

Reto Salesforce Predictive Modelling

MASTERCLASS



UNIVERSITYHACK 2018[®]
DATAATHON

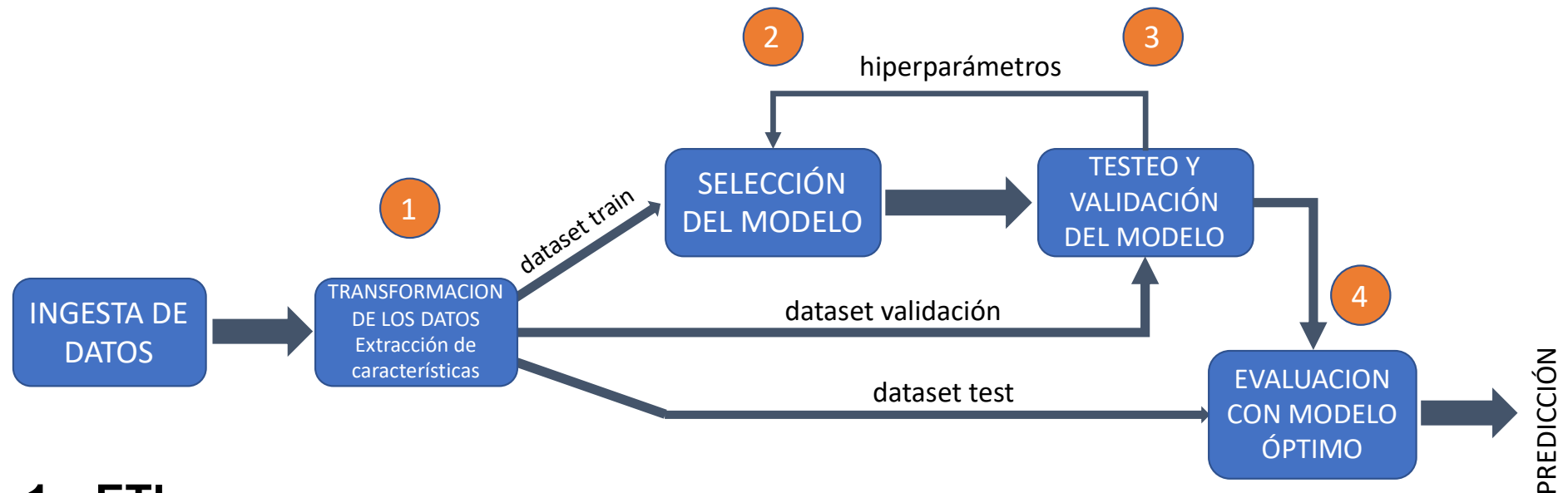
Raul Pingarrón
raul_pingarron@hotmail.es



Universidad
Europea Madrid

LAUREATE INTERNATIONAL UNIVERSITIES

Repasando las fases del Aprendizaje Automático



1. ETL

2. SELECCIÓN, TESTEO Y VALIDACIÓN DEL MODELO

- TRAIN: conjunto de datos para entrenamiento
- VALIDATION: conjunto de datos sobre los que evaluar el modelo (métricas rendimiento como RMSE, accuracy, etc)

3. EVALUACIÓN Y CLASIFICACIÓN/PREDICCIÓN

- TEST: conjunto de datos sin etiqueta (sobre los que realizar la predicción)

El Reto: Recomendaciones

1. Comprender el reto y entender los datos que nos proporcionan.

- Realizar un análisis exploratorio de los datos (examinar los datos, comprobar errores, etc).
- La fase de **ETL es muy importante**, tener los datos de manera correcta es la clave del éxito.
- Limpieza, enriquecimiento de los datos, selección/extracción de características y transformación (modifico los datos para colocarlos como mejor interese y adaptarlos al formato que necesita el algoritmo).
- **Documentarlo** incluyendo gráficas, segmentaciones, observaciones, etc.

2. Consideraciones al crear el modelo predictivo

- Modelo = Algoritmo + hiper-parámetros
- En el caso del reto será una regresión → ¿qué valor tendrá x ? siendo $x \in \mathbb{R}$
- Probar varios modelos (algoritmo e hiper-parámetros), realizar su evaluación (en base a una métrica) quedarnos con el mejor modelo para realizar la predicción.
- **Documentar todo el proceso**: qué modelos se han probado, cómo se ha validado cual(es) ha(n) sido la(s) métrica(s)

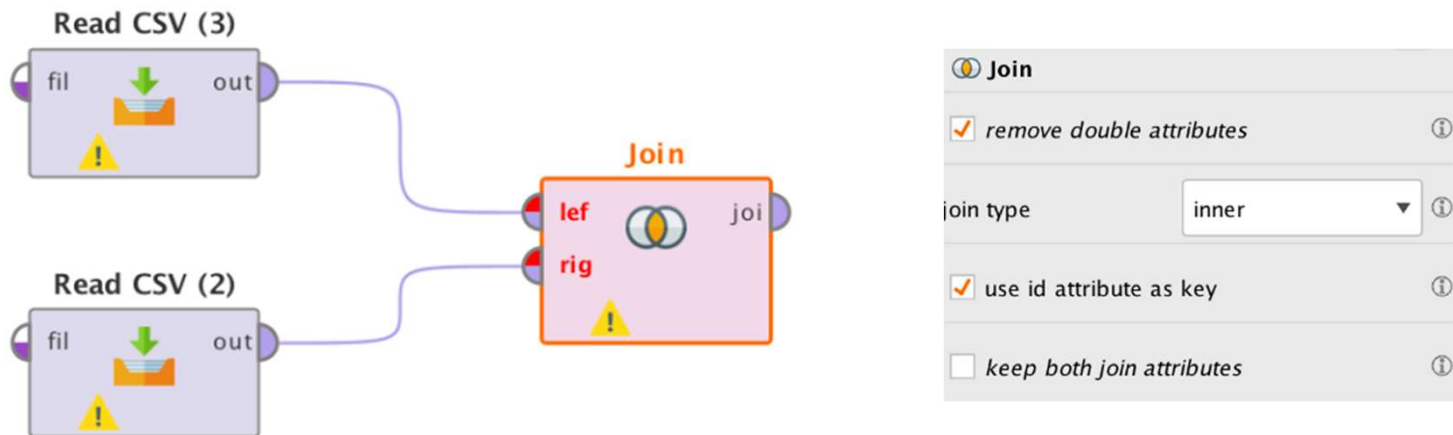
PASO 1:
EL PROCESO DE ETL



UNIVERSITYHACK 2018®
DATAATHON

Enriquecimiento de datos

- Se enriquece el dataset original con otro conjunto de datos (ej: valores del PIB por edad obtenidos de fuentes Open como el INE, etc).
- ATENCIÓN: hay que unir por un campo clave (código postal, edad) por lo que ambas tablas tiene que tener esta columna y el rol tiene que ser de tipo ID:



INNER JOIN: la unión resultante de los datasets de entrada contendrá solo aquellos ejemplos cuyo atributo clave coincida (en el ejemplo, el atributo ID)

RapidMiner

- Cargamos el dataset mediante el operador “Read CSV” y utilizamos el asistente de importación:

Data import wizard - Step 2 of 4

This wizard guides you to import your data.
Step 2: Please specify how the file should be parsed and how columns are separated.

File Reading

File Encoding:

☐ Trim Lines

☐ Skip Comments

Column Separation

☒ Comma ","

☐ Space

☐ Semicolon ";"

☐ Regular Expression

Escape Character:

☐ Use Quotes

Data import wizard - Step 4 of 4

This wizard guides you to import your data.
Step 4: RapidMiner Studio uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner Studio assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can also be defined here. Finally, you can rename attributes or deselect them entirely.

