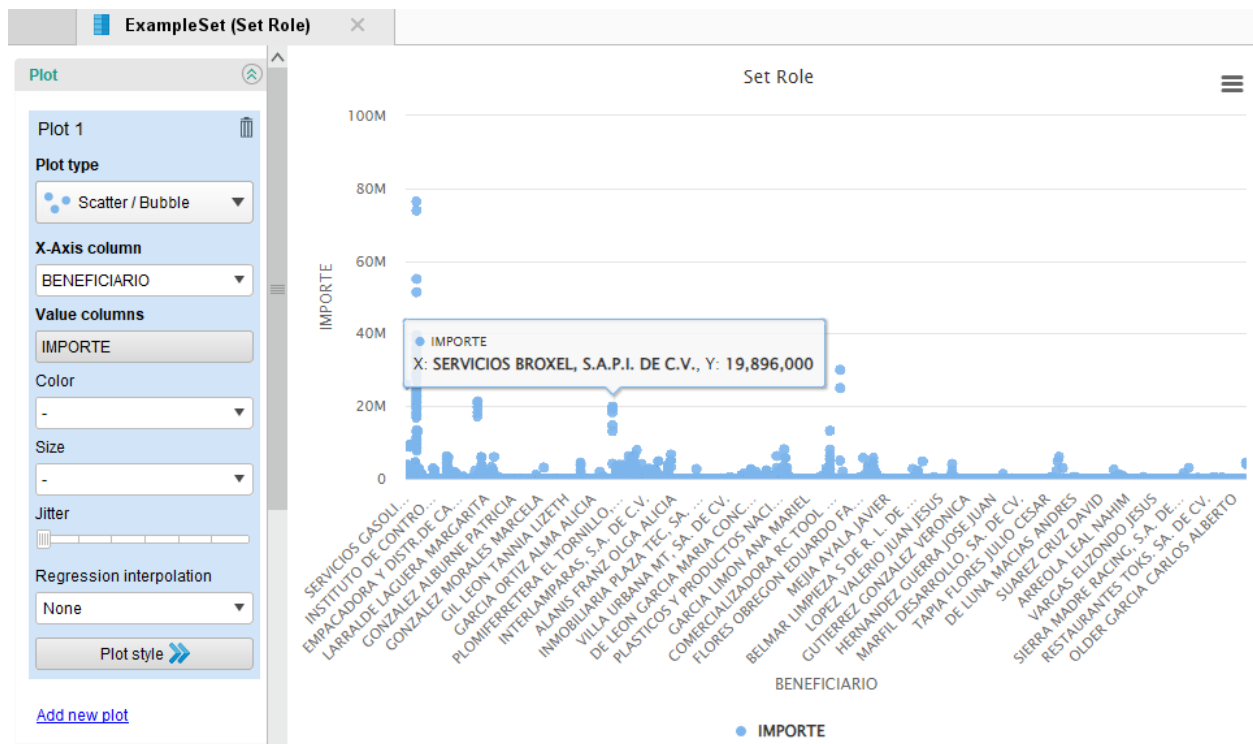


2 Distinct Values:

Value	Count	Percentage
TR	8,968	90.52%
CH	939	9.48%

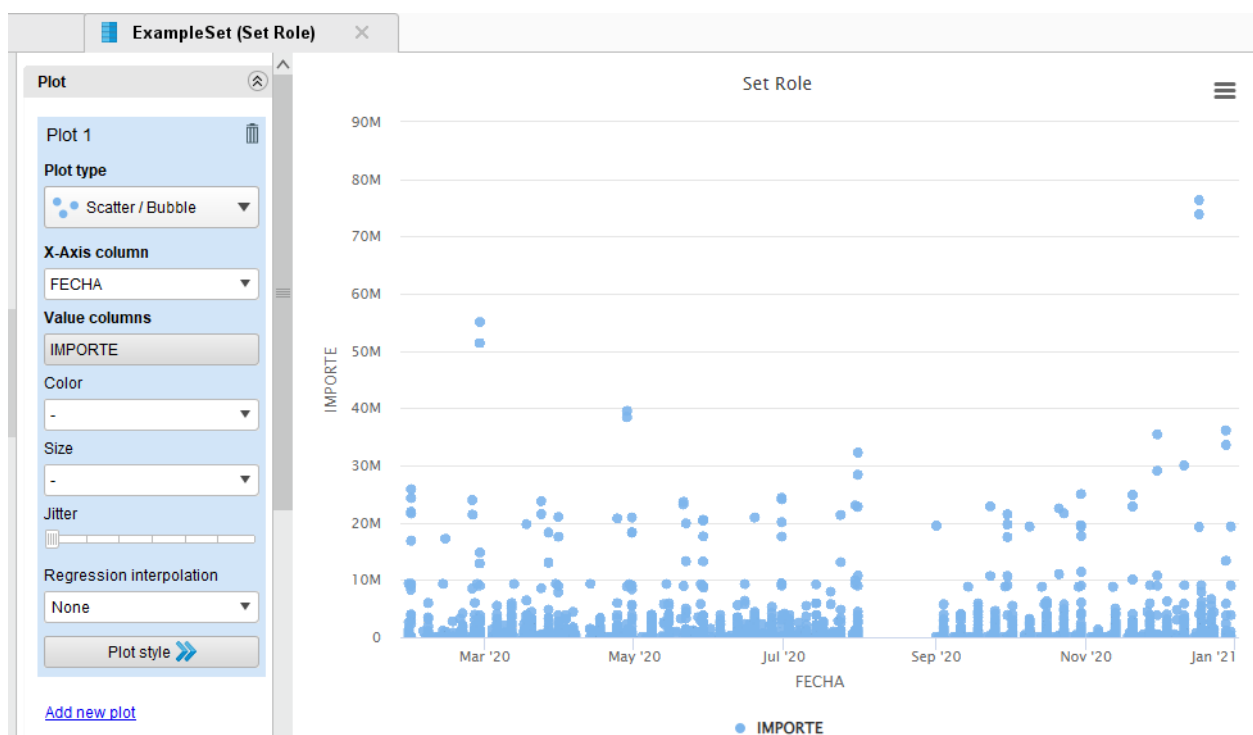
2) Análisis multidimensional de las variables

a) Identificar los proveedores o beneficiarios con mayor carga económica.