☒ Reload data ☒ Guess value types Date format:

☒ Preview uses only first 100 rows.

Socio_D...	Socio_D...	Socio_D...	Socio_D...	Socio_D...	Imp_Co...	Imp_Co...
Hombre	Socio_Demo_02	5	23038	126.0	117.5	
Hombre	Solter@	48	1	19155	123.0	12.5
Mujer	Solter@	55	4	24280	126.0	52.5
Hombre	Divorcia...	54	0	60401	39.0	25.0
Hombre	Divorcia...	57	1	43378	129.0	20.0
Mujer	Viud@	56	2	36974	115.5	92.5

Imp_Sal_08	Imp_Sal_09	Imp_Sal_10	Ind_Prod_01	Ind_Prod_02	Ind_Prod_03	Num_Oper_04	Num_Oper_05	Num_Oper_06	Poder_Adq...
integer	integer	integer	polyn...	polyn...	polyn...	integer	integer	integer	integer
attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute
6264	8527	25455	5	LE	F5	19	53	713	37925
10816	13557	3956	4	OB	F1	10	11	259	58425
14323	13007	2246	3	OB	F2	4	73	248	46125
2920	18246	29618	4	LE	F9	8	83	510	30750
13599	12531	18255	3	LE	F2	5	80	680	19475
2479	9448	28780	5	BO	F9	18	76	746	50225

☒ Ignore errors ☐ Show only errors

0 errors.

Row, Column	Error	Original value	Message
-------------	-------	----------------	---------

Previous Next Finish Cancel

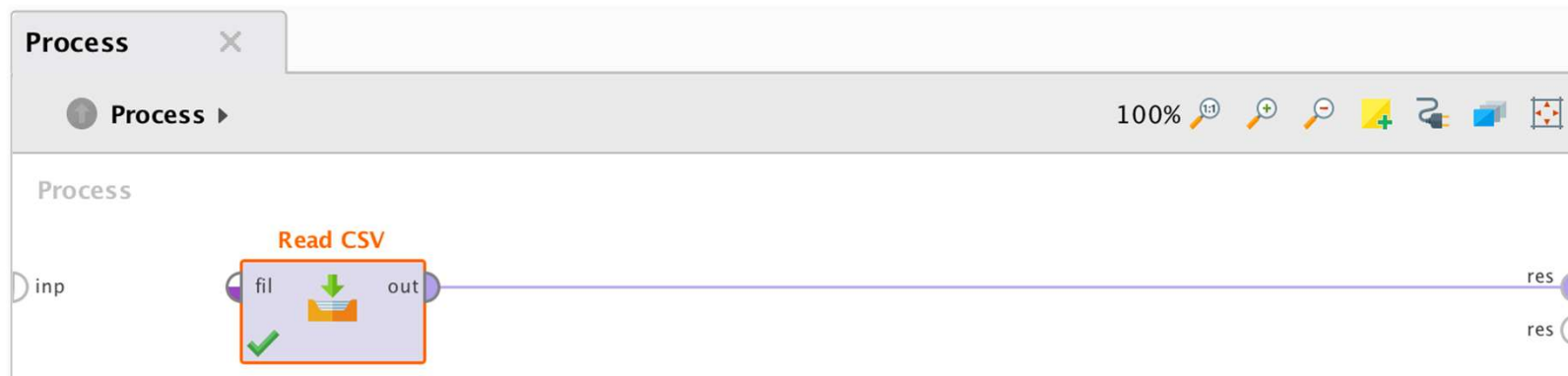
Variables o características categóricas

- Las variables categóricas suelen ser strings de dos o mas valores o estados
 - [“Hombre”, “Mujer”] [“ES”, “PT”, “UK”, “DE”, “AT”, “FR”]
- Tienen la propiedad de que **no existe una relación de orden** entre los distintos valores. La relación Hombre < Mujer no tiene sentido
- **Algunos algoritmos de regresión no admiten variables categóricas.**
- Solución: **codificarlas**
 - “One-hot-encoding” o codificación simple:
[“solter@”, “casad@”, “divorciad@”, “viud@”] → creamos 4 nuevos atributos con valores “0” o “1”

OJO: determinadas variables numéricas pueden “esconder” características categóricas: puede que la variable de situación laboral venga expresada por “0” (en paro), “1” (trabajador por cuenta propia), “2” (trabajador por cuenta ajena”, un número de socio, un código postal, etc

RapidMiner

- Conectamos el operador a la salida y verificamos los datos cargados (campos polinomiales, etc)



ExampleSet (Read CSV)						
ExampleSet (363834 examples, 1 special attribute, 31 regular attributes)				Filter (363,834 / 363,834)		
Imp_Sal_07	Imp_Sal_08	Imp_Sal_09	Imp_Sal_10	Ind_Prod_01	Ind_Prod_02	
8781	1691	12217	10646	5		F
7246	14139	5670	8461	2		F
2569	3953	7448	4190	4		F
9044	12021	8154	29810	3	LE	F

RapidMiner

- Hacemos un análisis exploratorio de los datos

¿Hay datos con errores? ¿Hay algún valor que falte?
¿Hay duplicados (usuarios repetidos)?

Result History

ExampleSet (Read CSV)

Data

Statistics

Charts

Advanced Charts

Annotations

Name	Type	Missing	Statistics		
<div>Id</div> <div>✓ ID_Customer</div>	Text	0	<div>Least</div> TR363834 (1)	<div>Most</div> TR000001 (1)	<div>Values</div> TR000001 (1), TR000002 (1), ...[363832 more]
<div>Label</div> <div>✓ Poder_Adquisitivo</div>	Real	0	<div>Min</div> 3600.960	<div>Max</div> 5040000	<div>Average</div> 16421.411
<div>✓ Imp_Cons_01</div>	Real	0	<div>Min</div> 0	<div>Max</div> 2659.110	<div>Average</div> 19.412
<div>✓ Imp_Cons_02</div>	Real	0	<div>Min</div> 0	<div>Max</div> 21097.327	<div>Average</div> 57.549
<div>✓ Imp_Cons_03</div>	Real	0	<div>Min</div> 0	<div>Max</div> 31080	<div>Average</div> 241.984
<div>✓ Imp_Cons_04</div>	Real	0	<div>Min</div> 0	<div>Max</div> 2070.837	<div>Average</div> 23.950
<div>✓ Imp_Cons_05</div>	Real	0	<div>Min</div> 0	<div>Max</div> 11975	<div>Average</div> 9.141

RapidMiner

- Visualización de datos

¿Cuáles son los rangos prevalentes en las características demográficas? ¿Qué relaciones hay entre las distintas características?