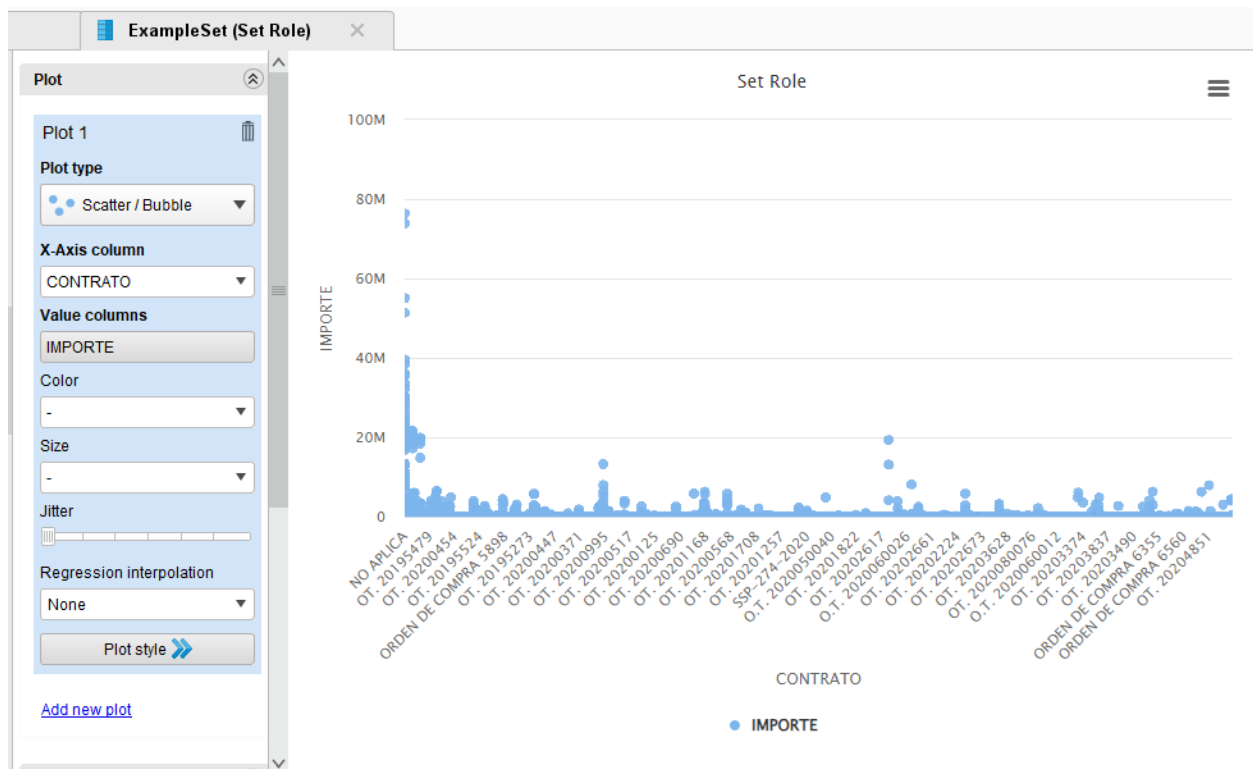


b) Determinar el gasto promedio por cada categoría.

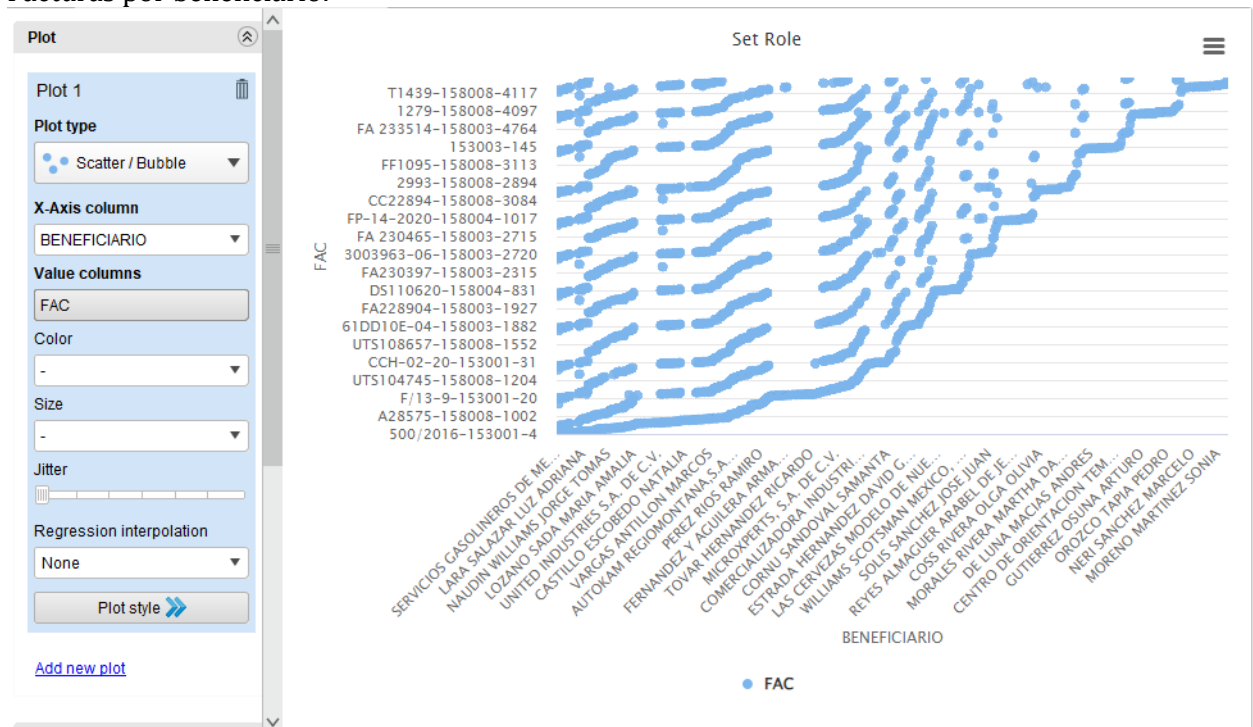
c) Determinar los gastos por fecha.



d) Realizar dos análisis multidimensionales adicionales.
Contratos con mayor Importe:



Facturas por beneficiario:



6. Modelación

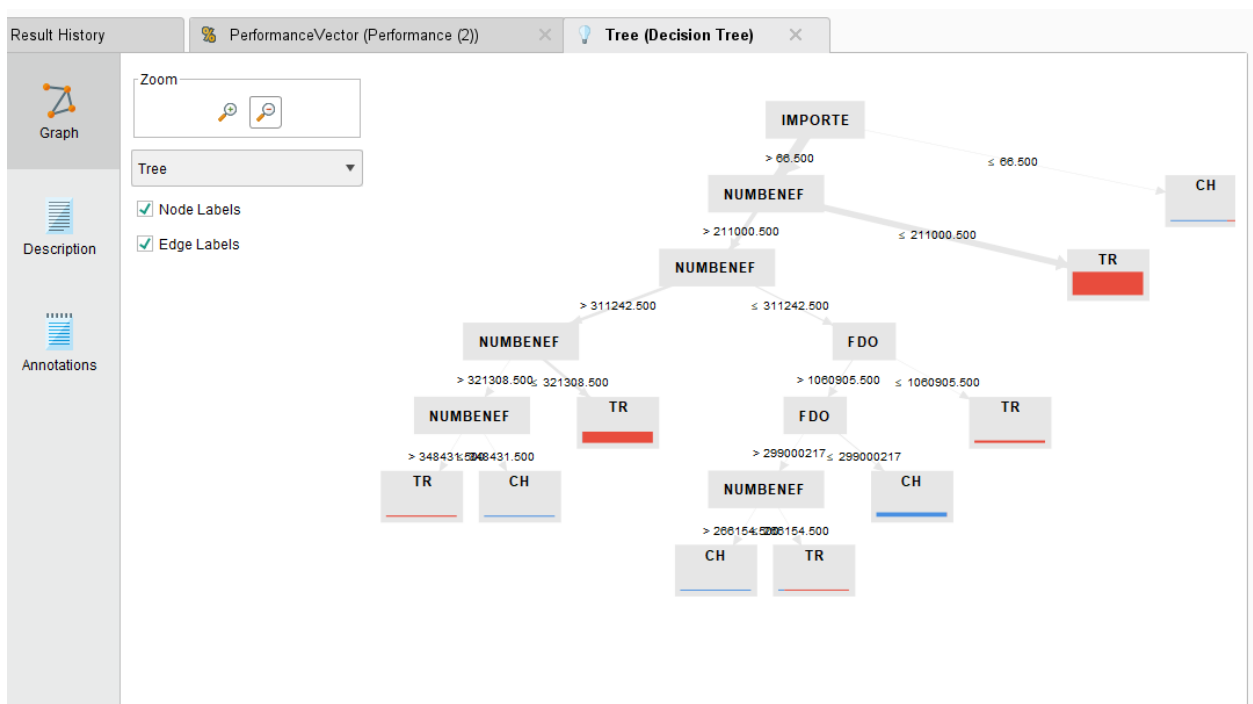
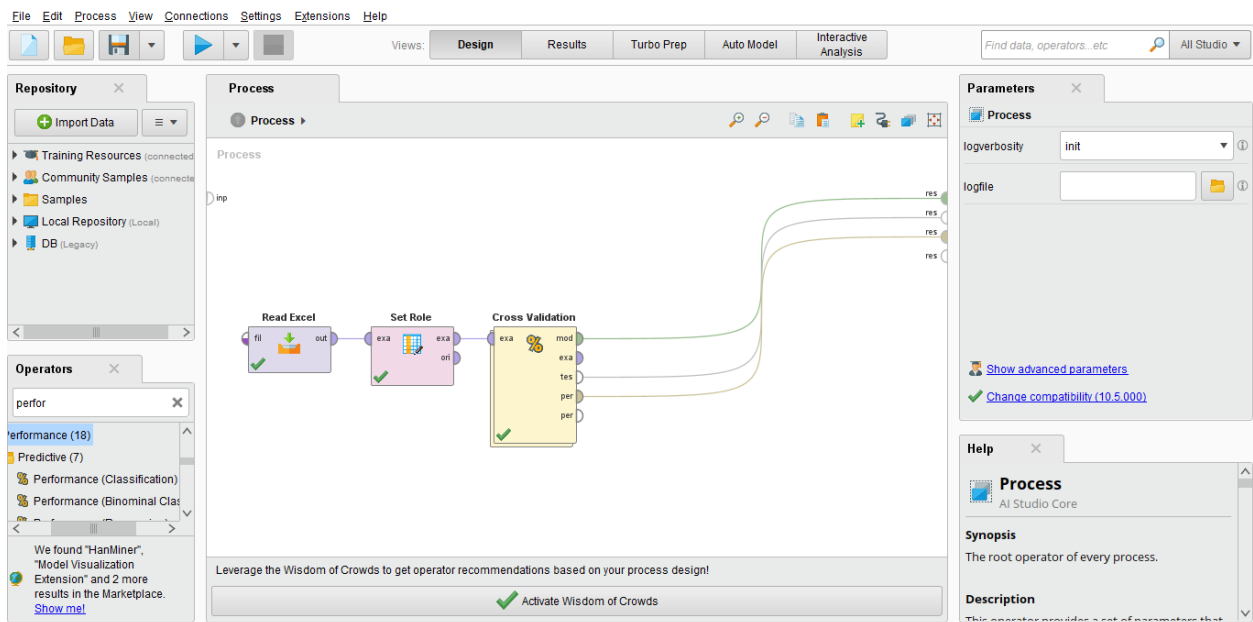
Modelo supervisado