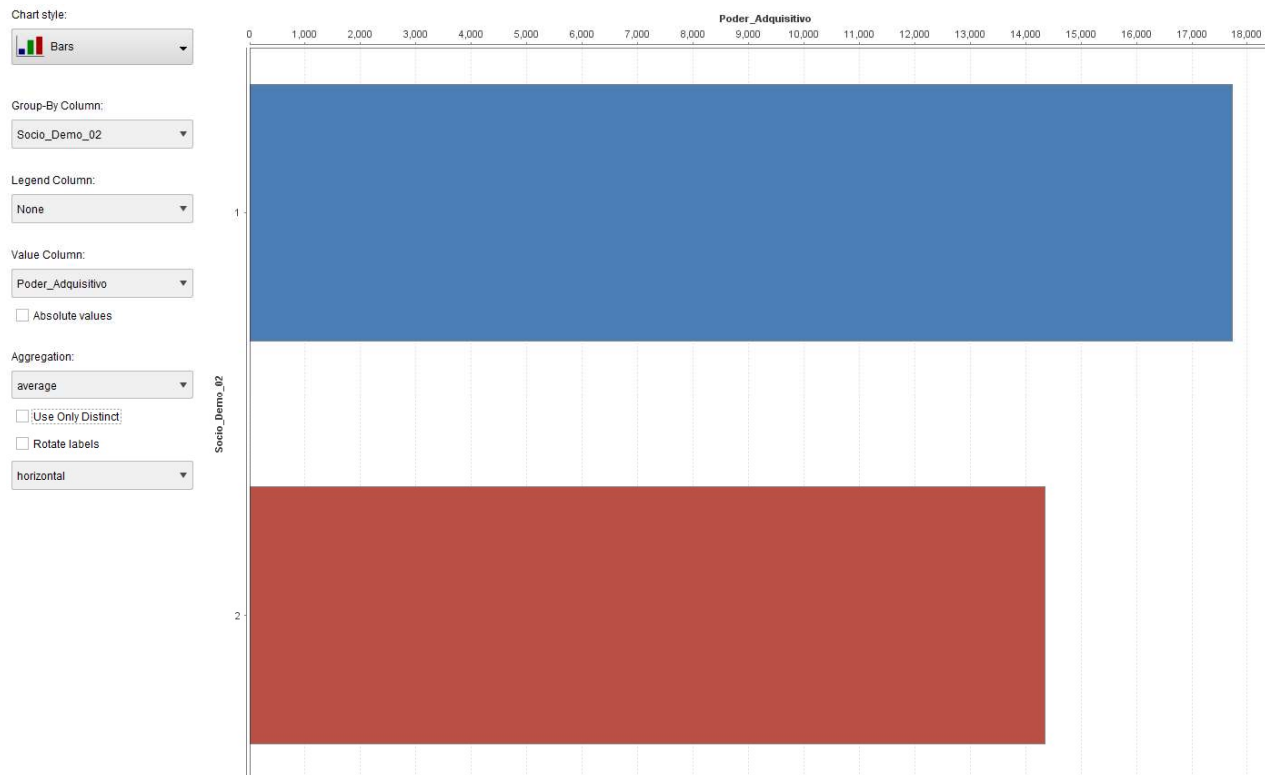
En el reto tenemos que analizar esto en detalle para hacer nuestra segmentación y explicarlo muy bien en el trabajo



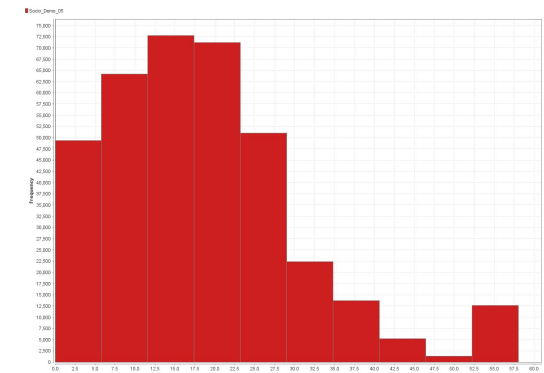
RapidMiner

- Visualización de datos

Podemos hacer correlaciones entre variables en Advanced Charts...edad e ingresos, etc

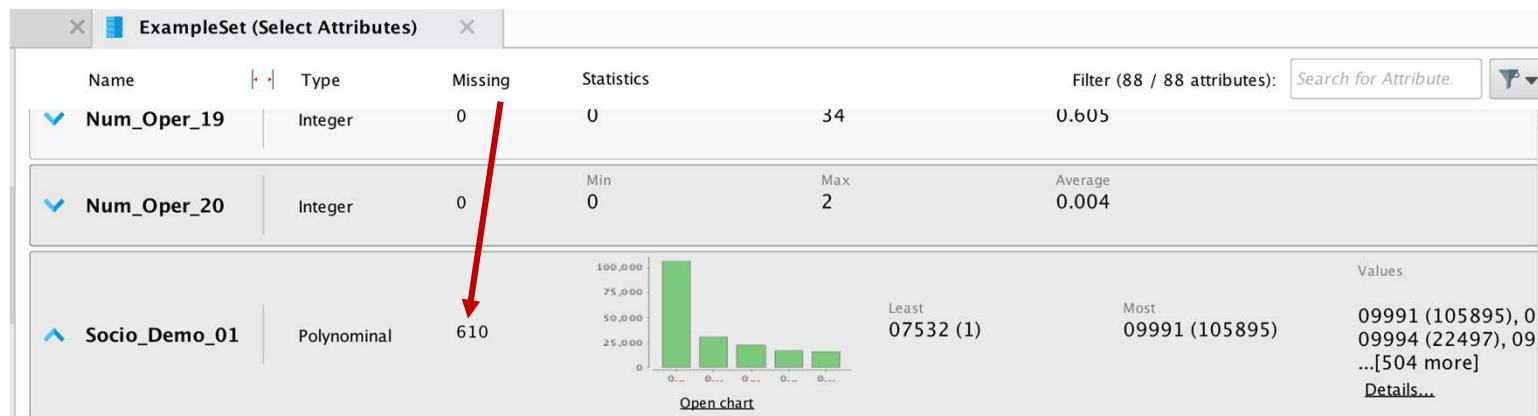


Media poder adquisitivo por sexo



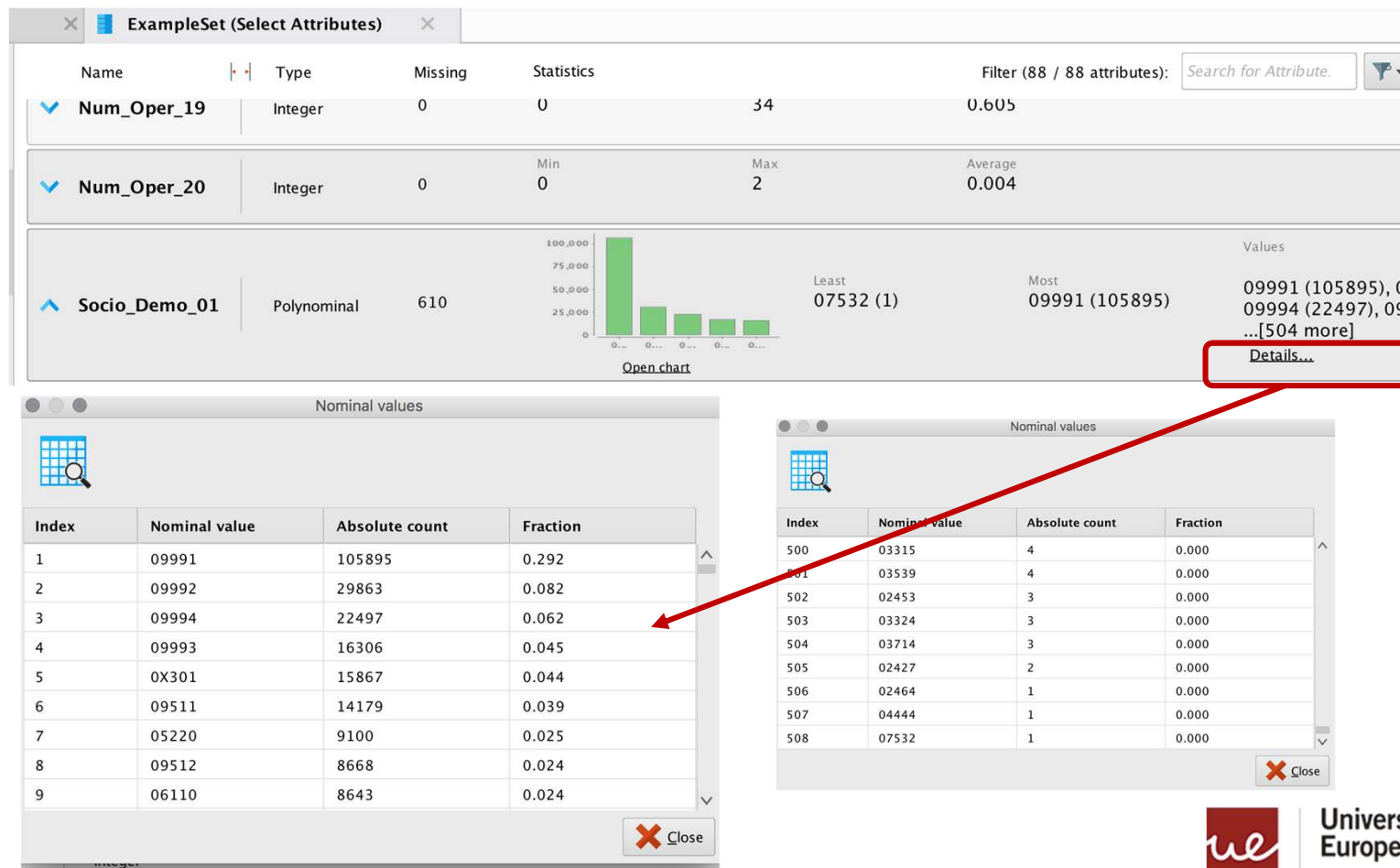
RapidMiner

- Missing values: ¿Qué hacer con los que faltan?
 - Si es menos de un 5% podemos eliminarlos (filter out)
 - Podemos sustituirlos por una palabra/etiqueta (ej: “missing”) si es nominal, o por el valor promedio si es numérico, etc.



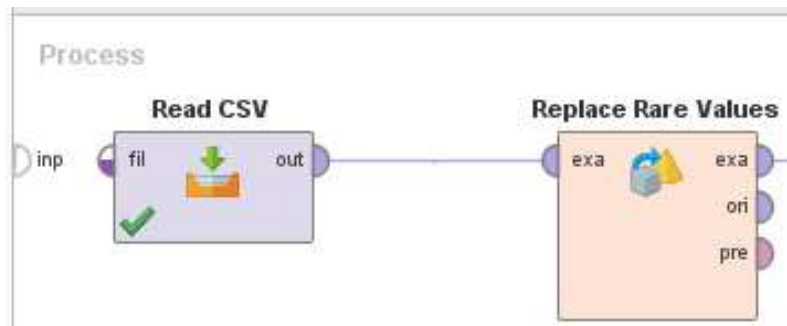
Variables polinómicas con elevada distribución

- Cuando un atributo nominal contiene multitud de posibles valores
 - Donde **muchos de esos valores son poco frecuentes**.
 - Al aplicar "one-hot-encoding" pasamos a inyectar cientos de nuevos atributos
 - El algoritmo no es capaz de aprender ninguna regla sobre los poco frecuentes



Variables polinómicas con elevada distribución

- SOLUCIÓN: Operador “**Replace Rare Values**”
 - Instalar extensión “Operator Toolbox” (v 0.9 en RM 8.1)



Replace Rare Values

☐ create view

attribute filter type: single

attribute: Socio_Demo_01

☐ invert selection

☐ include special attributes

☐ use relative threshold

threshold: 3142

replacement value: Otro

☒ replace if unknown

Variables polinómicas con elevada distribución

- RESULTADO:

- Nos quedamos con el 70% de los valores
- Hemos podado los ausentes
- Al aplicar "one-hot-encoding" pasaremos a inyectar 14 nuevos atributos


Nominal values			
Index	Nominal value	Absolute count ↓	Fraction
1	09991	105895	0.292
2	09992	29863	0.082
3	09994	22497	0.062
4	09993	16306	0.045
5	0X301	15867	0.044
6	09511	14179	0.039
7	05220	9100	0.025
8	09512	8668	0.024
9	06110	8643	0.024
10	06120	5727	0.016
11	07121	4782	0.013
12	04500	4550	0.013
13	05120	3966	0.011
14	08432	3143	0.009


Name	Type	Missing
▼ Nam_Open_20	Integer	
✓ Socio_Demo_01	Polynomial	0

Variables polinómicas con elevada distribución

OJO:

- Al aplicar "Replace Rare Values", si no hacemos nada mas ocurrirá esto:

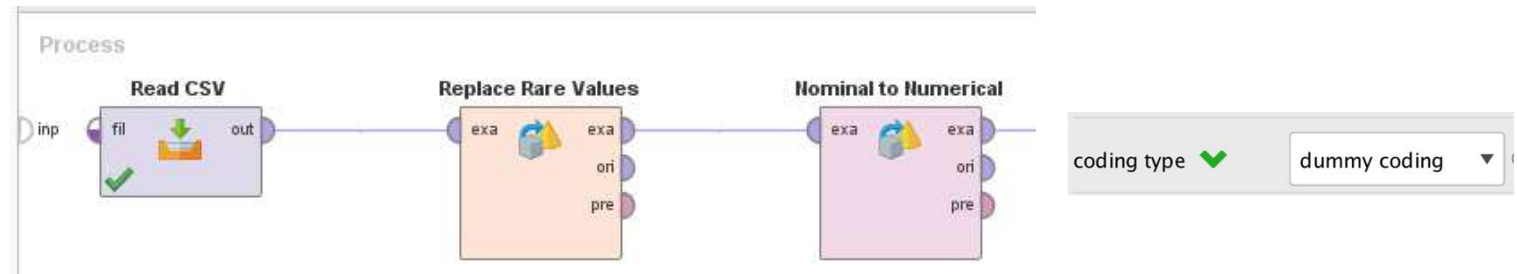
 Nominal values



Index	Nominal value	Absolute count	Fraction
1	Otro	110648	0.304
2	09991	105895	0.291
3	09992	29863	0.082
4	09994	22497	0.062
5	09993	16306	0.045
6	0X301	15867	0.044
7	09511	14179	0.039
8	05220	9100	0.025
9	09512	8668	0.024
10	06110	8643	0.024
11	06120	5727	0.016
12	07121	4782	0.013
13	04500	4550	0.013
14	05120	3966	0.011
15	08432	3143	0.009
16	00011	0	0
17	00012	0	0
18	00020	0	0
19	01111	0	0
20	01112	0	0
21	01113	0	0

Variables polinómicas con elevada distribución

- Y por tanto, al aplicar “one-hot-encoding” con el operador “Nominal to Numerical”:



ExampleSet (Nominal to Numerical)

ExampleSet (363834 examples, 2 special attributes, 641 regular attributes)

Row No.	ID_Customer	Poder_Adqu...	Ind_Prod_0...	Ind_Prod_0
2	TR000002	37497.492	1	0



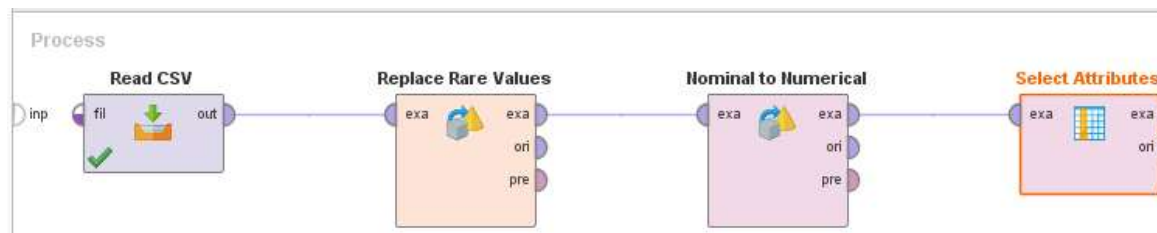
¡ Acarreamos atributos innecesarios !