El modelo supervisado que se va a utilizar es el árbol de clasificación

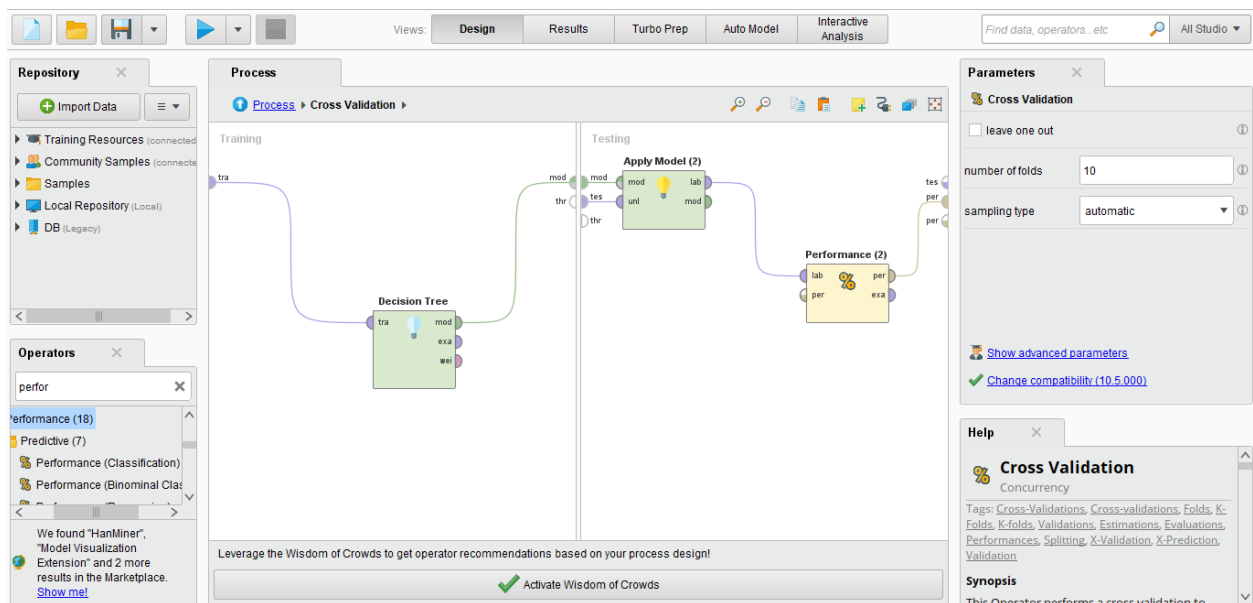
La variable objetivo es TIPO, las categorías son CH y TR

Las independientes son las otras variables presentes en el dataset.

i. Implementar un modelo supervisado



ii. Evaluar al modelo



iii. Medir el rendimiento

Result History

PerformanceVector (Performance (2))

Table View: Plot View

Criterion: accuracy

accuracy: 99.71% +/- 0.13% (micro average: 99.71%)

	true CH	true TR	class precision
pred. CH	920	10	98.92%
pred. TR	19	8958	99.79%
class recall	97.98%	99.89%	

iv. Generar y analizar la matriz de confusión para evaluar el desempeño

Se nota que el modelo tiene sobre ajuste ya que la predicción de la clase es superior al 98% en cada una de las categorías de la variable objetivo.