ExampleSet (Nominal to Numerical)

Name	Type	Missing	Statistics
✓ Socio_Demo_01_05894	Integer	0	Min 0, Max 0, Average 0
✓ Socio_Demo_01_05910	Integer	0	Min 0, Max 0, Average 0
✓ Socio_Demo_01_04309	Integer	0	Min 0, Max 0, Average 0
✓ Socio_Demo_01_07510	Integer	0	Min 0, Max 0, Average 0
✓ Socio_Demo_01_05931	Integer	0	Min 0, Max 0, Average 0

Variables polinómicas con elevada distribución

- ¿Qué hacemos con estos atributos innecesarios?
 - Los eliminamos filtrándolos



Select Attributes

attribute filter type ☒ numeric_value_filter

numeric condition

☒ invert selection

☐ include special attributes

ExampleSet (363834 examples, 2 special attributes, 146 regular attributes)

Row No.	ID_Customer	Poder_Adquisitivo	Ind_Prod_0...
1	TR000001	19709.915	1
2	TR000002	37407.402	1

Reducción de la dimensionalidad (para nota)

- Algunos algoritmos de ML no son muy efectivos con datasets de alta dimensionalidad (ruido, OOM, etc)
- Solución: Eliminamos características irrelevantes que no aportan nada o características redundantes que dicen lo mismo.

Algunos operadores en RapidMiner:

- Remove Correlated Attributes
- Select Weights by Correlation
- Select Weights by PCA
- Select Weights by SVM
- Wrapper Optimize selection (forward selection)
- Wrapper Backward elimination
- PCA, SVD

**PASO 2:
SELECCIÓN Y VALIDACIÓN
DEL MODELO**



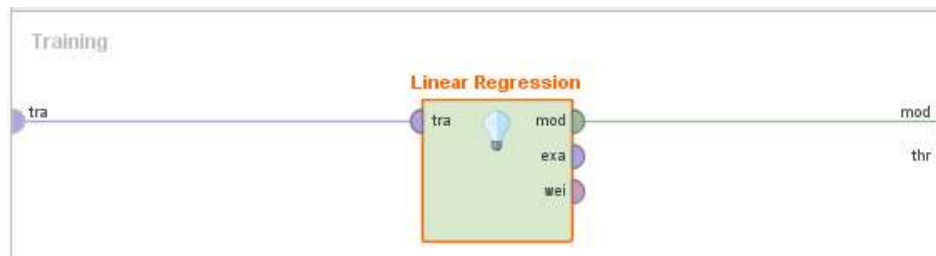
UNIVERSITYHACK 2018®
DATAATHON

Selección del Modelo

Algunos algoritmos de regresión en RapidMiner:

- Regresión lineal
- Random Forest
- Gradient Boosted Trees
- SVM
- Neural Net

Ejemplo de modelo en RapidMiner:

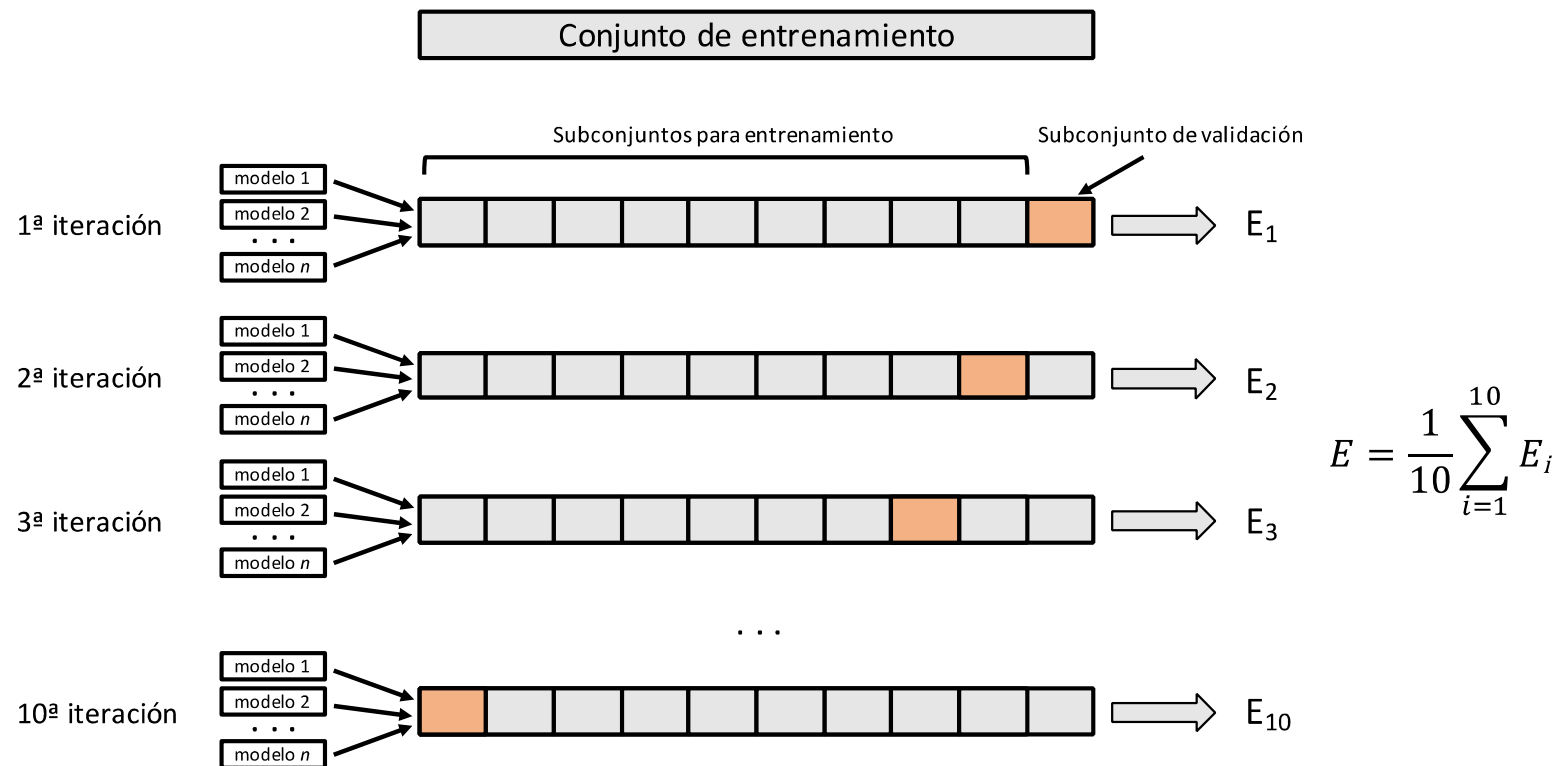


The image shows the 'Parameters' window for the 'Linear Regression' operator. It includes a dropdown for 'feature selection' set to 'none', a bar chart for '75% of users kept "M5 prime":' with options 'none' (13%), 'M5 prime (default)' (75%), 'greedy' (7%), 'T-Test' (3%), and 'Iterative T-Test' (2%), checkboxes for 'eliminate colinear features' and 'use bias', and input fields for 'min tolerance' (0.05) and 'ridge' (1.0E-8).

Parameter	Value
feature selection	none
75% of users kept "M5 prime":	
none	13%
M5 prime (default)	75%
greedy	7%
T-Test	3%
Iterative T-Test	2%
eliminate colinear features	<input checked="" type="checkbox"/>
min tolerance	0.05
use bias	<input checked="" type="checkbox"/>
ridge	1.0E-8

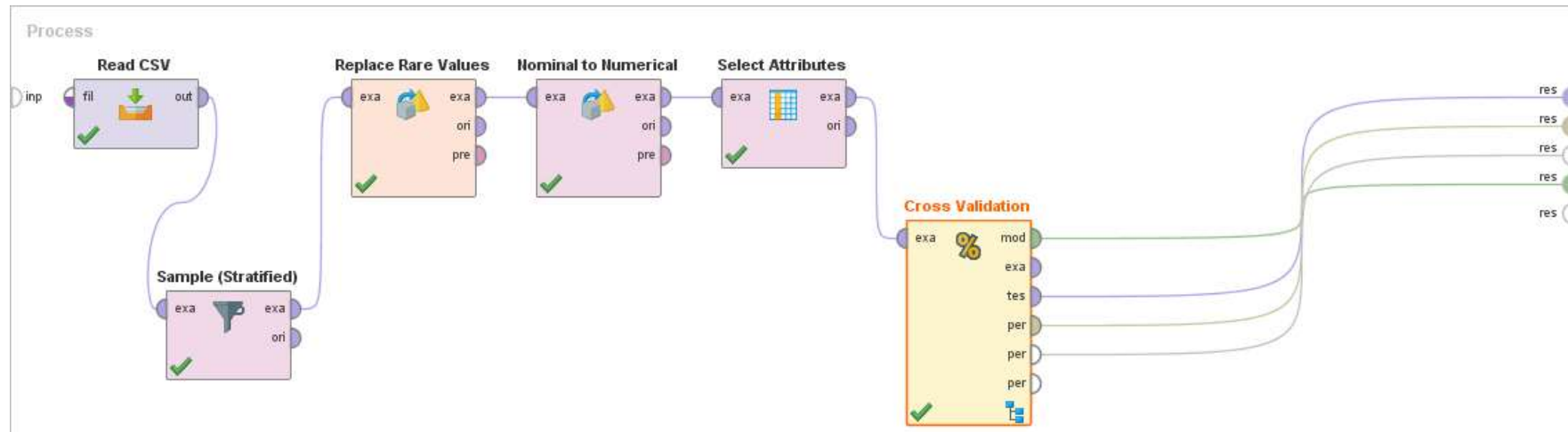
Validación del Modelo

VALIDACIÓN CRUZADA



Validación del Modelo

Validación Cruzada en RapidMiner



Parameters ✕

Cross Validation

☐ split on batch attribute

☐ leave one out

number of folds

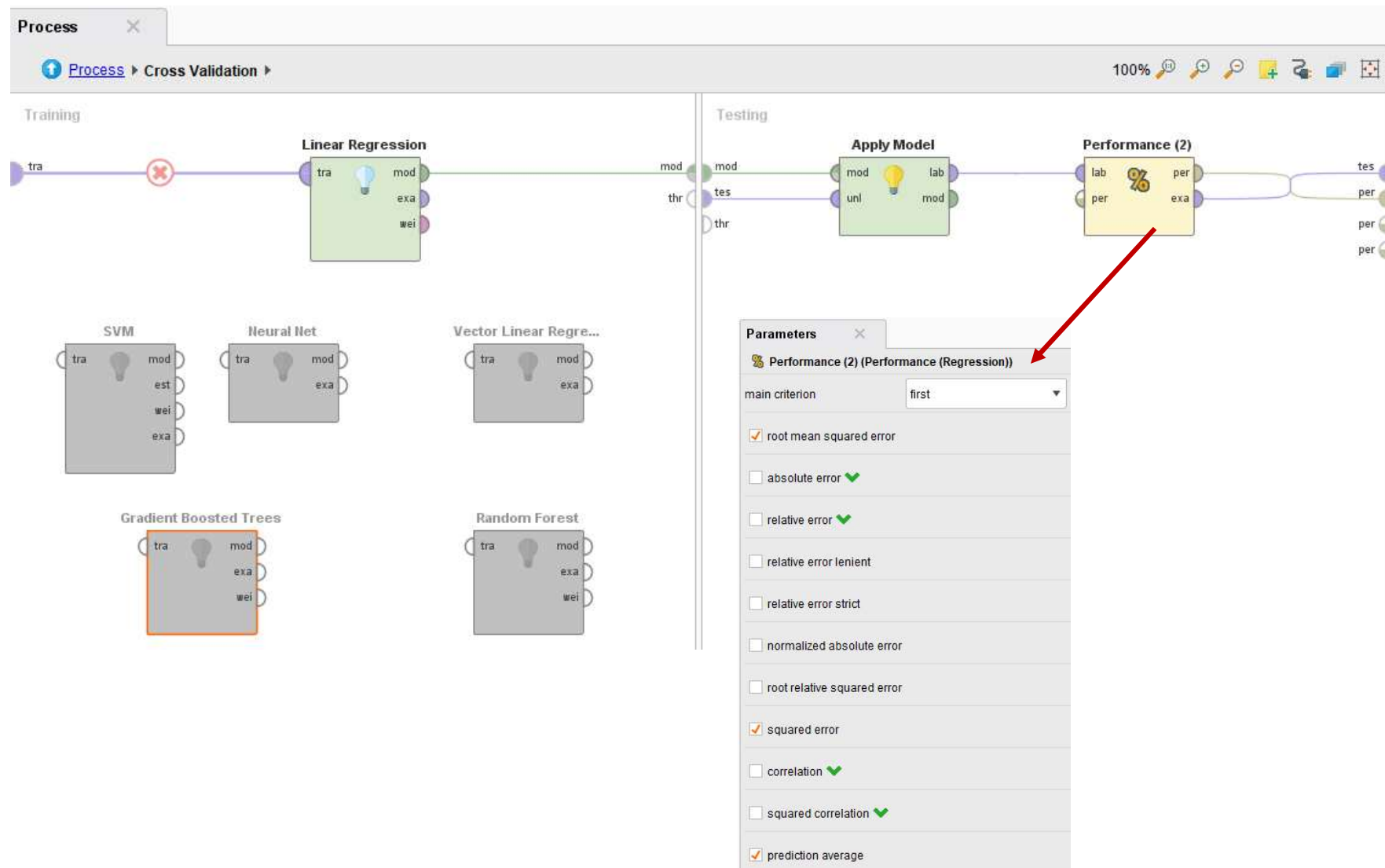
sampling type ☒ automatic

☐ use local random seed

☒ enable parallel execution

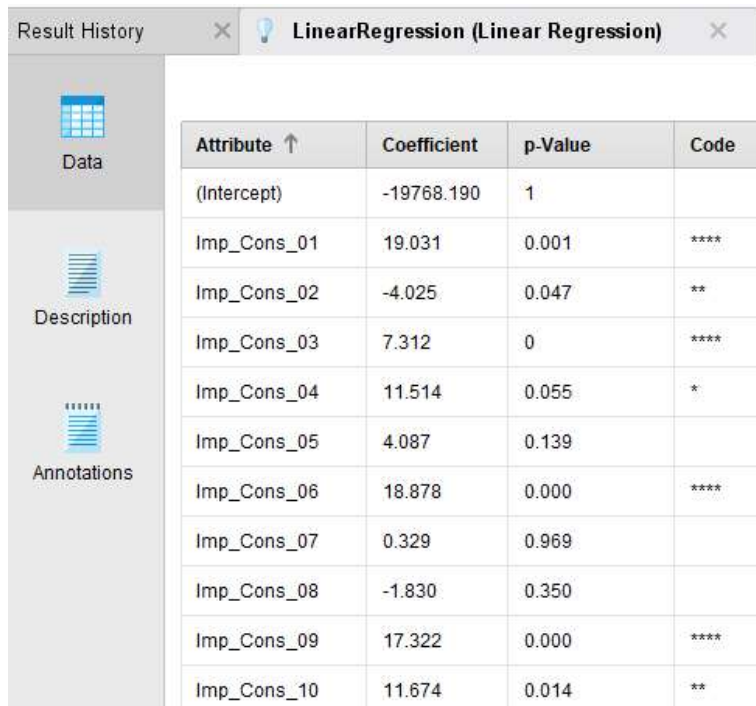
Validación del Modelo

Validación Cruzada en RapidMiner (dentro del wrapper)



Validación del Modelo

Resultados de nuestro modelo



Attribute ↑	Coefficient	p-Value	Code
(Intercept)	-19768.190	1	
Imp_Cons_01	19.031	0.001	****
Imp_Cons_02	-4.025	0.047	**
Imp_Cons_03	7.312	0	****
Imp_Cons_04	11.514	0.055	*
Imp_Cons_05	4.087	0.139	
Imp_Cons_06	18.878	0.000	****
Imp_Cons_07	0.329	0.969	
Imp_Cons_08	-1.830	0.350	
Imp_Cons_09	17.322	0.000	****
Imp_Cons_10	11.674	0.014	**

El signo del **coeficiente** indica la influencia del atributo sobre la etiqueta :

- Num_Oper_18 tiene una influencia positiva sobre el poder adquisitivo.
- A mayor valor Socio_Demo_05 mayor influencia negativa (menor será el poder adquisitivo)

El valor del predictor ha de estar dentro del umbral comprendido entre 0 y 0.5 de lo contrario no será una característica significativa:

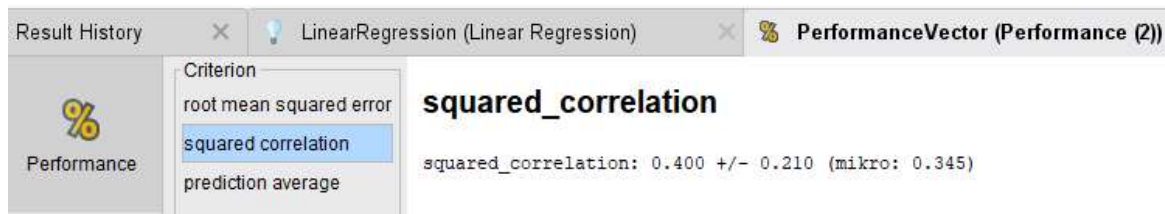
- Imp_Cons_07 es un pésimo predictor
- Las estrellas es la puntuación que da RapidMiner a las variables en función de lo significativas que son.
- Podríamos prescindir de Imp_Cons_07 y otras...

Validación del Modelo

Resultados de nuestro modelo



Mi modelo predice el poder adquisitivo con un rango de +/- 18.506€



El modelo explica el 40% de la varianza del poder adquisitivo

Result History | ExampleSet (Cross Validation)

ExampleSet (3638 examples, 3 special attributes, 125 regular attributes)

Row No.	ID_Customer	Poder_Adquisitivo	prediction(Poder_Adquisitivo)
1	TR000436	21271.630	37987.348
2	TR001167	6176.140	8071.885
3	TR002144	8414.140	17251.597
4	TR004094	14773.150	9648.070
5	TR005837	17049.520	11957.272
6	TR006060	29579.730	25788.005
7	TR006061	8696.160	8990.911
8	TR007603	10688.120	12768.577

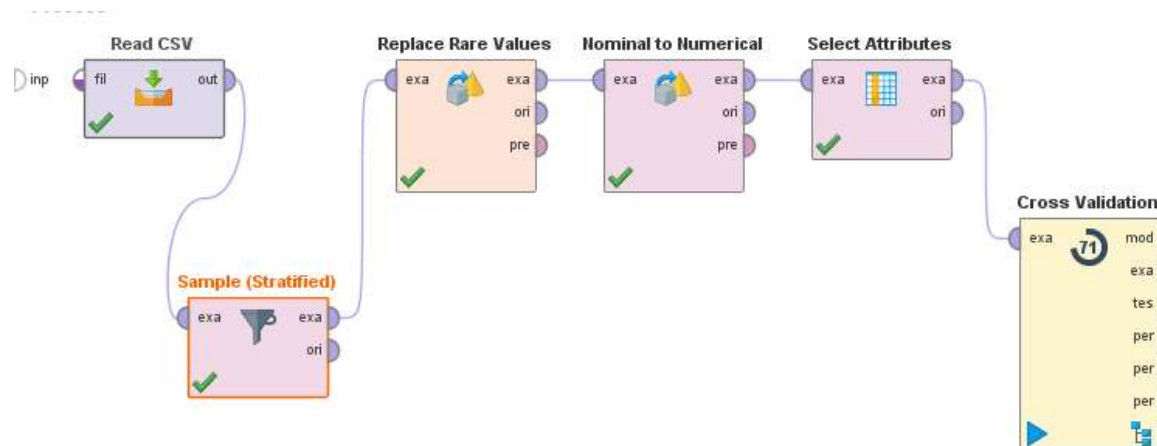


Validación del Modelo

Recomendación:

- Probar varios modelos
- Sacar las métricas de su rendimiento
- Compararlas y quedarse con el mejor modelo
- **Documentar** esta parte e incluir gráficas o valores.

NOTA: Puedes acelerar el proceso tomando solo un % del dataset gracias al operador "Sample":



PASO 3: EVALUACIÓN FINAL Y PREDICCIÓN

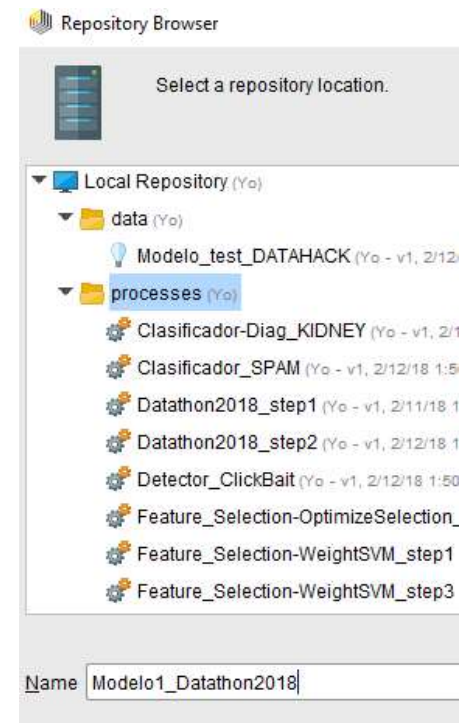
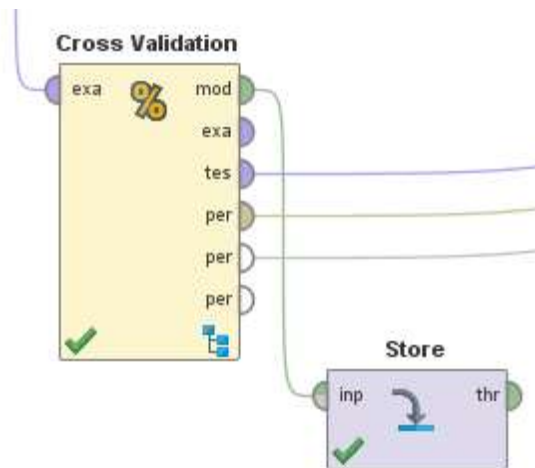


UNIVERSITYHACK 2018®
DATAATHON

Evaluación final y predicción

¿Cómo llevamos nuestro modelo a “producción”?

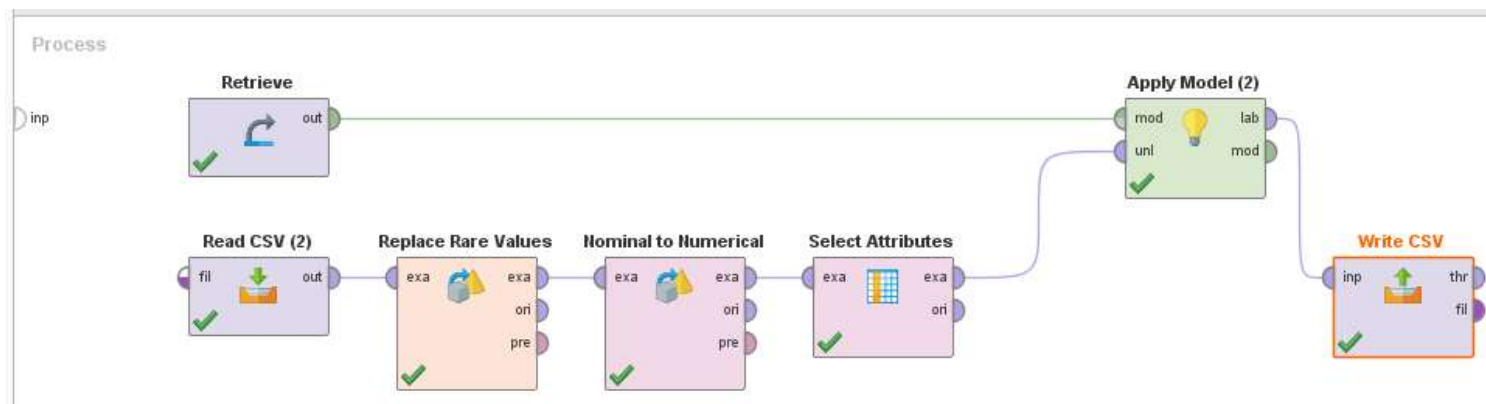
- Lo suyo es salvarlo para posteriores usos (entrenamiento online e inferencia/predicción online, etc.)



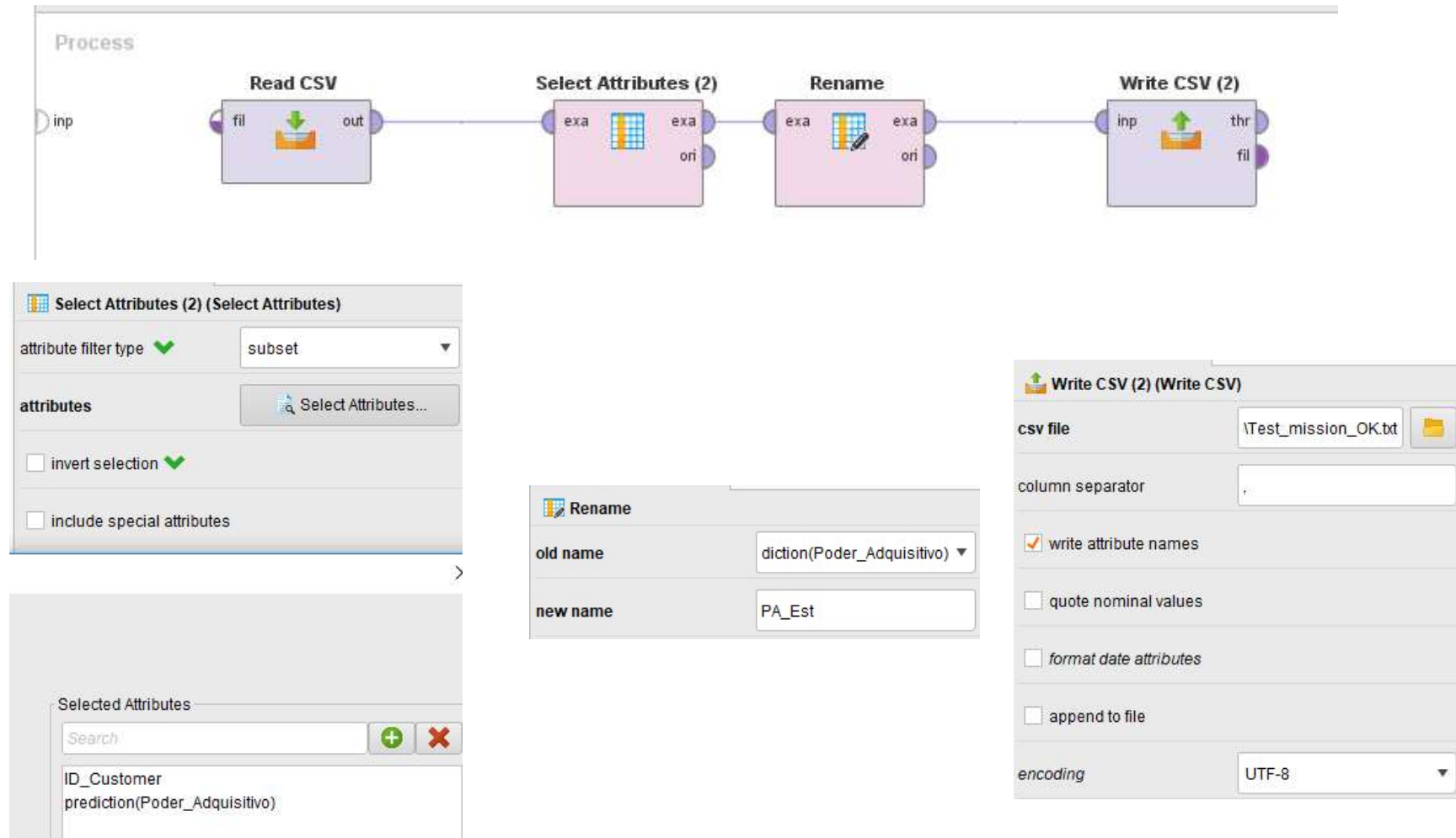
Evaluación final y predicción

¿Cómo llevamos nuestro modelo a “producción”?

- Salvamos el mejor modelo del paso anterior para posteriores usos (entrenamiento online e inferencia/predicción online, etc).
- Cargamos el modelo.
- Cargamos el dataset de TEST y lo transformamos igual que con el de entrenamiento.
- Salvamos los resultados (dataset transformado y su predicción) a un CSV



Entrega del dataset con la predicción



¡¡SUERTE!!



UNIVERSITYHACK 2018®
DATA Δ THON

Raul Pingarrón
raul_pingarron@hotmail.es



Universidad
Europea Madrid
LAUREATE INTERNATIONAL UNIVERSITIES